The role of common genetic variants for predicting the modulation of cardiovascular outcomes

Submitted by

Charli Elizabeth Harlow

To the University of Exeter

as a thesis for the degree of

Doctor of Philosophy in Medical Studies

in February 2022

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Charli Harlow

Signature: ……… ………….

Abstract

Attrition is a major issue in the drug development process with 79% of clinical failures due to safety and efficacy concerns. Genetic research can provide supporting evidence of a clear causal relationship between the drug target and disease or reveal unintended effects through associations with non-relevant phenotypes informing on potential drug safety. However, due to the underlying genetic architecture, it is often unclear which gene or variant in the loci identified through genetic analyses is driving the association. Due to recent advancements in CRISPR-Cas9 gene-editing, it is now possible to relatively easily perform whole gene knock-out studies and single base-edits to validate genetic findings of the most likely causal variant and gene. Utilising a combination of genetic approaches and functional studies can provide supporting evidence of the therapeutic profile and potential effects of drug therapies and improve our overall understanding of biological pathways and disease mechanisms.

The primary aim of this thesis is to provide genetic data to support the ongoing clinical development of hypoxia-inducible factor (HIF)-prolyl hydroxylase inhibitors (PHIs) for treating anaemia of chronic kidney disease (CKD). Genome-wide association studies (GWAS) were used to identify genetic variants lying within or nearby genes encoding the drug target (prolyl hydroxylase [PHD] enzymes). These identified variants were used in Mendelian Randomisation analysis and phenome-wide association studies to genetically mirror the pharmaceutical effects of PHIs and investigate cardiovascular safety. Functional validation studies were employed to functionally validate a genetic variant for use as a proxy and to obtain a better understanding of the downstream causal pathways and biological mechanisms of the drug target.

In summary, this thesis demonstrates how a combination of genetic analyses and functional validation studies is a powerful approach to validate GWAS results and further characterise therapeutic effects. This PhD project identified relevant genetic markers to genetically proxy therapeutic modulation of biomarker levels through PHD inhibition and could potentially inform further research using patient-level clinical data from Phase III trials.

Overview of the Data Chapters

Chapter 3 centres on the identification of a genetic variant, lying in *cis* with the *EPO* gene, associated with higher circulating erythropoietin (EPO) levels (the downstream effect of PHI treatment) by performing a GWAS meta-analysis. The identified genetic variant is validated as the most likely causal variant by testing its association with hepatic and renal gene expression and performing colocalisation analysis. The variant is then used as a genetic proxy for therapeutic modulation of endogenous EPO levels in Mendelian Randomisation analysis to predict the risk of cardiovascular disease (CVD) associated with therapeutically altered endogenous EPO levels.

Chapter 4 outlines the establishment of a whole *EPO* gene knock-out model using CRISPR-Cas9 gene-editing and whole transcriptomic profiling to better understand downstream transcriptomic changes and pathways involved in EPO signalling.

Chapter 5 utilises and further develops a protocol for seamless single base gene-editing. A relatively new technique combining CRISPR-Cas9 gene-editing with the *piggyBac™* transposon system is applied to establish a heterozygous knock-in model of a *cis-EPO* variant to functionally validate the variant as causal in controlling *EPO* expression levels.

Chapter 6 focuses on investigating the long-term effects of rises in circulating haemoglobin (Hgb) levels through therapeutic inhibition of the prolyl hydroxylase (PHD) enzymes with PHI treatment. Genetic variants associated with circulating Hgb levels lying within the genes encoding the PHDs are selected to genetically proxy therapeutic PHD inhibition and examine the long-term effects of higher circulating Hgb levels on risk of cardiovascular disease or any other unwanted effects.

Acknowledgments

A PhD would not be possible without the support and guidance from some of the very best in the field. I was very fortunate to have that in the form of my primary supervisor, Professor Tim Frayling. I would like to thank him for his continued help, support and encouragement throughout the last 4 years. He is always quick to respond, willing to help tackle any problems and has a huge amount of confidence in my ability. Despite being incredibly busy, he always makes time to support me and his knowledge as well as management skills are inspirational. I feel incredibly lucky that I was able to work under his supervision and as part of the Genetics of Complex Traits Team whom have all been helpful, friendly and supportive. A special thanks in particular to Dr Andy Wood for his help whenever needed; your statistical and coding skills never cease to amaze me. I would also like to thank my second supervisor, Dr Asami Oguro-Ando, for taking me on despite not originally being my supervisor and for supporting and facilitating the functional studies that I wanted to incorporate into the project. You have taught me a considerable amount and my PhD would not be what it is without the help of your lab. With that I owe a considerable amount to the O-A lab members, Rosie and Josan, who helped troubleshoot, assisted with experiments and made all those long and stressful lab hours bearable. Thank you also goes to my third supervisor, Dr Audrey Chu. Again, despite being handed over to you after the departure of my previous supervisors, you have been so incredibly uplifting, empowering, supportive and valuable. Thank you for making sure my placement at GSK happened, catching up on all the analysis related to this thesis, for getting my papers and thesis through all the loopholes at GSK and for helping me gain invaluable skills, insights and networks. Thank you to all at GSK who were so warming, welcoming and inspiring, particularly Nicholas Bowker and Chloe Robbins; it was an honour to learn from you and assist your projects in some way. A special mention also to my former supervisors, Dawn Watermouth and Vickas Patel who, despite leaving GSK, have always provided great, very timely feedback, words of encouragement and helped initiate this whole PhD project in the first place. Thank you to both the MRC and to GSK for funding this studentship.

Doing a PhD, buying a house, planning a wedding and getting a dog are probably a few of the most stressful things that one can endure in their lives, and for some reason, I don’t know why, I decided to face all of these in the last 4 years with a global pandemic thrown into the middle of it. Somehow, just somehow, I have managed it and for that I am incredibly proud of myself.

I would not have been able to face even one of these things, let alone all of them, without the incredible support network of my friends and family. To Chris, my best friend, my partner-in-crime and now finally my husband(!), thank you for always being so supportive, caring, patient, understanding (most of the time) and generous. Thank you for the endless love, cuddles and laughter along the way, for always listening and pretending to understand and for being my shoulder to cry and lean on. I know it hasn’t been the easiest ride and that I may often take my stress out on you, but together we have done it. I could not have made it through this roller-coaster without you and I cannot wait to see where the next adventure takes us. Mama and Papa, without you I wouldn’t be where I am today and I wouldn’t have the drive, determination, or motivation that I do. Thank you for always believing in me, pushing me, supporting me, and encouraging me to follow my dreams; I will forever be grateful for the sacrifices you’ve made and opportunities you have given me. Mum, thank you for picking up the phone whenever and wherever, for all the help and support no matter what; I know I don’t always show it, but I really do appreciate it. Dad, you will never know how much those lunch dates helped save my sanity and got me through to the end. I hope I have made you both proud. To Josie and Harry, sorry that I took all the intelligent genes, good job you both got the better-looking ones ey! Only kidding! I know I’m often quite annoying but thank you both for being my biggest supporters, for putting up with me, and for never failing to make me laugh. I hope I inspire you to do everything you want to do. To the one and only GDAD, I know how proud you will be of me in completing this PhD just as you have been through my other milestones. Thank you for the continuous faith and for all the texts, coffee, cakes, and lunches; you really are the best. To Gemma, how lucky was I that I met you on that very first, daunting day of our PhDs. I could not have asked for a better lunch buddy, listener, bingo player, netball teammate or Auntie to Miska. I have honestly made a friend for life in you, and I would have struggled to get through this PhD without you. To

Hayley, I’m so glad I made it past the 7-year mark! Thank you for the endless messages getting me through each week, for listening to my rants and worries, for the random surprises in the post and for all the fun weekends away from it all - you are one of a kind. To McGeevs, my longest and truest best friend. Who would have thought back in Year 7 French that I would eventually have a PhD under my belt?! Thank you for always being present, for always putting a smile on my face and for being there every step of the way through thick and thin. You will never know how much I value our friendship. Here’s to the next 15+ years of us. To Hobnob, my forever wine-drinking and stress-releasing bestie, I am so glad you moved back to Exeter. Thank you for always being there, for the never-ending good advice, for the belief you have in me and for always understanding. You are a keeper. And last but not least, to my fur babies, Miska & Jerry. Thank you for showing me unconditional love, for always being happy to see me, for keeping me company throughout the pandemic and the writing and for getting me out in the fresh air (even if that does often mean running after you Miska!).

It is all of you that have got me through this, believed in me when I didn’t believe in myself and continually encouraged me. I cannot even begin to thank you enough and will forever be grateful. I cannot wait to share the next chapter of my life with you.

Table of Contents

List of Figures

List of Tables

Publications arising from this thesis

Chapters 3-5

Currently under review at AJHG. Manuscript available upon request.

Harlow, CE. Gandiwijaya, J. Bamford, RA. Wood, AR. Van der Most, P. Verweij, N. [25 authors] & Frayling, TM. 2022. Identification and single-base gene-editing functional validation of a *cis-EPO* variant for use to mimic novel EPO-increasing therapies.

Chapter 6

Currently under internal review at GSK. Manuscript available upon request. Submission planned for PloS Genetics.

Harlow, CE. Patel, VV. Waterworth, DM. Wood, AR. Beaumont, R. Ruth, KS. Tyrell, J. Oguro-Ando, A. Chu, AY & Frayling, TM. 2022. Genetically proxied therapeutic inhibition of PHD enzymes and cardiovascular risk.

Other publications arising from this thesis:

Pulit, SL. **Stoneman, CE.** et al. 2019. Meta-analysis of genome-wide association studies for body fat distribution in 694,649 individuals of European ancestry. *Human Molecular Genetics*. PMID: 30239722.

Frayling, TM & **Stoneman, CE**. 2018. Mendelian randomisation in type 2 diabetes and coronary artery disease. *Current Opinion in Genetics & Development.* PMID: 29935421.

Author’s declarations

I was involved in the study design, analyses and manuscript preparation for all of the studies that are included as chapters in this thesis. For each study included, I was the first author on the corresponding paper. Co-authors of the publications arising from this thesis, as well as internal GSK reviewers, made suggestions of changes within the text of the papers and thus some of the text contained within this thesis was suggested by them.

All bioinformatics analyses were carried out by myself with a few exceptions:

Chapter 3

- The GWAS performed on circulating EPO levels within the three independent study cohorts, Health ABC, BLSA and PREVEND, were performed by the respective study analysts, Hampton Leonard, Niek Verweijj and Toshiko Tanaka.
- The hepatic eQTL analysis was performed by Peter Van der Most. The renal eQTL analysis was performed by Amy Etheridge.

Chapter 3 & Chapter 6

- The UK Biobank data used were generated within the Genetics of Complex Traits team. Dr Andy Wood and Dr Robin Beaumont were responsible for all quality controls checks, defining ancestries and performing GWAS. Dr Jessica Tyrell and Dr Kate Ruth were responsible for the curation of phenotypes.

All laboratory work and associated analyses were carried out by myself with the following exceptions:

Chapter 4

- Library preparation and RNA sequencing was performed by Audrey Farbos and Dr Karen Moore at the University of Exeter Sequencing Service.
- Yuriko Iizuka, an international placement student from the University of Tohoku that I directly supervised, assisted with some of the qRT-PCR experiments under my guidance.

Chapter 5

- Yuriko Iizuka assisted with the construction of the plasmids and with the growth and maintenance of cells under my guidance.

Abbreviations

A1	Allele 1
ACR	Albumin creatinine ratio
ATP	Adenosine Triphosphate
BLSA	Baltimore Longitudinal Study of Aging
bp	base-pair
BP	Blood pressure
CAD	Coronary Artery Disease
CADD	Combined Annotation Dependent Depletion score
Cas9	CRISPR-association 9 protein
cDNA	complementary DNA
chr	chromosome
CI	Confidence Intervals
CKD	Chronic kidney disease
CNV	Copy Number Variation
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
crRNA	CRISPR-RNA
CVD	Cardiovascular disease
DBP	Diastolic blood pressure
DD	Dialysis-dependent
ddH_20	Double distilled water
DEGs	Differentially Expressed Genes
DMEM	Dulbecco's Modified Eagle Medium
DNA	Deoxyribose nucleic acid
DSB	Double-stranded break
Dups	Duplications
EAF	Effect allele frequency
EDTA	Ethylenediaminetetraacetic acid
EGFP	Enhanced Green Fluorescent Protein
eGFR	Estimated glomerular filtration rate
EPO	Erythropoietin
EPOR	EPO receptor
eQTL	expression quantitative trait loci
ESA	Erythropoietin Stimulating Agent
ESRD	End-stage renal disease
FBS	Fetal Bovine Serum
FDR	False discovery rate
FIAU	Fialuridine
FIH	Factor inhibiting HIF
freq	frequency
GFR	Glomerular filtration rate
GO	Gene Ontology
gRNA	guide RNA
GRM	Genomic relationship matrix

GRS	Genetic risk score
GWAS	Genome-wide association study
GWS	Genome-wide significant
HAMP	Hepcidin gene
HCT	Haematocrit
HDR	Homology directed repair
Health ABC	The dynamics of Health, Aging and Body Composition
HEDI	Heterogeneity in dependant instruments
HEK-293	Human Embryonic Kidney 293 cells
hEPO	Human EPO
Hgb	Haemoglobin
HIF	Hydroxia inducible factors
HK1	Hexokinase
HR	Homologous recombination
HRC	Haplotype Reference Consortium
HRE	Hypoxia response element
InCHIANTI	Invecchiare in Chianti
ITRs	Inverted terminal repeats
IVW	Inverse variance weighted
JAK2	Janus Kinase 2
kb	kilobases
KO	knock-out
KOA	Knock-out A
KOB	Knock-out B
LD	Linkage disequilibrium
LDL	Low-density lipoprotein
LOF	Loss of function
MAC	Minor allele count
MAF	Minor allele frequency
MAPK	Mitogen-activate phosphokinase
mb	megabases
mCh	mCherry fluorescence
MDRD	Modification of diet in renal disease equation
MI	Myocardial Infarction
MODY	Maturity-onset diabetes of the young
MPRAs	Massively-parallel reporter assays
mQTL	methylation QTL
MR	Mendelian Randoisation
mRNA	messenger RNA
N	Number
NAFLD	Non-alcoholic fatty liver disease
ncRNA	non-coding RNA
NDD	Non-dialysis dependent
NEB	New England Biolabs
NGS	Next Generation Sequencing

NHEJ	Non-homology end joining
NIA	National Institute of Aging
OR	Odds Ratio
P-adj	P-adjusted
PAM	Protospacer adjacent motif
PB	*piggyBac^{TM}* transposon
PBS	Phosphate Buffered Saline
PBx	Excision-only *piggyBac^{TM}* transposase
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PCs	Principal Components
PHD	Prolyl hydroxylase enzymes
PheWAS	Phenome-wide association study
PHI	Prolyl hydroxylase inhibitors
PI3K	Phosphoinositide 3-kinase
PMSF	Phenylmethylsulphonyl fluoride
PNK	Polynucleotide Kinase
poly-A	3'polyadenylated
pQTL	protein QTL
pre-crRNA	Precursor CRISPR-RNA
PREVEND	Prevention of Renal and Vascular End-stange Diseases
QC	Quality Control
QQ	Quantile Quantile
qRT-PCR	quantitative reverse-transcription polymerase chain reaction
QTL	Quantitative Trait Loci
R&D	Research and Development
r1	Read 1
r2	Read 2
RCT	Randomised control trial
Ref	reference
REPCs	Renal EPO producing cells
rhEPO	Recombinant Human EPO
RIN	RNA Integrity Number
rlog	Regularised logarithm
RNA	Ribonucleic Acid
RNA-seq	RNA-sequencing
rRNA	ribosomal DNA
SB	Sleeping Beauty
SBP	Systolic blood pressure
SD	standard deviations
SE	Standard error
SEM	Standard error of the mean
SMR	Summary-data based MR
SNP	Single Nucleotide Polymorphism
SpCas9	Streptococcus Pyogenes Cas9

ssODNs	Single-stranded oligodeoxynucleotides
TALE	Transcription activator-like effector
TALENS	Transcription activator-like effector nucleases
tk	tyrosine kinase
TF	Transcription factor
tracrRNA	Trans-activating crRNA
UKB	UK Biobank
VEP	Variant Effect Predictor
VHL	Von Hippel-Lindau
VST	Variance Stabilising Transformation
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
WT	wild-type
ZFN	Zinc-finger nucleases

Chapter 1 General Introduction

This chapter provides a general introduction and background into the overarching aim of this thesis in using genetics to investigate the long-term effects of therapeutic modulation on cardiovascular risk and the different techniques employed.

Some sections have been taken directly from a literature review published in the Human of Molecular Genetics journal which I wrote during the first year of my PhD with the help of my primary supervisor Professor Tim Frayling;

Mendelian randomisation in type 2 diabetes and coronary artery disease. Timothy M Frayling, Charli E Stoneman*.* Human of Molecular Genetics, PMID 29935421, DOI 10.1016/j.gde.2018.05.010

1.1 Anaemia of Chronic Kidney Disease

1.1.1 Background

Chronic kidney disease (CKD) affects between 8-16% of people globally and is a progressive long-term condition characterised by a loss of kidney function (T. K. Chen et al., 2019). CKD most commonly occurs in the elderly population and people of African American or Hispanic ethnicities (Saran et al., 2019). CKD is assessed in terms of overall kidney function and the presence of kidney damage (Couser et al., 2011). Kidney function is measured using the estimated glomerular filtration rate (eGFR) with values less than 60 mL/min/1.73m^2 defining kidney disease (T. K. Chen et al., 2019; Eknoyan et al., 2013). Kidney damage is ascertained by the level of albuminuria (defined by a urine albumin/creatine ratio [ACR] > 30 mg/g), kidney biopsy, kidney transplant, structural abnormalities or haematuria (Eknoyan et al., 2013; Levin et al., 2013). CKD is classified in stages; stages 1 and 2 require the presence of proteinuria and reduced eGFR depicting some kidney function loss, stages 3 and 4 represent moderate CKD with less than 50% of kidney function and stage 5 is defined as end-stage renal disease (ESRD) where there is little to no kidney function (eGFR < 15 mL/min/1.73m^2) and dialysis is required (Couser et al., 2011). These patients are denoted dialysis dependent (DD) and those not yet receiving dialysis are denoted non-dialysis dependent (NDD). CKD is a heterogenous disorder with a range of causes. The most common causes are diabetes, hypertension, glomerulonephritis, infection and certain genetic risk factors, such as the presence of two *APOL1* risk alleles (Jha et al., 2013; A S Levey, 2021; O'Seaghdha et al., 2011; Shafi & Coresh, 2019; Tzur et al., 2010). CKD is associated with additional complications, such as an eight- to ten-fold increase in cardiovascular mortality and mineral or bone density disorders, with the severity of complications increasing as the CKD advances (T. K. Chen et al., 2019; Couser et al., 2011; R. Thomas et al., 2008).

Anaemia is one of the most common complications of CKD affecting one in seven CKD patients (T. K. Chen et al., 2019; Stauffer & Fan, 2014). Anaemia increases in prevalence as kidney disease progresses affecting the majority of patients with stage 5 CKD (Couser et al., 2011; KDOQI, 2006). Anaemia of CKD is typically normocytic, normochromic and hypo proliferative (Babitt & Lin, 2012). Anaemia is characterised by a reduced absolute number of circulating erythrocytes and reduced haemoglobin (Hgb) or haematocrit (HCT) levels reducing oxygen-carrying capacity and oxygen tissue delivery **(**Figure 1.1**)** (Chaparro & Suchdev, 2019). Reduced oxygen tissue delivery can lead to tissue damage, organ failure and eventually death (Guowen Liu et al., 2006)**.** Anaemia is associated with faster progression of CKD alongside poorer quality of life and increased risk of cardiovascular disease (CVD), thromboembolism, hospitalisation, cognitive impairment, morbidity and mortality (Di Lullo et al., 2015; Jankowski et al., 2021; Q. Zheng et al., 2021). Multiple mechanisms lead to the development of anaemia in CKD including low erythropoietin (EPO) levels, inflammation, bleeding and reduced iron availability all resulting in a reduced number of healthy circulating erythrocytes (Zumbrennen-Bullough & Babitt, 2013). The primary driver of anaemia in CKD is the relative deficiency of EPO **(**Figure 1.1**)** (Jelkmann, 2011).

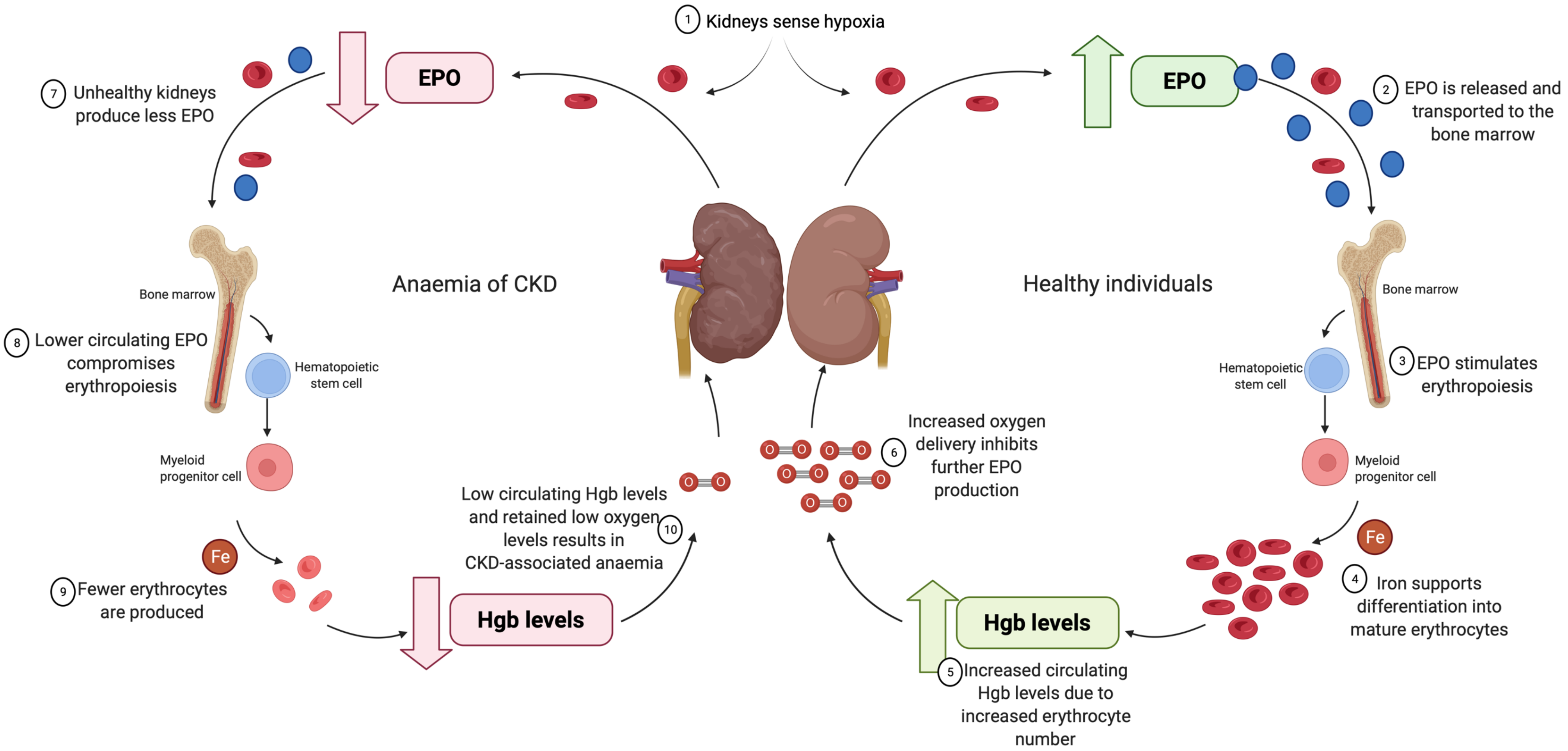

1 Kidneys sense hypoxia
EPO
7 Unhealthy kidneys produce less EPO
EPO
2 EPO is released and transported to the bone marrow
Anaemia of CKD
Healthy individuals
Bone marrow
8 Lower circulating EPO compromises erythropoiesis
Hematopoietic stem cell
Myeloid progenitor cell
Bone marrow
Hematopoietic stem cell
3 EPO stimulates erythropoiesis
Myeloid progenitor cell
Increased oxygen delivery inhibits further EPO production
6
O O
Low circulating Hgb levels and retained low oxygen levels results in CKD-associated anaemia
10
Fe
9 Fewer erythrocytes are produced
Hgb levels
Fe
4 Iron supports differentiation into mature erythrocytes
Hgb levels
5 Increased circulating Hgb levels due to increased erythrocyte number

Figure 1.1: The role of EPO in the development of anaemia in CKD. *(1) In the presence of low oxygen levels (hypoxia), EPO is released from the healthy kidneys and transported through the blood to the bone marrow (2). In the bone marrow, EPO binds to the EPO receptor (EPOR) and stimulates erythrocyte production and development (3). Iron is used to support the final stage of erythropoiesis producing mature erythrocytes (4). Increased number of mature erythrocytes results in increased circulating Hgb levels (5) leading to increased oxygen-carrying capacity within the cells and increased oxygen delivery to tissues. Increased oxygen delivery inhibits further EPO production through negative feedback mechanisms (6). In CKD, due to aberrant kidney function, less EPO is released by the kidneys in response to hypoxia (7). The lower circulating EPO levels results in decreased erythrocytosis (8) resulting in the production of fewer mature erythrocytes (9). Fewer erythrocytes result in reduced circulating Hgb levels and the retention of low oxygen levels resulting in the development of anaemia (10). Created by BioRender.com.*

1.1.2 The hypoxic response pathway

The hypoxic pathway is the primary regulator of EPO production through controlled transcription of the *EPO* gene alongside other hypoxic response genes, such as *VEGF, HAMP* (encoding hepcidin), and *HK1* (encoding hexokinase) (J. W. Lee et al., 2019; Masoud & Li, 2015; S. Ramakrishnan et al., 2014; Watts et al., 2020)**.** The hypoxic signalling pathway is an adaptive molecular mechanism activated in response to low oxygen levels (J. W. Lee et al., 2019). The master regulators of transcription in response to hypoxia is the family of transcription factors (TFs) termed the hypoxia-inducible factors (HIFs) (Figure 1.2**)** (Sormendi & Wielockx, 2018). The constitutively expressed HIFs consist of the nuclear HIFβ and the cytoplasmic oxygen-dependent HIFα subunits (Ziello et al., 2007). In the presence of oxygen, HIFα is hydroxylated at two prolyl residues by the oxygen- and iron-dependent HIF prolyl hydroxylase enzymes (PHD1-3 encoded by *EGLN1-3*) **(**Figure 1.2**)** (Rodriguez et al., 2021). Hydroxylation of HIFα allows the binding of the Von Hippel-Lindau (VHL) protein leading to ubiquitination and proteasomal degradation **(**Figure 1.2**)** (F. S. Lee & Percy, 2011). During normoxia, the GATA2 TF is also bound to the promoter region of *EPO* preventing transcription of the *EPO* gene **(**Figure 1.2**)** (Jelkmann, 2011). During hypoxia (a condition where oxygen levels are limited), the PHD enzymes are less active resulting in the stabilisation of HIFα subunits enabling the translocation to the nucleus and the formation of a heterodimeric complex through interaction with the nuclear HIFβ subunit and P300/CBP **(**Figure 1.2**)** (Haase, 2013; J. W. Lee et al., 2019). Due to the simultaneous decrease in GATA2 levels, the promoter region of hypoxic response genes becomes available for the binding of the HIF-P300/CBP complex subsequently increasing transcription of hypoxic response genes stimulating erythropoiesis and restoring oxygen homeostasis **(**Figure 1.2**)** (Jelkmann, 2011; Schönenberger & Kovacs, 2015; Shih et al., 2018)**.** This classic hypoxic response pathway and ability to restore oxygen homeostasis is significantly hampered in CKD patients due to the unhealthy kidneys releasing less EPO (Koury & Haase, 2015).

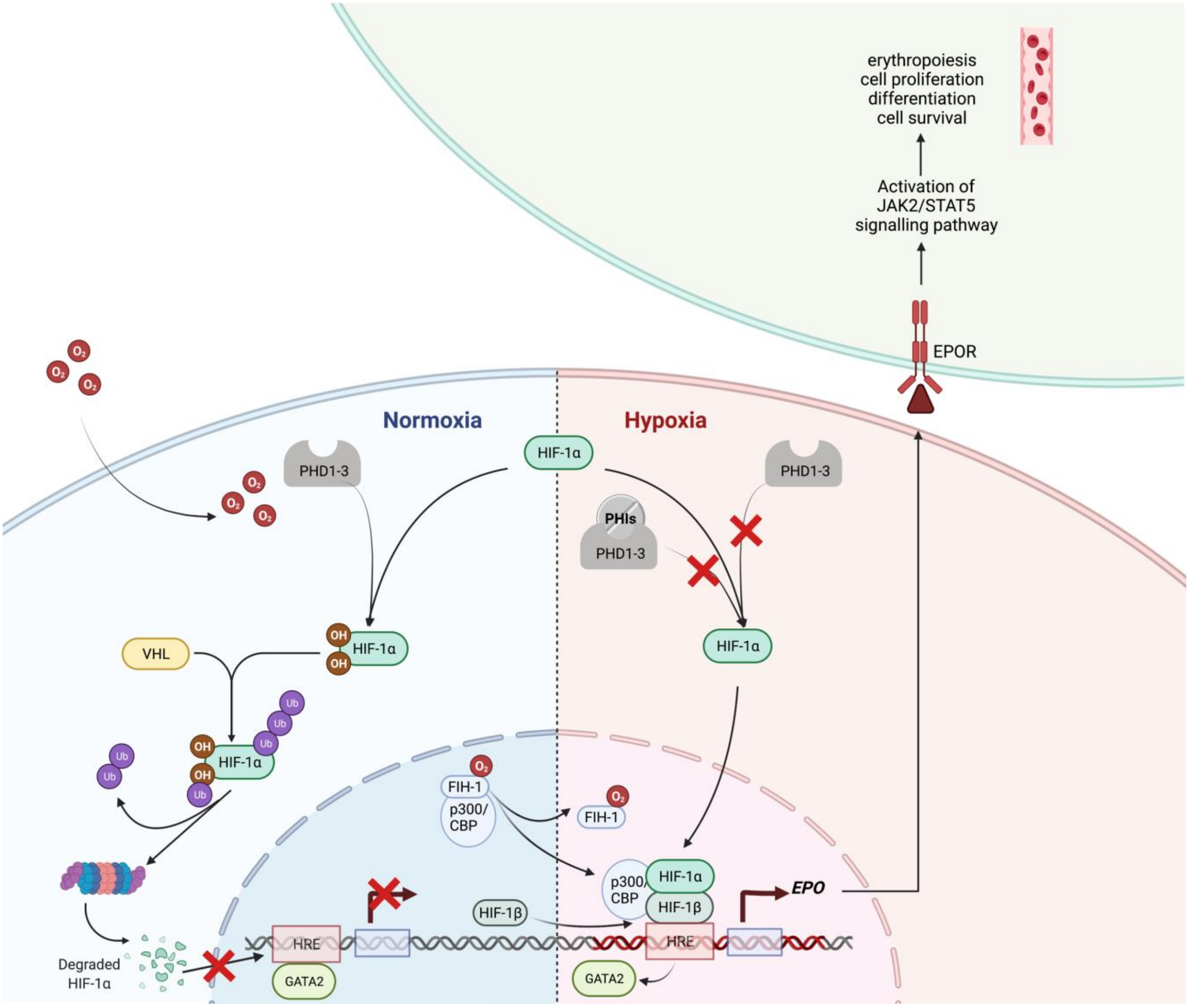


Figure 1.2: The hypoxic response pathway. *In the presence of normal oxygen levels (normoxia), the HIFα subunits, primarily HIF-1α, are hydroxylated by the PHD1-3 enzymes. Hydroxylation marks HIF-1α for proteasomal degradation through the ubiquitination by the VHL protein. The GATA2 repressive transcription factor is also bound to the hypoxia response element (HRE) during normoxia preventing binding of any activating transcription factors. During hypoxia, when oxygen levels are limiting, the PHD1-3 enzymes become inactive allowing stabilisation of HIF-1α and levels of GATA2 are repressed making the HRE available for binding. Stabilised HIF-1α is able to translocate to the nucleus where it forms a heterodimer complex with HIF-1β subunit enabling the recruitment of p300/CBP. The HIF1α-HIF1β-p300/CBP complex is able to bind to the HRE initiating transcription of the hypoxic response genes, primarily EPO. EPO is then secreted and binds to the EPO receptor (EPOR) where the JAK2-STAT5 signalling cascade is stimulated increasing erythropoiesis, cell signalling activity, proliferation, cell growth and differentiation. PHIs are a novel class of drugs for treating anaemia in CKD which act at the transcriptional level of the EPO gene. PHIs inhibit the PHD enzymes increasing activation of the hypoxic response pathway. HIF = hypoxia inducible factor; PHI = prolyl hydroxylase inhibitor; PHD = prolyl hydroxylase enzyme; VHL = Von Hippel-Lindau; HRE = hypoxic response element; JAK2 = Janus kinase 2; FIH-1 = factor inhibiting HIF. Created with BioRender.com.*

1.1.3 Erythropoietin

Erythropoietin (EPO), a glycoprotein cytokine, is the primary hormone responsible for effective erythropoiesis (A. K. Singh, 2018). The main sites of EPO production are the kidneys and the liver (Noguchi et al., 2008). Hepatic EPO production predominates in the foetal and perinatal periods, whilst renal EPO production predominates in adulthood (Shih et al., 2018; Suresh et al., 2020). In the kidney, EPO is produced by the interstitial fibroblasts, the peritubular capillary and the proximal convoluted tubule (Fisher et al., 1996; Nagai et al., 2014; Shih et al., 2018; Zeisberg & Kalluri, 2015). In the liver, EPO is produced in the perisinusoidal cells (K. U. Eckardt, 1996). EPO is regulated at the transcriptional level, through transcription of the *EPO* gene located on chromosome 7, in response to oxygen levels (Jelkmann, 2011). Low levels of EPO are constantly secreted to maintain a continuous turnover of erythrocytes (Suresh et al., 2020). During hypoxia, there is increased transcription of the *EPO* gene resulting in raised circulating EPO levels in an attempt to improve oxygen delivery by increasing the absolute number of erythrocytes (F. S. Lee & Percy, 2011; Souma et al., 2015).

EPO exerts its effects through binding to the EPO receptor (EPOR) initiating intracellular cell signalling through recruitment and activation of the Janus kinase 2 (JAK2) signalling cascade **(**Figure 1.2**)** (Koury & Haase, 2015). JAK2 signalling activates the STAT5, MAPK (mitogen activated protein kinase), and PI3K (phosphoinositide 3-kinase)-AKT pathways (Koury & Haase, 2015; F. S. Lee & Percy, 2011). Activation of these intracellular signalling pathways promote cell differentiation, cell survival, and proliferation by protecting against apoptosis (F. S. Lee & Percy, 2011). In this way, EPO regulates the differentiation, proliferation, and survival of erythrocytes. The EPOR has only shown to be expressed on erythroid progenitor cells and non-erythroid cells, such as neural cells, skeletal myoblast cells, and endothelial cells indicating a non-haematopoietic role for EPO which differs depending upon the tissue or cell-type (Broxmeyer, 2013; Lamon & Russell, 2013; Noguchi et al., 2008; Suresh et al., 2020). These non-haematopoietic roles are thought to include, but are not limited to, exerting cytoprotective, neuroprotective, and antiapoptotic effects, alongside energy metabolism and a response to inflammation or stress

(Jelkmann, 2011; Noguchi et al., 2008; Suresh et al., 2020; L. Wang et al., 2014; Yuanyuan Zhang et al., 2014). EPO also plays an important role in the regulation of internal iron stores (Batchelor et al., 2020). Iron is crucial for effective erythropoiesis and restoration of oxygen homeostasis by controlling differentiation of erythroblasts into reticulocytes and forming Hgb, the oxygen carrier (Batchelor et al., 2020). EPO influences the majority of iron stores through controlled regulation of erythrocyte destruction. EPO also influences iron uptake, mobility, and utilisation through hypoxia-controlled expression of the hepcidin gene (*HAMP*) by increasing secretion of erythroferrone from erythroid cells through activation of the JAK2-STAT5 pathway (Kaplan et al., 2018; Muckenthaler et al., 2017).

This tightly coordinated regulation of EPO and iron is important for maintaining effective erythropoiesis and subsequent Hgb levels and oxygen homeostasis (Haase, 2010; Watts et al., 2020). EPO levels can be substantially intervariable amongst individuals which is partly driven by genetics and the environment. Some individuals are able to adapt to low oxygen levels experienced at high altitudes better than others due to the speed of EPO production which is thought to be driven by epigenetic modifications, particularly DNA methylation (Childebayeva et al., 2019; Friedmann et al., 2005). Permanent residents of high altitudes have acclimatised to lower oxygen levels by increasing baseline EPO levels and are often protected against conditions resultant of sustained high circulating EPO or Hgb levels, such as polycythaemia (excessive number of erythrocytes) (Gardie et al., 2014). Mutations have been found within hypoxic response genes, *HIF2A, PHD2* and *PPARA* in permanent high-altitude residents as a result of genetic adaptation and selection (Haase, 2013). These individuals have higher resting ventilation, lower Hgb levels and more efficient oxygen utility but are less susceptible to chronic mountain sickness and associated adverse effects, such as hypertension and heart failure (Horscroft et al., 2017; Suresh et al., 2020). Mutations within the *EPO* gene in these individuals have not been reported. Rare gain of function mutations within the *EPO* gene and rare loss-of-function mutations within *EGLN1* (encoding PHD2) have previously been reported (Gardie et al., 2014; Zmajkovic et al., 2018). These mutations result in inherited or secondary polycythaemia due to sustained high levels of circulating EPO, and thus erythrocyte number and

Hgb/HCT levels. Polycythaemia leads to increased blood viscosity and adverse cardiovascular risk due to the excessive EPO and/or HCT levels (Gardie et al., 2014; Takeda et al., 2006).

The significant increase in oxygen uptake and improvement in oxygen utilisation in response to EPO has been exploited over the years by professional athletes in the form of blood-doping to improve endurance and performance by increasing the time before muscles fatigue (Tokish et al., 2004). Blood doping leads to supra-physiological EPO levels resulting in an abnormally high erythrocyte volume which is associated with increased risk of hypertension, stroke, blood clots, heart attacks, embolisms, and seizures in these athletes (Garimella et al., 2016; La Gerche & Brosnan, 2017). As the kidneys no longer retain full functioning in CKD, EPO is no longer released at sufficient levels resulting in increased hepcidin expression, reduced iron metabolism and mobility and reduced erythropoiesis reducing Hgb levels and leading to anaemia (Sugahara et al., 2017; Weiss et al., 2019).

1.1.4 Treatments for Anaemia in CKD

EPO deficiency is the primary cause for anaemia of CKD and therefore, the current standard of care for anaemia in CKD is the parenteral administration of recombinant human EPO (rhEPO) or its analogs, otherwise known as erythropoietin stimulating agents (ESAs) (Mikhail et al., 2017). Iron therapies or transfusions are also regularly used (Shepshelovich et al., 2016)**.** These treatments aim to increase erythrocyte production in an attempt to correct the anaemia by restoring oxygen tissue delivery. Despite potential benefits of these treatments including improvement in the quality of life and less need for blood transfusions, they do have limitations (Q. Zheng et al., 2021). Limitations of oral iron include prolonged treatment regimes and poor compliance due to gastrointestinal adverse effects, whilst limitations of ESAs or intravenous iron include inconvenient administration (by injection or infusion, respectively), pain at the injection site and risk of adverse effects such as hypersensitivity with intravenous iron or hypertension with rhEPO (Baird-Gunning & Bromley, 2016; Bonomini et al., 2016; Clement et al., 2010; Krapf & Hulter, 2009). Despite ESAs alleviating EPO deficiency, there are concerns regarding safety and efficacy due to the supra-physiological circulating EPO levels leading to sudden and/or excessive rises in circulating Hgb levels (Jelkmann, 2013)**.** Several studies have reported an increased risk of stroke, myocardial infarction (MI) and thromboembolism with supra-physiological EPO levels (Babitt & Lin, 2012; Locatelli & Del Vecchio, 2003; Pfeffer et al., 2009; Santhanam et al., 2010; A. K. Singh et al., 2006; Yi Zhang et al., 2004)**.** These safety concerns have led to ongoing efforts to develop novel treatments for anaemia in CKD. One class of treatments is oral HIF prolyl hydroxylase inhibitors (PHIs) which have recently completed Phase III clinical trials (Chertow et al., 2021; K.-U. Eckardt et al., 2021; Provenzano et al., 2021; A. K. Singh, Carroll, McMurray, et al., 2021; A. K. Singh, Carroll, Perkovic, et al., 2021)**.** PHIs stimulate endogenous EPO production within the physiological range by acting at the transcriptional level of the hypoxic response genes through inhibition of the PHD enzymes (A. K. Singh, Carroll, McMurray, et al., 2021). PHD inhibition prevents hydroxylation of HIFα subunits increasing stabilisation and allowing dimerization with the HIFβ subunit to initiate transcription of genes protecting against hypoxia, such as *EPO* **(**Figure 1.2**)** (Sugahara et al., 2017). PHIs also influence iron mobility and transport through acting upon this pathway by controlling expression of hepcidin (Kaplan et al., 2018). Clinical trials in DD- and NDD-CKD patients have already shown PHIs to be as effective as ESAs at maintaining Hgb levels over a 24-week period with small increases in circulating EPO levels (Brigandi et al., 2016; K.-U. Eckardt et al., 2021; Holdstock et al., 2019; Meadowcroft et al., 2019; Provenzano et al., 2021). Recent safety and efficacy data through completion of Phase III trials has emerged providing evidence that PHIs are noninferior to ESAs with respect to Hgb levels and risk of cardiovascular outcomes, primarily stroke, MI or CAD (Chertow et al., 2021; K.-U. Eckardt et al., 2021; Provenzano et al., 2021; A. K. Singh, Carroll, McMurray, et al., 2021; A. K. Singh, Carroll, Perkovic, et al., 2021). Certain PHIs have already received approval for anaemia in CKD treatment in Japan supporting continual ongoing development of PHIs for anaemic CKD patients worldwide (Akizawa, Nangaku, et al., 2020; Nangaku et al., 2021).

1.2 Genetics to aid the drug development process

1.2.1 The drug development process

The drug development process is resource-intensive, costly, time-consuming, and inefficient (Kaitin, 2010). For a drug candidate to be successful, it needs to pass through several phases, including discovery, pre-clinical development and clinical, before reaching the clinic with the pharmaceutical industry expending a considerable amount of time, effort, risk, and resources during each phase **(**Figure 1.3**)** (DiMasi et al., 2016; Kaitin, 2010). Although research and development (R&D) has improved in terms of productivity and technology over the years, the most difficult step is bringing potential therapeutic candidates out of discovery and through the cumbersome development process **(**Figure 1.3**)** (Kaitin, 2010). The cost of drug development continues to rise whilst the probability of drug candidates being approved remains low **(**Figure 1.3**)** (King et al., 2019). Despite success rates in late-stage development (Phase III to launch) improving from 50% to 66% in the most recent analysis of drug development, success rates of launch to Phase II and Phase II to Phase III remain static with <10% and 25% of potential drug candidates making it through these two milestones, respectively (Dowden & Munro, 2019). Attrition, therefore, remains a major issue with 79% of clinical failures being attributable to safety and efficacy issues (Dowden & Munro, 2019). For the drug development process to sustain its own growth, the number of clinical failures needs reducing whilst the number of successes needs increasing (King et al., 2019). The pharmaceutical industry has therefore been looking for other means to aid the prediction of safety and potential effects, aid drug candidate prioritisation, and highlight repurposing opportunities with the end goal of reducing clinical failures and increasing successes (Nelson et al., 2015).

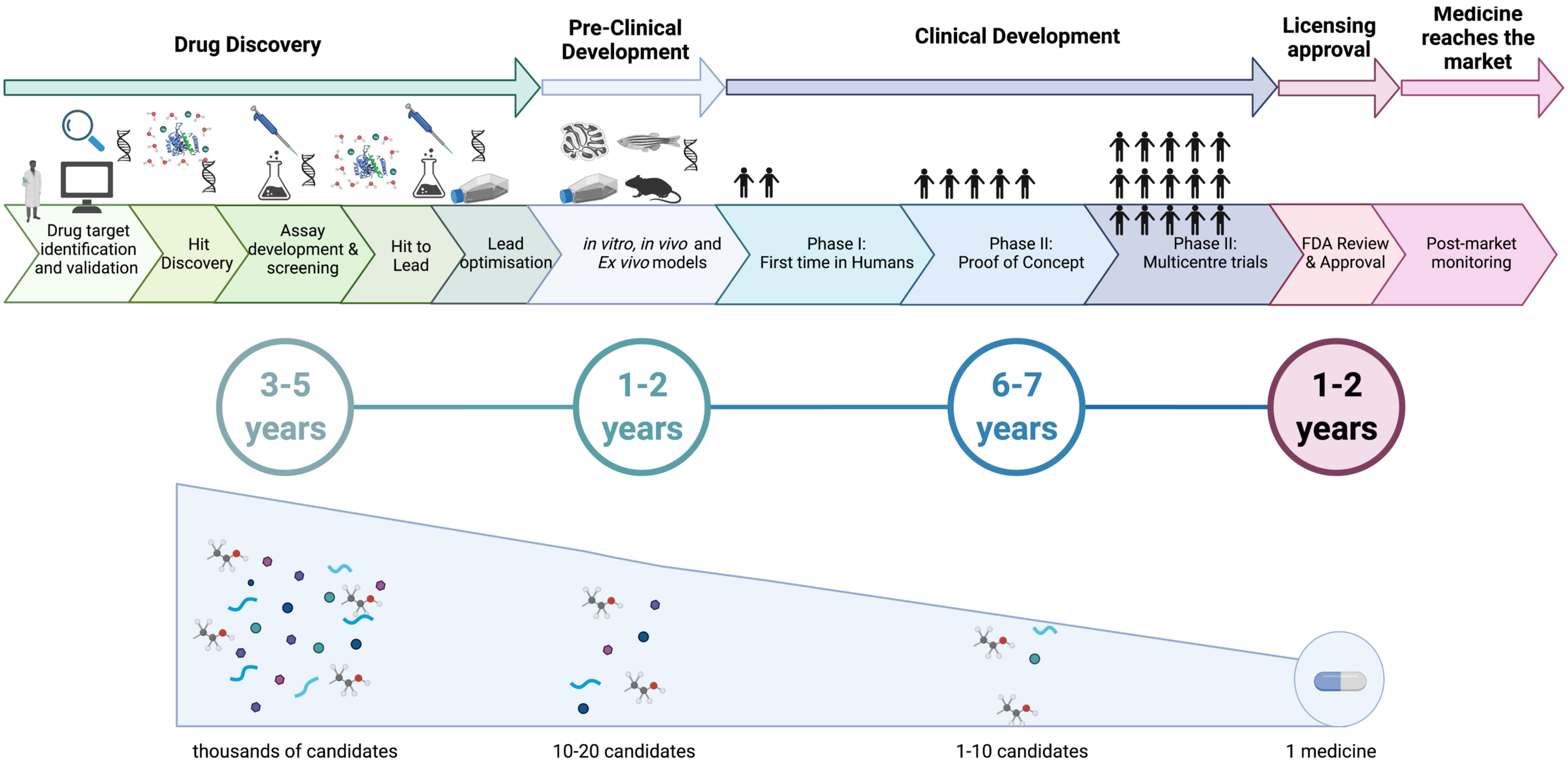


Figure 1.3: The drug development process. *See next page for figure legend.*

Figure 1.3: The drug development process. *The initial part of drug development involves the identification and validation of potential drug targets through identification of associations using genetic and expression data, structure-relationship analysis, over-expression experiments, transgenics, expression profiling, literature research, and competitor information. Potential drug targets then enter hit discovery where potential compounds (or 'hits') are identified. The compound then enters the lead discovery phase where it is undergoes high throughput screening and testing. This phase takes around 3-5 years and starts with thousands of potential candidates decreasing to only 10-20 which enter the pre-clinical phase. During the pre-clinical phase, which typically takes 1-2 years, candidates undergo in vitro and in vivo testing in animal models to assess toxicity, bioreactivity and determine the no-observed-adverse-effect levels. Successful candidates then enter clinical development where they are tested in human subjects. Phase 1 clinical trials are carried out in around 100 healthy volunteers to determine the safe dosage range and pharmacokinetic characteristics. Phase II trials are carried out in a few hundred patients with the disease. The goal of this phase is to determine drug efficacy and the minimum/maximum dose. Phase III clinical trials are then conducted in around a thousand individuals across several sites. These trials are randomised and are have a focus on intent-to-treat analysis and drug safety. The results from these trials are often published in peer reviewed journals. This process takes approximately 6-7 years and filters down the number of candidates to only a few (~1-10). Pharmaceutical companies then have the choice to move forward with submitting a new drug application to the FDA. The FDA review the submitted evidence, proposed labelling, patent information, directions for use and safety information before deciding whether the drug gets approved which can take 1-2 years. After approval, the compound enters the clinic and can be sold. Post-marketing monitoring is carried out for a couple of years. Created with BioRender.com.*

1.2.2 Human genetics and drug development

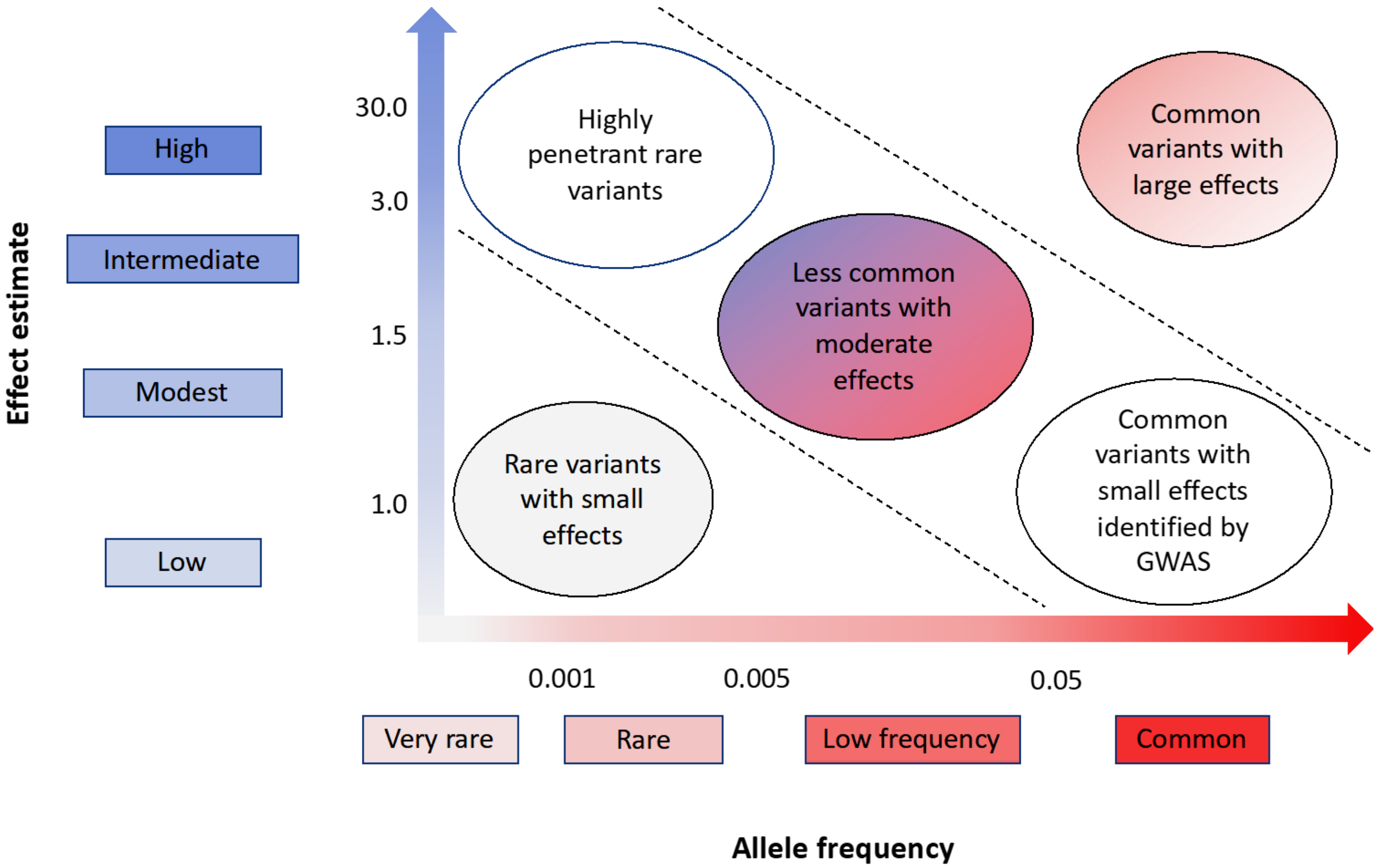
One field which has made considerable advances over the last decade in supporting drug development and continues to hold great promise in sustaining the drug development process is the field of human genetics (Nelson et al., 2015). The unprecedented increase in the amount of human genotypic and phenotypic information available has led to the identification of millions of genetic variants across the full frequency spectrum as associated with common diseases and traits (Visscher et al., 2017). This has revolutionised the field of complex diseases by providing insights into underlying genetic architecture and a better understanding of disease pathophysiology (Tam et al., 2019). Drug targets that are genetically informed are more likely to make it to Phase III clinical trials and the likelihood of the drug making it through to the clinic can be doubled if there is genetic evidence supporting drug safety and efficacy (King et al., 2019; Nelson et al., 2015). Therefore, one of the ultimate goals of genetic studies is to inform medicine by driving translational advances enabling safer and more effective strategies for disease prevention and treatment (Shuquan Rao et al., 2021). Although clinical trials are the best design for detecting small-to-moderate clinically important effects in the diseased population, they are limited by cost, time, ethical issues, the often underrepresentation of women, their power to detect adverse effects, and the bias towards the null due to failure of participant adherence (Bennett & Holmes, 2017; Carey et al., 2017; Mo et al., 2020). Genetic studies can overcome some of these limitations.

Genetic studies are often carried out at population level in large sample sizes using biobanks or cohorts (Uffelmann et al., 2021). Biobanks provide a more comprehensive understanding of the biological consequences of variants compared to case-control studies due to larger sample sizes, more robust and richer phenotyping including biomarker measurements and disease diagnosis and are often linked to electronic health records routinely updating the phenotypic data allowing for longitudinal analysis (Deaton et al., 2021; Denny et al., 2016; Diogo et al., 2018). Participants provide a wide-range of longitudinal phenotypic and genotypic data through questionnaires, electronic health records, biological samples, body measurements, clinical assessments and imaging (Sudlow et al., 2015). This wealth of readily available data provides the opportunity for investigation of practically any phenotype quickly, efficiently and in thousands of people increasing the chance of identifying associations which would not be possible in clinical trials (Bycroft et al., 2018; Conroy et al., 2019). Furthermore, in most cases, there is no under-representation of females due to increased willingness to participate meaning that sex-specific effects, which would be unattainable in clinical trials, can be investigated alongside potential for other dichotomising (e.g. by smoking status or body mass index) (Carey et al., 2017; L. Y. Liu et al., 2012; Randall et al., 2013). Genetic evidence can be used throughout the drug development process to anticipate the potential to be efficacious and the risk of unintended effects potentially reducing attrition rates and saving resources (Plenge et al., 2013). The identification of relevant genetic markers has been predicted to reduce the number of candidates entering clinical development enabling a higher fraction of R&D budget to later clinical phases and has potential to inform further research using patient-level clinical data from phase III trials (Hurle et al., 2016).

1.2.3 Genome-wide association studies

The development of high-throughput sequencing platforms has enabled large scale genome-wide association studies (GWAS) where single nucleotide polymorphisms (SNP) are tested for statistical associations with quantitative traits and diseases (Rohde et al., 2018). As of December 2021, 325,538 genome-wide significant (GWS) ($P \leq 5 \times 10^{-08}$) variant-trait associations in 5,527 publications were reported in the GWAS Catalog (Buniello et al., 2019). The majority of genetic variants identified through GWAS are common with a minor allele frequency (MAF) > 1% and have low-to-modest effect sizes (**Figure 1.4)** (Shuquan Rao et al., 2021). As the cost of sequencing technologies continue to decrease and larger, more extensive cohorts or biobanks emerge, the power to detect associations of smaller effect sizes and the ability to detect rare genetic variants with larger effect sizes through whole-genome or exome sequencing will increase (**Figure 1.4)** (Visscher et al., 2017).

GWAS primarily rely upon the principle of linkage disequilibrium (LD). LD describes the extent to which the allele of one SNP is correlated with the allele of another SNP within a given population as a result of population size, natural selection, mutations and recombination rates (Bush & Moore, 2012). Neighbouring genetic variants tend to be highly correlated and inherited together (Bush & Moore, 2012; Flister et al., 2013; Hormozdiari et al., 2014). Not all SNPs in high LD will reach formal levels of significance ($P \leq 5x10^{-08}$), but those that do can be used to identify the causal SNP. GWAS is often carried out in the context of a consortium where GWAS summary statistics from multiple independent studies are combined through meta-analysis to increase sample size and subsequent power (Uffelmann et al., 2021). Meta-analysis can be performed using a fixed-effects or a random-effects model assuming equal variance or testing for heterogeneity, respectively (Zeggini & Ioannidis, 2009). Meta-analysis improves power by increasing sample size or testing more variants identified through genotyping on different platforms. Meta-analysis also overcomes issues associated with data protection and allows for more precise effect estimates and significance levels as the results from each cohort can be weighted by sample size or the inverse variance weighted (IVW) method (Uffelmann et al., 2021; Zeggini & Ioannidis, 2009). GWAS meta-analyses have already proved useful in improving our understanding of complex diseases and quantitative traits through identification of novel genes and pathways (Frayling et al., 2007; Hirschhorn, 2009; Nikpay et al., 2015; Pulit et al., 2018). The utility of GWAS in clinical applications has also been highlighted. First, GWAS associations have advanced the prediction of individual risk of disease improving patient outcomes through early detection, prevention or treatment by use of genetic risk scores (GRS) (Tam et al., 2019; Uffelmann et al., 2021). Second, GWAS have improved disease classification and subtyping e.g. in the diagnosis of maturity-onset diabetes of the young (MODY) (Owen et al., 2010; Thanabalasingham et al., 2011). Third, GWAS have optimised treatments based on genotypes informing drug selection and preventing adverse effects (Giacomini et al., 2017; Muir et al., 2014; Tanaka et al., 2009). Fourth, GWAS have identified subgroups (e.g. ethnicities or genders) which may be at an increased risk of developing certain diseases (Tam et al., 2019; Visscher et al., 2017). These few examples highlight the potential of GWAS findings in improving our understanding of the underlying genetic architecture and biology of disease and positively impacting future disease prevention and treatment as sample sizes increase further and additional associations are identified.



Effect estimate
High
Intermediate
Modest
Low
30.0
3.0
1.5
1.0
Highly penetrant rare variants
Common variants with large effects
Less common variants with moderate effects
Rare variants with small effects
Common variants with small effects identified by GWAS
0.001
0.005
0.05
Very rare
Rare
Low frequency
Common
Allele frequency

Figure 1.4: Genetic predisposition and architecture of complex traits. Genetic variants associated with complex traits have a spectrum of effects based on allele frequency. GWAS typically identify common genetic variants (allele frequency > 0.01) with small effect sizes (bottom right) but can also identify common variants with large effect sizes (top right). Rare genetic variants (allele frequency < 0.005) often have much larger effects, but are harder to identify and rely upon alternative methods such as whole-genome or whole-exome sequencing (top left). These rare, large-effect variants are often the cause of Mendelian disorders. Rare variants of small effect size are hard to identify genetically (bottom left). The majority of genetic associations identified lie on the diagonal denoted by the dotted lines. Figure adapted from Bush & Moore (2012).

1.2.4 Using association data to investigate long-term effects of therapeutic modulation.

Several previous studies have shown the utility of using genetic variants, identified through GWAS, as a tool to identify causal associations between the drug target and intended therapeutic indication and potential unintended effects to inform possible drug safety (Gill et al., 2019; Lotta et al., 2016; Nelson et al., 2015; Nguyen et al., 2019; Plenge et al., 2013; Scott et al., 2016; Swerdlow et al., 2015). For example, genetic variants mimicking glucose-lowering GLP1R agonists are associated with lower glucose levels and decreased risk of type 2 diabetes as expected and are not associated with excess cardiovascular risk indicating the treatments likely safe (Scott et al., 2016), whilst variants mimicking low-density lipoprotein (LDL)-lowering agents (e.g. statins) are associated with lower LDL levels as expected but also higher risk of type 2 diabetes, providing an insight into the potential adverse effects of these LDL-lowering treatments (Lotta et al., 2016). These studies have altered the causal relevance of some biomarkers in relation to disease and highlight the importance of genetic studies in aiding prediction of drug safety and efficacy to inform drug development.

1.2.5 Mendelian Randomisation

Genetic variants lying within or nearby the gene encoding the drug target are most likely to have functional impact on the protein product. These genetic variants can be used as unconfounded, unbiased proxies for pharmacological action through the Mendelian Randomisation (MR) principle (Davey Smith & Hemani, 2014; Schmidt et al., 2020; Swerdlow et al., 2016; Walker et al., 2017). MR is an analytical method analogous to a randomised control trial (RCT) (**Figure 1.5A)**. MR relies upon the principal that if a modifiable exposure (e.g. a biomarker) is causal for disease, then a genetic variant associated with or mirroring the biological effects of the exposure will also be associated with the disease (Burgess et al., 2012) (**Figure 1.5A)**. MR relies on the identification of genetic variants associated with the exposure trait, i.e. the biomarker or drug target, and is based upon three main assumptions: 1) the genetic variant is associated with the exposure, 2) the genetic variant is not associated with confounders of the exposure-outcome relationship, and 3) the genetic variant only associates with the outcome through the exposure trait (**Figure 1.5B)** (Bennett & Holmes, 2017; Davies et al., 2018; Richmond & Davey Smith, 2021). Genetic variants can be used as instruments to test for causality between the exposure and outcome trait providing an estimate of long-term effects (**Figure 1.5B)** (reviewed by myself and Tim Frayling in (Frayling & Stoneman, 2018)). The causal estimate between exposure and outcome is calculated by dividing the SNP-outcome association by the SNP-exposure association which in the presence of a single genetic instrument is the Wald ratio (**Figure 1.5B)** (Burgess, Small, et al., 2017).

Similar to RCTs, MR exploits the power of randomisation through the random allocation of genetic variants at conception. The genetic variants randomly divide the study population into, on average, two identical groups apart from the levels of the exposure under investigation (**Figure 1.5A)** (Davey Smith & Ebrahim, 2003). The differences observed in the outcomes between the two groups can therefore be inferred to be a result of lifetime differences in exposure levels providing the MR assumptions are met (Ference et al., 2021). Compared to RCTs, MR limits participant risk, makes use of already available data decreasing cost, and provides an estimate of long-term effects (Frayling & Stoneman, 2018; Lawlor et al., 2008). MR can also overcome several types of confounding that may exist in RCTs (Bennett & Holmes, 2017; Richmond & Davey Smith, 2021; Smith & Ebrahim, 2005). First, as genetic variants are fixed at conception, they are non-modifiable and therefore reduce the risk of reverse causation (Lawlor et al., 2008; Richmond & Davey Smith, 2021). Second, due to inheritance of one trait being independent of inheritance of another trait, genetic variants should not be influenced by confounding (Smith et al., 2005). Third, genetic variants indicate lifetime differences in the exposure; associations will therefore not be attenuated by measurement error decreasing risk of regression dilution bias (Bennett & Holmes, 2017). Fourth, genetic variants reduce the risk of selection bias as they are unlikely to be influenced by how participants are selected (Smith & Ebrahim, 2005).

Careful consideration is needed when selecting genetic variants for proxies of drug treatments as violation of the MR assumptions can result in bias and unreliable estimates of causality (**Figure 1.5B-C)** (Bennett & Holmes, 2017;

Davey Smith & Ebrahim, 2003; Frayling & Stoneman, 2018; Nelson et al., 2015; Plenge, 2016). The presence of pleiotropy, a term used to describe associations where a genetic variant directly influences traits other than just the risk factor under investigation, limits the ability to make accurate casual inferences (**Figure 1.5A)**. Pleiotropy may be the result of correlation between the genetic variant being used and a nearby variant that alters another trait directly **(Figure 1.5C)** (VanderWeele et al., 2014). Pleiotropy can be graphically assessed and violation may not always be problematic (Sheehan et al., 2008). Another potential problem in MR studies is population stratification. Population stratification occurs when allele frequencies vary between subgroups of the background population who also vary in disease risk (Frayling & Stoneman, 2018). This problem can cause spurious associations and would occur in metabolic disease, if, for example, people of South Asian ancestry were in the same study as people of European ancestry, and these differences were not accounted for. Any allele that was more frequent in South Asians would be more likely spuriously associated with metabolic disease because of the higher frequency of these conditions in South Asians. However, population stratification is well-controlled using approaches such as genomic relationship matrices (GRM) to account for close and distant relatedness (Loh et al., 2015).

Several MR methodological advances are used to reduce risk of bias and increase chances of obtaining true causal estimates. Initial MR studies used a single variant to infer causality between a modifiable phenotype and outcome in a single sample (Frayling & Stoneman, 2018; Sheehan et al., 2008). However, the increasing availability of data from large-scale GWAS means multiple genetic variants can be combined into a GRS and used as an instrument. Multiple variants increase the specificity of the genetic instrument, minimise the risk of weak instrument bias, increase power and provide a more precise causal estimate (Burgess & Thompson, 2010; Richmond & Davey Smith, 2021). Due to the increasing number of consortia, summary-level data can be obtained from separate GWAS studies on the exposure and outcome trait (Sheehan & Didelez, 2019). This method, known as two-sample MR, further enhances power as the SNP-exposure and SNP-outcome associations can be taken from the largest available datasets and independent populations (Burgess et al., 2015; Sheehan & Didelez, 2019). MR tests of causality usually include sensitivity analyses to help verify the validity of the genetic instruments and the estimated causal associations (Bowden et al., 2015; Hemani, Bowden, et al., 2017; Hemani, Tilling, et al., 2017). These sensitivity analyses include Egger regression (Bowden et al., 2015), weighted-median methods (Bowden et al., 2016) and Steiger filtering (Hemani, Bowden, et al., 2017; Hemani, Tilling, et al., 2017). Steiger filtering is used to verify the validity of the instruments and reduces the risk of reverse causation by limiting variants to those with a greater effect on the exposure than the outcome (R^2 [exposure] > R^2 [outcome]) (Hemani, Bowden, et al., 2017; Hemani, Tilling, et al., 2017). Egger regression and weighted-median methods assess the validity of the causal association. Egger regression tests for pleiotropy and assumes that genetic variants more strongly associated with the exposure should more reliably estimate the causal effect on the outcome than weaker genetic variants (Bowden et al., 2015; Burgess & Thompson, 2017). Weighted-median assigns more weight to more precise genetic variants and assumes over 50% of the genetic variants are valid instruments (Burgess, Bowden, et al., 2017). Both these sensitivity analyses are robust to weaker assumptions than standard MR. For example, they can include, and account for the effects of, some variants that are not specifically influencing the exposure trait (Bowden et al., 2015).

Although MR cannot be used to replace RCTs, it can be used as a complementary approach to provide evidence on which drug targets to pursue, for what indication and evidence of drug safety or efficacy supporting clinical trial data (Ference et al., 2021). Several studies have already corroborated the utility of MR (Ference et al., 2016; Schmidt et al., 2017, 2020).

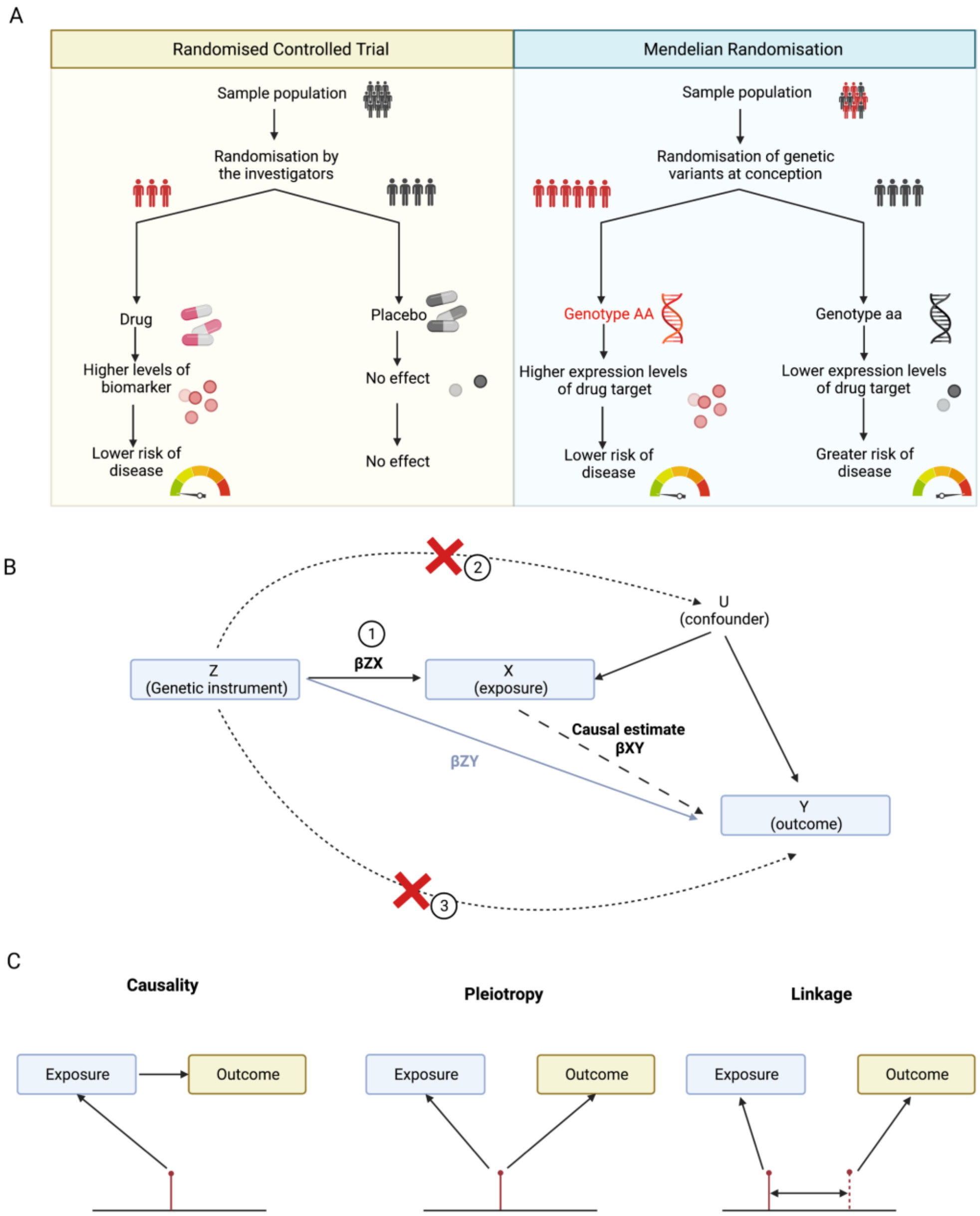


Figure 1.5: The principal of Mendelian Randomisation analysis. ***A:*** *Mendelian randomisation (MR) analysis is analogous to a randomised controlled trial (RCT) in that genetic variants are randomly allocated at conception similar way to random allocation of a drug treatment in RCTs. Genetic variants associated with expression levels of a drug target can be used to mimic the effects of a drug to predict the risk of adverse effects.* ***B:*** *MR uses genetic variants as instruments to assess the causal association between an exposure and outcome of interest. MR uses association data from published studies (ZX and ZY) to estimate an overall causal effect. The causal estimate (XY) is calculated by dividing the effect of the variant on the outcome (ZY) by the effect of the variant on the exposure (ZX). MR is based upon three assumptions (numbered in the diagram). (1) The genetic variant is associated with exposure. (2) The genetic variant is not associated with confounders of the exposure-outcome relationship. (3) The genetic variant is only associated with the outcome through the exposure trait.* ***C:*** *Three possible explanations for the observed association between an exposure and outcome through genetic variation; causality (measured by MR) where the effect of the variant on the outcome is mediated through the exposure, pleiotropy where the genetic variant has direct effects on the exposure and the outcome and linkage where there are two variants in high LD - one variant alters the exposure whilst the other variant alters the outcome. Created with BioRender.com.*

1.2.6 Phenome-wide association studies

Another method which is increasingly being employed to characterise the therapeutic profile of drugs is a phenome-wide association study (PheWAS). PheWAS is a cross-phenotype association approach that investigates the impact of genetic variants across a broad range of phenotypes (Verma & Ritchie, 2017). PheWAS infers an association in the reverse direction of GWAS by selecting a genetic variant of special interest, i.e. with known functional impact or prior disease association, and testing its association with any phenotype (Ye et al., 2015). PheWAS has the advantage that it allows investigation of known comorbidities of certain diseases, effects of environmental exposures, effects at particular life-stages and causal routes to elucidate mechanism (Bush et al., 2016; Verma & Ritchie, 2017). PheWAS offer the potential to 1) identify the risk of unwanted effects or secondary diseases that may not be considered as the primary area for concern in RCTs, 2) highlight safety concerns associated with drug targets early on in the drug development process 3) identify additional indications for disease expansion or drug repurposing and 4) improve our understanding of biological mechanism of action (Denny et al., 2016; Hebbring, 2014; Pulley et al., 2017; Robinson et al., 2018). Additionally, PheWAS can reveal violations of the exclusion restriction assumption (where a genetic variant is only associated with an outcome through the exposure trait [vertical pleiotropy]) and the independence assumption (where a genetic variant is not associated with confounders of the exposure-outcome association) made during MR (Davies et al., 2018; Richmond & Davey Smith, 2021; Sheehan & Didelez, 2019). PheWAS can be used to detect both horizontal and vertical pleiotropy by identifying associations between the genetic instrument and additional traits which may act through the same pathway (vertical pleiotropy) or different pathways (horizontal pleiotropy) (Hebbring, 2014; Pendergrass & Ritchie, 2015; Richmond & Davey Smith, 2021). Understanding pleiotropic effects and thus the shared genetic aetiology of many diseases can provide new insights into underlying pathophysiology highlighting potential new treatment strategies minimising research costs (Hebbring, 2014). However, in some cases, if power is inadequate or a trait on the pleiotropic pathway is absent, truly horizontal pleiotropic variants may be missed (Richmond & Davey Smith, 2021).

1.2.7 Limitations of using association data as proxies for therapeutic action.

Several problems can arise when using GWAS data to select genetic instruments for use to proxy pharmaceutical effects and test for causal associations and detect unintended effects (Porcu et al., 2019). First, it is often difficult to distinguish the true causal variant driving the association identified through GWAS due to underlying LD patterns (Flister et al., 2013; Hormozdiari et al., 2014; Pers et al., 2015; Schaid et al., 2018). Second, most genetic variants identified through GWAS do not directly affect the coding sequence due to the genomic location within non-coding regions (Dixon et al., 2007; Nica et al., 2010). These variants can lie within genomic regulatory elements, overlap promoters, enhancers or open-chromatin regions, and may affect gene expression by altering transcription factor binding (Lichou & Trynka, 2020). Non-coding variants can be highly cell-type, context- and disease-specific, and can bind to numerous transcription factors influencing gene expression in a microenvironment-dependent context (Broekema et al., 2020). Third, disease associated loci often contain multiple genes making it difficult to determine which gene is disease-relevant and is affected by the identified variant (Cano-Gamez & Trynka, 2020). These complexities make it challenging to not only determine which variant is the true driver of disease but to also interpret how GWAS loci influence their associated trait, and has hampered direct interpretation and clinical application of GWAS findings (Broekema et al., 2020; Cano-Gamez & Trynka, 2020).

1.3 Statistical validation of genetic variants

To address the complexities associated with interpreting GWAS findings and to obtain clearer biological insights, additional statistical genetic approaches have been developed to provide more confidence of the true casual variant and target gene driving the association (Broekema et al., 2020). These approaches, such as fine mapping, colocalisation, or quantitative trait loci (QTL) analysis, help refine the causal variant and gene (Benner et al., 2016; Giambartolomei et al., 2014; Nica et al., 2010; Nicolae et al., 2010; Porcu et al., 2019; Wallace, 2020). These post-GWAS approaches have importantly aided in understanding the link between the causal variant, target gene and molecular phenotype and in interpreting the biological impact on disease (Broekema et al., 2020; Shuquan Rao et al., 2021).

1.3.1 Fine mapping

Fine mapping aims to define the causal variant(s) and gene(s) responsible for a given trait and interpret their likely biological impact (Broekema et al., 2020). Fine mapping assumes there is at least one causal variant and uses a list of associated SNPs from GWAS to identify regions of interest (Schaid et al., 2018). The LD structure and genes mapped to each region are explored and statistical methods, primarily the Bayesian framework, are employed to determine the most likely causal SNPs and genes (Benner et al., 2016; Newcombe et al., 2016). These selected SNPs are evaluated for their likely function based on publicly available annotation data (Schaid et al., 2018). The majority of fine mapping approaches assume there is one true causal variant in each locus, which is often not the case due to additive or epistatic effects (Broekema et al., 2020). Due to the underlying LD structure responsible for highly correlated variants, fine mapping is challenging (Gao Wang et al., 2020). Novel approaches building upon the Bayesian framework have been developed to account for multiple causal variants and to assess the uncertainty of which highly correlated variants to select (Gao Wang et al., 2020). Several tools exist to implement fine mapping, such as FINEMAP and SuSIE (Benner et al., 2016; Gao Wang et al., 2020). These tools are continuously being refined and additional tools developed as statistical methods improve. Fine mapping is computationally fast, requires summary-level data and is useful in prioritising downstream functional studies (Hutchinson et al., 2020; Schaid et al., 2018).

1.3.2 Expression Quantitative trait loci analysis

As mentioned above (in **1.2.7**), the majority of identified GWAS signals reside in the non-coding regions of the genome (Nica & Dermitzakis, 2013). Several studies have shown these variants to be enriched in *cis*-regulatory elements regulating gene expression through altering transcription, splicing, chromatin accessibility, and mRNA stability (Gallagher & Chen-Plotkin, 2018). Trait-associated SNPs have been proposed to three times more likely be associated with messenger RNA (mRNA) expression suggesting that many associations are driven through gene expression changes (Hernandez et al., 2012; Nicolae et al., 2010; Porcu et al., 2019). SNPs associated with mRNA expression levels are termed expression quantitative trait loci (eQTL) and have already proved useful in refining GWAS results by providing additional supporting evidence of the target gene in particular tissues (Hormozdiari et al., 2016; Lawrenson et al., 2015; Nicolae et al., 2010). Standard eQTL analysis is performed by testing the association between a SNP and mRNA expression levels obtained from microarrays. The analysis can be performed proximally or distally to the gene without any prior knowledge of *cis*- or *trans*-acting regulatory regions (Albert & Kruglyak, 2015; Nica & Dermitzakis, 2013). eQTLs affect gene expression in *cis* or *trans*; the definition of *cis* is arbitrary but typically includes variants lying within 100 kilobases (kb) of the gene affected by the eQTL (Cookson et al., 2009). The majority of already identified eQTLs are *cis*-acting but this may be attributable to lack of power and computational complexity to detect *trans*-eQTLs across the genome (Cookson et al., 2009; Nica & Dermitzakis, 2013). As larger studies become available, the number of genes with eQTLs is expected to increase and the power to detect *trans*-eQTLs will increase (Nica & Dermitzakis, 2013). *Cis*-acting eQTLs predominantly have stronger effects than *trans*-acting eQTLs, despite the number of *trans*-eQTLs predicted to be greater (Hernandez et al., 2012). eQTL analyses can be performed in any tissue with the most common being liver, kidney, brain, blood and subcutaneous adipose tissue enabling the identification of cell-type specific and disease-relevant effects (Hernandez et al., 2012). Several eQTL datasets are publicly available, e.g. through the GTEx portal (https://gtexportal.org/home/), meaning it is relatively easy to interrogate the effects of GWAS signals in cell-types of interest; it is important to note that the majority of eQTL data is from bulk tissue samples as opposed to individual cell-types so some eQTLs may be missed (Choi et al., 2020). Previous studies have shown that eQTL effects are often detected in the expected disease-relevant cell-type providing validation of the most likely causal variant and target gene identified through GWAS and the link between the variant and the biological process (Albert & Kruglyak, 2015; Gallagher & Chen-Plotkin, 2018; Lawrenson et al., 2015; Raj et al., 2014). However, it is important to test the effects in a range of tissues as eQTLs can change dynamically during differentiation and in response to certain stimuli and some variants may be pleiotropic affecting different genes in different tissues (Albert & Kruglyak, 2015; Nica & Dermitzakis, 2013). Additional datasets measuring the association between genetic markers and additional molecular phenotypes, such as protein expression (pQTL) or DNA methylation (mQTL) levels, have been curated to further investigate how variants exert their effects and contribute to phenotypic changes (Albert & Kruglyak, 2015; Cookson et al., 2009). These datasets help bridge the gap between gene and phenotype by eluding to cell-type specific effects and provide an immediate understanding of the biological mechanisms driving associations (Cookson et al., 2009).

1.3.3 Colocalisation

Colocalisation analysis is used to integrate multiple association data, such as GWAS and eQTL analysis, to further nominate the most likely target gene and improve understanding of the molecular basis of these associations (Giambartolomei et al., 2013). Colocalisation explores whether two traits (e.g. disease and gene expression) are driven by the same causal variant in a given genomic region which may prove useful in understanding how variants lead to different disease risks (Wallace, 2021). Two main methods exist for performing colocalisation; summary-data based MR (SMR) and coloc. SMR, an extension of MR, tests if the effect of a variant on a trait (obtained from GWAS) is mediated by gene expression and uses heterogeneity measures to filter associations and detect pleiotropy (Zhu et al., 2016). SMR uses the Wald ratio (β [SNP-outcome] / β [SNP-exposure]) to calculate the causal estimate (βxy) and performs a heterogeneity in dependant instruments (HEDI) test to detect linkage from pleiotropy (Zhu et al., 2016). Consistent causal estimates imply a single shared causal SNP due to a greater likelihood of pleiotropy, whilst inconsistent causal estimates imply a greater likelihood of linkage suggesting distinct causal variants for the two traits (Zhu et al., 2018). SMR is unable to distinguish between linkage or pleiotropy if the two causal variants are in perfect LD (Hannon et al., 2017). Coloc employs a Bayesian framework by considering all possible configurations of causal variants for the two traits. Coloc utilises summary statistics and calculates an easily interpreted posterior probability (the probability of an event occurring after taking into account prior information) in support of each hypothesis (Giambartolomei et al., 2013, 2014). The five mutually-exclusive hypotheses tested are:

H_1: No association with either trait in the region
H_2: Association with trait one only
H_3: Association with trait two only
H_4: Association with both traits, but there are two independent SNPs
H_5: Association with both traits, and the same single causal SNP is shared

Initially, coloc assumed only one causal variant within a locus for each trait but has recently been updated to allow for multiple causal variants within a region (Wallace, 2021). Unlike SMR, coloc assesses a pair of causal variants at a time, avoids MR assumptions and through prior probabilities incorporates any expectation that the causal variants are likely shared (Wallace, 2020, 2021). Colocalisation analysis is essential for functional follow-up, for validating genetic studies and for identifying tissue-specific signals (Giambartolomei et al., 2013).

1.4 Functional studies to further validate genetic findings

An association between a variant at a genomic locus and a trait is not directly informative with respect to the underlying mechanism driving the phenotypic difference (Visscher et al., 2017). Functional studies are, therefore, still essential in bridging the gap between sequence and consequence. In order for genetic variants to aid the drug development process, functional studies of variants suspected to predispose to disease are necessary for a clearer and better understanding of the physiologically relevant, cell-type specific and microenvironment-dependant effects, for the correct interpretation of results in clinical diagnosis and the elucidation of therapeutic targets (Bonjoch et al., 2019; Cano-Gamez & Trynka, 2020). Functional approaches can provide additional validation of the links between the genomic architecture and phenotype improving understanding of cellular function and underlying biological mechanisms in health and disease as well as downstream effects of gene perturbations (Bonjoch et al., 2019; Shuquan Rao et al., 2021).

1.4.1 Traditional functional approaches to validate genetic findings

Several assays are well-established and widely used to explore the effects of regulatory genetic variants in cell-lines. One assay is the cell culture-based reporter assay where the candidate regulatory variant is cloned into a physiologically relevant position with respect to the reported gene and transfected into a cell-type of choice (Gallagher & Chen-Plotkin, 2018). The effect of different alleles on the reported gene is then compared. However, this technique can be laborious due to having to test constructs one-by-one. Massively-parallel reporter assays (MPRAs) have therefore been developed which can test the effect of thousands of variants in a single experiment (Choi et al., 2020). Despite these assays proving useful in determining the function of regulatory variants, MPRAs are not always reproducible due to transcriptional noise, are not representative of the true native genomic context as the variant is present in plasmid DNA so could produce false-positive or false-negative results, and unavoidable small differences in concentrations of the transfected plasmid DNA make inference of the effect difficult (Gallagher & Chen-Plotkin, 2018; Inoue & Ahituv, 2015). Alternative technologies for targeted gene knock-down include RNA interference which is fast and inexpensive but is limited in terms of not always being complete, risk of unpredictable off-target effects and short-term inhibition of gene function (Gaj et al., 2013).

1.4.2 Gene-editing

Gene-editing has come to the forefront of epidemiological molecular approaches and recent technological advancements have made functional validation of genetic findings more straightforward. Gene-editing enables investigation of variant effects in a physiologically relevant context and has increased chances of detecting true differences compared to previously used techniques (Bonjoch et al., 2019). Gene-editing can provide additional evidence to support the role of genetic variants in controlling expected gene expression and validate these variants as proxies for therapeutic action in MR (Lichou & Trynka, 2020; H. Wang et al., 2016; L. Yang et al., 2013). Recent advancements have made it possible to perform whole gene knock-out studies and SNP knock-in models relatively easily enabling validation of whether an expected allele alters gene expression, and to what extent, in relevant cell-types

(Courtney et al., 2016; H. Li et al., 2020; J. Lin & Musunuru, 2018; Okamoto et al., 2019; G. Zhao et al., 2018).

Since the identification of modifiable nucleases, there has been a rapid rise in the development and optimisation of gene-editing technologies enabling manipulation of any gene in a wide-range of *in-vivo* and *in-vitro* investigations (Gaj et al., 2013). Gene-editing relies upon the targeted introduction of a double-stranded break (DSB) by the nucleases which initiates one of the two major DNA repair mechanisms in mammalian cells; non-homologous end joining (NHEJ) or homology directed repair (HDR) resulting in targeted gene disruption, modification or insertion **(Figure 1.6)** (H. Li et al., 2020). In mammalian cells, DSBs are most commonly repaired by the efficient, simpler, error-prone NHEJ pathway which results in the loss of nucleotides from the ends of the DSBs (Chapman et al., 2012; Lino et al., 2018). This can result in the formation of large insertions or deletions (indels) which, if present in the coding sequence, can induce frameshift mutations and subsequent nonsense-mediated decay **(Figure 1.6)** (Chu et al., 2015; Hsu et al., 2013; Ran, Hsu, Wright, et al., 2013). Alternatively, the HDR pathway can be initiated in the presence of an exogenous repair template enabling the introduction of precise modifications such as specific mutations or desired insertions **(Figure 1.6)** (Hockemeyer et al., 2009).

The first targeted nuclease gene-editing technology was the zinc-finger nucleases (ZFNs) which comprise a specific trinucleotide DNA binding domain complementary to a site on the target DNA (Hockemeyer et al., 2009; Urnov et al., 2005, 2010). Multiple ZFNs can be combined to increase specificity (Carroll, 2011). The ZFNs are fused to the *FokI* endonuclease which introduces the site-specific DSB (Gaj et al., 2016). ZFNs function as dimers meaning two constructs need designing (one targeting the sense strand and one targeting the antisense strand) increasing time and resources (Carroll, 2011; H. Li et al., 2020). The major limitations of ZFNs are the poor targeting density, the risk of off-target mutations and the difficulty of constructing the ZFNs **(**Table 1.1**)** (Gabriel et al., 2011; Gaj et al., 2013; H. Kim & Kim, 2014; Pattanayak et al., 2011; A. M. Singh et al., 2015). The discovery of the transcription activator-like effector (TALE) protein led to the development of the next generation of gene-editing technologies called transcriptional activator-like effector nucleases (TALENs) (Becker & Boch, 2021; Boch et al., 2009). TALE proteins are naturally secreted by the Xanthomonas bacteria and comprise a DNA binding domain consisting of a series of highly conserved 33-35 base-pair (bp) domains which each recognise a single bp (Gaj et al., 2013, 2016; H. Kim & Kim, 2014). The specificity of TALEs to a single bp is determined by the hypervariable amino acids at position 12 and 13 (Boch et al., 2009; Joung & Sander, 2013). Similar to ZFNs, the TALEs can be linked together to recognise specific sites of the DNA and are typically fused with the *FOKI* endonuclease to enable the introduction of a DSB (Joung & Sander, 2013). TALEs can also be fused with site-specific recombinases and transcriptional activators to achieve targeted genomic rearrangements and regulated gene transcription, respectively (Becker & Boch, 2021). TALENs offer greater flexibility than ZFNs due to the single-base recognition and have been reported to be more specific and less toxic (Gaj et al., 2013, 2016). However, they are again limited by the difficult and time-consuming construction alongside the difficult cell delivery due to being large and needing a thymine (T) at the start of the binding site **(**Table 1.1**)** (Doudna & Charpentier, 2014; Lamb et al., 2013; H. Li et al., 2020; Nemudryi et al., 2014; A. M. Singh et al., 2015).

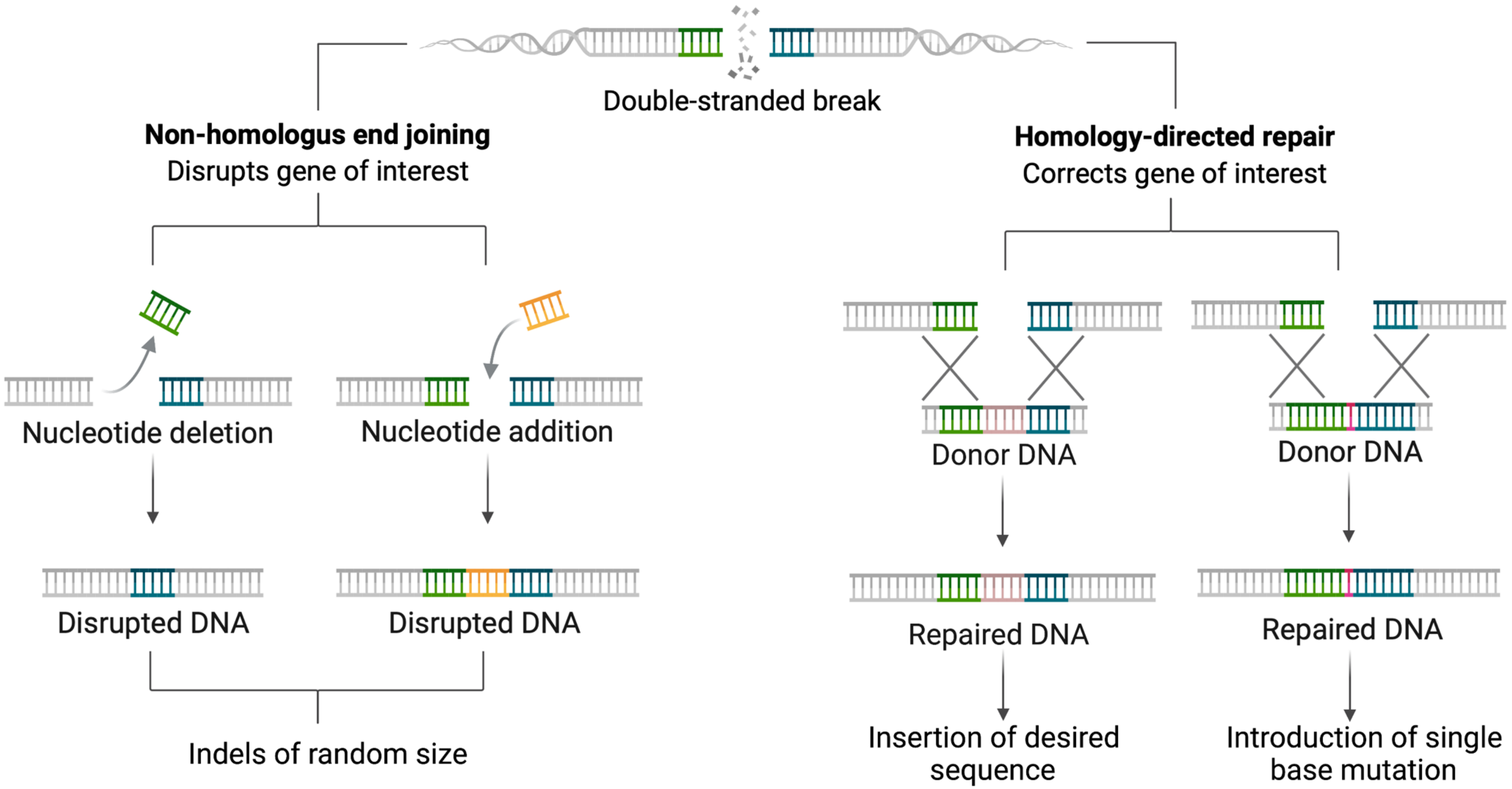


Figure 1.6: The two major DNA repair mechanisms in mammalian cells. *Upon the introduction of a double-stranded break (DSB) in the genomic DNA, one of the two DNA repair mechanisms is initiated. The non-homologous end joining (NHEJ) pathway is the most common pathway to be stimulated. NHEJ is error-prone and results in the random insertion or deletion of nucleotides surrounding the DSB. The deletion or addition of nucleotides can result in a frameshift mutation disrupting the coding sequencing through the introduction of a premature stop codon or the translation of a faulty protein. NHEJ often results in gene disruption and can lead to mRNA degradation. The alternative pathway is homology-directed repair (HDR). HDR is less common and relies upon the presence of a homologous sequence which is complementary to the flanking regions of the DSB. HDR results in the precise insertion of a desired genomic sequence or correction of a mutation. Created with BioRender.com.*

1.4.3 CRISPR-Cas9 gene-editing

The most recent addition to the gene-editing toolbox is the versatile CRISPR-Cas9 system (clustered, regularly interspaced, short palindromic repeats [CRISPR]–CRISPR-associated 9 [Cas9] protein) which enables the introduction of site-specific genomic DSBs by guide RNAs (gRNAs) (Mali et al., 2013; Ran, Hsu, Wright, et al., 2013). The CRISPR-Cas9 system has been derived from the most common type II system acquired by prokaryotes as a form of adaptive immunity protecting bacteria from viruses or phages through initiation of RNA-guided DNA cleavage (Gaj et al., 2016; Wiedenheft et al., 2012). Upon infection, viral DNA sections (termed spacers) are integrated into the CRISPR locus which are in turn transcribed into precursor CRISPR RNA (pre-crRNA) molecules alongside transcription of the *cas9* genes (Figure 1.7**A)** (Rath et al., 2015). The pre-crRNA is processed into crRNA by accessory factors where it can anneal to *trans*-activating crRNA (tracrRNA) to direct site-specific degradation of the foreign material by Cas9 (Figure 1.7**A)** (Jinek et al., 2012; Rath et al., 2015). Binding of the Cas9 endonuclease to the target DNA relies upon the presence of a 'seed' sequence (10-12 bp at the 3' end of the 20 bp sequence complementary to the target DNA) and a protospacer adjacent motif (PAM) upstream of the binding site (Jiang & Doudna, 2017). The PAM sequence, recognised by Cas9, is essential for DNA cleavage (Hsu et al., 2013). The CRISPR-Cas9 system has been simplified for use in gene-editing by incorporating only the Cas9 endonuclease and a single gRNA consisting of the essential tracrRNA and crRNA elements (Figure 1.7**B-D)** (Cong et al., 2013; Gaj et al., 2016). CRISPR-Cas9 can be efficiently engineered to target any site on the genomic DNA by modifying the ~20 bp crRNA of the gRNA providing the 20 bp are unique compared to the rest of the genome and is immediately adjacent to the PAM (Figure 1.7**C)** (Lino et al., 2018; X.-H. Zhang et al., 2015). The PAM sequence alters depending upon the Cas9 used; for the most popular *S. pyogenes* Cas9 (SpCas9), the PAM sequence is 5'-NGG-3' (Hsu et al., 2013; H. Li et al., 2020). Numerous resources have been developed to aid target site selection and the design of the most effective gRNAs by calculating on- and off-target scores using the latest algorithms to estimate the likelihood of introducing a DSB at the desired genomic locus making it user-friendly and relatively easy (C.-L. Chen et al., 2020; Concordet & Haeussler, 2018; Doench et al., 2014, 2016; Heigwer et al., 2014; Hsu et al., 2013). Additionally, the Cas9 and gRNA can easily and effectively be delivered into cells on the same (or separate) plasmids making the gene-editing platform advantageous over ZFNs and TALENs in terms of flexibility, robustness and ease of use **(**Table 1.1**)** (Cong et al., 2013; Mali et al., 2013). CRISPR-Cas9 has shown huge versatility and utility in modulating gene expression including genomic sequence alterations, epigenetic modifications, transcriptional modifications and multiplexing for disrupting multiple genes as well as functionally validating GWAS findings (Cong et al., 2013; Gilbert et al., 2013; Perez-Pinera et al., 2013; Pickar-Oliver & Gersbach, 2019; Qi et al., 2013; Vora et al., 2016). Alternative CRISPR-Cas9 systems have been identified and modifications made to the Cas9 enzyme, including conversion to a nickase or creation of a dead Cas9, which has extended the utility of CRISPR-Cas by enabling selective repression or activation, purification of target regions, precise modification of DNA or RNA through increased specificity and control of the DNA repair pathway alongside scalability for use in genome-wide screens (Adli, 2018; Doudna & Charpentier, 2014; Jiang & Doudna, 2017; Pickar-Oliver & Gersbach, 2019; Ran, Hsu, Lin, et al., 2013; Shuquan Rao et al., 2021; Rath et al., 2015; H. Wang et al., 2016; Tim Wang et al., 2014)**.**

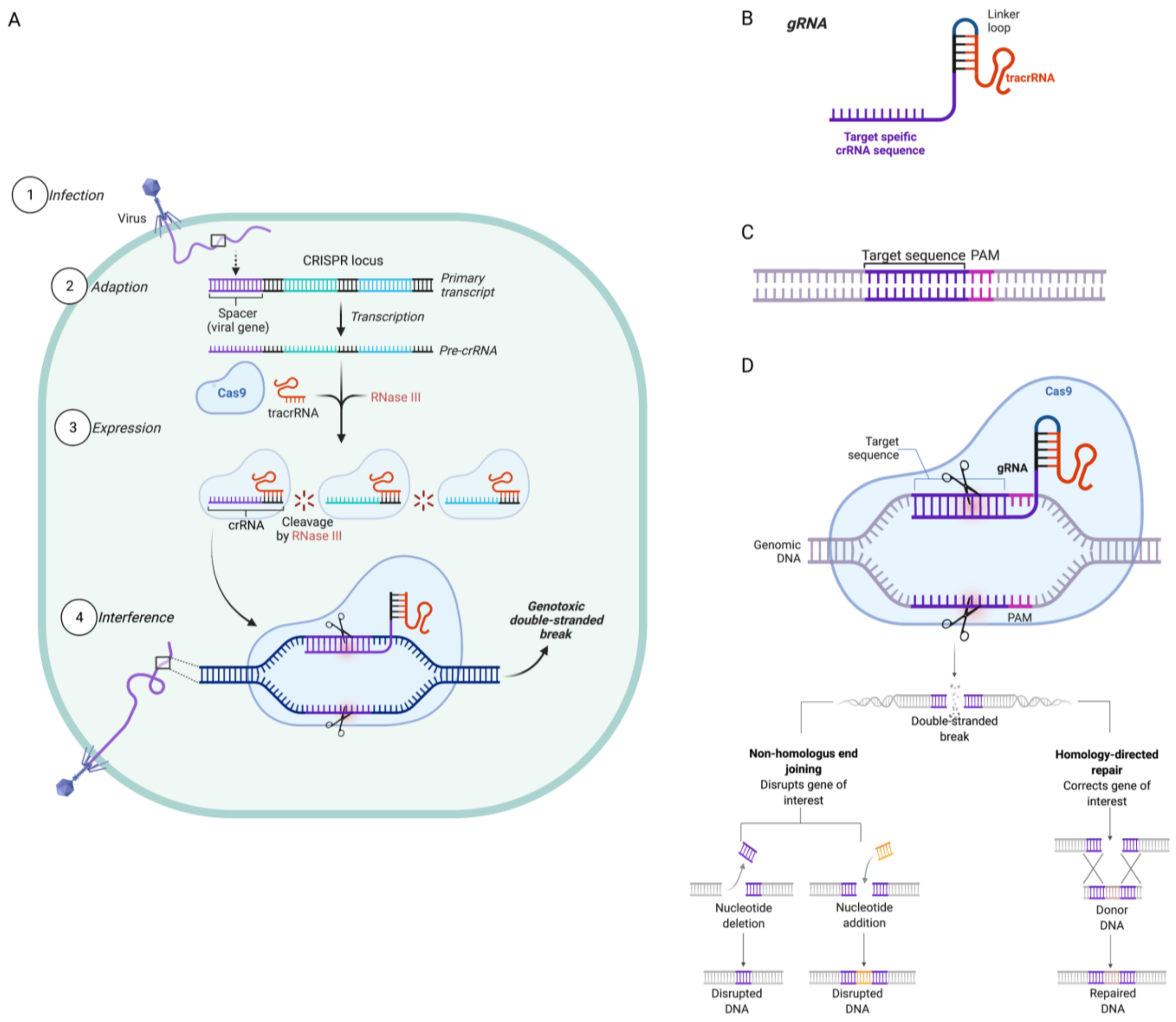


Figure 1.7: The CRISPR-Cas9 gene-editing system. ***A:*** *The CRISPR-Cas9 system forms the adaptive innate immune response in prokaryote cells. Upon infection of bacteria by viruses or phages (1), part of the foreign material becomes inserted into the CRISPR locus as a spacer (2). The CRISPR locus is then transcribed producing pre-crRNA (pre-CRISPR RNA) (3). The pre-crRNA is modified by adapter sequences and cleaved by RNase III to form separate mature crRNAs specific to each spacer. The mature crRNA can anneal tracrRNA (trans-activating crRNA) which targets the crRNA to the Cas9 proteins to form a ribonucleoprotein complex. If the bacteria is infected by the same virus, the Cas9-crRNA-tracrRNA complex recognises and binds to the viral sequence introducing a double-stranded break (DSB) in the viral DNA leading to degradation.* ***B:*** *The CRISPR-Cas9 system has been simplified and modified for use in gene-editing. The gRNA which can be designed to target virtually any region in the DNA consists only of the crRNA and the tracrRNA components.* ***C:*** *The gRNA is designed to target a unique 20 bp region of the target DNA. The target sequence must have a PAM sequence (5'-NGG-3' for S. pyogenesis Cas9) directly adjacent to it for successful targeting via CRISPR-Cas9. The PAM sequence is recognised by the Cas9 protein enabling Cas9-mediated cleavage.* ***D:*** *The CRISPR-Cas9 system is used in gene-editing techniques to enable the introduction of a DSB into the genomic DNA at the target site (complementary to the gRNA sequence). The DSB initiates either the non-homologous end joining (NHEJ) or homology directed repair (HDR) pathways in mammalian cells resulting in either the random insertion or deletion of nucleotides or the precise insertion of desired sequences. Created with BioRender.com.***Table 1.1: A comparison of the three approaches that have been developed for gene-editing through the use of modifiable nucleases.** *Scaling ranges from high (+++++) to low (+). ZFNs = Zinc finger nucleases; TALENs = Transcription Activator-Like Effector Nucleases; CRISPR = Clustered Regularly Interspaced Short*

Palindromic Repeats; TALE = Transcription Activator-Like Effector proteins; gRNA = guide RNA; bp = base-pairs; PAM = Protospacer Adjacent Motif; kb = kilobases;

	ZFNs	**TALENs**	**CRISPR-Cas9**
DNA targeting determinant	Zinc-finger proteins	TALE	gRNA
Length of DNA targeting determinant	18-36 bp	30 – 40 bp	23 bp (including PAM)
Requirements in target site	G-rich	Start with a T and end with an A	PAM sequence at the 3' end (5'-NGG-3' for *S. pyogenes* Cas9)
Design density	One per 100bp	One per 1 bp	One per 8 bp (for NGG)
Nuclease	*FokI*	*FokI*	Cas9
Size	2 x ~1 kb	2 x ~3 kb	4.3 kb
Cost	+++++	+++	++
Efficiency	+++	++++	+++++
Off-target effects	++	++	+++++
Design limitations	+++	+	++
Target design simplicity	+	+	+++++

Transfection difficulty	++	++++	+
Cytotoxicity	+++	+	+
Multiplex reactions	++	++	+++++

1.4.4 Single-base gene-editing using CRISPR-Cas9 and piggyBac^{TM} system.

The introduction of a precise mutation or desired gene insertion (e.g. the insertion of a fluorophore or selectable marker) is achievable by providing a homologous sequence on an exogenous pre-designed donor plasmid which upon transfection triggers homologous recombination (HR) (H. Li et al., 2020; Lino et al., 2018). The HDR repair template must contain the desired gene modification alongside additional homologous sequences (termed homology arms) up- and downstream of the target; the length of these homology arms is dependent upon the size of the desired edit (Au et al., 2019; Boel et al., 2018; Yusa et al., 2011). The standard approach of introducing homologous sequences is through electroporation (Sharan et al., 2009; Zwaka & Thomson, 2009). However, this approach is limited in terms of its efficiency with <10% of cells containing the modified allele (H. Kim & Kim, 2014).

Efficiency can be improved through the introduction of DSBs close to the desired gene-edit site by using genetically engineered nucleases, such as CRISPR-Cas9 (S. Liu et al., 2018; Maruyama et al., 2015; J.-P. Zhang et al., 2017). Traditionally, the simple and easy-to-construct singe stranded oligodeoxynucleotides (ssODNs) have been used as the donor template to initiate HDR after introduction of DSBs with CRISPR-Cas9. However, extensive screening of genome-edited cells is required due to the inability to perform drug selection and mutations are required in the PAM sequence to prevent re-cutting by Cas9 leaving behind unwanted marks in the genomic DNA which could affect transcriptional regulation of surrounding genes (Hendriks et al., 2015; A. M. Singh et al., 2016). The incorporation of a selection cassette, normally drug resistance, into the genome has therefore become the standard approach to improve the screening and/or isolation of correctly-modified cells (Ishida et al., 2018). Depending upon the technology used, the selection cassette is removed by Cre-loxP mediated recombination, site-specific nuclease mediated excision or *piggyBac^{TM}* transposon-based excision (Ishida et al., 2018). The benefit of combining CRISPR-Cas9 with the *piggyBac^{TM}* transposon system is that, unlike the former two, no genomic marks remain in the DNA after removal of the transposon resulting in seamless, footprint free modification (A. M. Singh et al., 2016; Gang Wang et al., 2017; Yusa et al., 2011). The *piggyBac^{TM}* system has already been widely used for transgenesis, engineering of pluripotent stem cells, gene therapy, production of recombinant proteins and the expression of multi-subunit protein complexes (Q. Chen et al., 2020; Z. Li et al., 2013; Schertzer et al., 2019; Yusa et al., 2009, 2021). Similar to all transposons, the *piggyBac*^{TM} system consists of a transposon and transposase. The *piggyBac*^{TM} transposon, which is held within an HDR template vector, contains a selection marker (often the puromycin [∆*puro*] resistance gene for positive selection and thymidine kinase [*tk*] gene for negative selection) facilitating the enrichment of correctly-modified cells (Yusa, 2013). The transposase enables the integration and removal of the *piggyBac*^{TM} transposon at any 'TTAA' site within the genomic DNA (Z. Li et al., 2013; Schertzer et al., 2019). These 'TTAA' sites are dispersed randomly throughout the genome at a rate of 1 every 246 bp (Yusa, 2013). The *piggyBac*^{TM} vector is designed to contain two homology arms (one with the desired gene-edit) complementary to the target DNA flanking the *piggyBac*^{TM} transposon **(Figure 1.8)** (Yusa, 2013). Upon site-specific cleavage of the genomic DNA at the target site by Cas9, the presence of the homology arms in the *piggyBac*^{TM} vector initiates HDR resulting in the incorporation of the desired gene-edit and the *piggyBac*^{TM} transposon into the genome at a 'TTAA' site nearby the target-site **(Figure 1.8)** (M. A. Li et al., 2013; Paquet et al., 2016; Yusa, 2013). A modified excision-only *piggyBac*^{TM} transposase is then used to remove the *piggyBac*^{TM} transposon from the genomic DNA **(Figure 1.8)** (X. Li et al., 2013). The excision-only transposase prevents potential random reintegration of the transposon elsewhere in the DNA (X. Li et al., 2013; A. M. Singh et al., 2016). After successful removal of the transposon, cells undergo a negative selection step killing cells retaining the *piggyBac*^{TM} transposon **(Figure 1.8)** (A. M. Singh et al., 2015). This results in relatively easy isolation of successful, footprint free, gene-edited clones (Fei Xie et al., 2014). The *piggyBac*^{TM} system has been found to prefer integration at promoter or exonic regions highlighting the potential for aiding single-base gene-editing and functional characterisation of GWAS signals that lie within the regulatory regions (M. A. Li et al., 2013). The only drawback with CRISPR-Cas9 and the *piggyBac*^{TM} system is the timeframe required for the construction of the HDR template and the selection steps. Despite this, fewer clones need screening to identify successfully gene-edited clones due to the selection steps and the actual 'hands-on' time is less (A. M. Singh et al., 2015). The combination of

CRISPR-Cas9 and *piggyBac*^{TM} provides one of the most robust, efficient and precise gene-editing approaches to achieve seamless single-base gene-edits and has already proved useful in functionally validating genetic variants as causal in controlling expected gene expression levels (S. Liu et al., 2018; Shuquan Rao et al., 2021; A. M. Singh et al., 2015, 2016; G. Zhao et al., 2018).

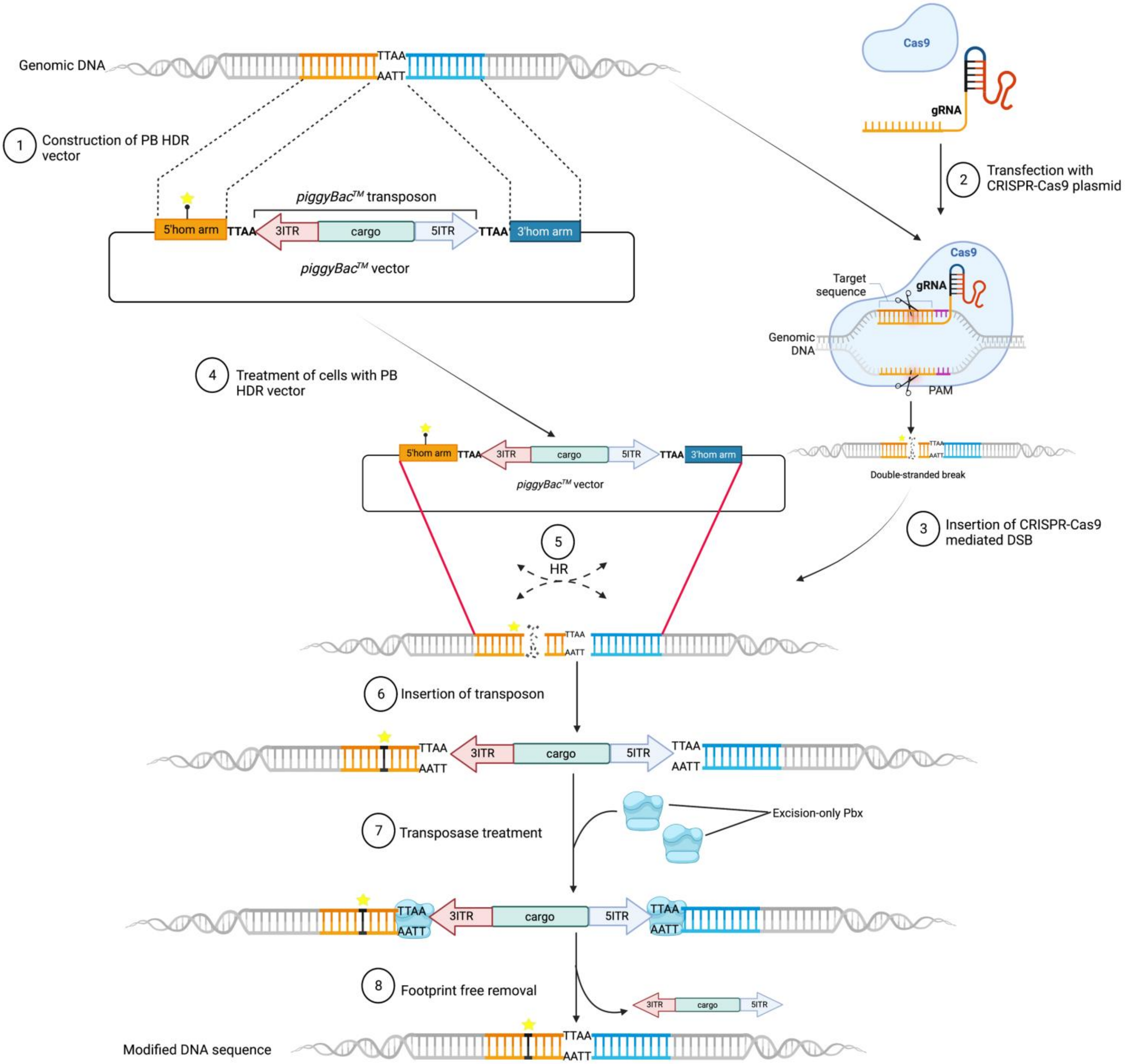


Figure 1.8: Site-specific footprint free gene-editing using CRISPR-Cas9 and the piggyBac^{TM} transposon system. *See figure legend on next page*

Figure 1.8: Site-specific footprint free gene-editing using CRISPR-Cas9 and the piggyBac^{TM} transposon system. ***(1)*** *Two homology arms (5’hom arm & 3’hom arm) flanking a ‘TTAA’ site nearby the target-site are designed with one containing the desired*

*gene-edit (whether that be a single base change or insertion of particular sequence [depicted here by the star]). The two arms are inserted into the piggyBac^{TM} vector either side of the piggyBac^{TM} transposon which consists of two inverted terminal repeats (5'ITR & 3'ITR) and cargo (normally positive and negative selection genes, e.g. ∆puro and tk. The ITRs are essential for recognition by transposases. (**2)** The CRIPSR-Cas9 targeting construct is designed as close as possible to the desired gene-edit site. **(3):** Upon transfection into cells, the CRISPR-Cas9 system introduces a site-specific double-stranded break (DSB) to the genomic DNA. (**4)** Cells are simultaneously treated with the piggyBac^{TM} vector containing the homology arms. (**5)** The presence of the homology arms in the piggyBac^{TM} vector stimulates DNA repair of the DSB via HDR resulting in homologous recombination (HR) between the genomic DNA and the homology arms replacing the genomic sequence with the desired gene-edit. (**6)** As a result of HR, the piggyBac^{TM} transposon is inserted into the genomic DNA at the nearby TTAA site enabling the positive selection of cells containing the transposon and potentially the desired gene-edit. (**7)** After positive selection, cells are treated with excision-only transposases (Pbx). (**8)** The transposases recognise the ITRs of the transposon resulting in footprint free removal of the transposon from the DNA sequence. Cells can then be put under negative selection to select for clones which no longer contain the transposon and have been successfully edited. Created with BioRender.com.*

1.5 Drug metabolism

Xenobiotic metabolism refers to the metabolic breakdown of foreign substances (called xenobiotics) that differ to an organism's normal biochemistry, such as a drug, by a living organism (Patterson et al., 2010). This process is carried out by a series of metabolic pathways controlled by specialised enzymes that modify the chemical structure of the foreign substance resulting in its breakdown and excretion from the body. Drug metabolism is divided into three main phases (Esteves et al., 2021). During phase I, reactive or polar groups are added to the xenobiotics by enzymes, in particular the cytochrome P450s (Omiecinski et al., 2011). Cytochrome P450s are the most important enzymes in xenobiotic metabolism as they can determine the rate xenobiotics are metabolised into inactive products (McGinnity & Grime, 2017). The P450s detoxify and/or bioactivate xenobiotics and catalyse functionalisation reactions by controlling N- and O-dealkylation, deamination, N- and S-oxidation, and hydroxylation (Omiecinski et al., 2011). During phase II, conjugation of the modified xenobiotics occurs rendering them less toxic. Conjugation includes sulphonation, methylation, glucuronidation, and amino acid conjugation. Specific classes of transferase enzymes are responsible for phase II e.g. UDP-glucuronosyltransferases, sulfotransferases, N-acetyltransferases, and glutathione S-transferases (Crocco et al., 2019). Conjugates are often more hydrophilic making them more excretable (McGinnity & Grime, 2017). The phase I and II enzymes work together to carefully control and regulate the metabolism, detoxification and bioactivation of xenobiotics. In phase III, conjugates are further metabolised ready for removal from the cell (Crocco et al., 2019). Conjugates and their metabolites are processed so they contain an anionic group which is recognised by efflux transporters which form the multidrug resistance protein family and catalyse, in an ATP-dependant manner, removal into the extracellular medium (Omiecinski et al., 2011).

The intensity and duration of drug action is reduced when activity of the enzymes and thus the rate of metabolism is increased. However, if the metabolising enzymes are responsible for converting a pro-drug into a drug, enzyme induction speeds up conversion and active drug levels, potentially leading to toxicity (Palleria et al., 2013). Several physiological and pathological factors influence drug metabolism including age (drugs are metabolised slower during foetal and neonatal periods and in the elderly), ethnicity, gender, nutrition, liver disease, kidney disease, and importantly pharmacogenetics (Ahmed et al., 2016). SNPs have been identified in genes encoding the drug metabolising enzymes which determines xenobiotic-related toxicity, adverse drug reactions, and the efficacy of drugs (Wormhoudt et al., 1999). Some of these variants have been found associated with altered cancer incidence and toxicity derived from chemical exposure (Kiyohara, 2000; Tomalik-Scharte et al., 2008). Polymorphisms are thought to be critical in controlling interindividual susceptibility to toxicity arising from exposure to certain drugs. For example, variants within the N-acetyltransferases have shown to have a considerable impact on the speed of acetylation with those carrying variants that lead to slow acetylation being at a higher risk of dose-dependent toxicity (Emilien et al., 2000; Hickman & Sim, 1991; Lazar et al., 2004).

1.6 General aims of this thesis

There is a growing body of evidence supporting the use of human genetics to proxy therapeutic modulation to provide genetic evidence of drug safety. Supporting genetic evidence increases the chances of therapies reaching the clinic reducing attrition rates and saving pharmaceutical companies considerable time and effort. Phase III noninferiority clinical trials of novel PHIs have recently been completed for the treatment of anaemia in CKD and show PHIs to be noninferior to ESAs for haematological efficacy and cardiovascular safety (Chertow et al., 2021; K.-U. Eckardt et al., 2021; Provenzano et al., 2021; A. K. Singh, Carroll, McMurray, et al., 2021; A. K. Singh, Carroll, Perkovic, et al., 2021). However, these clinical trials are only powered to achieve noninferiority on cardiovascular risk and not to detect risk of additional unwanted effects. There is also currently no genetic evidence supporting these trials. Therefore, in this thesis, I aim to identify genetic variants for use as proxies to investigate the long-term effects of therapeutic modulation of biomarker levels (EPO or Hgb) as a result of therapeutic PHD inhibition by PHI treatment. These results will further characterise the effects of long-term rises in Hgb or EPO levels and will provide additional evidence of drug safety to support the ongoing development of PHIs for treating anaemia in CKD. There are three overarching aims of my thesis:

1. To identify genetic variants associated with higher circulating EPO levels for use as genetic proxies for therapeutic modulation of endogenous EPO levels, the downstream effect of therapeutic PHD inhibition. I then use these variants to examine the cardiovascular risk and potential additional unintended effects associated with long-term higher endogenous EPO levels.

2. To functionally validate the identified genetic variants as causal in controlling EPO levels using gene-editing techniques.

3. To identify genetic variants associated with higher circulating Hgb levels to genetically proxy PHD inhibition and predict the long-term effects of higher circulating Hgb levels on cardiovascular risk or other unwanted effects.

These aims are addressed across four empirical chapters:

In my first empirical chapter (**Chapter 3),** I perform a GWAS meta-analysis of circulating EPO levels and identify a genetic variant associated with higher endogenous EPO levels lying in *cis* with the *EPO* gene. I use this *cis-EPO* genetic variant as a natural mimic for therapeutic increases in endogenous EPO levels in MR to investigate the risk of cardiovascular disease (CVD) with higher endogenous EPO levels. I also perform PheWAS to further investigate the effects of long-term higher circulating EPO levels.

In my second empirical chapter (**Chapter 4),** I perform whole *EPO* gene knock-out using CRISPR-Cas9 gene-editing to first establish a protocol for gene-editing in Human Embryonic Kidney (HEK-293) cells and to second establish a better understanding of the downstream biological mechanisms and pathways of EPO in a relevant human cell-line.

In my third empirical chapter **(Chapter 5),** I utilise CRISPR-Cas9 technology alongside the *piggyBac*^{TM} system to perform single-base gene-editing to functionally validate the *cis-EPO* variant (identified in **Chapter 3**) as causal in controlling EPO levels and therefore a valid proxy for long-term therapeutic modulation of endogenous EPO levels.

In my fourth empirical chapter (**Chapter 6**), I investigate the effects of higher circulating Hgb as a result of therapeutic PHD inhibition on cardiovascular risk and also test for any additional unwanted effects.

Figure 1.9 provides a graphical summary of the integration of the chapters in this thesis.

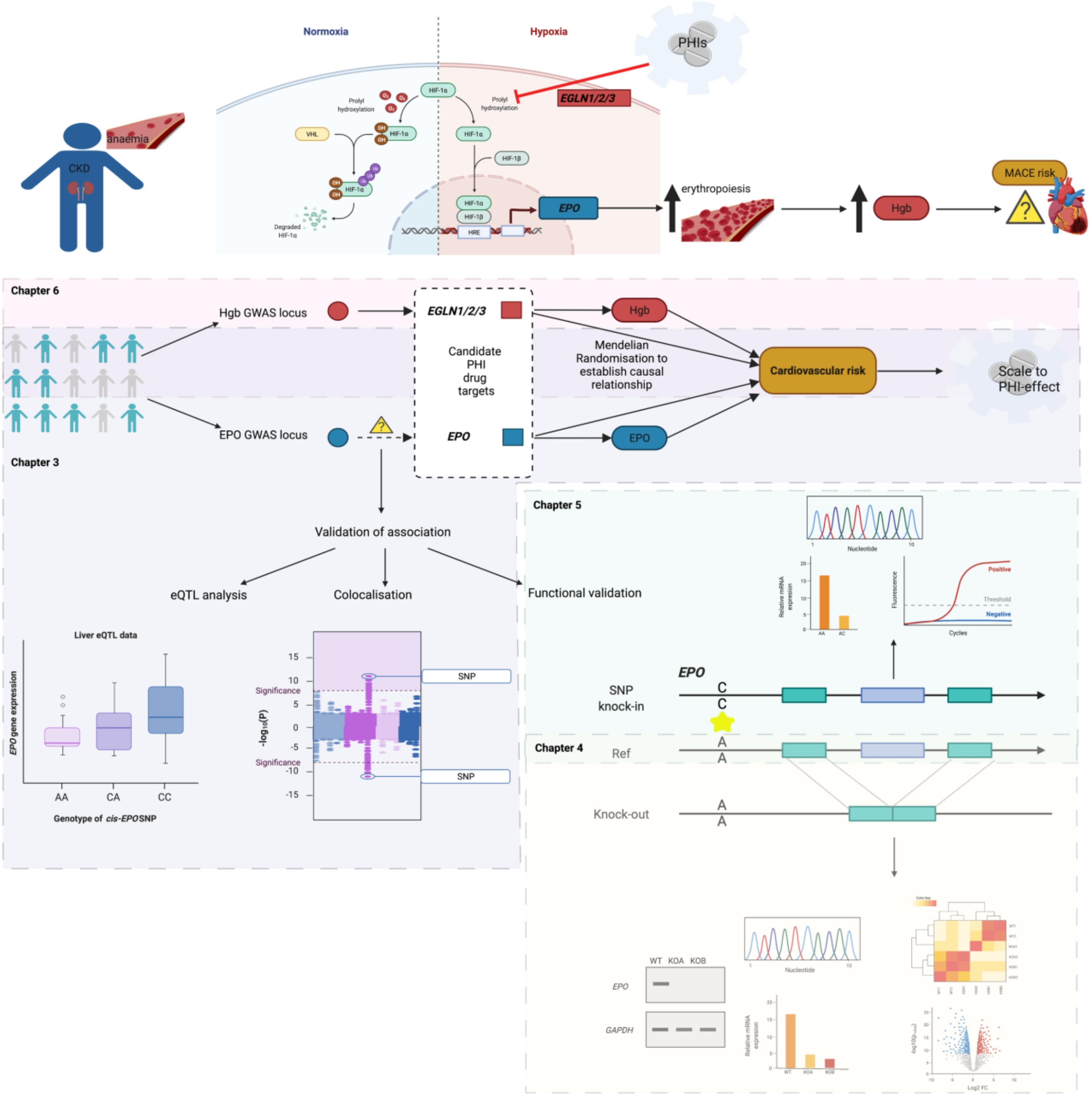


Figure 1.9: Graphical overview of the integration of the Chapters in this thesis. *The overarching aim of this thesis is to provide genetic evidence into the therapeutic profile of PHD inhibition to support the ongoing development of novel treatments (PHIs) for anaemia in CKD. CKD = chronic kidney disease; PHI = prolyl hydroxylase inhibitors; CVD = cardiovascular disease; Hgb = haemoglobin; EPO = erythropoietin; SNP = single nucleotide polymorphism; HIF = hypoxia inducible factor; VHL = von Hippel-Lindau; HRE = hypoxia response element; eQTL = expression quantitative trait loci; GWAS = genome-wide association study; Ref = reference; KO = knock-out; WT = wild-type. Created with BioRender.com.*

Chapter 2 General Methods

This chapter details the general materials and methods used throughout this thesis in multiple chapters. Detailed descriptions of methods and analysis specific to the chapter is given in the respective empirical chapter.

2.1 Statistical Genetic methods

2.1.1 GWAS

GWAS is one of the most commonly used approaches to associate genes with diseases. The method involves scanning the genome of thousands of individuals and identifying genetic markers where the allele frequencies differ between patients with or without a disease or between individuals with high or low levels of a biomarker (Marees et al., 2018). Identification of these trait-associated SNPs provides new insights into the genetic architecture and a better understanding of the biological mechanism underlying phenotypes potentially leading to personalised treatments (Visscher et al., 2017). The basic premise of a GWAS is the collection of DNA and phenotypic information (e.g. through medical health records, questionnaires and biological samples) from a group of individuals, genotyping of each individual using GWAS arrays or sequencing strategies, quality control of genotyping data, imputation of untyped variants using a reference panel, performing statistical association testing, conducting a meta-analysis, and interpretation of the results (Bush & Moore, 2012).

2.1.1.1 *Genotyping*

GWAS has only been made possible since the generation of large-scale genotyping platforms. Genotyping for common variants is typically done using microarrays or next generation sequencing methods when also detecting rare variants. For the purpose of GWAS, microarray-based technology is typically used due to the cost. SNP arrays are able to detect > 1 million SNPs (Marees et al., 2018). Self-assembled arrays are the most common type of array used for genotyping. The basic premise of DNA microarray technology is that DNA is synthesised onto small beads and these beads are then assembled randomly onto a glass surface. DNA is extracted from samples and labelled through incorporation of florescent dyes. The sample of labelled nucleic acids is washed over the microarray chip. Any sequences complementary to the sequence on the chip will hybridise and these are then visualised. The most commonly used approaches are allele discrimination, Illumina's Golden Gate Assay, the Infinium assay or array primer extension assays (Bumgarner, 2013). Allele discrimination by hybridisation is used by Affymetrix and involves the presence of known DNA oligo sequences complimentary to each allele being assembled on the array; the variant is placed within the centre of this oligo sequence as this position affects hybridisation the most (D. G. Wang et al., 1998). Labelled DNA fragments are washed over the array and hybridise to complementary sequences which are then visualised. Illumina's Golden Gate assay is based upon allele specific extension; two allele specific oligos are tailed with different universal primers and hybridised to genomic DNA (Bumgarner, 2013). Another oligo complementary to the same locus is tailed with a barcode sequence and a different universal primer (J. B. Fan et al., 2003). DNA polymerase extends the allele-specific primers across the genomic DNA sequence and these products are ligated to the third oligo (with the barcode sequence) (J. B. Fan et al., 2003). The barcode on the third oligo allows the PCR product to be uniquely detected on the array which contains oligo sequences complementary to the barcode sequences. The presence of multiple barcodes on the microarray enables multiplexing of many loci in one reaction (J. B. Fan et al., 2003). Array primer extension and Illumina's Infinium are similar (Bumgarner, 2013). In array primer extension assays, the array contains DNA that is attached to the chip at the 5' end whilst the 3' end stops one base before the SNP. Fragmented genomic DNA hybridises to the array and the array oligo is extended in a single nucleotide terminator sequencing reaction. The terminator is fluorescently labelled and then detected to determine the allele at this position (Kurg et al., 2000). In the Infinium assay, instead of the oligo being bound to the chip, it is on a bead and the SNP to be added is labelled with a hapten (a molecule that only elicits a response when bound to a larger molecule, such as a protein) which then binds a fluorescently labelled protein (Gunderson et al., 2006). The choice of genotyping platform depends on the specific needs, for example when genotyping samples from different ethnicities arrays with variants applicable to specific ethnicities are available, whilst when focusing on complex diseases the Axiom Biobank genotypic array may be better. Microarrays do not sequence every nucleotide in the genome but only those that are the array is designed to detect which typically include the most well-studied and well-known variants (Bumgarner, 2013).

2.1.1.2 *Imputation*

Imputation allows for the prediction of unmeasured genotypes in low-density datasets (e.g. those from SNP arrays) using densely genotyped datasets as references such as HapMap (Frazer et al., 2007) or 1000 Genomes (Jostins, Morley, & Barrett, 2011). This results in a substantial increase in power and allows for the meta-analysis of studies genotypes on difference SNP arrays. Several tools are available for imputation. Within this thesis, genotype data was imputed using the Michigan Imputation Server (Das et al., 2016) (https://imputationserver.sph.umich.edu/index.html#!) which uses Eagle2 (Loh et al., 2016) to phase haplotypes, and Minimac4 (https://genome.sph.umich.edu/wiki/Minimac4) with the most recent 1000 Genomes reference panel (phase 3, version 5) (Auton et al., 2015).

2.1.1.3 *Quality control checks*

Genotype data is inherently imperfect and therefore extensive quality control checks are required during GWAS in order to generate reliable results. Errors in the data can arise due to poor quality of DNA samples, poor DNA hybridization to the array, poorly performing genotype probes, and sample mix-ups or contamination (Marees et al., 2018). The typically used QC checks are outlined in **Table 2.1** and involve filtering out SNPs and individuals based on missingness, sex discrepancy, minor allele frequency (MAF), deviations from Hardy-Weinberg equilibrium (HWE), heterozygosity, relatedness and population stratification.

2.1.1.3.1 Missingness

Missingness refers to both SNP-levels missingness and individual-level missingness (Laurie et al., 2010). Individual-level missingness is the number of SNPs missing for a specific individual; this can indicate poor DNA quality or technical issues (Marees et al., 2018). SNP-level missingness refers to the number of individuals within the study for whom information on a specific SNP are missing; SNPs with a large amount of missingness can lead to bias (Laurie et al., 2010).

2.1.1.3.2 Sex discrepancy

Sex discrepancy is the difference between the assigned sex and the sex determined based on genotype and can only be checked when SNPs on the sex chromosome have been called. A discrepancy is likely indicative of sample mix-up (Marees et al., 2018).

2.1.1.3.3 Minor Allele Frequency

MAF is the frequency at which the second most common allele occurs at a given site in a given population. Minor alleles drive a considerable amount of selection and play a role in heritability (Manolio et al., 2009). The majority of GWAS are underpowered to detect associations with SNPs with a low MAF and therefore exclude these from the analysis (Marees et al., 2018).

2.1.1.3.4 Hardy-Weinberg equilibrium

Hardy-Weinberg equilibrium (HWE) assumes that in a population of infinite size, the genotypes and allele frequencies remain constant across generations and thus in equilibrium by assuming no natural selection, migration, or mutation (Lachance, 2016). The expected frequency of a genotype can be calculated using **Equation 2.1**. Violation of HWE occurs when the observed frequency of a genotype in a population is significantly different from the expected genotype frequency (Trikalinos et al., 2006). In GWAS, HWE violation may be indicative of genotyping errors. The threshold used in disease-specific populations are often less stringent as violations of HWE may actually be indicative of a true association with disease risk (Marees et al., 2018).

2.1.1.3.5 Heterozygosity

Heterozygosity refers to the presence of two different alleles at a specific variant. The rate of heterozygosity in an individual is the proportion of heterozygous genotypes, which is calculated using a list of SNPs that are not highly correlated by excluding high inversion regions (Samuels et al., 2016). The mean of heterozygosity is calculated across the population and any individuals with a heterozygosity more than 3 standard deviations from the mean are excluded (Marees et al., 2018). A high level of heterozygosity within an individual may be indicative of poor sample quality whilst low heterozygosity can indicate a high level of inbreeding (Anderson et al., 2010).

2.1.1.3.6 Relatedness

Relatedness is a measure of how strongly a pair of individuals is genetically related. Most GWAS' assume that all individuals are not more closely related than second-degree relative (Anderson et al., 2010). The inclusion of more closely related individuals in a GWAS can result in biased estimations (Marees et al., 2018). Several tools have been developed over the past few years, such as BOLT-LMM and GEMMA, which can account for relatedness through the inclusion of a genomic relatedness matrix (GRM) in the model (Loh et al., 2015; Zhou & Stephens, 2012).

2.1.1.3.7 Population Stratification

Many cohorts involve the inclusion of individuals from diverse ancestries known as population stratification. The inclusion of diverse ancestries can confound standard QC metrics due to different allele frequencies across subpopulations (Anderson et al., 2010). These differences can give rise to spurious associations and/or the masking of true associations. For example, genotypic differences between case/controls may be detected due to different populations rather than an effect on the disease (Cardon & Palmer, 2003). It is therefore important to perform association analysis separately in subpopulations before combining results through meta-analysis. Subtle population stratification can also exist in a single ethnic population and therefore it is important to test and control for population stratification in GWAS in order to reduce the amount of potential systematic bias (Abdellaoui et al., 2013).One method to determine different ancestries and to account for population stratification is by using principal component analysis (PCA) (Price et al., 2006). The method calculates the genome-wide average proportion of alleles shared between any pair of individuals to generate principal components (PCs) of the genetic variation for each individual. The PCs for each individual can be plotted to explore whether there are groups of individuals that are genetically more similar to each other than expected; for a study including Europeans and Asians, PCA would reveal a clustering of Europeans indicating that they are genetically more similar to each other than to the Asians (Marees et al., 2018). The PCs of the study participants under investigation are often plotted against those derived from a population of known ethnic structure, such as HapMap (Frazer et al., 2007) or 1000 Genomes (B. Howie et al., 2011), to determine any outliers and to determine participants from the same ancestry (Anderson et al., 2010). After exclusion of any outliers and inclusion of only those individuals from the same subpopulation, it is important to repeat PCA and to use these resulting PCs as covariates in association tests to correct for any remaining, underlying stratification which could lead to spurious associations (Marees et al., 2018). The inclusion of up to 10 PCs is typically accepted in GWAS.

$$\mathrm{p} + \mathrm{q} = 1$$

$$p^2 + 2pq + q^2 = 1$$

p: dominant allele frequency

q: recessive allele frequency

p^2: dominant allele homozygous frequency

2pq: Heterozygous frequency

q^2: recessive allele homozygous frequency

Equation 2.1: Hardy-Weinberg equilibrium equations

Table 2.1: Quality control checks typically performed in GWAS.

QC check	**Threshold**	**Explanation**
SNP-level missingness	Exclude SNPs with missingness > 20%	Excludes SNPs that are missing in a large proportion of the subjects. SNPs with low genotype calls are removed.
Individual-level missingness	Exclude individuals with missingness > 20%	Excludes individuals who have high rates of genotype missingness. Individual with low genotype calls are removed.
Sex discrepancy	Males should have an X chromosome homozygosity estimate >0.8 and females should have a value <0.2.	Checks for discrepancies between sex of the individuals recorded in the dataset and their sex based on X chromosome heterozygosity/homozygosity rates. Indicative of sample mix-ups
Heterozygosity	Remove individuals who deviate ±3 SD from the samples' heterozygosity rate mean.	Excludes individuals with high or low heterozygosity rates.
Minor allele frequency	Exclude SNPs with MAF <0.01 or <0.05 typically used.	SNPs with a low MAF are rare, therefore power is lacking for detecting SNP-phenotype associations. These SNPs are also more prone to genotyping errors. The MAF threshold should depend on your sample size, larger samples can use lower MAF thresholds.
Hardy-Weinberg equilibrium	Exclude SNPs with HWE $P < 1 \times 10^{-06}$	Excludes markers which deviate from Hardy–Weinberg equilibrium. Common indicator of genotyping error, may also indicate evolutionary selection. Deviations can indicate sample contamination, inbreeding.
Relatedness	Exclude any individuals closer than second-degree relatives or account for relatedness in statistical analysis e.g. using a GRM.	Cryptic relatedness can interfere with association analysis.

Population stratification	Include only individuals from the same subpopulation based on PCA.	As allele frequencies differ between subpopulations, biases can be introduced by including different ancestries leading to false-positive associations or masking of true causal association.

2.1.1.4 *Association testing*

Linear regression is a statistical method used to estimate the relationship between two variables assuming linearity. Linear regression is the standard approach for identifying whether genetic variants are associated with traits of interest in GWAS (Tao Wang et al., 2018). The most commonly used technique in GWAS is multiple linear regression which allows for the inclusion of confounding variables, such as environmental confounders like smoking, biological confounders like sex or age, and technical confounders like genotyping chip (Pourhoseingholi et al., 2012). The multiple linear regression model is denoted in **Equation 2.2.** Linear regression makes several key assumptions (Osborne & Waters, 2002; Uyanık & Güler, 2013);

1. There is a linear relationship between the outcome and the independent variables. This can be assessed by plotting a scatter plot.
2. The residuals are normally distributed. This can be visually assessed using a quantile-quantile (qq) plot whereby the theoretical quantiles are plotted against the standardised residuals.
3. The independent variables are not highly correlated to each other i.e. there are no multi-collinearity. This can be measured using a Pearson's correlation matrix among all independent variables. A correlation coefficient less than 0.8 indicates no multi-collinearity.
4. The data is homoscedastic meaning that the variance of error terms across values of the independent variable are similar. This can be assessed graphically using a qq plot

Violation of these assumptions can lead to increased error rates and biases and therefore it is standard practice to perform inverse normalisation on residuals. For binary or disease traits, logistic regression is used instead which generates an estimate of the log odds of disease in the presence of an additional copy of the minor allele (Marees et al., 2018). An advancement on the traditionally used linear/logistic regression approach is the use of a linear mixed model (LMM) which is better at controlling for population stratification and cryptic relatedness, correcting inflation of false-positives that may be caused by many small genetic effects, and increases statistical power by jointly modelling all SNPs (H. Chen et al., 2016; Widmer et al., 2014). LMMs estimate the genetic similarity between a pair of individuals to capture the genealogy within the population and enable the inclusion of both fixed effects (the overall mean, the effect of a SNP and the effect of covariates) and a random effect (reflecting the polygene background whereby a combination of genes influence the trait) (Uffelmann et al., 2021). The variance of the random effect is dependent on the kinship matrix which is a measurement of genetic similarity across individuals. (Eu-ahsunthornwattana et al., 2014; Lippert et al., 2011). The overall model for LMM methods is shown in **Equation 2.3**. LMMs have shown to perform favourably in population based and case-control cohorts and are effective at accounting the complete genealogy of the population, including population structure, family structure and cryptic relatedness, reducing the risk of false positives whilst maintaining power as all individuals from the same subpopulation can be included regardless of relatedness (Loh et al., 2015; Yu et al., 2006; Zhou & Stephens, 2012).

$$\boldsymbol{y = x\beta_s + z_n\beta}$$

y: a phenotype of interest
x: SNP genotype at a given locus
β_s: changes in y as a function of genotype at x
z_n: covariates with effect sizes β

Equation 2.2: Multiple linear regression model used in GWAS.

$$y = X\beta + g + \varepsilon$$

X: matrix of fixed effects including overall mean, SNP being tested and covariates with β denoting the coefficients of these fixed effects.
g: a random effect
ε: the random residual effect

Equation 2.3: Linear Mixed Model based approach used in GWAS.

2.1.1.5 *Meta-analysis*

The basic principal of meta-analysis is to combine the evidence of association from individual studies using appropriate weights. METAL (Willer et al., 2010) is a commonly used software and was used in this thesis when combining association statistics in **Chapter 3**. METAL involves two main approaches; 1) the conversion of *P*-values and effect estimates into a z-score where a very negative z-score represents a small *P*-value and a negative effect of the allele on disease risk or trait levels (i.e. the allele is associated with lower trait levels or decreased risk of disease) and a very positive z-score represents a small *P*-value and a positive effect of the allele on disease risk or trait levels (i.e. the allele is associated with higher trait levels or increased risk of disease), 2) combining z-scores across SNPs in a weighted sum with the weight proportional to the square-root of the study sample size meaning that largest studies are given more weight compared to smaller studies (Willer et al., 2010). This method enables results to be combined when effect estimates are not available in all studies or are in different units across studies. This method is equivalent to a fixed-effect inverse-variance weighted method where the weight is proportional to the inverse of the standard error providing that the trait distribution is identical across samples (Willer et al., 2010). The fixed-effect inverse-variance weighted method was performed in this thesis using metan (Harris et al., 2008) in Stata (StataCorp, 2019).

2.1.2 UK Biobank GWAS data

For all analysis using the UKB in this thesis, the imputed data released in 2017 was analysed. Genome-wide genotyping was performed on ~450,000 individuals using the UK Biobank Axiom Array and on ~50,000 individuals using the UKB BiLEVE array. The two SNP arrays were very similar with over 95% common marker content. The UK Biobank Axiom array was an updated version of the UK BiLEVE Axiom array, and included additional novel markers. Approximately 812,000 unique markers (SNPs and indels) were directly measured, with > 90 million variants being imputed using the Haplotype Reference Consortium (HRC) (Loh et al., 2016) and UK10K + 1000 Genomes (Auton et al., 2015) reference panels. This has been described in more detail elsewhere (Bycroft et al., 2018). Due to the reported technical error with non-HRC imputed variants, I focused solely on the set of ~40 million imputed variants from the HRC reference panel. PCA was performed to determine population stratification. PCs were generated in the 1000 Genomes cohort using high-confidence SNPs to obtain their individual loadings. These loadings were then used to project all of the UKB samples into the same PC space, and individuals were then clustered using PCs 1–4. To account for population structure and relatedness of individuals, a linear mixed model implemented in BOLT-LMM (Loh et al., 2015) was used to perform GWAS. Only autosomal SNPs with a MAF > 1 %, in Hardy Weinberg equilibrium ($P > 1 \times 10^{-06}$), passing QC in all 106 batches, and present on both genotyping arrays were included in the GRM. For all continuous traits, single inverse normalisation was performed to account for skewed distributions and the resulting residuals were adjusted for genotyping array, sex, age at baseline and centre alongside any trait-specific covariates.

2.1.3 eQTL analysis

eQTL analysis is used to identify genetic variants associated with gene expression on the premise that a proportion of transcripts are under genetic control (Nica & Dermitzakis, 2013). The basic approach for performing eQTL analysis is similar to that of GWAS but involves the testing of an association between a genotype and gene expression in a specific tissue or cell-type of interest (Marta et al., 2015). Individual transcript levels are determined in a selected tissue or cell-type from selected unrelated individuals of the same ancestry using microarrays. Briefly, RNA is extracted from the samples of interest, converted to cDNA and labelled before array hybridisation and scanning using a confocal scanner. Raw data files are then pre-processed including adjusting for differences between arrays, background estimation and correction, and performing normalisation before being mapped to genes. Measured gene expression is typically inverse rank normalised to account for skewed distributions. Each individual gene transcript is treated as a quantitative trait which differs to typical GWAS where a single or a few complex phenotypes are investigated. Association analysis is performed to identify SNPs significantly associated with expression using either a Kruskal-Wallis test for non-parametric data (Greenawalt et al., 2011; Schadt et al., 2008) or an additive linear model (Innocenti et al., 2011) adjusting for covariates such as age, sex, PCs and additional hidden factors identified through PEER (https://www.sanger.ac.uk/science/tools/PEER) (Stegle et al., 2010).

Similar to GWAS, eQTL data is often made publicly available, for example in GTEx (https://gtexportal.org), enabling further investigation of putative susceptibility loci identified through GWAS are investigated through eQTL analysis to determine whether these statistically significant associations hold in the relevant tissue or cell-type providing additional evidence of the true causal gene and pathway or to provide additional biological evidence supporting susceptibility loci that fail to reach genome-wide significance. In this thesis, I used extracted eQTL association statistics for variants lying within 500 kb either side of my SNP of interest. These association statistics were produced by collaborators using data that was not publicly available at the time (Damman et al., 2015; Etheridge et al., 2020; Greenawalt et al., 2011; Innocenti et al., 2011; Schadt et al., 2008)

2.1.4 Correction for multiple testing

Modern genotyping and imputation results in the analysis of millions of SNPs. Testing millions of SNPs for associations generates a large number of tests and thus a considerable multiple testing burden (Marees et al., 2018). However, due to the presence of LD, SNPs are highly correlated and are therefore not independent decreasing the number of independent tests being performed (Storey & Tibshirani, 2003). A Bonferroni-correction can be applied to calculate the adjusted *P*-value threshold controlling for the probability of having at least one false-positive association (Pe'er et al., 2008) . During GWAS, after accounting for LD, there are considered to be ~1 million independent SNPs. For this reason, a *P*-value of 5 x 10^{-08} (0.05 / 1 million) is typically accepted as the threshold for determining genome-wide significance and was used throughout this thesis (Dudbridge & Gusnanto, 2008). This was used as the threshold for determining genome-wide significance in **Chapter 3.** Alternatively, due to the inclusion of lower frequency variants since advancements in genotyping technologies increasing power, a lower *P*-value threshold of 6.6.x 10^{-09} can also be used (Fadista et al., 2016). This was the case for the Hgb GWAS used in **Chapter 6.** For the eQTL analysis in **Chapter 3**, due to scanning only 0.01% of the genome (a 1 mb region surrounding the SNP of interest), a lower p-value threshold was used to determine significance (0.05 / 180 SNPs = 2.78 x10^{-04}). For all MR analysis, the standard significance threshold of $P < 0.05$ was used which is the typically accepted threshold for interpreting MR studies. When performing PheWAS, the *P*-value significance threshold was determined using a Bonferroni correction by dividing 0.05 by the number of phenotypes investigated; in **Chapter 3,** the threshold used was 5.75 x 10^{-05} (0.05 / 869 traits) and in **Chapter 6,** the threshold used was 5.42 x 10^{-05} (0.05 / 923).

$$Adjusted\ P\ value\ threshold = \frac{0.05}{n}$$

n: number of independent tests

***Equation 2.4: Formula for the Bonferroni-corrected P-value threshold.** In the case of GWAS/eQTL analysis, n is equivalent to the number of independent SNPs being tested. For GWAS, 1 million SNPs are considered to be independent and thus often n = 1,000,000.*

2.1.5 Scaling genetic associations

Genetic associations between a variant and trait of interest are often very small due to common genetic variants indicating lifelong perturbations and having an accumulative effect alongside other variants on the trait of interest. For this reason, it is important to scale genetic estimates to the minimally clinically relevant effect, whether that be the effect of a drug on trait levels or a 1-unit increase in trait levels, to obtain a physiologically relevant and reliable estimate. This can be achieved by calculating the scaling factor. A similar approach has previously been used by Scott et al. (2016). The scaling factor is determined by calculating the magnitude of difference between the effect of the genetic variant on the trait of interest and the minimal clinically relevant effect (**Figure 2.1**). This scaling factor can then be used to convert the genetic associations measured between the variant and outcome or exposure and outcome in MR studies to determine the predicted effect at a physiologically relevant level. **Figure 2.2** shows a worked example of scaling genetic effect estimates obtained through MR to the effect of a drug which increases which increases biomarker levels by 10-fold. For binary traits, it is important to ensure the scaling is performed on the logarithmic scale before converting back into odds ratios to estimate the physiologically relevant effect of an increase in exposure on risk of disease (**Figure 2.2**).

Typically, during GWAS and/or meta-analysis, trait values are residualised and inverse normalised to account for skewed distributions and meet the normal distribution assumptions made during statistical testing. For this reason, genetic effect estimates obtained through GWAS also often need scaling back to the original trait units to determine whether the difference estimated through GWAS/PheWAS is clinically relevant. To achieve this, the standard deviation of the trait is taken from the study/population of interest and used to convert the effect estimate in SDs back into the original trait units by dividing (**Table 2.2**).

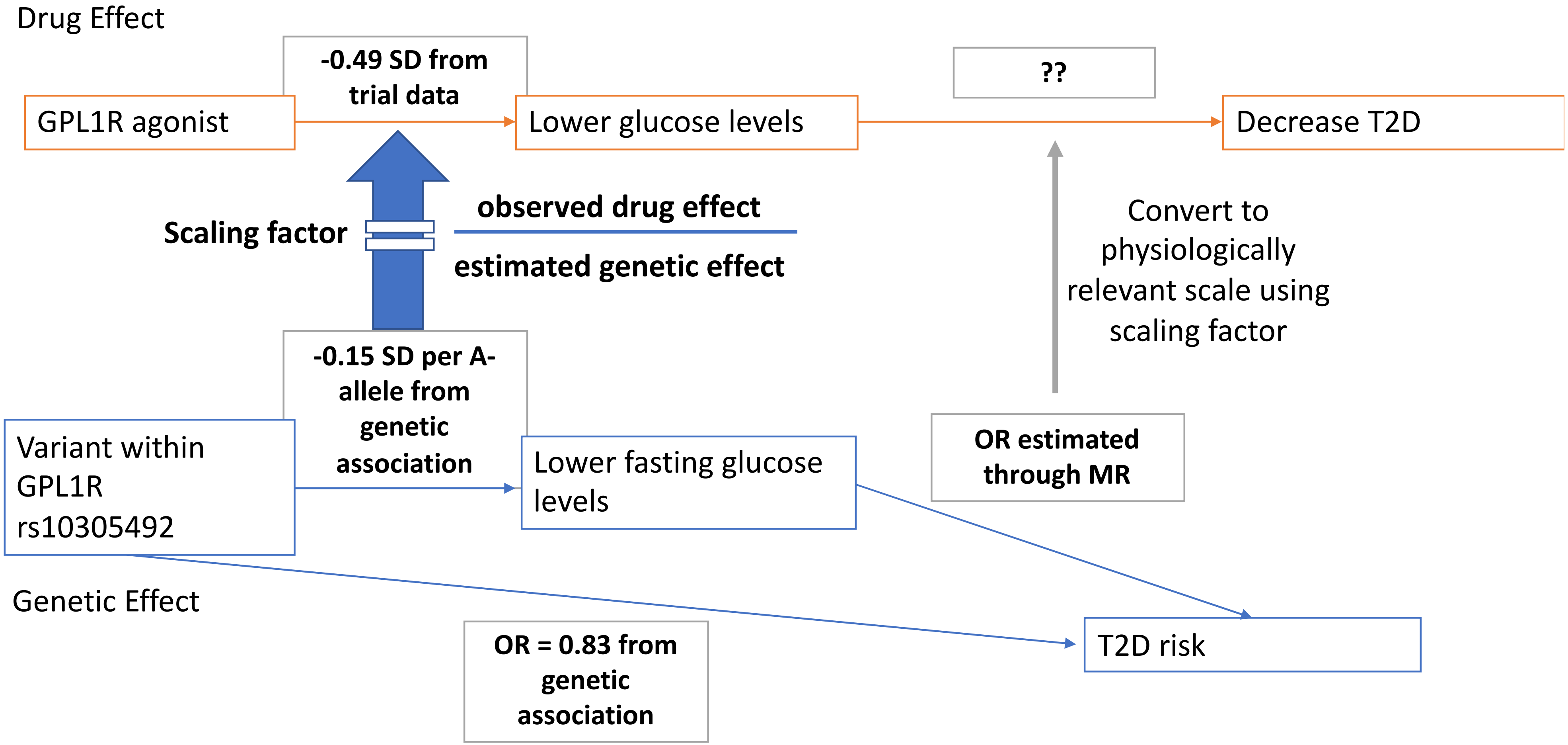
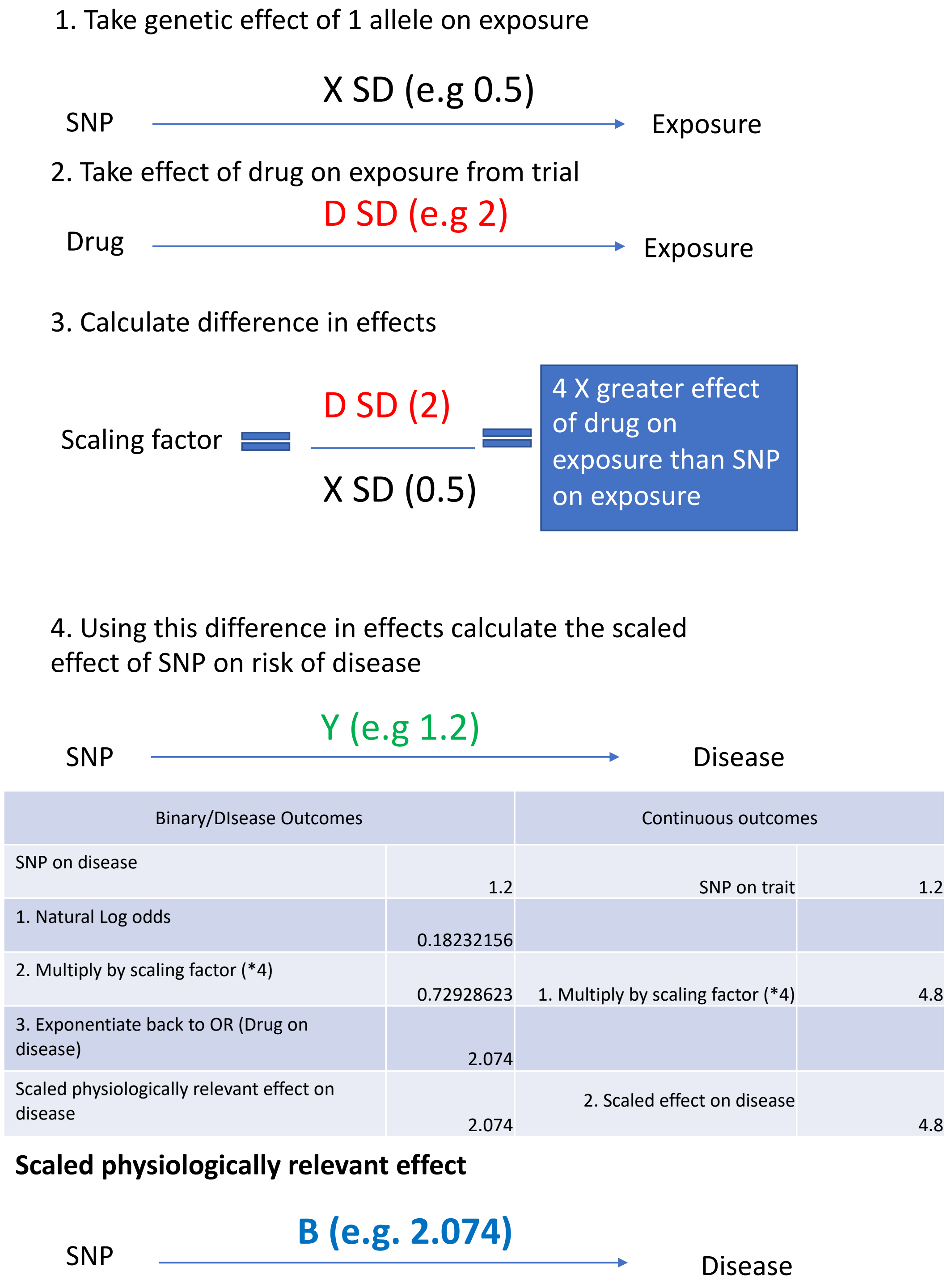

Drug Effect
GPL1R agonist
-0.49 SD from trial data
Lower glucose levels
??
Decrease T2D
Scaling factor
observed drug effect
estimated genetic effect
Convert to physiologically relevant scale using scaling factor
-0.15 SD per A-allele from genetic association
Variant within GPL1R rs10305492
Lower fasting glucose levels
OR estimated through MR
Genetic Effect
OR = 0.83 from genetic association
T2D risk

Figure 2.1: Diagram depicting the principal of rescaling estimated genetic associations to a clinically relevant effects to predict the likely physiologically relevant effect on outcomes. *Example and numbers have been taken directly from Scott et al. (2015) where scaling was also used.*



1. Take genetic effect of 1 allele on exposure

$$\text{SNP} \xrightarrow{\text{X SD (e.g 0.5)}} \text{Exposure}$$

2. Take effect of drug on exposure from trial

$$\text{Drug} \xrightarrow{\text{D SD (e.g 2)}} \text{Exposure}$$

3. Calculate difference in effects

$$\text{Scaling factor} = \frac{\text{D SD (2)}}{\text{X SD (0.5)}} =$$

4 X greater effect of drug on exposure than SNP on exposure

4. Using this difference in effects calculate the scaled effect of SNP on risk of disease

$$\text{SNP} \xrightarrow{\text{Y (e.g 1.2)}} \text{Disease}$$

Binary/DIsease Outcomes		Continuous outcomes	
SNP on disease	1.2	SNP on trait	1.2
1. Natural Log odds	0.18232156		
2. Multiply by scaling factor (*4)	0.72928623	1. Multiply by scaling factor (*4)	4.8
3. Exponentiate back to OR (Drug on disease)	2.074		
Scaled physiologically relevant effect on disease	2.074	2. Scaled effect on disease	4.8

Scaled physiologically relevant effect

$$\text{SNP} \xrightarrow{\textbf{B (e.g. 2.074)}} \text{Disease}$$

Figure 2.2: Worked example of rescaling genetic estimates to clinically relevant effects obtained from clinical trial data for both quantitative traits and disease outcomes.

Table 2.2: Conversion of genetic estimates in standard deviations into original units of the trait.

1. **Convert the effect of the SNP on trait of interest (in SDs) into raw units** **Raw units = genetic estimate / SD in the population**				
	Effect of SNP on trait of interest		Effect of SNP on trait of interest	
Trait	in SD (genetic estimate)	SD of trait in population of trait	raw units	
Hgb	0.05	1	0.05	
2. **Calculate the scaling factor to estimate a 1-unit, 2-unit or 0.5-unit increase in Hgb levels** **Scaling factor = desired unit increase / effect of SNP on trait of interest in raw units**				
Calculate the scaling factor for 1-unit increase	1/0.05 = 20			
Calculate the scaling factor for 2-unit increase	2/0.05 = 40			
Calculate the scaling factor for 0.5-unit increase	0.5/0.05 = 10			
3. **Scale the genetic estimate of SNP on PheWAS trait/outcome to desired-unit increase (in SDs) and then convert to the original units of the PheWAS trait/outcome** **Scaled effect (in SDs) = Genetic estimate of SNP on PheWAS trait x scaling factor** **Original units = Scaled effect x SD of PheWAS trait/outcome in population**				
	Effect of SNP on trait from PheWAS	Scaled to 1 unit increase in Hgb		Predicted effect of a 1 unit increase Hgb on EPO in raw EPO units
	in SD (genetic estimate)	in SD	SD of trait in population	in raw units
EPO	0.02	0.4	1.5	0.6

2.2 General wet-lab methods

2.2.1 Molecular cloning

All work with live bacteria was carried out under sterile conditions, using a category 2 biological safety cabinet, sterile consumables and sterile media. Molecular cloning was used to generate the plasmids used within this thesis. Molecular cloning involves the insertion of a segment of DNA into a backbone vector for propagation. Recombinant DNA (plasmid containing the DNA insert) can then be isolated allowing the expression and/or manipulation of gene expression in cell-lines. The basic molecular cloning process involves:

1. Restriction digest of backbone vector to linearise the circular plasmid and generate single stranded overhangs
2. Ligation of DNA segment into the backbone vector creating recombinant molecules (i.e. plasmids)
3. Transformation of plasmids into bacteria for propagation
4. Screening of bacteria containing the plasmid
5. Isolation of the plasmid for downstream use.

2.2.1.1 *Plasmid preparation*

Plasmids ordered from Addgene arrived as a live bacterial stab culture. Stab cultures were streaked out onto agar plates containing the appropriate antibiotic selection. Plates were incubated overnight at 37 ^{0}C for the formation of the colonies. Colonies were selected from the agar plate and individually cultivated in 5 mL of LB broth supplemented with the appropriate antibiotic in a 20 mL sterile bacterial tube, shaken at 220 rpm overnight at 37 ^{0}C. Plasmids were then isolated from bacteria using a QIAprep Spin Miniprep kit (Qiagen, Maryland, USA) which is explained below in **2.2.1.5.**

2.2.1.2 *Restriction digest and ligation*

Segments of DNA were inserted into the backbone vectors/plasmids in a single digestion and ligation reaction. Briefly, the circular plasmid is digested by a restriction enzyme generating a linearised plasmid with overhangs of single stranded DNA, known as sticky ends. The DNA to be inserted is designed to also contain these sticky ends and will therefore also be digested by the same restriction enzyme producing complementary overhangs which can anneal enabling the insertion of the insert into the backbone vector. The single digestion and ligation reaction mix contained a 3:1 ratio of insert:plasmid, 2 µl 10X FastDigest Buffer (New England BioLabs, Ipswich, UK), 1 µl dithiothreitol (DTT, 10 mM), 1 µl adenosine triphosphate (ATP, 10 mM), 1 µl FastDigest restriction enzyme (New England BioLabs, Ipswich, UK), 0.5 µl T4 ligase (New England BioLabs, Ipswich, UK) made up to a total volume of 20 µl with ddH$_2$0. The reaction mix was then incubated in a thermocycler under the conditions in **Table 2.3.**

Table 2.3: Thermocycler conditions for single ligation and digestion reaction using T4 ligase.

Temperature (^{0}C)	Time (minutes)	Number of cycles
37	5	6
21	5	
4	Infinite	

2.2.1.3 *Bacterial transformation*

To obtain multiple copies of the recombinant plasmid, bacteria were transformed with the ligation reaction before purifying and screening for the desired plasmid. 50 µl of sub-cloning efficiency competent *Escherichia Coli DH5alpha* were thawed on ice for one hour and subsequently mixed with 2 µl of ligation reaction and incubated for 15 minutes on ice. Bacteria were then heat shocked for 45 seconds at 42 ^{0}C and incubated immediately on ice for 2 minutes to promote horizontal gene transfer. 950 µl of Lysogeny Broth (LB) supplemented with 1X SOC media (10 mM NaCl, 2.5 mM KCl, 10 mM MgSO4, 20 mM Glucose, 10 mM MgCl2) was added to the bacteria and bacteria were placed in a shaker at 37 ^{o}C for 1 hour at 225 rpm to enable expression of the bacterial proteins. Bacteria were centrifuged at 4000 rpm for 2 minutes. The pellet was resuspended in 250 µl of LB broth media and plated onto an agar plate containing the appropriate antibiotic selection (100 µg/mL ampicillin). Plates were inverted and incubated overnight at 37 ^{0}C. Colonies containing the transformed plasmid should grow overnight.

2.2.1.4 *Overnight cultures*

Colonies on agar plates are not large enough to extract sufficient amount of plasmid DNA for downstream applications. Instead, single colonies are inoculated in LB containing antibiotic and grown overnight to enable the replication of larger volumes of bacteria containing the plasmid.

3 mL of LB prepared with 100 μg/mL ampicillin (or appropriate antibiotic) was aliquoted into a culturing tube. A single colony was selected using an inoculating loop and the culturing tube containing LB and antibiotic was inoculated by stirring the inoculating loop in the broth. Cultures were stored overnight at 37 ^{0}C shaking at 225 rpm. A negative control of LB with antibiotic but no bacteria was included to confirm aseptic technique. The following day, bacterial growth was confirmed by the presence of cloudy cultures whilst the negative control remained clear.

2.2.1.5 *Isolation of plasmid DNA*

To isolate pure plasmid DNA from the bacterial cultures, the overnight cultures were processed using the QIAprep Spin Miniprep Kit (Qiagen, Maryland, USA). The kit is designed for up to 20 μg of high-copy plasmid DNA and lyses the bacteria to release the plasmid DNA. RNase A was added to Buffer P1 and 100 % ethanol added to Buffer PE as instructed by the kit. Overnight cultures were centrifuged at room temperature at 13,000 rpm for 3 minutes and supernatant removed. The pellet was resuspended in 250 μl of Buffer P1 and transferred to a microcentrifuge tube. 250 μl of Buffer P2 was added and mixed by gentle inversion 6 times. 350 μl of Buffer N3 was added and the tube inverted 6 times to mix before centrifugation for 10 minutes at 13,000 rpm. 800 μl of the supernatant was transferred to a QIAprep 2.0 Spin Column (Qiagen, Maryland, USA) and centrifuged for 1 minute at 13,000 rpm. The plasmid DNA should now be bound to the column. The flow-through was discarded and 750 μl of Buffer PE added followed by centrifugation at 13,000 rpm for 1 minute. The flow-through was discarded and residual wash buffer was removed by centrifugation at 13,000 rpm for 1minute. The QIAprep 2.0 column (Qiagen, Maryland, USA) was placed into a 1.5 mL Eppendorf tube and 50 μl of ddH_20 was added to the centre of the column. The column was left to stand for 1 minute before centrifugation for 1 minute at 13,000 rpm to elute the plasmid DNA. The eluted plasmid DNA is quantified using the Nanodrop machine as described in **2.2.4** and stored at -20 ^{0}C.

2.2.1.6 *Confirmation of insertion of DNA into plasmid*

To confirm the insertion of the DNA into backbone vectors, a diagnostic double restriction digest was carried out. If the insert is successfully cloned into the vector, a restriction enzyme cut-site is disrupted resulting in a different digest pattern to that of the empty backbone vector. A double restriction digest was performed on 100ng of plasmid DNA using 0.5µl of each restriction enzyme (e.g. EcoRI and BbSI), and 1 µl of FastDigest Green Buffer in a total reaction volume of 10 µl. The digest reaction was run in a thermocycler at 37 ^{0}C for 60 minutes followed by 65 ^{0}C for 20 minutes. The resulting digest products were separated and visualised on a 0.5% (w/v) agarose gel in 1X TAE buffer containing SYBR Safe DNA gel stain (described in more detail in **2.2.6).** The agarose gel was imaged using the Licor Oddysey Imaging system (LI-COR Biosciences Ltd, Cambridge, UK). Once the correct digest pattern had been observed, positive plasmids were sent for Sanger sequencing. 20 µl of 100 ng/µl plasmid DNA was prepared in a 1.5 mL Eppendorf alongside 20 µl of 10 µM LKO.1 5' primer within the U6 promoter (5'-GACTATCATATGCTTACCGT-3'). Sanger sequencing was performed by Genewiz Ltd (Genewiz, Essex, UK). Sequence data was analysed using SnapGene software (from Insightful Science; available at snapgene.com) to confirm insertion of the insert in the correct orientation and location.

2.2.1.7 *Glycerol stocks*

Glycerol stocks were made for long-term storage of recombinant plasmids. 500µl of bacteria from overnight culture was mixed with 500µl of 50% glycerol in a cryovials and then stored at -80 ^{0}C.

2.2.2 Tissue culture

2.2.2.1 *HEK-293 cell-line*

The Human Embryonic Kidney-293 (HEK-293) cell-line was used for all functional work performed in this thesis. The cell-line was a gift from Dr John

Chilton. The HEK-293 cell-line is one of the most cited *in vitro* cell models as it is easy to culture, cheap to maintain, highly reproducible, easy to transfect and is highly characterised (P. Thomas & Smart, 2005). The HEK-293 cell-line is an immortalised cell-line generated by transforming and culturing human embryonic kidney cells from a female foetus with sheared adenovirus type 5 DNA (F. L. Graham et al., 1977). The adenovirus DNA prevents cell-cycle arrest enabling continuous propagation of the resulting cell-line (Shaw et al., 2002). HEK-293 cells are highly heterogenous, comprising of endothelial, epithelial and fibroblast cell-types, and therefore have a complex karyotype (Stepanenko & Dmitrenko, 2015). HEK-293 cells are hypotriploid routinely carrying 64 chromosomes. Chromosomal abnormalities include a total of three copies of the X chromosomes and four copies of chromosome 17 and chromosome 22.(Y.-C. Lin et al., 2014) This complex karyotype makes their behaviour different from primary human cells (Stepanenko & Dmitrenko, 2015). HEK-293 cells were chosen as the cell-line model of choice as *EPO* is highly expressed in the liver and the kidneys. However, as HEK-293 cells are from an embryonic kidney cell-line, they might not be fully representative of the adult kidney, particularly the diseased kidney where a lack of *EPO* has a significant effect. Careful consideration is therefore needed when interpreting results. Furthermore, due to the complex karyotype, it is difficult to know with certainty that the effects measured throughout this thesis are due to disrupting all copies of the expected genes and are not resultant of extra copies of some genes/residual expression due to extra chromosomal copy numbers.

2.2.2.2 *Cell maintenance*

Human Embryonic Kidney-293 (HEK-293) cells were cultivated in Dulbecco’s Modified Eagle’s Medium (DMEM) containing GlutaMAX^{TM} (ThermoFisher Scientific, Massachusetts, USA) supplemented with filtered 10% foetal bovine serum (FBS) (ThermoFisher Scientific, Massachusetts, USA) and incubated at 37^{0}C, 5% CO_2 to mimic *in vivo* conditions. Cells were routinely passaged at 80-90% confluency and media changed every 2-3 days or as required.

2.2.2.3 *Cell Passage*

Cells were passaged once they reached 80-90% confluency to maintain the cell-line. All media was pre-warmed to 37 ^{0}C in a bead bath prior to passaging. Media was removed from the cells using an aspirator. TryPLE (ThermoFisher Scientific, Massachusetts, USA) (3 mL for 10 cm plate, 1 mL for 6-well plate, 500 µl for 24-well plate) was added to enable break-down of the extra-cellular matrix for collection and cells returned to the incubator for 5 minutes. Cells were then collected into a sterile falcon tube and the remaining plate/well was washed with media to collect any remaining cells. The cell suspension was centrifuged for 5 minutes at 1,000 rpm to pellet the cells. The supernatant was removed and the pellet resuspended in 10 mLs of fresh pre-warmed media by pipetting up and down ~10 times using a sterile stripette. Cells were usually split at a 1:10 dilution (1 mL of cell resuspension added to 9 mL of media) unless a specific number of cells was required in which case cell counting (see **2.2.2.4**) was undertaken.

2.2.2.4 *Cell Counting*

Cells which required to be at a certain density underwent cell counting using a haemocytometer before being added to the plate/dish. Following resuspension in 10 mL of media during passaging, 10 µl of resuspension was added to a haemocytometer. The number of cells in the haemocytometer chamber can be determined directly by counting the number of cells present using a microscope. The number of cells within the chamber is then used to calculate the concentration (or density) of cells in the resuspension. The cell density can be calculated by dividing the number of cells in the chamber by the volume of the chamber, which is known beforehand, accounting for any dilutions made. **Equation 2.5** was used when counting the number of cells present in four large corner quadrants. Once the stock cell density has been calculated using **Equation 2.5**, the volume of resuspension needed to obtain the required density of cells was calculated using **Equation 2.6**

$$Stock\ cell\ density\ (Number\ of\ cells/mL) = \left(\frac{Number\ of\ cells\ counted}{4}\right) x\ 10^4\ x\ dilution\ factor$$

Equation 2.5: Counting the number of cells per mL.

$$C1\ x\ V1 = C2\ x\ V2$$

C1: Concentration of cells in the stock resuspension

V1: the volume of cell resuspension needed to make the working concentration

V2: the working volume required

C2: the concentration desired

Equation 2.6: The volume of resuspension needed to obtain the required density of cells.

Here is a worked example if 150 cells were counted over the four corner quadrants from 10 mLs of resuspension.

$$Stock\ cell\ density\ (Number\ of\ cells/mL) = \left(\frac{150}{4}\right) x\ 10^4\ x\ 1$$

$$Stock\ cell\ density\ (Number\ of\ cells/mL) = 375{,}000\ cells/mL$$

For a required cell density of 50,000 cells per mL in a required volume of 5 mLs:

$$V1 = \frac{50{,}000\ x\ 5}{375{,}000}$$

V1 (the volume of cell resuspension (mL)) = 0.67 mL

Therefore, to achieve a cell density of 50,000 cells per mL in 5 mLs, you need 670 µl of cell resuspension and the rest of media to make up to 5 mLs.

2.2.2.5 *Single cell isolation*

Using a Leica DMi8 Widefield microscope, single cells were isolated by single-cell picking using a 0.1-2 µl pipette and transferred to a 96-well culture plate. Single cells were then incubated at 37 ^{0}C and 5% CO$_2$.

2.2.2.6 *Clonal expansion*

Following single cell isolation, single isolated were cultured in 96-well plates for several weeks until 80-90% confluency had been achieved. Cells were then passaged into 24-well plates until confluency was reached and then subsequently passaged into 6-well plates and then 10-cm plates.

2.2.2.7 *Cryogenic preservation*

For long-term storage of cell-lines, cells underwent cryogenic preservation and were stored in liquid nitrogen. Cells were pelleted following the same protocol for passaging and then resuspended in 1 mL of media supplemented with 10% dimethyl sulfoxide (DMSO) (Sigma Aldrich, Missouri, USA) or resuspended in 1 mL of Cell Banker (Amsbio, Abingdon, UK). The 1 mL of resuspension was split between two cryovials and stored in Mr Frosty^{TM} (ThermoFisher Scientific, Massachusetts, USA) at -80 ^{o}C for at least 24 hours. Mr Frosty^{TM} (ThermoFisher Scientific, Massachusetts, USA) cools the cells down at a rate of -1 ^{o}C per minute to prevent cell lysis. Cells were then transported on dry ice to liquid nitrogen stores where they are stored for long-term at -196 ^{o}C.

2.2.2.8 *Recovery of cells from cryogenic preservation*

To recover cells from long-term storage, cells are defrosted quickly in a water bath at 37 ^{0}C. Cells are transferred to an Eppendorf and an equal volume of pre-warmed media is added to prevent cell death and dilute DMSO (Sigma Aldrich, Missouri, USA). Cells are then pelleted by centrifugation at 1,000 rpm for 5 minutes. The supernatant was removed and pellet resuspended in 10 mLs of media. The resuspended pellet was transferred to the same size plate from which it was originally frozen from and incubated at 37 ^{0}C and 5% CO_2 until confluent. Cells were passaged at least twice before being used for experiments.

2.2.3 Nucleic acid extraction

To extract nucleic acid (genomic DNA or total RNA), cultured cells were grown to confluency and then disassociated using TryPLE (ThermoFisher Scientific, Massachusetts, USA). Dissociated cells were split equally into two 15 mL falcon tubes and then centrifuged at 1,000 rpm for 5 minutes. The supernatant was removed. The pellet of one falcon tube was re-plated for continual growth (see 2.2.2.3), whilst the other pellet was either frozen at -20 ^{o}C until genomic DNA extraction, or resuspended in 250 µl of TRIzol^{TM} reagent (ThermoFisher Scientific, Massachusetts, USA) and stored at -80 ^{o}C for RNA isolation. All plasticware used for nucleic acid extraction were sterile and RNase/DNase-free. All surfaces and pipettes were cleaned with ethanol and/or RNaseZap (ThermoFisher Scientific, Massachusetts, USA) prior to extraction to prevent any degradation by RNAses.

2.2.3.1 *Genomic DNA isolation*

Genomic DNA was isolated from cell pellets using the PureLink Genomic DNA Extraction Kit (Invitrogen, Massachusetts, USA). A heat block was set to 55 ^{o}C prior to starting DNA extraction and 96-100% ethanol was added to Wash Buffer 1 and 2 according to instructions on the bottle. The whole cell pellet was resuspended in 200 µl of PBS and transferred to a 1.5 mL Eppendorf. 20 µl of Proteinase K and 20 µl of RNase A were added to the sample and mixed by vortexing. The sample was incubated at room temperature for 2 minutes before adding 200 µl of PureLink Genomic Lysis/Binding Buffer and mixing by vortexing. Samples were incubated for 10 minutes at 55 ^{o}C on the heat block. After incubation, 200 µl of 100% ethanol was added before mixing for 5 seconds by vortex. The sample was transferred to a PureLink Spin Column and centrifuged at 10,000 g for 1 minute at room temperature. The collection tube was discarded and the spin column placed in a clean PureLink collection column. 500 µl of Wash Buffer 1 was added to the to the column and centrifuged at 10,000g for 1 minute at room temperature. The PureLink collection tube was discarded and the spin column placed into a clean collection tube before added 500 µl of Wash Buffer 2 to the spin column and centrifuging for 3 minutes at 13,000 rpm at room temperature. The PureLink spin column was placed inside a 1.5ml Eppendorf tube. To elute the DNA, 25 µl of ddH$_2$O was added directly to the centre of the spin column, incubated for 1 minute at room temperature and then centrifuged at 13,000 rpm for 1 minute. A second elution step was performed adding an additional 25 µl of ddH$_2$O to the spin column and centrifuging at 13,000 rpm for 1.5 minutes. The purified DNA was stored at -20^{0}C for long-term storage. Quality and purity of DNA was measured by the Nanodrop Spectrometer as described in **2.2.4.1.**

2.2.3.2 *Total RNA isolation*

Total RNA was purified directly from samples preserved in TRIzol□ reagent using the Direct-zol^{TM} RNA Miniprep kit (Cambridge Biosciences, Cambridge, UK). All reagents were supplied with kit unless otherwise stated. Briefly, 500 µl 96 - 100% ethanol was added to the sample lysed in TRIzol^{TM} (ThermoFisher Scientific, Massachusetts, USA and mixed by reverse pipetting. The mixture was transferred to a Zymo-Spin^{TM} ICR column within a collection tube and centrifuged at 13,000 rpm for 30 seconds. The flow-through was discarded and the column placed in a clean collection tube. 400 µl of RNA Wash Buffer prepared with the recommended amount of 96 - 100% ethanol was added. In an RNase/DNase free tube, 5 µl of DNase I was mixed, by gentle inversion, with 75 µl of DNA digestion buffer per sample undergoing total RNA isolation. 80 µl of this digestion mix was then added directly to the spin column matrix. The column was then incubated at room temperature for 15 minutes to allow digestion of DNA. 400 µl of RNA PreWash Buffer prepared with the recommended amount of 96 - 100% ethanol was added to the column and centrifuged at 13,000 rpm for 30 seconds. The flow-through was discarded and the step repeated. 700 µl of RNA Wash Buffer was then added to the column and centrifuged at 13,000 rpm for 1 minute to ensure complete removal of the wash buffer. The column was transferred to an RNAse/DNase Free 1.5 mL Eppendorf tube. To elute higher concentrations of RNA, 30 µl of DNase/RNase Free-Water was added directly to the column matrix and centrifuged for 30 seconds at 13,000 rpm. 5 µl of RNA was aliquoted into a separate DNase/RNase Free tube for checking the quality of RNA to prevent risk of contamination. Eluted RNA was stored at -80 ^{0}C for long-term storage until needed. RNA was quantified and checked for quality, purity, and integrity using the Qubit^{TM} 2.0 Fluorometer (ThermoFisher Scientific, Massachusetts, USA), and an Agilent 2020 TapeStation with RNA ScreenTape (Agilent Technologies, California, USA), respectively as described in **2.2.4.**

2.2.4 Quantification and Quality check of nucleic acid

Following extraction of nucleic acid, samples were quantified using the Nanodrop ND-8000 spectrophotometer (ThermoFisher Scientific, Massachusetts, USA) for DNA, or the Qubit^{TM} 2.0 Fluorometer (ThermoFisher Scientific, Massachusetts, USA) for RNA. RNA quality, purity, and integrity were assessed using an Agilent 2020 TapeStation with RNA ScreenTape (Agilent Technologies, California, USA).

2.2.4.1 *Nanodrop*

After purification of DNA, the yield and purity were calculated using the Nanodrop ND-8000 spectrophotometer (ThermoFisher Scientific, Massachusetts, USA) by measuring the absorption at 260 nm (A260) of 1 µl of undiluted sample. The Nanodrop software automatically calculated the concentration (in ng/µl) using a modified Beer-Lambert equation. DNA purity was calculated by measuring absorption at 280 nm (A280) and 260 nm (A260); nucleic acids absorb ultra-violet (UV) light at 260 nm whilst proteins absorb UV at ~280 nm. A ratio of A260/A280 of ~1.8 indicates ‘pure’ DNA. A secondary measure of absorbance at 230 nm (A230) was also taken as common contaminants such as ethanol and phenol absorb UV at ~230 nm. A A260/A230 ratio between 1.8-2.2 indicates a high purity sample. The Nanodrop was blanked using 1 µl of ddH_20 before measuring samples.

2.2.4.2 *Qubit Fluorometer*

The Qubit^{TM} 2.0 Fluorometer (ThermoFisher Scientific, Massachusetts, USA) uses a fluorescent dye that emits signal only when bound to RNA even in the presence of free nucleotides or contaminants. 200 µl of Qubit Working solution was made for each sample by diluting the Qubit RNA BR Reagent 1:200 with Qubit RNA BR Buffer. 199 µl of working solution was then mixed with 1µl of RNA sample. Standards were also prepared using 190µl of Qubit Working Solution and 10 µl of Standard provided by the Qubit BR kit. Standards and samples were vortexed briefly and incubated at room temperature for 2 minutes before the RNA was quantified using the Qubit^{TM} 2.0 Fluorometer (ThermoFisher Scientific, Massachusetts, USA) by inserting the tubes directly into the machine.

2.2.4.3 *TapeStation*

RNA samples being sent for sequencing also underwent a further check for quality, purity and integrity using an Agilent 2020 TapeStation with RNA ScreenTape (Agilent Technologies, Calafornia, USA). The TapeStation software is an automated electrophoresis tool for analysis RNA quality and works by calculating an RNA integrity number (RIN) using an algorithm that considers several regions of the recorded electropherogram particularly certain peaks in the 18S and 28S subunits of ribosomal RNA (Scraeber, 2006). The RIN can take a number between 0-10 with 10 indicating high RNA integrity and 1 indicating degradation of RNA and thus very low RNA integrity. Samples were prepared in an 8 tube PCR strip by mixing 5 µl of RNA sample buffer with 1 µl of RNA sample by reverse pipetting. Samples were then spun down and vortexed for 1 minute at 2000 rpm. Samples were spun down again and then heated at 72 ^{0}C for 3 minutes followed by being placed on ice for 2 minutes. Samples were then analysed in the Agilent 2200 TapeStation instrument. Briefly, the tubes were placed in the sample block, loading tips were inserted into the loading tip holder on the instrument and the RNA ScreenTape device was inserted into the instrument, after being flicked gently to remove bubbles, with the label facing towards the front and the barcode to the right. The tubes wishing to be run were selected on the software and the run was started. The subsequent machine read-out provided the RIN score for each sample.

2.2.5 Polymerase Chain Reaction (PCR)

Polymerase Chain Reaction (PCR) is used to amplify a region of DNA. The reaction relies upon the presence of a DNA polymerase and primers specific to a certain region of the DNA enabling amplification of this specific region. The reaction undergoes several cycles of heating and cooling in a thermocycler. The first step of the PCR reaction is to heat the reaction activating the heat-sensitive polymerase. The following three steps are then repeated for a certain number of cycles to enable amplification of the desired sequence:

1. Denaturing: the reaction mix is heated to 95 ^{0}C to denature the double-stranded DNA
2. Annealing: the reaction mix is cooled to a primer specific temperature, called the melting temperature (usually between 50 ^{0}C – 65 ^{0}C), to

enable specific binding of the primers and prevent non-specific amplification.

3. Elongation: the reaction mix is heated to 72 ^{0}C to allow the polymerase to synthesise the complementary strand of DNA using the dNTPs. This elongation step is carried out for 30 seconds per 500 bp i.e. for 750 bp amplicon, elongation will be 45 seconds.

The number of cycles that these three steps are run for is typically between 35 and 45 and is dependent upon the application. The number of amplicons present in the reaction mix is equivalent to 2^n with n being the number of cycles. A final elongation step is then performed at 72 ^{0}C at the end to allow for final extension.

2.2.5.1 *Primer design*

DNA sequences were imported into Benchling from the UCSC Genome Browser (https://genome.ucsc.edu/). Primer pairs were then designed by using Primer3 (www.bioinformatics.nl) ensuring pairs adhered to the following rules:

1. Sequences between 20 – 25 bp
2. Similar melting temperature between both sequences
3. GC content between 45 – 55 % in both pairs
4. Few repetitive sequences in the sequences
5. *In silico* PCR analysis was performed using USCS Genome Browser (https://genome.ucsc.edu/cgi-bin/hgPcr) to ensure 100 % matching with the desired sequence and no matches elsewhere in the genome.

Primers were then purchased from IDT (Integrated DNA Technologies, Leuven, Belgium; https://eu.idtdna.com/) and resuspended at 100 µM with molecular biology grade water.

2.2.5.2 *Optimisation of PCR conditions*

The optimal annealing temperatures for primer pairs was determined using a temperature gradient PCR reaction in an Eppendorf Mastercycler thermocycler. PCR reactions were set-up for each primer pair altering only the annealing temperature by 2 ^{0}C increments across the PCR block between 52 ^{o}C and 64 ^{0}C. PCR products were visualised using gel electrophoresis as outlined below in **2.2.6.** The annealing temperature showing the strongest resulting band of the correct size was chosen for subsequent reactions.

2.2.5.3 *PCR reaction*

The PCR reaction used throughout this thesis when performing PCR is outlined in **Table 2.4** and the cycling conditions are shown in **Table 2.5** with the melting temperature and elongation time changing to suit the primers and the desired amplicon size.

Table 2.4: PCR reagents and volumes used in standard PCR throughout this thesis. *All reagents (except the primers) were from Solis Biodyne (Teaduspargi, Estonia).*

Reagent	**Volume (µl)**
10X Buffer B1	2
10 µM dNTPs	0.4
25 mM $MgCl_2$	1.5
10 µM Fwd primer	0.5
10 µM Rev primer	0.5
HotStart Taq Polymerase	0.2
100 ng DNA	X
ddH_20	Make up to 20µl
Total	**20µl**

Table 2.5: Thermocycling conditions used for standard PCR throughout this thesis. **The annealing temperature specific to the primer pair is determined through optimisation using a gradient PCR. This temperature is typically between 50 ^{0}C and 65 ^{0}C. ** the time for elongation is dependent upon the amplicon size and the enzyme being used. For HotStart Taq Polymerase it is chosen based upon synthesis of 500 bp every 30 seconds.*

Stage	**Temperature (^{0}C)**	**Time**	**Number of cycles**
Enzyme activation	95	15 minutes	1 cycle
Denaturation	95	30 seconds	35 cycles
Primer annealing	*	30 seconds	
Elongation	72	**	
Extension	72	10 minutes	1 cycle
	15	Infinity	

2.2.6 Agarose Gel electrophoresis

Gel electrophoresis is used to separate out different sized fragments of DNA based on their size and charge following a PCR reaction or restriction digest. A gel is a 3D matrix composed of pores that the negatively-charge DNA migrates through once an electric field is applied. Longer DNA molecules are unable to migrate as far and as quickly through the gel compared to shorter DNA molecules. An agarose gel was made by combining agarose powder with 1 X TAE (Tris-acetate-EDTA) Buffer (40 mM Tris, 20 mM acetic acid, 1 mM EDTA) to achieve different percentages which are dependent upon the size of the fragments being separated; low percentage gels have larger pores for larger fragments to move through more easily. The recommended gel percentage for different sized fragments is outlined in **Table 2.6**. 1 X SYBR DNA Gel Stain (ThermoFisher Scientific, Massachusetts, USA) was then added to the mixture to enable visualisation of the DNA bands. The agarose gel mixture was loaded into electrophoresis apparatus and left to set for around 20 minutes. 10 µl of PCR product was mixed with 5 X OrangeG (Sigma Aldrich, Missouri, USA), a loading dye, and 8 µl of this mixture was added to the wells of the agarose gel alongside a 1 kb or 100 bp Ready-to-Load DNA ladder (Solis BioDyne, Teaduspargi, Estonia). Gel electrophoresis was then run at 120 V for 45 minutes to allow DNA migration through the gel. Once the gel has finished running, the agarose gel was imaged and bands visualised using the Licor Oddysey Imaging system (LI-COR Biosciences Ltd, Cambridge, UK).

Table 2.6: Recommended agarose gel percentage for visualisation of different DNA fragments. *kb = kilobases, bp = base-pairs, w/v = weight/volume.*

Agarose gel percentage (w/v)	Range of effective separation
0.5 %	1 kb – 30 kb
0.7 %	800 bp – 12 kb
1.0 %	500 bp – 10 kb
1.2 %	400 bp – 7 kb
1.5 %	200 bp – 3 kb
2.0 %	50 bp – 2 kb

2.2.7 Enzymatic PCR purification

After PCR, the amplicons can be used for sequencing or restriction digestion. Therefore, it is good practice to clean-up the PCR reaction by removing any buffers, salts, unused dNTPs, or primers which may inhibit or disrupt the downstream applications. PCR products were therefore purified using the ExoSap-IT PCR Cleanup kit (ThermoFisher Scientific, Massachusetts, USA). 5 µl of PCR product was mixed with 2 µl of the Exo-Sap^{IT} reagent and run in a thermocycler for 15 minutes at 37 ^{o}C to degrade the remaining primers and nucleotides followed by 15 minutes at 80 ^{o}C to inactivate the reagent.

2.2.8 Quantitative reverse-transcription PCR (qRT-PCR)

Quantitative reverse-transcription PCR (qRT-PCR) is used to relatively quantify, in real-time, the levels of RNA transcripts within a cDNA sample. The real-time detection is made possible by the inclusion of a fluorescent molecule, commonly a DNA-binding dye (such as EVAgreen), in the reaction that reports an increase in the amount of DNA with a proportional increase in fluorescent signal. The fluorescence is measured by a specialised thermal cycler equipped with fluorescence detection modules and the amount of fluorescence reflects the amount of amplified product in each sample. During the PCR reaction, the amount of PCR product doubles during each cycle (the exponential phase) until the reaction components become limited and the reaction reaches a plateau phase. Initially, despite the PCR product doubling, the fluorescence remains at background level and is undetectable. Eventually, enough amplified product accumulates and the fluorescence is detectable; the cycle number at which this occurs is known as the threshold cycle (C_T). This is measured during the exponential phase when reagents are not limited and is therefore used to accurately and reliably calculate the initial amount of template in the reaction. If a large amount of template is present at the start of the reaction indicating that the gene is more highly expressed, fewer amplification cycles are required to provide a fluorescent signal above the background and therefore the reaction will have a lower C_T value. A lower expressed gene with less starting template present will require more cycles to provide a fluorescent signal above the background and will thus have a higher C_T value. The C_T values can be compared across samples to calculate the relative fold change in gene expression compared to a control group.

2.2.8.1 *cDNA synthesis*

Complementary DNA (cDNA) was synthesised from total purified RNAs using the PrimeScript^{TM} RT reagent Kit (Takara Bio Europe SAS, Saint-Germain-en-Laye, France). A master mix was prepared consisting of 2 µl 5X PrimeScript Buffer, 0.5 µl PrimeScript RT Enzyme Mix I, 0.5 µl of 50 µM OligodT Primer, 0.5 µl of 100 µM Random 6mers and mixed with 500 ng RNA. The final reaction volume was made up to 10 µl with RNase/DNase Free Water. The reaction mixture was then incubated in a thermocycler at 37 ^{o}C for 15 minutes to enable reverse transcription followed by 5 seconds at 85 ^{o}C to inactivate the reverse transcriptase enzyme. A 100% conversion RNA to cDNA efficiency was assumed to calculate the concentration of cDNA at 50 ng/µl.

2.2.8.2 *cDNA PCR*

A cDNA PCR reaction was run to check the specificity of the primers and ensure a product of the expected size was amplified. Where possible, primers were designed to cross exon-exon junctions to ensure they were specific to cDNA and would therefore only amplify cDNA. A PCR reaction of 2 µl 10 X Buffer I, 1.5 µl $MgCl_2$, 0.4 µl dNTPs (10 mM), 0.5 µl of forward primer (10 µM), 0.5 µl reverse primer (10 µM), 100 ng cDNA and ddH_20 to make a total reaction volume of 20 µl was prepared. Cycling conditions are outlined in **Table 2.7.** PCR products were separated and visualised using a 1.0 % agarose gel in 1 X TAE buffer containing SYBR DNA Gel Stain run at 120 V for 45 minutes.

2.2.8.3 *qRT-PCR*

qRT-PCR was performed using Hot FIREPol EvaGreen^{TM} qPCR Master Mix with ROX (Solis BioDyne, Teaduspargi, Estonia) which was then run using the QuantStudio 6 Flex qPCR machine (ThermoFisher Scientific, Massachusetts, USA). qRT-PCR reactions were set up in 384 well plates by mixing 1 µl of 5 X Hot FIREPol EvaGreen qPCR Mix Plus ROX (Solis BioDyne, Teaduspargi, Estonia), 0.125 µl forward primer (10 µM), 0.125 µl reverse primer (10 µM), 1 µl 1:15 diluted cDNA, 2.25 µl ddH_20 in a 5 µl reaction volume. Cycling conditions are listed in **Table 2.8.** Reactions were carried out on at least three biological replicates and three technical replicates.

2.2.8.4 *Data analysis*

The resulting SDS output files were uploaded to the ThermoFisher Cloud (ThermoFisher Scientific, Massachusetts, USA) and analysed using the Relative Quantification qPCR App within the software. This platform was used to correct C_T values for their efficiency and to ensure there were no apparent outliers before further analysis. Output was imported into Excel and the C_T values were used for analysis using the comparative C_T method (Schmittgen & Livak, 2008). The C_T values are the cycle numbers where the fluorescence generated by the PCR is distinguishable from the background noise. First the average C_T value was calculated for each technical replicate before determining the difference in gene expression between the gene interest and the housekeeping gene (**Equation 2.7).** This step is carried out to normalise expression of the gene of interest to a gene not affected by the experiment. The housekeeping gene used for normalisation was that with the most stable gene expression which was determined from the raw data using the RefFinder webtool (Xie et al. 2012). RefFinder returned the geometric mean value across all housekeeping genes measured as well as the geometric mean value across housekeeping genes combined and determined which gene/combination of gene were the most stable, and thus the most appropriate for the ΔC_T normalisation step. The difference between the experimental samples and the controls samples was then calculated before determining the relative fold gene expression level (**Equation 2.7).** Fold-changes were log transformed before performing statistical analysis of the gene expression values to account for skewed distributions. Differences in gene expression levels between samples and controls were investigated for statistical significance by a paired *t*-test carried out in RStudio version.3.6.1. A paired *t* test was chosen as I was comparing the means across two different groups (i.e. knock-out versus control) (Mishra et al., 2019). Statistically significant differences were determined using the typically accepted P-value threshold < 0.05 (Mishra et al., 2019).

$$Step\ 1{:}\ Average\ the\ Ct\ values\ for\ any\ technical\ replicate$$

$$\text{Step 2: } \Delta\text{Ct} = \text{Ct (gene of interest)} - \text{Ct (housekeeping gene)}$$

$$\text{Step 3: } \Delta\Delta\text{Ct} = \Delta\text{Ct (experimental sample)} - \Delta\text{Ct (untreated sample)}$$

$$Step\ 4{:}\ Fold\ Change = 2^{-\Delta\Delta\text{Ct}}$$

Equation 2.7 Delta Delta C_T method for calculating the fold-change of the relative mRNA expression in tested samples compared to wild-type controls. *The symbol Δ refers to delta, which is a mathematical term used to describe the difference between two numbers.*

Table 2.7: Thermocycling conditions for cDNA PCR of the qPCR primers to check specificity and for amplicons of the expected size.

Stage	**Temperature (^{0}C)**	**Time**	**Number of cycles**
Denaturing	95	15 minutes	1 cycle
Annealing	95	20 seconds	35 cycles
	Primer specific T_m	30 seconds	
	72	1 minute	
Elongation	72	5 minutes	1 cycle
	4	∞	

Table 2.8: Thermocycling conditions used for qPCR reactions.

Stage	**Temperature (^{0}C)**	**Time**	**Number of cycles**
Hold Stage	95	15 minutes	1 cycle
Denaturation	95	15 seconds	45 cycles
Annealing	Primer specific T_m	20 seconds	
Extension	72	20 seconds	
Melt Curve analysis	95	15 seconds	1 cycle
	60	1 minute	
	95	15 seconds	

Chapter 3 Genetically proxied therapeutic increases in endogenous EPO levels is not associated with increased cardiovascular risk.

This Chapter includes sections that have been taken directly from a pre-print paper in which I am the first author. This paper is currently undergoing reviews at GSK and AJHG.

Harlow, CE. Gandiwijaya, J. Bamford, RA. Wood, AR. Van der Most, P. Verweij, N. [25 authors] & Frayling, TM. 2022. Identification and single-base gene-editing functional validation of a *cis-EPO* variant for use to mimic novel EPO-increasing therapies.

I performed most of the data analysis in this Chapter with the support and guidance of my supervisor, Professor Tim Frayling. The GWAS and meta-analysis of circulating EPO described were carried out as part of the EPO consortium which consisted of four independent studies. I undertook the GWAS of EPO in study participants of the InCHIANTI study. Collaborators within the EPO consortium undertook GWAS in the other three studies (BLSA: Toshiko Tanaka, Health ABC: Hampton Leonard, PREVEND: Niek Verweij). I was the main analyst and undertook all the quality control steps and performed the meta-analysis. The eQTL data used was produced by collaborators who had access to the eQTL data (Hepatic: Amy Etheridge and Renal: Peter Van der Most). The GWAS analysis of UK Biobank was carried out inhouse within the Genetics of Complex Traits Team. Dr Andrew Wood and Dr Robin Beaumont prepared the imputed genotypes and defined ancestry based on Principal Component Analysis (PCA). Dr Jessica Tyrell and Dr Kate Ruth generated the phenotypes.

3.1 Introduction

Anaemia, one of the primary complications of chronic kidney disease (CKD), affects one out of every seven CKD patients (Hill et al., 2016; St Peter et al., 2018; Stauffer & Fan, 2014). Anaemia is not only associated with faster progression of CKD but also with increased risk of adverse events, particularly heart disease or stroke, two of the major causes of death in CKD patients (Cases et al., 2018; Q. Zheng et al., 2021). Current therapies used to treat anaemia in CKD include blood transfusions, intravenous iron therapies or parenteral injections of recombinant erythropoietin (rhEPO), the latter two which attempt to increase erythropoiesis and restore oxygen levels (Bonomini et al., 2016; Kaplan et al., 2018; Parfrey, 2021). However, these treatments are limited. Blood transfusions increase the risk of infection and alloimmunisation. Oral iron therapies have poor compliance due to gastrointestinal adverse effects and intravenous iron or rhEPO are inconvenient because they require injections, the need for cold-chain transport and storage and have increased risk of adverse side-effects including hypertension (with rhEPO) and hypersensitivity (with intravenous iron) (Babitt & Lin, 2012; Bonomini et al., 2016; Krapf & Hulter, 2009; Portolés et al., 2021; Q. Zheng et al., 2021). Furthermore, rhEPO or its analogues raises additional safety concerns since previous clinical trials and studies have indicated an increased risk of stroke, myocardial infarction (MI), venous thromboembolism, and heart failure possibly due to the sudden supra-physiological erythropoietin (EPO) levels causing an excessive rise in haemoglobin (Hgb) levels (Fishbane & Spinowitz, 2018; Jelkmann, 2013; Pfeffer et al., 2009). These safety concerns have led to the development of hypoxia-inducible factor (HIF) prolyl hydroxylase inhibitors (PHIs) as a novel class of treatment for anaemia in CKD.

PHIs work at the transcriptional level of the hypoxic response genes, including *EPO*, by stabilising HIFs through inhibition of the prolyl hydroxylase enzymes (PHD1-3) (Haase, 2013; Sugahara et al., 2017). In turn, by activating the hypoxic response pathway, PHIs increase endogenous EPO levels in a controlled manner, resulting in increased erythrocyte production and development, increased Hgb levels and oxygen tissue delivery (Kaplan et al., 2018). These novel therapies are hoped to be safer and more efficacious than current treatments at correcting the anaemia by maintaining EPO levels within the physiological range, reducing the risk of cardiovascular disease (CVD) or levels of clinical markers for CVD associated risk factors, primarily blood pressure and resting heart rate. Blood pressure (BP) is one of the most major modifiable risk factors for CVD and has the strongest evidence for causation (Fuchs & Whelton, 2020; Vasan et al., 2001), whilst high resting heart rate is one of the strongest predictors of overall cardiovascular morbidity and mortality (Perret-Guillaume et al., 2009).

Recent completion of Phase III clinical trials assessing cardiovascular safety and hematologic efficacy has indicated non-inferiority of PHIs compared to rhEPO and shown that PHIs can increase and maintain Hgb levels with small increases in circulating EPO levels (Akizawa et al., 2021; Akizawa, Iwasaki, et al., 2020; N. Chen, Hao, Liu, et al., 2019; N. Chen, Hao, Peng, et al., 2019; Chertow et al., 2021; K.-U. Eckardt et al., 2021; Fishbane et al., 2021; A. K. Singh, Carroll, McMurray, et al., 2021; A. K. Singh, Carroll, Perkovic, et al., 2021; Q. Zheng et al., 2021). Furthermore, PHIs have already received approval in Japan supporting their ongoing development elsewhere (Akizawa, Nangaku, et al., 2020; Chertow et al., 2021; Dhillon, 2020; Kanai et al., 2021; Parfrey, 2021; A. K. Singh, Carroll, McMurray, et al., 2021).

Several studies have shown that genetic data can provide supporting evidence of an association between the drug target and intended therapeutic indication as well as any potential unintended effects through associations with additional phenotypes to inform potential drug safety (Gill et al., 2019; Lotta et al., 2016; Nelson et al., 2015; Nguyen et al., 2019; Plenge et al., 2013; Scott et al., 2016; Swerdlow et al., 2015). Genetic variants that lie within or nearby the gene encoding the drug target are most likely to have functional impact on the protein product. Genetic variants can be used as unconfounded, unbiased proxies for pharmacological action through drug-target Mendelian Randomisation (MR) to explore the effect of long-term modulation of drug targets on disease outcomes (Davey Smith & Hemani, 2014; Swerdlow et al., 2016; Walker et al., 2017). MR is an analytical method analogous to a randomised control trial (RCT) which relies upon the principal that if a modifiable exposure (e.g. a biomarker) is causal for disease, then a genetic variant associated with or mirroring the biological effects of the exposure will also be associated with the disease

(Burgess et al., 2012). MR relies on the identification of genetic variants associated with the exposure trait, i.e. the biomarker or drug target, and makes several assumptions including that the genetic variant is only associated with the outcome through the exposure (Davies et al., 2018). The genetic variant can therefore be used to test for causality between the exposure and disease providing an estimate of long-term effect (Frayling & Stoneman, 2018). Genetic variation resulting in changes to circulating exposure levels is different to that of physiological variation. For example, focusing on endogenous EPO levels, genetic variation represents the long-term, consistent effects of exposure to elevated circulating EPO levels whilst physiological variation, such as during prolonged exposure to hypoxia in high altitudes, represents an individual's ability to respond to such changes resulting in temporary, short-term fluctuations in circulating levels that are often driven by epigenetic modifications (Childebayeva et al., 2019; Friedmann et al., 2005; Haase, 2013; Suresh et al., 2020).

Association data is often used to select valid genetic variants as proxies for pharmaceutical action. The power to detect a causal association between the exposure and outcome increases when obtaining variant-exposure and variant-outcome association data from independent studies (known as two-sample MR) due to increased sample sizes and reduced risk of bias from Winner's Curse (Burgess et al., 2016; Lawlor, 2016; Sheehan & Didelez, 2019). However, there are limitations with using association data to select valid genetic instruments (Davey Smith & Ebrahim, 2003; Nelson et al., 2015; Plenge, 2016; Porcu et al., 2019). First, neighbouring genetic variants tend to be inherited together and are highly correlated with each other due to linkage disequilibrium (LD) which makes it difficult to distinguish the causal variant driving the association (Bush & Moore, 2012; Flister et al., 2013; Hormozdiari et al., 2014). Second, most genetic variants identified through genome-wide association studies (GWAS) do not directly affect the coding sequence due to residing within the non-coding regions (Dixon et al., 2007; Nica et al., 2010). These variants, therefore, can lie within genomic regulatory elements, overlap promoters, enhancers or open-chromatin regions, and may affect gene expression by altering transcription factor binding (Lichou & Trynka, 2020). Disease associated loci identified through GWAS often contain multiple genes and therefore it is difficult to determine which gene is involved and is being affected by the identified variant (Cano-Gamez & Trynka, 2020). Additional approaches, such as fine mapping, quantitative trait loci (QTL) or colocalisation analysis can help refine the gene involved (Benner et al., 2016; Giambartolomei et al., 2014; Nica et al., 2010; Nicolae et al., 2010; Porcu et al., 2019; Wallace, 2020).

Phenome-wide association studies (PheWAS) are another method in which genetics can be used to help predict the safety and efficacy of drugs where a genetic variant, or combination of variants, associated with the drug function, are tested for associations with a wide-range of phenotypes in large sample sizes (Diogo et al., 2018). PheWAS can help further validate genetic instruments used in MR by identifying associations between genetic variants and other relevant phenotypes likely on the same pathway. PheWAS can also identify pleiotropic effects improving our understanding of 1) biological mechanism of action, 2) additional indications with potential for disease expansion or repurposing, 3) associations with conditions in the opposite direction compared to the primary indication indicating potential unwanted effects or 4) associations with additional indications which may be secondary to the primary indication (Denny et al., 2016; Pulley et al., 2017; Robinson et al., 2018).

3.2 Chapter Aims

In this Chapter, I aimed to use human genetics to examine the long-term effects of genetically proxied modulation of EPO levels on risk of CVD (coronary artery disease [CAD], stroke and MI) or any unwanted effects that may arise due to pharmacological manipulation of circulating EPO levels. To achieve this, I aimed to:

1. Identify a genetic variant that influences circulating EPO levels by performing the first and largest GWAS meta-analysis of circulating EPO levels.
2. Validate the genetic variant as a proxy for long-term therapeutic rises in circulating EPO levels.
3. Use the variant in drug-target MR as a natural mimic for therapeutically altered EPO levels to investigate the long-term effects of elevated endogenous EPO levels on risk of cardiovascular disease (CVD) or

levels of clinical markers (systolic blood pressure [SBP], diastolic blood pressure [DBP] or resting heart rate) predisposing to CVD risk factors (e.g. hypertension).

4. Perform PheWAS using this genetic variant to identify any additional unintended effects associated with long-term rises in endogenous EPO levels.

3.3 Methods

An overview of the steps performed and methods utilised to genetically assess the long-term effects associated with higher circulating EPO levels are outlined in **Figure 3.1.**

3.3.1 Genome-wide association study meta-analysis of circulating EPO levels

To identify genetic variants associated with circulating EPO levels, I performed a genome-wide association study (GWAS) meta-analysis of 6,127 individuals of European and African descent from four independent cohorts; InCHIANTI (N=1,210), PREVEND (N=2,954), BLSA (N=458) and HealthABC (N=1,505). Summary statistics for the four studies can be seen in **Table 3.1.** The analysis plan for the GWAS meta-analysis of circulating EPO levels is outlined in **Figure 3.2.**

3.3.1.1 *Invecchiare in Chianti (InCHIANTI)*

InCHIANTI is a prospective, population-based study of 1,453 individuals aged between 20 - 102 years (1,156 > 65 years) living in the Chianti region of Tuscany, Italy. Data was collected between 1998 and 2000 and included telephone interviews, medical examinations and blood samples. A detailed description of the study has been described previously (Ferrucci et al., 2000).

3.3.1.2 *Baltimore Longitudinal Study of Aging (BLSA)*

BLSA is a longitudinal cohort study conducted by the Intramural Research Program of the National Institute of Aging (NIA) which started in 1958 (Shock, 1984). Healthy volunteers aged above 17 are enrolled in the study and participate in follow-up assessment visits of health, physical and psychological performance every 2 years. Currently, the study population has over 3,200 active participants. An independent institutional review board approved the BLSA study protocol, and participants provided informed consent for all analyses included in this report.

3.3.1.3 *Prevention of Renal and Vascular ENd-stage Disease (PREVEND)*

The PREVEND study (Pinto-Sietsma et al., 2000) is a prospective, observational cohort of 8,592 Groningen inhabitants aged between 28 - 75 years. The main aim of the study is to assess the long-term impact of elevated urinary albumin levels on cardiac-, renal- and peripheral vascular end-stage diseases. Upon enrolment, participants agreed to give a urine sample and answered a questionnaire. Participants are followed up every 2-3 years for a survey on cardiac-, renal- and peripheral vascular morbidity.

3.3.1.4 *The Health, Aging and Body Composition Study (HealthABC)*

HealthABC is a prospective, longitudinal study of 3,075 individuals aged between 70 - 79 years in 1997 and 1998 living in Memphis, Tennessee or Pittsburgh. 42% of participants were of African-American ancestry and 52% were of Caucasian ancestry. Participants were enrolled in the study if they had no disabilities, no life-threatening conditions or difficulties walking quarter of a mile and climbing 10 steps. The study consisted of yearly clinical examinations for 6 years, primarily taking measurements of body composition, strength and function, and biannual phone calls to update health status, followed by bi-annual telephone interviews up until year 16 and examination in year 16 (Simonsick et al., 2001).

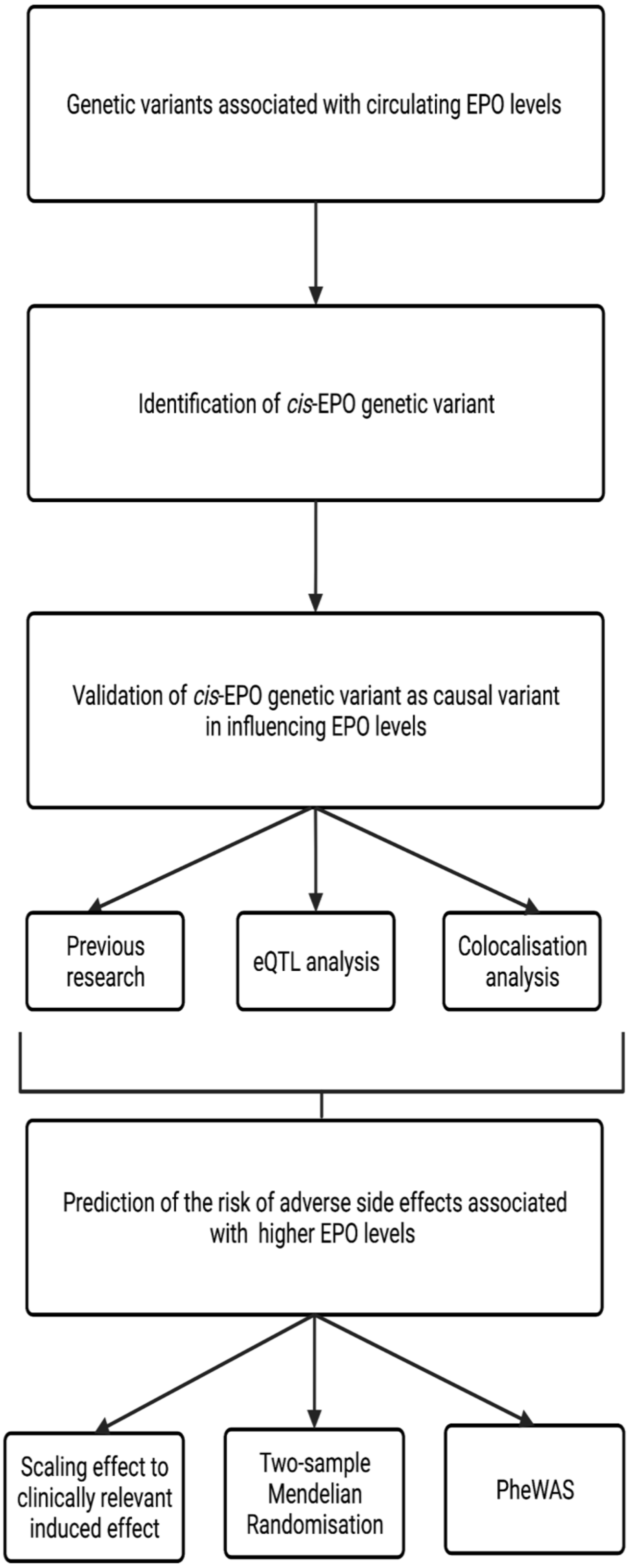

Genetic variants associated with circulating EPO levels
Identification of *cis*-EPO genetic variant
Validation of *cis*-EPO genetic variant as causal variant in influencing EPO levels
Previous research
eQTL analysis
Colocalisation analysis
Prediction of the risk of adverse side effects associated with higher EPO levels
Scaling effect to clinically relevant induced effect
Two-sample Mendelian Randomisation
PheWAS

Figure 3.1: Schematic outline of steps performed in Chapter 2*to identify a genetic variant lying in cis with the EPO gene for use as a partial proxy for therapeutically-altered endogenous EPO levels as a result of activation of the hypoxic pathway through therapeutic PHD inhibition to explore the risk of long-term effects associated with higher endogenous EPO levels.*

Table 3.1: Study characteristics of the four independent studies included in the EPO meta-analysis. *Individuals containing valid genomic data, haemoglobin level data and EPO level data were included in the analysis. SD: standard deviations, PCs: principal components.*

Cohort	**Sample size**	**% Men**	**Mean Age years (SD)**	**Mean EPO IU/L (SD)**	**Mean haemoglobin g/dL (SD)**	**Mean eGFR mL/min/1.73m^2 (SD)**	**Software GWAS implemented in**	**Covariates adjusted for**	**Other sample exclusion criteria based on GWAS data**
InCHIANTI	1210	44.63	66.7(15.3)	9.7 (5.1)	14.1 (1.1)	75.8 (16.0)	GEMMA 0.94.1	Age and Sex	Genotype of phenotype missing data
PREVEND	2954	51.76	53.69 (11.92)	9.03 (14.94)	13.76 (1.23)	80.95 (13.95)	SNPtest v2.5.4	Age, Sex, PC 1-10	Genotype of phenotype missing data, sex mismatch, <95% call rate, PC outliers.
HealthABC Europeans	969	51.7	73.68 (2.77)	12.94 (6.47)	14.16 (1.08)	78.46 (15.95)	Rvtests 2.1.0	Age, Sex, Study site, and PCs	Excess heterozygosity, missingness > 5%, sex mismatch, population outliers, and related individuals were excluded by a GRM cut-off of 0.125 (no closer than cousin)
Health ABC African Americans	536	41.2	73.25 (2.86)	13.63 (1.01)	13.63 (1.01)	88.33 (20.83)	Rvtests 2.1.0	Age, Sex, Study site, and PCs	Excess heterozygosity, missingness > 5%, sex mismatch, population outliers, and related individuals were excluded by a GRM cut-off of 0.125 (no closer than cousin)
BLSA	458	50	69.1 (13.6)	15.2 (1.48)	14.0 (1.1)	72.1 (13.8)	GEMMA 0.94.1	Age, sex and PCs	Phenotype or GWAS missing data

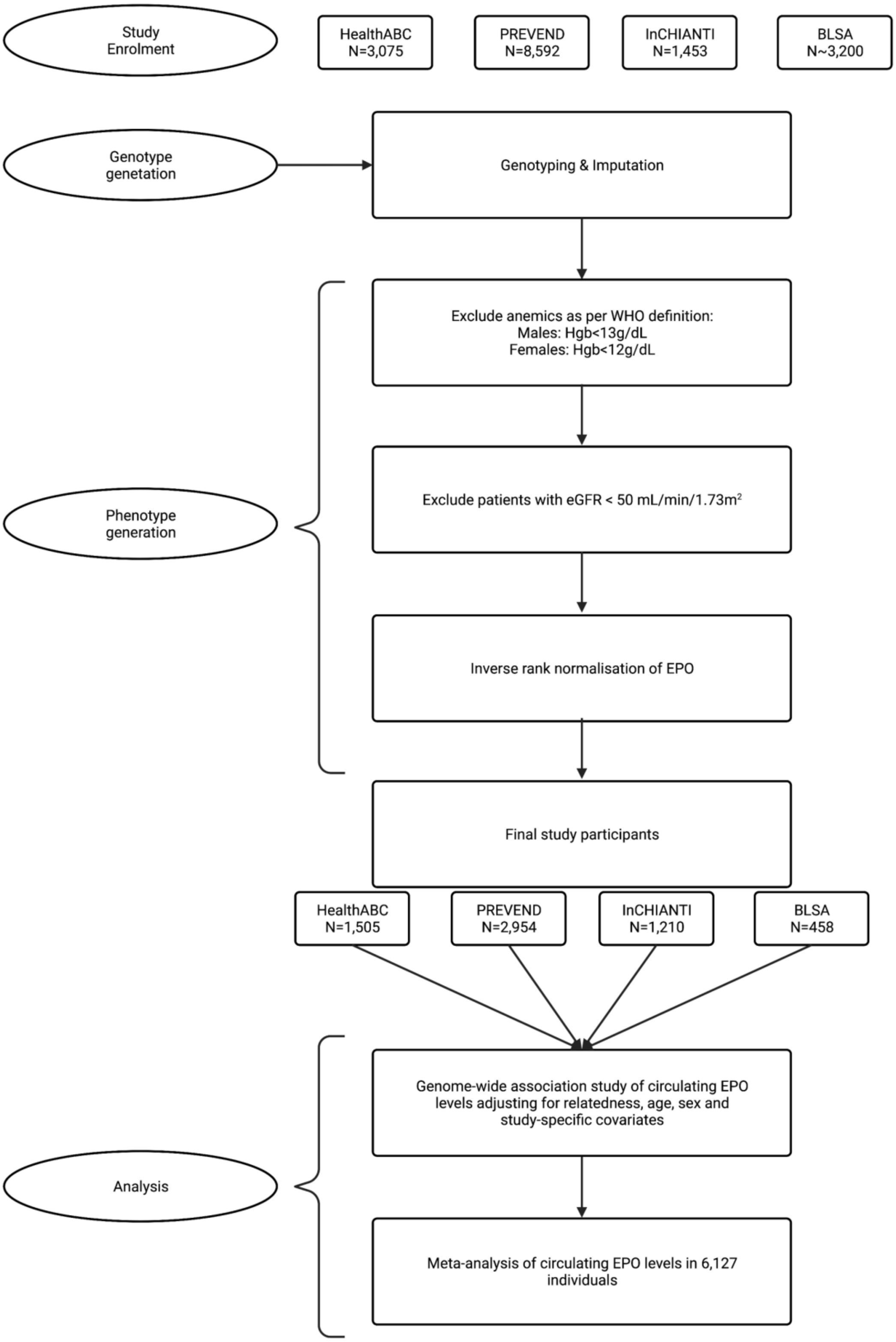

Study
Enrolment
HealthABC
N=3,075
PREVEND
N=8,592
InCHIANTI
N=1,453
BLSA
N~3,200
Genotype
genetation
Genotyping & Imputation
Exclude anemics as per WHO definition:
Males: Hgb<13g/dL
Females: Hgb<12g/dL
Phenotype
generation
Exclude patients with eGFR < 50 mL/min/1.73m2
Inverse rank normalisation of EPO
Final study participants
HealthABC
N=1,505
PREVEND
N=2,954
InCHIANTI
N=1,210
BLSA
N=458
Genome-wide association study of circulating EPO levels adjusting for relatedness, age, sex and study-specific covariates
Analysis
Meta-analysis of circulating EPO levels in 6,127 individuals

Figure 3.2: Analysis plan for the meta-analysis of circulating EPO levels in 6,127 individuals of European and African ancestry. *Hgb: haemoglobin, eGFR: estimated glomerular filtration rate, N: number of individuals.*

3.3.1.5 *Generation of the EPO phenotype*

To generate the EPO phenotype, I included all individuals with valid genomic data, Hgb level data and EPO level data from four independent cohorts (InCHIANTI, PREVEND, BLSA, HealthABC). I excluded anaemic patients as per the WHO definition (Males: Hgb levels < 13 g/dL, Females: Hgb < 12 g/dL) and patients with renal dysfunction based on an estimated glomerular filtration rate (eGFR) threshold of 50 mL/min/1.73m^2 resulting in a final sample size of 6,127 individuals of European and African American descent (InCHIANTI: N = 1,210, PREVEND: N = 2,954, BLSA: N = 458 and HealthABC: N = 1,505) (**Table 3.1).**The standard cut-off for renal dysfunction is 60 mL/min/1.73m^2 but as the study cohorts were on average older than the general population (**Table 3.1)** a lower threshold was used due to lower eGFR rates not being unusual in older populations (Wetzels et al., 2007). Values between 50 - 60 mL/min/1.73m^2 also remained within the normal distribution of each cohort **(Figure 3.3).** The eGFR was calculated using the Modification of Diet in Renal Disease (MDRD) equation **(Equation 3.1)** (Andrew S Levey et al., 2007). I regressed EPO measures on sex and age and performed rank inverse normalisation on the resulting residuals to account for skewed data.

$$\begin{aligned}\boldsymbol{eGFR}\ (\boldsymbol{mL/min\,/m^2}) \\ &= \mathbf{175} \times \left(\boldsymbol{serum\ creatinine}^{-1.154}\right) \times \left(\boldsymbol{Age}^{-0.203}\right) \times \left(\mathbf{0.742}\ \boldsymbol{if\ female}\right) \\ &\times \left(\mathbf{1.212}\ \boldsymbol{if\ African\ American}\right)\end{aligned}$$

Equation 3.1: Estimation of the glomerular filtration rate (eGFR) using the MDRD equation.

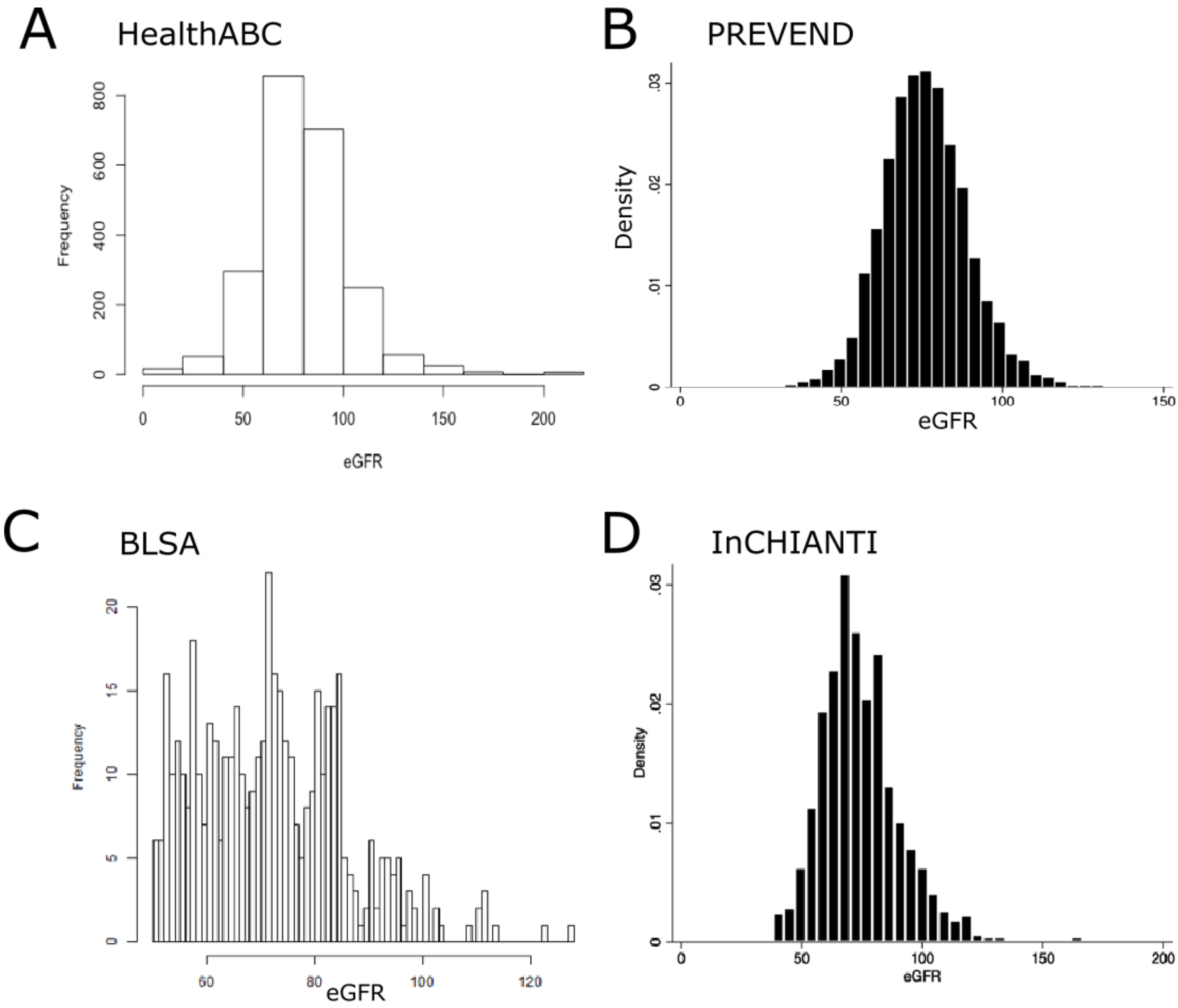


***Figure 3.3: Distribution of estimated GFR (eGFR) in the four independent cohorts**that contributed to the EPO meta-analysis (HealthABC **(A),** PREVEND **(B),** BLSA **(C)**, InCHIANTI **(D)** prior to exclusion of individuals with renal dysfunction (eGFR < 50mL/min/m^2). In each cohort, the eGFR was estimated using the MDRD equation. Those with an eGFR 50 mL/min/1.73m^2 were excluded from the analysis.*

3.3.1.6 *Imputation and Phasing*

Chromosomes 1-23 were included and genotype data was reported using NCBI b37 (hg19) coordinates. For Europeans, imputation was carried out to the Haplotype Reference Consortium (HRC) version 1.1 using MiniMac3 (http://genome.sph.umich.edu/wiki/Minimac), whilst for African Americans, imputation was carried out to CAAPA (Das et al., 2016). Phasing was carried out using Eagle version 2.3 (Das et al., 2016).

3.3.1.7 *Association analysis*

GWAS was performed in GEMMA (Zhou & Stephens, 2012) using an additive linear mixed model adjusting for any study-specific covariates, such as study site and PCs, alongside a genomic relationship matrix (GRM) to account for all types of relatedness (**Table 3.1**). After performing GWAS, quality controls checks were undertaken and any single nucleotide polymorphisms (SNPs) with allele frequencies > 4 standard deviations (SD) or < -4 SD from the HRC allele frequency were excluded (McCarthy et al., 2016). Study-specific estimates were combined and an inverse variance weighted fixed-effects meta-analysis on ~25.1 million imputed SNPs in 6,127 unrelated individuals of European and African American descent was performed using METAL (Willer et al., 2010) with the following filters: minor allele count (MAC) > 3, effect allele frequency (EAF) > 1, EAF < 0, info >= 0.3. After performing meta-analysis, SNPs with a minor allele frequency (MAF) < 0.01 were excluded and a multi-SNP-based step-wise conditional and joint association analysis was employed using GCTA-COJO (J. Yang et al., 2011, 2012) to select SNPs independently associated with endogenous EPO levels (defined as $P < 5 \times 10^{-08}$).

3.3.2 Identification and validation of *cis-EPO* genetic variant

To identify a genetic variant most likely to have direct functional impact on the protein product for use as a genetic proxy for therapeutically altered endogenous EPO levels, I analysed the GWAS data around the *EPO* gene specifically to identify any *cis*-acting genetic variants. I selected variants previously identified to have a functional effect on EPO levels (Amanzada et al., 2014; Kästner et al., 2012; Khabour et al., 2012; Tong et al., 2008). I converted the genetic effect estimate (in SD) to original units (IU/L) using the standard deviation reported in the InCHIANTI study (5.1 IU/L, from **Table 3.1**).

3.3.2.1 *Expression Quantitative trait loci analysis*

Having identified a *cis-EPO* variant (rs1617640) associated with circulating EPO protein levels (*P* = 9.32 x 10^{-04}), I tested its *cis*-effects (+/- 500 kb) with gene expression in a meta-analysis of hepatic gene expression from 861 livers from European individuals in three datasets (Etheridge et al., 2020) and 236 kidneys from 134 individuals in one renal gene expression dataset (Damman et al., 2015). I selected the liver and kidney as EPO is highly expressed in both tissues (GTEx: https://gtexportal.org/).

3.3.2.2 *Hepatic eQTL data-sets*

The eQTL analysis within the liver included three human liver datasets of genotype and gene expression microarray data (see **Table 3.2**, for demographics, details of platforms used, and GEO accession numbers). The three data sets were then meta-analysed and the association between SNPs lying within a 1 mega base (mb) region of the *cis-EPO* SNP (500 kb either side) and *EPO* gene expression or nearby *TFR2* gene expression were extracted.

3.3.2.2.1 Data set 1: Innocenti et al. (2011)

Innocenti et al. (2011) (data set 1) profiled 205 normal (non-diseased) post-mortem liver samples from European organ donors. Analysis began with 240 normal (non-diseased) livers that were collected from unrelated donors of self-reported European and African descent (Innocenti et al., 2011). Gene expression levels were analysed using Agilent. Probe intensity was adjusted by subtracting background intensity using the minimum method, and quantile normalized between arrays. Dixon's outlier test was used to remove 13 arrays (out of a total of 517) based on total number of flagged probes, intra-array variance, inter-array variance, biological replicate variance, and spike-in linearity (Innocenti et al., 2011). Genotyping was performed using the Illumina quad-610 array (Illumina, San Diego, CA, USA). One sample was removed because it had a no call rate >10%. The initial marker set comprised 620,901 markers. 8,300 markers were removed because they showed significant deviation from Hardy-Weinberg equilibrium. 29,705 SNPs were removed from the analysis because they had a no call rate in more than >10% of the samples. The final marker set was comprised of 583,073 SNPs. Identity by descent analysis, performed in Plink, revealed 14 pairs of duplicated samples. Erroneous, redundant sample collection was later confirmed by the tissue bank. Genotype and expression data for these samples were merged for all downstream analyses. The sex of each sample was imputed by K-means clustering of Y-linked gene expression levels and X- and Y-linked genotypes; 3 samples had mismatched imputed and annotated sexes and were therefore removed. PCA was performed using the African and European populations from the Human Genome Diversity panel as reference populations. Four samples were flagged as outliers and thus removed. The first principal component separated African from non-African individuals. Only European samples (N = 205) were included in this analysis. A linear mixed model was run adjusting for age, sex, PCs, probe intensity, subtle variations in the hybridisation protocols, random effects of the probes, and random effects of each individual. To further control for the effects of outliers and population stratification, prior to eQTL mapping, the distribution of estimated individual effects, for each gene expression trait, was normal quantile transformed (Innocenti et al., 2011). Genotypes with minor allele frequencies less than 1% were excluded (Innocenti et al., 2011).

3.3.2.2.2 Data set 2: Schroder et al. (2013)

Schroder et al.(2013) (data set 2) profiled 149 samples from normal noncancerous liver tissue resected from patients with liver cancer. All tissue samples were examined by a pathologist, and only histologically noncancerous tissues were used for analysis (Schröder et al., 2013). Genotyping was performed using the Illumina HumanHap300 Genotyping BeadChip (Illumina, San Diego, CA, USA) with 318,237 SNPs. The raw data was pre-processed using Illumina BeadStudio version 3.0 (Illumina, San Diego, CA, USA). SNPs with a low call rate (< 95%), MAF < 4% and not in in HWE (false discovery rate ≤ 0.2) were excluded from further analysis. Identity-by-state distances were calculated to identify possible related individuals resulting in the exclusion of one individual and PCA was applied to detect further population substructures of which none were detected (Schröder et al., 2013).. Gene expression of > 48 000 mRNA transcripts was assessed using Illumina Human-WG6v2 Expression BeadChip (Illumina, San Diego, CA, USA). Low-level pre-processing, including background estimation and correction, normalisation, and probe-set summary was performed using Illumina BeadStudio version 3.0 (Illumina, San Diego, CA, USA). 9,875 genes with high detection *P*-value (> 0.1) or > 10% missing values were filtered out and removed from the data set. The remaining missing signal intensities were estimated using the ‘k nearest neighbour' algorithm and the resulting data set was subsequently log2 transformed. The raw data of 48,701 probe signal intensities were mapped and reduced to signal intensities corresponding to 15,439 unique genes. The finally processed data set was from 149 livers (71 males and 78 females) and consisted of 299,352 SNPs and 15,439 gene expression levels (Schröder et al., 2013). GenABEL (Aulchenko et al., 2007) was used to test all 4.6 billion possible combinations of SNPs and expression traits (299,352 SNPs x 15,439 traits) for significant associations. A genetic model was assumed in which both alleles contribute to gene expression in an additive manner.

3.3.2.2.3 Data set 3: Greenawalt et al. 2011

Greenawalt et al. (2011) profiled 960 liver samples (data set 3) collected at the time of Roux-en-Y gastric bypass surgery. Demographic information including age, race, gender, height, type of surgery, and year of surgery was collected for each patient. The cohort was predominantly (88%) self-reported “white” and female (75%). 59 % of surgeries were performed laparoscopically, and the rest were performed open. Gene expression was arrayed on a custom 44K DNA oligonucleotide microarray manufactured by Agilent Technologies as described previously (Hughes et al. 2001; Schadt et al. 2008). The custom array consists of 4,720 control probes and 39,820 non-control oligonucleotides. Samples with a 3′ bias >1 or <−1.5 from the mean of all samples were removed from the analysis, to prevent any bias from cRNA yield. Gene expression profiling results were successfully collected from 651 livers. A normalisation method based on control probes present on the microarrays was used to remove bias in expression profiles related to potential latent variables. The control probes were separated into two classes: specialty probes, such as spike-in probes or other probes designed to monitor the quality of the microarrays, and border probes used to describe the geometry of the microarray. PCA was performed and 14 liver PCs were identified which were added into a linear model to obtain residuals of the liver expression data for each probe. 950 DNA samples were genotyped on the Illumina 650Y BeadChip array. Sex was confirmed using PLINK (Purcell et al., 2007) and identity by state (IBS) analysis was performed to identify related individuals. 28 individuals were removed due to being identified as related leaving 922 samples for use in analysis. EIGENSTRAT was used to confirm reported race (Price et al., 2006).

3.3.2.2.4 *cis*-eQTL analysis

Genotypes were imputed to the 1000 Genomes Project Phase 1 reference panel with Minimac (http://genome.sph.umich.edu/wiki/Minimac) and expression probe sequences were mapped to ENSEMBL genes. To test for associations between genotype and gene expression, an additive (codominant) linear model was employed in the Matrix eQTL software package (http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/). A 1 mb window flanking the transcriptional start/stop sites was used to identify cis-eQTLs. For each data set, minor allele dosage, filtered to exclude variants with MAF < 0.01, was used to examine genotype association with rank inverse quantile normalised gene expression. Covariates included sex and age, the first 1 (data sets 1 & 2) or 5 (data set 3) principal components from genetic ancestry analysis, and 15–35 hidden factors identified using PEER (https://www.sanger.ac.uk/science/tools/PEER) (Stegle et al., 2010). Following identification of cis-eQTLs in each individual data set, cis-eQTLs identified in at least two data sets were included in the combined analysis. This resulted in the meta-analysis of 861 liver samples from individuals of European ancestry (**Table 3.2)** (additional methods and results have been reported in (Etheridge et al., 2020)). The T statistics from the additive linear model for each *cis*-eQTL within each data set were used to generate a meta-T-statistic using **Equation 3.2**. The meta-T-test-statistic was assumed to be normally distributed under the null due to the large sample size and was thus used as a measure of the effect size of eQTLs and also to calculate the associated *P*-values (Etheridge et al., 2020). Limitations include the fact that tissue samples for each data set were collected using different sample collection and storage protocols. Patient populations also differed in health status and exposure to clinical interventions prior to tissue collection. Attempts to control for this variability have been made by adjusting the expression analysis for hidden biological and technical variation that might affect gene expression. A fixed effect model was utilised which has been shown to increase power of detection in the presence of heterogeneity among data sets. This statistical approach allows for greater discovery of eQTLs. However, the cis-eQTLs with the strongest signals were those that were common across data sets, indicating the robustness of these signals to heterogeneity among the data sets. Furthermore, one data set was much larger than the other (data set 3) and in many cases this may drive the association detected due to the increased power and therefore other true associations may exist which were not statistically significant due to the lower sample size of the other two data sets.

$$tmeta\ = \frac{\sum witi}{\sqrt{\sum wi2}}, w = \sqrt{n - (\#covariates) - 1}$$

Equation 3.2: The modified meta test-statistic for the meta-analysis of renal gene expression. i = datasets 1-3 and n = sample size.

3.3.2.3 *Renal eQTL data-set*

The TransplantLines eQTL cohort used for the kidney eQTL analysis is part of a donor cohort for which gene expression results have been described previously (Damman et al., 2015). The dataset includes kidneys from living donors (*N* = 37), kidneys donated after brain death (*N* = 82) and kidneys donated after cardiac death (non-heart-beating) (*N* = 38). All organ donors and recipients were white. Time of biopsy (that is, before transplantation, before reperfusion and after reperfusion) was recorded as well. For some donors, multiple biopsies from different time points were taken and biopsies from both kidneys were available. Gene expression was arrayed using the Illumina HumanHT-12 v4 Expression BeadChips. The raw data files were processed using GenomeStudio Software and further analysed using GeneSpring GX 12.0 software (Agilent, Santa Clara, CA). Data normalisation was performed using the default GeneSpring GX 12.0 median shift normalization to the 75th percentile (applying a log2 transformation) and baseline transformation using the median of all samples. Samples were genotyped on the Illumina CytoSNP

12 v2 array and imputed using the 1000 Genomes Phase 1 ALL reference panel (Auton et al., 2015) using Impute2 (B. Howie et al., 2011; B. N. Howie et al., 2009). Expression and genotype data were available for 236 kidney biopsies from 134 donors. A mixed model eQTL analysis adjusting for sex, age, donor type, time of biopsy, first three PCs, and sample ID was run to account for multiple samples from a donor. Limitations with this data set include the small sample size and the fact that samples came from living and deceased donors which could result in differences in expression pattern.

3.3.2.4 *Colocalisation analysis*

I performed colocalisation analysis to assess the likelihood that the hepatic *EPO* eQTL was the same signal as the circulating protein level association. I obtained summary data for hepatic *cis*-eQTLs associated with *EPO* expression (FDR < 0.1) ± 500 kb of rs1617640 and extracted the summary statistics for these SNPs from our circulating EPO meta-analysis. I performed approximate Bayes Factor colocalisation analyses using the R coloc package (Giambartolomei et al., 2014; Wallace, 2020). I performed analysis using *P*-values and obtained overall estimates of the posterior probability that both our EPO meta-analysis and the liver eQTL share the same causal variant.

Table 3.2: Description of the three hepatic eQTL data sets and patient demographics. *Number of samples from each study utilised in the combined analysis following removal of individuals with non-European genetic ancestry, sex mismatches, and related samples within and between data sets*

Dataset	**Sample Size**	**Expression**	**Genotyping**	**PMID**
Data set 1	161	Agilent-014850 Whole Human Genome. 4x44K gene expression (NCBI GEO accession: GSE25935)	Illumina Human610-Quad v1.0 BeadChip (NCBI GEO accession: GSE26105)	21637794
Data set 2	145	Illumina Human Whole Genome-6 v2.0 (NCBI GEO accession: GSE32504)	Illumina HumanHap300-Duo v2.0 Genotyping (NCBI GEO accession: GSE39036)	22006096
Data set 3	555	Agilent Technologies (NCBI GEO accession: GSE9588)	HumanHap 650Y	21602305

Table 3.3: Association between rs1617640 and EPO and TFR2 expression in human liver. *Within three liver eQTL data sets, linear regression was carried out to model EPO and TFR2 expression levels with adjustment for relevant covariates. Results from the three liver datasets were combined by meta-analysis. EPO and TFR2 expression levels were determined using microarray and only included patients of European ancestry. The data was coded such that a positive beta (tmeta) means that as the number of minor alleles (C-alleles) increases, there is an increase in EPO or TFR2 expression. SD: standard error.*

Dataset	**Sample Size**	**Expression**	**Genotyping**	**PMID**	***EPO***			***TFR2***		
					Beta	**SE**	**P-value**	**Beta**	**SE**	**P-value**
Dataset1	161	Agilent-014850 Whole Human Genome. 4x44K gene expression (NCBI GEO accession: GSE25935)	Illumina Human610-Quad v1.0 BeadChip (NCBI GEO accession: GSE26105)	21637794	0.059	0.101	0.558	0.083	0.08	0.304
Dataset2	145	Illumina Human Whole Genome-6 v2.0 (NCBI GEO accession: GSE32504)	Illumina HumanHap300-Duo v2.0 Genotyping (NCBI GEO accession: GSE39036)	22006096	0.105	0.099	0.291	0.265	0.073	3.98 E-4
Dataset3	555	Agilent Technologies (NCBI GEO accession: GSE9588)	HumanHap 650Y	21602305	0.284	0.050	1.35 E-8	0.246	0.037	5.43 E-11
Meta-analysis	**861**				**T_{mets} = 5.39**	**0.149**	**6.86 E-8**	**T_{mets} = 7.38**	**0.149**	**1.56 E-13**

3.3.3 Drug-target two-sample Mendelian Randomisation

To investigate the causal association between higher circulating EPO levels and risk of CVD, I performed two-sample MR using the *cis-EPO* variant as the genetic instrument **(Figure 3.4).** I obtained genotype-exposure association statistics from the EPO meta-analysis (N = 6,127). Primary outcomes were CAD, MI or stroke using GWAS data consisting of 60,801, 42,561 and 40,585 cases respectively **(**Error! Reference source not found.**)** (Nikpay et al., 2015; Malik et al., 2018). I also performed internal GWAS using UK Biobank (UKB) on CAD, MI or stroke in 37,741, 105,90 and 9,092 cases respectively (Error! Reference source not found.**; see 3.3.5.1** for additional information on UKB). Where I had genotype-outcome association data from both UKB and publicly available GWAS, I performed an inverse variance weighted, fixed-effects meta-analysis using metan (Harris et al., 2008) to obtain an overall effect estimate for the genotype-outcome association. As only one genetic variant was used as an instrument, I calculated an overall causal estimate between the exposure and outcome using the Wald ratio **(Figure 3.4)** (Burgess, Small, et al., 2017). I also performed MR to investigate the effect of higher circulating EPO levels on levels of clinical markers (systolic blood pressure [SBP], diastolic blood pressure [DBP] and resting heart rate) predisposing to CVD risk factors using a meta-analysis of publicly available GWAS data and GWAS on UKB in 678,320, 677,567 and 514,695 individuals **(Figure 3.4,** Error! Reference source not found.**)** (Wain et al., 2017; Den Hoed at al., 2013).

3.3.4 Comparing clinical trial effects and genetic effects to estimate likely impact of therapeutically altered endogenous EPO levels on cardiovascular risk.

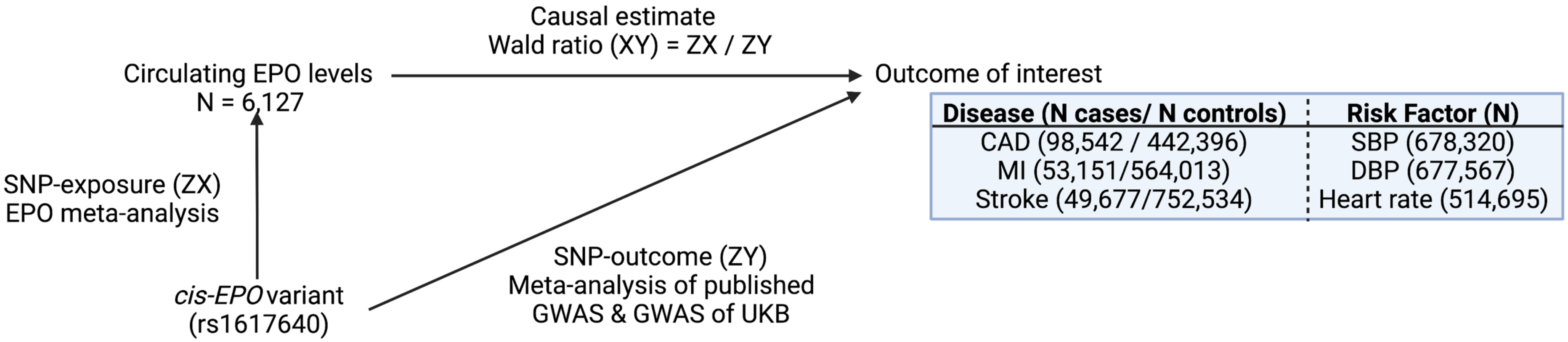
To scale the genetic effect estimates to a more representative, therapeutically relevant effect, I obtained the effects of a PHI in dialysis-dependent (DD) patients on endogenous EPO levels from a Phase II fixed dose randomised control trial (RCT) (Meadowcroft et al., 2019). The RCT provided an estimate of the effect of a fixed dose of PHI on EPO levels during the first four weeks (median “maximum” change in EPO levels from baseline at week 4 [27.1] / SD at baseline [61] = 0.44 SD) (Meadowcroft et al., 2019). I calculated the scaling factor by dividing the PHI-induced effect (0.44) by the effect of the *cis-EPO* SNP on circulating EPO levels (0.063). I used this value (7.05) to scale the genetically instrumented effect estimates and 95% confidence intervals of the *cis-EPO* SNP on CVD or the clinical markers of associated risk factors to the effect of the PHI on endogenous EPO levels.

3.3.5 Phenome-wide association study

To further investigate the therapeutic profile of modulated EPO levels, I tested the association of rs1617640 near the *EPO* gene with 869 traits in up to 451,099 UKB individuals of European ancestry. Genotype-phenotype associations were generated using BOLT-LMM (Loh et al., 2015) as previously described in Frayling et al. (2018). Traits were selected as described in Frayling et al. (2018). For continuous traits, I performed inverse normalisation prior to regression analysis to account for skewed distributions. I determined statistical significance using a genome-wide threshold of $P < 5 \times 10^{-08}$ as well as a Bonferroni corrected threshold of $P < 5.75 \times 10^{-05}$.

3.3.5.1 *UK Biobank (UKB) Cohort*

Briefly, UKB is a large-scale biomedical and research resource, containing genetic and health information from half a million UK participants. UKB recruited more than 500,000 individuals aged 37 - 73 years between 2006 and 2010 from across the UK (R. Collins, 2012). Participants provided a range of information via questionnaires and interviews (e.g. health status, lifestyle) and measurements (e.g. anthropometric, BP); this has been described in detail by Sudlow et al. (2015). SNP genotypes were generated from the Affymetrix Axiom UK Biobank array and the UK BiLEVE array and underwent extensive central quality control (http://biobank.ctsu.ox.ac.uk). My analysis was based on 451,099 individuals of European descent as defined by principal component analysis (PCA) (Frayling et al. 2018). 111 participants who withdrew from the study and 348 individuals whose self-reported sex did not match their genetic sex on the basis of relative intensities of X and Y chromosome SNP probe intensity were removed. Genotype-phenotype associations were generated using BOLT-LMM (Loh et al., 2015) which uses an LD score regression approach to account for structure caused by relatedness adjusting for SNP chip type, age, sex and test centre. For continuous traits, residuals were created by adjusting for age and sex and phenotypes were inverse normalised to account for skewed distributions. The UKB has approval from the North West Multicenter Research Ethics Committee (https://www.ukbiobank.ac.uk/ethics/), and these ethics regulations cover the work in this thesis. Written informed consent was obtained from all participants.



Disease (N cases/ N controls)	Risk Factor (N)
CAD (98,542 / 442,396)	SBP (678,320)
MI (53,151/564,013)	DBP (677,567)
Stroke (49,677/752,534)	Heart rate (514,695)

Figure 3.4: Schematic representation of two sample MR. *The cis-EPO genetic variant was used as an instrument to assess the causal association between higher circulating EPO levels and risk of CVD or levels of clinical markers for associated CVD risk factors.*

Table 3.4: cis-EPO SNP-outcome association statistics. *Summary statistics were obtained from previously published and publicly available GWAS and from GWAS on 451,099 UK Biobank individuals of European ancestry. For the categorical disease traits, effects are given as odds ratios. For the continuous traits, betas are measured in terms of the number of standard deviations. A fixed-effects inverse-variance weighted meta-analysis of the summary statistics from the GWAS on UK Biobank and previously published, publicly available GWAS was performed using metan to obtain an overall effect estimate of the*

effect of the genetic variant on the outcome of interest. All effects have been aligned to the EPO-increasing allele (A). CAD – Coronary artery disease, MI – myocardial infarction, SBP – systolic blood pressure, DBP – diastolic blood pressure, A1 freq – Allele 1 frequency.

Outcome	Study	A1	A2	A1 freq	OR	Lower 95%	Upper 95%	P-value	N (Cases/Controls)
CAD	UK Biobank	A	C	0.6	1.001	0.986	1.016	0.82	37741 / 318892
CAD	Nikpay et al. (2015)	A	C	0.59	1.005	0.985	1.025	0.643	60801 / 123504
	Meta-analysis				**1.002**	**0.99**	**1.01**	**0.72**	**98542 / 442396**
MI	UK Biobank	A	C	0.6	0.99	0.964	1.018	0.48	10590 / 440509
MI	Nikpay et al. (2015)	A	C	0.57	1.004	0.983	1.026	0.706	42561 / 123504
	Meta-analysis				**0.999**	**0.98**	**1.02**	**0.889**	**53151/564013**
Stroke	UK Biobank	A	C	0.6	0.963	0.935	0.993	0.014	9092 / 346423
Stroke	Malik et al. (2018)	A	C	0.6	1.01	0.993	1.031	0.344	40585 / 406111
	Meta-analysis				**0.995**	**0.98**	**1.01**	**0.553**	**49677/752534**
Continuous Outcomes									
Outcome	**Study**	**A1**	**A2**	**A1 freq**	**Beta (95% CI)**	**Lower 95%**	**Upper 95%**	**P-value**	**N**
SBP	UK Biobank	A	C	0.6	0.04	-0.05	0.12	0.8	450075
SBP	Wain et al. (2017)	A	C	0.59	0.024	-0.11	0.16	0.725	228245
	Meta-analysis				**0.03**	**-0.04**	**0.11**	**0.38**	**678320**
DBP	UK Biobank	A	C	0.6	-0.07	-1.20E-01	-0.02	1.00E-05	449322
DBP	Wain et al. (2017)	A	C	0.59	-0.03	-0.11	0.05	0.469	228245
	Meta-analysis				**-0.06**	**-0.1**	**-0.02**	**0.006**	**677567**
Heart rate	UK Biobank	A	C	0.6	-0.07	-0.13	-0.02	0.019	423846

Heart rate	den Hoed et al. (2013)	A	C	0.59	-0.021	-0.13	0.09	0.703	90849
	Meta-analysis				**-0.06**	**-0.11**	**-0.02**	**0.01**	**514695**

3.4 Results

3.4.1 Identification of three genomic loci associated with EPO levels at genome-wide significance.

With the aim of identifying human genetic variants associated with circulating EPO levels and using these variants as genetic proxies for therapeutically altered endogenous EPO levels, I performed a GWAS meta-analysis of circulating EPO. I used 6,127 individuals of European and African American descent. I identified three genomic loci containing three conditionally independent signals associated with circulating EPO ($P < 5 \times 10^{-8}$) (**Table 3.5, Figure 3.5**). The most strongly associated genetic variant rs4895441 (6q23, *HBS1L-MYB* locus) was previously identified as associated with EPO levels in a GWAS of 2,691 individuals (Grote Beverborg et al., 2015). This variant has a stronger primary effect on other erythrocyte phenotypes in previously published GWAS and a PheWAS on UKB European individuals **(Table 3.6).** The remaining two conditionally independent genomic loci (rs855791 and rs112631630) represent novel associations with circulating EPO levels. However, rs855791, located in the *TMPRSS6* locus, has primary, stronger effects on several other erythrocyte phenotypes compared to the effect on circulating EPO in a PheWAS on UKB European individuals **(Table 3.6)** and has been previously identified as associated with other erythrocyte phenotypes and iron homeostasis biomarkers in published GWAS studies (Benyamin et al., 2014; M.-H. Chen et al., 2020). The variant (rs112631630), located in the *NRAP* locus, is only present in African Americans and I was unable to test in additional datasets. Therefore, these variants were not deemed specific instruments for use in subsequent MR analysis to genetically proxy therapeutic modulation of endogenous EPO levels.

Table 3.5:Summary statistics for the lead genetic variants ($P < 5 \times 10^{-08}$) identified in the meta-analysis of circulating EPO in 6,127 individuals of European and African descent.

GWAS was performed using GEMMA in four independent cohorts (PREVEND, HealthABC, InCHIANTI, BLSA). Summary association statistics were combined in a meta-analysis using METAL. Conditional analysis was performed using GCTA-COJO to identify conditionally independent genetic variants associated with circulating EPO levels (at $P < 5 \times 10^{-08}$). EAF: effect allele frequency. SE: standard error.

Lead SNP	Chromosome	Position	Genomic Locus	Effect allele	EAF	Effect size	SE	P-value	Sample size
rs4895441	6	135426573	HBS1L	A	0.74	-0.24	0.021	1.45E-30	6127
rs112631630	10	115407228	NRAP	A	0.996	3.598	0.65	3.09E-08	536
rs855791	22	37462936	TMPRSS6	A	0.42	0.113	0.019	2.47E-09	6529

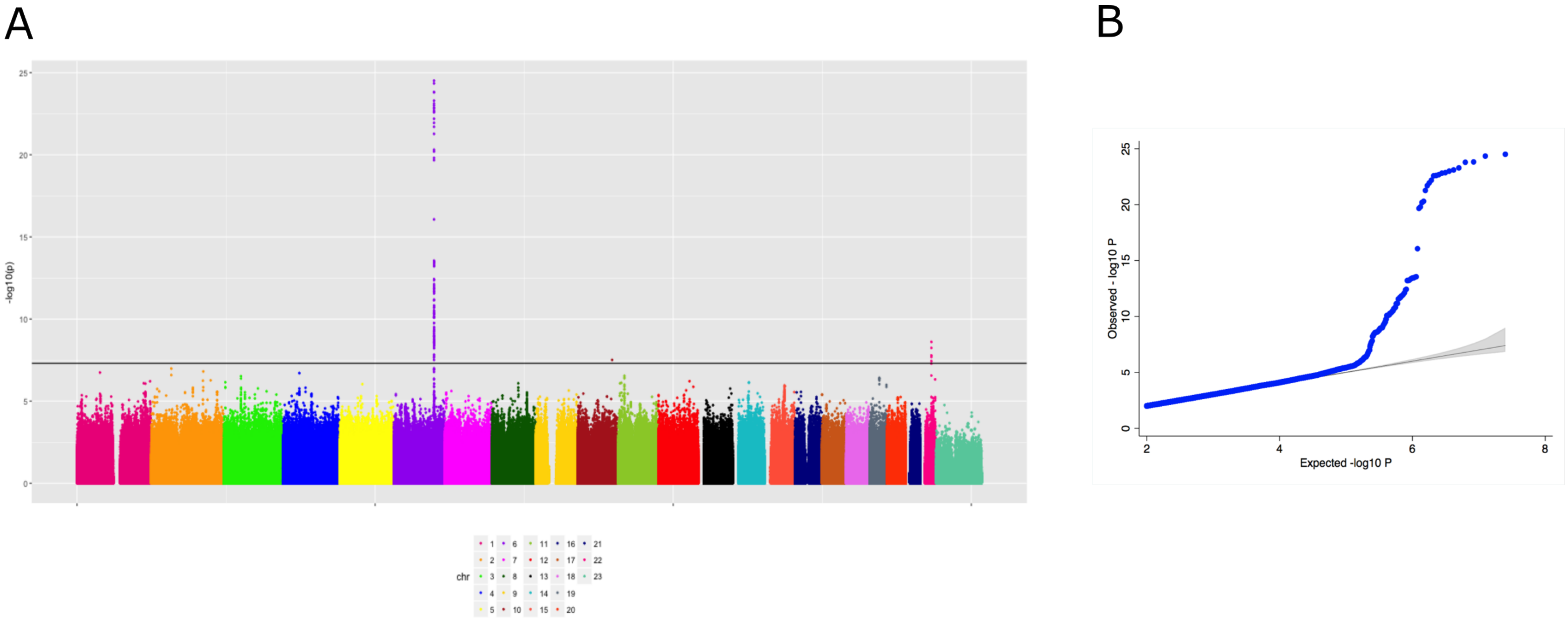


Figure 3.5: Meta-analysis of circulating EPO levels in 6,127 individuals of European and African American descent. ***A:*** *Manhattan plot showing the –$\log_{10}$(P-value) of the association between circulating EPO levels and 25.1 million single-nucleotide polymorphisms (SNPs) in 6,127 individuals following correction for sex, age and study-specific covariates. The black line shows the indicative suggestive genome-wide significance threshold of P < 5 x 10^{-8}. Each individual dot represents a SNP. Genetic markers are ranked by chromosome and positions.* ***B:*** *QQ plot for the EPO meta-analysis showing the observed versus expected p-values. The black diagonal line represents the expected distribution. Points to the left of the diagonal represent associations that are more significant than expected.*

Table 3.6: Association between rs4895441 and rs855791, two of the three lead genetic variants identified in the EPO meta-analysis, with other phenotypes in 451,099 UK Biobank unrelated, European individuals. *Associations reaching genome-wide significance (P < 5 x 10^{-08}) are shown in the table below. The EPO-increasing alleles of rs4895441 and rs855791 have a primary, stronger association with several other blood cell phenotypes compared to its effect on EPO levels. Betas have been aligned to the EPO-increasing alleles.*

Genetic variant	Phenotype	Gender	Beta	SE	P-value
rs4895441	Erythropoietin	Combined	0.218	0.021	1.45E-30
	Corpuscular haemoglobin	Combined	0.179	0.002	4.3E-1635
	Corpuscular Volume	Combined	0.157	0.002	6.2E-1353
	Mean corpuscular volume	Combined	0.157	0.002	6.7E-1270
	Red blood cell count	Combined	-0.139	0.002	2.3E-1224
	Mean corpuscular volume (anaemics excluded)	Combined	0.163	0.002	2.7E-1148
	Platelet crit	Combined	0.12	0.002	5.4E-790
	Platelet Count	Combined	0.107	0.002	1.8E-635
	Fibrosis-4 Score	Combined	-0.079	0.002	1.9E-399
	Red blood cell distribution width	Combined	-0.092	0.002	1.3E-386
	Red blood cell distribution width (anaemics excluded)	Combined	-0.098	0.002	4.5E-378
	Non-alcoholic fatty liver disease fibrosis score	Combined	-0.085	0.002	2.6E-367
	Corpuscular haemoglobin concentration	Combined	0.075	0.002	4.70E-261
	Sphered cell volume	Combined	0.07	0.002	7.60E-257
	Haematocrit percentage	Combined	-0.062	0.002	5.30E-255
	Non-alcoholic fatty liver disease fibrosis score	Female	-0.093	0.003	4.00E-220
	fibrosis-4 Score	Female	-0.082	0.003	1.30E-214
	Reticulocyte Volume	Combined	0.065	0.002	3.60E-209
	fibrosis-4 Score	Male	-0.078	0.003	1.30E-171

	Non-alcoholic fatty liver disease fibrosis score	Male	-0.077	0.003	3.00E-141
	Haemoglobin Concentration	Combined	-0.038	0.002	1.10E-103
	Reticulocyte Percentage	Combined	0.046	0.002	1.50E-101
	Eosinophil Count	Combined	-0.042	0.002	2.40E-83
	High Light Scatter reticulocyte percentage	Combined	0.041	0.002	8.40E-81
	Glycated haemoglobin	Combined	-0.036	0.002	3.60E-66
	Lymphocyte Count	Combined	-0.034	0.002	1.00E-52
	Monocyte Count	Combined	-0.031	0.002	9.70E-51
	Neutrophil Count	Combined	-0.03	0.002	1.40E-42
	Eosinophil Percentage	Combined	-0.029	0.002	2.10E-39
	Glycated haemoglobin	Female	-0.035	0.003	2.10E-35
	Glycated haemoglobin	Male	-0.037	0.003	2.50E-31
	Cholesterol corrected for statin use	Combined	-0.024	0.002	7.10E-30
	Low density lipoprotein corrected for statin use	Combined	-0.024	0.002	7.60E-30
	Cholesterol corrected for statin use	Male	-0.028	0.003	1.30E-19
	Platelet distribution width	Combined	0.018	0.002	2.60E-19
	Aspartate Aminotransferase	Combined	-0.019	0.002	3.40E-19
	Low density lipoprotein corrected for statin use	Male	-0.027	0.003	4.30E-18
	Albumin	Combined	0.02	0.002	4.00E-17
	Cholesterol	Combined	-0.017	0.002	9.10E-15
	High Light Scatter Reticulocyte Count	Combined	0.016	0.002	2.20E-14
	Low density lipoprotein corrected for statin use	Female	-0.021	0.003	2.70E-14
	Cholesterol corrected for statin use	Female	-0.021	0.003	6.40E-14
	Low density lipoprotein	Combined	-0.015	0.002	2.40E-12
	Aspartate Aminotransferase	Female	-0.021	0.003	7.70E-12

	Immature reticulocyte	Combined	0.014	0.002	6.40E-11
	Microalbumin	Combined	0.013	0.002	1.30E-10
	Reticulocyte Count	Combined	0.013	0.002	1.60E-10
	Albumin	Male	0.022	0.004	4.60E-10
	Albumin	Female	0.019	0.003	2.20E-09
	Apolipoprotein B	Combined	-0.013	0.002	2.40E-09
	Microalbumin	Female	0.019	0.003	3.25E-03
	Cholesterol levels	Male	-0.019	0.003	3.47E-03
	Aspartate Aminotransferase	Male	-0.019	0.003	3.49E-03
	High Density Lipoprotein	Male	-0.018	0.003	3.45E-03
rs855791	EPO	Combined	0.113	0.019	2.47E-09
	Corpuscular Haemoglobin	Combined	-0.151	0.002	3.6E-1399
	Corpuscular Volume	Combined	-0.129	0.002	3.0E-1102
	Mean corpuscular volume	Combined	-0.13	0.002	9.0E-1038
	Mean corpuscular volume (anaemics excluded)	Combined	-0.133	0.002	2.5E-932
	Red blood cell distribution width (anaemics excluded)	Combined	0.113	0.002	1.4E-480
	Haemoglobin concentration	Combined	-0.073	0.002	3.5E-446
	Red blood cell distribution width	Combined	0.113	0.002	1.0E-416
	Corpuscular Haemoglobin concentration	Combined	-0.067	0.002	9.20E-251
	Glycated haemoglobin levels	Combined	0.064	0.002	2.00E-244
	Haematocrit Percentage	Combined	-0.053	0.002	2.10E-224
	Sphered cell volume	Combined	-0.055	0.002	1.10E-193
	Glycated haemoglobin levels	Female	0.062	0.003	8.10E-129
	Glycated haemoglobin levels	Male	0.065	0.003	7.10E-111
	Reticulocyte Percentage	Combined	-0.04	0.002	6.60E-90

	Reticulocyte Count	Combined	-0.035	0.002	4.70E-72
	Total Bilirubin	Combined	-0.029	0.002	9.10E-67
	Platelet Count	Combined	0.027	0.002	5.00E-52
	Platelet crit	Combined	0.027	0.002	7.30E-52
	Fibrosis-4 score	Combined	-0.022	0.002	1.60E-38
	High light scatter reticulocyte percentage	Combined	-0.025	0.002	5.60E-37
	Total Bilirubin	Female	-0.031	0.003	1.40E-35
	Total Bilirubin	Male	-0.032	0.003	8.80E-34
	High light scatter reticulocyte count	Combined	-0.022	0.002	1.80E-28
	Non-alcoholic fatty liver disease fibrosis score	Combined	-0.021	0.002	3.40E-28
	Platelet distribution width	Combined	-0.019	0.002	8.90E-24
	Direct Bilirubin	Combined	-0.019	0.002	4.80E-23
	Red blood cell count	Combined	0.016	0.002	2.20E-22
	Fibrosis-4 score	Female	-0.024	0.002	4.20E-22
	Non-alcoholic fatty liver disease fibrosis score	Female	-0.025	0.003	1.40E-19
	Direct Bilirubin	Male	-0.023	0.003	2.20E-16
	Fibrosis-4 score	Male	-0.020	0.003	2.80E-15
	Immature reticulocytes	Combined	0.015	0.002	5.20E-13
	Non-alcoholic fatty liver disease fibrosis score	Male	-0.018	0.003	1.10E-10
	Phosphate levels	Combined	0.012	0.002	1.20E-08
	Direct Bilirubin	Female	-0.014	0.003	4.00E-08

3.4.2 Identification of *cis*-SNP lying in the promoter region of *EPO* for use to genetically proxy therapeutic alteration of endogenous EPO levels.

As the variants identified by GWAS were not deemed sufficiently specific instruments to act as genetic proxy tests for the long-term effects of therapeutic rises in endogenous EPO levels, I looked more closely at the *EPO* locus where associations have been previously reported (Amanzada et al., 2014; Tong et al., 2008). I identified rs1617640, a *cis*-SNP which has been previously shown to have a functional role at controlling EPO levels, lying in the promoter region of the *EPO* gene, 1,125bp upstream of the transcription start site (Tong et al., 2008). Based on our meta-analysis, each copy of the A-allele of rs1617640 was associated with 0.063 SD (equivalent to 0.32 IU/L) higher endogenous EPO levels ($P = 9.32 \times 10^{-04}$) (**Table 3.7),** which is consistent with the direction of effect previously reported in patients with diabetic retinopathy or hepatitis C (Amanzada et al., 2014; Tong et al., 2008). The EPO-increasing A allele has been previously reported to create a binding-site for the AP1 or EV11/MEL1 transcription factors highlighting the functional role rs1617640 may have at regulating *EPO* expression levels (Tong et al., 2008).

3.4.3 The *cis-EPO* SNP is a hepatic eQTL for *EPO* expression and nearby *TFR2* expression

To provide additional insight into the rs1617640-EPO association and further evaluate its utility as an instrument, I tested the association of the *cis-EPO* SNP with gene expression in the kidney and the liver as *EPO* is highly expressed in these two tissues (Franklin, 2013). In the liver, I found that the C-allele at rs1617640 was associated with higher *EPO* gene expression (ß = 0.22 [0.14, 0.3], $P = 6.86 \times 10^{-08}$) and also *TFR2* expression (ß = 0.23 [0.17, 0.29], $P = 1.56 \times 10^{-13}$), a gene which lies upstream of the *EPO* gene and is involved in iron metabolism (Benyamin et al., 2014; Forejtnikovà et al., 2010) (**Table 3.8).** No effect of rs1617640 on renal expression of *EPO* (ß = 0.16 [-2.46, 2.78], $P > 0.05$) or *TFR2* (ß = 1.33 [-0.59, 3.26], $P > 0.05$) was found (**Table 3.8**). I therefore preceded with hepatic results only.

Table 3.7: Summary statistics for association between the cis-EPO genetic variant (rs1617640) and circulating EPO levels obtained from a meta-analysis in 6,127 individuals of European and African descent.

Analysis	RSID	Chromosome	Position	A1	A2	Freq A1	Beta	SE	P-value	Sample size
EPO meta-analysis	rs1617640	7	100317298	A	C	0.62	0.063	0.02	9.32E-04	6127

Table 3.8: Association between rs1617640 and EPO and TFR2 expression in human kidneys and liver.

Analysis of the association between rs1617640 and EPO or nearby TFR2 expression was performed in the kidneys and the liver. No association was found in the kidneys (P > 0.05), but there was evidence for an association between rs1617640 and EPO or TFR2 gene expression (P < 0.05). The data was coded such that a positive beta (tmeta) means that as the number of minor alleles (C-alleles) increases, there is an increase in EPO or TFR2 expression.

Dataset	Sample size	Tissue	EPO				TFR2			
			T-statistic	Beta	SE	P-value	T-statistic	Beta	SE	P-value
Liver meta-analysis	861	Liver	5.39	0.218	0.1488	$6.86\text{x}10^{-08}$	7.38	0.226	0.1488	$1.56\text{x}10^{-13}$
Kidney TransplantLines	286	Kidney	0.117	0.157	1.338	0.907	1.358	1.333	0.981	0.177

3.4.4 Colocalisation analysis confirms the *cis-EPO* variant is shared between the EPO meta-analysis and the hepatic eQTL data

Colocalisation analysis provided evidence that the variant associated with circulating EPO levels in the meta-analysis and hepatic *EPO* mRNA expression has a 71% posterior probability of being the same causal variant **(*Figure 3.6*)**

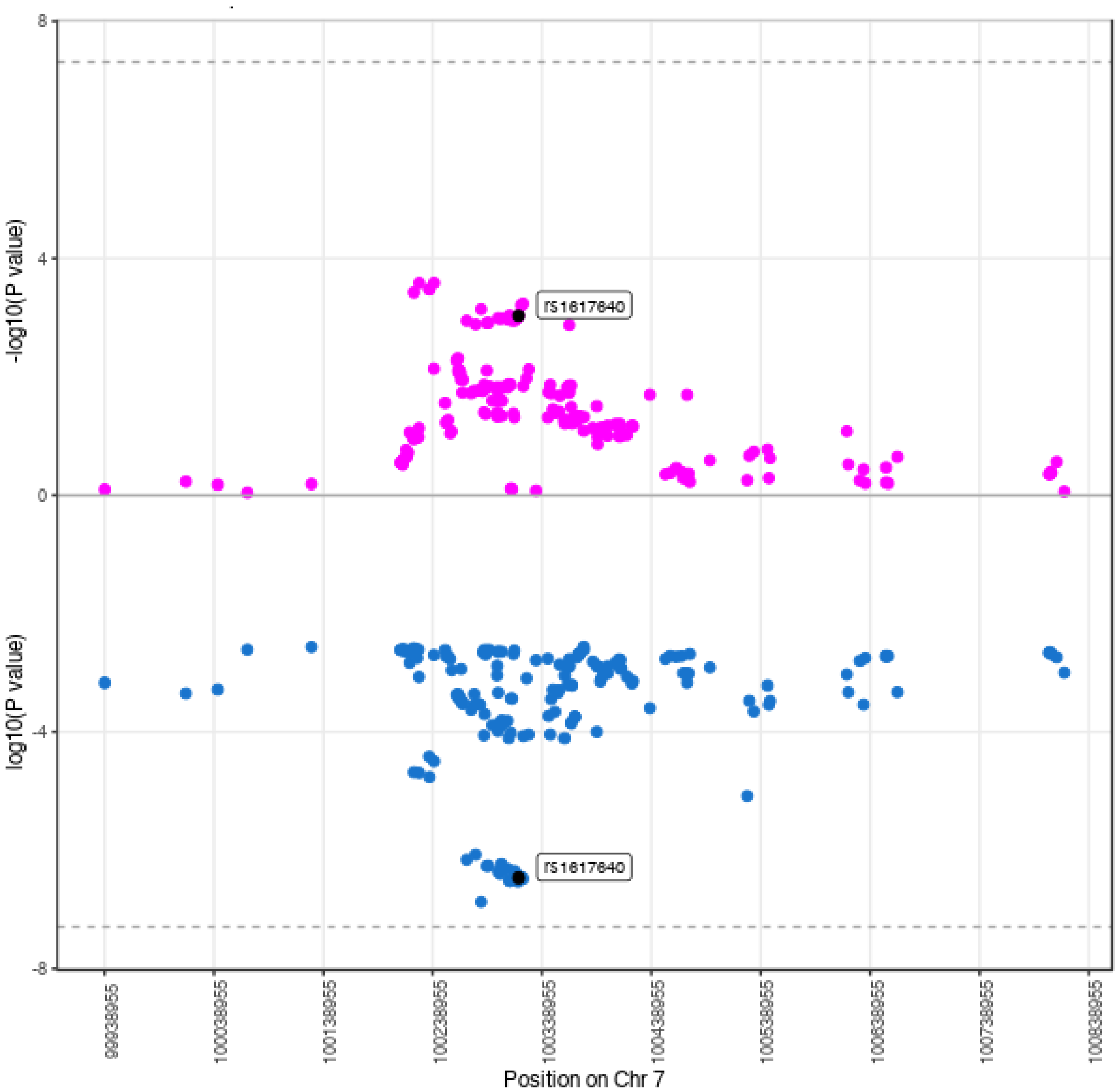


Figure 3.6: Colocalisation analysis of circulating endogenous EPO levels and liver EPO gene expression to identify whether the same causal variants are shared. *I tested whether the same genetic variant is associated with both circulating EPO levels and liver EPO gene expression levels within a 500kb flanking window of the cis-EPO genetic variant (rs1617640). I found evidence to support the same causal variant being shared by the meta-analysis of circulating endogenous EPO levels and liver EPO gene expression levels (Posterior probably = 71%). The top half of the Miami plot (pink dots) represents -log_{10}(P-values) for the circulating EPO levels obtained from the EPO meta-analysis whilst the bottom half of the Miami plot (blue dots) represents log_{10}(P-values) for EPO gene expression in the liver obtained from eQTL analysis. The dashed lines represent genome-wide significance levels of P = 5 x 10^{-08}.*

3.4.5 Genetically proxied long-term higher endogenous EPO levels are not associated with increased cardiovascular risk

I used the *cis-EPO* SNP as an instrument in two-sample MR as a genetic proxy for the effects of therapeutically altered EPO levels on risk of CVD or levels of clinical markers predisposing to CVD risk factors. I found no evidence of a causal association between 1 SD (equivalent to 5.1 IU/L) higher endogenous EPO levels and increased risk of CAD (OR [95% CI] =1.03 [0.85, 1.25], *P* = 0.72), stroke (OR [95% CI] = 0.92 [0.70, 1.21], *P* = 0.55) or MI (OR [95% CI] = 0.98 [0.75, 1.29], *P* = 0.89) **(Figure 3.7*Figure 3.7*, Table 3.9**). I found evidence of a causal association between 5.1 IU/L higher endogenous EPO levels and lower resting heart rate (ß [95% CI] = -0.996 [-1.74, -0.25], *P* = 0.01) and lower DBP (ß [95% CI] = -0.98 [-1.67, -0.29], *P* = 0.006) but not with SBP (ß [95% CI] = 0.53 [-0.65, 1.71], *P* = 0.38) (**Figure 3.7, Table 3.9**).

Table 3.9: Causal estimates of the association between higher circulating EPO levels and risk of CVD or levels of clinical markers for CVD associated risk factors. *Causal estimates were calculated using the Wald ratio due to the single genetic variant being used as the instrument. Two-sample MR implemented in the MRBase package.*

Disease Outcomes							
Exposure	**Outcome**	**Odds ratio**	**Lower 95%**	**Upper 95%**	**P-value**	**N Cases**	**N Controls**
EPO	CAD	1.03	0.85	1.25	0.72	98,542	442,396
EPO	Stroke	0.92	0.70	1.21	0.55	49,677	387,008
EPO	MI	0.98	0.75	1.29	0.89	53,151	564,013
Associated risk factors							
	Risk Factors	**Effect estimate**	**Lower 95%**	**Upper 95%**	**P-value**	**Total Sample size**	
EPO	SBP	0.53	-0.65	1.71	0.38	586,080	
EPO	DBP	-0.98	-1.67	-0.29	0.0057	585,325	
EPO	Heart rate	-0.996	-1.74	-0.25	0.0097	514,706	

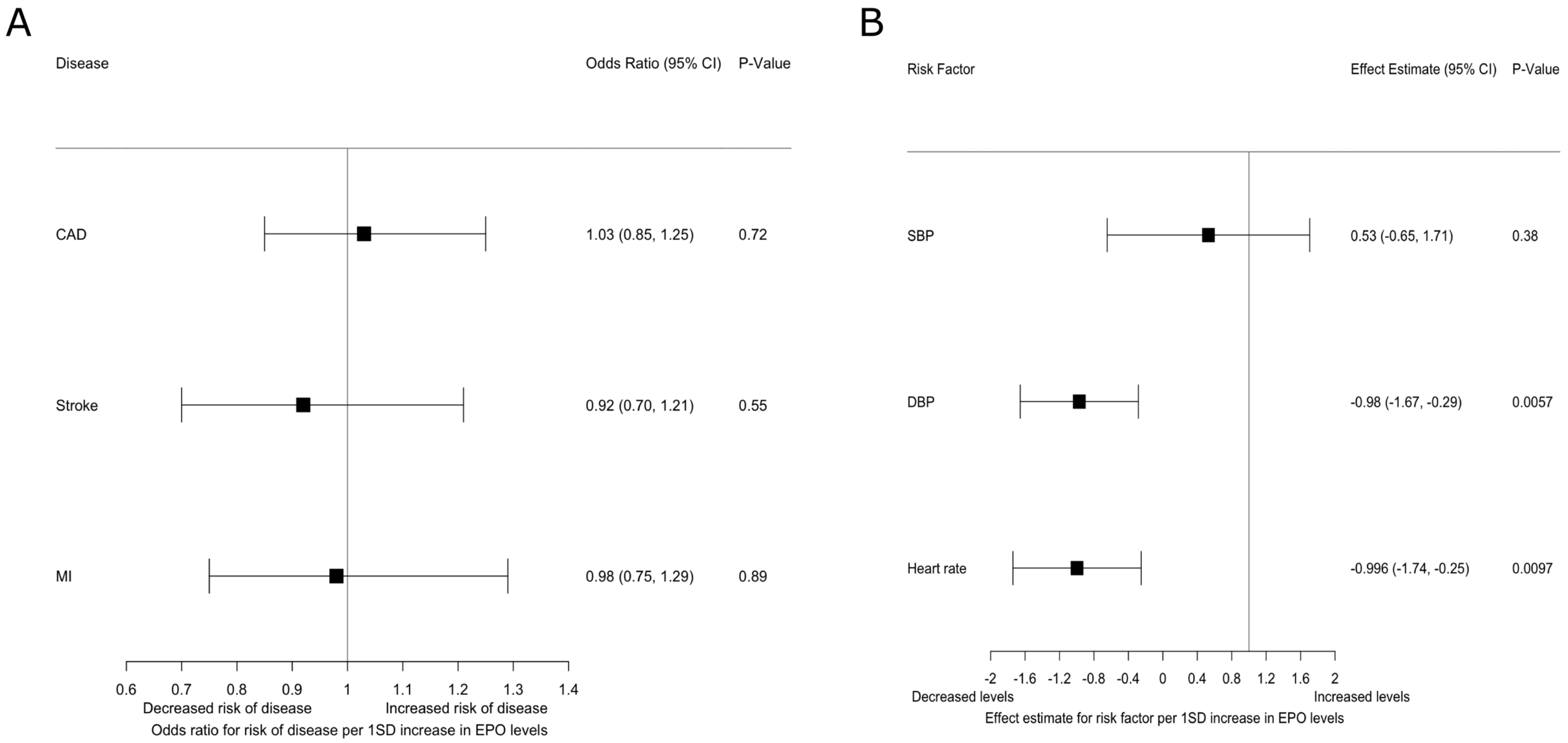


Figure 3.7: Genetically proxied higher endogenous EPO levels is not associated with an increased risk for CVD or increased levels of clinical markers for CVD risk factors. ***A:*** *Two-sample MR analysis using the cis-EPO variant as a genetic instrument showed no evidence of an association between a 1 SD (equivalent to a 5.1 IU/L) higher EPO level and increased risk of CAD, stroke or MI.* ***B:*** *Two-sample MR analysis using the cis-EPO variant as a genetic instrument showed evidence of a causal association between a 1 SD (equivalent to 5.1 IU/L) higher EPO level and decreased DBP and heart rate but no association with SBP.*

3.4.6 Rescaling genetic effects to clinically relevant effects of therapeutic rises in endogenous EPO levels shows no increase in risk of CVD.

In a RCT, patients receiving a PHI (daprodustat) had circulating EPO levels 0.44 SD (27.1 / 61), equivalent to 2.2 IU/L, higher than baseline EPO levels (Meadowcroft et al., 2019). Given that the per allele effect of rs1617640 on endogenous EPO was 0.063 SD, I rescaled the genetic association by multiplying by 7.05 (0.44 / 0.063) to obtain a clinically relevant estimate of the likely impact of therapeutically altered rises in genetically proxied endogenous EPO on cardiovascular risk. Using this scaling factor allowed quantification of the upper and lower bounds of the predicted genetically proxied effects of long-term endogenous EPO-rises on CVD (**Table 3.10, Figure 3.8**). By rescaling the genetic associations to the PHI-induced effect, genetically proxied therapeutic rises in endogenous EPO levels were not associated (at *P* < 0.05) with increased odds of CAD (OR [95% CI] = 1.01 [0.93, 1.07]), MI (OR [95% CI] = 0.99 [0.87, 1.15]) or stroke (OR [95% CI] = 0.97 [0.87, 1.07]) (**Table 3.10, Figure 3.8)**. I could exclude a 1.07, 1.15 and 1.07 increased odds of CAD, MI or stroke respectively (**Table 3.10, Figure 3.8)**. For the clinical markers predisposing to CVD, I did not observe an association between higher genetically proxied therapeutic rises in endogenous EPO levels and SBP (ß [95% CI] = 0.21 [-0.28, 0.78]), DBP (ß [95% CI] = -0.42 [-0.71, -0.14]) or resting heart rate (ß [95% CI] = -0.42 [-0.78, -0.14]) (**Table 3.10, Figure 3.8). **I could exclude 0.78 mmHg increased SBP levels and could exclude any increase in DBP or resting heart rate (**Table 3.10, Figure 3.8).**

To compare the effects of PHIs to that of the current rhEPO treatments, I also rescaled the genetic effects to that of rhEPO by multiplying the genetic associations by 822 (0.063 / 51.8) due to patients receiving rhEPO having EPO levels 51.8 SD higher than controls (Meadowcroft et al., 2019). However, due to the effect of rhEPO being so large and thus the upper confidence intervals being so big, I was unable to meaningfully exclude a damaging effect (**Table 3.10)**.

*__Table 3.10: Rescaling the causal estimates to the PHI-induced effect.__ The causal estimate of the effect of higher endogenous EPO levels on risk of CVD using the cis-EPO SNP as an instrument were rescaled to the PHI-induced or rhEPO-induced effect on circulating EPO levels from a recent RCT (Meadowcroft et al. 2019). This provided a clinically relevant estimate on the physiological scale of the likely impact of novel or current treatments for anaemia in CKD on CVD*__.__

Treatment	**Disease**				
PHI-induced effect	**Exposure**	**Outcome**	**Odds ratio**	**Lower 95%**	**Upper 95%**
	EPO	CAD	1.014	0.932	1.073
	EPO	MI	0.993	0.867	1.150
	EPO	Stroke	0.965	0.867	1.073
	Risk Factors				
	Exposure	**Outcome**	**Effect estimate**	**Lower 95%**	**Upper 95%**
	EPO	SBP	0.212	-0.282	0.78
	EPO	DBP	-0.423	-0.705	-0.14
	EPO	Heart rate	-0.423	-0.776	-0.14
rhEPO-induced effect	**Disease**				
	Exposure	**Outcome**	**Odds ratio**	**Lower 95%**	**Upper 95%**
	EPO	CAD	1.04	0.81	3573.79
	EPO	MI	0.44	$6.11x10^{-08}$	11782904
	EPO	Stroke	0.02	$6.11x10^{-08}$	3573.79
	Risk Factors				
	Exposure	**Outcome**	**Effect estimate**	**Lower 95%**	**Upper 95%**
	EPO	SBP	24.67	-32.89	90.44
	EPO	DBP	-49.33	-82.22	-16.44
	EPO	Heart rate	-49.33	-90.44	-16.44

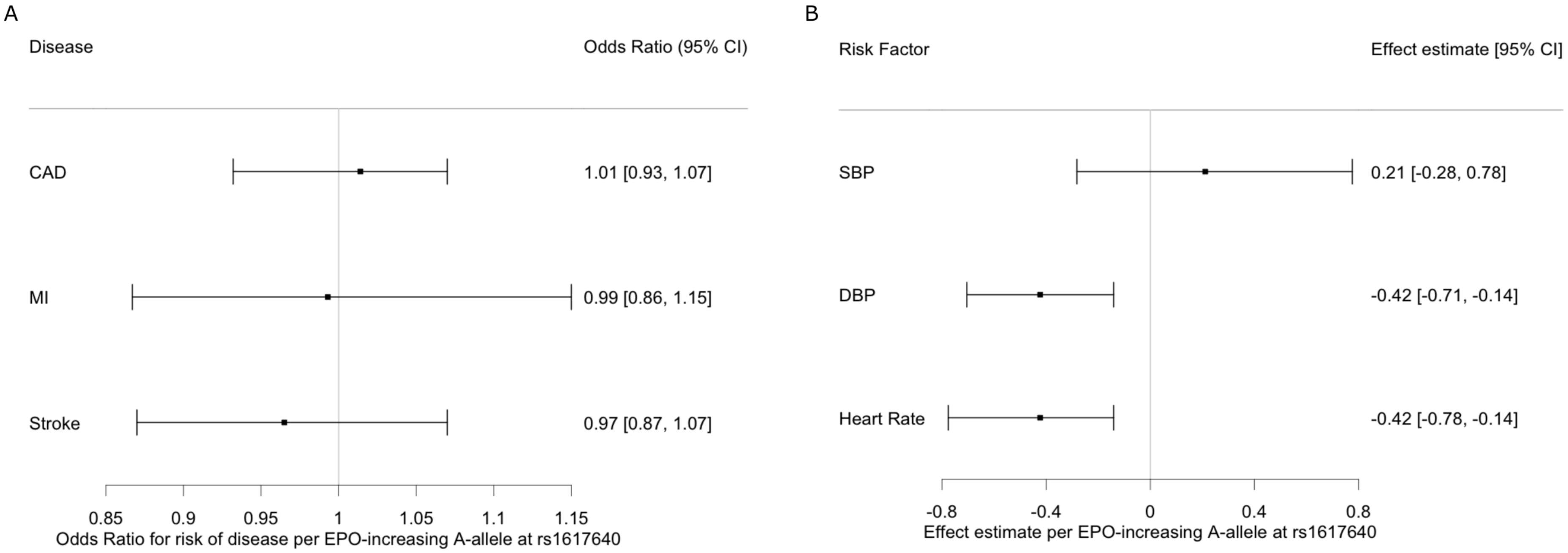


Figure 3.8: Genetically proxied therapeutic rises in endogenous EPO levels are not associated with an increased risk of CVD or clinical CVD risk factors. *The genetic estimates obtained using two-sample MR were rescaled to the PHI-induced effect observed in a RCT.* ***A:*** *Based on the upper confidence intervals, I was able to exclude an increased 1.07, 1.15 and 1.07 odds of CAD, MI or stroke respectively.* ***B:*** *Based on the upper limit, I could exclude levels higher than 0.78 mmHg for SBP and no increase for DBP or resting heart rate.*

3.4.7 The *cis-EPO* SNP is associated with several relevant erythrocyte phenotypes with similar effect sizes to that on circulating EPO and no unintended effects or diseases

To further determine the specificity of the *cis-EPO* SNP as a genetic instrument for endogenous, physiological EPO levels and identify any additional, unintended effects that may be associated with long-term rises in endogenous EPO levels, I tested the association between the *cis-EPO* SNP and 869 traits in up to 451,099 unrelated European UKB individuals. I found that the *cis-EPO* SNP was associated with 18 relevant erythrocyte traits ($P < 5 \times 10^{-08}$) with similar effect (0.01 - 0.06 SD) to that of the A-allele on circulating EPO levels (0.063 SD) (**Figure 3.9*, *Table 3.11**). I also found evidence for an association between the EPO-increasing A-allele of rs1617640 and decreased fibrosis-4 score (β = -0.01, $P = 4.7 \times 10^{-17}$) and non-alcoholic fatty acid liver disease (NAFLD) fibrosis score (β = -0.02, $P = 2.20 \times 10^{-25}$) **(Figure 3.9, Table 3.11).** However, these associations were not clinically significant (equivalent to a 0.06 and 0.07 change in fibrosis-4 or NAFLD for 1 IU/L increase in EPO levels). These associations are likely driven by the strong association with higher platelet counts (β = 0.02, $P = 4.7\text{x}10^{-39}$) **(Table 3.11).** I did not find evidence for an association between genetically proxied higher endogenous EPO levels and any other unintended effects or diseases.

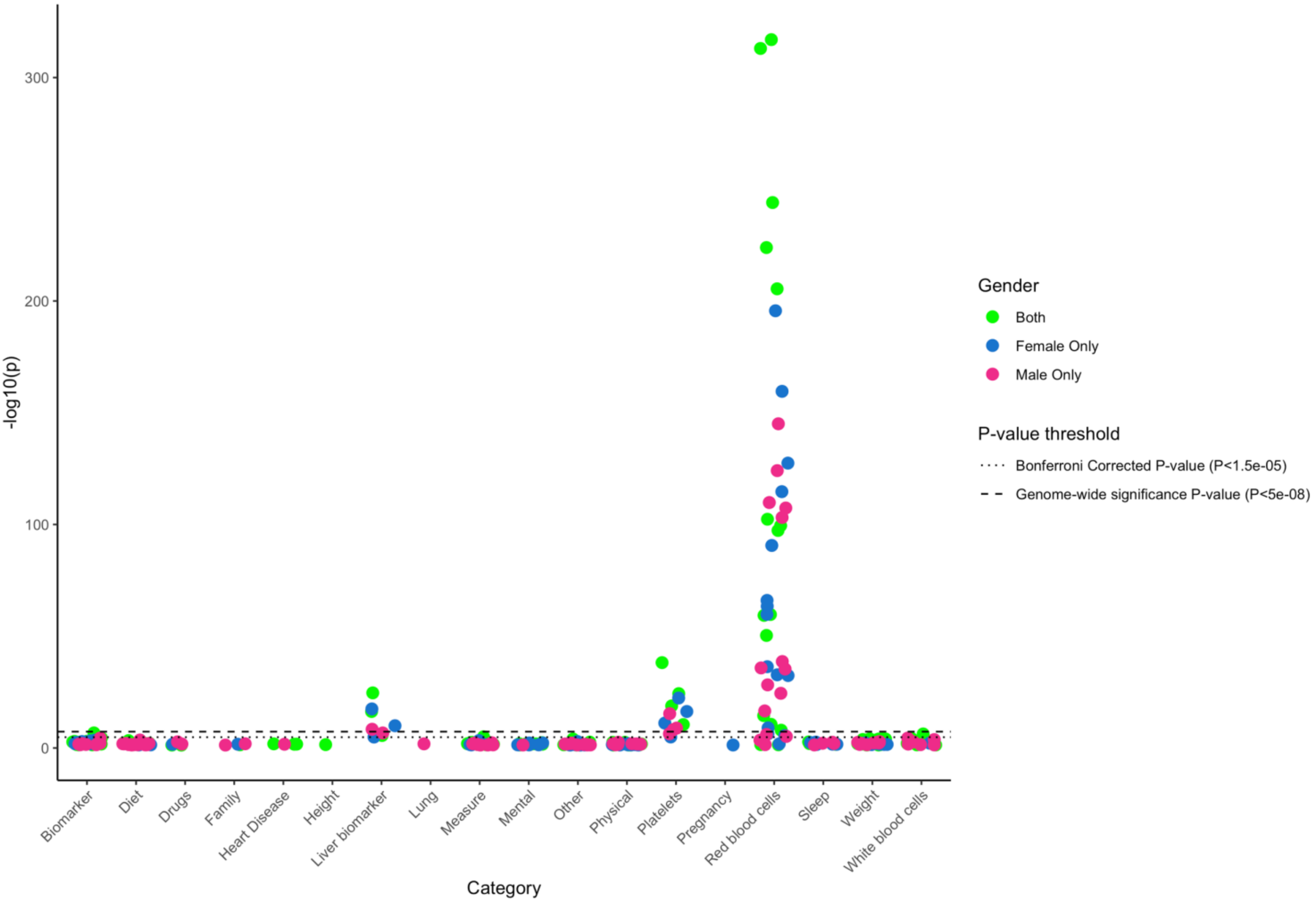


Figure 3.9: PheWAS of the cis-EPO SNP with 869 traits in up to 451,099 individuals from UK Biobank. *The cis-EPO SNP was most strongly associated with relevant erythrocyte phenotypes and liver biomarkers indicating that the cis-EPO SNP is a valid proxy for endogenous EPO levels. Plot represents the -log*$_{10}$*(P-values) (y-axis) for all traits passing a P-value threshold of 0.05. Analysis was performed in both males and females combined (green dots), females only (blue dots) and males only (pink dots). The dotted line highlights associations passing a Bonferroni corrected P < 1.5 x 10*$^{-05}$ *and the dashed line highlights associations passing genome-wide significance P-value threshold < 5 x 10*$^{-08}$*. Traits have been clumped together into categories which are represented on the x-axis.*

Table 3.11: Association between rs1617640 and traits passing genome-wide significance identified through PheWAS. *PheWAS was performed to investigate the effect of rs1617640 with 869 traits in up to 451,099 European, unrelated UKB individuals. The majority of traits passing genome-wide significance ($P < 5 \times 10^{-08}$) are erythrocyte phenotypes with a similar effect size to the effect of rs1617640 on EPO levels (beta=0.06). The EPO-increasing A allele is associated with increased corpuscular volume, sphered cell volume, reticulocyte volume and percentage but decreased haemoglobin concentration, haematocrit percentage and red blood cell distribution width. All reported effect sizes are aligned to the EPO-increasing A allele of rs1617640.*

Phenotype	**Beta**	**SE**	**P-value**
Red blood cell count	-0.0643	0.0017	2.3E-315
Corpuscular haemoglobin	0.0721	0.002	3.3E-313
Corpuscular volume	0.0615	0.0019	8.60E-245
Mean corpuscular volume	0.0612	0.002	1.20E-224
Mean corpuscular volume (anaemics excluded)	0.0634	0.0021	3.50E-206
Red blood cell distribution width	-0.0429	0.0021	4.50E-103
Red blood cell distribution width (anaemics excluded)	-0.0454	0.0022	3.40E-100
Haematocrit percentage	-0.0349	0.0017	4.70E-98
Corpuscular haemoglobin concentration	0.0328	0.002	1.90E-60
Sphered cell volume	0.0308	0.0019	4.60E-60
Haemoglobin concentration	-0.0244	0.0016	4.40E-51
Platelet Count	0.0234	0.0019	6.40E-39
Non-alcoholic fatty acid liver disease fibrosis score	-0.0202	0.0019	2.20E-25
Platelet volume	-0.0178	0.0018	4.80E-25
Platelet crit	0.0165	0.0019	1.50E-19

Fibrosis-4 score	-0.0141	0.0017	4.70E-17
Reticulocyte volume	0.0152	0.002	3.30E-15
Reticulocyte percentage	0.013	0.002	2.70E-11
Platelet distribution width	-0.0124	0.0019	3.50E-11
High light scatter reticulocyte percentage	0.0114	0.002	1.10E-08

3.5 Discussion

In this chapter, I have shown how a human genetic variant can be used to assess the therapeutic profile and effects of long-term genetically mediated alterations in drug target levels. At the time of analysis, this was the largest GWAS meta-analysis of circulating EPO levels in individuals of European and African American descent. Through GWAS meta-analysis, I identified a genetic variant lying in *cis* with the *EPO* gene and used this variant as a partial proxy for therapeutic modulation of EPO levels to test the associated risk of CVD with long-term rises in endogenous EPO. Several lines of evidence indicate the *cis-EPO* variant is an excellent proxy for predicting the long-term cardiovascular risk associated with genetically mediated therapeutic rises in endogenous EPO levels. First*,* I found the A-allele of rs1617640 to increase circulating EPO levels in a meta-analysis of circulating EPO levels consistent with the direction of effect reported in previous studies in patients with diabetic retinopathy or hepatitis C receiving antiviral therapies (**Table 3.7)** (Amanzada et al., 2014; Tong et al., 2008). Second, I found the *cis-EPO* SNP to be an eQTL for hepatic *EPO* gene expression with evidence of colocalisation for being the same causal variant in the circulating EPO meta-analysis and hepatic *EPO* eQTL data (**Table 3.8*, *Figure 3.7**). Third, by performing a PheWAS, I found the *cis-EPO* variant to be associated with relevant erythrocyte phenotypes with similar effects to the effect on circulating EPO levels indicating that the SNP is having an effect through similar, expected, and relevant pathways **(Figure 3.9, Table 3.11).**

Having provided genetic evidence that a *cis-EPO* SNP is a valid genetic instrument and likely causal in altering *EPO* expression levels, I used this variant as a genetic proxy for long-term therapeutic modulation of endogenous EPO levels to show that higher endogenous EPO levels (equivalent to 5.1 IU/L) are not associated with an increased risk of CVD or elevated levels of clinical markers (SBP, DBP or resting heart rate) predisposing to CVD risk factors (e.g. hypertension) **(Figure 3.7).** To obtain a more representative and physiologically relevant effect of therapeutically altered endogenous EPO levels, I rescaled these genetic effects to the PHI-induced effects on endogenous EPO levels from a fixed dose Phase II trial (Meadowcroft et al., 2019) and was able to exclude an increased odds of 1.07, 1.15 and 1.07 for CAD, MI or stroke respectively and 0.78 mmHg increased levels of SBP **(Figure 3.7).** I was able to exclude adverse effects on DBP and heart rate **(Figure 3.7).** I found nominal evidence (*P* < 0.05) for an association between higher endogenous EPO levels and lower heart rate and DBP indicating that higher EPO levels may in fact have a protective effect on CVD risk factors (i.e. hypertension) **(Figure 3.7).** It is important to translate these findings at the clinical level to see if the reduction in heart rate or DBP (decrease between 0.14 - 0.78 beats per minute for resting heart rate and 0.14 - 0.71 mmHg for DBP) caused by 1 SD (equivalent to 5.1 IU/L) higher endogenous EPO levels is enough to potentially restore previously high BP levels within the normal range reducing strain on the heart and risk of CVD.

These results are consistent with the hypothesis that PHIs are likely safe for treating anaemia of CKD when only increasing circulating EPO levels within the physiological range. Due to the existence of outliers and/or skewed data in the Phase II trial (Meadowcroft et al., 2019, from where I extracted the effect of PHI on EPO), the actual standard deviation of change in EPO from baseline after PHI treatment is likely to be smaller meaning that the scaled estimates will be similarly affected by these outliers and have wider confidence intervals. Although I also scaled the genetic effects to that of rhEPO, I was unable to make any meaningful interpretations due to the large effect of rhEPO on circulating EPO levels and subsequent large confidence intervals (**Table 3.10).** This is likely due to rhEPO resulting in supra-physiological circulating EPO levels which is not accurately mimicked by genetic variants acting at the transcriptional level on physiological levels which have smaller effect-sizes. Scaling genetic estimates to the effect of a fixed dose after four weeks of treatment may not be the most clinically relevant since PHIs require titration to a Hgb target, but this was the best data I could obtain as changes in EPO levels were only measured in the one fixed dose Phase II trial (Meadowcroft et al., 2019). Since PHIs work through the same mechanism, these results would be supportive of all PHI compounds (daprodustat, vadadustat and roxadustat) with respect to the genetically predicted effect of therapeutically altered endogenous EPO levels. Any slight differences in effects on EPO levels and CVD would likely be related to independent biochemical properties of the compound and variations in dosages and the conductance of clinical trials. PHIs are also likely to affect transcription of other hypoxic response genes and have off-target effects which were not investigated in this study. This data provides genetic evidence supporting the Phase III clinical results that PHIs are likely noninferior than rhEPO for CVD. Further longitudinal clinical analysis will help decipher the uncertainty behind the cardiovascular risk associated with these treatments and will further emphasise the utility of using human genetics to provide an insight into drug safety and efficacy.

Our findings support previous studies which have also found rs1617640 to have an allele-specific effect on *EPO* expression levels. Some of these studies have reported the A-allele to be associated with higher EPO concentrations (Amanzada et al., 2014; Tong et al., 2008), whilst others reported conflicting evidence with the C-allele being associated with higher EPO concentrations and promoter activity (Y. Fan et al., 2016; Kästner et al., 2012). These findings together with the fact I did not find the SNP to be a renal eQTL but a hepatic eQTL suggest that the rs1617640 polymorphism has different effects depending upon cell-type, physiological condition, state and timing (**Table 3.8**) (Renner et al., 2020). The lack of association in the renal eQTL analysis may be attributable to the small study size (*N* = 286) and therefore lack of power. The emergence of larger eQTL datasets in relevant tissue types would provide more power and could strengthen associations. The renal dataset is also based on bulk tissue sample as opposed to single cell-type data and it may be that the *cis-EPO* variant only influences *EPO* gene expression in certain renal cell-types, such as the interstitial cells (the predominant site of EPO production). The direction of effect was also inconsistent to the effect obtained in the meta-analysis with the C-allele associated with higher *EPO* mRNA expression as opposed to the A-allele. Again, this may be due to the eQTL analysis being in bulk tissues as opposed to specific single cell-types. Functional investigation into the effects of this polymorphism in relevant tissues and cell-types during different developmental stages would further reveal the exact functions and downstream effects of the *cis-EPO* variant and help provide additional validation that the variant is a valid proxy. I also found evidence supporting the *cis-EPO* variant as a renal eQTL for neighbouring *TFR2* gene expression (ß = 0.23 [0.17, 0.29], *P* = 9.33 x 10^{-14}), a gene involved in iron transport (**Table 3.8**). This is likely an example of coordinate regulation; iron is required to make more Hgb so if *EPO* expression is upregulated to increase Hgb levels then iron transport, and hence *TFR2* gene expression, would be too (Forejtnikovà et al., 2010; Haase, 2013). Further investigation into these genes and pathways and the associated risk of adverse effects could be warranted to gain a better understanding of the downstream effects of PHIs acting on the hypoxic pathway.

There are some limitations to our study. Our results, as with any study using human genetics as predictors, cannot rule out any adverse effect, but instead can provide upper bounds on their probability (Page, 2014). First, genetic analyses are often performed in a general, 'healthy' population as opposed to a diseased cohort in whom treatment is used. Diseased populations may respond differently to what is estimated by the genetic association due to variable baseline levels or additional underlying conditions (Mokry et al., 2015; Sofianopoulou et al., 2021). Despite rescaling the genetic effect to the PHI-induced effect to try and overcome this, I still assume linearity which may not be the case in reality. For example, causal estimates could change dependent upon baseline levels and therefore inferences about the likely effect at individual anaemic CKD patient level need careful consideration particularly when doses are titrated (Sofianopoulou et al., 2021). As larger studies become available, particularly in disease-specific cohorts, our power to detect associations and ability to perform stratified analyses will increase and we will become more confident about the conclusions drawn from these types of investigations (Sofianopoulou et al., 2021; Visscher et al., 2017). Second, common genetic variants differ from clinical trials in that they represent subtle, lifelong perturbations whereas clinical trials test more acute larger changes (Burgess et al., 2012; Pulley et al., 2017). Additionally, the biomarker or drug may only be efficacious in specific physiological states and it is difficult to represent these different states using genetic approaches (Burgess et al., 2012; Mokry et al., 2015; H.-C. Yang et al., 2010).

MR also has its limitations; it is important to consider the strength and validity of the genetic instrument and that it meets the MR assumptions (outlined clearly in **1.2.5**). Previous studies have shown that the best proxies to mirror pharmacological effects are those variants lying within or close to the gene encoding the drug target which are therefore most likely to have functional impact on the protein product (Lauridsen et al., 2015; Matías-García et al., 2021; Melzer et al., 2008; J. Zheng et al., 2020). As I was interested in investigating the causal effect of altering *EPO* gene expression through PHI use, I decided to focus specifically on the *EPO* gene and identified the *cis-EPO* variant for use as a partial proxy therapeutic modulation of EPO levels rather than using genetic variants passing genome-wide significance in the circulating EPO meta-analysis. The variants passing genome-wide significance were found to have a stronger, primary effect on other erythrocyte phenotypes rendering them weak EPO instruments (**Table 3.6**). The association between the identified *cis-EPO* variant and EPO levels, however, did not reach formal levels of significance in our meta-analysis ($P < 5 \times 10^{-08}$), a key limitation, and the effect estimate was small which may indicate weak instrument bias and the tendency for causal estimates to shift towards the observational data (**Table 3.7**) (Burgess, Small, et al., 2017; Lawlor, 2016). Furthermore, the *cis-EPO* variant did not appear to be the most strongly associated variant with circulating EPO levels within the 500 kb window either side **(Figure 3.10).** However, all other leading variants had p-values within the same order of magnitude ($\times 10^{-04}$) to the *cis-EPO* SNP and were in strong LD (**Figure 3.10**). As I was using a single genetic instrument, it is difficult to fully test for horizontal pleiotropy as I was unable to use established methods such as MR Egger (Bowden et al., 2015). However, I did perform colocalisation analysis between our meta-analysis and liver eQTL data and found evidence that the same causal variant is shared (posterior probability = 71%) **(Figure 3.6).** I was therefore able to eliminate some unreliable associations. The posterior probability was likely influenced by the weak significance of the genetic signal and therefore larger sample size could further improve evidence for colocalisation. Furthermore, I found rs1617640 to be associated with several relevant erythrocyte phenotypes in a PheWAS performed in a study of much greater sample size (UKB; N = 451,099). Unfortunately, I did not have data on EPO measures in this large number. As larger cohorts containing relevant genetic data, biomarker measurements and disease outcome classifications become available, potential concerns regarding weak instrument bias and power are likely to be overcome (Spencer et al., 2009; Visscher et al., 2017).

The *cis-EPO* SNP lies at the biomarker level (i.e. at the EPO level). This may mean I am missing unintended drug effects that may occur through alternative mechanistic pathways (e.g. EPO-independent pathways). For this reason, investigation using other genetic variants to mirror the effects of PHIs, such as those lying within *EGLN* genes (the targets of PHIs), would be useful in further predicting unintended drug effects and provide additional genetic evidence into potential safety to support drug development. Moreover, understanding the functional consequence of the genetic variant can aid our understanding and drug target validation. For example, previous studies have shown how loss-of-function (LOF) variants can be used to mirror and predict the consequences of antagonistic drug treatments (Stitziel et al., 2014). Further investigation into the functional role of rs1617640 in controlling endogenous EPO levels and understanding the mechanism of action would consolidate the use of the *cis-EPO* SNP as an instrument.

In summary, this Chapter indicates that genetically proxied long-term rises in endogenous EPO levels do not increase cardiovascular disease risk, with upper limits of 1.07, 1.15 and 1.07 for CAD, MI and stroke respectively given a 2.2 unit rise in endogenous EPO levels. Understanding the relationship between EPO and CVD is an important and unresolved question, and the identification of a relevant genetic marker that can test the long-term effect of therapeutic action could potentially inform further research using patient-level clinical data from Phase III trials.

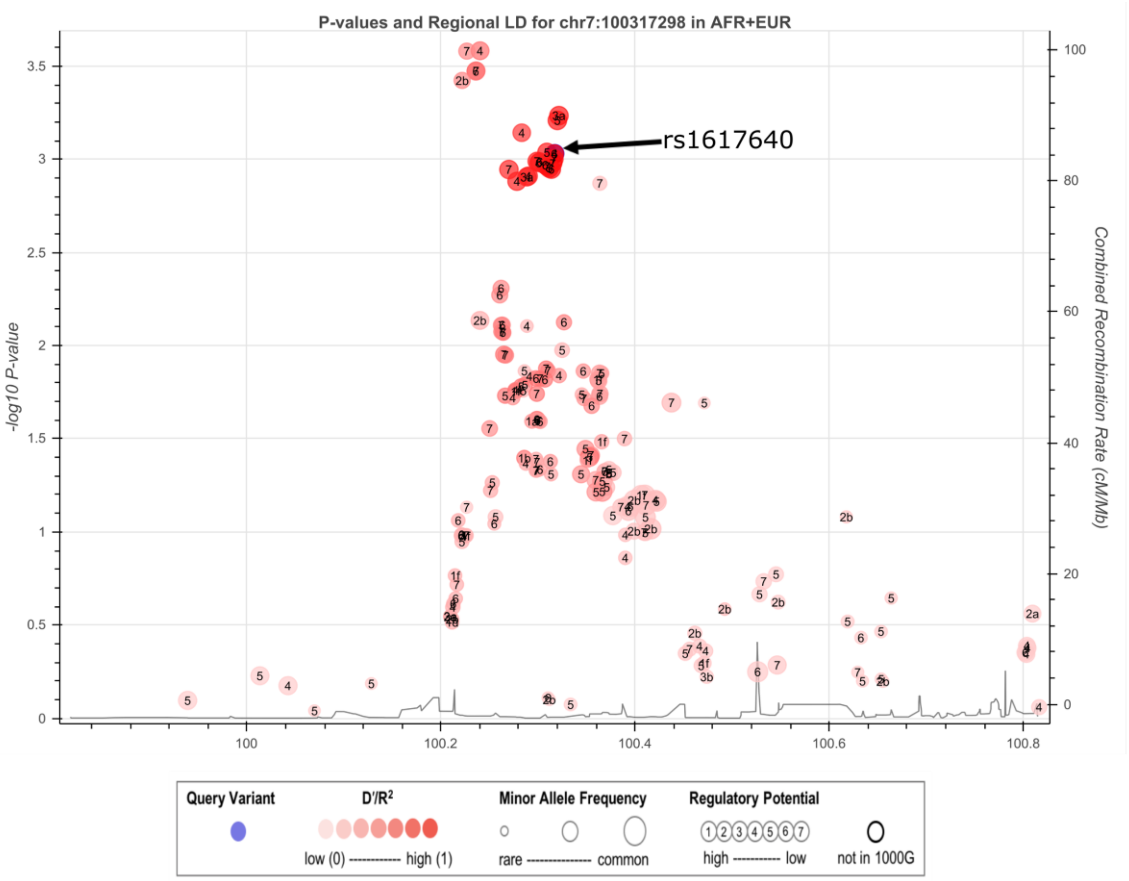


Figure 3.10: Regional LD pattern for the cis-EPO SNP within a 1 mb region. *Each circle represents an individual genetic variant. The one in blue and labelled is the cis-EPO SNP. The size of the circle represents the minor allele frequency (smallest being rare, largest being common). The shading pattern represents the correlation between the query variant (rs1617640) and the nearby variant. The numbers within each circle (1-7) highlight the regulatory potential with 1 being high and 7 being low. The -log$_{10}$ p-values obtained in the meta-analysis of circulating EPO levels in 6,127 individuals of African-American and European descent are shown on the left-hand-side y-axis. The pattern of LD is shown on the right-hand-side y-axis. The x-axis shows the chromosomal position on chromosome 7. Figure produced using LDLink (https://ldlink.nci.nih.gov/?tab=ldassoc).*

Chapter 4 Establishment of whole *EPO* gene knock-out cell model using CRISPR-Cas9 gene-editing and whole transcriptomic profiling to better characterise molecular pathways involved in EPO signalling

This Chapter includes sections that have been taken directly from a pre-print paper in which I am the first author. I have reformatted and expanded on sections for the purpose of this thesis. This paper is currently undergoing reviews at GSK and AJHG.

Harlow, CE. Gandiwijaya, J. Bamford, RA. Wood, AR. Van der Most, P. Verweij, N. [25 authors] & Frayling, TM. 2022. Identification and single-base gene-editing functional validation of a *cis-EPO* variant for use to mimic novel EPO-increasing therapies.

I planned all of the experiments and performed the majority of the laboratory works carried out in this Chapter with the help, support and guidance of my supervisors, Professor Tim Frayling and Dr Asami Oguro-Ando, and my fellow Oguro-Ando lab members, primarily Mr Josan Gandawijaya and Dr Rosie Bamford. This is the first time an *EPO* knock-out cell-line has been established in HEK-293 cells and that extensive whole transcriptomic profiling has been performed comparing transcriptomic changes between wild-type and *EPO* knock-out cell-lines. The CRISPR-Cas9 plasmids used in these experiments were kindly gifted; pU6-(BbsI)_CBh-Cas9-T2A-mCherry was a gift from Ralf Kuehn (Addgene plasmid # 64324 ; http://n2t.net/addgene:64324 ; RRID:Addgene_64324) and pSpCas9(BB)-2A-GFP (PX458)) was a gift from Feng Zhang (Addgene plasmid #41838; http://n2t.net/addgene:48138; RRID:Addgene_48138). I would like to acknowledge and thank the Exeter Sequencing Service at the University of Exeter for performing the RNA-sequencing for me and for Dario Parciti for his initial guidance with the RNA-seq analysis. This Chapter utilised equipment funded by the Wellcome Trust Institutional Strategic Support Fund (WT097835MF), Wellcome Trust Multi User Equipment Award (WT101650MA) and BBSRC LOLA award (BB/K003240/1).

4.1 Introduction

Erythropoietin (EPO) is a glycoprotein cytokine that is primarily responsible for the production, development and survival of erythrocytes (Yuanyuan Zhang et al., 2014). EPO is primarily produced by the kidney and foetal liver in response to oxygen levels and exerts its effects by binding to the EPO receptor (EPOR) (Suresh et al., 2020). During homeostasis, low levels of basal EPO are predominantly released by the peritubular capillary endothelial cells of the kidney to maintain a constant supply of oxygen to tissues (Jelkmann, 2011). In the presence of hypoxia or low erythrocyte numbers, the expression of *EPO* is tightly upregulated via the hypoxia-inducible factors (HIFs) which in turn increases erythropoiesis resulting in higher haemoglobin (Hgb) levels and restored oxygen-carrying capacity and delivery (Haase, 2013; Noguchi et al., 2008). Both EPO and EPOR have recently been identified in tissues other than the kidneys and liver including the brain, retina, and myoblasts indicating a role for EPO in both the haematopoietic and non-haematopoietic systems (Lamon & Russell, 2013). Non-haematopoietic roles of EPO include, but are not limited to, exerting neuroprotective effects, antiapoptotic effects, regulating angiogenesis, a response to inflammation and stress, energy homeostasis, metabolism, and an immune response (Suresh et al., 2020; L. Wang et al., 2014; Yuanyuan Zhang et al., 2014). These effects highlight the pleiotropic activities of EPO which differ depending upon the tissue or cell-type (Broxmeyer, 2013). Further understanding of the exact cellular pathways that are activated by EPO in specific tissues and cells is needed.

Therapies increasing circulating EPO levels are used to treat anaemia in CKD due to low EPO levels being one of the main drivers of anaemia in CKD patients (Fishbane & Spinowitz, 2018; Jelkmann, 2013; Pfeffer et al., 2009). However, it is very difficult to establish an animal or cell model with a phenotype similar to anaemia of CKD, particularly the low EPO levels (Yuanyuan Zhang et al., 2014). Therefore, few studies establishing *EPO* knock-out or knock-down have been reported due to embryonic lethality, which has hampered functional analysis of the mechanisms and downstream pathways of *EPO* in different systems (C. S. Lin et al., 1996; Wu et al., 1995; Yuanyuan Zhang et al., 2014). One *EPO* knock-out model in zebrafish investigated the role *EPO* plays in the kidneys and conditional knock-out studies in mice investigated regulatory regions of the *EPO* gene with the hope of highlighting how these models could have clinical utility rather than focusing on biological pathways and functions (Yamazaki et al., 2021). The ability to extrapolate the findings in animal models to the human specific effects of EPO is limited; animals have different systems than humans, a human disease phenotype and/or the disease causation cannot always be replicated in animals and there are fundamental differences between species in terms of genetics, physiology and molecular biology (Leenaars et al., 2019; McGonigle & Ruggeri, 2014). It is therefore important to develop a better understanding of the human specific signalling cascades and functional impact of influencing *EPO* expression, particularly in relation to disease.

Several cell-lines, such as HepG2, erythroid cell-lines (TF-1 and UT-7) and renal EPO-producing cells (REPCs), have therefore been used to investigate the regulation and downstream effects of EPO (Chamorro et al., 2013; Chin et al., 2019; Frede et al., 2011; Imeri et al., 2019; Obara et al., 2008; Udupa, 2006). However as mentioned above, there are very few studies establishing a knock-out. Instead, many studies have explored the impacts of treatment with EPO, or hypoxia-inducible compounds like cobalt chloride, in a wide-range of different models to assess the effects of exogenous EPO on different pathways (Berlian et al., 2019; Cavadas et al., 2016; Chamorro et al., 2013; X.-H. Liu et al., 1999; Park et al., 2015; Rana et al., 2019; Tani et al., 2020). Despite, these models providing a better understanding of the pleiotropic functions of EPO and replicating the effects stimulated by current treatments for anaemia in CKD, the models are not always physiologically relevant and it is difficult to elude which pathways are directly affected by *EPO* gene expression itself and not a response to high exogenous levels. Moreover, several studies also focus on other genes involved in the hypoxic response pathway as opposed to the effects of *EPO* itself. For example, several studies have performed knock-out experiments on the HIFs or EPOR to investigate the effect of these on downstream biological mechanisms but not the direct effects of EPO (Cimmino et al., 2019; Q. Liu et al., 2018; Luk et al., 2013; Paliege et al., 2010; Våtsveen et al., 2016). For these reasons, the downstream molecular functions, biological pathways and signalling cascades of endogenous EPO in different cell-types and systems remain elusive. It is therefore essential to establish a knockout of the *EPO* gene in a relevant human cell-line.

Over the last decade, CRISPR-Cas9 gene-editing has come to the forefront making it more straightforward to edit the genome of cells and animals (Agrotis & Ketteler, 2015; Doudna & Charpentier, 2014; Ran, Hsu, Wright, et al., 2013). CRISPR-Cas9 has aided the ability to correlate genes or genetic variants to disease, gene expression, and biological pathways (H. Li et al., 2020).

Whole transcriptomic profiling followed by subsequent gene ontology (GO) and enrichment analysis can be used to identify widespread gene expression differences and provide a better understanding of the downstream biological pathways, mechanisms, and networks altered by the cell model (E. Y. Chen et al., 2013; Kuleshov et al., 2016; Z. Xie et al., 2021). Traditionally, gene expression microarrays were used but this has since been replaced with highly parallel high-throughput next generation sequencing (NGS) technologies, such as RNA-sequencing (RNA-seq). RNA-seq has made it possible to identify widespread gene expression changes reliably and accurately with an unbiased insight into all transcripts (M. S. Rao et al., 2019; Z. Wang et al., 2009). RNA-seq analysis allows transcript and isoform level quantification (given adequate sequencing depth), alongside identification of novel transcripts, gene fusions and single nucleotide variants in a single assay without needing prior knowledge; a clear advantage compared to pre-existing technologies (Mortazavi et al., 2008; Z. Wang et al., 2009; Wilhelm & Landry, 2009). Additionally, RNA-seq can quantify gene expression changes over a larger dynamic range than microarray technologies, has higher specificity and sensitivity, and can detect rare and low-abundance transcripts (J. Li et al., 2016; Y. Liu et al., 2015; C. Wang et al., 2014; Z. Wang et al., 2009; S. Zhao et al., 2014).

A typical RNA-seq experiment involves isolation of RNA, conversion to complementary DNA (cDNA), preparation of sequencing libraries through addition of adapters and sequencing using NGS platforms, such as Illumina's HiSeq 2500 system which was used in this Chapter (**Figure 4.1**) (Kukurba & Montgomery, 2015). As ribosomal RNA (rRNA) accounts for over 95% of the cellular RNA, different techniques are employed prior to reverse transcription to deplete rRNA and isolate alternative RNA species (Conesa et al., 2016). For whole transcriptomic profiling, messenger RNA (mRNA) can be isolated by selecting for the presence of 3’polyadenylated (poly-A) tails using poly-T oligos that are covalently attached to magnetic beads (**Figure 4.1**) (Kukurba & Montgomery, 2015). The majority of high throughput NGS platforms employ the *sequencing-by-synthesis* method to allow for parallel sequencing of millions of cDNA molecules (**Figure 4.2***)* (Fuller et al., 2009). Briefly, cDNA fragments are immobilised on a flow-cell by complementary binding of the adapters to surface bound probes. Fragments are then clonally amplified through bridge amplification to create clusters of identical copies of DNA molecules (**Figure 4.2**). Clonal molecules are then sequenced by Illumina platforms using the ensemble-based sequencing-by-synthesis method which involves the detection of single bases through the binding of DNA polymerase and the incorporation of fluorescently-labelled reversible-terminator nucleotides (each nucleotide has its own fluorescent marker) into the growing DNA sequence (**Figure 4.2**) (Bentley et al., 2008; C.-Y. Chen, 2014). As molecules are clonal, this approach provides relative RNA expression levels of genes (Kukurba & Montgomery, 2015). The ensemble-based *sequencing-by-synthesis* approach, as opposed to the single-molecule-based approach, has low sequencing error rates (Fuller et al., 2009). Downstream computational analysis can then be performed on the resulting RNA-seq reads to check for sequencing quality, map reads to the reference genome and enable quantification of the number of reads aligning to particular genes. Biases that may be incurred during sequencing, such as polymerase chain reaction (PCR)-amplification bias, can be controlled for during these steps (Mandelboum et al., 2019; Roberts et al., 2011; W. Zheng et al., 2011). Several RNA-seq analysis pipelines have been developed and numerous packages are available depending upon the research question being proposed (Conesa et al., 2016; Love et al., 2015). Once RNA-seq analysis has been performed and a list of differentially expressed genes (DEGs) across samples obtained, it is important to validate the findings using an independent technique, such as quantitative reverse-transcription polymerase chain reaction (qRT-PCR) (Lowe et al., 2017)**.**

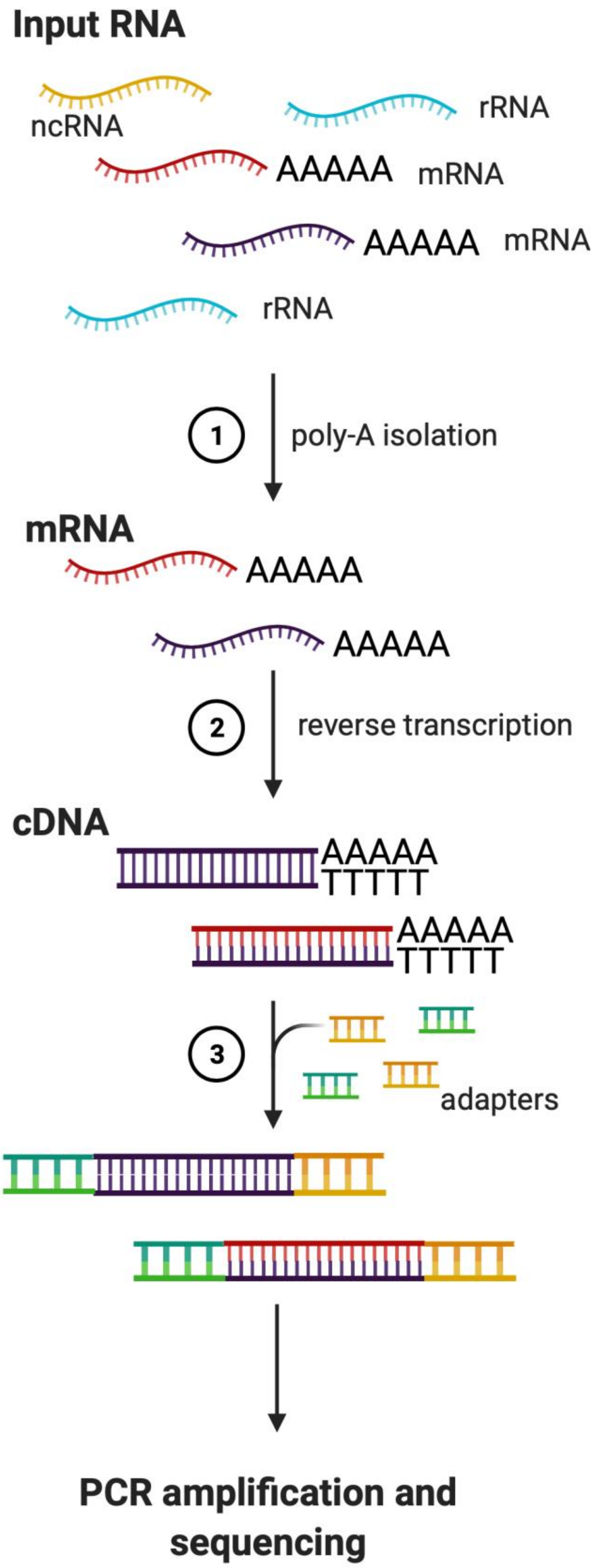


Figure 4.1: Schematic representation of a typical RNA-seq library preparation protocol for isolation of mRNA.*Total RNA is extracted from the biological material of interest e.g. cells or tissue. 1) Particular RNA molecules are isolated using specific protocols, such as poly-A selection to enrich for polyadenylated transcripts. 2) RNA is converted to complementary DNA (cDNA) using reverse transcription. 3) Sequencing adapters, specific to the platform being used, are ligated to the ends of the cDNA fragments. Following PCR amplification of these fragments, the RNA-seq library is ready for sequencing. ncRNA = non-coding RNA; rRNA = ribosomal RNA. Created with BioRender.com.*

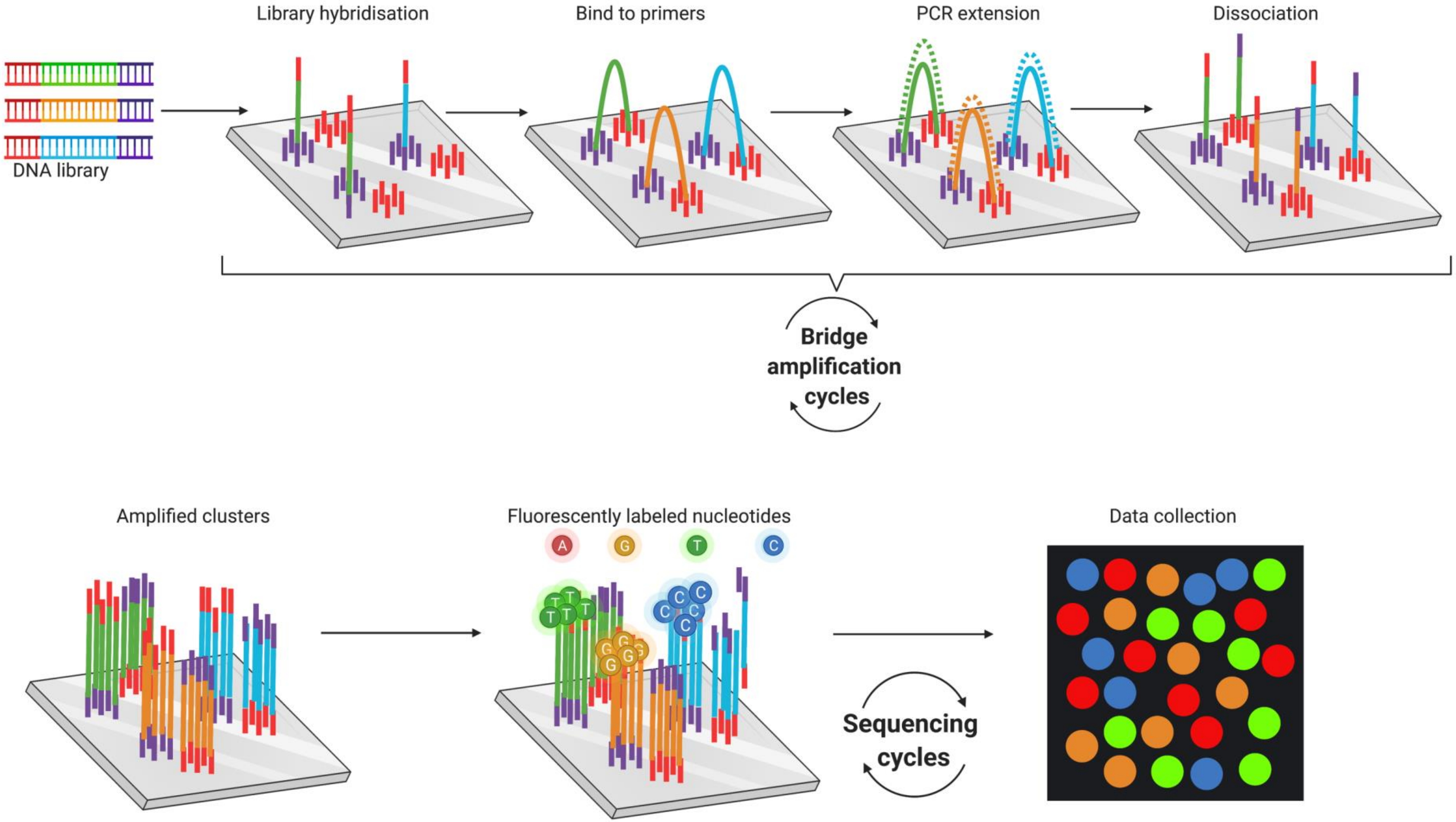

Library hybridisation
Bind to primers
PCR extension
Dissociation
DNA library
Bridge amplification cycles
Amplified clusters
Fluorescently labeled nucleotides
A
G
T
C
Data collection
Sequencing cycles

Figure 4.2: Schematic representation ensemble-based sequencing-by-synthesis used by the Illumina platforms. See next page for legend.***Figure 4.2: Schematic representation ensemble-based sequencing-by-synthesis used by the Illumina platforms.*** *Adapter sequences on the ends of the cDNA fragments bind to complementary probes on the flow-cell resulting in library hybridisation. Bridge PCR reactions amplify each bound fragment producing clusters of clonal fragments. During each sequencing cycle, due to the presence of DNA polymerase, one fluorescently-labelled reversible-terminator nucleotide is incorporated into the growing DNA strand. A laser excites the attached fluorophore in each cluster being sequenced. An optic scanner collects the signal from each fragment cluster. As each nucleotide is labelled with a separate fluorophore, the nucleotide at that position in the sequence can be determined. The sequencing terminator with the fluorophore is removed from each fragment cluster and the next sequencing cycle starts. Created with BioRender.com.*

4.2 Chapter Aims

In this Chapter, I took advantage of CRISPR-Cas9 gene-editing technology and high throughput RNA-seq to test the *in vitro* effects of changes to EPO levels. I hoped to gain further insight into the cell-line specific downstream causal genes and signalling cascades as a result of the largest change in EPO levels i.e. EPO or no EPO. This would provide a better understanding of the genes and downstream pathways implicated in response to EPO changes, and potentially enable the identification of novel genes and pathways that are altered in response to a lack of *EPO in vitro.* The findings could then be used when elucidating the role and functionally validating the effects of the rs1617640 variant on controlling *EPO* expression levels (see **Chapter 5**).

The specific aims for this Chapter were to:

1. Establish a whole *EPO* gene knock-out in a human cell-line using CRISPR-Cas9 gene-editing technology.
2. Perform RNA-seq experiments to identify a list of differentially expressed genes (DEGs).
3. Perform gene ontology (GO) analysis to identify downstream biological pathways and molecular functions of EPO.
4. Functionally validate the findings from RNA-seq analysis.

4.3 Methods

An overview of the series of experiments undertaken to establish a whole *EPO* gene knock-out and identify transcriptomic changes can be seen in **Figure 4.3.** The cell-line used throughout was the Human Embryonic Kidney (HEK)-293 cell-line as *EPO* is highly expressed in the kidneys. Standard cell culture methods were used to grow and maintain the cell-line (see **Chapter 3**).

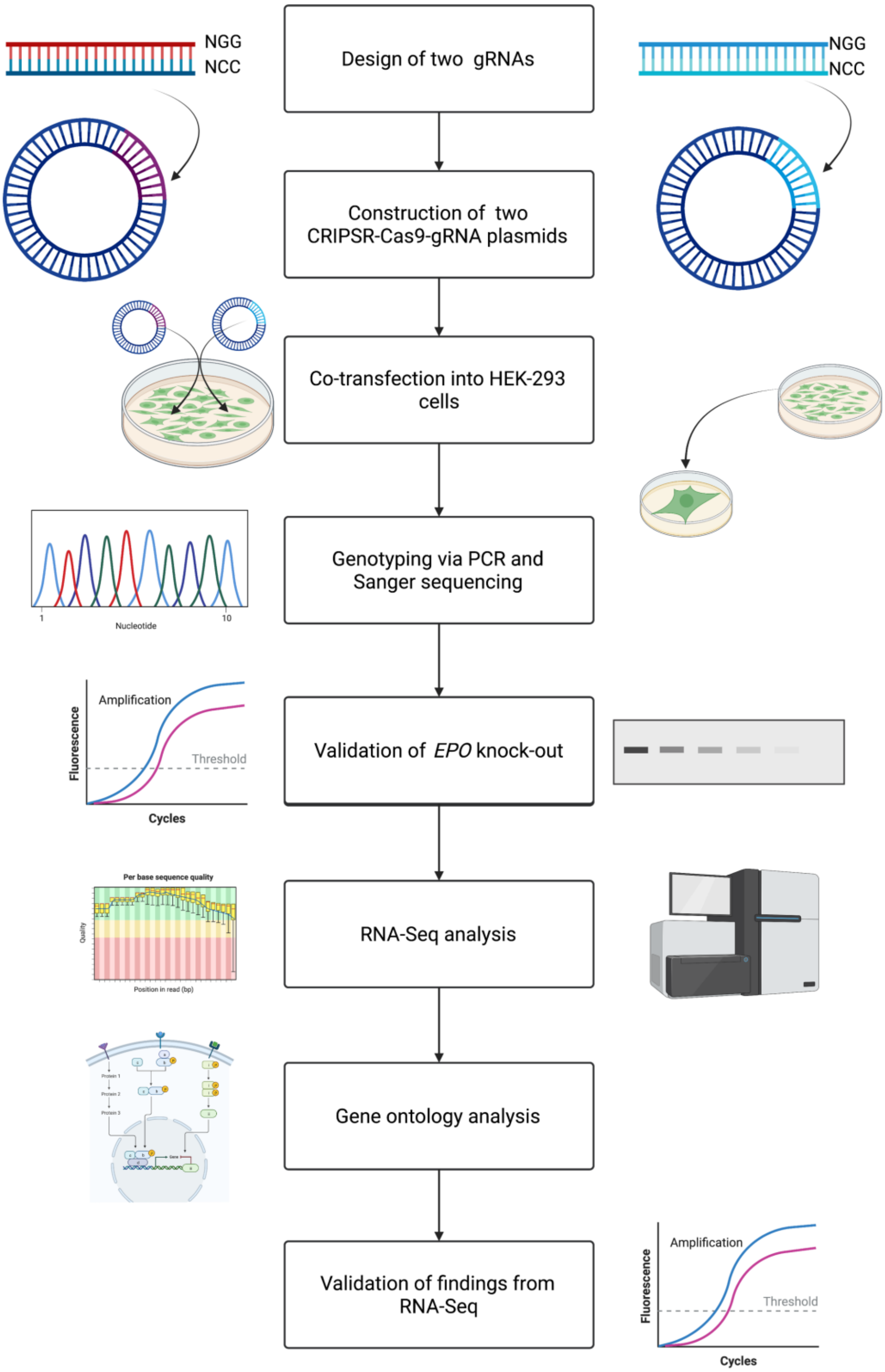
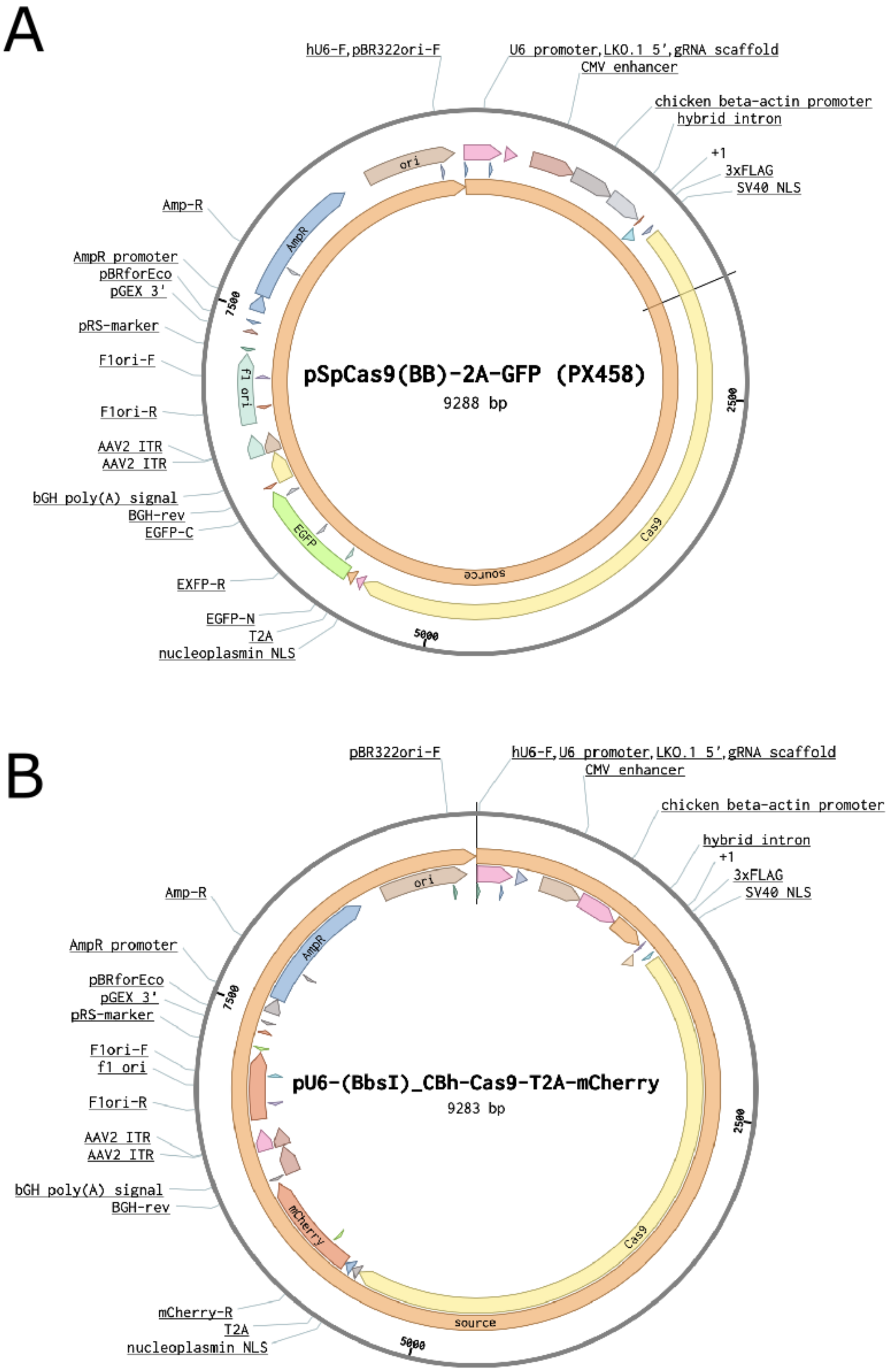

NGG
NCC
Design of two gRNAs
NGG
NCC
Construction of two CRIPSR-Cas9-gRNA plasmids
Co-transfection into HEK-293 cells
Genotyping via PCR and Sanger sequencing
1
Nucleotide
10
Amplification
Fluorescence
Threshold
Cycles
Validation of *EPO* knock-out
Per base sequence quality
Quality
Position in read (bp)
RNA-Seq analysis
Gene ontology analysis
Validation of findings from RNA-Seq
Amplification
Fluorescence
Threshold
Cycles

Figure 4.3: Overview of the experimental plan for the establishment of a whole EPO knock-out cell-line and whole transcriptomic analysis. *CRISPR-Cas9 gene-editing technology with paired gRNAs was used to target the EPO gene and result in EPO gene disruption. After confirmation that the EPO gene had successfully been disrupted through Sanger sequencing, PCR, qRT-PCR and western blotting, knock-out cell-lines were sent for whole transcriptomic analysis using RNA-seq. An RNA-seq analysis pipeline was established and gene ontology analysis performed to determine the downstream effect of dysregulating EPO expression. Findings from RNA-seq were validated by qRT-PCR. Created with BioRender.com*

4.3.1 Plasmids

Two commercially available genomic CRISPR-Cas9 plasmids containing fluorescent markers of either EGFP (enhanced green fluorescent protein) (pSpCas9(BB)-2A-GFP (PX458), Addgene: #48138) or mCherry (red fluorescent marker) (pU6-(BbsI)_CBh-Cas9-T2A-mCherry, Addgene: #64324) to enable positive selection after transfection were purchased from Addgene (http://www.addgene.org) (**Figure 4.4**



).

Each plasmid contains resistance to ampicillin for use in cloning (Chu et al., 2015; Ran, Hsu, Lin, et al., 2013).

4.3.2 Construction of CRISPR-Cas9 plasmid

4.3.2.1 *Design of paired gRNAs targeting the EPO gene*

To maximise efficiency of achieving full *EPO* gene knock-out, paired gRNAs were designed. gRNA sequences were identified using the online CRISPR design tool (available at https://www.benchling.com/crispr/) by screening the exonic regions conserved across *EPO* transcripts. Sequences with the highest predicted off-target (>60) and on-target (>50) scores and least matches to other genomic locations through a BLAST search were chosen. One gRNA targeted exon 2 and the other gRNA targeted exon 4 **(Figure 4.5, Table 4.1).** Overhangs for cloning into a BbsI restriction site were added to the two ends of the gRNA sequences **(Table 4.1).** Final gRNA sequences were ordered from IDT (Integrated DNA Technologies, Leuven, Belgium; https://eu.idtdna.com/). gRNA sequences were annealed and phosphorylated using T4 Polynucleotide Kinase (PNK) (New England BioLabs, Ipswich, UK) and then ligated into the CRISPR-Cas9 plasmid in a single digestion and ligation reaction with an insert:plasmid ratio of 3:1 (calculated using the New England BioLab calculator - https://nebiocalculator.neb.com/#!/ligation) using BbsI restriction enzyme and T4 ligase (New England BioLabs, Ipswich, UK). The gRNA targeting exon 2 was cloned into the CRISPR-Cas9 plasmid with the EGFP fluorescent marker (pSpCas9(BB)-2A-GFP (PX458)) whilst the gRNA targeting exon 4 was cloned into the CRISPR-Cas9 plasmid with the mCherry fluorescent marker (pU6-(BbsI)_CBh-Cas9-T2A-mCherry).

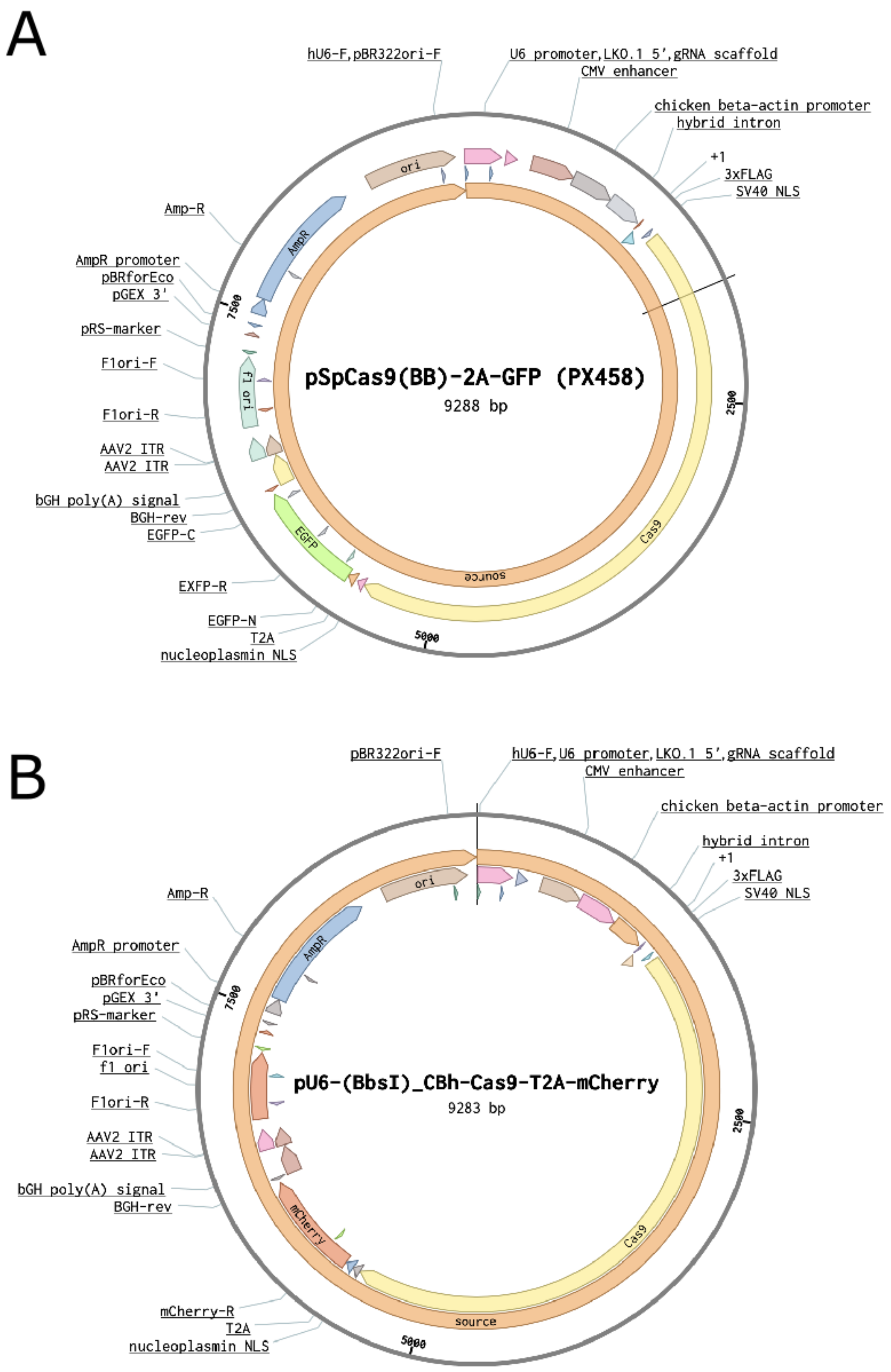


Figure 4.4: **Plasmid map of the CRISPR-Cas9 vector backbones.** ***A:*** *Plasmid map of the vector backbone of the pSpCas9(BB)-2A-GFP (PX458) (Addgene: #48138) CRISPR-Cas9 plasmid.* ***B:*** *Plasmid map of the vector backbone of the pU6-(BbsI) CBh-Cas9-T2A-mCherry (Addgene: #64324) CRISPR-Cas9 plasmid. Both plasmids contain ampicillin resistance genes to enable cloning. The gRNAs were ligated into the gRNA scaffolds of the plasmids. Plasmid maps obtained from Benchling.com.*

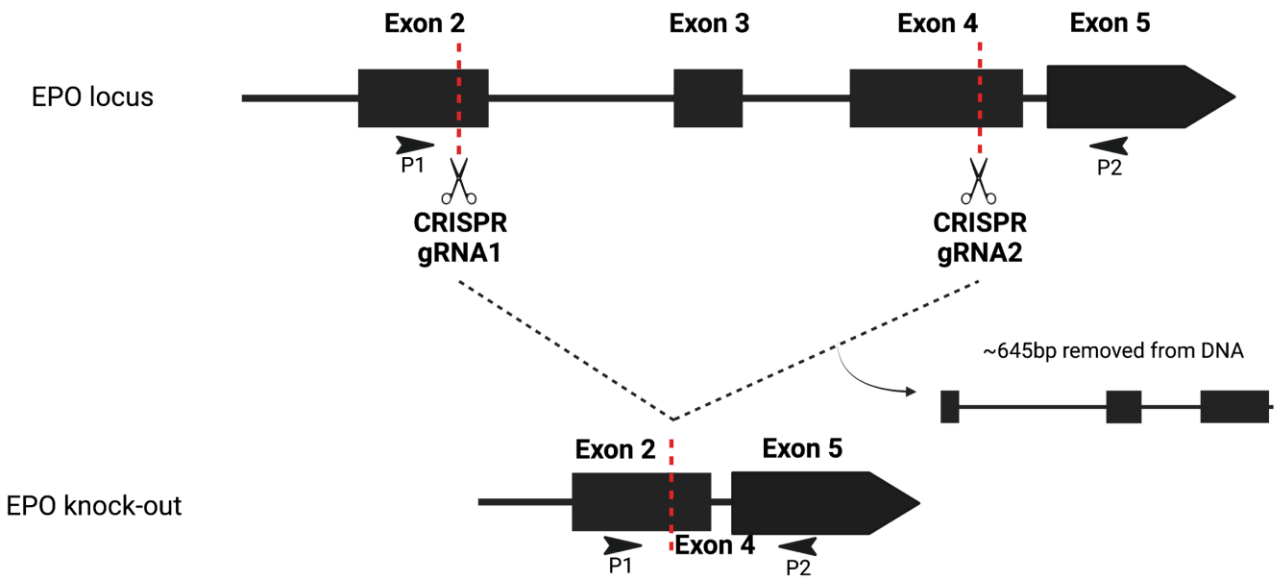


Figure 4.5: Schematic of the genomic location of the two gRNA sequences targeting the EPO gene and the expected region to be removed from the DNA upon successful cleavage by the Cas9 endonuclease. *The location of primer sequences (P1 & P2) for genotyping are indicated by the arrows. Created with BioRender.com.*

Table 4.1: The design of the paired gRNA sequences for targeting the EPO gene.

*One gRNA was designed targeting exon 2 and the other was designed targeted exon 4. Overhangs complementary to the BbsI restriction enzyme cut site (shown in **bold**) were added to the ends of the forward and reverse gRNA sequence to enable cloning of the gRNA into the CRISPR-Cas9 plasmid (pSpCas9(BB)-2A-GFP (PX458), Addgene: #48138 for gRNA targeting exon 2; pU6-(BbsI)_CBh-Cas9-T2A-mCherry, Addgene: #64324, for gRNA targeting exon 4). A 'G' was added after the 'CACC' sticky end if not already present on the 5' sequence, and a complementary 'C' to the 3' end.*

gRNA	**Chromosomal Location**	**Strand**		**Sequence**
EPO_exon2_gRNA_Fwd	7:100721655-100721674	-	Raw gRNA sequence	5'-AGAGGTACCTCTCCAGGACT**CGG**-3'
			Calculate reverse complement of raw gRNA sequence without PAM	5'-AGTCCTGGAGAGGTACCTCT-3'
			Add BbsI overhangs	**5'-CACCG**AGAGGTACCTCTCCAGGACT-3' **5'-AAAC**AGTCCTGGAGAGGTACCTCT**C**-3'
			Final oligo sequences for ordering	5'-CACCGAGAGGTACCTCTCCAGGACT-3' 3'- CTCTCCATGGAGAGGTCCTGACAAA-5'
EPO_exon4_gRNA_Fwd	7:100722778-100722797	+	Raw gRNA sequence	CATGTGGATAAAGCCGTCAG**TGG**
			Calculate reverse complement of raw gRNA sequence without PAM	CTGACGGCTTTATCCACATG
			Add BbsI overhangs	5'-**CACCG**CATGTGGATAAAGCCGTCAG-3' 5'-**aaac**CTGACGGCTTTATCCACATG**C**-3'
			Final oligo sequences for ordering	5'-CACCGCATGTGGATAAAGCCGTCAG-3' 3'- CGTACACCTATTTCGGCAGTCCAAA-5'

4.3.2.2 *Cloning of the recombinant plasmids into bacteria*

To obtain multiple copies of the recombinant plasmid, I first transformed bacteria with the ligation reaction before purifying and screening for the final CRISPR-Cas9-gRNA constructs. 50 µl of sub-cloning efficiency competent Escherichia Coli DH5α were thawed and mixed with 2 µl of ligation reaction (or a pUC19 control DNA as negative control). The reaction mix was incubated on ice for 15 minutes before being heat shocked for 45 seconds at 42^{0}C and incubated on ice for 2 minutes. Bacteria were mixed with 950 µl of Lysogeny Broth (LB) (ThermoFisher Scientific, Massachusetts, USA) supplemented with 1X SOC media (10 mM NaCl, 2.5 mM KCl, 10 mM MgSO4, 20 mM Glucose, 10 mM MgCl2) and incubated for 1 hour at 37^{0}C shaking at 225 rpm. Bacteria were pelleted by centrifugation at 4000 rpm for 2 minutes. Pellets were resuspended in 250 µl LB (ThermoFisher Scientific, Massachusetts, USA) and plated on an agar plate containing 100 µg/ml ampicillin (Merck Life Science, Watford, UK). After incubation overnight at 37^{0}C, single colonies containing transformed plasmids were selected and inoculated in LB containing ampicillin (Merck Life Science, Watford, UK). Cultures were then incubated overnight at 37 ^{0}C, shaking at 225 rpm. Plasmids were purified the following day from the overnight cultures using the QIAprep Spin Miniprep Kit (Qiagen, Maryland, USA) following manufacturer's instructions. The yield and purity of purified plasmids was measured using the Nanodrop ND-8000 spectrophotometer (ThermoFisher Scientific, Massachusetts, USA).

4.3.2.3 *Screening of the plasmid DNA for successful integration of the gRNA*

To confirm successful integration of the paired gRNAs into the CRISPR-Cas9 plasmids, a double diagnostic digest was used. Upon successful integration, a restriction enzyme cut-site is disrupted resulting in a different digest pattern compared to that of the empty backbone vector (**Figure 4.6A**). Purified plasmid DNA (100 ng) was digested with BbsI and EcoRI restriction enzymes using the FastDigest Green Buffer (ThermoFisher Scientific, Massachusetts, USA). Digested products were visualised by gel electrophoresis (**Figure 4.6B**) and positive plasmids were sent for Sanger sequencing using the LKO.1 forward primer (5'-GACTATCATATGCTTACCGT-3') to confirm insertion of the paired gRNAs in the correct location and orientation within the CRISPR-Cas9 plasmids (**Figure 4.6C**).

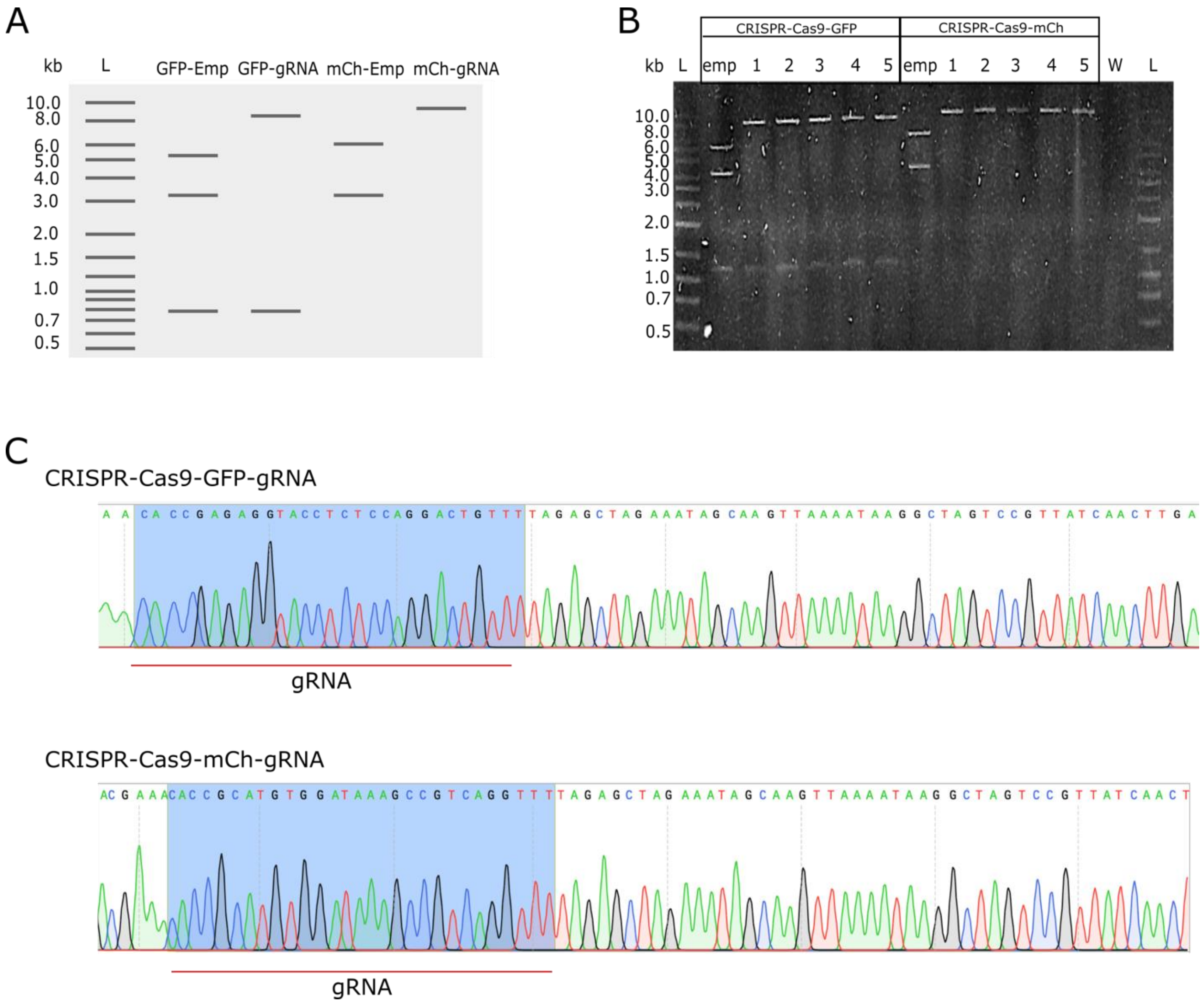


Figure 4.6: Construction of the CRISPR-Cas9 plasmids for targeting the EPO gene. ***A:*** *Virtual double diagnostic digest of the paired CRISPR-Cas9 plasmids (GFP = pSpCas9(BB)-2A-GFP (PX458), mCh = pU6-(BbsI)_CBh-Cas9-T2A-mCherry). Lanes 2 & 4 represent double diagnostic digest of empty plasmid backbones. Lanes 3 & 5 depict the resulting band patterning after digestion of plasmids containing the gRNA. When a gRNA is ligated into the plasmid backbone, a restriction enzyme cut-site is lost and the resulting banding pattern differs. Emp = Empty backbone. L= NEB 2-Log.* ***B:*** *Gel electrophoresis image of the double restriction enzyme diagnostic digest confirming that the gRNA had successfully been inserted into the 10 plasmids that were screened. L=Solis Biodyne 1kb ladder. Emp = Empty plasmid backbone. CRISPR-Cas9-GFP = pSpCas9(BB)-2A-GFP (PX458), CRISPR-Cas9-mCh = pU6-(BbsI)_CBh-Cas9-T2A-mCherry. W = plasmid replaced with water in the digest reaction for use as a negative control.* ***C:*** *Sanger Sequencing of CRISPR-Cas9-GFP-gRNA1 and CRISPR-Cas9-mCh-gRNA4 confirmed successful insertion of each gRNA in the correct orientation and location in the correct plasmid.*

4.3.3 CRISPR-Cas9 gene-editing of the *EPO* gene

4.3.3.1 *Co-transfection of HEK-293 cells with the paired gRNAs within the CRISPR-Cas9 plasmids*

Prior to gene targeting, HEK-293 cells were seeded at a density of 1 million cells in a 10 cm plate and incubated at 5% $C0_2$ and 37 ^{0}C for 24 hours. Cells were transfected with 6 µg of the CRISPR-Cas9-gRNA targeting exon 2 and 6 µg of the CRISPR-Cas9-gRNA targeting exon 4 using lipofectamine transfection reagent following manufacturer's instructions (ThermoFisher Scientific, Massachusetts, USA). Cells were incubated for 24 hours before being visualised under the Leica DMi8 Widefield microscope (Leica, Milton Keynes, UK) microscope to confirm successful transfection.

4.3.3.2 *Isolation of single cells*

48 hours after successful transfection, double positive fluorescent (red and green) single cells were isolated by single-cell picking into 96-well plates using the EVOS FLoid Imaging system (ThermoFisher Scientific, Massachusetts, USA). Single cells were clonally expanded for around 2 weeks and passaged from 96-well plates into 24-well plates and then 6-well plates when at 80-90% confluency.

4.3.3.3 *Genotyping*

To screen for cell-lines with successful deletion of the expected region of the *EPO* gene between the paired gRNAs **(Figure 4.5)**, clonally expanded cells were pelleted and DNA was extracted from half of the pellet using the PureLink Genomic DNA Extraction Kit (Invitrogen, Massachusetts, USA). The remaining half of the pellet was re-plated for continual growth. DNA concentration and purity were measured using the Nanodrop ND-8000 Spectrophotometer (ThermoFisher Scientific, Massachusetts, USA). 100 ng of DNA was subject to genomic PCR using HOT FIREPol DNA polymerase according to manufacturer's protocol (Solis BioDyne, Teaduspargi, Estonia). To identify successful targeting of the *EPO* gene and to distinguish between homozygous or heterozygous knock-outs, specific primers either side of the paired gRNAs were designed; epo-forward (P1): 5'-TCTAGAATGTCCTGCCTGGC-3', epo-reverse (P2): 5’-GGCCCTGTGACATCCTTAGA-3’ **(Figure 4.5).** Resulting PCR products were visualised using gel electrophoresis. PCR amplicons showing the expected banding pattern were purified using ExoSAP-IT PCR Product Cleanup reagent following manufacturer’s instructions (ThermoFisher Scientific, Massachusetts, USA). Samples were subsequently sent for Sanger sequencing using Genewiz (Genewiz, Essex, UK) with the epo-forward (P1) primer to confirm removal of the expected region between the paired gRNAs in the genomic DNA. Successfully targeted clones were further propagated for downstream analysis.

4.3.4 Validation of disruption to the EPO gene

Two homozygous *EPO*$^{-/-}$ knock-out cell-lines were identified through genotyping via PCR and Sanger sequencing. These are denoted KOA and KOB. These knock-out cell-lines were continually passaged before being subjected to downstream analyses to confirm *EPO*$^{-/-}$ knock-out had affected *EPO* mRNA expression and EPO protein levels. Frozen stocks were made for long-term preservation.

4.3.4.1 *Over-expression of EPO*

For use as a positive control when confirming successful *EPO*$^{-/-}$ knock-out through Western blot analysis and qRT-PCR*,* HEK-293 cells were seeded at a density of 1 million cells per 10 cm plate and 24 hours later transfected with 12 µg of an *EPO* overexpression (hEPO) construct (pLV-EF1alpha-EPO-218-PDGFR) using Lipofectamine LTX reagent (ThermoFisher Scientific, Massachusetts, USA) following manufacturer’s protocol. pLV-EF1alpha-EPO-218-PDGFR was a gift from Tao Liu (Addgene plasmid # 139057 ; http://n2t.net/addgene:139057 ; RRID:Addgene_139057) (T. Liu et al., 2017).

4.3.4.2 *qRT-PCR*

Isogenic *EPO*$^{-/-}$ knock-outs, wild-type (WT) HEK-293 cells transfected with hEPO (positive control) and WT HEK-293 cells treated with empty CRISPR-Cas9 plasmids were subjected to qRT-PCR to assess the effects of knock-out on *EPO* mRNA expression levels. Cells were pelleted and RNA was isolated using the Direct-zol^{TM} RNA Miniprep kit following manufacturer’s protocol

(Cambridge Biosciences, Cambridge, UK). 500 ng of RNA was converted to cDNA using PrimeScript^{TM} RT reagent kit (Takara Bio Europe SAS, Saint-Germain-en-Laye, France) following manufacturer's protocol. qRT-PCR was performed on at least three biological replicates using Hot FIREPol EvaGreen^{TM} qPCR Master Mix with ROX (Solis BioDyne, Teaduspargi, Estonia) on the QuantStudio 6 Flex qPCR machine (ThermoFisher Scientific, Massachusetts, USA). Any samples with Ct values greater than two standard deviations (SD) from the mean were removed. Gene expression levels were standardised against the reference gene *GAPDH* mRNA levels using the $2^{-\Delta\Delta CT}$ method (Livak & Schmittgen, 2001). I also checked expression of alternative housekeeping genes (*UBC* and *Pol2ra*) and assessed the most stable gene or combination of genes for use as an endogenous control using RefFinder (https://www.heartcure.com.au/reffinder/) (Fuliang Xie et al., 2012). *GAPDH* showed the most stable expression and was therefore chosen as the reference gene for standardisation **(Figure 4.7).** Differences in gene expression levels between WT and *EPO*$^{-/-}$ cell-lines were investigated for statistical significance by a paired t-test carried out in RStudio version.3.6.1 (RStudio Team, 2018). Primer sequences are listed in **Table 4.2.**

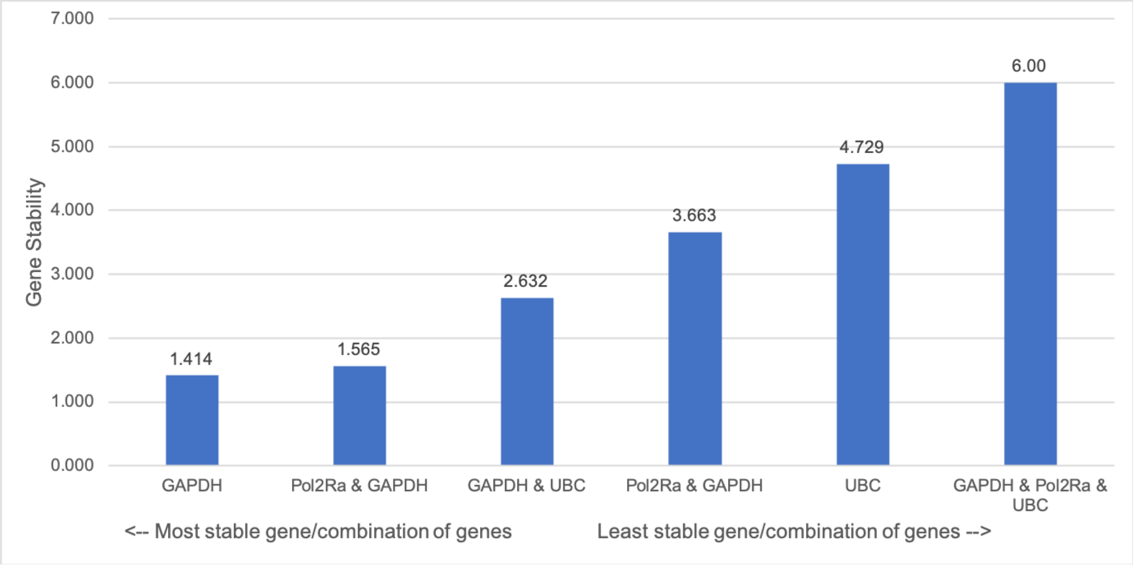


Figure 4.7: The stability of housekeeping genes for use as the reference gene in qRT-PCR analysis. *The stability of the genes was determined using the RefFinder web-based tool (https://www.heartcure.com.au/reffinder/) by inputting the raw Ct values obtained during qRT-PCR for each gene (or the average of these Ct values when combining reference genes). GAPDH was chosen as the housekeeping gene for all analysis due to being the most stable in HEK-293 cells.*

Table 4.2: Primer sequences used for qRT-PCR to validate the down-regulation of EPO mRNA expression in the EPO$^{-/-}$ cell-lines. *Primer sequences were designed using the Primer3 online tool (https://primer3plus.com) and by performing a BLAST search to check for matches elsewhere on the genome. Primer sequences were designed to target exon-exon junctions if possible to increase chances of amplifying only cDNA, not genomic DNA.*

Primer name	**Target Gene**	**Sequence (5’-3’)**
EPO-forward	*EPO*	CCTTCGCAGCCTCACCACT
EPO-reverse	*EPO*	TGTACAGCTTCAGCTTTCCCC
GAPDH-forward	*GAPDH*	TCCTCTGACTTCAACAGCGAC
GAPDH-reverse	*GAPDH*	GCTGTAGCCAAATTCGTTGTCA
UBC-forward	*UBC*	ATTTGGGTCGCGGTTCTTG
UBC-reverse	*UBC*	TGCCTTGACATTCTCGATGGT
Pol2ra-forward	*POL2RA*	CCATCAAGAGAGTCCAGTTCG
Pol2ra-reverse	*POL2RA*	ACCCTCCGTCACAGACATTC

4.3.4.3 *Western Blotting*

EPO protein levels were assessed using western blot analysis. Isogenic *EPO*$^{-/-}$ HEK-293 cells generated by CRISPR-mediated genome-editing, WT controls transfected with empty CRISPR-Cas9 cells and HEK-293 cells transfected with hEPO overexpression vector (T. Liu et al., 2017) were subjected to western blot analysis. Briefly, protein lysate was extracted from HEK-293 cells by washing cells with ice-cold PBS and then centrifuging at 2000 rpm, 4 ^{0}C for 5 minutes. The supernatant was removed and pellets were resuspended in 50 µl lysis Buffer (0.5 µl PMSF (ThermoFisher Scientific, Massachusetts, USA), 0.5 µl Protease inhibitor (Sigma Aldrich, Missouri, USA), 0.5 µl Phosphatase inhibitor 2 (Sigma Aldrich, Missouri, USA), 0.5 µl Phosphatase inhibitor 3 (Sigma Aldrich, Missouri, USA), 48 µl RIPA Buffer). The mixture underwent vortexing for 5 seconds followed by incubation on ice for 5 seconds (repeated 5 times) before incubation on ice for 15 minutes. Samples were again subject to vortexing 5 seconds and incubation on ice for 5 seconds (repeated 5 time) before being centrifuged at 13,200 rpm at 4 ^{0}C for 20 minutes. The supernatant was then transferred to a clean 1.5 mL Eppendorf (ThermoFisher Scientific, Massachusetts, USA). Protein concentration was measured using the PHERAstar (BMG LABTCH, Bucks, UK) by performing a BCA Assay following manufacturer's protocol (ThermoFisher Scientific, Massachusetts, USA). For protein migration, 10% running and stacking gels were prepared. 50 µg of protein was loaded into wells and electrophoresis was run at 90 V for 15 minutes for entry into the stacking gel and then increased to 150 V for 90 minutes for subsequent migration through the running gel. Gels were transferred onto Immobolin PVDF membranes (Merck Life Sciences, Watford, UK) in transfer buffer for 120 minutes at 250 mA. Following 3 x 5 minute washes in TBS-T, membranes were blocked overnight in blocking buffer (5% skimmed-milk power, TBS-T) rocking at 4 ^{0}C. The following day, membranes were equilibrated to room temperature by rocking for an hour before being washed 3 times in TBS-T for 5 minutes. Membranes were incubated in monoclonal mouse anti-EPO antibody (1:1000, 5% skimmed-milk power, TBS-T; MAB2871; R&D systems, Abingdon, UK) for 1.5 hours rocking at room temperature before being washed in TBS-T (3 x 5 minutes) and incubated in goat anti-mouse IgG (H+L) Cross-Absorbed Alexa Fluor^{â} 680 (1:5000, TBS-T, ThermoFisher Scientific, Massachusetts, USA) rocking at room temperature for 1 hour. Membranes were visualised on the LI-COR Odyssey CLx system (LI-

COR Biotechnologies, Nebraska, USA). Next, I probed for the expression of the housekeeping gene, *GAPDH*, by incubating membranes in mouse anti-GAPDH antibody (1:1000, TBS-T; sc-47724; Santa Cruz Biotechnology, Texas, USA) for 1 hour rocking at room temperature, washing in TBS-T (3 x 5 minutes) and then goat anti-mouse IgG (H+L) Cross-Absorbed Alexa Fluorâ 680 (1:5000, TBS-T, ThermoFisher Scientific, Massachusetts, USA) secondary antibody for 1 hour rocking at room temperature. Membranes were visualised on the LI-COR Odyssey CLx system (LI-COR Biotechnologies, Nebraska, USA). Experiments were performed at least three times.

4.3.5 Whole transcriptome NGS

Isogenic *EPO*$^{-/-}$ knock-out and WT *EPO*$^{+/+}$ control cells were pelleted and RNA was isolated using the Direct-zol^{TM} RNA Miniprep kit following the manufacturer's protocol (Cambridge Biosciences, Cambridge, UK). All WT cells were treated with empty CRISPR-Cas9 backbone vectors and underwent the same experimental conditions and treatments as the knock-in cell-lines. The concentration of RNA was measured accurately using the Qubit^{TM} 2.0 Fluorometer (ThermoFisher Scientific, Massachusetts, USA) following manufacturer's protocol by mixing 1 µl RNA with 200 µl Qubit^{TM} working solution (ThermoFisher Scientific, Massachusetts, USA). RNA samples were checked for quality, purity and integrity using an Agilent 2020 TapsStation with RNA ScreenTape (Agilent Technologies, California, USA) following manufacturer's protocol. Four RNA samples per cell-line (WT, KOA & KOB) with high concentration and an RNA integrity number (RIN) > 8 were chosen for RNA-seq analysis (Schroeder et al., 2006) **(Table 4.3).** 1000 ng of RNA was prepared and sent for RNA-seq using the Exeter Sequencing Service. The following steps were performed by the Exeter Sequencing service: library preparation using the TruSeq DNA HT Library Preparation Kit using the 3' poly-A tail primer Oligo(dT) from Illumina (Illumina, California, USA), RNA-seq using the Illumina HiSeq 2500 high-throughput sequencing system (Illumina, California, USA) resulting in 75 bp paired-end sequences. Sequencing data were provided as raw FastQ files.

Table 4.3: The RNA quality and quantity of the 12 samples that were sent off for RNA-seq analysis. *Four WT RNA samples were sequenced alongside four KOA and four KOB samples. The WT cells were treated exactly the same as the EPO$^{-/-}$ cells but had been*

transfected with empty CRISPR-Cas9 plasmids. Each sample from the corresponding cell-line is a biological replicate. RIN = RNA Integrity Number.

Sample	**RNA concentration (ng/µl)**	**RIN**	**Area (28S/18S)**
WT1	960	9.1	1.4
WT2	614	9.6	1.3
WT3	856	9.5	1.8
WT4	624	9.8	1.3
KOA-1	478	9.5	2.1
KOA-2	254	9.8	1.8
KOA-3	846	8.6	2.1
KOA-4	480	9.3	2.0
KOB-1	756	10	2.9
KOB-2	322	9.4	1.8
KOB-3	418	9.9	2.2
KOB-4	308	9.6	2.4

4.3.6 Bioinformatic analysis of RNA-seq data

The raw reads (read 1 and read 2) were downloaded from the Exeter Sequencing Service. Bioinformatic analysis was performed on a Unix-based operating system server or using RStudio version.3.6.1 (RStudio Team, 2018). The workflow developed and followed for RNA-seq analysis is outlined in **Figure 4.8**. All scripts generated for the analysis of the RNA-seq data are avaliable in a repository on Github and can be seen in **Appendix 1: Scripts for RNA-seq analysis**

4.3.6.1 *Quality control of sequence reads*

Quality control (QC) checks were undertaken on the raw reads using MultiQC (Ewels et al., 2016). The main QC metrics considered were the quality per base, GC content per sequence, per sequencing quality score, sequence length distribution, sequence duplication levels and the adapter sequence content **(Figure 4.9).** Adapter sequences (as defined by Illumina), nucleotides with poor quality from the 3' end and reads shorter than 25 bp were removed using CutAdapt version 1.13 **(Figure 4.10, Table 4.4)** (Martin, 2011). MultiQC (Ewels et al., 2016) was again used to check the trimmed reads (**Figure 4.11**).

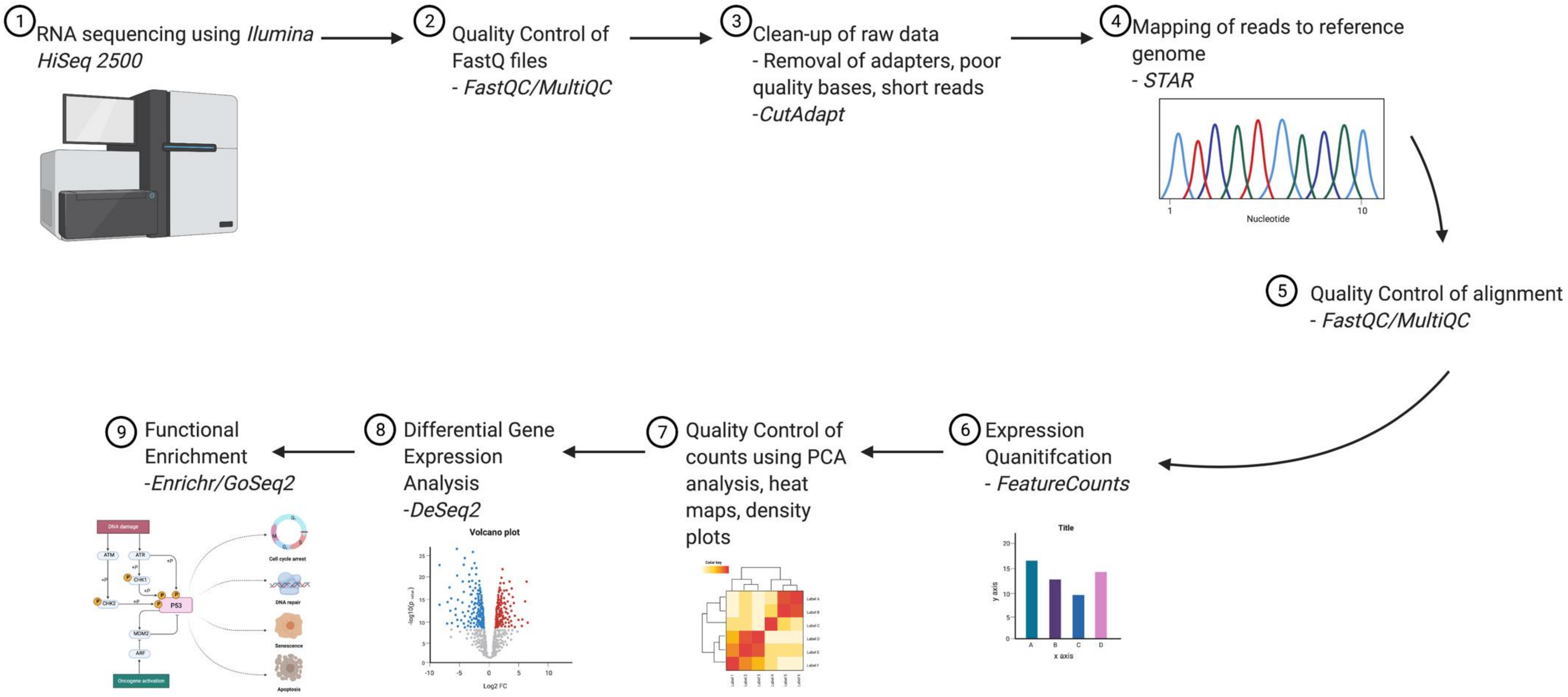
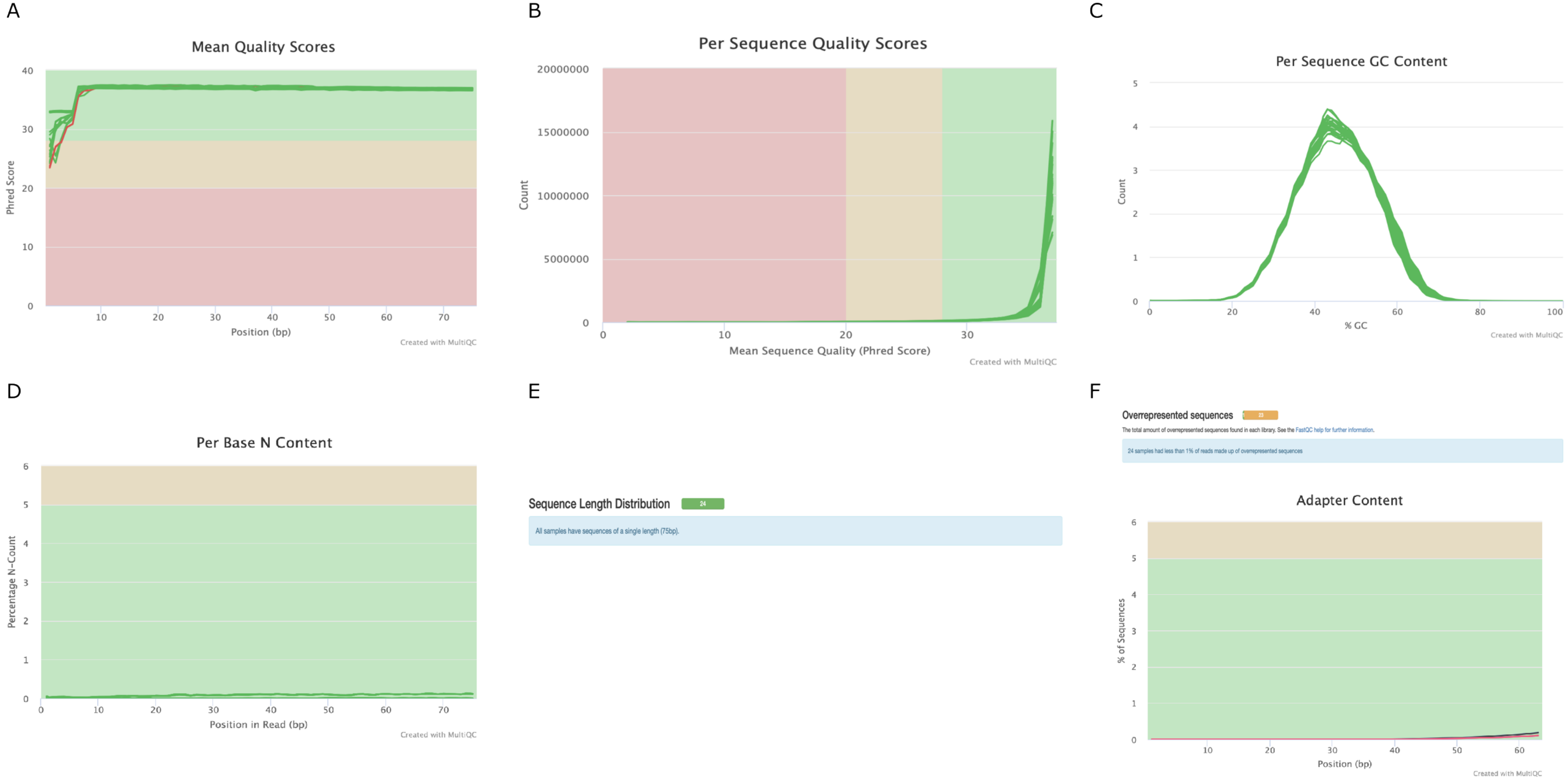

1 RNA sequencing using *Ilumina HiSeq 2500*
2 Quality Control of FastQ files - *FastQC/MultiQC*
3 Clean-up of raw data - Removal of adapters, poor quality bases, short reads -*CutAdapt*
4 Mapping of reads to reference genome - *STAR*
1
10
Nucleotide
5 Quality Control of alignment - *FastQC/MultiQC*
6 Expression Quanitifcation - *FeatureCounts*
Title
y axis
x axis
A
B
C
D
7 Quality Control of counts using PCA analysis, heat maps, density plots
Color key
Label A
Label B
Label C
Label D
Label E
Label F
8 Differential Gene Expression Analysis -*DeSeq2*
Volcano plot
-log10(p value)
Log2 FC
9 Functional Enrichment -*Enrichr/GoSeq2*
DNA damage
ATM
ATR
CHK1
CHK2
P53
MDM2
ARF
Oncogene activation
Cell cycle arrest
DNA repair
Senescence
Apoptosis

Figure 4.8: Schematic representing an overview of the pipeline followed to perform RNA-seq analysis. The different software packages used in each step are shown in italics. Created with BioRender.com.
:



All samples have sequences of a single length (75bp).

The total amount of overrepresented sequences found in each library. See the FastQC help for further information.

24 samples had less than 1% of reads made up of overrepresented sequences

Figure 4.9: Quality Control checks of the raw RNA-seq raw reads. ***A:*** *Plot of the Phred Scores for the 12 trimmed sequencing reads. All Phred scores are above 20 with the lower quality for the first five bases.* ***B:*** *Per sequence quality plot. All sequencing reads had an average quality > 30.* ***C:*** *Plot showing the GC content per sequence.* ***D:*** *The number of bases read as 'N' along each sequencing read.* ***E:*** *The distribution of the sequence lengths across reads.* ***F:*** *Quality control check to see if adapter sequences were present and if any sequences were over-represented. Images were produced using MultiQC (Ewels et al.,2016).*

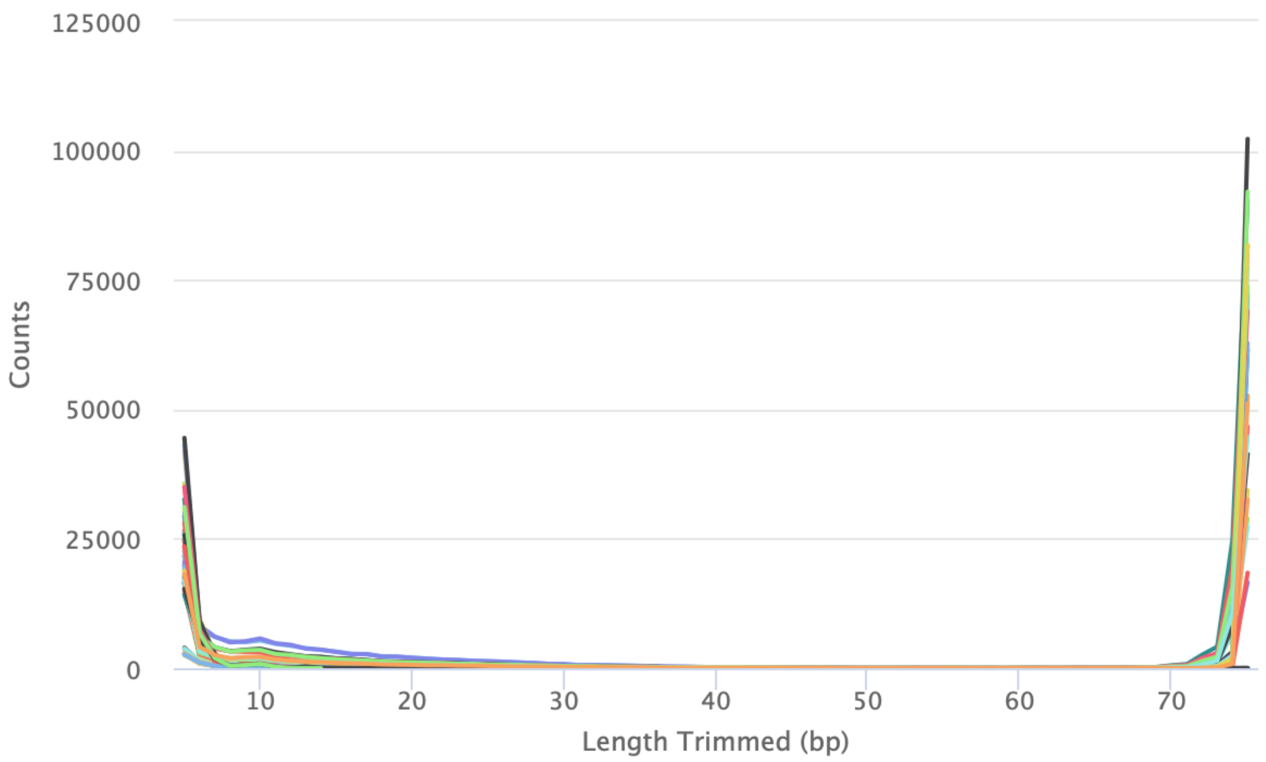
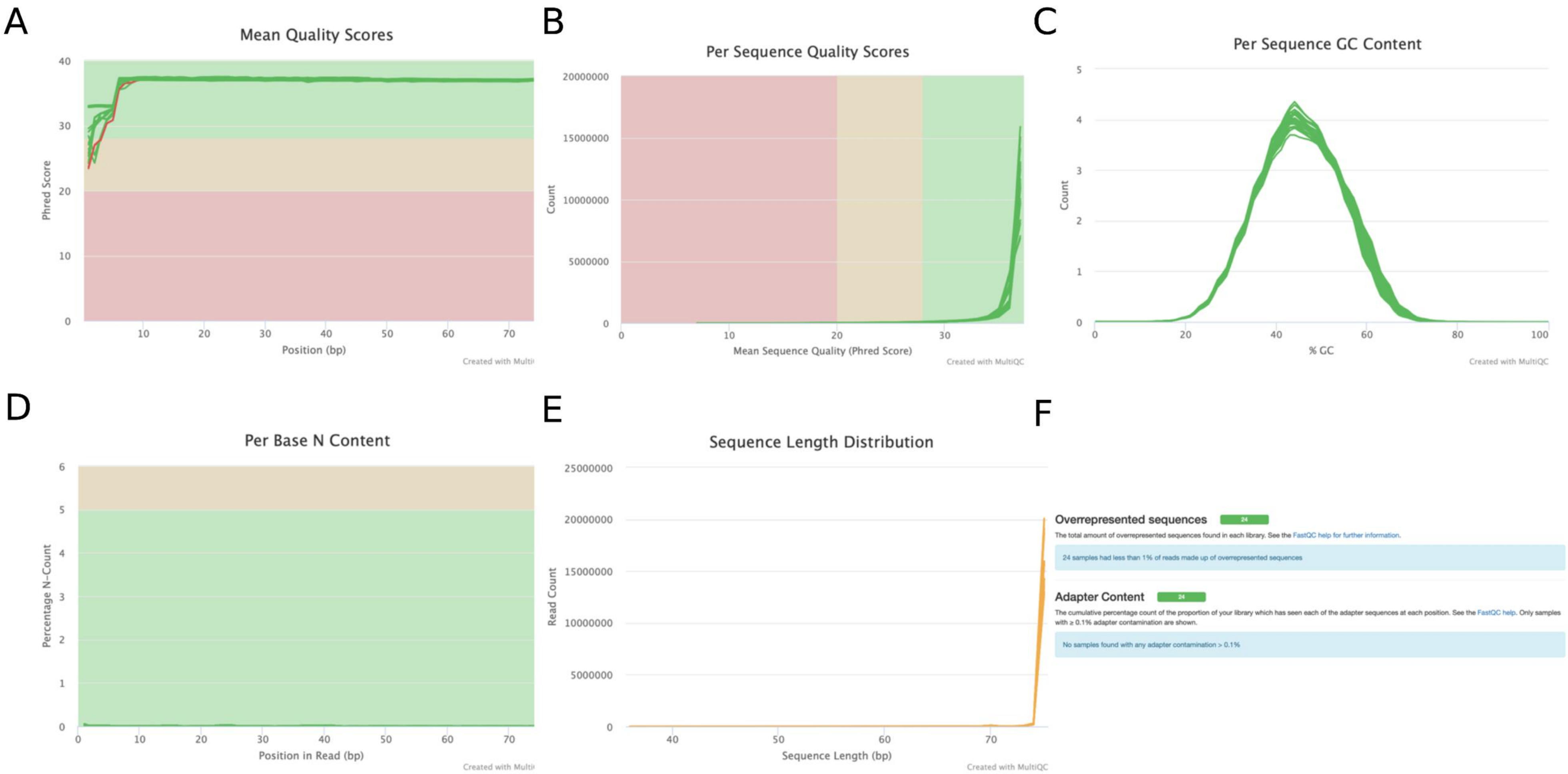


Figure 4.10: Lengths of trimmed sequences and the number of counts in each sample. *The highest number of counts are within the first 10 bp and at the end of the reads due to the inclusion of adapter sequences and poorer quality nucleotides at the 3’ end.*

Table 4.4: General statistics table showing the percentage of reads trimmed for the RNA-seq experiment after carrying out QC checks using MultiQC. *Dups = Duplicate reads. M Seqs = Total sequences (millions)*

Sample Name	% Trimmed	% Dups	% GC	Length	% Failed	M Seqs
WT1_r1		46.2%	45%	75 bp	33%	13.6
WT1_r2	0.9%	45.6%	45%	75 bp	8%	13.6
WT1_trimmed_r1		45.3%	45%	75 bp	33%	13.5
WT1_trimmed_r2		44.5%	45%	75 bp	8%	13.5
WT2_r1		49.1%	45%	75 bp	25%	16.5
WT2_r2	0.7%	46.6%	45%	75 bp	17%	16.5
WT2_trimmed_r1		48.3%	45%	75 bp	25%	16.4
WT2_trimmed_r2		45.8%	45%	75 bp	17%	16.4
WT3_r1		48.9%	45%	75 bp	25%	14.8
WT3_r2	1.0%	45.6%	45%	75 bp	17%	14.8
WT3_trimmed_r1		47.9%	45%	75 bp	25%	14.7
WT3_trimmed_r2		44.6%	45%	75 bp	17%	14.7
WT4_r1		47.9%	44%	75 bp	25%	16.3
WT4_r2	0.9%	46.0%	45%	75 bp	17%	16.3
WT4_trimmed_r1		46.9%	44%	75 bp	25%	16.2
WT4_trimmed_r2		45.0%	45%	75 bp	17%	16.2
KOA-1_r1		52.9%	44%	75 bp	33%	19.5
KOA-1_r2	1.2%	48.1%	44%	75 bp	8%	19.5
KOA-1_trimmed_r1		51.8%	44%	75 bp	33%	19.4
KOA-1_trimmed_r2		47.3%	44%	75 bp	17%	19.4
KOA-2_r1		54.3%	44%	75 bp	33%	19.9
KOA-2_r2	1.3%	48.9%	45%	75 bp	8%	19.9
KOA-2_trimmed_r1		53.1%	44%	75 bp	33%	19.7
KOA-2_trimmed_r2		48.1%	45%	75 bp	8%	19.7
KOA-3_r1		47.8%	45%	75 bp	25%	14.0

KOA-3_r2	0.8%	46.9%	45%	75 bp	8%	14.0
KOA-3_trimmed_r1		47.1%	45%	75 bp	25%	13.9
KOA-3_trimmed_r2		46.3%	45%	75 bp	8%	13.9
KOA-4_r1		47.1%	45%	75 bp	25%	14.5
KOA-4_r2	1.0%	46.8%	45%	75 bp	8%	14.5
KOA-4_trimmed_r1		46.2%	45%	75 bp	25%	14.3
KOA-4_trimmed_r2		46.0%	45%	75 bp	8%	14.3
KOB-1_r1		55.2%	45%	75 bp	33%	20.8
KOB-1_r2	1.0%	48.4%	45%	75 bp	17%	20.8
KOB-1_trimmed_r1		54.2%	45%	75 bp	33%	20.6
KOB-1_trimmed_r2		47.4%	45%	75 bp	17%	20.6
KOB-2_r1		44.1%	45%	75 bp	33%	13.4
KOB-2_r2	1.4%	43.4%	45%	75 bp	8%	13.4
KOB-2_trimmed_r1		43.0%	45%	75 bp	33%	13.2
KOB-2_trimmed_r2		42.4%	45%	75 bp	8%	13.2
KOB-3_r1		49.2%	44%	75 bp	25%	16.3
KOB-3_r2	1.1%	47.1%	45%	75 bp	17%	16.3
KOB-3_trimmed_r1		48.2%	45%	75 bp	25%	16.1
KOB-3_trimmed_r2		46.2%	45%	75 bp	17%	16.1
KOB-4_r1		48.2%	45%	75 bp	25%	14.8
KOB-4_r2	0.9%	47.3%	45%	75 bp	17%	14.8
KOB-4_trimmed_r1		47.3%	45%	75 bp	25%	14.6
KOB-4_trimmed_r2		46.5%	45%	75 bp	17%	14.6



Overrepresented sequences

The total amount of overrepresented sequences found in each library. See the FastQC help for further information.

24 samples had less than 1% of reads made up of overrepresented sequences

Adapter Content

The cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position. See the FastQC help. Only samples with ≥ 0.1% adapter contamination are shown.

No samples found with any adapter contamination > 0.1%

Figure 4.11: Quality Control checks of the RNA-sequencing raw reads after trimming had been performed using CutAdapt to remove poor quality sequences, adapter sequences and reads < 25 bp. ***A:*** *Plot of the Phred Scores for the 12 trimmed sequencing reads. All Phred scores are above 20 with the lower quality for the first five bases.* ***B:*** *Per sequence quality plot. All sequencing reads had an average quality > 30.* ***C:*** *Plot showing the GC content per sequence.* ***D:*** *The number of bases read as 'N' along each sequencing read.* ***E:*** *The distribution of the sequence lengths across reads.* ***F:*** *Quality control check to see if all adapter sequences had been removed and if any sequences were over-represented. Images were produced using MultiQC (Ewels et al.,2016).*

4.3.6.2 *Alignment of reads to the reference genome*

Reads were aligned to the Homo sapiens GRCh38/hg38 reference genome using STAR version 2.7.1 **(Figure 4.12)** (Dobin et al., 2013)**.** The reference genome included the nucleotide sequence, haplotypes, chromosomes, scaffolds and patches. First, a genome index file was generated using the ‘genomeGenerate’ function. To generate the index files, the reference genome sequence was provided alongside the corresponding annotation file with information on gene names and transcripts from Ensembl (Howe et al., 2021) (http://ftp.ensembl.org/pub/release-105/fasta/homo_sapiens/dna/). The --sjdbOverhang option was set at 74 as this length should be equal to the read length minus 1. This option specifies the length of the genomic sequence around the annotated junction used in constructing the splice junction database. Second, mapping of the reads to the reference genome was run using the ‘alignReads’ function. Default options were used and the output produced were binary alignment files in BAM format which were sorted by coordinate.

4.3.6.3 *Gene Quantification*

To calculate the changes in gene expression as a result of *EPO*$^{-/-}$ knock-out, the number of reads mapping to a gene were counted. Gene quantification was performed using the featureCounts subread package (Liao et al., 2013, 2014) based on Ensembl GRCh38/hg38 annotation release version 2.0.0 **(Figure 4.13)** (Howe et al., 2021).

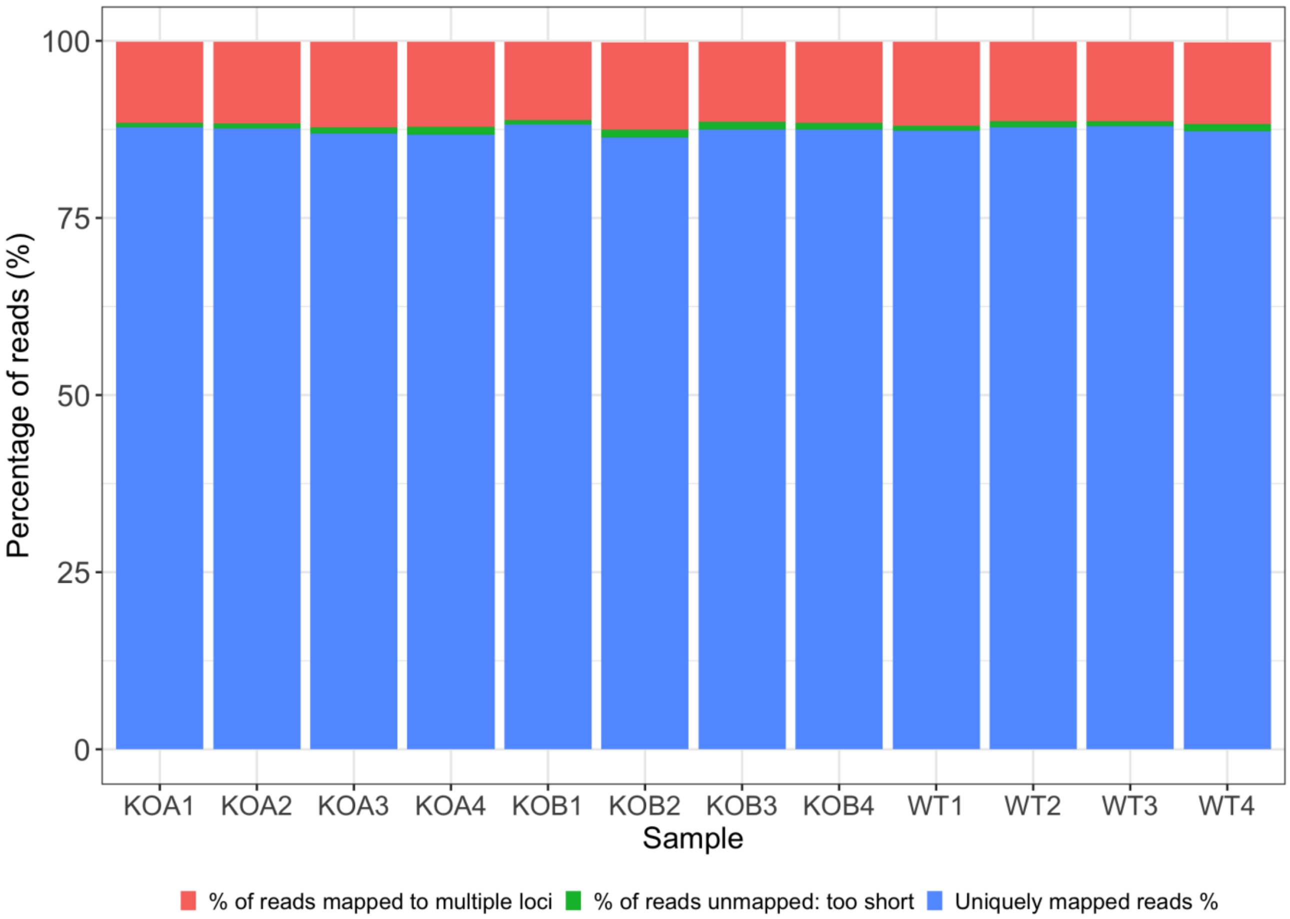


Figure 4.12: The percentage of reads mapping to the reference genome*. The percentage of reads (y-axis) per sample (x-axis) aligning to either unique positions (blue), multiple loci (red) or no loci due to being too short (green) on the GRCh38/hg38 reference genome. Alignment of sequencing reads to the reference genome (Homo sapiens GRCh38/hg38) was performed using STAR version 2.7.1 (Dobin et al., 2013). Plot produced in RStudio version.3.6.1.*

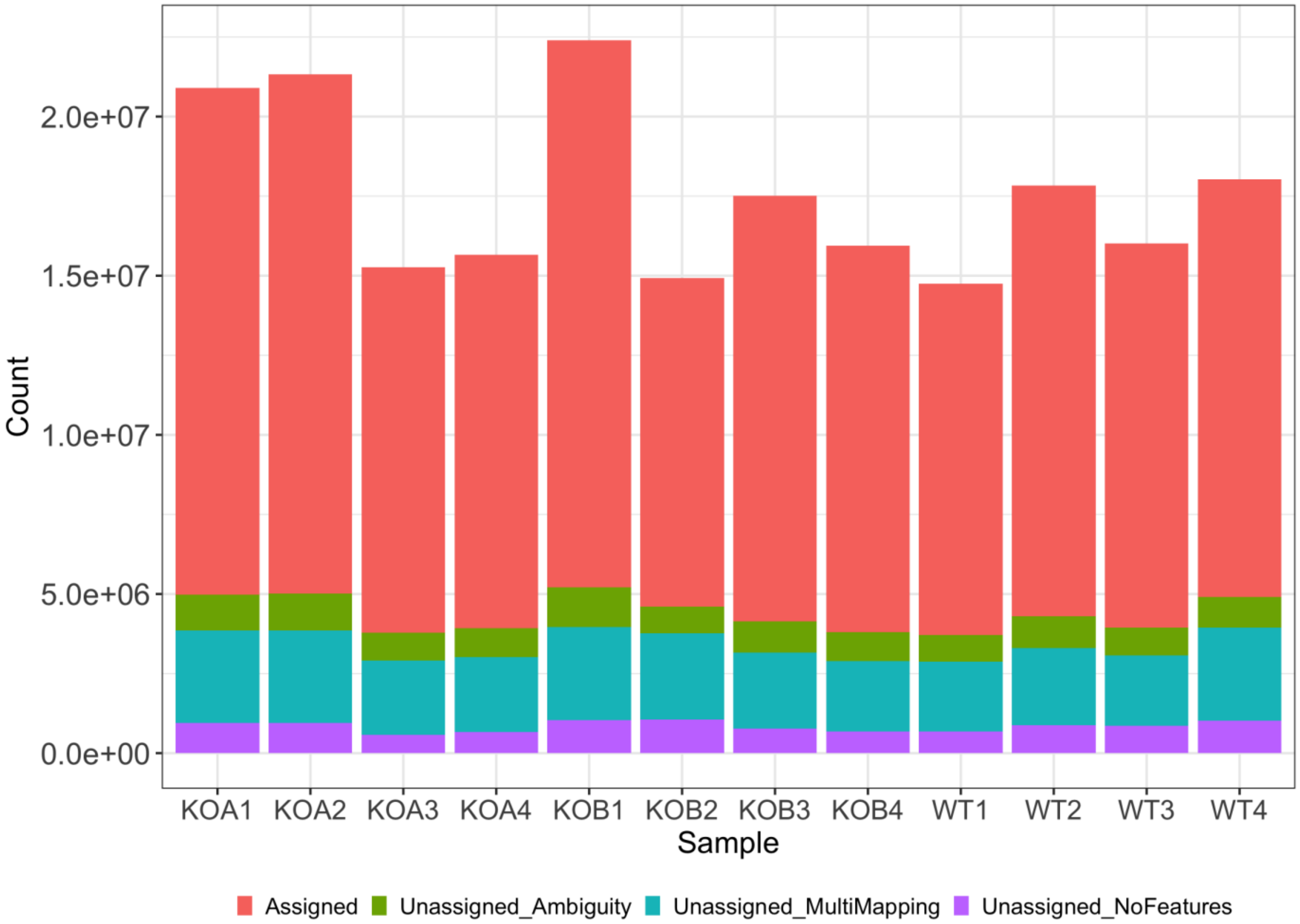


Figure 4.13: The number of reads assigned to genomic features. *The number of reads (y-axis) per sample (x-axis) being assigned to genomic features (exons). Gene quantification was performed using the featureCounts subread package (Liao et al., 2013, 2014) based on the Ensembl GRCh38/hg38 annotation release version 2.0.0. Red: reads that were successfully assigned to a genomic feature. Green: reads unassigned to a genomic feature due to ambiguity. Turquoise: reads unassigned to a genomic feature due to multi-mapping. Purple: reads unassigned to a genomic feature due to not overlapping any genomic feature. Plot produced in RStudio version.3.6.1.*

4.3.6.4 *Raw count normalisation and data distribution analysis*

All further RNA-seq analysis was performed in RStudio version.3.6.1 unless otherwise stated (RStudio Team, 2018). Transcripts whose mean count across all samples were less than 10 were removed. Counts were normalised using the median of ratios method implemented in the DeSeq2 package (Love et al., 2014). The variance in the data was assessed by generating a dispersion plot and a mean-SD plot to check for heteroskedasticity using the *‘plotDispEsts’* and *‘meanSdPlot’* functions in DeSeq2 respectively (**Figure 4.14A)** (Love et al., 2014)**.** Some variability in RNA-seq data is expected, but due to the clear hump to the left-hand-side of the mean-sd plot (**Figure 4.14A**) violating the assumption of homoscedasticity, data was subsequently transformed to stabilise the variance. Counts were transformed using the two methods offered by DeSeq2 (Love et al., 2014); variance stabilising transformation (VST, ‘*vst’* function) (Anders & Huber, 2010) or the regularised-logarithm transformation (rlog, *‘rlogTransformation’* function) (Love et al., 2014). The transformed data showed homoscedasticity highlighted by the flatter trend in the mean-sd plots (**Figure 4.14B-C**). I used the rlog transformation for all downstream analysis including clustering and data visualisation. The distribution of the transformed normalised counts was visualised via a box-and-whisker plot alongside violin plots using the ‘*boxplot’* and ‘*ggplot2’* functions respectively. Dissimilarities between samples was checked using a dendrogram which was created using Euclidean distance implemented by the *‘dist’* function and Ward’s linkage implemented by the *‘hclust’* function

4.3.6.5 *Principal Component Analysis*

Principal component analysis (PCA) was performed using the R ‘*prcomp’* function to check for similarity between samples. Bi-plots were generated to compare the top five eigenvectors (principal components [PCs]) on a pairwise basis. PCA was used to identify genes that best segregate the knockout samples from the wild-type controls. Gene-to-eigenvector eigenvalues were derived to identify the genes responsible for the variation along the different PCs.

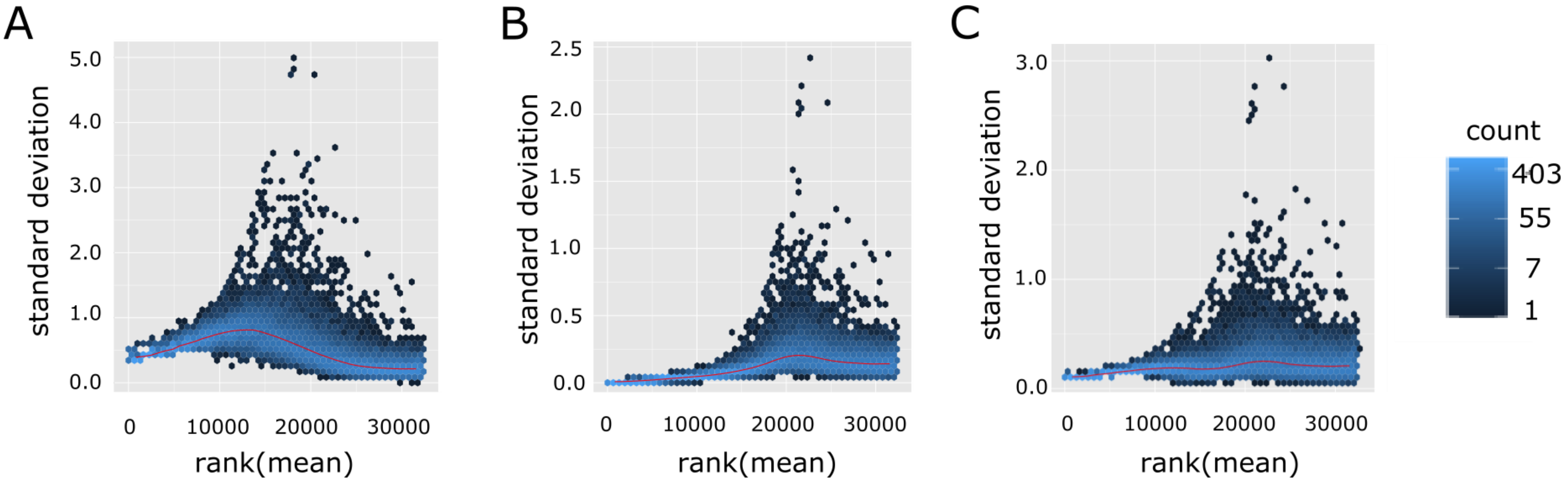


Figure 4.14: Mean-SD plots after performing normalisation and transformation on the raw counts to check for heteroskedasticity. ***A:*** *Mean-SD plot for the normalised raw count data transformed on the $\log_2$ scale.* ***B:*** *Mean-SD plot for the normalised count data having undergone rlog transformation.* ***C:*** *Mean-SD plot of the normalised counts after variance stabilising transformation.*

4.3.6.6 *Differential Gene Expression analysis*

To identify genotype-specific gene expression changes, I performed differential gene expression analysis using DeSeq2 (Love et al., 2014). DeSeq2 (Love et al., 2014) uses a negative binomial model to fit the observed read counts and estimate the difference in expression between the knock-outs and controls. The $\log_2$ fold-change was calculated using **Equation 4.1.**

$$Log_2\, fold - change = \frac{Mean\ normalised\ counts\ for\ knock - out\ samples}{Mean\ normalised\ counts\ for\ control\ samples}$$

Equation 4.1: Calculation of the $\log_2$fold-change.

P-values were calculated using the Wald test and a Benjamini-Hochberg correction was applied to account for multiple testing. Statistically significant differentially expressed genes (DEGs) were determined by an adjusted p-value (*P*-adj) ≤ 0.05. I determined strong differential expression when genes were regulated by at least 2-fold. As RNA-seq was performed on two *EPO*$^{-/-}$ knock-out cell-lines (KOA & KOB), to obtain the most accurate list of DEGs most likely to be differentially expressed due to the effect of *EPO*$^{-/-}$, differential expression analysis was performed comparing *EPO*$^{+/+}$ to each *EPO*$^{-/-}$ knock-out respectively (i.e. WT vs KOA and WT vs KOB). MA plots and volcano plots were generated to check for different gene expression. The MA plot was generated by plotting the $\log_2$ fold-change against the natural log of the mean of the normalised counts + 1 and the volcano plot was generated using '*ggplot'* comparing the -$\log_{10}$(*P*-adj values) against the $\log_2$ fold-change. DEGs (*P*-adj ≤ 0.05) overlapping in both comparisons were identified using the '*venn.diagram'* function. The mean PC values were calculated from the original PCA for the overlapping DEGs.

4.3.6.7 *Supervised Clustering*

Supervised clustering was performed by filtering transcripts from the differential expression analysis at *P*-adj ≤ 0.05 and absolute $\log_2$ fold-change ≥ 2. Regularised log counts for the transcripts were converted to the z-scale and then clustered using the 1-Pearson correlation distance and Ward's linkage using the '*heatmap'* function of the ComplexHeatMap package (Gu et al., 2016). Transcripts were clustered whilst samples were fixed to maintain the order.

4.3.6.8 *Gene set enrichment*

Gene set enrichment analysis was performed on the list of overlapping DEGs using Enrichr (E. Y. Chen et al., 2013; Kuleshov et al., 2016; Z. Xie et al., 2021). Enrichment was performed with default parameters. Bar charts displaying results were obtained from Enrichr. The GoSeq2 package (Young et al., 2010) implemented in RStudio version.3.6.1 (RStudio Team, 2018) was also used to compare the results obtained by Enrichr. A weight for each gene was calculated using the '*probability weighted'* function to account for length bias. Over- and under-expressed GO categories among the DEGs were calculated using the Wallenius equation. A Benjamini and Hochberg adjustment was used to correct for multiple testing. Bar plots representing the results obtained from GoSeq2 were generated using '*ggbarplot'* function.

4.3.7 Validation of differential gene expression

Top DEGs were subsequently subjected to qRT-PCR, as described above in **4.3.4.2** to validate differential expression. As several candidate genes appeared to be involved in the Notch signalling pathway, additional genes involved in the canonical Notch signalling pathway were also selected for analysis by qRT-PCR. Primer sequences for all genes are listed in **Table 4.5.** Reactions were carried out on at least three biological replicates. Any samples with Ct values greater than two SD from the mean were removed. Gene expression levels were standardised against the reference gene *GAPDH* mRNA levels using the $2^{-\Delta\Delta CT}$ method (Livak & Schmittgen, 2001). Differences in gene expression levels between WT *EPO*$^{+/+}$ and *EPO*$^{-/-}$ knock-out cell-lines were investigated for statistical significance by a paired t-test carried out in RStudio version.3.6.1 (RStudio Team, 2018).

Table 4.5: Primer sequences for use in qRT-PCR to validate expression of candidate genes in the EPO$^{-/-}$ knock-out compared to EPO$^{+/+}$.

Primer name	**Target Gene**	**Sequence (5’-3’)**
EPO-forward	*EPO*	CCTTCGCAGCCTCACCACT
EPO-reverse	*EPO*	TGTACAGCTTCAGCTTTCCCC
GAPDH-forward	*GAPDH*	TCCTCTGACTTCAACAGCGAC
GAPDH-reverse	*GAPDH*	GCTGTAGCCAAATTCGTTGTCA
MLH1-forward	*MLH1*	*GAAGTTATCCAGCGGCCAG*
MLH1-reverse	*MLH1*	*TGAATCAACTTCAGGCCTCC*
HEY1-forward	*HEY1*	*TGCGGACGAGAATGGAAACT*
HEY1-reverse	*HEY1*	*TCGTCGGCGCTTCTCAATTA*
PARP9-forward	*PARP9*	*GCCTCATACATCTCTTCCACGT*
PARP9-reverse	*PARP9*	*GCCTCATACATCTCTTCCACGT*
DTX3L-forward	*DTX3L*	*GCCTCATACATCTCTTCCACGT*
DTX3L-reverse	*DTX3L*	*ACTCTCTCCTTAGCTGCCCT*
Notch1-forward	*NOTCH1*	*CGCACAAGGTGTCTTCCAG*
Notch1-reverse	*NOTCH1*	*AGGATCAGTGGCGTCGTG*
Hes1-forward	*HES1*	*AAGAAAGATAGCTCGCGGCA*
Hes1-reverse	*HES1*	*TACTTCCCCAGCACACTTGG*
Hey2-forward	*HEY2*	*CTTCCACGGAGCTCAGGTAC*
Hey2-reverse	*HEY2*	*CTTCCACGGAGCTCAGGTAC*
LRATD2-forward	*LRATD2*	*GCCGAGCCTACACCTTCAAA*
LRATD2-reverse	*LRATD2*	*CGAAACCAACTCCAGGGTCA*
LCP1-forward	*LCP1*	*GGTGTTAACCCTCGAGTCAA*
LCP1-reverse	*LCP1*	*AGTTTGGGGTATGGCGGTTT*
ZNF331-forward	*ZNF331*	*GGTCTCACTGGATTTGGAGT*
ZNF331-reverse	*ZNF331*	*AGCGTACCTTCACATATCCAG*
IFITM2-forward	*IFITM2*	*TTCATGAACACCTGCTGCCT*
IFITM2-reverse	*IFITM2*	*AGATGTTCAGGCACTTGGCG*

4.4 Results

4.4.1 Establishment of the *EPO*$^{-/-}$ knock-out human cell-line model

To determine the downstream causal genes and signalling cascades of EPO, I generated an *EPO*$^{-/-}$ knock-out cell-line in HEK-293 cells using the CRISPR-Cas9 gene-editing technique with paired gRNAs. The paired gRNAs were designed to target conserved sequences on protein-coding exons of the *EPO* gene to achieve effective gene disruption **(**Figure 4.5**).** The gRNAs target the Cas9 endonuclease to the desired locations on the genomic DNA to introduce two DSBs **(**Figure 4.5**).** These DSBs are repaired via NHEJ resulting in the hypothetical removal of 645 bp from the *EPO* gene sequence **(**Figure 4.5**).** Fluorescent imaging confirmed successful co-transfection of the two CRISPR-Cas9-gRNA plasmids into HEK-293 cells **(Figure 4.15A**). Single cells expressing both red and green fluorescence were isolated and clonally expanded before being screened for disruption of the *EPO* gene using genomic PCR. Two homozygous *EPO*$^{-/-}$ cell-lines were identified as potential knock-outs (KOA & KOB) (**Figure 4.15B**). Sanger sequencing of the two confirmed deletion of the region between the paired gRNAs; KOA had an excision of 1,188 bp whilst KOB had an excision of 1,137 bp (**Figure 4.15C**).

4.4.2 *EPO*$^{-/-}$ knock-out resulted in reduced *EPO* mRNA expression levels

The two knock-out cell-lines were subjected to qRT-PCR to confirm that gene disruption had resulted in a reduction in *EPO* mRNA expression levels. Both cell-lines had significantly reduced mRNA expression levels ($P < 0.01$) compared to WT control and HEK-293 cells treated with an *EPO* over-expression vector confirming that gene disruption had altered *EPO* mRNA expression levels (**Figure 4.16A)**.

4.4.3 *EPO*$^{-/-}$ knock-out resulted in reduced EPO protein levels

EPO protein levels were analysed by western blotting and revealed that gene disruption had also reduced EPO protein expression in both knock-out cell-lines **(Figure 4.16B).** For both KOA and KOB, no EPO protein was detected whilst GAPDH protein was detected in all samples (**Figure 4.16B**). These findings confirmed that the gene disruption caused by CRISPR-Cas9 gene-editing had altered the genomic sequence and led to a down-regulation of EPO expression at the mRNA and protein level validating successful disruption to the *EPO* gene.

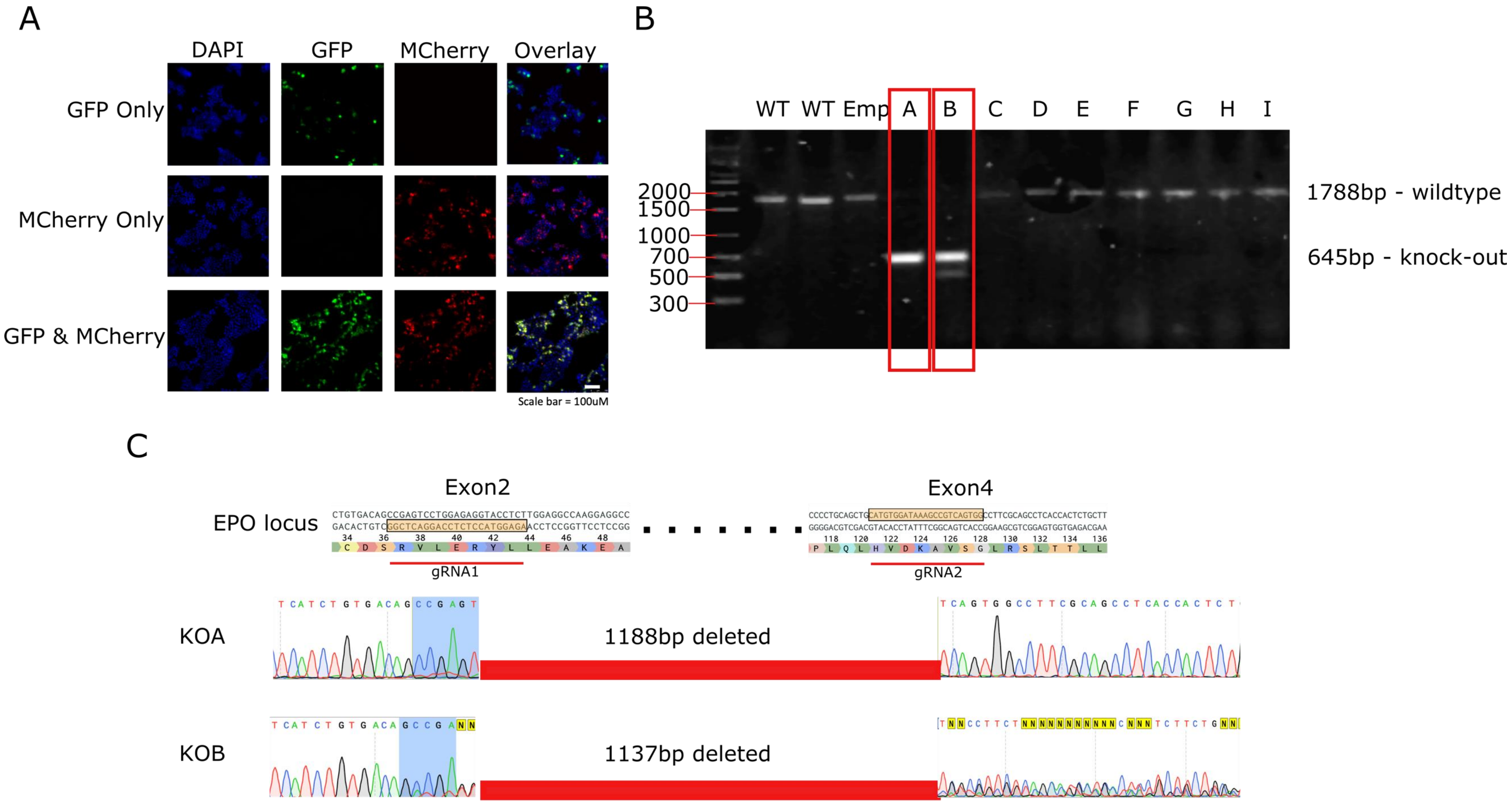

A
DAPI
GFP
MCherry
Overlay
GFP Only
MCherry Only
GFP & MCherry
Scale bar = 100uM
B
WT
WT
Emp
A
B
C
D
E
F
G
H
I
2000
1500
1000
700
500
300
1788bp - wildtype
645bp - knock-out
C
Exon2
Exon4
EPO locus
CTGTGACAGCCGAGTCCTGGAGAGGTACCTCTTGGAGGCCAAGGAGGCC
GACACTGTCGGCTCAGGACCTCTCCATGGAGAACCTCCGGTTCCTCCGG
34 36 38 40 42 44 46 48
C D S R V L E R Y L L E A K E A
gRNA1
CCCCTGCAGCTGCATGTGGATAAAGCCGTCAGTGGCCTTCGCAGCCTCACCACTCTGCTT
GGGGACGTCGACGTACACCTATTTCGGCAGTCACCGGAAGCGTCGGAGTGGTGAGACGAA
118 120 122 124 126 128 130 132 134 136
P L Q L H V D K A V S G L R S L T T L L
gRNA2
KOA
TCATCTGTGACAGCCGAGT
1188bp deleted
TCAGTGGCCTTCGCAGCCTCACCACTCT
KOB
TCATCTGTGACAGCCGANN
1137bp deleted
TNNCCTTCTNNNNNNNNNNNCNNNTCTTCTGNN

Figure 4.15: Establishment of two EPO$^{-/-}$ knock-out cell-lines.

A: *Fluorescent Imaging of HEK-293 cells after transfection with CRISPR-Cas9-gRNA plasmids confirmed successful transfection of both plasmids. Double positive single cells were isolated and clonally expanded.* ***B:*** *After clonal expansion, single cells were genotyped by PCR using primers either side of the gRNA. Cell-lines A & B appeared to be knock-outs due to the presence of the band at the lower amplicon size of 645 bp. WT = wild-type HEK-293 cells used as a positive control. Emp = wild-type HEK-293 cell treated with empty CRISPR-Cas9 plasmids used as a positive control.* ***C:*** *Sanger sequencing of cell-lines A and B confirmed that the expected region of the EPO gene between the paired gRNAs had successfully been removed from the genomic DNA. The blue highlighted region is the 3' end of gRNA where the PAM site is located.*

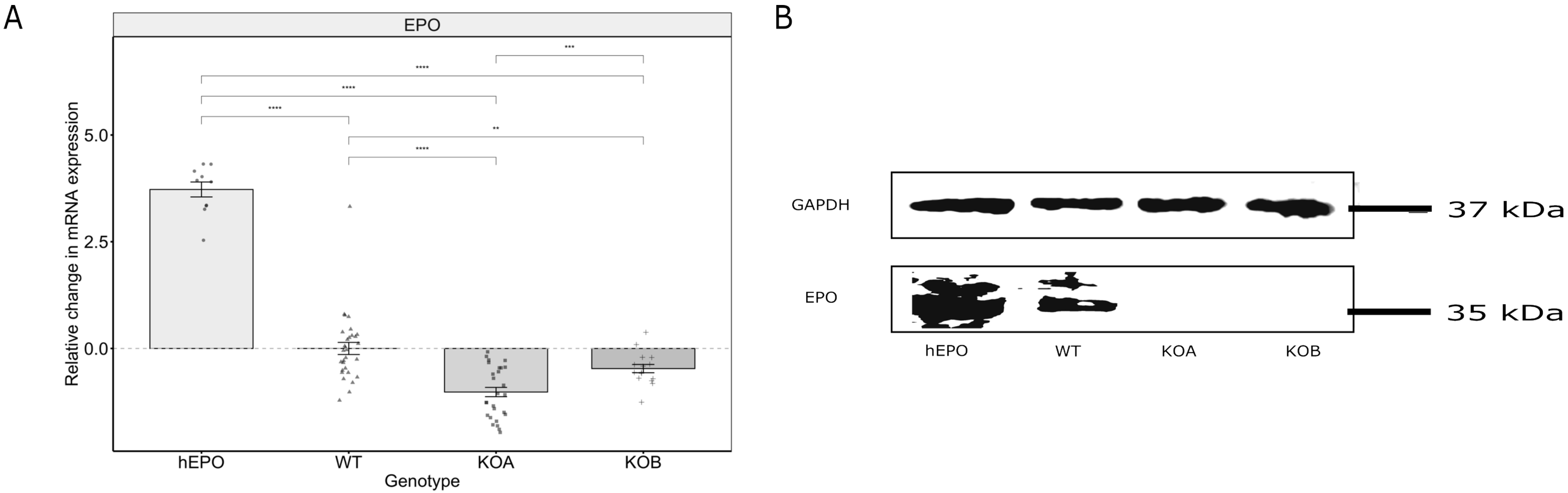


Figure 4.16: EPO mRNA expression and protein expression are disrupted in both EPO$^{-/-}$ knock-out cell-lines compared to WT controls. ***A:*** *EPO mRNA expression is significantly reduced in the knock-out cell-lines (KOA & KOB) compared to wild-type (WT) controls. HEK-293 cells treated with an over-expression EPO construct were used as a positive control (hEPO). Data is shown as mean ± SEM. Paired t-test was performed to test for levels of significance. *P≤0.05, **P≤0.01, ***P≤0.001, ****P≤0.0001.* ***B:*** *EPO expression was reduced at the protein level in the knock-outs (KOA & KOB) compared to wild-type (WT) controls and the HEK-293 cells treated with an EPO over-expression construct (hEPO). GAPDH was used as a positive control and shows consistent pattern of expression across all samples. EPO ~ 34 kDa. GAPDH ~ 37 kDa.*

4.4.4 Distribution and variability checks revealed clear transcriptional differences between controls and knock-outs

A normalised dataset with no sample outliers was shown and consistent gene expression across samples was seen with a peak frequency of gene expression at 17 transcripts **(**Figure 4.17**A)**. Investigation into the variability of the data through the generation of a dispersion plot showed the expected distribution indicating the data was a good fit for the DeSeq2 model **(**Figure 4.17**B**). The data generally scattered around the curve which represents the expected dispersion value for genes of a given expression strength. Genes with a high level of expression represented by a higher mean value showed decreased levels of dispersion compared to those genes with a lower level of expression **(**Figure 4.17**B**). From unsupervised hierarchal clustering, it was evident that the knock-outs exhibited large transcriptional differences from the control samples due to the segregation pattern **(**Figure 4.18**)**. Further branching segregated the two knock-out cell-lines from each other indicating cell-line specific transcriptional differences. Biological replicates showed the highest degree of correlation within cell-lines as expected. The top 50 most highly expressed genes were plotted on a heat map across samples (**Figure 4.19).** The top genes with highest expression were well-known ubiquitously expressed genes including *EEF1A1, GAPDH, TUBB,* mitochondrial genes and ribosomal protein genes as expected (**Figure 4.19)**.

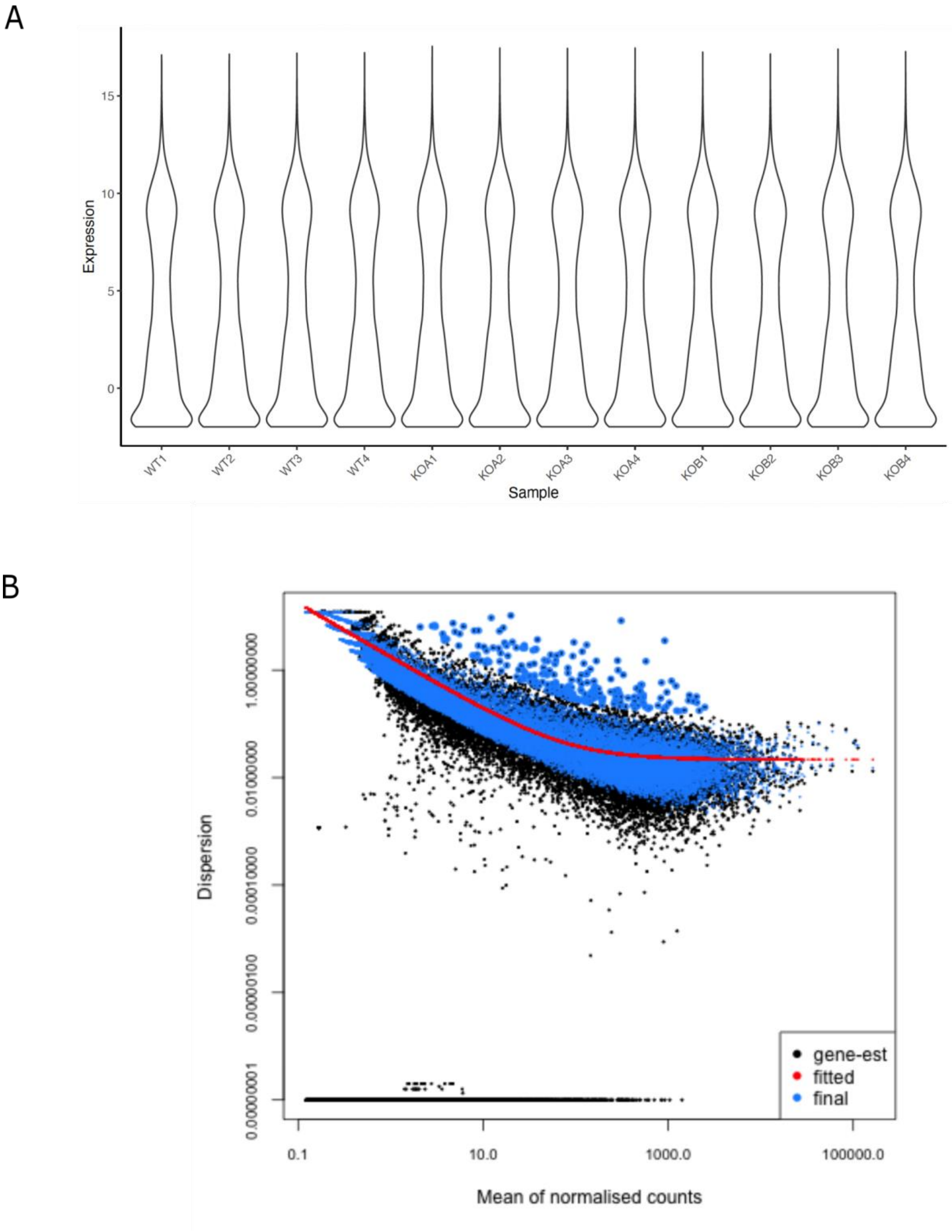


Figure 4.17: Assessing the distribution and variability of the read counts.
A: *Violin plot showing the distribution and variability of gene expression between wild-types and knock-outs. Each violin represents a histogram and box-and-whisker plot. The y-axis represents the regularised log counts of transcripts. The x-axis represents each sample. There were no outliers present in the data after performing normalisation on the raw read counts.* ***B:*** *Dispersion plot. Black dots represent the maximum-likelihood estimate of transcript dispersion from normalised counts; blue dots represent the shrunk dispersion estimates and the red line represents the fitted model generated using logistic regression. There was little evidence of heteroskedasticity in the data as assessed by plotting a dispersion plot on the mean of normalised counts.*

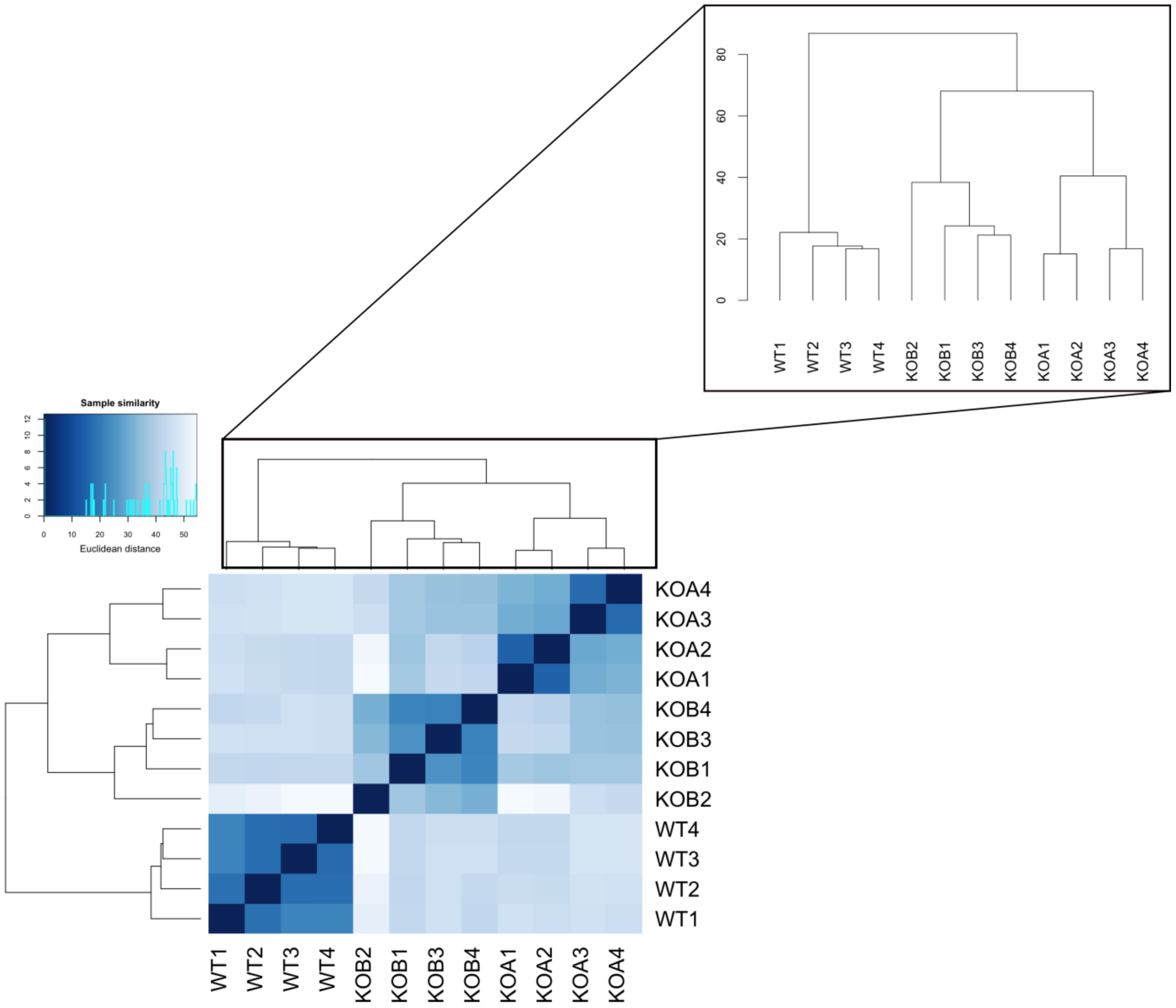


***Figure 4.18: Distance between samples measured by unsupervised hierarchical clustering*.** *Clustering was performed with Euclidean distance and Ward's linkage for all WT and EPO*$^{-/-}$ *knock-out samples. The darker the shade of blue or the shorter the vertical line on the dendrogram, the more similar the gene expression between samples.*

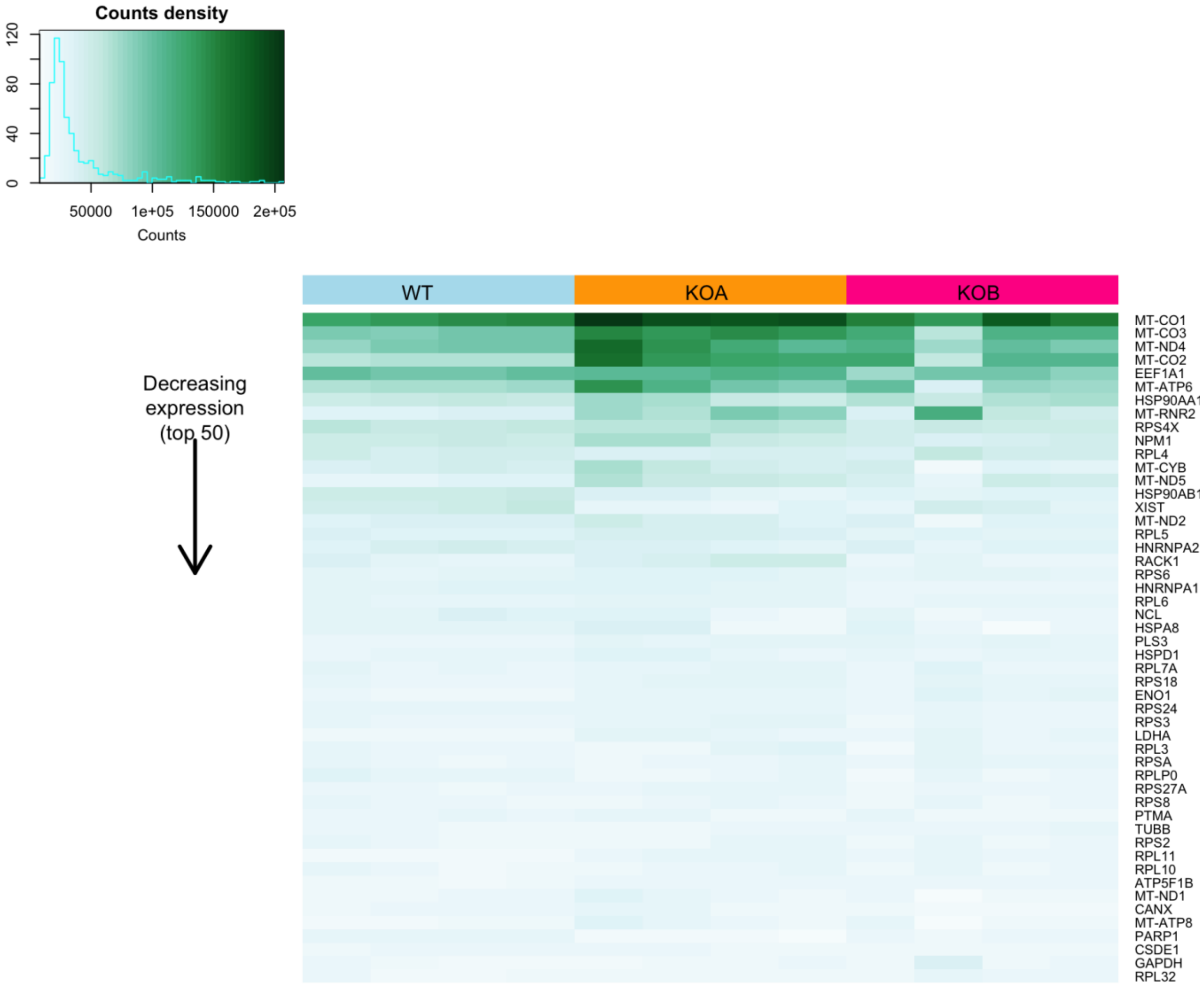


Figure 4.19: Heat map for the top 50 most highly expressed genes across all samples sent for RNA-seq analysis. *The heat map was generated comparing normalised counts of each gene across each sample.*

4.4.5 PCA analysis revealed clear segregation between knock-outs and controls

PCA analysis revealed large transcriptional differences between *EPO*$^{-/-}$ knock-out samples and WT samples with 43% of the variation in the dataset being explained by the differences in the knock-outs compared to controls. PC1 entirely segregated *EPO*$^{+/+}$ WT from *EPO*$^{-/-}$ knock-out and indicated that a reasonable proportion of the transcriptome is different between genotypes **(Figure 4.20**). The top 500 genes responsible for the variation along this axis were checked by ordering the absolute eigenvalues for PCA. The top 20 genes identified included *MAGE-A* genes, *GABRA3* and *MLH1* **(**Table 4.6**).** The differential expression of these genes was evaluated across samples. A strong and significant differential expression was seen between knock-outs and control samples for each of these genes ($\log_2$fold-change ≥ |2| & *P*-adj ≤ 0.05) **(**Table 4.6**).**

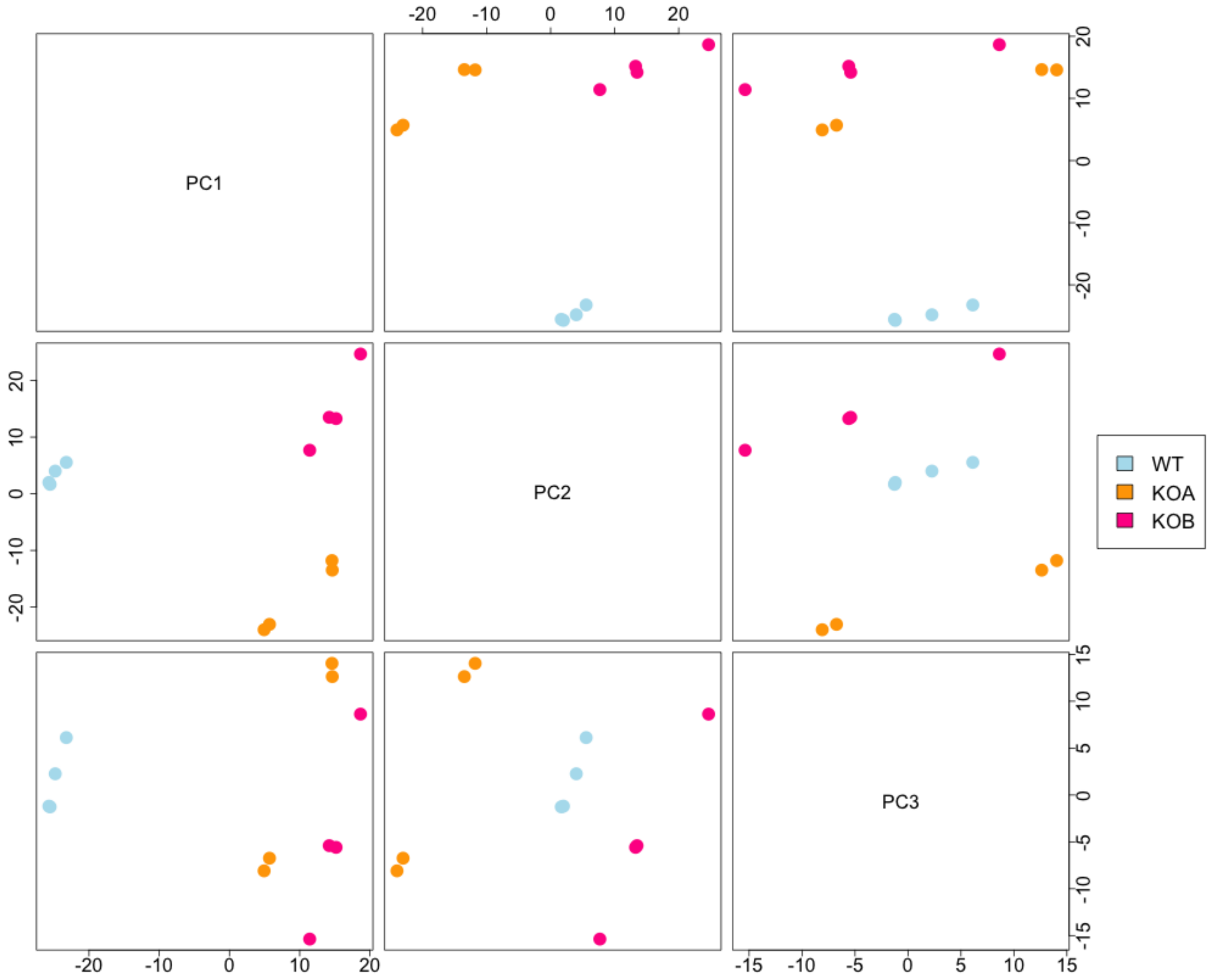


Figure 4.20: Pairwise PCA bi-plots to evaluate the clustering of samples.

Analysis of the expression variation patterns using PCA revealed that a substantial portion of the transcriptome differs between EPO$^{+/+}$ WT cell-lines (blue dots) and EPO$^{-/-}$ (orange or pink dots) knock-out cell-lines. When observing PC1, a high variation can be observed (43%) with a clear segregation between wild-type and knock-out cell-lines.

Table 4.6: The top 20 genes responsible for the segregation between wild-type and knock-outs based on PC1 values. *Top 20 genes and respective eigenvalues derived from PCA analysis and the differential expression of these top 20 genes. Logarithmic fold-changes indicate differential expression in knock-outs.*

Geneid	**GeneSymbol**	**Chromosome**	**Class**	**PC1**	**Log$_2$ fold-change**	**P-adjusted**
ENSG00000198681	MAGEA1	X:153179284-153183880	protein_coding	0.115	-12.16	1.13E-59
ENSG00000221867	MAGEA3	X:152698767-152702347	protein_coding	0.109	-11.81	8.54E-56
ENSG00000197172	MAGEA6	X:152766136-152769747	protein_coding	0.107	-11.33	1.34E-51
ENSG00000136167	LCP1	13:46125920-46211871	protein_coding	0.105	-6.40	1.02E-17
ENSG00000170627	GTSF1	12:54455950-54473602	protein_coding	0.105	-11.53	1.52E-53
ENSG00000130844	ZNF331	19:53519527-53580269	protein_coding	0.078	-5.73	3.41E-167
ENSG00000011677	GABRA3	X:152166234-152451315	protein_coding	0.061	-9.34	4.63E-34
ENSG00000143320	CRABP2	1:156699606-156705816	protein_coding	0.061	3.99	3.80E-14
ENSG00000133169	BEX1	X:103062651-103064171	protein_coding	0.059	4.92	3.69E-72
ENSG00000101160	CTSZ	20:58995185-59007254	protein_coding	0.058	-7.04	3.12E-40
ENSG00000251381	LINC00958	11:12961541-12989597	lncRNA	0.056	-6.02	5.34E-58
ENSG00000249568	AC104793.1	4:161378953-161388417	lncRNA	0.055	5.51	1.83E-42
ENSG00000076242	MLH1	3:36993332-37050918	protein_coding	0.055	7.65	7.02E-23
ENSG00000198185	ZNF334	20:46499630-46513559	protein_coding	0.054	-9.30	1.62E-33
ENSG00000188511	C22orf34	22:49414524-49657542	lncRNA	0.051	-9.12	1.50E-31
ENSG00000168672	LRATD2	8:126552443-126558478	protein_coding	0.051	3.20	0.00E+00
ENSG00000224817	AC010789.1	10:101701994-101730037	lncRNA	0.050	-9.09	1.97E-31
ENSG00000204389	HSPA1A	6:31815543-31817946	protein_coding	0.050	8.93	3.53E-06
ENSG00000181007	ZFP82	19:36383120-36418644	protein_coding	0.050	-8.36	1.31E-26
ENSG00000099399	MAGEB2	X:30215563-30220089	protein_coding	0.050	-3.28	5.11E-02

4.4.6 Differential gene expression analysis identified 3,501 DEGs as a result of *EPO*$^{-/-}$ knock-out.

To obtain the best list of DEGs specific to knocking out the *EPO* gene, differential gene expression analysis was performed comparing controls to KOA and KOB separately. 6,470 DEGs were identified in KOA whilst 6,674 DEGs were identified in KOB (P-adj $\leq$ 0.05). 693 and 657 of these DEGs had the strongest amount of differential expression ($\log_2$ fold-change $\geq |2|$) in KOA and KOB respectively. MA and volcano plots (**Figure 4.21A-B**) revealed an abundance of genes up- or down-regulated in the knock-outs compared to the controls with similar numbers of DEGs (*P*-value $\leq$ 0.05) with similar levels of differential expression identified. Several of the genes which showed the highest amount of differential expression were shared between the two knock-outs. These results indicated that differential expression was likely due to the effect of disrupting the *EPO* gene. When combining the list of DEGs identified in the two separate analyses (*P*-adj $\leq$ 0.05), 3,722 were found to be shared with 314 showing evidence of strong differential expression ($\log_2$ fold-change $\geq |2|$) (**Figure 4.22A**). The direction of effect of these 3,501 DEGs was consistent between knock-out cell-lines **(Figure 4.22B,** Pearson's correlation coefficient, r=0.9, P-value < $2.2\text{x}10^{-16}$). 1,750 of the consistent overlapping DEGs were down-regulated whilst 1,741 were up-regulated. A strong agreement of expression pattern across knock-out samples was seen causing clear segregation between knock-out and controls **(Figure 4.23).** The 221 genes showing inconsistent directions of effect were removed from further analysis. A list of the top 20 up**-** and down-regulated genes are listed in **Table 4.7.** The mRNA expression levels of two of the top up- and two of the top down-regulated DEGs which had highest predicted expression in HEK-293 cells were validated by qRT-PCR to confirm the same expression patterns were seen within the cells as estimated by RNA-seq. The results confirmed the differential expression identified by RNA-seq with those up-regulated showing higher mRNA expression levels and those down-regulated showing lower mRNA expression levels compared to controls (**Figure 4.23**).

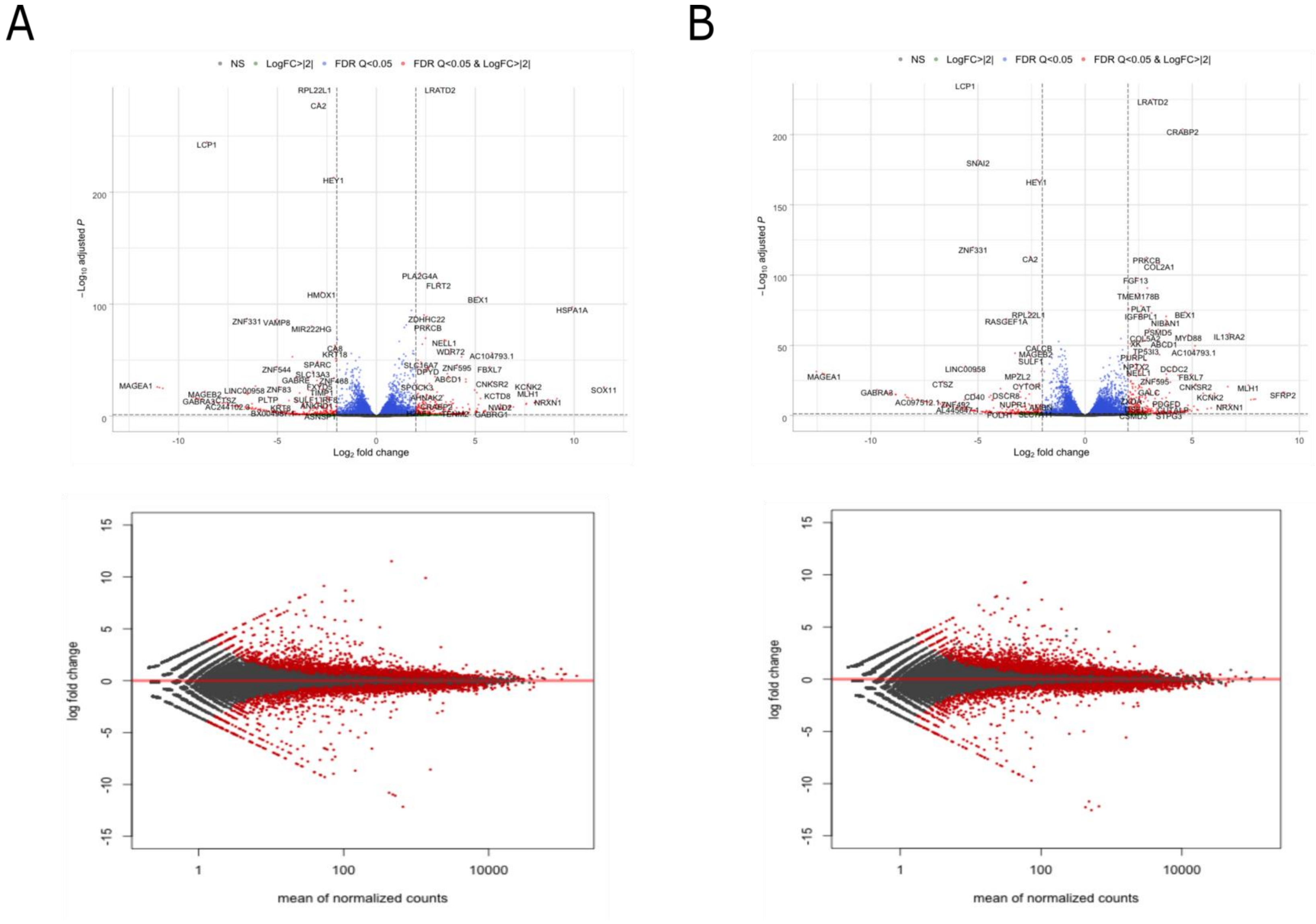


Figure 4.21: Differential gene expression analysis of EPO knock-outs compared to wild-type controls. ***A:*** *Volcano (top plot) and MA (bottom plot) plot for the differential expression of KOA compared to wild-type.* ***B:*** *Volcano (top plot) and MA (bottom plot) plot for the differential expression of KOB compared to wild-type*. *In the volcano plots, the green dots represent absolute $\log_2$ fold-change ≥ 2; blue dots represent P-adj ≤ 0.05; red dots represent P-adj ≤ 0.05 and absolute $\log_2$ fold-change ≥ 2. In the MA plots, the red dots represent genes passing P-adj threshold of 0.05.*

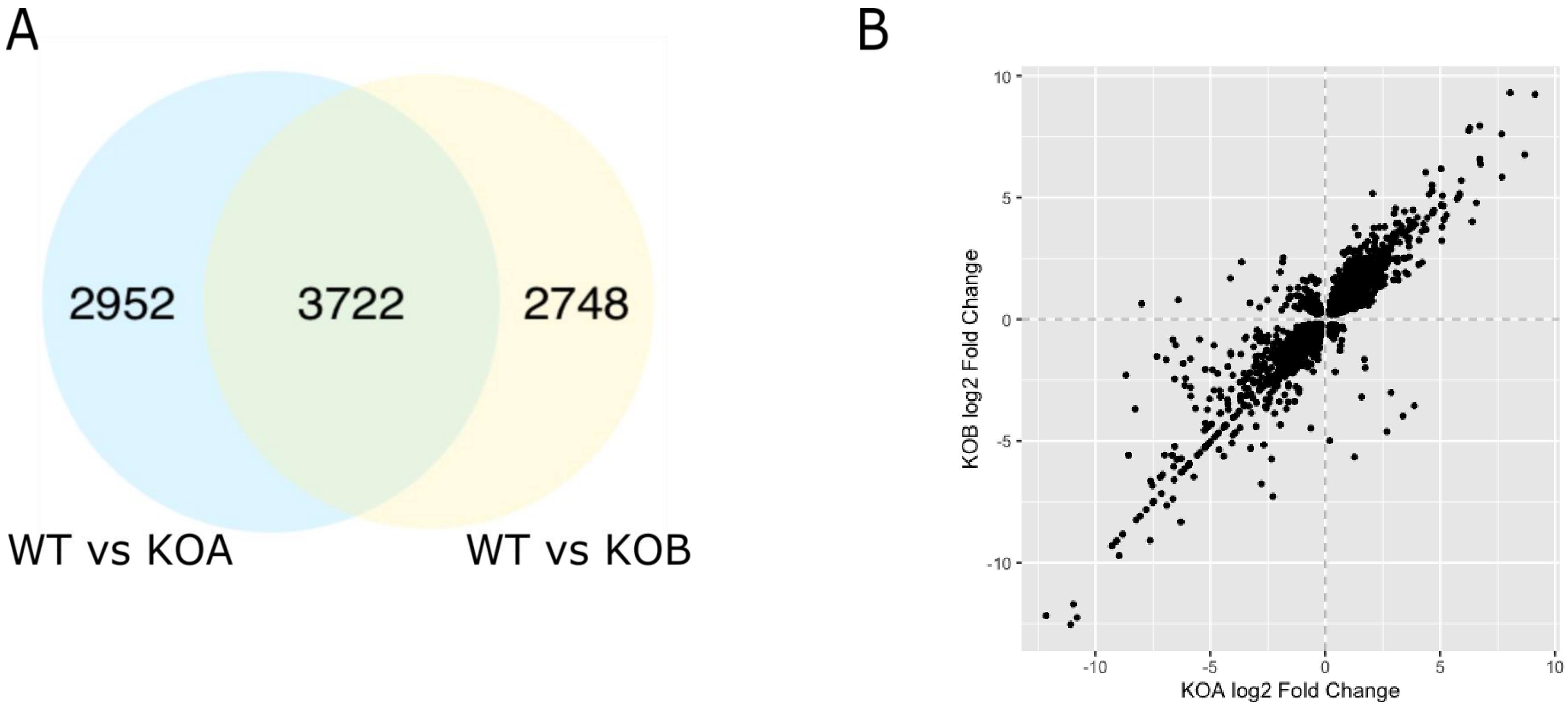


Figure 4.22: Identification of 3,722 overlapping differentially expressed genes in both knock-outs compared to wild-type controls. A: *Venn diagram of the differentially expressed genes identified in differential expression analysis comparing KOA to wild-type and KOB to wild-type. 3,722 of the DEGs identified in each analysis were shared and likely to be differentially expressed due to disruption of the EPO gene.* ***B:*** *Comparison of the* $\log_2$ *fold-change of the 3,722 DEGs in KOA (x-axis) and KOB (y-axis). There is a high degree of consistency in the* $\log_2$ *fold-changes in both knock-outs. Pearson's correlation coefficient, r=0.9, P-value <* $2.2x10^{-16}$.

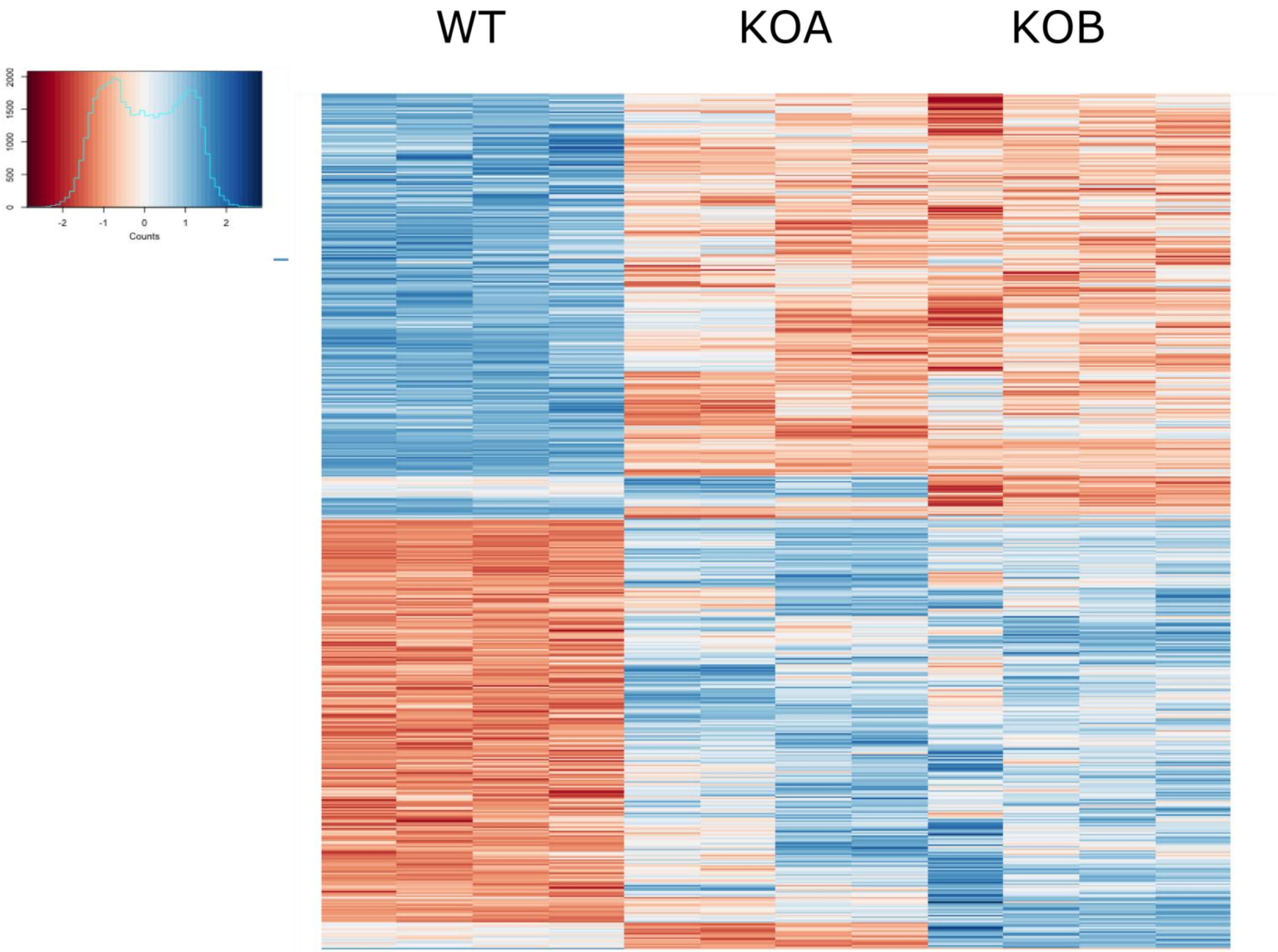


Figure 4.23: Heat map of the 3,722 overlapping differentially expressed genes. *Clustering was performed on the 3,722 overlapping DEGs which were found to be statistically significantly differentially expressed by passing a P-adj $\leq$ 0.05 and absolute log$_2$ fold-change $\geq$ 2. The regularised log counts for the differentially expressed genes were converted to a z-score by scaling across rows. Rows were clustered via Euclidean distance and Ward's linkage. Samples were fixed to match the cellular genotype. Altered expression profiles are seen between knock-outs and wild-types.*

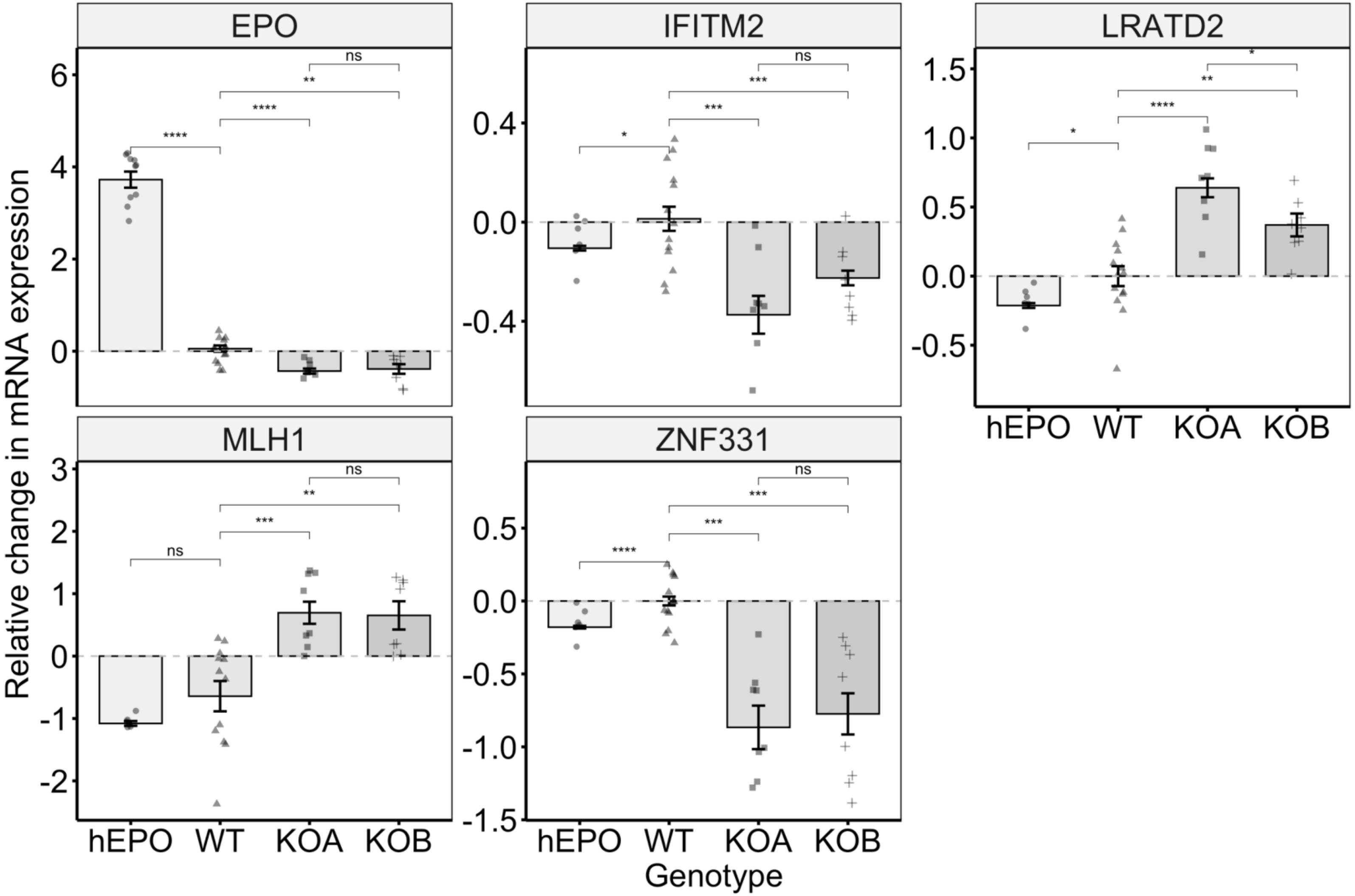


Figure 4.24: Relative change in mRNA expression levels of four of the most DEGs identified through RNA sequencing. *qRT-PCR was performed to validate differential expression of a four of the most strongly DEGs that have high expression in HEK-293 cells. Cells transfected with an EPO over-expression construct were used as positive controls (hEPO). EPO was repeated as a control experiment. Data is shown as mean ± SEM. Paired t-test was performed to test for levels of significance. *P≤0.05, **P≤0.01, ***P≤0.001, ****P≤0.0001.*

Table 4.7: Differential expression of the top 20 up- and down-regulated genes.

Log$_2$ fold-changes and P-adj values are summarised for each of the analyses; WT vs KOA and WT vs KOB. The average of the PC1 eigenvalues was calculated from the PC eigenvalues for WT vs KOA and PC eigenvalues for WT vs KOB.

			WT vs KOA			WT vs KOB			
Direction	**Geneid**	**GeneSymbol**	**Log$_2$ Fold-change**	**SE**	**padj**	**Log$_2$ Fold-change**	**SE**	**padj**	**Average PC1**
Down-regulated	ENSG00000136167	LCP1	-8.56	0.25	1.95E-245	-5.58	0.12	0	0.11
	ENSG00000198681	MAGEA1	-12.15	1.02	2.64E-30	-12.18	1.02	2.34E-30	0.10
	ENSG00000221867	MAGEA3	-11.08	0.99	9.72E-27	-12.54	1.03	4.55E-32	0.09
	ENSG00000197172	MAGEA6	-10.97	0.99	3.20E-26	-11.71	1.03	5.13E-28	0.09
	ENSG00000170627	GTSF1	-10.80	0.99	1.60E-25	-12.26	1.02	9.50E-31	0.09
	ENSG00000130844	ZNF331	-6.55	0.32	8.62E-88	-5.23	0.22	1.20E-120	0.07
	ENSG00000099399	MAGEB2	-8.67	0.84	8.10E-23	-2.30	0.16	1.70E-46	0.06
	ENSG00000100979	PLTP	-5.47	0.60	6.33E-18	-0.83	0.18	5.07E-05	0.06
	ENSG00000137573	SULF1	-3.55	0.39	4.53E-18	-2.53	0.18	2.43E-41	0.06
	ENSG00000163584	RPL22L1	-3.11	0.08	0	-2.63	0.14	2.07E-74	0.06
	ENSG00000170421	KRT8	-4.84	0.70	1.15E-10	-1.07	0.17	4.29E-09	0.05
	ENSG00000104267	CA2	-2.92	0.08	3.08E-280	-2.56	0.11	6.42E-114	0.05
	ENSG00000113140	SPARC	-2.97	0.20	2.33E-49	-0.45	0.11	0.00026248	0.05
	ENSG00000179455	MKRN3	-7.33	0.74	2.34E-21	-1.53	0.22	9.83E-11	0.05
	ENSG00000011677	GABRA3	-8.97	1.03	1.91E-16	-9.72	1.03	3.85E-19	0.05
	ENSG00000198131	ZNF544	-5.04	0.35	8.34E-45	-3.27	0.23	3.86E-45	0.05
	ENSG00000101160	CTSZ	-7.61	0.85	2.22E-17	-6.64	0.62	6.09E-25	0.05
	ENSG00000100292	HMOX1	-2.77	0.12	2.89E-111	-1.91	0.16	4.63E-32	0.05
	ENSG00000222041	CYTOR	-6.11	0.54	1.70E-27	-2.72	0.26	1.48E-23	0.04
	ENSG00000102265	TIMP1	-2.75	0.26	2.59E-24	-0.73	0.17	0.00012571	0.04
	ENSG00000198185	ZNF334	-9.29	1.03	1.55E-17	-9.31	1.04	1.51E-17	0.04
Up-regulated	ENSG00000133169	BEX1	5.15	0.23	4.59E-107	4.66	0.25	1.94E-74	0.07
	ENSG00000082482	KCNK2	7.69	0.66	2.02E-29	5.84	0.69	1.64E-15	0.06
	ENSG00000168672	LRATD2	3.23	0.08	0	3.16	0.10	6.99E-226	0.06
	ENSG00000076242	MLH1	7.67	0.74	2.98E-23	7.61	0.75	2.38E-22	0.06
	ENSG00000179915	NRXN1	8.68	1.03	2.06E-15	6.75	1.09	8.75E-09	0.06
	ENSG00000183580	FBXL7	5.73	0.40	1.20E-44	4.94	0.42	4.48E-30	0.05
	ENSG00000101986	ABCD1	3.64	0.28	4.28E-36	3.69	0.23	3.20E-53	0.05
	ENSG00000166415	WDR72	3.76	0.22	1.29E-60	3.78	0.21	2.44E-71	0.05

	ENSG00000143320	CRABP2	3.04	0.42	1.46E-11	4.55	0.15	1.31E-204	0.05
	ENSG00000165973	NELL1	3.44	0.19	1.17E-68	2.50	0.20	5.05E-33	0.05
	ENSG00000149970	CNKSR2	5.86	0.48	1.35E-31	5.16	0.50	4.25E-23	0.05
	ENSG00000095261	PSMD5	3.52	0.20	4.77E-68	3.41	0.20	5.57E-62	0.05
	ENSG00000138744	NAAA	3.70	0.23	6.79E-56	3.80	0.23	5.62E-60	0.04
	ENSG00000272602	ZNF595	4.07	0.28	1.62E-46	3.25	0.29	7.56E-27	0.04
	ENSG00000166501	PRKCB	2.61	0.13	7.10E-82	2.87	0.12	2.63E-113	0.04
	ENSG00000177108	ZDHHC22	2.55	0.13	1.73E-89	1.68	0.19	4.29E-17	0.04
	ENSG00000154162	CDH12	3.33	0.21	2.30E-56	3.80	0.22	8.50E-65	0.04
	ENSG00000136160	EDNRB	4.52	0.36	9.07E-34	5.13	0.33	2.20E-50	0.04
	ENSG00000145423	SFRP2	9.13	1.05	2.46E-16	9.23	1.04	2.93E-17	0.04

4.4.7 Gene enrichment analysis emphasises the pleiotropic effects of EPO by revealing an enrichment of DEGs in several metabolic, stress-related and DNA repair pathways.

The 3,501 DEGs with consistent differential expression were found enriched in multiple biological processes involved in DNA repair, mRNA processing, protein binding and degradation, cellular respiration and mitochondrial function using Enrichr indicating the important role EPO signalling has within the cell (Figure 4.25**, Appendix 2**: Gene ontology analysis**)** (E. Y. Chen et al., 2013; Kuleshov et al., 2016; Z. Xie et al., 2021). Several other biological processes enriched with the DEGs involved energy metabolism including fatty acid oxidation, mitochondrial functions and cellular respiration pathways (**Appendix 2**: Gene ontology analysis**)**. The main signalling pathway identified through gene enrichment analysis was the Notch signalling pathway including its receptor and ligand (**Appendix 2**: Gene ontology analysis**).** Focusing on molecular functions, the DEGs were enriched in ATPase activity and mannosyl-oligosaccharide mannosidase activity (**Appendix 2**: Gene ontology analysis**)**. Notch signalling was also identified when looking for an enrichment of genes in KEGG pathways alongside other pathways involved in metabolism, such as thermogenesis and AMPK signalling (**Figure 4.27**, **Appendix 2**: Gene ontology analysis). Both Enrichr and GoSeq2 gave very similar results.

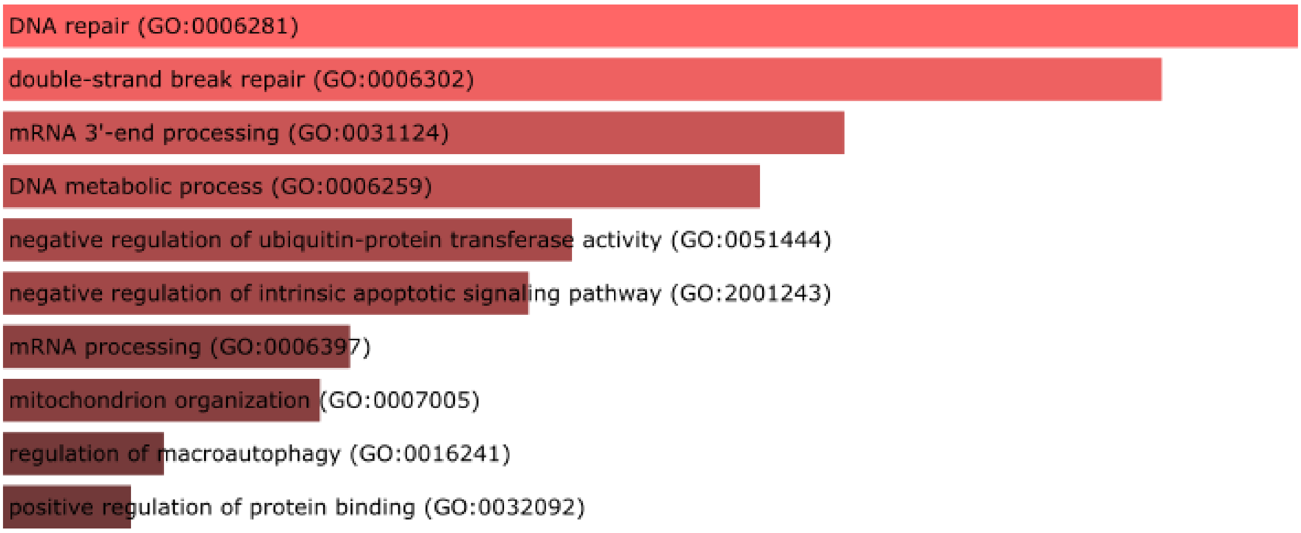


***Figure 4.25: The top 10 biological processes enriched with genes differentially expressed as a result of EPO$^{-/-}$* knock-out.** *Gene enrichment analysis was performed on the 3,501 DEGs with consistent patterns of differential expression using Enrichr. The bars are sorted by p-value with the length of the bar representing the p-value (lowest P-value = longest bar).*

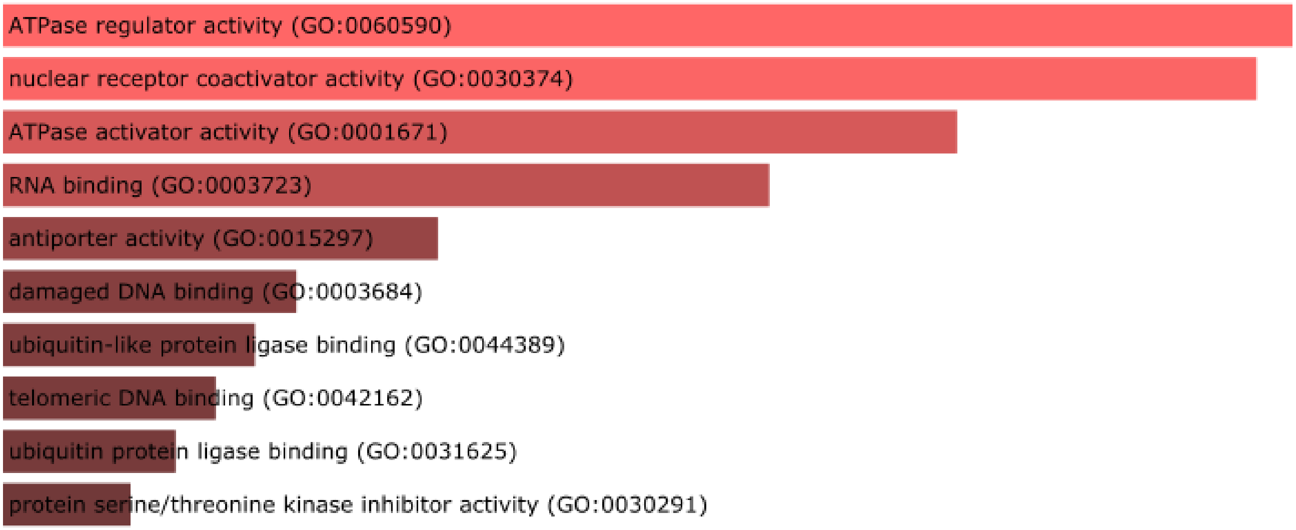


Figure 4.26: The top 10 molecular functions enriched with genes differentially expressed as a result of EPO$^{-/-}$ knock-out.

Gene enrichment analysis was performed on the 3,501 DEGs with consistent patterns of differential expression using Enrichr. The bars are sorted by p-value with the length of the bar representing the p-value (lowest P-value = longest bar).

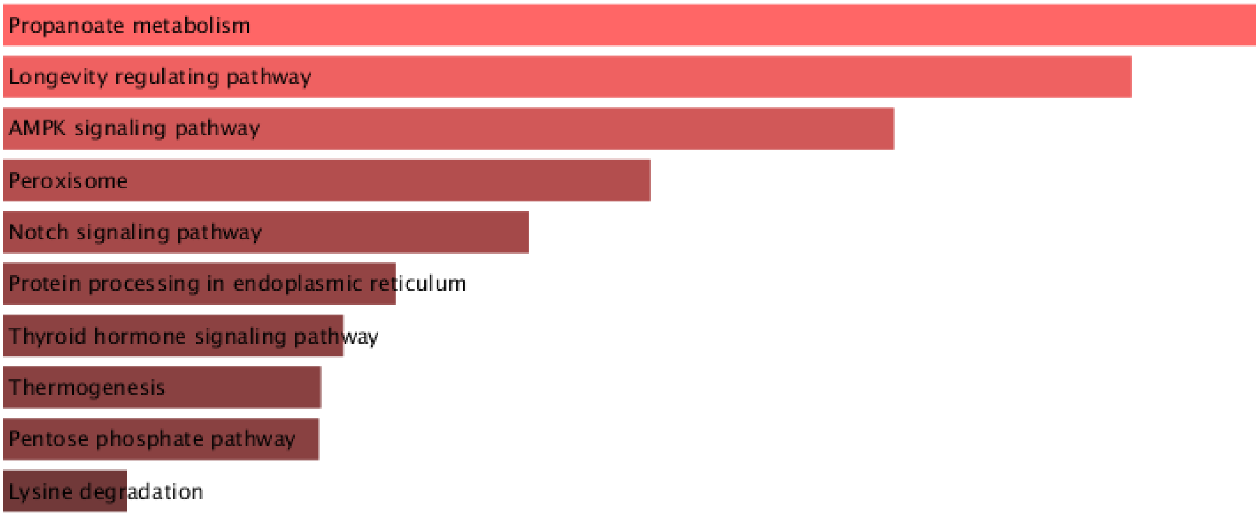


Figure 4.27: The top 10 KEGG pathways enriched with DEGs identified through RNA-seq analysis of the WT controls compared to the EPO$^{-/-}$knock-out.*Gene enrichment analysis was performed on the 3,501 DEGs with consistent patterns of differential expression using Enrichr. The bars are sorted by p-value with the length of the bar representing the p-value (lowest P-value = longest bar).*

4.4.8 Differential gene expression analysis and gene enrichment analysis reveals a role for EPO in the Notch signalling within HEK-293 cells

Gene enrichment analysis revealed an enrichment of DEGs, such as *DTX3L Parp9,* and *Hey1,* involved in the Notch signalling pathway alongside downstream molecular functions and biological processes of Notch signalling, such as cell-to-cell communications and cell-fate decisions (Figure 4.25**, Figure 4.26, Figure 4.27, Appendix 2**: Gene ontology analysis) (Mercher et al., 2008). Previous studies have also indicated a role for increased Notch signalling activity in response to hypoxia (Borggrefe et al., 2016; Diez et al., 2007; Fischer et al., 2004). I therefore used qRT-PCR to further validate dysregulated expression of genes involved in the Notch signalling pathway in the knock-outs. The Human Cell Atlas (Regev et al., 2017) was used to identify candidate DEGs within the Notch signalling pathway which have high expression in HEK-293 cells. I also identified other genes (*VEGFR3, Notch1, Hey2, Hes1)* involved in the Notch signalling pathway which had high expression in HEK-293, but did not show strong significant differential expression to confirm that several parts of the Notch signalling cascade showed altered mRNA expression as a result of *EPO* disruption (**Figure 4.29).** qRT-PCR confirmed dysregulated expression levels of seven genes (P < 0.05) implicated in the Notch signalling confirming Notch signalling is likely down-regulated as a result of a loss of *EPO* function ***(*****Figure 4.29)**. For *Notch1* and *Parp9,* cells transfected with the over-expression *EPO* construct also showed similar pattern of altered mRNA expression levels as the knock-outs (**Figure 4.29)**. This highlights the complex cross-talk occurring within cells and suggests a negative feedback loop mechanism to prevent excessive hypoxic gene induction (Diez et al., 2007). Overall, my results indicate a role for EPO in controlling the Notch signalling pathway.

4.4.9 Differential gene expression analysis and qRT-PCR reveals a role for EPO in the BMP/SMAD signalling pathway within HEK-293 cells

Differential gene expression analysis revealed several DEGs (with evidence of strong differential expression [$\log_2$fold-change > |2|]) involved in the BMP/Smad pathway, such as *BMP5* and *RPL22L1* **(Figure 4.30**). The BMP/Smad pathway has been shown to play a role haematopoiesis, regulation of circulatory iron and protection against renal disease (Goh et al., 2015; Prestigiacomo & Suter-Dick, 2018; Shuyun Rao et al., 2012; W. Wang et al., 2005; Yong Zhang et al., 2013).

I, therefore, hypothesised that EPO signalling plays a role in regulating this pathway and used qRT-PCR to further investigate mRNA expression levels of the identified BMP/Smad pathway DEGs and other genes within this signalling cascade **(Figure 4.30**). qRT-PCR revealed statistically significant downregulation (P < 0.001) of mRNA expression levels for *BMP5* and *RPL22L1*, consistent with findings from RNA-seq analysis **(Figure 4.31).** qRT-PCR of other genes implicated in the BMP pathway (*Smad5, RPL22, NRF2*) showed a trend towards down-regulated mRNA expression levels in *EPO*$^{-/-}$ knock-outs but this did not reach statistical significance (P > 0.05) **(Figure 4.31).**

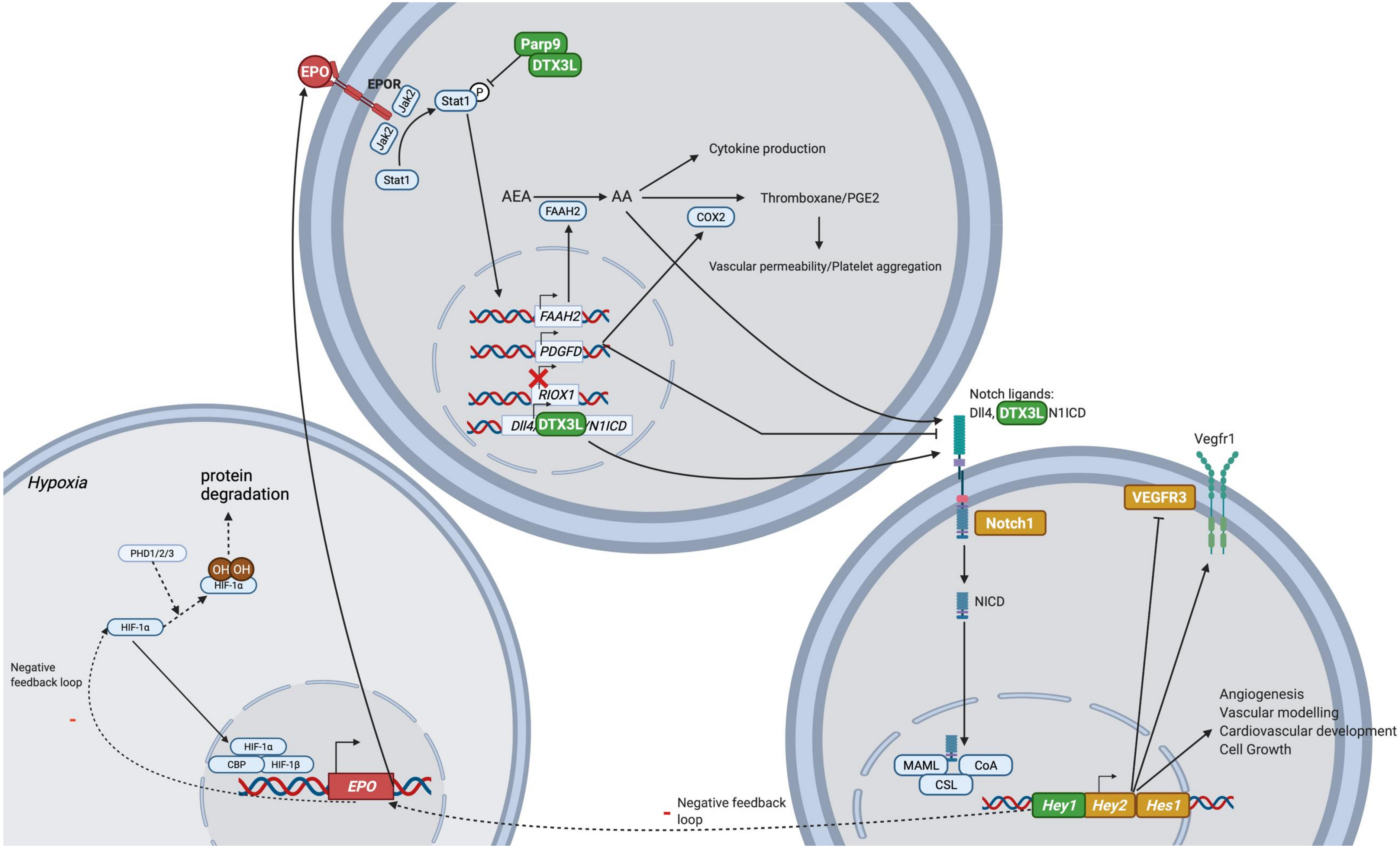


Figure 4.28: Schematic outline of the hypothesised link between EPO and the Notch signalling pathway. *Through differential gene expression analysis of RNA-seq data in EPO$^{-/-}$ knock-out cell-lines, several genes were identified that are implicated in the Notch signalling pathway (genes highlighted in green). Gene ontology analysis also revealed an enrichment of DEGs in pathways related to Notch signalling. qRT-PCR was used to investigate mRNA expression levels of these genes as well as other genes implicated in the Notch signalling pathway (highlighted in orange) to further elucidate a role for EPO in regulating the Notch signalling cascade. Created with BioRender.com.*

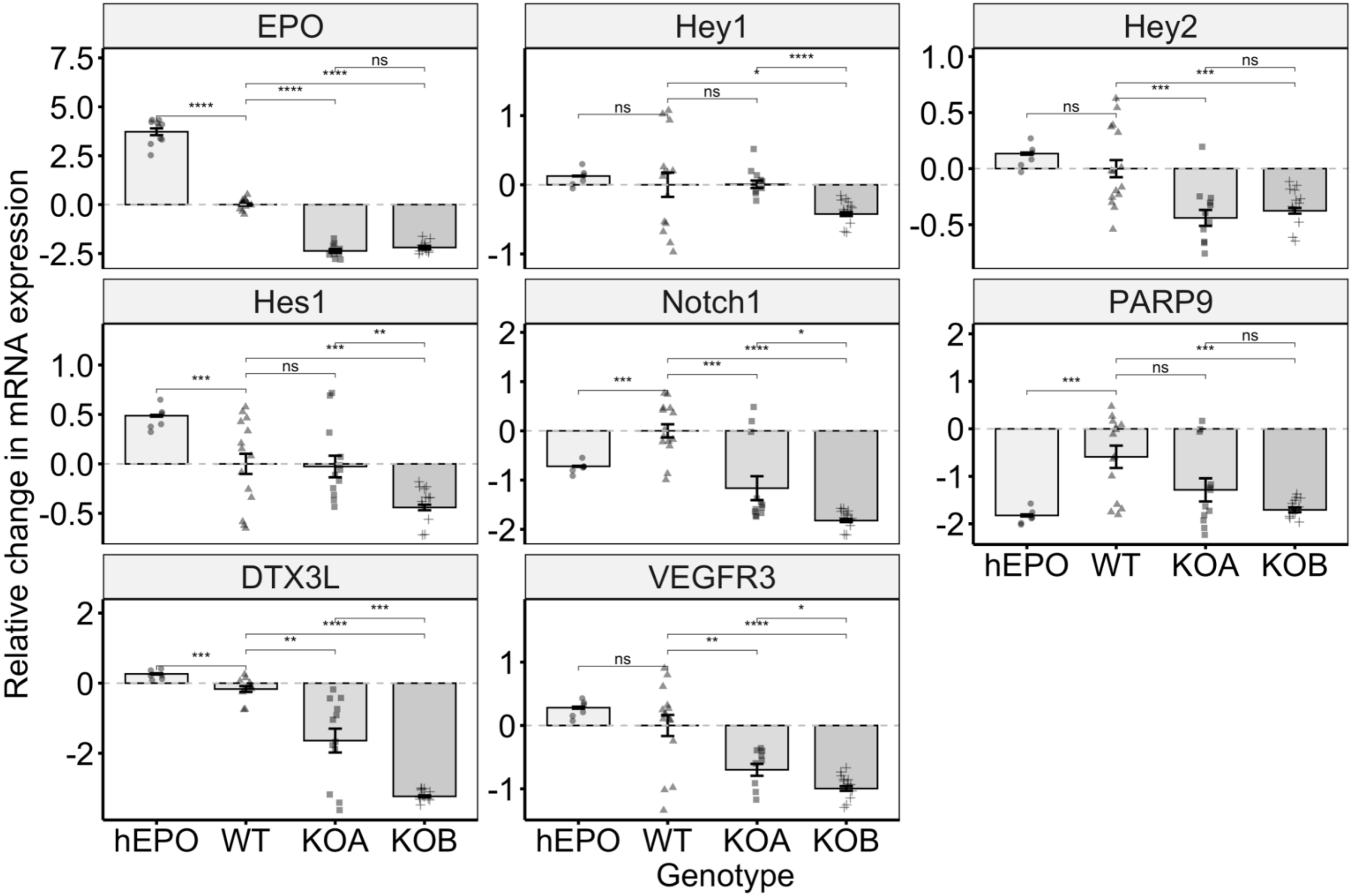


*Figure 4.29: Relative change in mRNA expression levels of several genes involved in the Notch signalling pathway. qRT-PCR was performed to validate the differential expression of genes involved in the Notch signalling pathway which had high expression in HEK-293 cells. Cells transfected with an EPO over-expression construct were used as positive controls (hEPO). EPO was repeated as a control experiment. Data is shown as mean ± SEM. Paired t-test was performed to test for levels of significance. *P≤0.05, **P≤0.01, ***P≤0.001, ****P≤0.0001.*

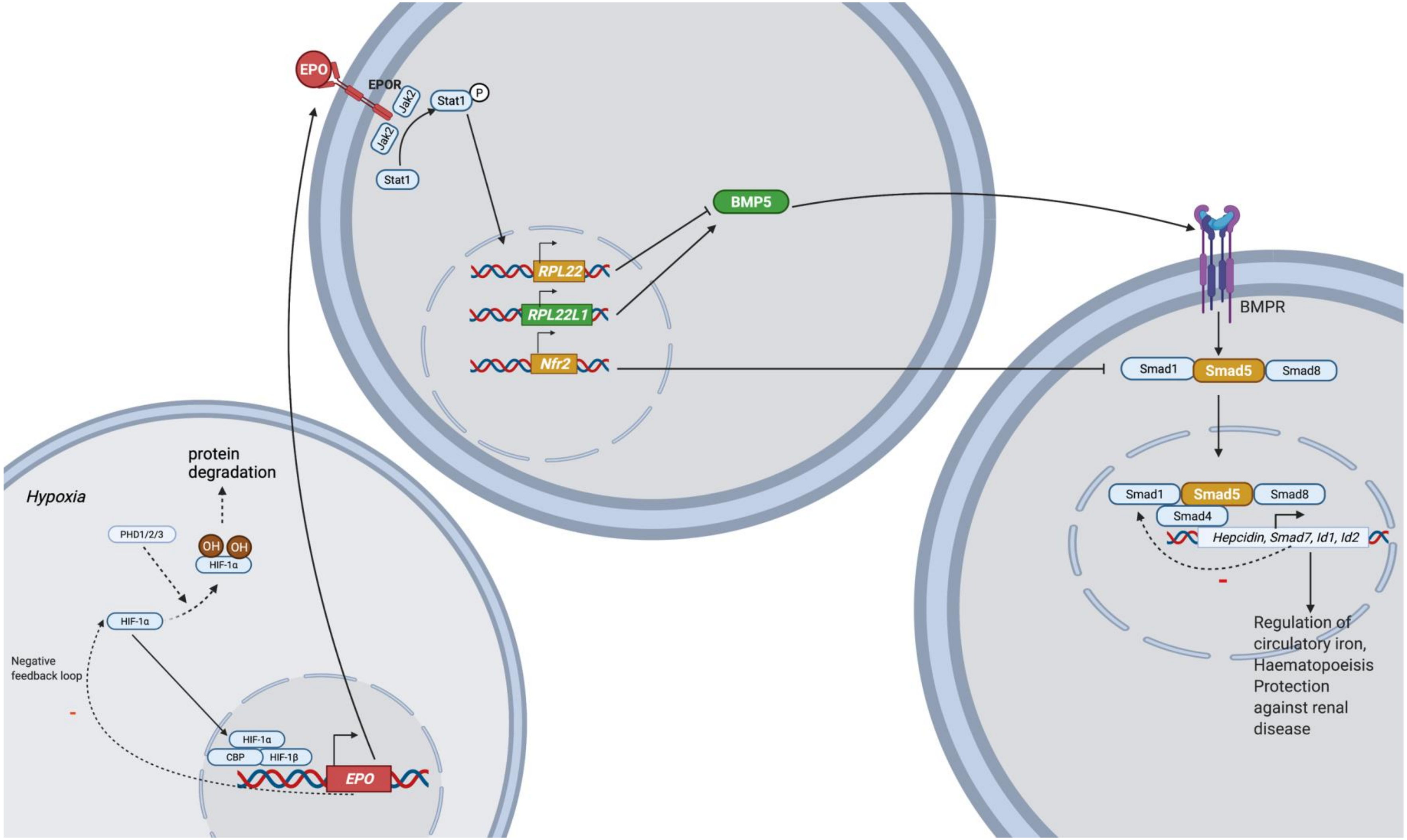


Figure 4.30: Schematic outline of the hypothesised link between EPO and the BMP/Smad signalling pathway. *Through differential gene expression analysis of RNA-seq data in EPO$^{-/-}$ knock-out cell-lines, several genes were identified that are implicated in the BMP/Smad signalling pathway (genes highlighted in green). The BMP/Smad pathway has previously been shown to be involved in regulating iron homeostasis, haematopoiesis and provide protection against renal disease and therefore it was thought EPO might play a role in controlling this pathway. qRT-PCR was used to investigate mRNA expression levels of the genes identified as differentially expressed through RNA-seq (highlighted in green) as well as other genes implicated in the BMP/Smad signalling pathway (highlighted in orange) to further elucidate a role for EPO in regulating the BMP/Smad signalling cascade. Created with BioRender.com*

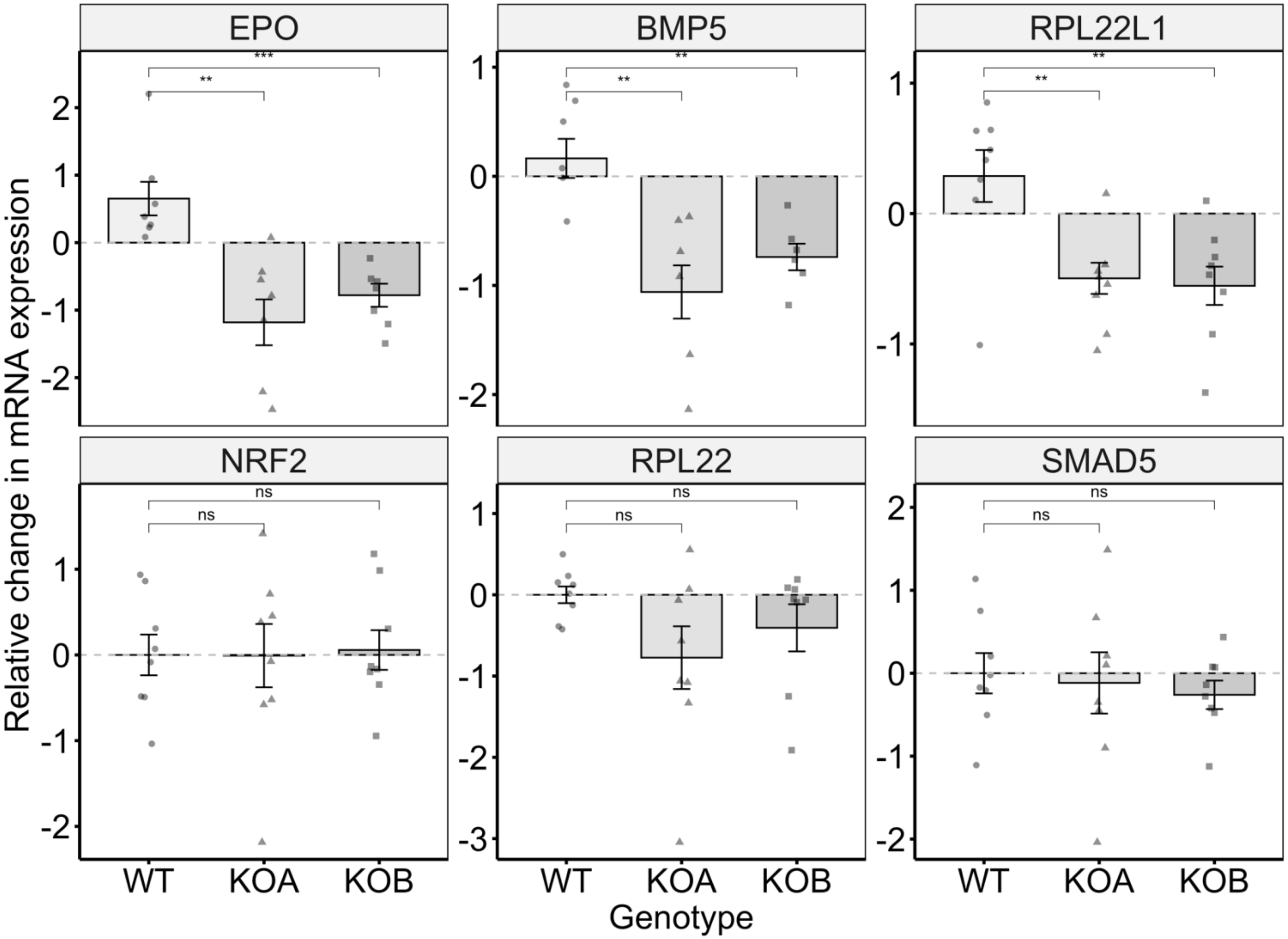


Figure 4.31: Relative change in mRNA expression levels of several genes involved in the BMP/Smad signalling pathway. *qRT-PCR was performed to validate the differential expression of genes involved in the BMP/Smad signalling pathway which had high expression in HEK-293 cells to confirm whether EPO has a role in this pathway. EPO was repeated as a control experiment. Data is shown as mean ± SEM. Paired t-test was performed to test for levels of significance. *P≤0.05, **P≤0.01, ***P≤0.001, ****P≤0.0001.*

4.5 Discussion

In this Chapter, a whole *EPO*$^{-/-}$ gene knock-out was established using CRISPR-Cas9 gene-editing with a paired gRNA approach. Paired gRNAs were used to increase efficiency and reduce risk of off-target effects (X.-H. Zhang et al., 2015). The paired gRNAs target two conserved sites on the protein-coding exons resulting in the deletion of a large portion (~1,100 bp) of the genomic sequence efficiently disrupting target gene function **(Figure 4.15)**. The loss of this region from the genomic DNA sequence results in a frameshift and a subsequent reduction in mRNA and protein expression of the target gene which was confirmed by qRT-PCR and western blotting **(Figure 4.16).** Transcriptomic-wide analysis was performed and 3,501 DEGs likely attributable to a lack of *EPO* were identified with 314 showing strong differential expression ($\log_2$fold-change ≥ |2|) **(Figure 4.22).** These DEGs were found enriched in several important cellular processes including Notch signalling emphasising the mitogenic effects of EPO, DNA repair, mitochondrial function and energy metabolism which support previous studies suggesting a pleiotropic role for EPO (Suresh et al., 2020). This comprehensive transcriptomic profiling of gene expression based on RNA-seq has generated a robust set of genes of biological significance in relation to the downstream effects of EPO in HEK-293 cells. I have established a knowledge base of up- and down-regulated genes capturing a wide-range of changes attributable to *EPO* expression that can be used as a baseline in future studies to investigate the effects of genetic variants on *EPO* expression levels. To my knowledge, this study is the first to establish an *EPO*$^{-/-}$ knock-out model in human cell-lines and to quantify cell-line gene expression changes resultant of *EPO* gene disruption*.* The established cell-line model can be used in future research to further understand the downstream causal pathways that have been suggested here and provide a better understanding of the aetiology of diseases characterised by low EPO levels, such as anaemia.

Through functional work, I have provided a better understanding of the molecular functions and biological pathways in which EPO is involved in a relevant human cell-line. The results demonstrate the effects of the largest change in EPO levels *in vitro* and can be used for comparisons when trying to functionally validate the effect of variants or gene modifications, identified through genetic analyses, on *EPO* expression levels. This work shows the feasibility of introducing gene-edits to the *EPO* gene via CRISPR-Cas9 in a relevant cell-line. Similar approaches can be used to establish SNP knock-in cell models and similar expression profiles can be investigated to determine if these SNPs are causal in controlling *EPO* gene expression and to what degree genetic variants have the same effect compared to the largest change in EPO levels i.e. whole gene knock-out. These findings can then be used to support genetic studies providing functional evidence for correlation between variants and disease.

Transcriptomic analysis revealed several genes implicated in the Notch signalling pathway (*HEY1, DTX3L, PARP9)*. Differential expression of these genes, alongside additional genes implicated up- and downstream of Notch (*NOTCH1, HEY2, HES1, VEGFR3)*, were validated using qRT-PCR indicating an important role for EPO in this signalling cascade and downstream effects, such as angiogenesis, cardiovascular development and cell survival **(Figure 4.28, Figure 4.29).** Notch signalling has previously been reported to be activated in response to hypoxia in several other cell-types and to play a role in erythroid homeostasis by regulating apoptosis and these findings suggest that Notch activity could be mediated by *EPO* expression (Borggrefe et al., 2016; Diez et al., 2007; Fischer et al., 2004; Gustafsson et al., 2005; Robert-Moreno et al., 2007). The Notch signalling pathway plays a primary role in several key cellular process that regulate cell fate specification including the maintenance and differentiation of haematopoietic stem cells in the early embryo emphasising how the link between EPO and Notch signalling could be important in controlling erythrocyte development (Duarte et al., 2018; C. Huang et al., 2021). A role for EPO in initiating Notch signalling has previously been implicated in breast cancer where pharmacological rhEPO increased the number of breast-cancer initiating cells through increased activation of Notch signalling impacting overall survival and local tumour control highlighting the complex downstream adverse effects of increased Notch signalling through increased EPO levels (Phillips et al., 2007). I provide further evidence of the interplay between EPO and Notch signalling.

I also indicate a role for EPO in the BMP/Smad pathway as several keys genes of the BMP/Smad pathway, such as *BMP5* and *RPL22L1*, were found differentially expressed in the knock-outs (*P*-adj $\leq$ 0.05 & $\log_2$ Fold-change $\geq$ |2|) and this differential expression was validated through qRT-PCR **(Figure 4.30, Figure 4.31).** The BMP/Smad pathway is known to influence circulating iron levels and haematopoiesis through *HAMP* expression and this could be driven by alterations in *EPO* expression levels (Goh et al., 2015; Prestigiacomo & Suter-Dick, 2018; Yong Zhang et al., 2013). However, I did not find altered mRNA expression levels of other genes implicated in this pathway that had high expression in HEK-293 cells, e.g. *RPL22, Nfr2* and *Smad5* (**Figure 4.31)**. This does not mean that this pathway is not affected by *EPO* expression but could highlight the complicated feedback mechanisms within the cell and effects on expression of the other tested genes may only be detected by intracellular signalling between different cell-types which is not possible to investigate using single-cell culture methods. It would be worth following up these findings by investigating effects in co-culture or upon exposure of cells to hypoxia to further elucidate the link between EPO and the BMP/Smad pathway.

Through GO analysis, I found an enrichment of DEGs in several biological processes including DNA repair, metabolic processes, mRNA processing, cell-cycle activity, fatty acid oxidation, and control of signalling pathways through protein binding, protein degradation, coactivator recruitment and receptor activity. These results highlight the pivotal role EPO plays in a wide-range of molecular functions and biological activities, and signifies that EPO sits in a network of genes that are involved in integral cellular functions. These findings support previous studies which show that EPO influences cell signalling pathways, cellular proliferation, cell-fate, and protection against cellular stresses further emphasising the pleiotropic effects of EPO in systems other than just the haematopoietic system (Suresh et al., 2020; L. Wang et al., 2014). The top biological processes enriched in the DEGs were DNA repair pathways indicating that *EPO*$^{-/-}$ knock-out results in aberrant DNA repair which is consistent with previous studies finding a role for EPO in regulating p53-dependent pathways **(**Figure 4.25**)** (Pham et al., 2019). These biological processes related to DNA repair, such as DSB repair, mRNA 3'end processing and base-excision repair, could have also been identified due to the CRISPR-Cas9 gene-editing mechanism introducing DSBs and initiating the DNA repair pathway and therefore enrichment of DEGs involved in these pathways in knock-outs is not surprising (**Appendix 2**: Gene ontology analysis**)**. Additionally, the enrichment of DEGs involved in mannosidase activity is reassuring; the mannosidase enzymes are involved in the synthesis of glycoproteins of which EPO is one and in the knock-out model reduced functioning of these enzymes is found indicating the RNA-seq results are specific *EPO* disruption (**Appendix 2**: Gene ontology analysis**)**.

Several of the GO terms, including cellular response to amino acid stimulus, mitotic sister chromatid segregation, cAMP-dependent protein kinase inhibitor activity, ATPase activity, and RNA binding, are related to processes and functions which occur in the kidney and are similar to those identified previously by transcriptomic analysis of biopsy samples in CKD patients (**Appendix 2**: Gene ontology analysis**)** (Guo et al., 2019). The kidneys are one of the most energy-demanding organs in the human body with an abundance of mitochondria, ATP utility and oxygen consumption (Console et al., 2020). The top terms include ATPase regulator and activator activity, mitochondrial organisation, fatty acid oxidation, aerobic respiration, and regulation of oxidative phosphorylation which are crucial for ATP production**.** ATP production requires high oxygen levels and therefore EPO could influence ATP production by regulating oxygen levels and altering these biological processes (Solaini et al., 2010). ATPases release energy for all cellular activities and the enrichment in ATPase functions indicate that the amount of energy released for other cellular processes and pathways in kidneys cells is altered in the *EPO* knock-outs (Bonora et al., 2012; Neupane et al., 2019)**.** These results provide a better understanding of the role EPO plays within the kidney but also indicates the role EPO may have in other organs and tissues, such as the heart or skeletal muscle, which also require high levels of ATP (Console et al., 2020). The transcriptomic findings presented could be instrumental in the wider field investigating the impacts of EPO on disease aetiology.

Transcriptomic data can be used to improve the understanding of EPO mechanism of action, aid evaluation of drug target candidates, aid prioritisation of likely successful drug targets and aid in early prediction of likely adverse drug target effects (Paananen & Fortino, 2020). Currently, EPO is used to treat anaemia in CKD due to the role EPO plays in the haematopoietic system

(Jelkmann, 2011, 2013). The RNA-seq findings, however, identify additional non-haematopoietic pathways and networks for EPO involvement and therefore highlight the potential for EPO to be used as a treatment for other diseases which may be caused by aberrations in these pathways. Further research into the specific effects of EPO on these identified pathways and the impact in different diseases is warranted.

As with any functional work there are a number of limitations that need considering. Firstly, our model was established in HEK-293 cells. Despite, HEK-293 cells being biologically relevant for the investigation of EPO due to it being highly expressed in the kidneys and HEK-293 being easy to grow, easy to transfect, and able to be utilised for both stable and transient expression, there are some caveats to using this cell-line (P. Thomas & Smart, 2005). HEK-293 cells have a complex karyotype carrying two or more copies of each chromosome and therefore it is difficult to know with certainty that *EPO* has been deleted in all chromosomal copies (Stepanenko & Dmitrenko, 2015)**.** EPO is also not that highly expressed in HEK-293 cells and therefore another cell-line displaying higher basal EPO levels would be worth considering for future investigations, such as HepG2 cells (Regev et al., 2017). Furthermore, although kidney cells produce and secrete high levels of EPO, it does not mean that the cells themselves express high levels of the protein and therefore a cell-line may not be the best for investigating the functional roles of EPO. Instead, co-culture could be used to investigate the amount of EPO produced and the impact in other cells of relevance to investigate the cross-talk between cells and the secretory effects of EPO (Vis et al., 2020). Furthermore, HEK-293 cells originate from an embryonic kidney cell, and therefore might not be fully representative of an adult, diseased kidney. Secondly, I established two *EPO*$^{-/-}$ cell-lines. Despite confirming that both cell-lines have reduced *EPO* expression at the mRNA level and protein level, there are slight differences between the two regarding the region of the genome that was excised by CRISPR-Cas9. This difference resulted in different premature stop codons being introduced and could explain some of the differences seen between knock-outs in and why KOA segregates clearly from KOB in PCA analysis. The knock-outs are not exactly homologous to each-other despite having undergone the exact same experimental conditions and both showing successful *EPO* disruption.

Consideration of these differences is needed when interpreting results. Although every effort was made to ensure that each cell-line was treated under the exact same experimental conditions and grown to similar confluency (80%), there may be some experimental variation which could have impacted the resultant gene expression profiles. Thirdly, CRISPR-Cas9 gene-editing comes with its own limitations primarily the ability to control and test for off-target effects (Doench et al., 2016; Naeem et al., 2020; Ran, Hsu, Wright, et al., 2013)**.** Despite using the paired gRNA approach to reduce risk of off-target effects, it is hard to know whether this is the only place the genome has been disrupted without sequencing the whole genome. Generating two *EPO*$^{-/-}$ cell-lines provides more confidence that the results are likely due to the *EPO* disruption and no off-target effects as similar results are seen in both knock-out cell-lines. Fourthly, care is required when performing RNA-seq analysis due to the risk of potential biases. The expression of hundreds of genes tends to be over- or under-estimated and many of these genes are relevant to human disease and this may be the case for EPO (Robert & Watson, 2015). From RNA-seq analysis, few reads were aligned to the *EPO* gene itself – this may be attributable to low basal *EPO* expression in HEK-293 cells or may be due to an under-estimation of *EPO* gene expression. Alternative alignment methods could be investigated. RNA-seq involves several processing steps including library generation, PCR amplification and these steps are error-prone and can lead to the inclusion of biases (Shi et al., 2021). Although, steps have been taken to eradicate risk of these errors and potential biases, such as trimming and transformation, they can remain. In this Chapter, I only looked at total gene expression changes and although significant gene expression changes were identified, there could be transcripts that are differentially expressed within the non-DEGs which have been missed.

In conclusion, this study is the first to achieve whole *EPO*$^{-/-}$ knock-out in HEK-293 cells and to perform downstream transcriptomic analysis. I have identified an enrichment of genes involved in several biological pathways and molecular functions further highlighting the pleiotropic activities of EPO. The current work highlights the feasibility of carrying out CRISPR-Cas9 gene-editing of the *EPO* gene in HEK-293 cells and provides a better understanding of the downstream pathways of EPO signalling.

Chapter 5 Single-base gene-editing to functionally validate the *cis-EPO* variant as causal in controlling EPO levels

This Chapter includes sections that have been taken directly from a pre-print paper in which I am the first author. I have reformatted and expanded on sections for the purpose of this thesis. This paper is currently undergoing review at GSK and AJHG.

Harlow, CE. Gandiwijaya, J. Bamford, RA. Wood, AR. Van der Most, P. Verweij, N. [25 authors] & Frayling, TM. 2022. Identification and single-base gene-editing functional validation of a *cis-EPO* variant for use to mimic novel EPO-increasing therapies.

I planned and set out all experiments described in this Chapter. I performed the majority of the laboratory works and analysis with the help, support and guidance of my supervisors, Professor Tim Frayling and Dr Asami Oguro-Ando. I also had the help and support of fellow Oguro-Ando lab members, Dr Rosemary Bamford and Mr Josan Gandawijaya for troubleshooting throughout. This is the first time that this type of experiment using CRISPR-Cas9 and the piggyBac^{TM} system to introduce a single SNP gene-edit has been carried out at Exeter and therefore it took time to develop, adapt and optimize the protocol outlined by Yusa et al. (2013). I would like to thank Dr. Akshay Bhinge for providing me with the piggyBac^{TM} multivector plasmid and the *piggyBac^{TM}* transposase plasmid. The CRISPR plasmid used (pSpCas9(BB)-2A-GFP [PX458]) was a gift from Feng Zhang (Addgene plasmid #41838; http://n2t.net/addgene:48138; RRID:Addgene_48138).

5.1 Introduction

Since the unprecedented increase in the amount of publicly available human genotypic and phenotypic data, the power to detect causal associations has improved and the number of genetic variants identified as associated with complex traits has increased (Cano-Gamez & Trynka, 2020; Visscher et al., 2017). This has led to rapid advancements in the understanding of the molecular basis of disease and facilitated the discovery of novel therapeutics (Shu et al., 2018). Despite these successes, the interpretation of GWAS associations remains limited particularly when trying to provide clinical insight into complex traits (Cano-Gamez & Trynka, 2020). It is often very difficult to distinguish the causal variant that is driving the association due to the complex underlying genetic architecture and the presence of LD meaning that neighboring genetic variants are inherited together and highly correlated with each other (Bush & Moore, 2012; Flister et al., 2013; Hormozdiari et al., 2014). It is also often difficult to determine which gene is involved and directly affected by the identified genetic variant (Shuquan Rao et al., 2021). This is because disease-associated loci, identified through GWAS, often contain multiple genes (Cano-Gamez & Trynka, 2020). Moreover, the majority (greater than 90%) of genetic variants identified through GWAS do not directly affect the coding sequence due to lying within the non-coding regions of the genome (Dixon et al., 2007; Nica et al., 2010). These variants could therefore affect expression of several genes within the locus due to lying within regulatory elements or overlapping promoters, enhancers and open-chromatin regions making determination of the affected gene and downstream pathway increasingly difficult (Lichou & Trynka, 2020). Despite additional approaches, such as fine mapping, colocalisation analysis and eQTL analysis helping refine the most likely involved genes, problems with LD remain making it unclear whether that variant is the true causal variant driving the association (Benner et al., 2016; Giambartolomei et al., 2014; Nica et al., 2010; Nicolae et al., 2010; Porcu et al., 2019; Wallace, 2020). Experimental data of how well these tools work is also lacking (Brandt et al., 2020).

Functional studies are becoming a powerful complementary approach to further dissect the direct effects of genetic variants and aid in the validation of genetic variants for use as proxies in predicting the long-term effects of therapeutic modulation in the drug development process (Lichou & Trynka, 2020; H. Wang et al., 2016; L. Yang et al., 2013). Despite high through-put methods, such as massively parallel reporter assays (MPRAs), proving useful in finding active regulatory variants or variants affecting splicing, these approaches are limited by only being able to validate non-coding variants in the enhancer or promoter regions and they cannot test the effect of variants that may alter gene expression through post-transcriptional modifications or by influencing nonsense-mediated decay (Brandt et al., 2020; J. Lin & Musunuru, 2018). Alternative methods are therefore needed to validate effects of single variants on expression levels in the native genomic context (Shuquan Rao et al., 2021).

The introduction of a desired gene-edit is achievable by providing a homologous sequence within an exogenous pre-designed donor plasmid which upon transfection triggers homologous recombination (HR) (Hsu et al., 2014; Maruyama et al., 2015; J.-P. Zhang et al., 2017). The development of the CRISPR-Cas9 system has revolutionized the field of genome-editing in mammalian cells and has provided a means of introducing single variants in a more straightforward and relatively easy manner (Ng et al., 2020). CRISPR-Cas9 has simplified and increased the speed at which genetic variants and candidate genes can be functionally validated by bypassing the need for protein engineering to develop a site-specific nuclease or the need to generate a new germline or conditional alleles (Brandt et al., 2020; J. Lin & Musunuru, 2018; Lino et al., 2018). The introduction of a site-specific double-stranded break (DSB) by the highly specific and efficient RNA-guided Cas9 nuclease means a precise gene-edit can be introduced in the presence of a donor DNA template triggering DNA repair via homology directed repair (HDR) as opposed to non-homology end joining (NHEJ) (Courtney et al., 2016).

CRISPR-Cas9 has been shown to be effective at introducing single-base mutations particularly in the present of a donor vector containing a selection cassette. The presence of a selection cassette dramatically improves the screening process of identifying clones with the desired gene-edit and reduces the hands-on experimental time (Courtney et al., 2016; H. Li et al., 2020; J. Lin & Musunuru, 2018; Okamoto et al., 2019; G. Zhao et al., 2018). Traditionally, the *Cre/loxP* or Flp/*FRT* systems have been used alongside CRISPR-Cas9 to achieve precise gene modifications (van der Weyden et al., 2002). However, these systems are limited regarding the site in the genome they can be inserted into and the fact that single *loxP* or *FRT* sites are left in the genome after excision of the selection cassette (A. M. Singh et al., 2016; Yusa, 2013). The retention of these small redundant sequences may disturb functional elements and affect the investigated phenotype making it difficult to determine whether the observed effects are due to the redundant sequences or the gene-edit itself (Meier et al., 2010).

An alternative method for removing selection cassettes from the genome is through the use of a transposon containing the cassette. Two main transposable elements exist; retrotransposons and DNA transposons. Retrotransposons employ a 'copy-and-paste' mechanism by which a RNA-mediated intermediate is reverse transcribed into single stranded cDNA molecule before being converted into a double-stranded molecule and integrated into the genome (Tipanee et al., 2017). Retrotransposons actively spread through the human genome and have been used to enable transgene insertion (Muñoz-López & García-Pérez, 2010). DNA transposons employ a 'cut-and-paste' mechanism where two inverted terminal repeat (ITRs) either side of the transposon are recognized and cleaved by a transposase enzyme leaving free sticky DNA ends (Ivics et al., 2009). The transposon is then integrated into the new genomic region at a particular site recognised by the same transposase (Tipanee et al., 2017). DNA transposons are simply organized and several systems, including the *piggyBac^{TM},* Sleeping Beauty (SB), *Tol2* and *Frog Prince,* have been developed for insertional, precise, germline mutagenesis (Ivics et al., 2009; Wen et al., 2017; Woodard & Wilson, 2015). The *piggyBac^{TM}* transposon, originally isolated from the cabbage looper moth *Trichoplusia ni*, has several advantages over the latter including higher transposition efficiency, a tolerance for cargo of any size, no evidence for local hopping, a preference to land nearby the donor sequence and transposon removal without leaving redundant sequences (Ivics et al., 2009; M. A. Li et al., 2013; Liang et al., 2009). The *piggyBac^{TM}* system relies upon the identification of an endogenous 'TTAA' site within the genome for insertion of the transposon (Ding et al., 2005). These sites occur regularly (around every 246 bp) enabling the potential for precise genetic modification anywhere in the genome (Yusa, 2013). The *piggyBac*^{TM} system also enables positive and negative drug-selection based screening of genetic modifications to determine whether the transposon has been successfully integrated and excised from the genome which makes it easier and simpler to obtain targeted clones, particularly heterozygotes (X. Li et al., 2013; A. M. Singh et al., 2015). The utilisation of the *piggyBac*^{TM} transposon system enables the precise and scarless modification of the human genome and has been shown to be a flexible and efficient approach at achieving heterozygous or homozygous gene-editing (S. Liu et al., 2018; Woodard & Wilson, 2015; Fei Xie et al., 2014; Yusa et al., 2011; G. Zhao et al., 2018)**.** It therefore has great potential for introducing desired genetic changes within the genomic DNA to aid functional validation of causal variants and candidate genes.

In **Chapter 3,** I identified a genetic variant lying in the promoter region of the *EPO* gene and used the variant as a partial proxy for therapeutic rises in endogenous EPO levels. I found the A-allele to be associated with higher circulating EPO levels supporting previous studies that have investigated the allele-specific effects (Amanzada et al., 2014; Tong et al., 2008). However, I was limited in my analysis as the *cis-EPO* variant was a relatively weak genetic instrument due to not reaching formal levels of genome-wide significance ($P < 5 \times 10^{-08}$), was only an eQTL for hepatic gene expression and did not show the strongest evidence for colocalisation (Posterior probability = 71%). Therefore, I aimed to use a molecular approach of CRISPR-Cas9 targeted gene-editing with the *piggyBac*^{TM} transposon system to establish human cells with the *cis-EPO* variant to further elucidate the direct effects of the *cis-EPO* variant on *EPO* expression levels and confirm that this variant can serve as a valid genetic proxy for assessing the therapeutic effects of higher endogenous EPO levels.

5.2 Chapter Aims

The primary aim of this Chapter was to functionally validate the *cis-EPO* variant (identified in **Chapter 3**) as causal in controlling *EPO* expression levels. To accomplish this, the specific aims were to:

1. Generate a CRISPR-Cas9 vector and homology directed repair *piggyBac*^{TM} vector targeting the *cis-EPO* variant to introduce the desired gene-edit into HEK-293 cells.
2. Isolate single-cells with the polymorphism at rs1617640.
3. Validate whether the *cis-EPO* genetic variant alters *EPO* mRNA expression levels and is associated with abnormal expression of similar genes identified by RNA-seq analysis of *EPO* knock-outs.

5.3 Methods

An overview of the protocol established to obtain a single-base gene-edit at rs1617640 is outlined in **Figure 5.1.** This protocol was adapted from the one described by Yusa et al. (2013). Human Embryonic Kidney (HEK)-293 cells were used throughout following standard cell culture methods for growth and maintenance.

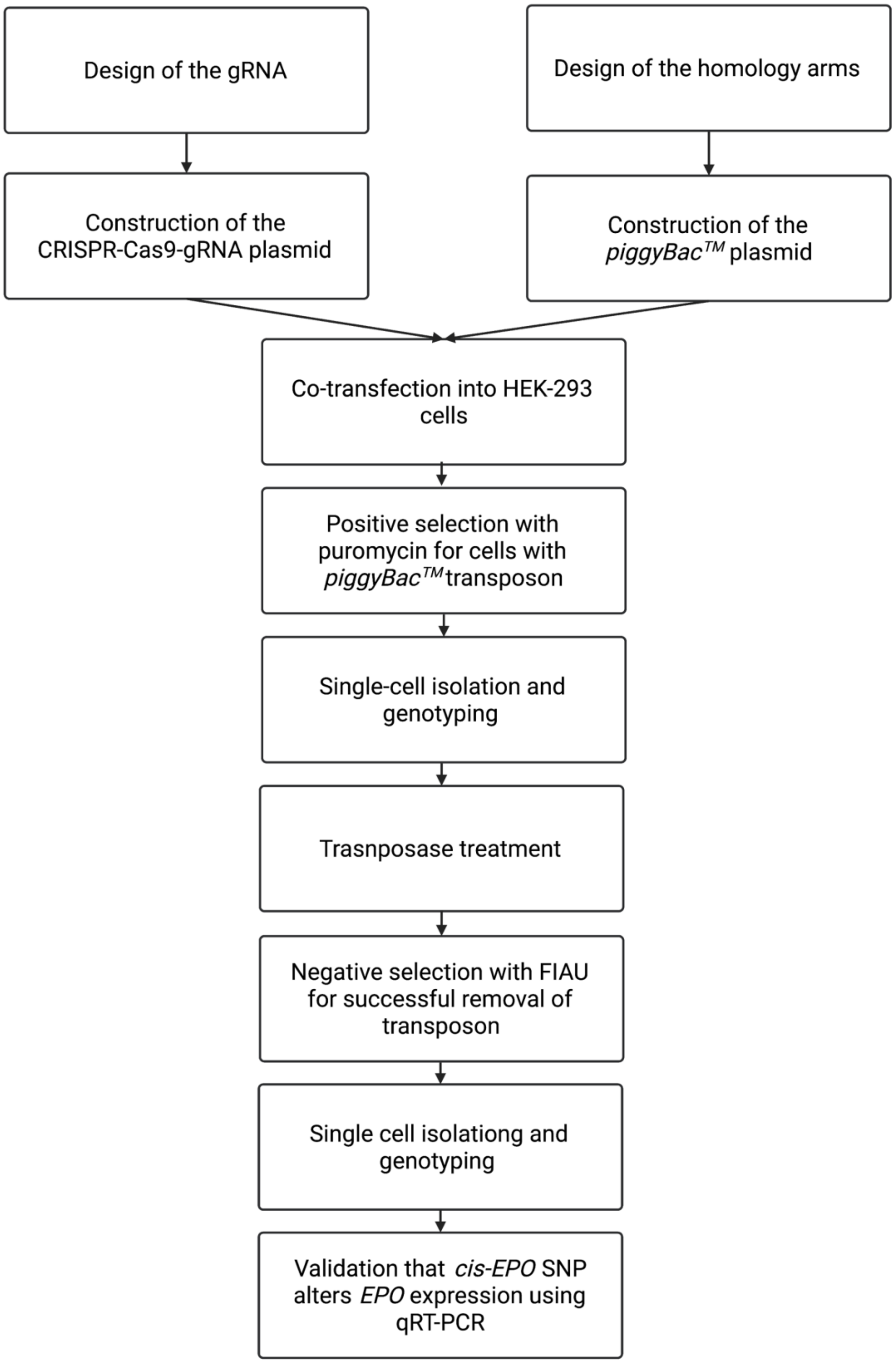
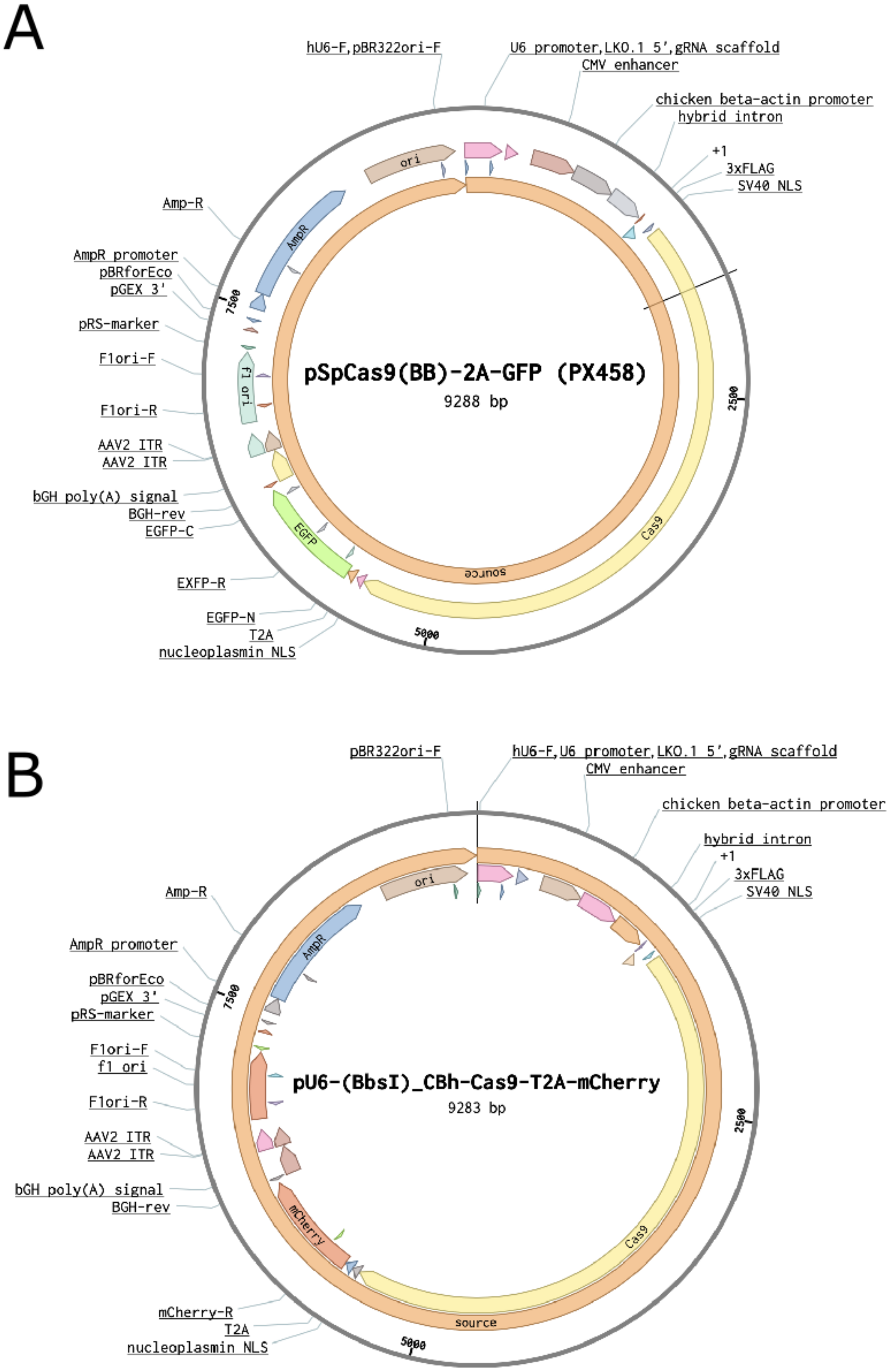

Design of the gRNA
Design of the homology arms
Construction of the CRISPR-Cas9-gRNA plasmid
Construction of the *piggyBac*™ plasmid
Co-transfection into HEK-293 cells
Positive selection with puromycin for cells with *piggyBac*™ transposon
Single-cell isolation and genotyping
Trasnposase treatment
Negative selection with FIAU for successful removal of transposon
Single cell isolationg and genotyping
Validation that *cis-EPO* SNP alters *EPO* expression using qRT-PCR

Figure 5.1: Overview of the methodology to establish a heterozygous knock-in model of the cis-EPO genetic variant. *CRISPR-Cas9 gene-editing was employed with the piggyBac™ transposon providing the donor sequence containing the desired gene edit at rs1617640 (A > C). The following protocol was adapted and optimized for use in HEK-293 cells from Yusa et al. (2013). gRNA, guide RNA; FIAU, Fialuridine; qRT-PCR, quantatiative reverse transcription polymerase chain reaction.*

5.3.1 Plasmids

The same commercially available genomic CRISPR-Cas9 plasmid containing green fluorescent marker (pSpCas9(BB)-2A-GFP (PX458), Addgene: #48138) used in the generation of the whole *EPO* gene knock-out in **Chapter 4** (**4.3.1)** was used (**Figure 4.4** **A)**.



This plasmid contained resistance to ampicillin for use in cloning and was first described in Ran et al., (2013). The *piggyBac*^{TM} multivector plasmid (SGK:005, MV-PGK-Puro-TK) (Hera BioLabs, Kentucky, USA) was used to provide the homology directed repair template sequences (**Figure 5.2**). The *piggyBac*^{TM} multivector contains a drug-selection marker (*puro*Δtk) to enhance HR efficiency (X. Li et al., 2013; S. Liu et al., 2018; Woodard & Wilson, 2015). To enable bacterial propagation, the *piggyBac*^{TM} transposase was cloned from the SBP-002 PBx vector (Hera BioLabs, Kentucky, USA) into the pUC19 vector (Addgene: #50005) using the BamHI and HindIII restriction sites.

5.3.2 Determining the genotype of wild-type HEK-293 cells

HEK-293 cells were maintained in Gibco's Dulbecco's Modified Eagle Medium (DMEM) (ThermoFisher Scientific, Massachusetts, USA) supplemented with 10% fetal bovine serum (FBS) (ThermoFisher Scientific, Massachusetts, USA) at 37 ^{0}C and 5% CO_2. When at 90% confluency, cells were pelleted and DNA was extracted using the PureLink Genomic DNA Extraction Kit following manufacturer's instructions (Invitrogen, Massachusetts, USA). To determine the wild-type genotype of rs1617640 in HEK-293 cells, primers (epo_snp-forward [5'-3']; CTGAATGGGATAGGCTGGTAGT, epo_snp-reverse [5'-3']; ATGGGGGCAAATAGGGCAAG) were designed either side of rs1617640 (**Figure 5.3).** PCR was performed using HOT FIREPol DNA polymerase following manufacturer's protocol (Solis BioDyne, Teaduspargi, Estonia). 10 μl of subsequent PCR product was analysed by gel electrophoresis on a 1.5% agarose gel. The remaining PCR product was purified using the ExoSAP-IT PCR Product Cleanup reagent following manufacturer's instructions (ThermoFisher Scientific, Massachusetts, USA) before being sent for Sanger sequencing using Genewiz (Genewiz, Essex, UK) with the forward primer (5'-CTGAATGGGATAGGCTGGTAGT-3') to determine the genotype of rs1617640 in HEK-293 cells.

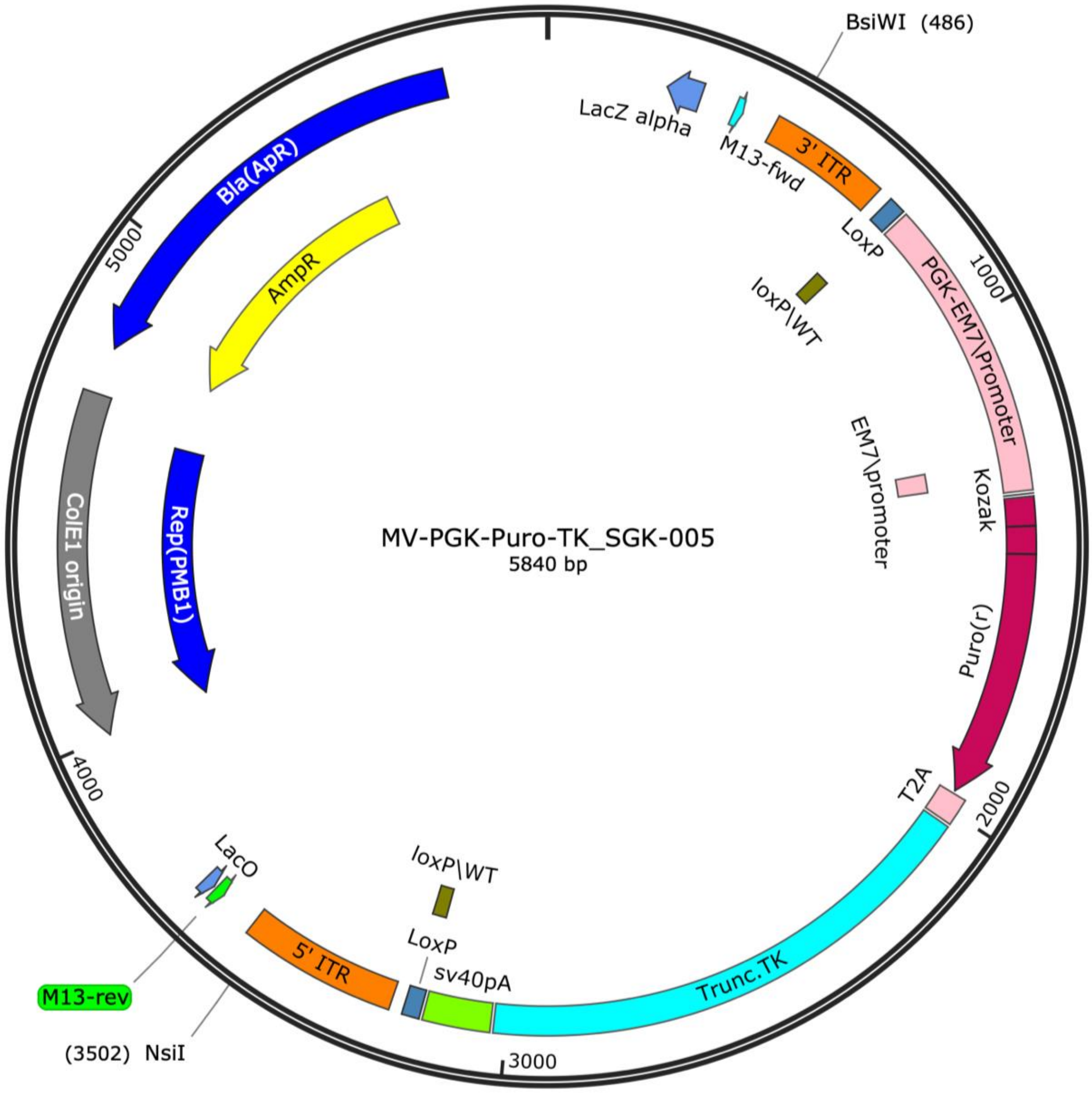


Figure 5.2: Plasmid map of the backbone of the piggyBac^{TM} multivector. *The plasmid backbone contains the piggyBac transposon sequences (5'ITR and 3'ITR) flanking the selection cassette which contains the tk gene conferring sensitivity to FIAU and the puromycin resistance gene conferring resistance to puromycin. Image was produced using SnapGene Viewer available at snapgene.com.*

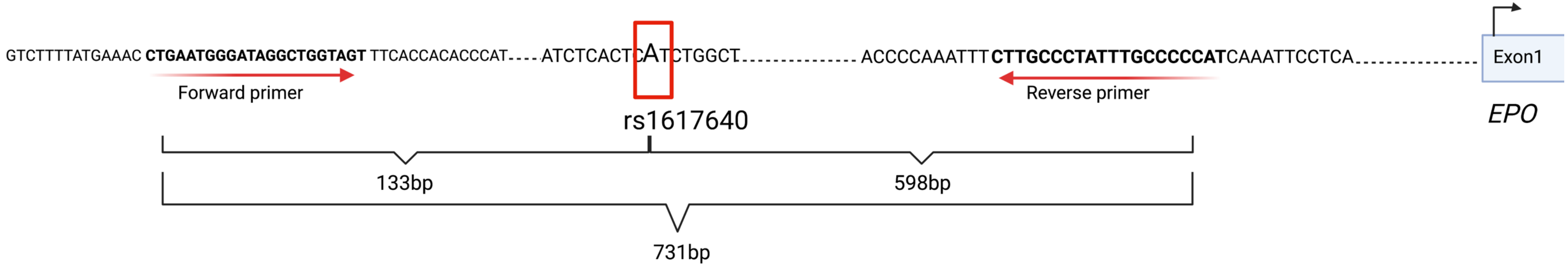

GTCTTTTATGAAAC CTGAATGGGATAGGCTGGTAGT TTCACCACACCCAT----ATCTCACTCATCTGGCT---------------ACCCCAAATTT CTTGCCCTATTTGCCCCCATCAAATTCCTCA-----------------
Forward primer
rs1617640
Reverse primer
Exon1
EPO
133bp
598bp
731bp

Figure 5.3: Schematic of the genomic location of the cis-EPO SNP. *The cis-EPO variant (rs16167640; chr7:10031729) lies 1,127 bp upstream of the transcription start site of the EPO gene. The position of the primers for the genotyping of wild-type HEK-293 cells have been highlighted in bold and depicted by the red arrows. PCR was performed using the two primers resulting in an amplicon of 731 bp. The resulting PCR amplicon was purified and sent for Sanger sequencing using the forward primer.*

5.3.3 Construction of the CRISPR-Cas9 plasmid targeting rs1617640

5.3.3.1 *Design of gRNA*

The online CRISPR design tool (https://www.benchling.com/crispr/) was used to design a gRNA sequence targeting the *cis-EPO* variant (Benchling [Biology Software], 2021). For SNP knock-in experiments, previous studies have recommended designing a gRNA around 10 bp away from the desired gene-edit location to maximise efficiency (Yusa, 2013; Yusa et al., 2011). A 20 bp gRNA sequence within the vicinity of rs1617640 (cut-site 6 bp from rs1617640) with the highest off-target (> 50) and on-target score (> 50) and least matches with other genomic loci through a BLAST (Boratyn et al., 2013) search was chosen. Overhangs complementary to the BbsI restriction enzyme cut-site were added (**Table 5.1**) to enable cloning into the CRISPR-Cas9 plasmid backbone. The final gRNA sequence (**Table 5.1**) was ordered through IDT (Integrated DNA Technologies, Leuven, Belgium; https://eu.idtdna.com/). The gRNA sequences were annealed and phosphorylated using T4 Polynucleotide Kinase (PNK) (New England BioLabs, Ipswich, UK) and ligated into the CRISPR-Cas9 plasmid in a single digestion and ligation reaction with an insert:plasmid ratio of 3:1 (calculated using the New England BioLab calculator - https://nebiocalculator.neb.com/#!/ligation) using the BbsI restriction enzyme and T4 ligase (New England BioLabs, Ipswich, UK).

5.3.3.2 *Cloning of recombinant plasmids in bacteria*

Ligated plasmids were transformed into DH5α E. coli and overnight cultures of single colonies were grown as described in the General Methods (**Chapter 2**) and Section:**4.3.2**. Plasmid DNA was purified and isolated from overnight cultures following the QIAprep Spin Miniprep Kit protocol (Qiagen, Maryland, USA).

5.3.3.3 *Screening of the plasmid DNA for successful integration of the gRNA*

To confirm successful integration of the gRNA into the plasmid, a double diagnostic digest was used (**Figure 5.4A**). Purified plasmid DNA was digested with BbsI and EcoRI restriction enzymes using the FastDigest Green Buffer (ThermoFisher Scientific, Massachusetts, USA). Digested products were visualised by gel electrophoresis (**Figure 5.4B**) and positive plasmids were sent for Sanger sequencing using the LKO.1 forward primer (5’-GACTATCATATGCTTACCGT-3’) to confirm insertion of the gRNA in the correct location and orientation (**Figure 5.4C).**

Table 5.1: The design the of gRNA sequence for targeting the cis-EPO variant. *Once the closest raw gRNA sequence to the cis-EPO variant with the highest off-target and on-target scores and least matches to other genomic loci has been identified, overhangs complementary to the BbsI restriction enzyme cut site (shown in **bold**) were added to the ends of the forward and reverse gRNA sequence to enable cloning of the gRNA into the CRISPR-Cas9 plasmid (pSpCas9(BB)-2A-GFP (PX458), Addgene: #48138).*

Steps	**Sequence**
Raw gRNA sequence (6bp away from *cis-EPO* variant)	5’ – GGAATCTCACTCCTCTGGCTCA**GGG** – 3’
Calculate reverse complement of raw gRNA sequence without PAM	5’ – TGAGCCAGAGGAGTGAGATTCC - 3’
Add BbsI overhangs	5’ - **CACC**GGAATCTCACTCCTCTGGCTCA - 3’ 5’ - **AAAC**TGAGCCAGAGGAGTGAGATTCC - 3’
Final oligo sequences for ordering	5’ - **CACC**GGAATCTCACTCCTCTGGCTCA - 3’ 3’- CCTTAGAGTGAGGAGACCGAGT**CAAA** - 5’

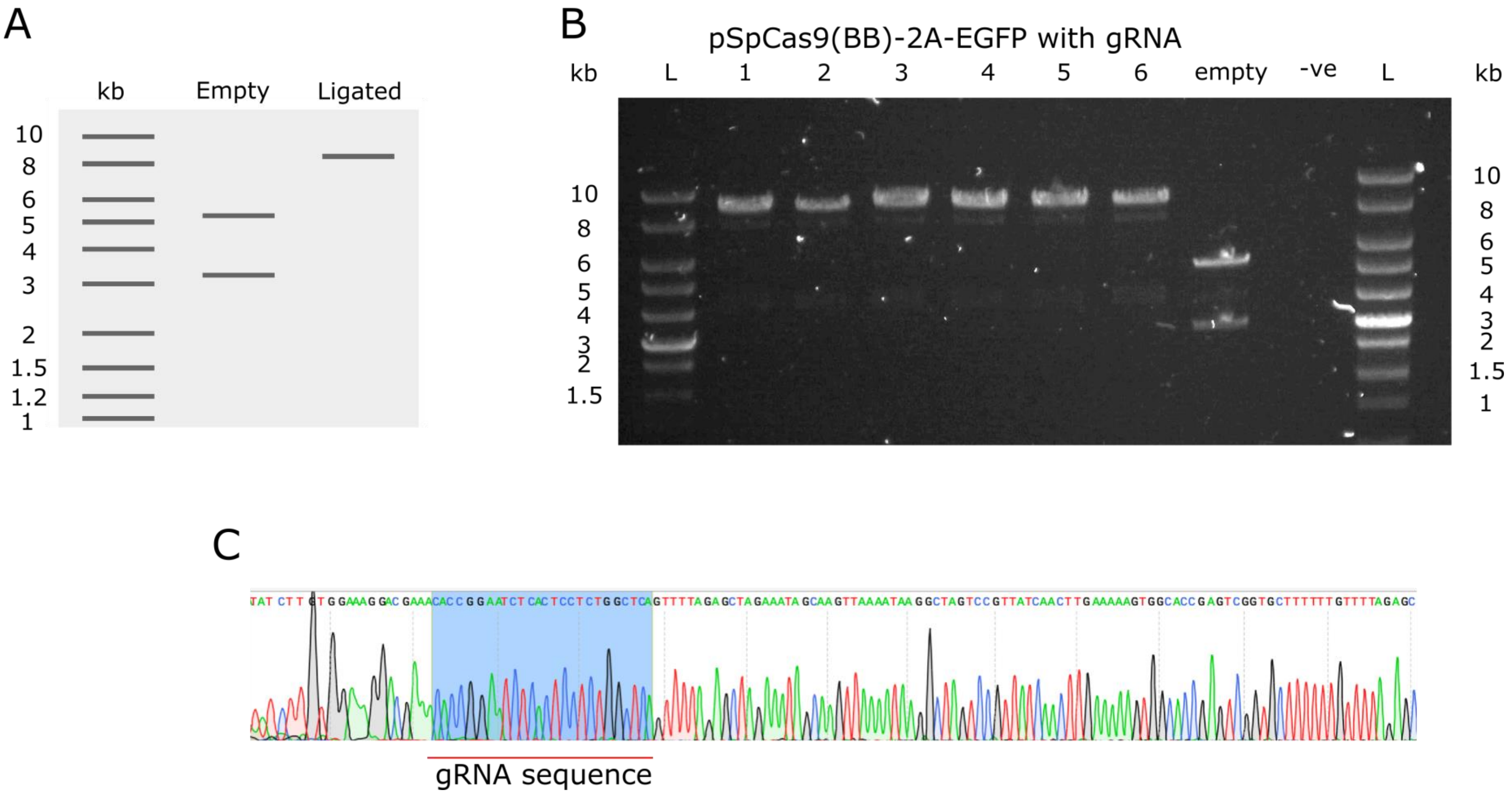


Figure 5.4: Construction of the CRISPR-Cas9 plasmid with the gRNA targeting the cis-EPO SNP. A: *Schematic showing the expected band size after a double diagnostic digest with BbsI and EcoRI restriction enzymes to confirm the ligation of the gRNA into the CRISPR-Cas9 plasmid. Lane 1 = NEB 2-log DNA ladder; Lane 2 = Empty, pSpCas9(BB)-2A-GFP CRISPR-Cas9 backbone vector with no gRNA inserted; Lane 3 = pSpCas9(BB)-2A-GFP CRISPR-Cas9 plasmid with gRNA inserted.* ***B:*** *Gel electrophoresis image of a double diagnostic restriction enzyme digest with BbsI and EcoRI. Lane L = Solis Biodyne 1 kb ladder; Lanes 1-6 = pSpCas9(BB)-2A-GFP CRISPR-Cas9 plasmid with gRNA targeting cis-EPO; Lane 7 = Empty, pSpCas9(BB)-2A-GFP CRISPR-Cas9 backbone vector with no gRNA inserted as digestion control; Lane 8 = negative control where water replaces the plasmid in the digestion reaction.* ***C:*** *Sanger sequencing confirmed the insertion of the correct gRNA sequence (highlighted in blue) in the correct location and orientation within the pSpCas9(BB)-2A-GFP CRISPR-Cas9 plasmid. Sequencing was performed using the LKO.1 primer.*

5.3.4 Design and Construction of the *piggyBac^{TM}* targeting vector

An outline of the steps taken to construct the *piggyBac^{TM}* expression vector targeting the *cis-EPO* SNP can be seen in **Figure 5.5.**

5.3.4.1 *Design of homology arms*

To generate the rs1617640 *piggyBac^{TM}* targeting vector, 500 bp of sequence either side of the TTAA closest to rs1617640 were identified. The homology sequence upstream of the TTAA contained the *cis-EPO* genetic variant and therefore included the desired polymorphism. To achieve footprint free excision, the BsiWI (within the 3'ITR of the *piggyBac^{TM}* transposon) and the Nsi1 (within the 5'ITR of the *piggyBac^{TM}* transposon) restriction enzymes were used when cloning the homology arms into the *piggyBac^{TM}* targeting vector*.* Using these restriction enzymes for cloning meant the homology arms had to include the remainder of the ITRs; the remainder of the 3'ITR sequence after the cut-site was added to the 3' end of the 5' homology arm and the remainder of the 5'ITR sequence after the cut-site was added to the 5' end of the 3' homology arm. Gibson cloning, outlined in **Figure 5.6,** was used to insert the homology arms into the *piggyBac^{TM}* holding vector and therefore sequences complementary to the plasmid also had to be added to the homology arms. For the 5' homology arm (which contained the desired SNP change at rs1617640, c.-1306 A>C), 20 bp of sequence complementary to the plasmid directly before the BsiWI cut-site was added to 5' end followed by the cut-site overhang. 20 bp of sequence complementary to the plasmid directly after the cut-site was added to the 3' end of the arm **(Figure 5.7A**). An additional 5' homology arm containing the wild-type sequence (A-allele at rs1617640) was also created for use as a control that had undergone the same experimental procedures as the potential knock-ins **(Figure 5.7B**). For the 3' homology arm, 20 bp of sequence complementary to the plasmid directly before the NsiI cut-site was added to the 5' end followed by the cut-site overhang to enable annealing into the plasmid. 20 bp of sequence complementary to the plasmid directly after the cut-site was added to the 3' end of the arm **(Figure 5.7C**). These sequences were ordered as MiniGene sequences directly from IDT (Integrated DNA Technology, Leuven, Belgium; https://eu.idtdna.com/). On arrival, the MiniGene homology arm sequences are held within pUC-IDT holding vectors and needed amplifying out of the holding vector via PCR before they can subsequently be cloned into the *piggyBac^{TM}*

plasmid. To do this, PCR primers were designed **(Table 5.2)** and PCR was performed using HOT FIREPol DNA polymerase (Solis BioDyne, Teaduspargi, Estonia). PCR products were purified using ExoSAP-IT PCR Product Cleanup reagent following manufacturer’s instructions (ThermoFisher Scientific, Massachusetts, USA). The purified PCR products were sent for Sanger sequencing (Genewiz, Essex, UK) to confirm the resulting homology arms were the expected sequence and that the 5’ homology arm contained the desired SNP edit (c.-1306A>C).

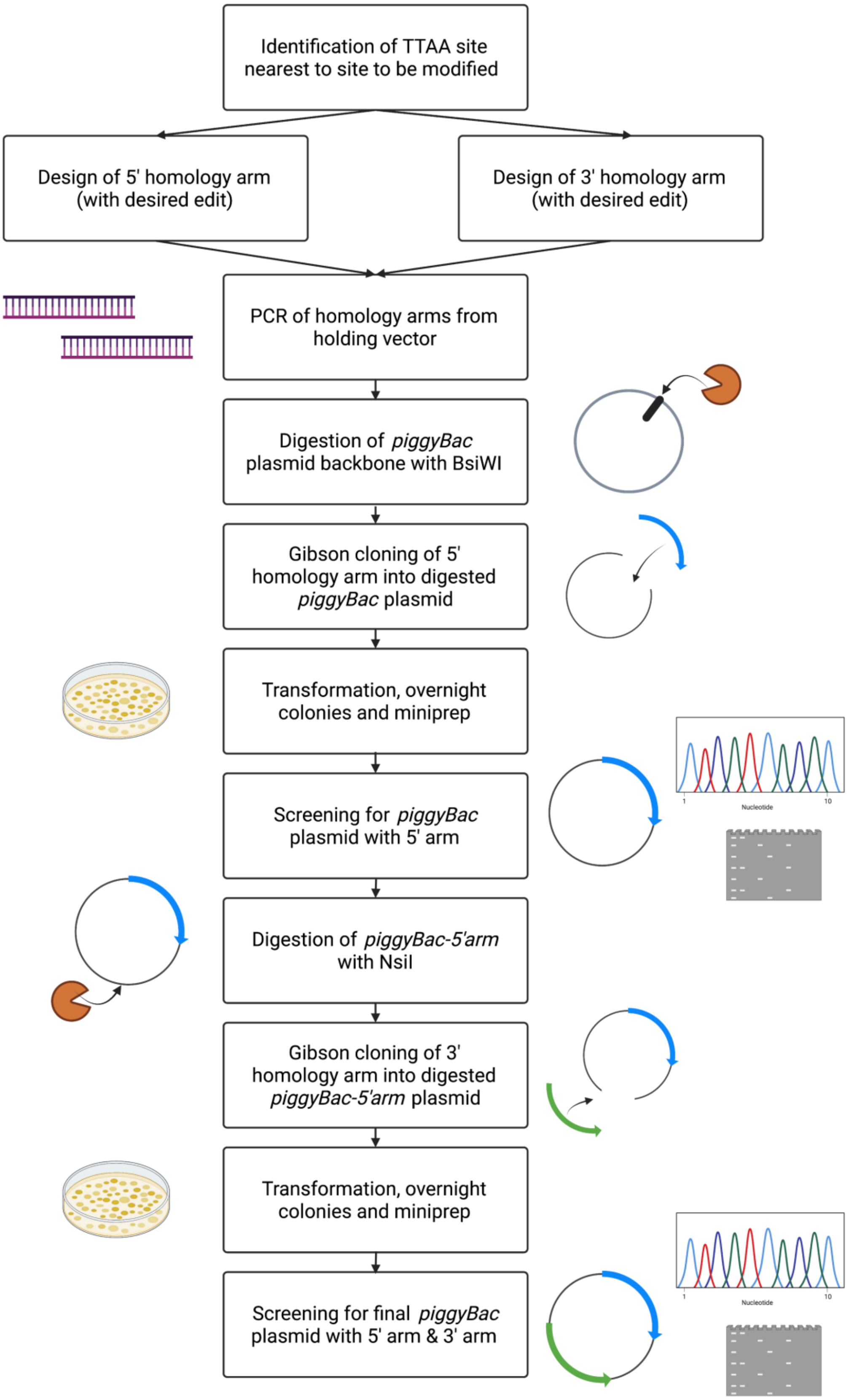

Identification of TTAA site nearest to site to be modified
Design of 5' homology arm (with desired edit)
Design of 3' homology arm (with desired edit)
PCR of homology arms from holding vector
Digestion of *piggyBac* plasmid backbone with BsiWI
Gibson cloning of 5' homology arm into digested *piggyBac* plasmid
Transformation, overnight colonies and miniprep
Screening for *piggyBac* plasmid with 5' arm
1
Nucleotide
10
Digestion of *piggyBac-5'arm* with NsiI
Gibson cloning of 3' homology arm into digested *piggyBac-5'arm* plasmid
Transformation, overnight colonies and miniprep
Screening for final *piggyBac* plasmid with 5' arm & 3' arm
1
Nucleotide
10

Figure 5.5: Experimental steps to construct the final piggyBac^{TM} targeting vector.

Homology arms either side of the nearest TTAA site to the desired gene-edit location were designed. The 5' homology arm contained the desired gene-edit. These arms were cloned into the piggyBac^{TM} holding vector either side of the piggyBac^{TM} transposon sequentially using Gibson Cloning. Restriction enzyme digests and PCR were used to determine that the arms had been successfully integrated into the plasmid in the correct location and orientation.

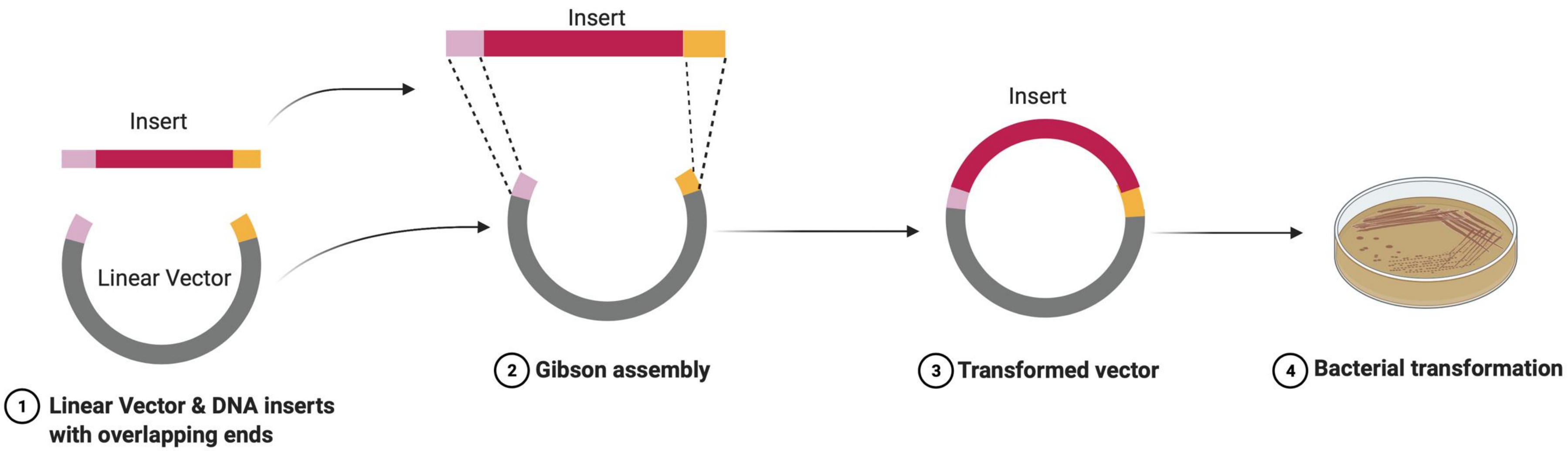

Insert
Linear Vector
Insert
Insert
1 Linear Vector & DNA inserts with overlapping ends
2 Gibson assembly
3 Transformed vector
4 Bacterial transformation

Figure 5.6: Gibson Cloning Assembly Mechanism. *A vector is linearised through a restriction enzyme digest. At least 20 bp of complementary sequence is added to either end of the sequence fragment for insertion. Gibson cloning is performed by incubating the insert, the linearised vector and Gibson Cloning Master Mix for 15 minutes at 50 ^{0}C. The resulting vector containing the desired insert is transformed into bacteria. Overnight colonies are selected and plasmids purified before being screened for correct ligation of the insert into the backbone vector. Created with BioRender.com.*

A

5’ homology arm sequence for SNP change at rs1617640 (C/C)

ATAATCATATTGTGACGTACTATTTATTTATTTATTTTAGAGACAAGGTCTTGCCATGTTGTCCGGGCTGGTCTCGAACTCCTGGGCTCAAA
GGATCTTCCTGCCTTGGTCTCCCAAAGTGCTGGGATTATAGGTGTCAGCTGCGGCGCCTGGACCTTTCCTGTCTTTTATGAAACCTGAAT
GGGATAGGCTGGTAGTTTCACCACACCCATTTGACAGATGAGGACATTGAGGGCTCAAGGACGAGGCCACTTTCTAAGGTGTGAGAGAC
CAGCTAGTCTTGGTCTCCTGCTCTGGGAATCTCACTCCTCTGGCTCAGGGTTTCCAGAAGCCATAAAACCTTAGCTGTAAATCCCAGCCC
CCATCACTCTTGGTGTTAGCTGTATTTCAGTGTTCTTAA**CCCTAGAAAGATAATCATATTGTGAC**GTACGTTAAAGATAATCATG

B

5’ homology arm sequence for wild type (WT) sequence at rs1617640 (A/A)

Figure 5.7: Homology sequences designed for the targeting of rs1617640 for insertion into the piggyBac^{TM} plasmid. ***A:*** *The 5' homology arm sequence containing the desired SNP change at rs1617640 (C – highlighted in red) for insertion upstream of the 3'ITR in the piggyBac^{TM} plasmid.* ***B:*** *The 5' homology arm sequence containing the HEK-293 WT sequence at rs1617640 (A – in red) for insertion upstream of the 3'ITR in the piggyBac^{TM} plasmid.* ***C:*** *The 3' homology arm sequence for insertion downstream of the 5'ITR in the piggyBac^{TM} plasmid. Underlined sequences represent the 20 bp sequence complementary to the piggyBac^{TM} multivector sequence to enable integration into the plasmid via Gibson Cloning. The bases highlighted in* ***bold*** *are the remainder of the 3'ITR or 5'ITR after the restriction cut-sites to enable seamless removal of the piggyBac^{TM} transposon from the DNA. The TTAA site is emphasised in blue and rs1617640 is highlighted in red.*

5.3.4.2 Cloning of the homology arms into the piggyBac^{TM} expression vector

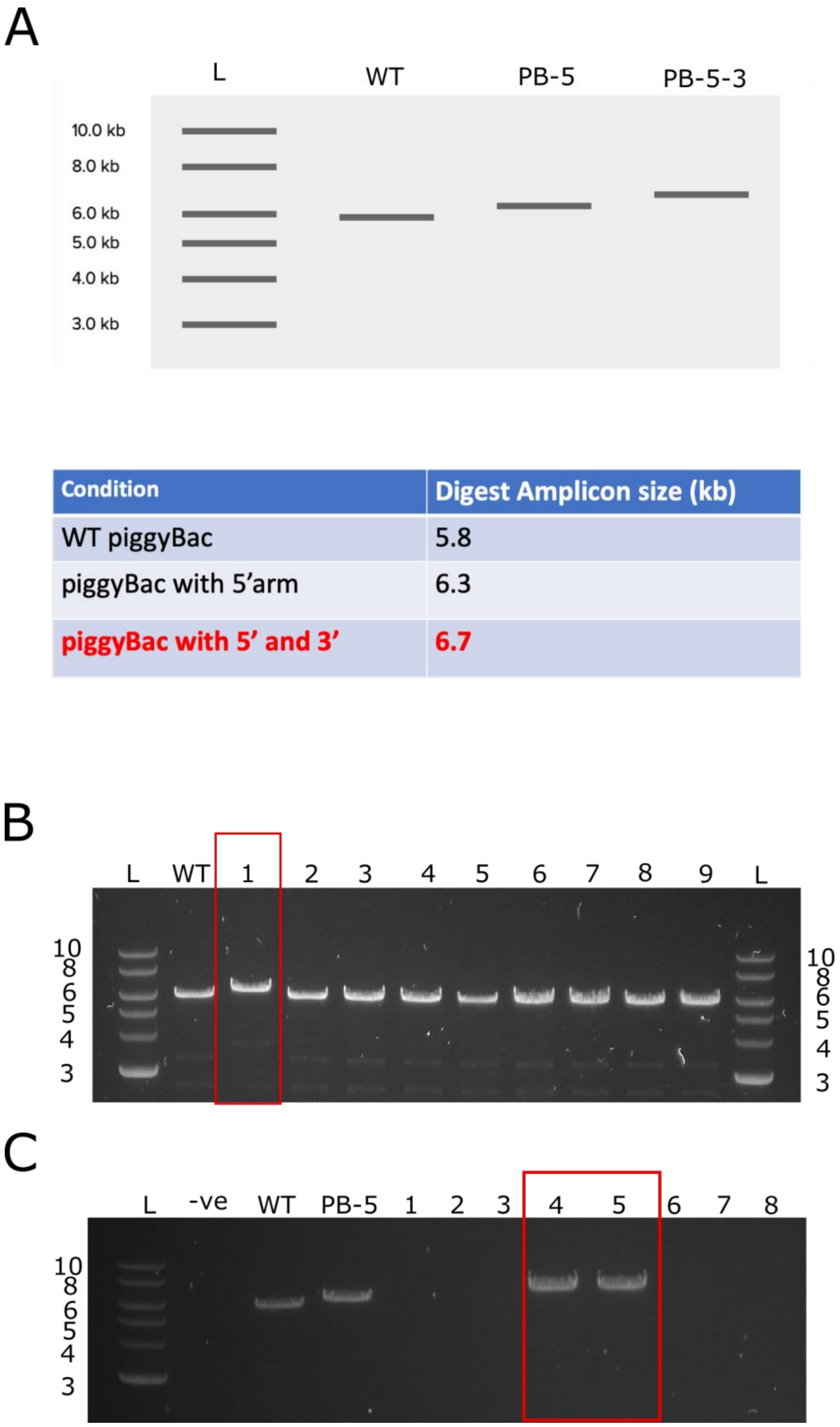
The Gibson Cloning Assembly Kit (New England BioLabs, Ipswich, UK) was used to clone the two homology arms into the *piggyBac^{TM}* expression vector (**Figure 5.6**). I first performed Gibson Cloning to clone the 5' homology arm into the *piggyBac^{TM}* vector. I digested 1000 ng of the *piggyBac^{TM}* plasmid backbone with BsiWI and then incubated the digested plasmid with 2-fold molar excess of 5' homology arm (calculated using the New England BioLab calculator, https://nebiocalculator.neb.com/#!/ligation) and 10 μl Gibson Master Mix in a total volume of 20 μl for 15 minutes at 50 ^{0}C. I then transformed 2 μl of the ligation reaction into NEB 5-alpha competent *E.coli* cells provided with the Gibson Assembly Cloning Kit following the NEB transformation protocol (New England BioLabs, Ipswich, UK). Overnight cultures of single colonies were set-up and plasmids were purified following the QIAprep Spin Miniprep Kit protocol (Qiagen, Maryland, USA). Plasmids containing the 5' homology arm were screened for using a diagnostic digest with BsiWI **(Figure 5.8)**. 1000 ng of plasmid containing the 5' homology arm was subsequently digested with NsiI and Gibson cloning was repeated with the 3' homology arm as described above to insert the 3' homology arm into the *piggyBac^{TM}* plasmid. Transformation, overnight colonies and plasmid purification was repeated as above. Plasmids containing both homology arms were screened for using a diagnostic digest with the NsiI restriction enzyme **(Figure 5.8)**. PCR was performed on positive plasmids containing both homology arms and plasmids were sent for Sanger sequencing to confirm 1) the insertion of the correct sequences either side of the *piggyBac^{TM}* selection cassette and 2) the presence of the desired SNP change at rs1617640 (c.-1306, A>C) in the 5' homology arm (**Figure 5.9**). For primer sequences see **Table 5.3**.

Table 5.2: Primer sequences used for PCR amplification of the 5' and 3' homology arms from the holding vector.

. Primer	Sequence (5'-3')
5' arm forward	ATAATCATATTGTGACGTAC
5' arm reverse	CATGATTATCTTTAACGTAC
3' arm forward	AGCAATATTTCAAGAATGCA
3' arm reverse	CGTAAAATTGACGCATGCAT

Table 5.3: Primer sequences used to confirm the insertion of the homology arm sequences into the correct location in the piggyBac^{TM} plasmid.

Primer	Purpose	Sequence (5'-3')
M13_fwd	Screening for insertion of 5'arm in correct location by PCR & sequencing	TGTAAAACGACGGCCAGT
5arm_3ITR_rev	Screening for insertion of 5'arm in correct location by PCR	CGTCAATTTTACGCATGATTATCTTTAAC
3arm_5ITR_fwd	Screening for insertion of 3' arm in correct location by PCR	GCGACGGATTCGCGCTATTTAGAAAG
M13_rev	Screening for insertion of 3' arm in correct location by PCR & Sequencing	CAGGAAACAGCTATGACCATG



Condition	Digest Amplicon size (kb)
WT piggyBac	5.8
piggyBac with 5'arm	6.3
piggyBac with 5' and 3'	6.7

Figure 5.8: Diagnostic digest to confirm the insertion of the homology arms into the piggyBac^{TM} plasmid. ***A:*** *Virtual diagnostic digest highlighting the expected fragment size following digestion with BsiWI. L= NEB 2-log, WT = empty piggyBac^{TM} vector, PB-5 = piggyBac^{TM} plasmid with the 5'homology arm inserted. PB-5-3 = piggyBac^{TM} plasmid with the 5' homology arm and 3' homology arm inserted.* ***B:*** *Diagnostic digest with BsiWI to identify plasmids with the 5'-homology arm inserted. Plasmid 1 had the 5'-homology arm successfully inserted and was taken forward to insert the 3' homology arm. L=Solis Biodyne 1 kb ladder, WT = empty piggyBac^{TM} plasmid.* ***C:*** *Diagnostic digest with BsiWI to identify plasmids with the 5' and the 3' homology arms inserted. Plasmids 4 & 5 appeared to have both arms inserted and were sent for sequencing. L= Solis Biodyne 1 kb ladder, WT=empty piggyBac^{TM} plasmid, PB-5= piggyBac^{TM} plasmid with 5'arm*

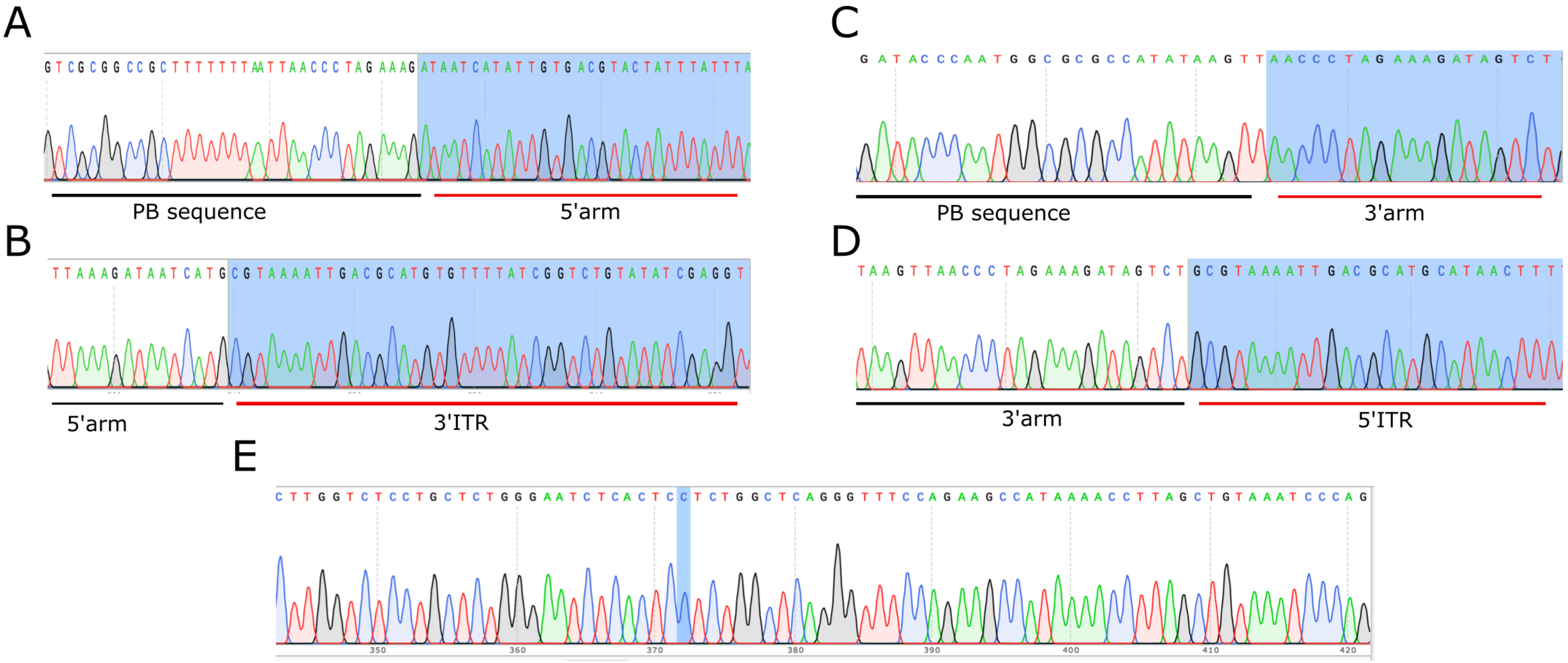


Figure 5.9: Sanger sequencing to confirm the presence of the correct homology arm sequences in the correct location and orientation in the piggyBac^{TM} transposon plasmid. A-B: *The 5' homology arm was correctly inserted into the correct location upstream of the 3'ITR in the piggyBac^{TM} plasmid* ***C-D:*** *The 3' homology arm was correctly inserted into the correct location downstream of the 5' ITR in the piggyBac^{TM} plasmid.* ***E****: Sanger sequencing confirmed the desired SNP-edit at rs1617640 (A>C) was present in the 5' homology arm that was inserted into the piggyBac^{TM} plasmid.*

5.3.5 Genetic modification of the *cis-EPO* SNP in HEK-293 cells with CRISPR-Cas9 gene-editing

5.3.5.1 *Co-transfection of HEK-293 cells*

Prior to gene-targeting, HEK-293 cells were seeded in a 10-cm dish at a density of 100,000 cells per mL. 24 hours later, cells were co-transfected with 6 μg of *piggyBac*^{TM} HDR template plasmid (containing the desired SNP edit or the wild-type sequence at rs1617640) and 6 μg of CRISPR-Cas9-gRNA plasmid using lipofectamine LTX following manufacturer's protocol (ThermoFisher Scientific, Massachusetts, USA). Cells were incubated for 48 hours before being visualised under the Leica DMi8 Widefield microscope (Leica, Milton Keynes, UK) to confirm successful transfection.

5.3.5.2 *Selection of cells containing piggyBac*^{TM} *transposon*

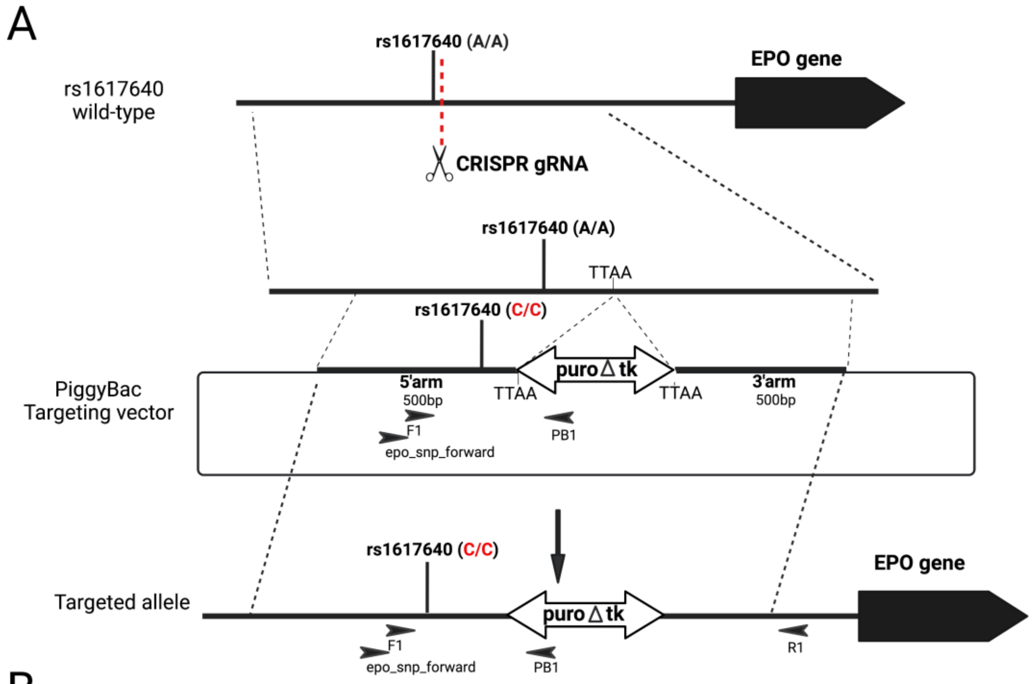
48 hours after transfection, cells were placed under 1 μg/mL puromycin selection (Sigma-Aldrich, Missouri, USA) for 14 days to select cells containing the *piggyBac*^{TM} selection cassette (Iwaki & Umemura, 2011; Yusa, 2013). Media was replaced every 2-3 days with fresh DMEM (ThermoFisher Scientific, Massachusetts, USA) supplemented with 10% FBS and 1 μg/mL puromycin. Following selection, single puromycin-resistant cells were isolated into 96-well plates via single cell picking using a 2 μl pipette under the EVOS FLoid Imaging system (ThermoFisher Scientific, Massachusetts, USA). Single cells were clonally expanded for around 2 weeks and gradually moved from 96-well plates to 24-well plates and then 6-well plates.

5.3.5.3 *Genotyping*

To confirm successful insertion of the *piggyBac*^{TM} transposon in the correct genomic location and correction of the SNP at rs1617640, clonally expanded cells were pelleted and DNA was extracted from half the pellet using the PureLink Genomic DNA Extraction Kit (Invitrogen, Massachusetts, USA). The other half of the pellet was plated for continual growth. DNA concentration was measured using the Nanodrop ND-8000 Spectrophotometer (ThermoFisher Scientific, Massachusetts, USA) and 100 ng DNA was subjected to genomic PCR using HOT FIREPol DNA polymerase according to manufacturer's instructions (Solis BioDyne, Teaduspargi, Estonia). To screen for successful homozygous targeting and integration of the *piggyBac*^{TM} transposon, two pairs of primers were designed (**Figure 5.10, Table 5.4**). PCR products were visualised using gel electrophoresis (**Figure 5.10**). PCR amplicons for the samples showing the expected banding pattern were purified using ExoSAP-IT PCR Product Cleanup reagent following manufacturer's instructions (ThermoFisher Scientific, Massachusetts, USA). Purified products were sent for Sanger sequencing using Genewiz (Genewiz, Essex, UK) with the forward primer (epo_snp-forward; **Table 5.4**) to determine the genotype at rs1617640 and check for insertion of the *piggyBac*^{TM} transposon in the correct location.

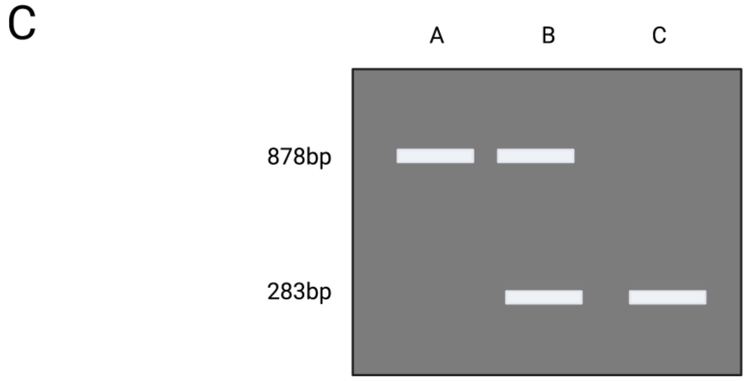
Table 5.4: Primer sequences used for screening correctly mutated clones at rs1617640.

Primer	**Sequence (5’-3’)**
F1	CTGGTAGTTTCACCACACCCA
R1	TTGGGCGGAGACTCAGAGAT
epo_snp_forward	CTGAATGGGATAGGCTGGTAGT
PB1	CGTCAATTTTACGCATGATTATCTTTAAC
F2	TGAGCCACCACACCTGACTA
R2	TTCTTCCTCCCCACCTCACT
PB-F	GGCATAGTATATCGGCATAG
PB-R	GTTAGAAGACTTCCTCTGC



B

Genotype	Primer combination	Amplicon size (bp)
Original allele	F1 - R1	878
Targeted allele (with PB inserted)	epo_snp_forward - PB1	283



Possible Genotypes
A: Two nontargeted alleles random integrations elsewhere
B: One allele targeted (heterozygous), one targeted and small indel in other allele
C: Two targeted alleles (homozygous)

Figure 5.10: PCR-based screening of targeted clones.

***A:** Schematic of gene targeting upon successful integration of transposon into genomic DNA. Arrowheads represent primers used in genotyping to screen for successful homozygous targeting. **B:** Primer combinations and expected amplicon sizes. Primer pair F1-R1 would result in a PCR amplicon of 878 bp if transposon has not been integrated. Primer pair F2-PB1 would result in a PCR amplicon of 283 bp upon successful transposon integration into the DNA at the correct location. **C:** PCR-based genotyping. Homozygote colonies with successful transposon integration at both alleles will have a single band at 283 bp whilst heterozygotes can be identified by the presence of two bands.*

5.3.6 Removal of *piggyBac*^{TM} transposon cassette

5.3.6.1 *Transposase treatment*

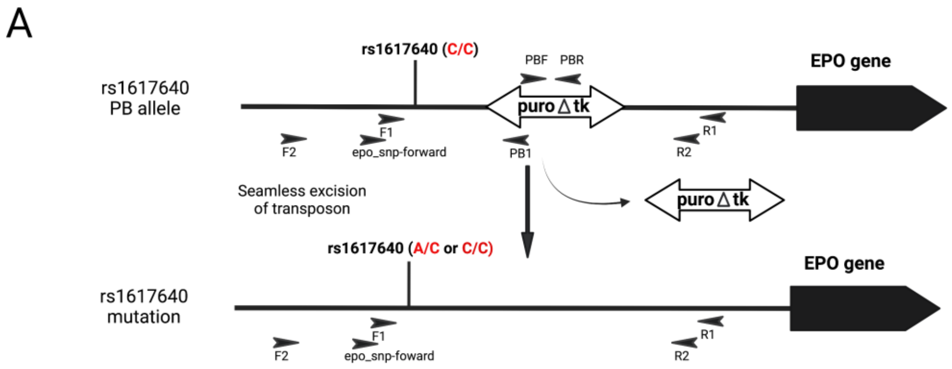
A biallelic Cas9-targeted clone containing the desired SNP edit at rs1617640 (c.-1306, A>C) was selected for treatment with transposase to remove the *piggyBac*^{TM} selection cassette from the genomic DNA. Cells were seeded at a density of 100,000 cells per ml in a 6-well plate and the following day were transfected with 2500 ng of the *piggyBac*^{TM} excision-only transposase using Lipofectamine LTX reagent following manufacturer's protocol (ThermoFisher Scientific, Massachusetts, USA).

5.3.6.2 *Selection of cells for successful excision of piggyBac*^{TM}

To select for cells no longer carrying the *piggyBac*^{TM} transposon cassette, 48 hours after transposase treatment, cells were subject to a second selection step under Fialuridine (FIAU) treatment (200 nM, Sigma-Aldrich, Missouri, USA) for 10 days. Media was replaced every 2-3 days with fresh DMEM supplemented with 10% FBS and 200 nM FIAU. Single cells were isolated by single cell picking using a 2 μl pipette under the EVOS FLoid Imaging system (ThermoFisher Scientific, Massachusetts, USA). Single cells were clonally expanded for around 2 weeks and gradually moved from 96-well plates to 24-well plates and then 6-well plates.

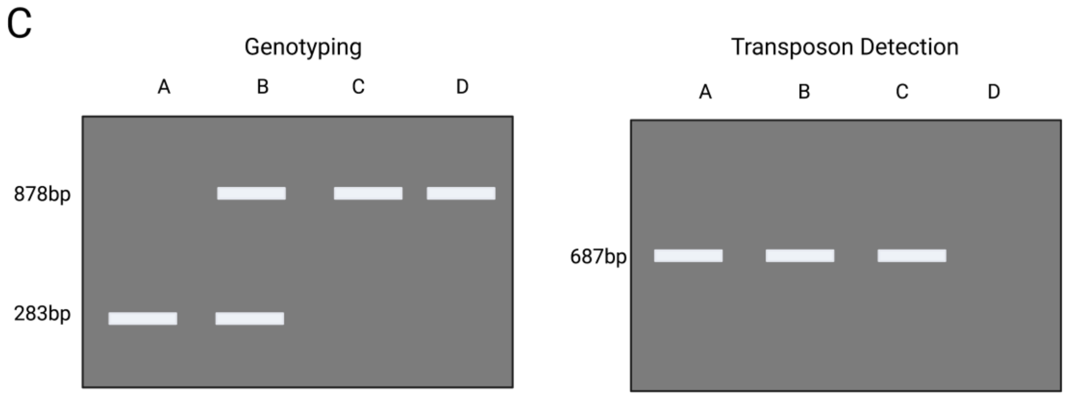
5.3.6.3 *Genotyping*

Single clones were propagated for genotyping by PCR analysis to determine successful transposon excision using the same two sets of primers as before for detecting transposon integration (**Table 5.4**; F1-R1; epo_snp-foward-PB1). An additional pair of primers were also used (**Table 5.4**; PBF-PBR; F2-R2) to check for reintegration of the *piggyBac*^{TM} cassette elsewhere in the genome (**Figure 5.11**). PCR amplicons showing the expected banding pattern were purified using ExoSAP-IT PCR Product Cleanup reagent following manufacturer's instructions (ThermoFisher Scientific, Massachusetts, USA) and were subsequently sent for Sanger sequencing using Genewiz (Genewiz, Essex, UK) to confirm seamless removal of the *piggyBac*^{TM} cassette from the genomic DNA and correct modification of the *cis-EPO* variant (**Table 5.4**; F1 primer). Successful clones were further propagated for downstream analysis.



B

Genotype	Primer combination	Amplicon size (bp)
Targeted allele (with PB inserted)	epo-snp_forward - PB1	283
Modified allele	F1 - R1	878
Modified allele	F2 - R2	974
Transposon detection	PBF - PBR	687



Possible Genotypes
A: Transposon remains in both alleles
B: Transposon remains in one allele
C: Transposon removed from both alleles but reintegration
D: Transposon removed from both alleles but no reintegration

Figure 5.11: PCR-based screening of transposon-excised clones. ***A:*** *Schematic of transposon excision. Arrowheads represent primers used in genotyping to determine whether transposon had successfully been removed.* ***B:*** *Primer combinations and expected amplicon sizes during genotyping. Primer pair F1-R1 would result in a PCR amplicon of 878 bp if transposon has been successfully excised. Primer pair F2-PB1 would result in a PCR amplicon of 283 bp upon unsuccessful transposon excision. PBF-PBR was used to check for reintegration and would result in amplicon size of 687 bp if transposon is still present anywhere in the genomic DNA* ***C:*** *Examples of PCR results for different genotypes. Homozygote colonies with successful transposon excision at both alleles and no reintegration will have a single band at 878 bp and no band present for PBF-PBR (Lane D). If the transposon has been reintegrated elsewhere in the genome in homozygote targeting, a band will be present at 687 bp (Lane C). Heterozygotes targeted only at one allele will have two bands present when genotyping and a band present when screening for transposon (Lane B).*

5.3.7 Validation that the *cis-EPO* variant alters *EPO* gene expression levels and downstream causal pathways

Isogenic HEK-293 cells, containing either the wild-type (A/A) or mutant (A/C) genotype at rs1617640, were subjected to quantitative reverse transcription PCR (qRT-PCR) to validate the *cis-EPO* variant as causal in controlling *EPO* messenger RNA (mRNA) expression levels. Cells were pelleted and RNA was isolated using the Direct-zol^{TM} RNA Miniprep kit following the manufacturer's protocol (Cambridge Biosciences, Cambridge, UK). 500 ng of RNA was converted to cDNA using PrimeScript^{TM} RT reagent kit (Takara Bio Europe SAS, Saint-Germain-en-Laye, France) following manufacturer's protocol. qRT-PCR was performed using Hot FIREPol EvaGreen^{TM} qPCR Master Mix with ROX (Solis BioDyne, Teaduspargi, Estonia) using the QuantStudio 6 Flex qPCR machine (ThermoFisher Scientific, Massachusetts, USA) on at least three biological replicates. Any samples with Ct values greater than 2 standard deviations (SD) from the mean were removed. Gene expression levels were standardised against the reference gene (*GAPDH*) mRNA levels using the $2^{-\Delta\Delta CT}$ method (Livak & Schmittgen, 2001). Genes involved in the Notch signalling pathway and identified as dysregulated through whole transcriptomic profiling of the *EPO* gene knock-outs in **Chapter 4** were also subjected to qRT-PCR to investigate whether the *cis-EPO* SNP has a similar effect to whole gene knock-out. All primer sequences used were the same as those in **Table 4.2** and **Table 4.5.**

5.3.8 Statistical analysis

All data are presented as the mean ± standard error (SE). Statistical analysis was performed using RStudio version 3.6.1 (RStudio Team, 2018). Comparisons between the two genotypes, wild-type and mutant, were analysed by paired Student's t-tests. Differences with $P < 0.05$ were considered statistically significant.

5.4 Results

5.4.1 HEK-293 cells are homozygous for the A-allele at rs1617640.

To determine the genotype of wild-type HEK-293 cells, I amplified the region surrounding the *cis-EPO* variant and sent the subsequent amplicon for Sanger sequencing. Sanger sequencing revealed that HEK-293 cells were homozygous for the A-allele at rs1617640 (**Figure 5.12).**

5.4.2 Construction of the CRISPR-Cas9 and *piggyBac^{TM}* vectors targeting the *cis-EPO* variant

To introduce the naturally occurring polymorphism of rs1617640, I adapted a molecular strategy, combining CRISPR-Cas9 and the *piggyBac^{TM}* system, to precisely generate a *cis-EPO* mutant cell-line model without leaving any marks in the targeted genomic DNA. Upon examination of the genomic sequence surrounding the *cis-EPO* variant, I designed a single gRNA targeting Cas9 to the DNA inducing a DSB 6 bp away from the *cis-EPO* variant (**Figure 5.13A**). I identified a 'TTAA' site 87 bp away which could be used for the successful integration and footprint free excision of the *piggyBac^{TM}* transposon (**Figure 5.13A**). The closest 'TTAA' site to rs1617640 was chosen as the efficiency of inserting the modification decreases as the distance between the desired modification site and 'TTAA' site increases; 80% of clones will carry the modification when the site is 200 bp away from the 'TTAA' site, whilst only 70% of clones will carry the modification when the edit-site is 300 bp away from the 'TTAA' site (Yusa, 2013). I designed two homology arms complementary to 500 bp either side of the identified 'TTAA' site. The homology sequence upstream of the 'TTAA' site (5' homology arm) contained the desired SNP change at rs1617640 (either wild-type A/A genotype [as a control] or mutant C/C genotype) (**Figure 5.13A**). The original donor vector contains two inverted terminal repeat (ITR) *piggyBac*^{TM} transposon sequences and a bi-functional hybrid *puroΔtk* gene for positive selection through puromycin treatment and negative selection through FIAU treatment. I designed a targeting vector carrying the homologous sequences inserted either side of the two ITRs flanking the *PGK-puroΔtk* cassette (**Figure 5.13B**). When the CRISPR-Cas9 plasmid introduces a DSB in the vicinity of rs1617640, the presence of the homology arms in the *piggyBac^{TM}* plasmid initiates homologous recombination and the *piggyBac*^{TM} cassette can be inserted precisely at the ‘TTAA’ site (**Figure 5.13C**). Due to the presence of the puromycin resistance gene, genome-edited cells can be selected with puromycin. After treatment with excision-only transposase, the *piggyBac*^{TM} transposon can be seamlessly excised from the genome (**Figure 5.13D**). As the genome no longer contains the *tk* gene (held in the *piggyBac*^{TM} transposon), cells can be selected for successful excision using FIAU.

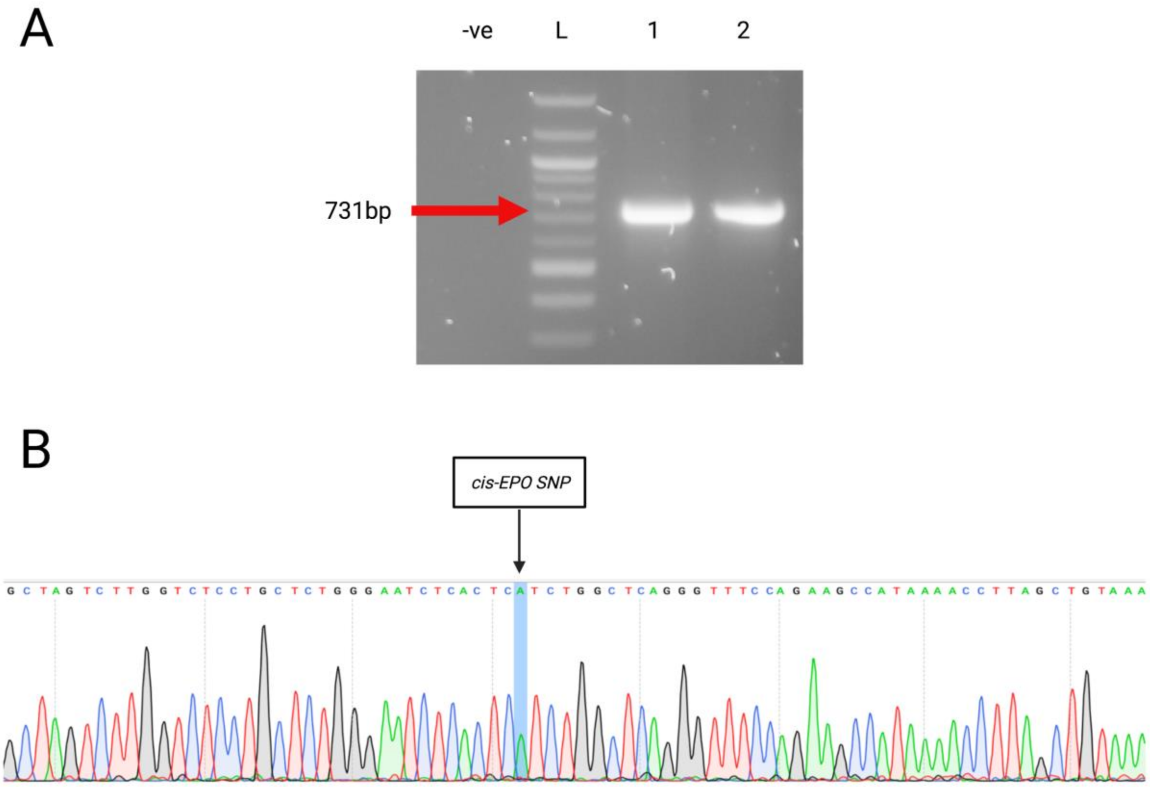


Figure 5.12: Genotyping of wild-type HEK-293 cells at the cis-EPO SNP. ***A:*** *A fragment of 731 bp was amplified surrounding the cis-EPO SNP****.*** ***B:*** *Sanger sequencing was performed on purified PCR product confirming that HEK-293 cells are homozygous for the A-allele at rs1617640.*

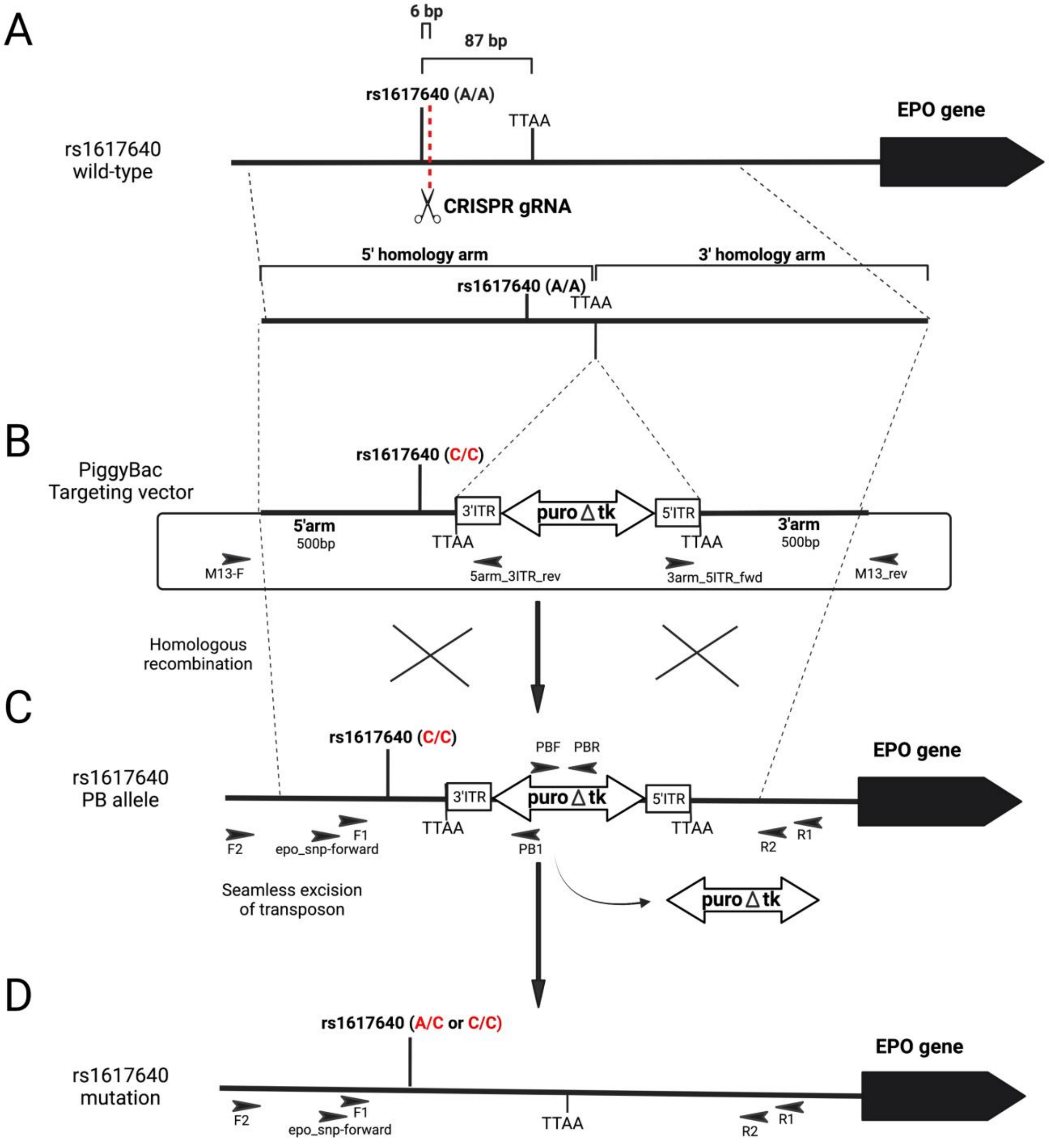


Figure 5.13: Schematic diagram of site-specific gene-editing of rs1617640 using CRISPR-Cas9 and the piggyBac^{TM} transposon system. ***A:*** *Targeting of the cis-EPO SNP upstream of the EPO gene. A gRNA was designed to introduce CRISPR-Cas9 mediated cleavage 6 bp downstream from the cis-EPO SNP. The nearest 'TTAA' site for integration of the piggybac^{TM} plasmid was 87 bp away.* ***B:*** *The piggyBac^{TM} targeting construct carrying the two homology arms (the 5' homology arm with the desired SNP-edit) flanking the selectable markers, puroΔtk, within the transposon.* ***C:*** *Insertion of the piggyBac^{TM} transposon following HR at the desired TTAA site nearby the cis-EPO SNP. Following puromycin selection, clones containing the piggyBac^{TM} transposon confirming successful HR were identified by PCR and sequencing using epo_snp_forward-PB1 and F1-R1.* ***D:*** *Transfection with excision-only transposase resulted in the seamless removal of the piggyBac^{TM} cassette. Clones were identified by PCR amplification and sequencing using epo_snp-forward-PB1, F1-R1 and F2 -R2.*

5.4.3 Site-specific homologous recombination CRISPR-Cas9-mediated gene-targeting of the *cis-EPO* SNP in HEK-293 cells

HEK-293 cells were co-transfected with the CRISPR-Cas9-gRNA plasmid and a targeting vector containing the *piggyBac*^{TM} selection cassette. Fluorescent microscopy imaging confirmed successful transfection (**Figure 5.14**). To confirm site-specific HR and successful integration of the *piggyBac*^{TM} selection cassette*,* I selected the transfected cells for two weeks with puromycin and screened resulting single cells by genomic PCR and Sanger sequencing. Puromycin-resistant cells contained the *piggyBac*^{TM} cassette as detected by PCR (Figure 5.15**A).** Sanger sequencing revealed site-specific HR at the targeted site, the insertion of the *piggyBac*^{TM} cassette in the correct genomic location downstream of the *cis-EPO* variant and the introduction of the desired mutation at the *cis-EPO* variant (c.-1306, A>C). One screened cell-line appeared to be biallelic for the C-allele at rs1617640 and was therefore clonally expanded and used for all subsequent experiments **(**Figure 5.15**B-C**).

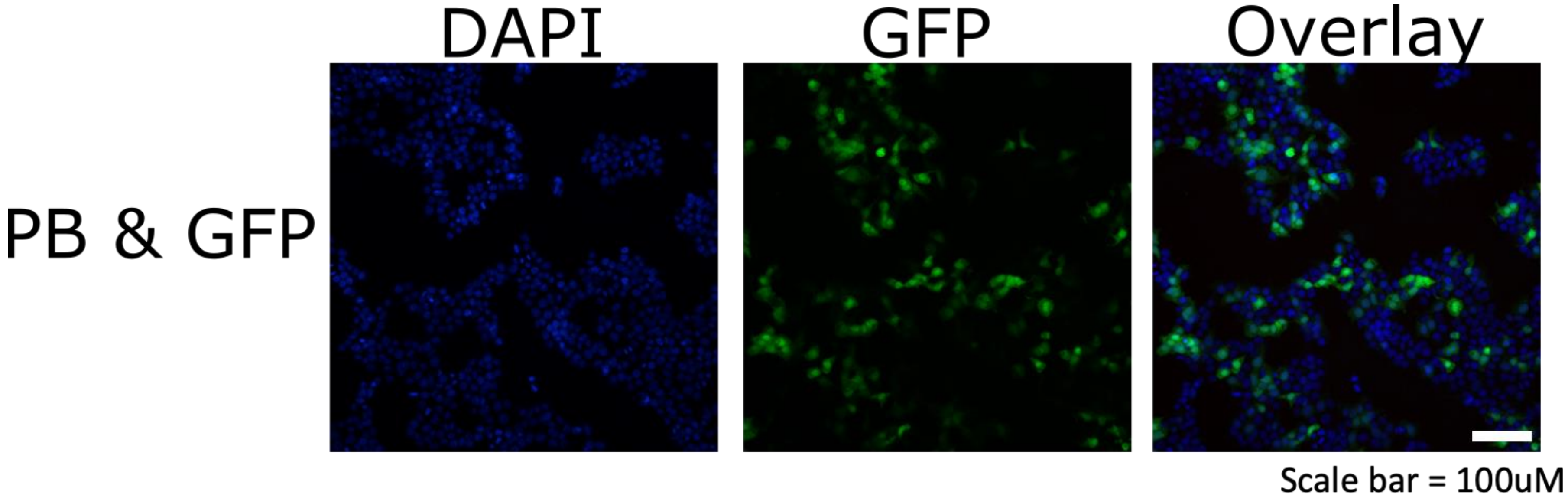


Figure 5.14: Microscopy imaging confirmed successful co-transfection of the CRISPR-Cas9-GFP (pSpCas9(BB)-2A-GFP) and piggyBac^{TM} plasmid into HEK-293 cells. *Images were taken on the Leica DMi8 Widefield microscope. Scale bar represents 100 μM. PB = final piggyBac^{TM} plasmid (containing the 5' and 3' homology arms). GFP = final CRISPR-Cas9-GFP plasmid (containing the gRNA).*

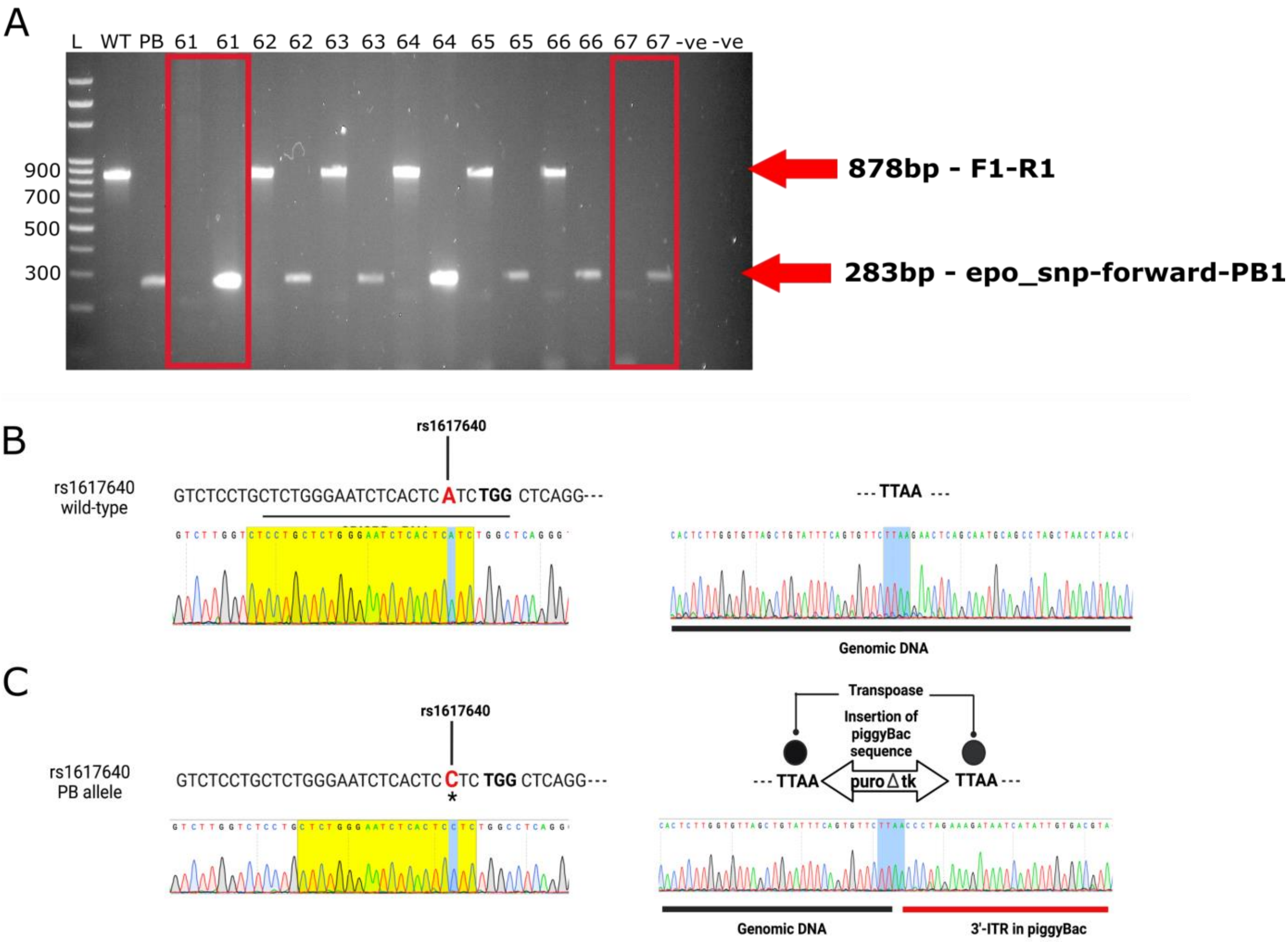

A
L WT PB 61 61 62 62 63 63 64 64 65 65 66 66 67 67 -ve -ve
900
700
500
300
878bp - F1-R1
283bp - epo_snp-forward-PB1
B
rs1617640
rs1617640 wild-type
GTCTCCTGCTCTGGGAATCTCACTCATCTGG CTCAGG---
--- TTAA ---
Genomic DNA
C
rs1617640
rs1617640 PB allele
GTCTCCTGCTCTGGGAATCTCACTCCTC TGG CTCAGG---
*
Transpoase
Insertion of piggyBac sequence
--- TTAA
puro Δ tk
TTAA ---
Genomic DNA
3'-ITR in piggyBac

Figure 5.15: Identification of clone that had successfully undergone HR and contained the desired point mutation at rs1617640. ***A:*** *Genotyping via PCR to detect the integration of the piggyBac^{TM} transposon within the genomic DNA at the correct location. Clones 61 & 67 appeared to be homozygous for the piggyBac^{TM} transposon. -ve = DNA replaced by water in the PCR reaction. WT = HEK-293 wild-type DNA used as a positive control to check F1-R1 primers amplified a band of correct size. PB = final piggyBac^{TM} targeting vector (with homology arms) used as a positive control to check epo_snp-forward-PB1 primers amplified a band of correct size. L = Solis Biodyne 100 bp ladder.* ***B:*** *Sanger sequencing of wild-type HEK-293 cells to double check genotype at rs1617640 (left plot; star and red letter emphasises cis-EPO allele, yellow highlighted sequence represents the gRNA sequence) and to show the sequence at the 'TTAA' site before integration of the piggyBac^{TM} transposon (right plot).* ***C:*** *Sanger sequencing of the targeted cells confirmed the successful gene-editing of the cis-EPO SNP (left plot; A>C, star and red letter emphasise the mutation of the cis-EPO allele, yellow highlighted sequence represents the gRNA). Both alleles of the cis-EPO SNP had been targeted as highlighted by the clean band and integration of the piggyBac^{TM} transposon (shown in the right plot).*

5.4.4 Generation of heterozygous *cis-EPO* SNP knock-ins in HEK-293 cells

Isogenic cell-lines containing the *piggyBac*^{TM} transposon with the selection cassette and homozygous for the wild-type A-allele or for the mutant C-allele were treated with a plasmid encoding the excision-only *piggyBac*^{TM} transposase. PCR amplification of the genomic DNA upstream of the 5' homology arm and downstream of the 3' homology arm resulted in three distinct genotypes in cell-lines containing the *cis-EPO* knock-in; clones showing transposon removal from only one allele, clones showing transposon removal from both alleles and clones showing retention of the transposon in both allele (**Figure 5.16A**). Clones showing transposon removal from both alleles underwent subsequent genomic PCR analysis using primers within the *piggyBac*^{TM} transposon to confirm that the *piggyBac*^{TM} transposon had not been reintegrated into the genomic DNA (**Figure 5.16B)**. Clones lacking a band representing no reintegration of the *piggyBac*^{TM} transposon were sent for Sanger sequencing. The results showed that heterozygous targeting of the *cis-EPO* SNP had been successfully achieved modifying the A-allele to a C-allele (**Figure 5.16C)**. Homozygotes for the mutant allele of rs1617640 were unable to be identified. However, for the purpose of this study, heterozygote editing was adequate to determine if the genetic variant altered *EPO* expression levels.

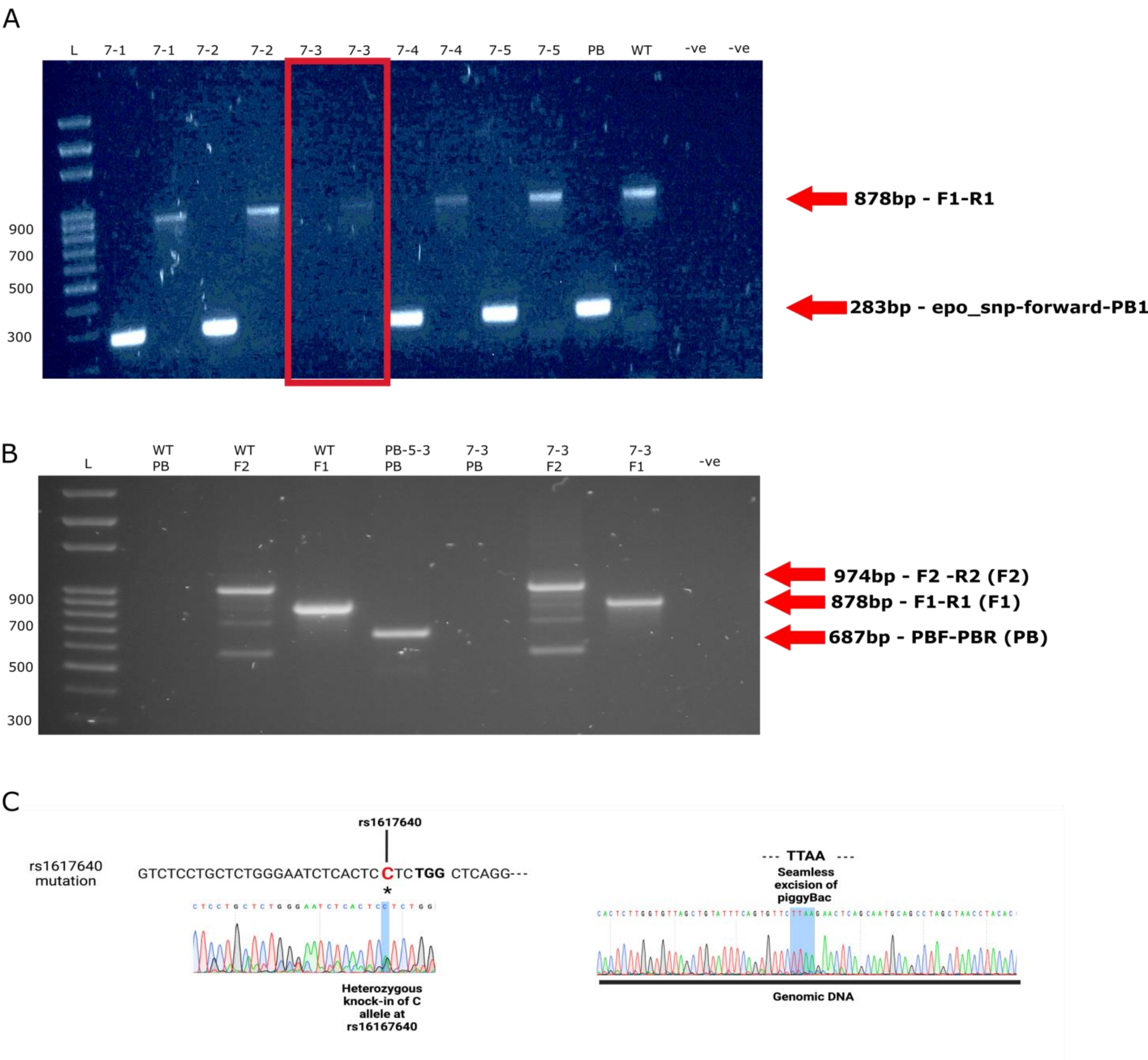


Figure 5.16: Successful removal of the piggyBac^{TM} and single-site base-editing at rs1617640. ***A:*** *Genotyping via PCR to detect clones no longer carrying the piggyBac^{TM} transposon downstream of rs1617640. Clone 7-3 appeared to have the piggyBac^{TM} transposon successfully removed at both alleles. WT = wild-type HEK-293 cells used as a positive control to check F1-R1 primer pair was working. PB = final piggyBac^{TM} with homology arms used as a positive control to check that the epo_snp-forward-PB1 primer pair was working. -ve = DNA is replaced by water in the PCR reaction. L = Solis Biodyne 100 bp ladder.* ***B:*** *Genotyping via PCR to check no reintegration of the piggyBac^{TM} transposon in Clone 7-3. WT = wild-type HEK-293 cells used as a positive control to check F1-R1 and F2-R2 primer pairs were amplifying a band of correct size. PB-5-3 = final piggyBac^{TM} plasmid with homology arms used as a positive control to check PBF-PBR primers worked correctly. PB = PBF-PBR primer pair, F2 = F2-R2 primer pair, F1=F1-R1 primer pair. -ve = DNA is replaced by water in the PCR reaction. L = Solis Biodyne 100 bp ladder.* ***C:*** *Sanger sequencing of Clone 7-3 confirmed that the piggyBac^{TM} transposon had seamlessly been removed from the genomic DNA on both alleles (right plot) and that the desired point mutation had been made at the cis-EPO SNP (A > C). The double peak seen at the cis-EPO SNP indicated that heterozygous gene-editing had been achieved.*

5.4.5 Heterozygotes for the A-allele at the *cis-EPO* SNP have reduced *EPO* expression levels compared to homozygotes for the A-allele

For functional validation of the *cis-EPO* SNP as the most likely causal variant in controlling *EPO* expression levels and therefore a valid instrument for use to genetically proxy therapeutic rises in endogenous EPO levels (analysis performed in **Chapter 3),** I performed qRT-PCR to assess the allele-specific effects of rs1617640 on *EPO* expression levels in HEK-293 cells. The results showed that *EPO* mRNA expression levels were higher in the cells homozygous for the A-allele of rs1617640 compared to heterozygotes for the A-allele confirming that the *cis-EPO* polymorphism has an allele-specific effect on *EPO* gene expression levels (**Figure 5.17**). These results are consistent with my genetic findings that the A-allele is associated with higher circulating EPO levels **(Chapter 3).**

5.4.6 Heterozygotes for the A-allele at the *cis-EPO* SNP have altered expression levels of Notch signalling genes.

qRT-PCR was also performed on three Notch signalling genes (*HEY2, DTX3L, PARP9)* which showed differential expression in the whole *EPO* gene knock-outs (identified by whole transcriptomic analysis in **Chapter 4**) to see if specific alteration of the *cis-EPO* variant also resulted in dysregulated Notch signalling. Homozygotes (A/A) of the *cis-EPO* SNP had down-regulated expression of the three Notch signalling genes compared to heterozygotes (A/C) of the *cis-EPO* SNP (**Figure 5.18**). Negative control experiments with two additional housekeeping genes (*Pol2RA* and *PPIA)* expected to not show differential expression confirmed that the *cis-EPO* SNP-editing (c.-1306, A>C) was most likely specific to the EPO pathways (**Figure 5.18**).

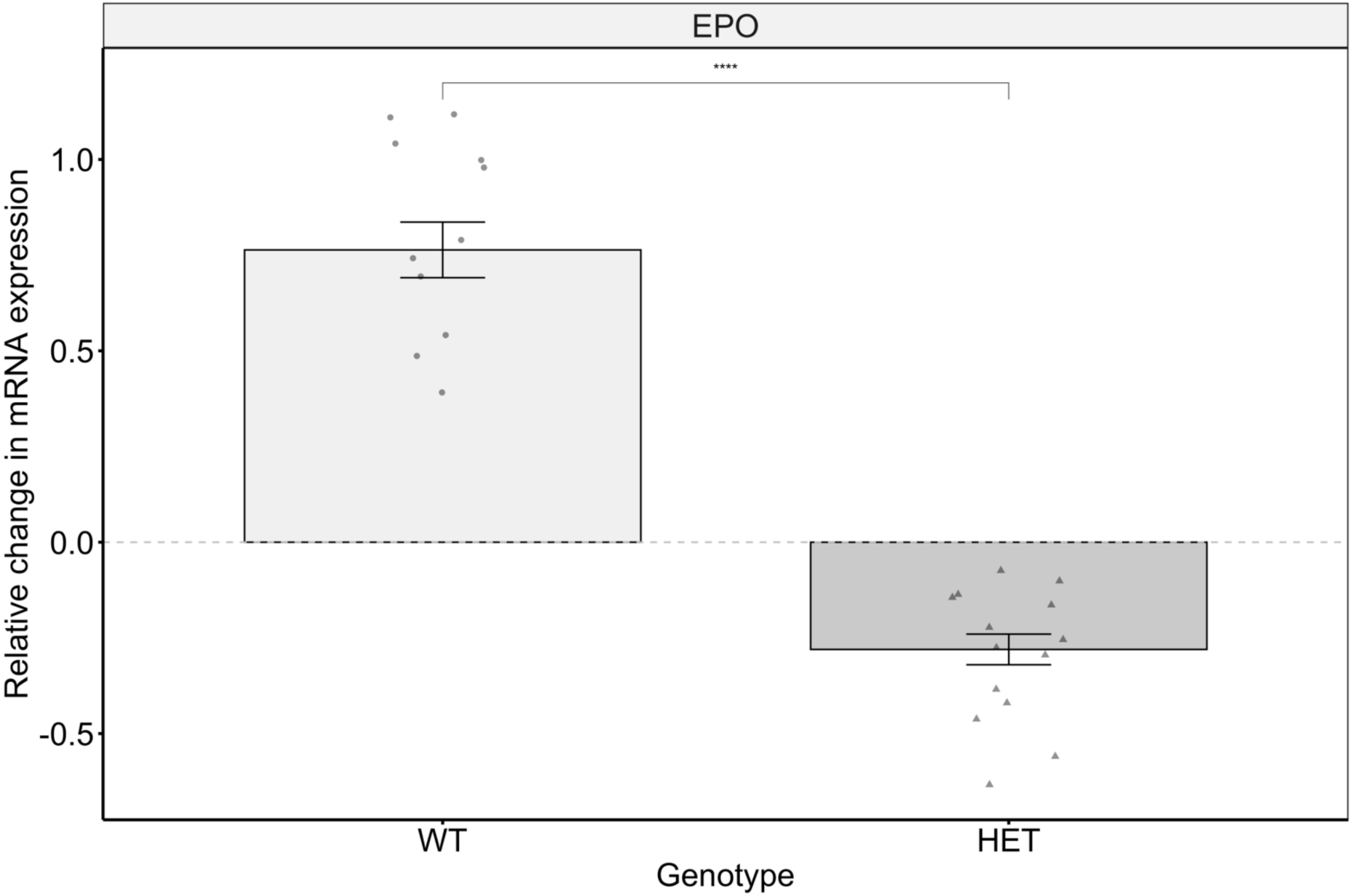


Figure 5.17: The cis-EPO variant has an allele-specific effect on EPO mRNA expression levels. *Homozygotes for the A-allele at rs1617640 showed significantly higher EPO mRNA expression levels than heterozygotes for the A-allele at rs1617640 indicating that the cis-EPO variant is important in controlling EPO mRNA expression levels. EPO mRNA expression levels were measured by performing qRT-PCR in heterozygotes (A/C) and homozygotes (A/A) of rs1617640. The graph shows the relative change in mRNA expression levels (±SEM) between genotypes. HET = HEK-293 cells heterozygous for the A-allele at rs1617640. WT = wild-type HEK-293 cells with A/A genotype at rs1617640. Columns and error bars represent mean and standard error (SEM) values. Paired t-test was performed in RStudio. ****P≤0.0001.*

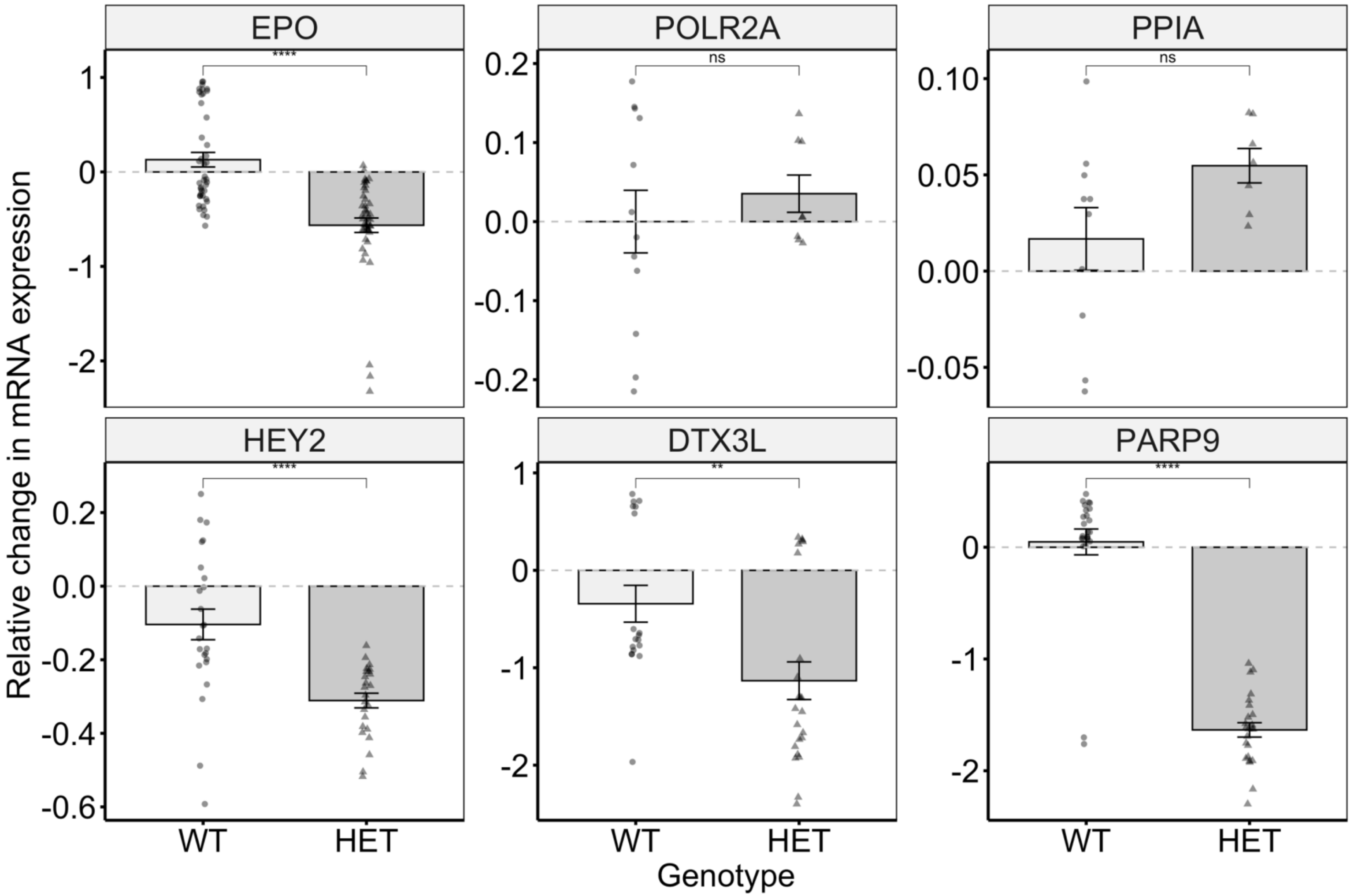


Figure 5.18: The A-allele at rs1617640 is associated with down-regulation of genes involved in the Notch signalling pathway.

*mRNA expression levels were measured by performing qRT-PCR in heterozygotes (A/C) and homozygotes (A/A) of rs1617640. Three Notch signalling genes (HEY2, DTX3L & PARP9), identified as dysregulated in EPO$^{-/-}$ knock-outs, were checked for expression. Expression levels of two controls genes (PPIA and POLR2A) were also checked as negative controls and EPO was repeated again as a positive control for altered expression. The graph shows the relative change in mRNA expression levels (± SEM) between genotypes. Columns and error bars represent mean and standard error (SEM) values. Paired t-test was performed. ns = non-significant, *P≤0.05, **P≤0.01, ***P≤0.001, ****P≤0.0001.*

5.5 Discussion

In this study, I generated a heterozygous knock-in model of the functional polymorphism at rs1617640 (c.-1306, A>C) in HEK-293 cells to functionally validate the genetic variant as causal in controlling *EPO* expression levels. I further developed and optimised an approach using CRISPR-Cas9 with the *piggyBac*^{TM} transposon system to alter a non-coding region within the transcription start site of the *EPO* locus. First, by selecting an appropriate gRNA to target the *cis-EPO* variant and introduce a DSB within the vicinity of the variant, the desired single-base mutation, carried by the *piggyBac*^{TM} template, was inserted into the genome through DSB-mediated HR **(Figure 5.13,** Figure 5.15**).** Second, following puromycin selection to isolate cells containing the *piggyBac*^{TM} transposon and FIAU selection after transposase-guided excision of the selection marker, I achieved heterozygous modification of the A-allele at rs1617640 (Figure 5.15**, Figure 5.16**). Third, I observed that no residual exogenous sequence remained at the targeted site and the remaining genome appeared undisturbed **(Figure 5.16)**. Heterozygous modification of the A-allele at rs1617640 showed reduced *EPO* expression levels compared to wild-type homozygotes of the A-allele indicating that rs1617640 is the most likely causal variant at controlling EPO expression levels (**Figure 5.17).** The impact of the alteration at rs1617640 on Notch signalling genes found dysregulated in transcriptomic profiling of *EPO* knock-outs (see **Chapter 4**) was also investigated to see if the allele has similar effects to that of the gene. Heterozygotes of the A-allele of rs1617640 were found to have altered expression levels of Notch signalling genes further indicating that rs1617640 has an allele-specific effect on controlling EPO levels and downstream signalling pathways of EPO (**Figure 5.18**). This study provides functional evidence that the *cis-EPO* variant is causal in controlling *EPO* expression and therefore supports the use of the *cis-EPO* variant as a partial proxy for therapeutically altered endogenous EPO levels.

I chose the *cis-EPO* genetic variant as a functional target to provide experimental data to validate the genetic findings in **Chapter 3** and to provide an example of how gene-editing can be employed to provide functional evidence to improve our understanding of likely candidate genes and downstream pathways of variants identified through genetic analyses. The results obtained in this study support the genetic association identified in **Chapter 3** and previous research investigating the functional effects of the rs1617640 polymorphism (Amanzada et al., 2014; Renner et al., 2020; Tong et al., 2008). Previous studies have investigated the functional effects of the rs1617640 SNP in patients with different pathologies and have reported conflicting results. Studies investigating the effect of the *cis-EPO* variant in patients with diabetic retinopathy and hepatitis C patients on antiviral therapies provide evidence consistent with our findings indicating that the A-allele is associated with higher EPO levels (Amanzada et al., 2014; Tong et al., 2008). Other studies investigating the effect of the A-allele in Neuro2a cells, in Chinese patients with diabetic retinopathy or in male Jordanian blood donors, have found the A-allele to be associated with lower EPO concentrations (Y. Fan et al., 2016; Kästner et al., 2012; Khabour et al., 2012). These conflicting results indicate that the allele-specific effect of the *cis-EPO* variant on EPO levels is likely dependent upon the physiological state, the cell-type, the cellular localisation, the pathological state and perhaps the physiological timing (Renner et al., 2020). It is difficult to replicate these different conditions *in vitro.* These findings highlight the importance of studying allele-specific effects in tissue-relevant cell-lines and the careful consideration needed when translating findings. Future studies utilising the described methods in model animals, such as mice or zebrafish, and human organoid models would be worthwhile to assess the allele-specific effects in living organisms. Despite the conflicting results, all studies, including this one, indicate that the *cis-EPO* variant plays an important role in controlling EPO levels and this study is the first investigating the impact of the variant in a human cell model.

The introduction of single-base gene-edits to cell-lines is often limited by the low efficiency of HR (Fei Xie et al., 2014). However, this study shows how a customised gRNA with the CRISPR-Cas9 gene-editing system efficiently generated a site-specific DSB directly targeting the promoter region of the *EPO* gene. The simultaneous introduction of the *piggyBac*^{TM} targeting vector to cells promoted efficient DNA repair of this site-specific DSB via HR as shown by the presence of the piggyBac^{TM} transposon in all screened clones after puromycin selection **(**Figure 5.15**).** The drug-selection based enrichment of genetically modified clones makes screening easier and faster than alternatively used approaches, such as ssODNs, but the efficiency of HDR altering both alleles remains a limitation (Y. Huang et al., 2012; Ishida et al., 2018)**.**

I was only able to identify a successfully edited heterozygote clone. Heterozygote targeting was enough to validate the genetic variant as likely causal in controlling *EPO* expression levels and therefore did not affect the study outcome. It would be useful, however, to generate a model representing the complete allelic-series (i.e. for the *cis-EPO* SNP: C/C, A/C, A/A) to enable investigation of whether the allele has an additive effect as suggested by the genetic study (**Chapter 3)** or a recessive effect. Complete allelic-series models would comprise the gold standard for validating genetics variants as causal (J. Lin & Musunuru, 2018). The models generated in this study of mutant homozygotes (C/C), wild-type homozygotes (A/A) and heterozygotes (A/C) containing the *piggyBac*^{TM} cassette could be used to investigate the allelic-series effects providing control experiments investigating whether the integration of the *piggyBac*^{TM} transposon altered *EPO* expression was carried out (by testing for no difference in *EPO* expression levels between A/A homozygotes with and without the *piggyBac*^{TM} transposon). Further screening and optimisation of the protocol for use in HEK-293 cells is needed to improve the efficiency of obtaining a homozygote after transposase treatment.

It is important when using the *piggyBac*^{TM} and CRISPR-Cas9 systems to consider the distance of the 'TTAA' site and the cleavage site from the desired gene-edit site (Bialk et al., 2015; O'Brien et al., 2019; Paquet et al., 2016)**.** The closer the cut-site and the 'TTAA' site to the modification site, the higher the efficiency of HDR and chance of achieving a homozygous mutation (Doench et al., 2016; Kondrashov et al., 2018; Paquet et al., 2016; Yusa, 2013). It has previously been shown that 80% of clones carry single modifications when the 'TTAA' site is 200 bp from the modification site and this reduces to 70% when the distance increases to 300 bp (Yusa, 2013). The frequency of obtaining homozygous modifications therefore reduces further and thus the distance between the 'TTAA' site and modification site in this study may be the reason so few homozygous clones were identified after puromycin selection and none after transposase treatment. Furthermore, previous studies have shown that the homology arm junction should be no longer than 100 bp from the cut-site (Ran, Hsu, Wright, et al., 2013; J.-P. Zhang et al., 2017) and the gRNA should be as close to the modification site as possible to increase chances of point mutations (Bollen et al., 2018). It is a balancing act between obtaining a gRNA as close to the intended modification site as possible with a high off- and on-target score (Bollen et al., 2018; Doench et al., 2016). Several gRNAs could be chosen and tested for efficiencies (Cong et al., 2013). I was able to identify many puromycin-resistant clones targeted on one allele or both alleles retaining the wild-type sequence (A/A), but only found one successfully targeted clone mutated at both alleles highlighting the low frequency of true correctly modified homozygotes. Upon removal of the transposon and negative screening, the genotype of the isogenic homozygous mutant (C/C) cell-line (containing the *piggyBac*^{TM} selection cassette*)* returned to that of a heterozygote (A/C). It may be that the mutant homozygous cell-line with the transposon was not a true homozygote before transposase treatment or that upon transposon removal, the cell-line underwent spontaneous genomic modification returning the genotype to that of a heterozygote. This warrants further investigation and could indicate that the C/C genotype of rs1617640 is not viable within HEK-293 cells.

The risk of off-target effects are always an area for concern when using CRISPR-Cas9 gene-editing techniques (Tycko et al., 2019). The benefit of combining the *piggyBac*^{TM} system with CRISPR-Cas9 is that effective screening for the incorporation of the transposon into the expected genomic location can be undertaken reducing the chances of off-target effects. Cells can also easily be screened for reintegration of the transposon after transposase treatment by PCR. The use of the excision-only transposase increases the chance of footprint free removal as the transposase lacks integration activity and is therefore unlikely to be reintegrated elsewhere (X. Li et al., 2013). However, the concern regarding off-target cleavage that may have unpredictable phenotypic consequences still remains (Fu et al., 2013; H. Li et al., 2020). The most practical approaches to overcome these risks currently are careful design of gRNA using the most up-to-date algorithms and online tools alongside additional computational analysis to identify potential off-target sites and screening for these, or the generation of multiple independent cell-lines and checking for the same phenotype in each (Guanqing Liu et al., 2020; Yusa, 2013). As the field advances, further development of tools and methods for a more comprehensive analysis of off-target effects as well as the generation of readily available modified Cas9 proteins optimised for reducing off-target effects will help overcome these concerns (C.-L. Chen et al., 2020; Kang et al., 2020; Naeem et al., 2020; D. Wang et al., 2019). Although alternative approaches to introduce a single-base gene-edit could have been used, the combination of CRISPR-Cas9 with the *piggyBac*^{TM} system was chosen over other transposons, like Sleeping Beauty, due to the *piggyBac*^{TM} having a preference for integration around transcription units which is where the *cis-EPO* SNP is located (i.e. within the *EPO* gene transcription start site) (Ivics et al., 2009). This made the method an appropriate and efficient approach to introduce a point mutation at rs1617640.

This chapter has further developed and optimised the protocol first described by Yusa et al. (2013) for the introduction of single-base gene-edits into HEK-293 cells. I have shown how CRISPR-Cas9 targeted gene-editing and HR induced by the piggyBac^{TM} transposon system can be used to achieve precise and footprint-free modification of genetic variants within the regulatory region of genes. This relatively novel approach can be used to provide experimental evidence of the most likely causal variant and candidate gene further improving the interpretation and understanding of the underlying biological mechanisms identified through genetic analyses. These findings provide validation of the choice of instrument for use as a genetic proxy for therapeutic modulation of endogenous EPO levels and provide additional insight into the regulatory region of the *EPO* gene.

Chapter 6 Genetically proxied therapeutic inhibition of PHD enzymes and cardiovascular risk.

The work presented in this chapter is currently available as a pre-print which is under internal review at GSK and will be submitted to BioRxiv and PloS Medicine. I have reformatted and expanded on sections taken directly from the pre-print for the purpose of this thesis.

Harlow, CE. Patel, VV. Waterworth, DM. Wood, AR. Beaumont, R. Ruth, KS. Tyrell, J. Oguro-Ando, A. Chu, AY & Frayling, TM. 2022. Genetically proxied therapeutic inhibition of PHD enzymes and cardiovascular risk.

For this Chapter, I helped develop the analysis plan and performed all analysis with the support and guidance of my supervisors, Professor Tim Frayling and Dr Audrey Chu. I worked on the work presented in this Chapter alongside GSK whilst doing my placement to provide additional support to the pharmacogenomics work being undertaken into the effects of PHI treatments.

6.1 Abstract

Novel treatments for anaemia of chronic kidney disease (CKD), prolyl-hydrolase inhibitors (PHIs), have recently completed large-scale Phase III clinical trials to assess the safety and efficacy compared to the current standard of care. Current treatments for anaemia of CKD, such as recombinant human EPO (rhEPO) and its analogues, are associated with increased risk of adverse effects on the heart and vasculature, specifically major adverse cardiovascular events that can include coronary artery disease (CAD), myocardial infarction (MI) and stroke. One hypothesis is that these adverse effects are caused by excessively high haemoglobin (Hgb) levels. PHIs act through inhibition of the prolyl hydroxylase enzymes (PHDs) increasing physiological Hgb levels through activation of the hypoxic pathway. Naturally occurring genetic variation in or near genes encoding the PHDs can be used to help understand the potential effect of long-term therapeutic PHD inhibition. I performed two-sample Mendelian Randomisation (MR) to test the genetically proxied effects of PHD inhibition on stroke (40,585 cases; 406,111 controls), coronary artery disease (CAD; 60,801 cases; 123,504 controls) and myocardial infarction (MI; 42,561 cases; 123,504 controls) using data from GWAS meta-analyses. I used eight genetic variants in or near the three PHI target genes *(EGLN1/2/3)* to partially mimic the effects of PHD inhibition. To identify potential other effects of long-term rises in Hgb levels, I performed a phenome-wide association study (PheWAS) using GWAS data on up to 451,099 UK Biobank individuals. Genetically proxied therapeutic PHD inhibition, equivalent to a 1.00 g/dL increase in circulating Hgb levels, was not associated (at P < 0.05) with increased odds of CAD (OR [95% CI] = 1.04 [0.91, 1.20]), MI (OR [95% CI] = 1.03 [0.88, 1.20]) or stroke OR [95% CI] = 0.98 [0.85, 1.12]). Similar results were found when using a larger but less specific set of 515 variants associated with circulating Hgb levels. By performing PheWAS, I found no associations between genetically proxied PHD inhibition and other diseases or risk factors with clinically meaningful differences. These results suggest that general long-term raising of circulating Hgb levels through PHD inhibition shows no increase in cardiovascular risk; I could exclude (at $P < 0.05$) odds ratios of 1.35 for a minimally clinically significant increase in Hgb, of 1.00 g/dL. The main limitation is that common genetic variants proxy effects within the normal range and in the healthy population so may be less relevant to severely ill patients.

6.2 Introduction

Hypoxia-inducible factor (HIF)-prolyl hydroxylase (PHD) inhibitors (PHIs) have recently completed Phase III clinical trials for treating anaemia in CKD (N. Chen, Hao, Liu, et al., 2019; N. Chen, Hao, Peng, et al., 2019; Chertow et al., 2021; K.-U. Eckardt et al., 2021; A. K. Singh, Carroll, McMurray, et al., 2021; A. K. Singh, Carroll, Perkovic, et al., 2021; Q. Zheng et al., 2021). PHIs act at the transcriptional level of the hypoxic-response genes by inhibiting the PHD enzymes (PHD1-3) leading to an accumulation of HIF-α activating the hypoxic response pathway (Gupta & Wish, 2017; Yeh et al., 2017). Increased transcription of the hypoxic-response genes, including *EPO*, results in increased erythropoiesis and subsequent elevated circulating haemoglobin (Hgb) levels restoring tissue oxygen delivery and correcting the anaemia (Haase, 2013; F. S. Lee & Percy, 2011; Watts et al., 2020). By acting at the transcriptional level, PHIs maintain endogenous EPO levels within the physiological range preventing sudden and/or excessive Hgb level elevations potentially reducing the risk of cardiovascular events, thromboembolism, and heart failure compared to current treatments (Gupta & Wish, 2017). Phase II trials indicate that PHIs can produce dose-dependent changes in Hgb levels and maintain target Hgb levels with small increases in EPO levels in patients either receiving or not receiving dialysis treatment (Brigandi et al., 2016; Holdstock et al., 2019; Meadowcroft et al., 2019). Phase III trials show PHIs to be non-inferior compared to rhEPO in terms of cardiovascular safety supporting ongoing development (Akizawa et al., 2021; N. Chen, Hao, Peng, et al., 2019; Chertow et al., 2021; K.-U. Eckardt et al., 2021; A. K. Singh, Carroll, McMurray, et al., 2021; A. K. Singh, Carroll, Perkovic, et al., 2021).

Genetic studies can be used to support clinical trial data by providing additional evidence that drug targets are associated with the intended therapeutic indication and not associated with unintended opposite direction effects further characterising the therapeutic profile (Nelson et al., 2015; Nguyen et al., 2019; Plenge et al., 2013). Several examples already corroborate the power of genetic studies in providing supporting evidence of drug safety (Lauridsen et al., 2015; Lotta et al., 2016; Swerdlow et al., 2015). Mendelian Randomisation (MR) is one approach in which genetics can be used to help identify causal relationships between intended (e.g. higher biomarker levels or disease) and unintended drug effects (e.g. disease or unintended effects) (Davey Smith & Ebrahim, 2003; Walker et al., 2017). Genetic variants lying within or nearby the gene encoding the drug target, or associated with the drug's intended effects, are used as unconfounded, unbiased proxies for pharmacological action, providing evidence of lifelong exposure on risk of disease (Swerdlow et al., 2016). Phenome-wide association studies (PheWAS) are another method by which genetics can help characterise on-target therapeutic profile. In these studies, a genetic variant, or combination of variants, associated with the intended drug effects are tested for associations with a wide-range of phenotypes in large sample sizes, to identify potential unexpected effects that may have not been considered in clinical trials (Diogo et al., 2018).

PHIs target the hypoxic pathway through inhibition of the PHD enzymes encoded by the *EGLN* genes (*EGLN1/2/*3). Studies of rare genetic variants and *in vivo* models provide some insight into the potential effects of targeting the *EGLN* pathways (Gardie et al., 2014). Some studies have shown that rare loss-of-function variants lying in *EGLN1* give rise to polycythaemia (pathogenic erythrocyte numbers) and inappropriate EPO production which is potentially linked to cardiovascular risk (e.g. hypertension or thrombotic events) in patients carrying these variants (Gardie et al., 2014). Additionally, mice lacking *EGLN1* show embryonic lethality due to heart and placental defects (Minamishima et al., 2008; Takeda et al., 2006). However, these studies are limited by the small number of patients studied and the differences between humans and mice. Common *EGLN* gene variants with modest effects can therefore provide insight into the potential long-term effect of therapeutically altering Hgb in the physiological range through PHD inhibition. In this study, I aimed to use common genetic variants, lying within or near the *EGLN* genes to partially mimic PHD inhibition and assess the associated risk of cardiovascular disease (CVD; defined here as coronary artery disease [CAD], myocardial infarction [MI] and stroke) with lifelong exposure to circulating Hgb level elevations through genetically proxied therapeutic PHD inhibition or other potential effects of long-term Hgb levels through targeting the *EGLN* genes.

6.3 Chapter Aims

The primary aim of this Chapter is to investigate the effects of long-term higher circulating Hgb levels through genetically proxied therapeutic PHD inhibition. To achieve this, I aim to:

1. Identify genetic variants associated with circulating Hgb levels lying within the genes encoding the drug-target genes (*EGLN1/2/3).*
2. Use these genetic variants to partially proxy therapeutic PHD inhibition and predict the cardiovascular risk associated with long-term higher circulating Hgb levels.
3. Perform PheWAS to examine whether PHD inhibition leading to higher long-term Hgb levels is likely to increase risk of additional other effects.
4. Perform secondary analysis with all genetic variants associated with long-term rises in Hgb levels to test the effect of general Hgb raising on cardiovascular risk.

6.4 Methods

6.4.1 Selection of Hgb-associated genetic variants

Using the most recent published GWAS of Hgb, I selected 515 conditionally independent single nucleotide polymorphisms (SNPs) (minor allele frequency [MAF] > 1%) associated with Hgb levels at $P < 5 \times 10^{-09}$ (Vuckovic et al., 2020) **(Appendix 3**: Selection of 515 Hgb-associated SNPs**).** I extracted the publicly available summary association statistics for these genome-wide Hgb-associated variants from Vuckovic et al. (2020) and aligned effect sizes to the Hgb-increasing allele. These statistics were based on 408,112 Europeans studied in UK Biobank (UKB).

6.4.2 Selection of drug-target specific Hgb-associated SNPs

From the list of 515 conditionally independent Hgb-associated genetic variants identified by Vuckovic et al. (2020), eight SNPs annotated to three *EGLN* genes (*EGLN1, EGLN2, EGLN3)* encoding PHI drug targets (PHD1-3) were selected **(Table 6.1)*.*** A gene symbol was provided for the Hgb-associated SNPs by Vuckovic et al. (2020) based on the variant effect predictor (VEP) annotation tool (McLaren et al., 2016), assigning the gene symbol(s) for the most serious predicted consequence **(Table 6.1)**. One of the variants lies within an exon and disrupts the coding sequence (rs61750953; serine [TCG] > Leucine [TTG]), and one lies within the 5’ UTR of *EGLN2* (rs184088518 G>T) **(Table 6.1)**. Summary statistics for the association between these eight *EGLN*-specific SNPs and circulating Hgb levels were extracted from Vuckovic et al. (2020).

Table 6.1: Association between the eight genetic variants annotated to the EGLN genes, the target of PHIs, and Hgb levels. *Effect sizes were obtained from Vuckovic et al., 2020.* *Annotations were provided using VEP.*

Gene	RSID	Chr	Pos	Type of variant	Evidence	Ref allele	Effect allele	Effect allele Freq	Effect estimate for effect allele	SE	P-value	N
EGLN1	rs999010	1	231495316	Downstream gene variant	eQTL	A	G	0.63	0.03	0.002	1.17E-36	408122
EGLN1	rs61835223	1	231562228	Upstream gene variant	Nearest gene	A	G	0.02	0.12	0.008	2.55E-55	408122
EGLN2	rs73047068	19	41297106	Intron variant	Nearest gene	C	G	0.84	0.02	0.003	2.25E-10	408122
EGLN2	rs192191487	19	41305065	Intron variant	Nearest gene	G	A	0.02	0.08	0.009	2.84E-18	408122
EGLN2	rs184088518	19	41305138	5 prime UTR variant	Nearest gene	T	G	0.98	0.12	0.007	3.65E-60	408122
EGLN2	rs61750953	19	41306650	Missense variant	Coding	T	C	0.99	0.10	0.009	2.74E-28	408122
EGLN3	rs797343	14	34646269	Intron variant	eQTL	C	T	0.68	0.02	0.002	4.43E-19	408122
EGLN3	rs12897414	14	34724550	Intron variant	Nearest gene	T	C	0.38	0.01	0.002	4.12E-11	408122

6.4.3 Definition of CVD

I selected three cardiovascular diseases (CVD) - stroke, MI, or CAD - given their relevance to PHIs and the availability of GWAS data from very large samples. I obtained summary association statistics for the 515 Hgb-associated SNPs on CAD, MI or stroke from recently published, publicly available GWAS in European individuals which did not include UKB individuals to ensure estimates came from independent cohorts increasing statistical power and reducing risk of ‘winner’s curse’ (whereby the true causal estimate can be underestimated) (Lawlor, 2016). For MI and CAD, the GWAS performed by Nikpay et al. (2015) in 42,561 and 60,801 cases respectively and 123,504 controls was used. CAD was defined by a record of MI, acute coronary syndrome, chronic stable angina or coronary stenosis > 50% (based on coronary angiographic evidence) obtained from patient and death registers (see Nikpay et al. (2015) for additional details). For stroke, the GWAS performed by Malik et al. (2018) in 40,585 cases and 406,111 controls was used. Stroke was defined as ischemic stroke or intracerebral haemorrhage based on clinical and imaging criteria (Malik et al., 2018). Subarachnoid haemorrhages were excluded (Malik et al., 2018). I did not look for proxies for the SNP (rs192191487) which was missing in the stroke GWAS (Malik et al. 2018) **(Table 6.2).**

Table 6.2: The association between the eight EGLN-specific Hgb-associated genetic variants and risk of stroke, MI or CAD. *Association statistics have been obtained from previously published, publicly available GWAS. Summary statistics for stroke were extracted from Malik et al. (2018) and summary statistics for CAD and MI were extracted from Nikpay et al (2015). Betas are aligned to the effect allele. EAF = effect allele frequency. Ref = reference. Chr = chromosome.*

							Stroke			MI			CAD		
Gene	**RSID**	**Chr**	**Pos**	**Ref_allele**	**Effect_allele**	**EAF**	**beta**	**se**	**p-value**	**beta**	**se**	**p-value**	**beta**	**se**	**p-value**
EGLN1	rs999010	1	231495316	A	G	0.63	-0.0076	0.0097	0.4309	0.0090	0.0105	0.3921	0.0115	0.0095	0.2259
	rs61835223	1	231562228	A	G	0.02	0.0668	0.0399	0.0941	0.0872	0.0511	0.0876	-0.0977	0.0471	0.0379
EGLN2	rs73047068	19	41297106	C	G	0.84	-0.0066	0.0123	0.5899	0.0072	0.0155	0.6429	0.0015	0.0141	0.9146
	rs192191487	19	41305065	G	A	0.02				-0.0076	0.0487	0.8764	-0.0242	0.0442	0.5836
	rs184088518	19	41305138	T	G	0.98	-0.0645	0.0401	0.1083	0.0350	0.0364	0.3361	0.0311	0.0331	0.3480
	rs61750953	19	41306650	T	C	0.99	-0.0725	0.0476	0.1275	-0.0355	0.0437	0.4161	-0.0182	0.0405	0.6534
EGLN3	rs797343	14	34646269	C	T	0.68	0.0003	0.0104	0.9759	-0.0036	0.0125	0.7708	-0.0007	0.0111	0.9491
	rs12897414	14	34724550	T	C	0.38	0.0083	0.0095	0.3863	-0.0010	0.0108	0.9230	-0.0005	0.0097	0.9612

6.4.4 Two-sample Mendelian Randomisation

I performed two-sample MR using the MRBase package (Hemani et al., 2018) implemented in R (RStudio Team, 2018). Palindromic SNPs with intermediate allele frequencies were removed by the package. I first performed drug-target two-sample MR using the eight drug-target-specific Hgb-associated *EGLN* SNPs as instruments and then performed secondary analysis using the 515 Hgb-associated SNPs. Five two sample MR methods were performed; inverse-variance weighted (IVW); MR-Egger (Bowden et al., 2015); Weighted median (Bowden et al., 2016); weighted mode (Bowden et al., 2016); simple mode (Hartwig et al., 2017). I have presented the IVW approach as our main analysis method, with the latter four representing sensitivity analyses to account for unidentified pleiotropy which may bias our results. IVW assumers there is no horizontal pleiotropy (where genetic variants influence the outcome independently of the exposure) and that the SNP-exposure association is not correlated with the path from SNP-outcome that is independent of the exposure (InSIDE assumption) (Bowden et al., 2019; Burgess & Thompson, 2017). I tested for pleiotropic effects using the MR-Egger intercept obtained through the ‘mr_pleiotropy_test’ function and for heterogeneity using the ‘mr_heterogeneity’ function (Bowden et al., 2015). When there was evidence of pleiotropy (as indicated by $P < 0.05$), I placed more weighting on the MR-Egger estimate, which partially accounts for pleiotropic effects and provides unbiased estimates.

6.4.5 Steiger-filtering

To obtain the most specific Hgb genetic instrument, Steiger filtering (Hemani, Bowden, et al., 2017; Hemani, Tilling, et al., 2017) was implemented using the MRBase package (Hemani et al., 2018) in R (RStudio Team, 2018) on the 515 Hgb-associated SNPs. Steiger filtering uses a statistical method to select those genetic variants which explain more variance in the exposure than the outcome (R^2 [exposure] > R^2 [outcome]) (Hemani, Bowden, et al., 2017). I filtered the 515 Hgb-associated genetic variants to obtain a more specific instrument with primary effects on the exposure using Steiger_direction = true and Steiger *P*-value < 0.05. I repeated two-sample MR using this filtered Hgb-specific set of 288, 237 and 284 genetic variants to obtain a more reliable estimate of the relationship between long-term genetically proxied Hgb levels and cardiovascular risk. I did not perform Steiger filtering when performing MR using the *EGLN*-specific genetic instruments.

6.4.6 Testing the association of the eight conditionally independent *EGLN*-specific SNPs with circulating EPO levels

I investigated the association between each of the eight *EGLN*-specific Hgb-associated genetic variants and circulating EPO levels to determine if the genetic variants were likely acting through the hypoxic pathway and influencing circulating Hgb levels through increased EPO activity. Summary statistics for the association between the eight *EGLN*-specific Hgb-associated genetic variants and endogenous EPO levels were obtained from the EPO meta-analysis in 6,127 individuals of European and African American descent performed in **Chapter 3.**

6.4.7 Comparison of effect estimates to typical Hgb levels in the general population

To obtain a more representative, physiologically relevant effect, I scaled the genetic estimates on disease outcomes by a factor of 0.81 (1 / 1.23) where 1 is the desired unit change of Hgb in raw units (g/dL) and 1.23 is the standard deviation of Hgb in UK Biobank (UKB) (*N* = 437,573). This value provided an estimate of the genetically proxied odds of disease for a 1-unit increase in long-term circulating Hgb levels. A 1-unit increase in Hgb levels is the minimally clinically significant difference.

6.4.8 PheWAS of an *EGLN*-specific genetic risk score

To investigate the potential pleiotropic effects of the eight *EGLN*-specific SNPs or identify potential other effects downstream of Hgb through targeting the *EGLN* genes, I performed a PheWAS on 923 traits in up to 451,099 unrelated, European UKB individuals using a weighted genetic risk score (GRS) consisting of the eight *EGLN*-specific Hgb-associated SNPs. Traits were selected following the same approach as Frayling et al. (2018). I generated the weighted Hgb GRS by extracting the dosages of the *EGLN*-specific SNPs from 437,573 unrelated European UKB individuals, as defined by principal component (PC) analysis (method details in (Frayling et al., 2018)), with phenotypic and genotypic information. Alleles were aligned to the Hgb-increasing alleles. The weighted GRS was created using **Equation 6.1** .

$$Weighted\ Hgb\ GRS = \sum dosage \times \ |\beta(Hgb - increasing\ allele)|$$

Equation 6.1: Calculation of the weighted Hgb GRS

To obtain all genotype-phenotype associations, regression analysis of the weighted GRS on 923 traits adjusting for age, sex, chip, centre and PCs 1-5 was performed. Continuous traits were inverse normalised prior to regression to account for skewed distributions. Traits were stratified by sex as well to investigate sex-specific effects. I highlight associations reaching a Bonferroni-adjusted *P*-value < 5.42 x 10^{-05} (0.05 / 923). Effect estimates were converted back to original units to determine whether statistically significant associations were clinically significant using the standard deviation of phenotypes in UKB individuals.

6.5 Results

6.5.1 Genetically proxied therapeutic PHD inhibition resulting in long-term higher circulating Hgb levels is not associated with cardiovascular risk

To genetically proxy the effects of therapeutic PHD inhibition, I used eight Hgb-associated SNPs in three genes, *ELGN1, EGLN2* and *EGLN3,* encoding PHDs (**Table 6.1**). Using these variants as instruments in drug-target two-sample MR, I found no evidence of a causal association with any of the three cardiovascular diseases tested **(Figure 6.1**, **Table 6.3**). There was no evidence of pleiotropy or heterogeneity in the genetic instruments **(Table 6.4**). As common genetic variants tend to have subtle effects on phenotypes, it can be helpful to scale their effects to provide estimates in a more physiologically relevant range (Bovijn et al., 2020; Scott et al., 2016; Yarmolinsky et al., 2022). I therefore present results of the estimated effect of a 1 g/dL increase in Hgb on CVD outcomes, based on genetic instrumentation of PHD inhibition (**Table 6.5).** I found no evidence (at $P < 0.05$) for increased odds of CAD (OR [95% CI] = 1.06 [0.84, 1.35]), MI (OR [95% CI] = 1.02 [0.79, 1.33] or stroke (OR [95% CI] = 0.91 [0.66, 1.24]) for a 1 unit increase in Hgb levels in the physiological range (e.g. from 14.2 g/dL to 15.2 g/dL) through genetically proxied PHD inhibition (**Figure 6.2, Table 6.5)**. Based on the upper confidence intervals, I could statistically exclude an odds of 1.35, 1.33 and 1.124 for CAD, MI or stroke respectively (**Figure 6.2, Table 6.5).**

Table 6.3: Using two sample MR with the eight EGLN-specific SNPs as instruments, there was no evidence for an association between genetically proxied higher Hgb levels and increased risk of CVD. *Values in the table represent the MR causal estimates obtained using each of the following approaches; MR Egger, weighted median, inverse variance weighted, simple mode and weighted mode. Only seven variants went into the genetic instruments when testing the association between Hgb and stroke as one SNP was missing from the GWAS. The inverse variance weighted method was used for the main analysis. The other methods were performed as sensitivity analyses.*

Exposure	**Outcome**	**Method**	**N_SNPs**	**OR**	**OR_L95**	**OR_U95**	**P-value**
Hgb	CAD	MR Egger	8	1.017	0.622	1.663	0.948
		Weighted median	8	0.922	0.607	1.399	0.702
		Inverse variance weighted	8	1.080	0.803	1.452	0.612
		Simple mode	8	0.879	0.440	1.756	0.726
		Weighted mode	8	0.805	0.423	1.530	0.529
	MI	MR Egger	8	0.956	0.577	1.585	0.868
		Weighted median	8	0.872	0.562	1.353	0.541
		Inverse variance weighted	8	1.029	0.743	1.424	0.865
		Simple mode	8	0.834	0.394	1.765	0.649
		Weighted mode	8	0.773	0.423	1.411	0.429
	Stroke	MR Egger	7	0.820	0.421	1.597	0.585
		Weighted median	7	0.764	0.485	1.202	0.244
		Inverse variance weighted	7	0.888	0.605	1.303	0.543
		Simple mode	7	0.679	0.324	1.422	0.344
		Weighted mode	7	0.660	0.331	1.318	0.284

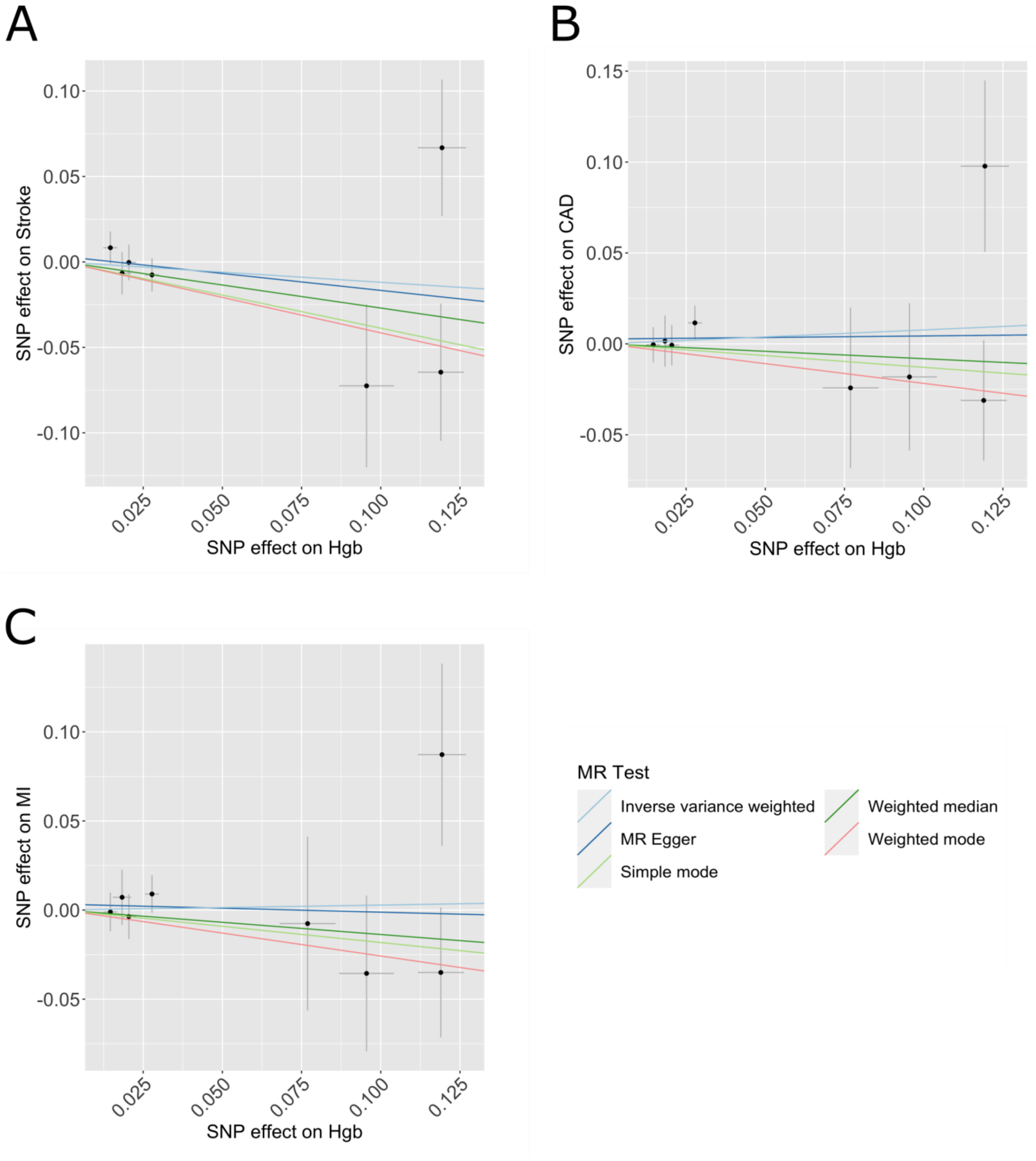


Figure 6.1: Genetically proxied therapeutic PHD inhibition shows no evidence of adverse cardiovascular risk with long-term higher circulating Hgb levels. *Two sample MR was carried out using the eight EGLN-specific Hgb-associated variants to genetically mimic therapeutic PHD inhibition. MR causal estimates were consistent across all methods, including sensitivity methods. There was no evidence of heterogeneity or pleiotropy. Plots were produced using the TwoSampleMR package in R.*

Table 6.4 There was no evidence of pleiotropy and heterogeneity when using the eight EGLN-specific SNPs as genetic instruments in two-sample MR.

Test for pleiotropy and heterogeneity was carried out in MRBase when using the eight EGLN-specific SNPs as genetic instruments to assess the causal association between higher Hgb levels and risk of CVD. df= degrees of freedom. Q=Measure of heterogeneity. IVW=Inverse weighted variance method.

Outcome	Exposure	Egger_intercept	SE	P-value	Q_MR_Egger	Q_df_MR_Egger	Q_pval_MR_Egger	Q_IVW	Q_df_IVW	Q_pval_IVW
MI	Hgb	0.0033	0.009	0.72	5.40	6	0.49	5.54	7	0.59
CAD	Hgb	0.003	0.009	0.77	6.81	6	0.34	6.92	7	0.44
Stroke	Hgb	0.003	0.011	0.78	8.68	5	0.12	8.83	6	0.18

Table 6.5: Rescaling the genetic estimates to the effect of a 1 unit increase in Hgb levels in the general population. *The standard deviation of typical Hgb was obtained from UKB individuals and used to rescale the genetic estimates to the effect of a 1 unit change in Hgb levels (14.2 g/dL to 15.2 g/dL) for a more physiologically relevant estimate of the effect. I could exclude an increased odds of 1.35, 1.33 and 1.24 for CAD, MI and stroke, respectively, based on the upper bounds.*

Trait	Effect	OR	L95	U95
CAD	Genetic effect	1.080	0.803	1.452
	Scaled to 1 g/dL unit increase	1.064	0.837	1.354
MI	Genetic effect	1.029	0.743	1.424
	Scaled to 1 g/dL unit increase	1.023	0.785	1.333
Stroke	Genetic effect	0.888	0.605	1.303
	Scaled to 1 g/dL unit increase	0.908	0.665	1.240

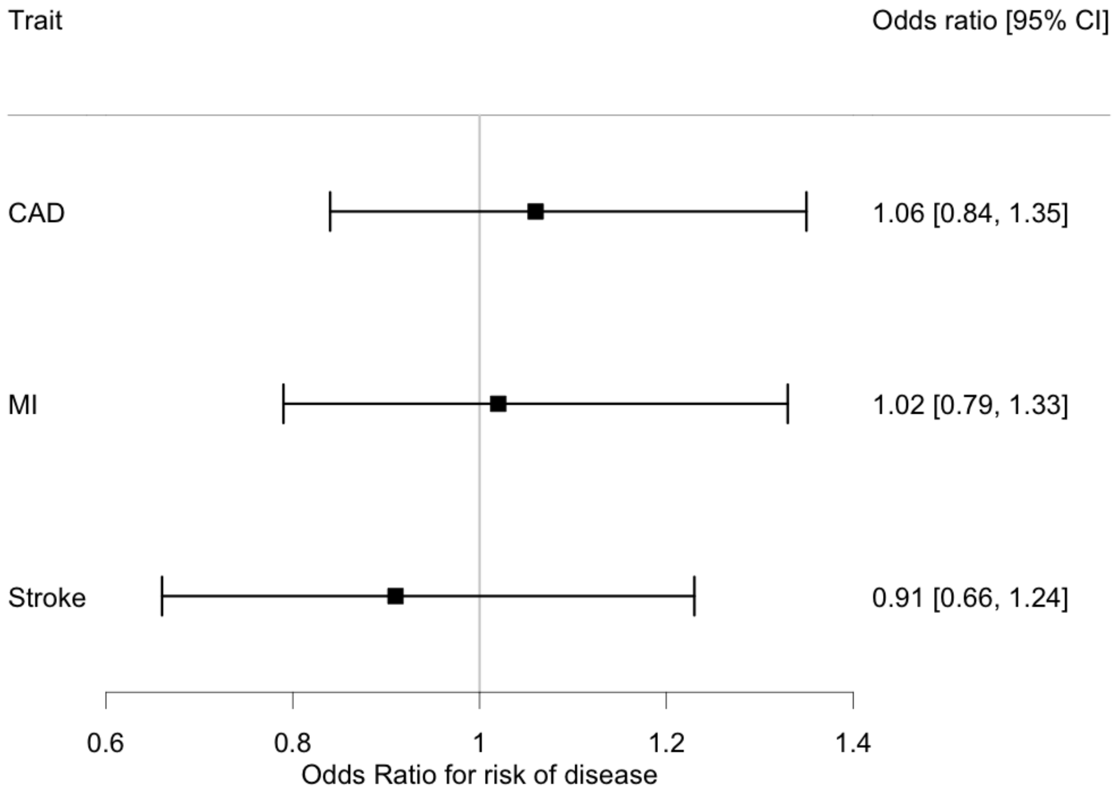


Figure 6.2: The effect of a 1 unit-change in Hgb levels on risk of CVD in the general population as genetically proxied by therapeutic PHD inhibition. *The genetic effects per 1 SD increase in circulating Hgb levels were rescaled to the typical Hgb level in the general population by multiplying the effects by the scaling factor (0.81) to represent a 1 unit increase in Hgb levels (e.g. going from 14.2 g/dL to 15.2 g/dL). Based on the upper bound of the estimate, I could exclude increased odds of 1.35 for CAD, increased odds of 1.33 for MI and increased odds of 1.24 for stroke with genetically proxied therapeutic PHD inhibition. Plot was produced using forestplot package in R.*

6.5.2 *EGLN*-specific SNPs are not associated with circulating EPO levels

To determine whether the eight *EGLN*-specific genetic variants were also likely to be influencing EPO levels (Gupta & Wish, 2017), I investigated the association between the eight *EGLN*-specific Hgb-associated SNPs and circulating EPO using the EPO GWAS meta-analysis in 6,127 individuals of European and African descent performed in **Chapter 3.** I found no association between the *EGLN*-specific SNPs and EPO **(Table 6.6).**

Table 6.6: The eight EGLN-specific Hgb-associated SNPs were not associated with circulating EPO levels. *The association statistics between the eight EGLN-specific Hgb associated SNPs and EPO were obtained from GWAS meta-analysis of circulating EPO levels in 6,127 individuals of European and African-American descent* ***(Chapter 2).*** *Statistical significance was determined using the standard genome-wide threshold of P < 5 x* 10^{-08}.

Gene	RSID	Chr	Pos	Effect_allele	Non_effect_ allele	EAF	Beta	SE	P-value	N
EGLN1	rs999010	1	231495316	A	G	0.355	-0.023	0.0204	0.2576	5591
	rs61835223	1	231562228	G	A	0.992	0.0005	0.0653	0.9934	6127
EGLN2	rs73047068	19	41297106	G	C	0.831	0.0523	0.0264	0.0474	6127
	rs192191487	19	41305065	A	G	0.016	0.1128	0.1053	0.2841	4917
	rs184088518	19	41305138	G	T	0.976	0.0481	0.0727	0.5079	5591
	rs61750953	19	41306650	C	T	0.982	-0.0523	0.0831	0.5291	5133
EGLN3	rs797343	14	34646269	T	C	0.704	-0.0045	0.0214	0.8323	5591
	rs12897414	14	34724550	T	C	0.629	-0.0038	0.0203	0.8497	5591

6.5.3 Genetic risk for higher Hgb levels is associated with relevant erythrocyte traits and biomarkers related to kidney function

To determine the specificity of the *EGLN*-specific SNPs as instruments for Hgb levels, I generated a weighted Hgb GRS consisting of the eight SNPs. I then used this GRS to perform a PheWAS on 923 traits in up to 451,099 unrelated European UKB individuals regardless of CKD status. The weighted Hgb GRS was associated with 0.05 SD (SE = 0.002, $P = 8 \times 10^{-168}$) higher circulating Hgb levels, equivalent to a per allele 0.062 unit increase in Hgb in the general population **(Table 6.7).** I found the *EGLN*-specific Hgb GRS was most strongly associated with erythrocyte phenotypes including red blood cell count (β [SE] = 0.05 [0.002], $P = 5.00 \times 10^{-120}$), haematocrit percentage (β [SE] = 0.05 (0.002), $P = 2.00 \times 10^{-168}$), reticulocyte count (β [SE] = 0.01 [0.002], $P = 1.98 \times 10^{-09}$), platelet crit (β [SE] = -0.01 [0.002], $P = 1.68 \times 10^{-06}$) and platelet count (β [SE] = -0.01 [0.002], $P = 5.64\text{-} \times 10^{-06}$) **(**Figure 6.3**, Table 6.7)**. I also found associations between the *EGLN*-specific Hgb GRS and traits related to kidney function, including estimated creatinine-based glomerular filtration rate (eGFR: β [SE] =-0.01 [0.002], $P = 3.63 \times 10^{-08}$) and microalbumin (β [SE] = -0.01 [0.002], $P = 4.88 \times 10^{-09}$) and liver function related traits, such as bilirubin (total bilirubin: β [SE] = 0.02 [0.002], $P = 7.40 \times 10^{-12}$), a biomarker indicative of erythrocyte disorders (Gazzin et al., 2016) (Figure 6.3**, Table 6.7)**. Despite being statistically significant, these associations were not clinically significant (equivalent to a 2.22, 0.02, and 1.08 unit change in eGFR, microalbumin and total bilirubin per 1 g/dl higher Hgb respectively). Stronger associations, passing the Bonferroni P-value threshold ($P < 5.42 \times 10^{-05}$), were found in women compared to men for bilirubin, microalbumin, creatinine, and eGFR, although the direction and magnitude of effects remained consistent (Figure 6.3**, Table 6.7)**.

6.5.4 Long-term rises in circulating Hgb levels through genetically proxied therapeutic PHD inhibition is unlikely to severely increase risk of other comorbidities.

To identify potential additional unwanted effects associated with long-term increases in Hgb levels through genetically proxied therapeutic PHD inhibition, I tested the *EGLN*-specific Hgb GRS for association with 923 traits in up to 451,099 unrelated European UKB individuals regardless of CKD status. I found evidence for an association with reduced sitting-to-standing height ratio (β [SE] = -0.01 (0.002), *P* = 5.54x10^{-10}), and increased risk of non-alcoholic fatty liver disease (NAFLD) fibrosis score (β [SE] = 0.01 [0.002], *P* = 1.12 x 10^{-06}) with higher genetically mediated Hgb levels (Figure 6.3**, Table 6.7)**. I also observed an association with family history of diabetes in siblings (OR [95% CI]: 1.04 [1.02, 1.06], *P* = 3.71 x 10^{-06}) but this was not consistent with the result of type 2 diabetes risk in participants (OR [95% CI]: 0.99 [0.97, 1.03], *P* = 0.998). Overall, these results indicate that long-term higher circulating Hgb levels through therapeutic PHD inhibition are unlikely to confer an increased risk of any secondary conditions **(**Figure 6.3**, Table 6.7)**.

Table 6.7: Phenome-wide association study of the weighted Hgb EGLN-specific GRS with 923 traits in up to 451,099 unrelated European UK Biobank individuals. *Genetically higher Hgb levels are most strongly associated with relevant erythrocyte traits and biomarkers. Only the associations passing Bonferroni P-value threshold (0.05 / 923) are shown in this table. The PheWAS was also carried out stratified by sex and only those sex-stratified traits passing Bonferroni significance are shown in this table.*

Phenotype	**Beta**	**SE**	**P-value**	**Gender**
Haematocrit percentage	0.051	0.002	2.00E-168	Combined
Haemoglobin concentration	0.050	0. 002	8.00E-168	Combined
Red blood cell count	0.045	0. 002	5.00E-120	Combined
Total Bilirubin	0.015	0.002	7.40E-12	Combined
Direct bilirubin	0.020	0.003	3.62E-10	Female
Standing to sitting height ratio	-0.014	0.002	5.54E-10	Combined
High light intensity reticulocyte count	0.014	0.002	1.35E-09	Combined
Reticulocyte count	0.014	0.002	1.98E-09	Combined
Microalbumin	-0.012	0.002	4.88E-09	Combined
Direct bilirubin	0.014	0.002	1.62E-08	Combined
CKD derived eGFR	-0.011	0.002	3.63E-08	Combined
Total Bilirubin	0.017	0.003	4.05E-08	Female
Microalbumin	-0.018	0.003	6.39E-08	Female
MDRD derived eGFR	-0.012	0.002	6.75E-08	Combined
Creatinine	0.010	0.002	1.29E-07	Combined
Creatinine	0.016	0.003	4.97E-07	Female
CKD derived eGFR	-0.013	0.003	5.54E-07	Female
MDRD derived eGFR	-0.015	0.003	6.03E-07	Female
Non-alcoholic fatty liver disease fibrosis score	0.011	0.002	1.12E-06	Combined
Platelet crit	-0.010	0.002	1.68E-06	Combined

Sibling Diabetes	0.042	0.009	3.71E-06	Combined
Direct bilirubin	0.016	0.004	5.15E-06	Male
Platelet count	-0.010	0.002	5.64E-06	Combined
Non-alcoholic fatty liver disease fibrosis score	0.013	0.003	1.60E-05	Female
Total Bilirubin	0.014	0.003	3.25E-05	Male

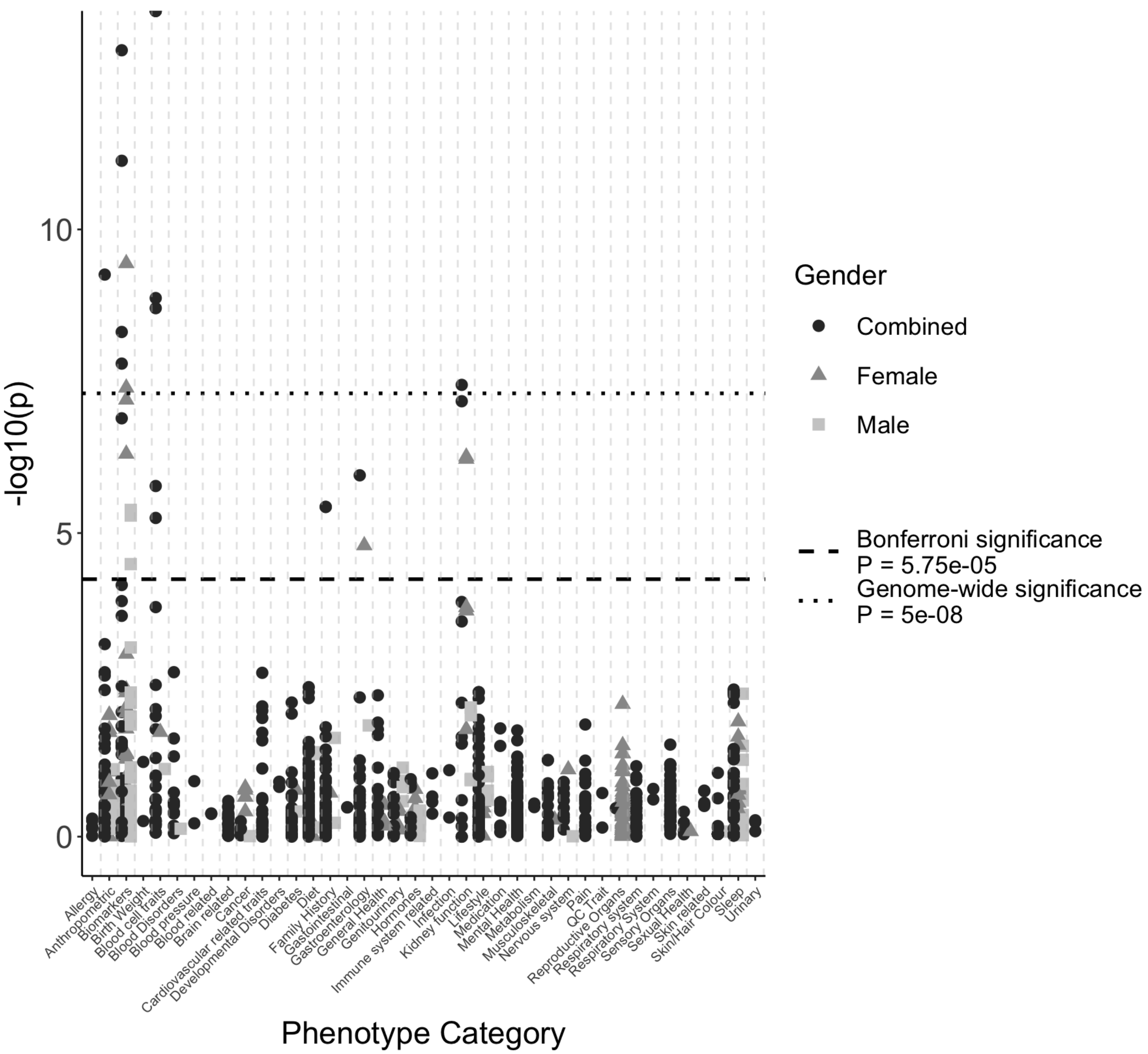


Figure 6.3: PheWAS of the EGLN-specific Hgb GRS reveals long-term higher Hgb levels through genetically proxied PHD inhibition are unlikely to increase risk of other comorbidities.

PheWAS was performed using the *EGLN*-specific variants on 923 traits in up to 451,099 unrelated, European UKB individuals. *The Hgb GRS was most strongly associated with erythrocyte phenotypes indicating that the EGLN-specific variants are involved in relevant pathways and are valid and specific instruments for mimicking therapeutic PHD inhibition. PheWAS traits have been clustered into relevant categories.*

6.5.5 Secondary analysis focusing on overall genetically proxied long-term rises in Hgb levels showed no increase in cardiovascular risk

To understand the causal association between general genetically proxied higher Hgb levels and cardiovascular risk, I performed two-sample MR using 515 Hgb-associated SNPs as instruments. I selected 515 conditionally independent genetic variants associated (at $P < 5 \times 10^{-09}$) with circulating Hgb levels from the most recent, publicly available GWAS on blood cell traits (Vuckovic et al., 2020) **(Appendix 3**: Selection of 515 Hgb-associated SNPs**).** Summary statistics for 409, 407 and 410 of the Hgb-associated variants were available in the publicly available GWAS of the three CVD of interest, CAD, MI, or stroke, respectively (Malik et al., 2018; Nikpay et al., 2015). Using these SNPs as instruments in two-sample MR, I found no evidence (at $P < 0.05$) that a 1-unit increase in genetically mediated Hgb levels in a physiological range leads to an increased risk of stroke (OR [95% CI]: 1.04 [1.00, 1.08], $P = 0.08$), or CAD (OR [95% CI]: 1.05 [1.00, 1.11], $P = 0.07$) in the general population **(Figure 6.4, Table 6.8)**. I found nominal evidence for an association between a 1 unit increase in genetically mediated Hgb levels and increased risk of MI (OR [95% CI]: 1.08 [1.02, 1.14], $P = 0.01$) **(Figure 6.4, Table 6.8)**, but there was strong evidence of pleiotropy and heterogeneity for both the CAD and MI estimates (Egger intercept *P*-value: CAD = 1.68×10^{-05}, MI = 3.80×10^{-05}, Heterogeneity *P*-value IVW: CAD = 2.11×10^{-45}, MI = 2.81×10^{-34}, from **Table 6.9).** I therefore decided to place more weight on the MR-Egger estimate which partially accounts for pleiotropic effects and found suggestive evidence (at $P < 0.05$) for a 0.84 (95% CI: 0.74, 0.95) and 0.86 (95% CI: 0.75, 0.98) decreased odds of CAD or MI with higher genetically mediated circulating Hgb levels **(Figure 6.4, Table 6.8)** (Bowden et al., 2015)**.**

6.5.6 Steiger filtering strengthens the results

To reduce the level of pleiotropy and heterogeneity when using the 515 Hgb-associated variants, I performed Steiger filtering (Hemani, Tilling, et al., 2017). By applying a Steiger filtering FDR threshold of 0.05 to limit the selected variants to those with a greater effect on the exposure than the outcome, the number of variants used to assess the relationship between higher Hgb levels and risk of CAD, MI, or stroke reduced by 107, 156, and 114, respectively

(**Table 6.10**). After applying Steiger filtering, the direction of effect of the causal estimates between MR methods were more consistent, and the amount of heterogeneity and pleiotropy decreased, but the confidence intervals were wider (**Figure 6.4, Table 6.11, Table 6.12)**. Using these filtered Hgb-associated variants, I again found no evidence of a causal association between higher genetically mediated circulating Hgb levels and increased risk of CAD, MI, or stroke (**Figure 6.4, Table 6.11).** I could exclude odds of 1.06, 1.08, and 1.08 for CAD, MI, and stroke with a 1-unit increase in Hgb levels.

Table 6.8: MR causal estimates for the association between higher Hgb levels and risk of CVD as instrumented by 515 conditionally independent Hgb-associated SNPs identified from Vuckovic et al., (2020).*The direction of effect of causal estimates were inconsistent between methods. Two-sample MR was carried out using the MRBase package in R. Summary association statistics obtained from recently published large-scale GWAS meta-analysis were used to provide the SNP-Hgb and SNP-CVD associations for the 515 conditionally independent Hgb association SNPs. Five MR methods were carried out; inverse variance weighted was the primary analysis method and the rest were used as sensitivity.*

Exposure	Outcome	Method	OR	L95	U95	P-value
Hgb	MI	MR Egger	0.858	0.750	0.982	0.027
		Weighted median	1.064	0.971	1.166	0.186
		Inverse variance weighted	1.097	1.022	1.177	0.010
		Simple mode	1.384	1.012	1.893	0.042
		Weighted mode	0.885	0.784	0.998	0.047
	Stroke	MR Egger	1.095	0.990	1.211	0.079
		Weighted median	1.054	0.972	1.142	0.205
		Inverse variance weighted	1.048	0.994	1.104	0.084
		Simple mode	1.045	0.842	1.297	0.688
		Weighted mode	1.057	0.953	1.173	0.293
	CAD	MR Egger	0.835	0.735	0.948	0.006
		Weighted median	0.969	0.892	1.053	0.462
		Inverse variance weighted	1.064	0.995	1.138	0.068
		Simple mode	1.145	0.881	1.489	0.312
		Weighted mode	0.856	0.767	0.955	0.006

Table 6.9: There was strong evidence of pleiotropy and heterogeneity in the 515 Hgb-associated variants. *Test for pleiotropy and heterogeneity was carried out in MRBase when using the 515 Hgb-associated SNPs as genetic instruments to assess the causal association between higher Hgb levels and risk of CVD. df = degrees of freedom. Q=Measure of heterogeneity. IVW=Inverse weighted variance method.*

Exposure	Outcome	Egger Intercept	SE	P-value	Q_MR_Egger	Q_df_MR_Egger	Q_pval_MR_Egger	Q_IVW	Q_df_IVW	Q_pval_IVW
Hgb	MI	0.007	0.0018	3.80E-05	800.47	391	2.23E-30	836.02	392	2.81E-34
Hgb	CAD	0.007	0.0017	1.68E-05	887.75	393	2.87E-40	930.64	394	2.11E-45
Hgb	Stroke	-0.001	0.0014	0.316	628.04	396	8.83E-13	629.63	397	8.27E-13

Table 6.10: Steiger filtering reduced the number of genetic instruments by 107, 156, and 114 for CAD, MI, and stroke, respectively. *Genetic variants were selected as valid instruments after Steiger filtering if the variance explained in the exposure was greater than the variance explained in the outcome determined by dir =TRUE and if they passed an FDR P-value threshold < 0.05.*

Exposure	Outcome	N SNPs before steiger filtering	N SNPs after steiger filtering
Hgb	MI	393	237
Hgb	CAD	395	288
Hgb	Stroke	398	284

Table 6.11: MR causal estimates for the association between higher Hgb levels and risk of CVD after Steiger filtering had been applied to select the most specific set of Hgb-associated SNP as instruments.

After Steiger filtering, the direction of effect of causal estimates between MR methods were more consistent but the confidence intervals were wide.

Exposure	Outcome	Method	OR	L95	U95	P-value
Hgb	MI	MR Egger	0.894	0.791	1.010	0.073
		Weighted median	0.973	0.880	1.075	0.585
		Inverse variance weighted	1.032	0.967	1.100	0.342
		Simple mode	1.307	0.917	1.864	0.140
		Weighted mode	0.852	0.739	0.981	0.027
	Stroke	MR Egger	1.090	0.997	1.192	0.060
		Weighted median	1.052	0.968	1.143	0.230
		Inverse variance weighted	1.047	0.999	1.098	0.053
		Simple mode	1.034	0.833	1.283	0.765
		Weighted mode	1.063	0.955	1.183	0.265
	CAD	MR Egger	0.884	0.801	0.975	0.015
		Weighted median	0.916	0.841	0.998	0.045
		Inverse variance weighted	1.014	0.962	1.068	0.615
		Simple mode	1.169	0.892	1.533	0.258
		Weighted mode	0.829	0.741	0.928	0.001

Table 6.12: After Steiger filtering, the level of pleiotropy and heterogeneity in the Hgb instruments reduced. *Test for pleiotropy and heterogeneity was carried out in MRBase when using the 515 Hgb-associated SNPs as genetic instruments to assess the causal association between higher Hgb levels and risk of CVD. df = degrees of freedom. Q=Measure of heterogeneity. IVW=Inverse weighted variance method.*

Exposure	Outcome	Egger Intercept	SE	Pval	Q_MR_Egger	Q_df_MR_Egger	Q_pval_MR_Egger	Q_IVW	Q_df_IVW	Q_pval_IVW
Hgb	CAD	0.007	0.002	0.002	352.7825	286	0.00429613	365.4307	287	0.00117225
Hgb	MI	0.008	0.0038	0.032	333.2471	235	2.56E-05	343.5695	236	5.89E-06
Hgb	Stroke	-0.0012	0.0022	0.566	312.8545	282	0.09983959	314.0194	283	0.09907926

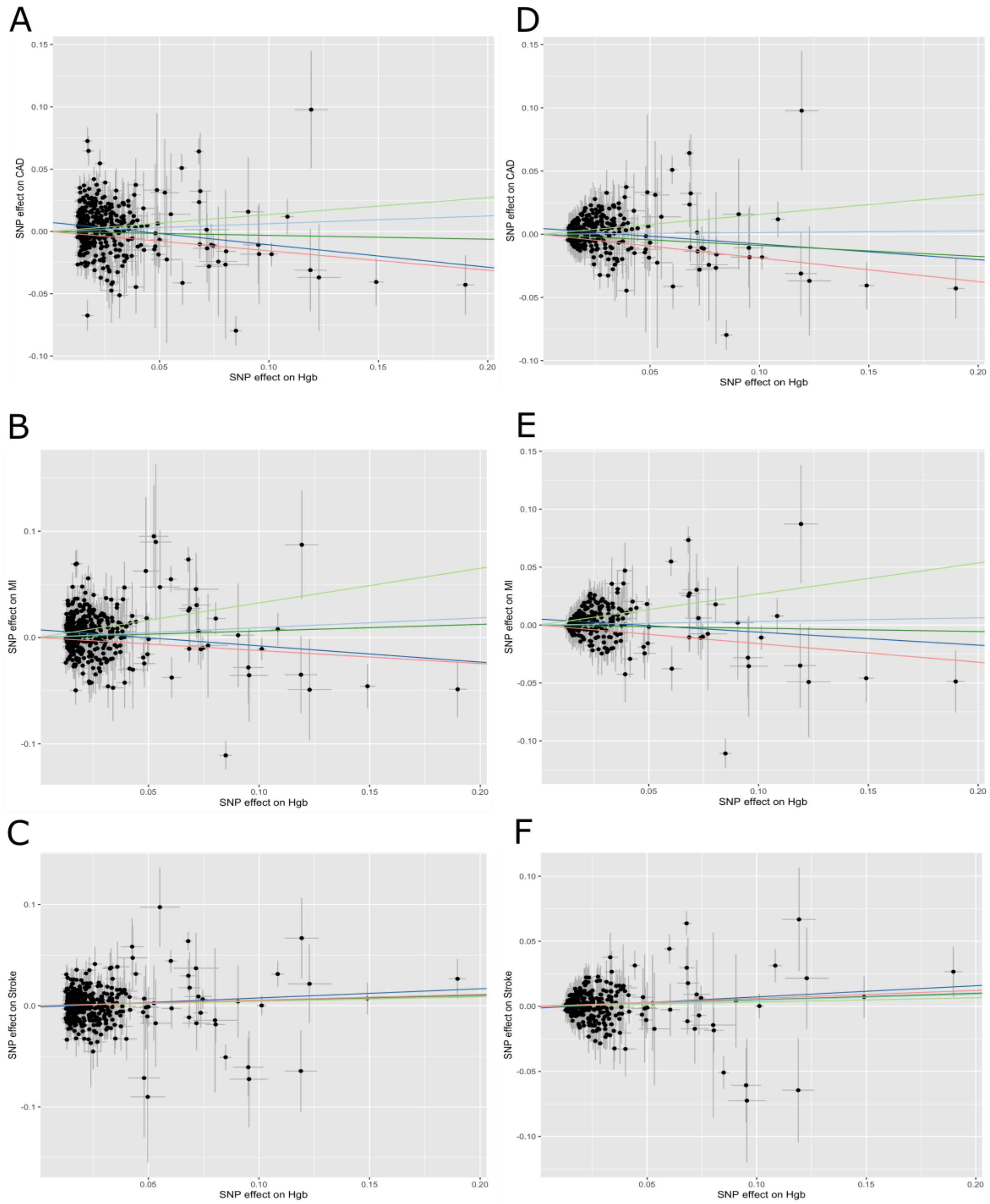


Figure 6.4: MR estimates for the association between Hgb levels and cardiovascular risk using 515 Hgb-associated SNPs as genetic instruments before (A-C) and after (D-F) applying Steiger filtering. *Before Steiger filtering, there was evidence of pleiotropy and heterogeneity in the 515 Hgb-associated genetic variants. More weighting was therefore placed on the MR-Egger causal estimate. Before and after applying Steiger filtering, I found no evidence of a causal association between higher genetically proxied circulating Hgb levels and increased odds of CAD, MI or stroke. During Steiger filtering, I filtered for SNPs which explained a higher variance in Hgb levels compared to the disease outcomes and passed a Steiger P-value threshold < 0.05. Estimates across the five methods became more consistent after Steiger filtering increasing reliability of the true causal estimate. Plots were produced using the TwoSampleMR package in R. The different colour lines represent the five different MR tests (light blue: inverse variance weighted, dark blue: MR Egger, light green: simple mode, dark green: weighted median, red: weighted mode).*

6.6 Discussion

Previous research has shown how human genetics can be used to further characterise therapeutic profiles and help anticipate the risk of unintended effects. PHIs have recently completed phase III clinical trials to treat anaemia in CKD (Akizawa, Iwasaki, et al., 2020; N. Chen, Hao, Liu, et al., 2019; N. Chen, Hao, Peng, et al., 2019; Chertow et al., 2021; K.-U. Eckardt et al., 2021; Fishbane et al., 2021; A. K. Singh, Carroll, McMurray, et al., 2021; A. K. Singh, Carroll, Perkovic, et al., 2021). These Phase III trials have shown non-inferiority for hematologic efficacy, and some non-inferiority for cardiovascular safety, with PHI treatment compared to rhEPO (K.-U. Eckardt et al., 2021; A. K. Singh, Carroll, McMurray, et al., 2021; A. K. Singh, Carroll, Perkovic, et al., 2021). In this Chapter, I used human genetic variants associated with circulating Hgb levels as genetic proxies for the pharmaceutical effect of PHIs and investigated the effect of lifelong exposure to higher circulating Hgb levels on cardiovascular risk and potential other effects. I provide genetic evidence to support cardiovascular safety of PHIs and further inform on potential risk of other effects with therapeutic PHD inhibition which may not be tested in clinical trials. I used a drug-target specific (*EGLN1/2/3)* Hgb genetic instrument to partially mimic the direct effects of therapeutic PHD inhibition through PHI treatment and found no evidence of a causal association between higher Hgb levels and increased cardiovascular risk **(Figure 6.1).** I rescaled the genetic estimates obtained using the *EGLN*-specific instrument to the Hgb levels typically found in the general population to obtain a more relevant effect estimate on the physiological scale. I did not observe (at $P < 0.05$) increased odds of CAD (OR [95% CI] = 1.06 [0.84, 1.35]), MI (OR [95% CI] = 1.02 [0.79, 1.33], or stroke (OR [95% CI] = 0.91 [0.66, 1.24]) with a 1 g/dL long-term higher Hgb level genetically proxied by therapeutic PHD inhibition. Based on the upper bound, I could exclude a 1.35, 1.33, and 1.24 increased odds of CAD, MI, or stroke, respectively with long-term therapeutic rises in Hgb levels (**Figure 6.2**). As all PHIs work through the same mechanisms (i.e. PHD inhibition), these results are supportive of all PHIs. Any differences seen between PHI compounds would be likely related to the biochemical and physical properties of the compounds and way the treatment is used particularly regarding dosing.

As with all uses of common variants as genetic proxies of drug interventions, there are limitations. First, genetic variants tend to represent subtle lifelong changes rather than the more acute and stronger changes from therapies (Pulley et al., 2017; Stitziel et al., 2014). Second, and most importantly, the genetic effects are based on estimates of Hgb alterations in the general population, regardless of CKD status, whereas PHI therapies are given only to anaemic CKD patients with Hgb levels towards the lower end of the range. CKD patients are likely to have variable biomarker levels at baseline which could alter the causal estimates and the presence of other underlying conditions which could alter the way they respond to therapeutic PHD inhibition than that estimated by the genetic association (Mokry et al., 2015; Sofianopoulou et al., 2021). Furthermore, the genetic association does not mimic what is happening at the metabolic level particularly in relation to drug metabolism. Drug metabolism is affected by dose, frequency, administration route, tissue distribution and protein binding, all of which can alter individual response to certain levels of a drug. Genetic variants are only able to proxy the long-term effects of rises in subsequent biomarker levels but not the acute effect at an individual level in terms of how quickly the drug is absorbed by the body, how quickly the drug may be metabolised or excreted, where the drug is likely to have the greatest impact and how the drug may interact with other treatments which are being used to treat comorbidities likely present in anaemic CKD patients. Many comorbidities present in anaemic CKD patients, such as type 2 diabetes, urinary tract infections, or coronary heart disease, will be treated with drugs (pioglitazones, statins, trimethoprim respectively) which impact the cytochrome p450 enzymes. The p450 enzymes, alongside flavin monooxygenases and hydroxyacid oxidase, are the drivers of Phase I of the drug metabolism process during which PHIs get metabolised determining the duration and intensity of PHI pharmacological action (Omiecinski et al., 2011). Genetic polymorphisms play an important role in controlling interindividual variability in drug metabolism and determining variability in drug-related toxicity, adverse drug reactions, alongside drug efficacy (Wormhoudt et al., 1999). Using human genetic variants in the way presented here is unable to replicate and predict these effects amongst treated patients whose genome will differ. Some patients may have variants results in faster metabolism and excretion of the PHIs, whilst other may have variants resulting in slower metabolism leading to acute levels of the drug which could increase risk of adverse side effects (Wormhoudt et al., 1999). Caution is therefore required when interpreting these genetic results at an individual patient level. Additionally, the gut microbiome has been found to play an important role in drug metabolism and differences in bacterial genes can influence pharmacokinetics which may alter an individual's response to a treatment (S. L. Collins & Patterson, 2020). The genetic findings presented here can only be used in conjunction with ongoing clinical trials which ultimately provide the best line of evidence of the long-term and short-term effects of treatments in diseased patients whom will be receiving PHIs and are also on other medications or have other underlying conditions which may impact PHI response or action.

The majority of genetic analyses assume linearity which is not always the case, particularly in relation to PHI treatment which is titrated at an individual patient level to achieve a target Hgb level meaning patients will have different baseline Hgb levels and subsequent increases in Hgb levels from baseline (Sofianopoulou et al., 2021). It is difficult to estimate this kind of effect using genetics. Third, it is often difficult to represent the efficacious, physiologically relevant state or representative cellular concentration of a drug target using genetics (Burgess et al., 2012). The ability to transfer these findings to the target patient population are therefore limited. As more extensive genetic studies become available, particularly in disease-relevant populations, the power to detect associations and ability to perform stratified analyses at different baseline levels will improve (Sofianopoulou et al., 2021; Visscher et al., 2017). This will increase the ability to accurately predict the risk of any potential unintended effects and further enhance the ability of genetics to inform therapeutic profiles and support drug development.

Despite rescaling the genetic effect to the minimally clinically significant difference in Hgb levels (i.e. a 1-unit increase) to try to overcome these limitations, this effect is not necessarily clinically relevant as it is only scaled to the standard deviation of Hgb levels in the general population, not to the effect of PHIs on Hgb levels. Therefore, inferences about the likely effects at individual anaemic CKD patient level need careful consideration. Ongoing clinical trials, which ultimately provide the strongest data, will help decipher these uncertainties and further emphasise the utility of human genetics in providing insights into drug safety and efficacy, with the potential to accelerate the drug development process.

Previous studies have shown that genetic variants lying close to the gene encoding the protein of interest are most likely to have functional impact and influence circulating levels of the protein product or drug target and are therefore the best proxies for mirroring therapeutic effects (Melzer et al., 2008; Swerdlow et al., 2016). For this reason, the primary analysis was performed using Hgb-associated SNPs annotated to the three target genes of PHIs (*EGLN1/2/3)* (Vuckovic et al., 2020). Using these drug-target specific variants increased my ability to directly and more specifically mimic the therapeutic effects of PHIs and the chance of identifying any potential cardiovascular risk attributable to long-term rises in circulating Hgb levels through therapeutic PHD inhibition. I performed a PheWAS to provide additional evidence that the variants were specific and valid proxies for therapeutic PHD inhibition and further insight into the potential effects of PHD inhibition. PheWAS has potential for improving or validating our understanding of biological mechanism, identifying additional indications with potential for repurposing, or indicating potential unwanted effects through associations with other conditions other than the primary indication (Denny et al., 2016; Pulley et al., 2017; Robinson et al., 2018). Through PheWAS, I found the weighted Hgb GRS to be most strongly associated with relevant erythrocyte phenotypes, such as platelet count and red blood count, indicating that these variants are strong, valid genetic instruments as they appear to influence circulating Hgb levels through altered erythropoiesis, the downstream effect of PHD inhibition **(**Figure 6.3**, Table 6.7).** I also found additional associations with relevant kidney- and liver-function related biomarkers, such as eGFR, microalbuminuria and bilirubin. Although these did not reach clinical significance, they further indicate that these instruments are likely acting through the hypoxic pathway in relevant tissue types (where EPO is predominantly produced) and are thus valid proxies for pharmaceutical inhibition of PHDs **(**Figure 6.3**, Table 6.7)** (Watts et al., 2020; Weidemann & Johnson, 2009). However, the direction of effect of higher genetically determined Hgb via the *EGLN* genes on these biomarkers appear counterintuitive; higher Hgb levels are associated with lower eGFR indicative of worse kidney function but with lower microalbuminuria which is a marker of healthier kidneys. Higher Hgb levels are also associated with increased bilirubin, which may be indicative of haemolysis leading to lower Hgb, not higher Hgb. As these investigations were carried out on the general population, it is unclear whether there is some sort of feedback mechanism or confounding impacting these findings and whether inference can be made to a CKD population. It would, therefore, be worth investigation in a CKD-specific population. Sex-specific PheWAS revealed stronger associations (based on *P*-values) between higher Hgb levels and several of the biomarkers, such as bilirubin, creatinine and eGFR, in women compared to men which suggests that higher Hgb levels have a greater effect in women **(**Figure 6.3**, Table 6.7).** Women, in general, have lower Hgb levels than men so increasing Hgb in women is expected to have a larger effect than in men who already have higher Hgb baseline levels (Murphy, 2014). Women are often underrepresented in clinical trials, so this study, using genetics as proxies for drug effects is a useful additional way of increasing relevance to a wider range of patients (Carey et al., 2017; L. Y. Liu et al., 2012; Randall et al., 2013).

When looking for associations with potential secondary diseases or unintended effects, I found evidence for an association between the *EGLN*-specific Hgb GRS and shorter legs and risk of NAFLD and family history of sibling diabetes **(**Figure 6.3**, Table 6.7).** The *EGLN* genes are known to play a role in glucose metabolism through activation of HIF-2a and this likely explains the association found between the *EGLN* SNPs and NAFLD (from metabolic syndrome) or family history of sibling diabetes (Holzner & Murray, 2021; S. K. Ramakrishnan & Shah, 2017; M. Yang et al., 2014). NAFLD is also prevalent in CKD patients and is a clinical marker of poor response to EPO treatments and could therefore be used to determine response to therapeutic PHD inhibition (Orlić et al., 2014). I did not find any association with type 2 diabetes which might be expected but this may be a result of sample size or indicates that the association with sibling diabetes is because of chance **(**Figure 6.3**, Table 6.7).** The association found with shorter legs may be spurious due to UKB being slightly older (37 - 73 years at age of recruitment) and height declining with age in the general population (Cline et al., 1989; Sorkin et al., 1999; Sudlow et al., 2015).

I also investigated the association between the *EGLN* SNPs and EPO to further validate the variants as strong proxies for PHI treatment. However, there was no evidence for an association between the SNPs and circulating EPO levels **(Table 6.8).** This lack of association is likely due to the lack of power from the small sample size in the EPO meta-analysis (*N* = 6,127) (**Chapter 3).** Although there was no evidence for an association, the direction of effect of the majority of the Hgb-increasing allele (75%) also increase circulating EPO levels which coincides with what is expected biologically. However, the EPO-Hgb relationship is complicated and often shows a J-shaped relationship due to the compensatory feedback mechanisms and this alters further through stages of CKD (Lundby et al., 2007; Panjeta et al., 2017). It would be beneficial to repeat the EPO meta-analysis as larger cohorts become available and the power to detect associations increase to further improve our understanding of this complicated feedback loop (Spencer et al., 2009).

My results provide genetic support of the findings from clinical trials in that PHIs are non-inferior for CVD than rhEPO for treating anaemia in CKD. These findings could also be used to support development of treatments for other diseases which act by increasing Hgb levels through the hypoxic pathway, highlighting the translational ability of these types of genetic studies to help predict the risk of potential unintended effects or benefits of any treatment for any disease undergoing clinical development (Plenge et al., 2013). However, it is important to consider the validity of the genetic instrument used in terms of how well the SNP mimics the pharmacological action of the drug and the strength of the variant as an instrument (Walker et al., 2017). MR analysis makes several assumptions and violation of these assumptions can lead to bias in the causal estimates (Davies et al., 2018). Here, when using all the Hgb-associated SNPs to assess the causal relationship between higher Hgb and risk of CVD, I found evidence of pleiotropy (Egger-intercept *P*-value < 0.05, **Table 6.9**) (Bowden et al., 2015), but showed that limiting the variants to those with larger effects on the exposure compared to the outcome (through Steiger filtering) reduced the pleiotropy and heterogeneity increasing power to detect the true causal direction **(Table 6.11, Table 6.12, Figure 6.4)** (Hemani, Bowden, et al., 2017).

In conclusion, my results suggest that general long-term elevated circulating Hgb levels through genetically proxied therapeutic PHD inhibition does not increase risk of CVD or additional complications. I have identified relevant genetic markers for testing the pharmaceutical effects of therapeutic PHD inhibition which could potentially inform further research using patient level clinical data from Phase III trials. I show additional evidence of how human genetics can be used to partially mimic pharmacological action and provide additional insight, alongside clinical trial data, into the long-term therapeutic effects of Hgb level elevations.

Chapter 7 General Discussion

In this PhD, I have utilised several genetic approaches alongside functional validation methods to assess the potential effects of therapeutically altered circulating EPO and Hgb levels. I have provided genetic evidence further characterising the therapeutic profile of PHI treatment particularly in relation to cardiovascular safety. In this final discussion, I present an overview of my primary findings and relate them to existing literature investigating impacts of higher EPO or Hgb levels on risk of disease. I also summarise the strengths and limitations, alongside implications and the future directions of my research.

7.1 Key Findings

7.1.1 Genetically proxied therapeutic rises in circulating EPO levels are not associated with increased risk of CVD or unwanted effects.

In **Chapter 3,** I used genetics to investigate the risk of CVD with therapeutic rises in endogenous EPO levels. Using a GWAS meta-analysis of circulating EPO levels, which at the time was the largest GWAS of circulating EPO, alongside gene expression and colocalisation analysis, I identified a *cis-EPO* variant associated with circulating EPO levels. This *cis-EPO* variant had been previously reported by two studies to be associated with circulating EPO levels in patients with diabetic retinopathy or hepatitis C. I used this *cis-EPO* variant as a proxy to test the effects of therapeutically altering endogenous EPO levels to mimic the downstream effects of PHIs. Using drug-target two-sample Mendelian Randomisation, I found no evidence for an association between therapeutically higher EPO levels, equivalent to 5.1 IU/L, and increased cardiovascular risk (CVD) or clinical markers for CVD associated risk factors (SBP, DBP or heart rate). Instead, I found nominal evidence (at $P < 0.05$) for a protective effect of genetically proxied therapeutic rises in EPO on DBP and resting heart rate. To obtain a clinically relevant effect estimate, I rescaled the genetic associations to the PHI-induced effect on circulating EPO levels and could exclude odds of 1.07, 1.15, and 1.07 for CAD, MI, or stroke respectively with a 2.2-unit increase in circulating EPO. I could also exclude levels higher than 0.78 mmHg for SBP and any increase in DBP or resting heart rate with a 2.2-unit increase in circulating EPO. To further characterise the therapeutic profile of higher EPO levels and check for any unintended effects, I performed a PheWAS of the *cis-*

EPO SNP and found associations with other relevant erythrocyte phenotypes with similar effect sizes and decreased liver function biomarker levels. I therefore provided genetic evidence that therapeutically increased EPO levels by novel treatments for anaemia in CKD are unlikely to infer a substantially increased risk of CVD or increased risk of any unintended effects.

7.1.2 Establishment of *EPO* gene knock-out and subsequent transcriptomic analysis aids characterisation of downstream causal genes and pathways.

In **Chapter 4,** I used CRISPR-Cas9 technology to generate an *EPO* knock-out in HEK-293 cells and then performed RNA-seq analysis to assess the transcriptional changes as a result of *EPO* knock-down. No previous studies have generated a whole *EPO* knock-out cell-line, only *EPO*-deficient animal models, and have not investigated whole transcriptomic changes in *EPO* knock-outs.

I designed paired gRNAs to knock-out the region between the conserved exons 2 and 4 of *EPO* rendering the *EPO* transcript non-functional. I generated two homozygous *EPO* knock-outs confirmed by PCR, Sanger sequencing, qRT-PCR and western blotting. I then performed whole transcriptional profiling following *EPO* knock-out to obtain a better understanding of the downstream causal genes and signaling pathways in HEK-293 cells. Several studies have assessed transcriptional changes in response to rhEPO but none in *EPO* knock-outs highlighting the novelty of this research. Using RNA-seq, I identified differentially expressed genes (DEGs) in the two knock-out cell-lines compared to wild-type controls. Over 3000 of these DEGs overlapped in both knock-outs and showed a strong correlation in their direction of effects. These overlapping DEGs were enriched in pathways and functions related to cell fate, DNA repair, metabolic processes including fatty acid oxidation, ATPase activity and aerobic respiration, as well as control of signaling pathways through protein binding, protein degradation, and receptor activity. These findings support previous literature highlighting the important role EPO has in the body and emphasise the pleiotropic effects of EPO in systems other than just the hematopoietic system. As Notch signalling and its related pathways featured prominently in the gene ontology (GO) analyses, I selected several genes involved in different parts of the canonical Notch signalling pathway. I validated differential expression of these genes using qRT-PCR highlighting the important role EPO plays in controlling Notch signalling activity and its downstream effects and functions. These findings add to literature which suggests a link between hypoxia and Notch signalling and warrants further investigation particularly in relation to disease aetiology.

The RNA-seq dataset produced in this chapter will be made publicly available for others to use (currently being added to a GitHub repository – https://github.com/CharliHarlow/EPO_metaanalysis_rnaseq_MR_phewas). I am the first to perform RNA-seq analysis within the Oguro-Ando lab and have generated a detailed bioinformatic pipeline which I will share with others aiding their analysis of transcriptional changes (added to a GitHub repository https://github.com/CharliHarlow/EPO_metaanalysis_rnaseq_MR_phewas).

7.1.3 Heterozygous SNP knock-in of the C-allele at rs1617640 using CRISPR-Cas9 functionally validates rs1617640 as causal in controlling *EPO* gene expression levels.

In **Chapter 5,** I established a heterozygous cell model of the *cis-EPO* variant in HEK-293 cells and functionally validated the variant as important in controlling *EPO* expression levels. I optimised a published protocol by Yusa et al. (2013) to perform CRISPR-Cas9 gene-editing alongside the *piggyBac^{TM}* transposon system to introduce a single-base change at rs1617640 in HEK-293 cells. Combining CRISPR-Cas9 and the *piggyBac^{TM}* system improved efficiency of activating homology directed repair and aided in selection of successfully single-base gene-edited clones. Through this protocol, I successfully incorporated the *piggyBac^{TM}* transposon into the genomic DNA at the expected location with the desired SNP edit at rs1617640 (A>C) confirmed by PCR and Sanger sequencing. I then excised the *piggyBac^{TM}* transposon from the genomic DNA without leaving any marks in the genomic DNA. I successfully generated a heterozygous knock-in cell-line of rs1617640 (A/C) confirmed via PCR and Sanger sequencing. I used this knock-in model to assess the effect of an allele change at rs1617640 on *EPO* expression levels. I found rs1617640 does have an allele-specific effect on *EPO* expression levels with heterozygotes of the A-allele showing down-regulated *EPO* mRNA levels compared to homozygotes of the A-allele. I also investigated mRNA expression levels of dysregulated Notch signaling genes identified through whole *EPO* gene knock-out **(Chapter 4**); I found heterozygotes of the A-allele to have lower expression of the three Notch signaling genes tested - *Hey2, DTX3L and PARP9*. These findings were consistent with the direction of effect seen in genetic studies and the *EPO* knock-outs providing functional evidence that rs1617640 does affect *EPO* expression and downstream causal genes.

This is the first time the combination of CRISPR-Cas9 gene-editing and the *piggyBac*^{TM} transposon system has been employed at Exeter to introduce a single-SNP base change. The same protocol can be adapted by others to introduce desired gene-edits, whether that be single-base changes, deletions or small indels at any site within the genome and can be easily adapted to other cell-lines, genes or diseases.

7.1.4 Genetically proxied therapeutic PHD inhibition leading to long-term higher Hgb levels is not associated with cardiovascular risk or potential unwanted effects.

In **Chapter 6,** I used human genetic variants to genetically proxy therapeutic PHD inhibition and investigated the effects of long-term higher Hgb levels on risk of CVD or additional unintended effects. First, I identified eight variants associated with Hgb levels annotated to the three genes encoding the PHD enzymes with little evidence of pleiotropy. I used these drug-target specific variants as proxies for therapeutic PHD inhibition and found no evidence for an increased risk of CVD with long-term higher Hgb levels. By rescaling the effects to the minimal clinically significant difference of a 1-unit increase in Hgb levels, I could exclude odds of 1.35, 1.33, and 1.24 for CAD, MI, or stroke respectively. Second, I performed PheWAS to further characterise the therapeutic effects of long-term rises in Hgb levels and found evidence for an effect on NAFLD and several relevant liver- and kidney-function biomarkers. The *EGLN* genes are known to play a role in glucose metabolism and NAFLD is a common manifestation found in diabetic patients and is a prevalent cardiovascular risk factor in CKD patients. Third, I performed secondary two-sample MR analysis using 515 variants associated with circulating Hgb levels to investigate cardiovascular risk with general long-term rises in Hgb. There was strong evidence for pleiotropy and heterogeneity in the genetic variants, so I performed

Steiger filtering to obtain a more specific and powerful genetic instrument consisting of variants with a greater effect on Hgb levels than the disease outcome. Steiger filtering provided a better understanding of the true directionality. I found no increased risk of CVD with genetically proxied long-term rises in circulating Hgb levels and was able to exclude odds of 1.06, 1.08, and 1.08 for CAD, MI, and stroke respectively with a 1-unit increase in Hgb levels. Steiger filtering reduced the amount of pleiotropy and heterogeneity in the instruments and strengthened the results providing a better understanding of the true causal estimate.

These results indicate that elevated Hgb levels as induced by therapeutic PHD inhibition does not infer an increased risk of CVD or additional complications. The associated biomarkers identified could be used to predict CKD patients most at risk of developing CVD with therapeutic rises in Hgb levels.

7.2 Implications and Integration of findings

Over the last decade, researchers have demonstrated the utility of human genetics in aiding the drug development process. Genetic evidence can support the use of treatments for disease, contradict findings found in RCTs highlighting safety concerns, highlight repurposing opportunities, and implicate potential unintended effects that are not the primary concern in clinical trials (Bovijn et al., 2020; Lotta et al., 2016; Okada et al., 2014; Scott et al., 2016; Yarmolinsky et al., 2022). All of these findings emphasise how genetics can be used to increase the likelihood of a drug succeeding (King et al., 2019; Nelson et al., 2015; Plenge et al., 2013). There are several conflicting clinical studies regarding the cardiovascular safety of current treatments for anaemia in CKD that cause supra-physiological rises in circulating EPO levels (Di Lullo et al., 2015; Gupta & Wish, 2017; Heuberger et al., 2013; Krapf & Hulter, 2009; Lundby et al., 2007). These safety concerns have led to the development of several PHI compounds which act at the transcriptional level of EPO (Akizawa, Iwasaki, et al., 2020; Chertow et al., 2021; Dhillon, 2020). Successful noninferiority clinical trials have already shown that PHIs are able to maintain EPO levels within the physiological range and are noninferior to ESAs in terms of cardiovascular safety and hematologic efficacy (Chertow et al., 2021; K.-U.

Eckardt et al., 2021; Provenzano et al., 2021; A. K. Singh, Carroll, McMurray, et al., 2021; A. K. Singh, Carroll, Perkovic, et al., 2021). To my knowledge, there is no genetic evidence investigating the effects of therapeutic modulation of circulating EPO or Hgb levels to support the ongoing clinical development of PHIs. When performing genetic analyses, it is becoming increasingly difficult to know which variant is the most likely causal variant and whether this variant is having an effect on the expected gene. For genetics to support the drug development pipeline, additional evidence validating the genetic proxies is required to increase confidence of the conclusions drawn. One way of providing strong evidence to support a genetic variant as causal is to use functional studies. My study is novel because very few studies have performed functional validation of variants used in MR.

The findings presented in this thesis have the potential to be used in corroboration with results obtained from completed clinical trials into PHI treatment to impact patient care and improve treatment of anaemia in CKD. Understanding the relationship between EPO and/or Hgb levels and cardiovascular risk is an important and unresolved question and this research using genetic markers provides a valuable contribution to the field. The identification of relevant genetic markers could potentially inform further research using patient level clinical data from the phase III trials. The addition of these findings to the field could expedite the time for the drug to reach the market.

Although EPO is known to have a series of pleiotropic effects playing a key role in a wide-range of different organs, tissues and systems, the actual genes and downstream pathways remain elusive. The majority of current models for investigating EPO focus on the effects of rhEPO which results in supra-physiological levels. These supra-physiological levels are not biologically relevant and due to the complex negative feedback loops, do not provide the best understanding of physiological EPO. Through whole transcriptomic profiling of the *EPO* knock-out cell-lines, a clear list of DEGs strongly implicated in Notch signaling were identified. These findings add to the current literature which suggests a role for the Notch signaling pathway in response to hypoxia (Duarte et al., 2018; Gustafsson et al., 2005; Phillips et al., 2007). Notch signaling is an important pathway involved in a range of molecular and cellular functions including determination of cell fate, cell cycle activity and DNA repair (Kopan, 2012; Marignol et al., 2013). It could be through this pathway that EPO exerts its pleiotropic and mitogenic effects. Understanding the link between these pathways could provide additional insight into the potential effects of higher EPO levels and aid future treatment of anaemia in CKD or any additional diseases which involve EPO or Notch signaling. Additional biological pathways and molecular functions involving ATP production, cellular oxidation, aerobic respiration and mitochondrial function were also revealed by RNA-seq analysis and provide a better understanding of the potential internal cellular signaling effects of EPO particular in high energy-dependent organs such as the kidneys, heart, brain, liver, and skeletal muscle (Console et al., 2020). Previous studies have suggested EPO to be protective against cellular stresses and injury by influencing several of these functions (Hernández et al., 2017). These findings could be important in highlighting potential opportunities of PHI drug repositioning for additional diseases caused by dysregulation of these pathways and functions, such as ischemia, neurodegenerative diseases in the brain, reperfusion injury, and fibrosis (Console et al., 2020; Hernández et al., 2017; Junk et al., 2002; Suresh et al., 2020; X. Wang et al., 2020).

The methods and approaches outlined throughout this thesis can be translated to the identification and functional validation of genetic variants for use as partial proxies to mimic the effects of any drug treatment or target. Integrating this kind of methodological and functional approach into the epidemiology field could have an important and impactful use in better refining genetic findings and aiding the drug development process. The molecular approaches used, particularly those to establish the heterozygous polymorphism at rs1617640, could become the gold standard for testing whether a variant is causal for a given phenotype, especially if models of the complete allelic series (major/major, major/minor and minor/minor) were generated (J. Lin & Musunuru, 2018).

7.3 Future Directions

As the number and size of genetic studies increases further, the number of common genetic variants associated with complex traits will increase, but many of these are likely to have even smaller effect sizes than those already identified

(Visscher et al., 2017). Future studies will need to focus on combining genetics with additional sequence-based data and experimental perturbations. Technologies used during this PhD will therefore require upscaling, including genome-wide CRISPR screens, to overcome the challenges associated with small effect sizes and the identification of the most likely causal gene (Bodapati et al., 2020; McGuire et al., 2020).

Making use of the increasing amount of genetic data to aid our decisions in clinical practice and developing tools which can help the integration of genetics into the drug development process is the ultimate goal. Tools are already being developed to use genetics during the drug discovery phase. For example, the GREP software aids quantification of whether GWAS signals are enriched for drug targets capturing potential repositionable drugs (Sakaue & Okada, 2019). The development of these tools highlights how genetics is becoming the forefront of aiding the drug development process emphasising the importance and timely nature of my work.

Previous studies have shown how gain-of-function variants can be used to mimic agonistic (Lotta et al., 2019) drug effects whilst loss-of-function variants can be used to mimic antagonistic effects (Jørgensen et al., 2014; Minikel et al., 2020). When focusing on identifying valid genetic proxies for therapeutic action, it is therefore important to also consider the functional consequence of the variant. I, and many others, have until now primarily focused on common genetic variation for use to mimic pharmaceutical action (Gill et al., 2019; Lotta et al., 2016; Okada et al., 2014; Scott et al., 2016; Swerdlow et al., 2015). However, common variation is not the only type of variation to give rise to disease. Rare variants, structural variants (e.g. copy number variations, insertions or deletions) and alterations in epigenetic marks could also drive associations (Eichler, 2019). Investigation of these other forms of variants is important to refine causal regions and improve understanding of disease. The larger effect size, rare variants, especially those leading to loss-of-function, provide a natural setting to mimic antagonistic therapeutic effects and have already proven useful in assessing clinical consequences (Cohen et al., 2006; Jørgensen et al., 2014; Stitziel et al., 2014). Large-scale whole-exome and whole-genome sequencing data (WES and WGS) has recently been released by UKB to enable identification of novel rare variant-trait associations with large effect sizes bridging the gap between common and rare variation (Backman et al., 2021; Halldorsson et al., 2021). Exome data identifies rare variants lying within the protein-coding regions and aids in understanding gene function, disease mechanisms and phenotypic consequences of protein-altering variation. WES data can inform the medical actionability of rare variants and corroborate the link between disease associations and biomarker levels enabling translation to the effects of clinical-stage drug targets (Backman et al., 2021; Sun et al., 2021). A preliminary look at the WES data in the first UKB release (in 200,000 individuals) revealed associations between rare loss-of-function variants or deleterious missense variants (CADD > 30) in the *EPO* and *EGLN* genes with Hgb levels and erythrocyte number in the expected direction. These rare variants were not associated with CVD providing further evidence that drugs acting through these genes are unlikely to infer added cardiovascular risk than current treatments. These preliminary findings emphasise the utility of rare variants in acting as proxies for testing therapeutic effects. Future analysis in larger sample sizes (e.g. in the most recent 450,000 WES release) will increase the power to detect rare variant-trait associations and will provide valuable additional insight into the long-term therapeutic effects of modulating circulating Hgb or EPO levels. WGS will also provide insight into the role of non-coding rare variants, particularly microsatellites and structural variants which are more likely to have functional impact through affecting non-coding genes, RNA and protein expression (Sudmant et al., 2015; Weischenfeldt et al., 2013; J. Zheng et al., 2020). Interrogating this data has great potential in identifying genetic variants that can better mimic therapeutic PHD inhibition improving our understanding of gene function and mechanisms and providing additional genetic evidence to support clinical trial data.

In this work, I performed *in vitro* molecular techniques to validate the *cis-EPO* variant as having an allele-specific effect on EPO levels. However, as the variant lies within the *EPO* regulatory region, it is likely that the variant alters transcription factor binding to change gene expression. Previously Tong et al. (2008) used *in silico* tools to predict the effect of allele changes at rs1617640 on potential transcription factor binding. As these computational methods to predict potential transcription factor binding sites have since improved (C. Chen et al., 2021; Jayaram et al., 2016; Zeng et al., 2020), it would be worth repeating to see if a transcription factor motif lies around the *cis-EPO* SNP and whether polymorphisms at this position affect the binding of any tissue-relevant transcription factors. These kinds of investigations would help guide future functional work on the knock-in model and would provide insight into the regulatory factors important for controlling *EPO* gene expression in different tissues which could lead to identification of novel drug targets for treating anaemia in CKD.

As I primarily focused on using genetic variants which lie at the biomarker level (i.e. the EPO level) or at the drug target gene level (*EGLN1/2/3*), I could have missed unintended drug effects. Alternative mechanistic pathways which are EPO-dependent but associated with additional biomarker levels, such as iron, could also be impacted by PHD inhibition and drive the cardiovascular risk found with supra-physiological levels. Identification of genetic variants associated with iron or other relevant biomarkers which lie within the PHI drug target genes would be worth investigation to see if there are any potential unintended effects through these alternative pathways. The same limitation goes with focusing only on certain genes; variants lying within other hypoxia-response genes, such as the *hepcidin* gene which increases iron levels in an attempt to restore oxygen availability, could also be worth investigation as these are also impacted during anaemia and by PHI treatment. This would further improve our current understanding of the pathophysiology of anaemia. I have also only focused on the therapeutic effects of PHIs and therefore my results do not necessarily apply to other treatments that may target the hypoxic pathway through other mechanisms e.g. HIF asparaginyl hydroxylase inhibitors. Future genetic analyses of genes involved in these mechanisms would provide an increased understanding of the underlying disease mechanism and may elucidate to other potential safe and efficient therapeutic targets.

When focusing on the *EPO* gene, I only identified one genetic variant lying nearby to the *EPO* gene associated with circulating EPO levels. However, I was limited by power due to sample size. As larger studies become available with deeper phenotyping and additional biomarkers measured and more summary statistics are made publicly available, the potential to perform additional GWAS and/or meta-analysis on EPO will be possible. Increased sample sizes will increase the power to identify additional variants that could be used as proxies to mimic therapeutic elevations of endogenous EPO levels and overcome the potential weak instrument bias with a single variant (Burgess & Thompson, 2010). A GWAS on cytokine levels, including EPO levels, in 10,000 Danish individuals has recently been published (Y. Wang et al., 2020) and therefore a meta-analysis of this GWAS with the GWAS meta-analysis performed in this thesis would be one way of increasing power for detecting additional EPO associations imminently.

As gene-editing continues development, there will be more opportunities for functional studies to validate genetic findings. Here, I have shown how one approach combining CRISPR-Cas9 gene-editing and the *piggyBac^{TM}* transposon system can be used to knock-in alternate alleles. It is also possible to validate genetic variants as causal by performing SNP knock-out studies which may be easier and would take advantage of the increased efficiency of NHEJ-mediated disruption compared to HDR (J.-P. Zhang et al., 2017). This could be done with or without precision by either introducing an insertion or deletion of a random size into the SNP-site disrupting gene function or the regulatory element, or by using two gRNAs that generate two DSBs flanking the SNP, respectively (J. Lin & Musunuru, 2018). These approaches could be relatively easily integrated into standard practice to validate genetic variants as causal particularly as larger gRNA libraries become available and companies begin to offer generation of particular cell-line models (Sanson et al., 2018). Furthermore, alternative methods for introducing single-base gene-edits such as base editors composed of a catalytically impaired Cas9 and base modification enzymes will further improve enabling any modification, not just C/G into T/A and vice versa (Shuquan Rao et al., 2021).

7.4 Limitations

As discussed in each of the individual chapters, there are several limitations that need to be taken into consideration of which I will address in more detail below.

7.4.1 Genetics to aid drug development

When using genetic variants as proxies for drug treatments to anticipate potential effects of pharmacological manipulation, it is important to identify specific variants that accurately mimic therapeutic actions. First, it is important to consider the choice of variant in terms of how well the SNP mimics the therapeutic effects. I used GWAS data and limited variants to those most likely to have a functional impact due to lying nearby or within the drug target gene. However, this variant may not be the best proxy for the drug. GWAS only highlight correlations between genetic variants and phenotypes and do not inform on which associated genetic variants are the true causal variants (Visscher et al., 2017). Due to linkage disequilibrium, highly correlated neighboring genetic variants tend to be inherited together making it difficult to distinguish the true causal variant (Pers et al., 2015; Schaid et al., 2018). Additionally, GWAS identified loci frequently contain multiple genes and therefore it can be difficult to determine which gene is being impacted by the identified variant (Shu et al., 2018). The fact that over 90% of common variants lie within non-coding regions does not help linking associated variants to a candidate gene because of the complex genomic structure (Cano-Gamez & Trynka, 2020). Hence, future research integrating all forms of sequence-based omic data is needed.

Although I selected Hgb-associated variants annotated to the *EGLN* genes in **Chapter 6** and PheWAS revealed strong associations with relevant phenotypes indicating that they likely impact Hgb levels through the hypoxic pathway, these variants may not be the causal variants actually affecting *EGLN* expression and therefore could be invalid instruments to proxy PHD inhibition. Further validation, such as colocalisation or functional analysis, could be useful. A major challenge of GWAS is the inability to detect all causal variants or all SNPs correlated with the causal variant (Rohde et al., 2018). This is commonly due to a lack of power as a result of sample size alongside the inclusion of a stringent significance threshold to control for false-positives (Rohde et al., 2018). The majority of SNPs have small effect sizes and these will therefore remain undetected unless sample sizes increase substantially (>100,000) (Visscher et al., 2017). The EPO meta-analysis performed in **Chapter 3** only included 6,127 individuals and therefore sample size was likely a limiting factor in detecting causal variants. Furthermore, the use of the typically accepted stringent *P*-value threshold of 5 x 10^{-08} (which is based upon a Bonferroni correction of 1 million independent tests) limits the power to detect associations (Dudbridge & Gusnanto, 2008). The number of tests carried out in some GWAS may be considerably lower than this due to being limited to European individuals and to common SNPs with a frequency greater than 1% of which many will be in high LD. This does therefore not necessarily translate to an increased number of independent tests (Auton et al., 2015). The choice of this benchmark for determining statistical inference remains under debate and requires careful consideration especially when focusing on low-frequency variants (Fadista et al., 2016).

Second, the strength and validity of the variant need considering when implementing drug-target specific MR. MR is based upon three assumptions; the genetic variant is associated with exposure, the genetic variant is only associated with the outcome through the exposure, and the genetic variant is not associated with any confounders of the exposure-outcome association (Davies et al., 2018). Ensuring that these assumptions are met can be difficult and despite several methods being developed to test these assumptions (Bowden et al., 2015; Hemani, Tilling, et al., 2017; Verbanck et al., 2018; Q. Zhao et al., 2018), violations can still happen. Biases can occur in causal estimates if variant-exposure estimates and variant-outcome estimates have been obtained from overlapping study sample participants resulting in an underestimation of the true causal effect, known as Winner's curse (Lawlor, 2016). I have attempted to reduce the risk of Winner's curse by obtaining association statistics from independent studies. I have, however, used the same individuals for discovery of the genetic instrument and extraction of the variant-association statistic resulting in potential bias towards the null (Burgess et al., 2016). For the EPO analysis (**Chapter 3).** I used one genetic instrument and despite performing two-sample MR to overcome the risk of inflated Type 1 errors, the estimate may be biased towards the null leading to lower power to detect a causal effect (Davey Smith & Hemani, 2014). This bias is less serious than a bias in the direction of the observational effect but could increase the chance of Type 2 error (Lawlor, 2016).

The most important limitation when using genetics to proxy therapeutic effects is that genetic variants indicate lifelong perturbations of smaller and subtler effects compared to short-term, larger effects of a drug at a particular time in an individual's life (Pulley et al., 2017). Therefore, the effect size of the genetic estimate may not be physiologically relevant. Despite rescaling the estimated genetic effects to the drug-induced effects to try and overcome this, from a clinical perspective, the absence of a statistically significant difference can be of limited value (Page, 2014). Although I found no statistical evidence for increased cardiovascular risk with genetically proxied higher Hgb or EPO levels, it does not automatically imply that the treatment will be clinically safe and effective. This is due to small sample sizes and measurement variability potentially influencing statistical results. For this reason, the upper bounds of the confidence intervals, which provide more information regarding directions and magnitude, were used to draw conclusions about the likely impact of biomarker levels on risk of disease. However, limitations remain in the ability to draw a strong and clear indication of the long-term effects of treatments by these upper confidence intervals so I cannot completely rule out an adverse effect.

Genetic variants are also studied at population level in relatively healthy individuals compared to the target patient population (Mokry et al., 2015). There is uncertainty about the effects higher levels may have in the target population due to varying baseline biomarker levels, the titration to a particular level resulting in different individual-level biomarker increases, the presence of other underlying comorbidities and the use of other medications which may alter risk (Burgess et al., 2012). Stratifying genetic analyses by biomarker levels can alter causal estimates and may provide more precise, clinically relevant effect estimates (Sofianopoulou et al., 2021).

7.4.2 Functional validation studies using cell-line models

To establish the cell-lines of interest, I have adapted CRISPR-Cas9 gene-editing techniques. Irrespective of ongoing advancements in CRISPR-Cas9 gene-editing, concerns remain regarding the validity of findings. The primary concern is risk of off-target effects associated with the gRNA sequences binding to other non-specific regions of the genomic DNA and introducing unwanted edits (Fu et al., 2013; Hsu et al., 2013). For the *EPO*$^{-/-}$ knock-outs, I used paired gRNAs targeting the exonic regions of *EPO*. Using paired gRNAs has shown to improve the efficiency of obtaining homozygous knock-outs and reduce the risk of off-target effects due to needing both gRNAs to bind to introduce the most significant impact on gene expression (Ran, Hsu, Lin, et al., 2013). Ensuring gRNAs have few matches to other exonic regions and the highest possible off-target score (highest score represents less off-target cutting) also improves the specificity and reduces disruption elsewhere (D. B. Graham & Root, 2015). A more comprehensive genome-wide screening would be needed to definitively assess off-target effects (D. Kim et al., 2015).

For the single SNP knock-in, I was limited in my choice of potential gRNAs by the location of the genetic variant in the non-coding upstream region of *EPO*. Targeting non-coding regions of the genome increases the risk of off-target effects due to the presence of highly repetitive sequences and few potential gRNA sequences (Tycko et al., 2019). However, the use of the *piggyBac*^{TM} system with CRISPR-Cas9 reduced potential off-target effects as I could screen for integration of the transposon in the correct genomic location (L. Yang et al., 2013). Off-target effects could be introduced through this combination of approaches after the excision of the *piggyBac*^{TM} transposon as the transposon may become reintegrated elsewhere in the genome (M. A. Li et al., 2013). Moreover, despite making every effort to ensure footprint-free removal of the transposon and screening for these using PCR and Sanger sequencing*,* some marks could be introduced during homologous recombination further up- or downstream of the sequenced region. As the field advances, tools for designing gRNAs and predicting or screening for potential off-target effects will improve (Kang et al., 2020; D. Wang et al., 2019; H. Wang et al., 2016). Modified Cas9 proteins and other enzymes that have been optimised for reducing off-targets will also become readily available. As CRISPR screens become more commonly used introducing specific single-base gene-edits for functional validation studies will become easier (C.-L. Chen et al., 2020; Naeem et al., 2020).

The *piggyBac*^{TM} transposon system is a relatively novel technique and although it increases the chance of obtaining precise gene-edits, the technique relies upon several selection processes and the introduction of a transposase (Yusa, 2013). Screening through surviving clones makes it a relatively inefficient method of isolating precisely edited cell-lines (Steyer et al., 2018). The resistance gene could be randomly integrated anywhere in the genome and thus those clones surviving selection may not necessarily be edited in the desired place. The use of puromycin and FIAU is widely used for selection of genetically-engineered cells, however, these drugs can cause changes to the transcriptome and it is difficult to screen for these resulting changes (Aviner, 2020).

I have used HEK-293 cell-lines as the model of choice due to the cells being from a biologically relevant tissue type. HEK-293 cells are widely used, easy to transfect and enable transient and stable expression (P. Thomas & Smart, 2005). However, they have a complex karyotype making it difficult to fully establish whether all chromosomal copies have been affected and contain the desired gene-edit (Stepanenko & Dmitrenko, 2015). HEK-293 cells may not have been the best and more representative cell-line model to use as HEK-293 cells are from an embryonic kidney cell-line, and therefore may not be fully representative of the adult kidney where EPO exerts its primary effects and in particular a diseased kidney. Alternative cell-lines could be used to repeat this work to validate the findings, such as renal EPO producing cells (REPCs) or peripheral blood stem cells. Cells were also grown in a 2D culture with only one cell-type present. The human body is more complex than a single monolayer of cells and there are a multitude of different pathways and cells working together to elicit an effect and a particular phenotype. EPO is a cytokine hormone primarily released in response to hypoxia and exerts its pleiotropic effects in a wide-range of different cell-types, tissues and systems (Hernández et al., 2017). EPO is involved in number of signaling networks between different systems; it is difficult to represent these complex networks in a cell model. As *EPO* expression relies upon several cellular responses, basal levels are relatively low in many cell-lines, including HEK-293 cells, and this is worth considering when attempting to investigate alterations in EPO and may be the reason most current studies investigate the effects of exposure to exogenous EPO. It would be beneficial to study changes related to *EPO* knock-out or SNP knock-in in additional relevant cell-lines for comparison, such as HepG2 or peripheral blood stem cells. Whole organisms could also be used, such as mouse, but the translational ability of these models is limited due to not being human. Furthermore, my results, alongside previous studies, indicate an allele-specific effect of the *cis-EPO* variant which differs in different tissues, cell-types and diseases. Expression may only be affected at a particular developmental timing, such as in the embryo or foetus, at difference stages of the cell-cycle or in response to external/environmental factors. This could explain why I found altered *EPO* expression in HEK-293 cells with an allele change at rs1617640 **(Chapter 5)** but no association of the variant with renal *EPO* expression in the eQTL analysis (although this may be due to lack of statistical power) **(Chapter 3).** These limitations highlight the need for further investigation in a multitude of different models, such as those at different developmental stages and exposed to different external stresses.

7.5 Summary

The work presented in this thesis shows how a combination of genetic and functional approaches can be used to better understand the therapeutic profile of pharmaceutical treatments. Focusing on novel treatments for anaemia in CKD and risk of cardiovascular disease, I have provided genetic evidence indicating that therapeutic modulation of Hgb or EPO levels do not increase cardiovascular disease risk, with upper limits of 1.35 and 1.07 for Hgb and EPO respectively. I have shown how genetic analyses combined with functional validation studies is a powerful approach to identify relevant genetic markers that can investigate the long-term effect of therapeutic action. The combination of these approaches can be used across the field to help refine genetic findings improving our ability to genetically proxy long-term therapeutic modulation and our understanding of the mechanisms underlying complex traits and disease aetiology.

Chapter 8 Appendix

Appendix 1: Scripts for RNA-seq analysis

8.1.1 Trimming of reads using Trimmomatic

```
#!/bin/bash

## Trimming using Trimmomatic
## CHARLI HARLOW

# Download Trimmomatic

wget
http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/Trim
momatic-0.39.zip

# Unzip Trimmomatic
unzip Trimmomatic-0.39.zip

# Load in java module needed for Trimmomatic
module load java/1.8.0_92

# may need to edit depending on settings required
# For more details about the trimmomatic package please see
http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/Trim
momaticManual_V0.32.pdf

# Paired-End
for ID in Empty1 Empty2 Empty3 Empty4 KOA1 KOA2 KOA3 KOA5 KOB1
KOB211 KOB3 KOB5
do
java -jar ../Trimmomatic-0.39/trimmomatic-0.39.jar
PE \ #specify paired-end or single-end
-phred33 #specifies the base quality encoding. Could be changed to
phred64 if using older sequencing machines
-trimlog trim_empty1_101219.log \ # creates a log file of all the
trimming
-basein /gpfs/mrc0/projects/Research_Project-
MRC158833/cs660/EPO_project/EPO_RNA_Seq/ftp1.sequencing.exeter.ac.u
k/H0253/11_trimmed/3064_Empty1_trimmed_r1.fq.gz \ # specifies the
input file name. Use this option if all input names have the same
common naming pattern so the reverse read file can automatically be
detected. If not then remove this option and just specify the two
files e.g input_filename_r1.fq.gz input_filename_r2.fq.gz
-baseout 3064_Empty1_trimmed_r1_filtered.fq.gz \ # specifies the
output file names. Four files will be produced 1-Paired forward
reads, 2-Unpaired forward reads 3-Paired reverse reads 4-Unpaired
reverse reads
ILLUMINACLIP:/gpfs/mrc0/projects/Research_Project-
MRC158833/cs660/EPO_project/EPO_RNA_Seq/ftp1.sequencing.exeter.ac.u
k/H0253/11_trimmed/Trimmomatic-0.39/adapters/TruSeq3-
PE.fa:2:30:10:2:keepBothReads \
HEADCROP:10 \ #Cut the specified number of bases from the start of
the read. Particular important to add this step in if looking at
the QC report and realising that the first 10 bases for example
have a phred score lower than
```

```
LEADING:3 \ #Cut bases off the start of a read, if below a
threshold quality
TRAILING:3 \ # Cut bases off the end of a read, if below a
threshold quality
SLIDINGWINDOW:4:15 \ #Performs a sliding window trimming approach.
It starts scanning at the 5' end and clips the read once the
average quality within the window falls below a threshold.
MINLEN:36 \ # Drop the read if it is below a specified length

# Single-End
# Remove the hashtags if wanting to run for single ended
# for ID in Empty1 Empty2 Empty3 Empty4 KOA1 KOA2 KOA3 KOA5 KOB1
KOB211 KOB3 KOB5
# do
# java -jar /gpfs/mrc0/projects/Research_Project-
MRC158833/cs660/EPO_project/EPO_RNA_Seq/Trimmomatic-
0.39/trimmomatic-0.39.jar SE -phred33 -trimlog
trim_{$ID}_101219.log /gpfs/mrc0/projects/Research_Project-
MRC158833/cs660/EPO_project/EPO_RNA_Seq/ftp1.sequencing.exeter.ac.u
k/H0253/11_trimmed/3064_{$ID}_trimmed_r1.fq.gz
3064_{$ID}_trimmed_r1_extratrimming.fq.gz
ILLUMINACLIP:/gpfs/mrc0/projects/Research_Project-
MRC158833/cs660/EPO_project/EPO_RNA_Seq/Trimmomatic-
0.39/adapters/TruSeq3-SE.fa:2:30:10 HEADCROP:10 SLIDINGWINDOW:4:15
MINLEN:36
# done
```

8.1.2 Quality controls checks using FastQC and/or MultiQC

```
#!/bin/bash

## Quality Control using FASTQC & MULTIQC
## CHARLI HARLOW

# create directory to store the fastqc output in - one for each
sample and each read

for ID in Empty1 Empty2 Empty3 Empty4 KOA1 KOA2 KOA3 KOA5 KOB1
KOB211 KOB3 KOB5
do
mkdir -p /gpfs/mrc0/projects/Research_Project-
MRC158833/cs660/EPO_project/RNA_sequencing_analysis/ftp1.sequencing
.exeter.ac.uk/H0253/11_trimmed/additional_trimming/additional_trimm
ing_fastqc/${ID}_r1/
mkdir -p /gpfs/mrc0/projects/Research_Project-
MRC158833/cs660/EPO_project/RNA_sequencing_analysis/ftp1.sequencing
.exeter.ac.uk/H0253/11_trimmed/additional_trimming/additional_trimm
ing_fastqc/${ID}_r2/

# Load in module to run FastQC
module load FastQC/0.11.7-Java-1.8.0_162

# Run FastQC
# Read1
fastqc /gpfs/mrc0/projects/Research_Project-
MRC158833/cs660/EPO_project/RNA_sequencing_analysis/ftp1.sequencing
.exeter.ac.uk/H0253/11_trimmed/additional_trimming/${ID}_additional
_trimming_r1.fq.gz --extract -o
./additional_trimming_fastqc/${ID}_r1/
# Read2
```

```
fastqc /gpfs/mrc0/projects/Research_Project-
MRC158833/cs660/EPO_project/RNA_sequencing_analysis/ftp1.sequencing
.exeter.ac.uk/H0253/11_trimmed/additional_trimming/${ID}_additional
_trimming_r2.fq.gz --extract -o
./additional_trimming_fastqc/${ID}_r2/

done

# Run MultiQC
# Change directory to faatQC directory
cd ./additional_trimming_fastqc/


# Load in module
module load MultiQC/1.2-intel-2017b-Python-2.7.14

# Run multiQC
multiqc . --dirs --interactive -o ./multiQC/


# Combine all picture files together to make one pdf

# For all the ones which have warnings or have failed
for ID in Empty1 Empty2 Empty3 Empty4 KOA1 KOA2 KOA3 KOA5 KOB1
KOB211 KOB3 KOB5
do
grep -v PASS ${ID}/3064_${ID}_trimmed_r1_fastqc/summary.txt
|montage txt:-${ID}/3064_${ID}_trimmed_r1_fastqc/Images/*png -tile
x3 -geometry +0.1+0.1 -title ${ID} ${ID}.png done < file_names.txt

convert *png fastqc_summary_warnings.pdf

# For all those which have PASSED
for ID in Empty1 Empty2 Empty3 Empty4 KOA1 KOA2 KOA3 KOA5 KOB1
KOB211 KOB3 KOB5
do
grep PASS ${ID}/3064_${ID}_trimmed_r1_fastqc/summary.txt |montage
txt:-${ID}/3064_${ID}_trimmed_r1_fastqc/Images/*png -tile x3 -
geometry +0.1+0.1 -title ${ID} ${ID}.png done < file_names.txt

convert *png fastqc_summary_passed.pdf
```

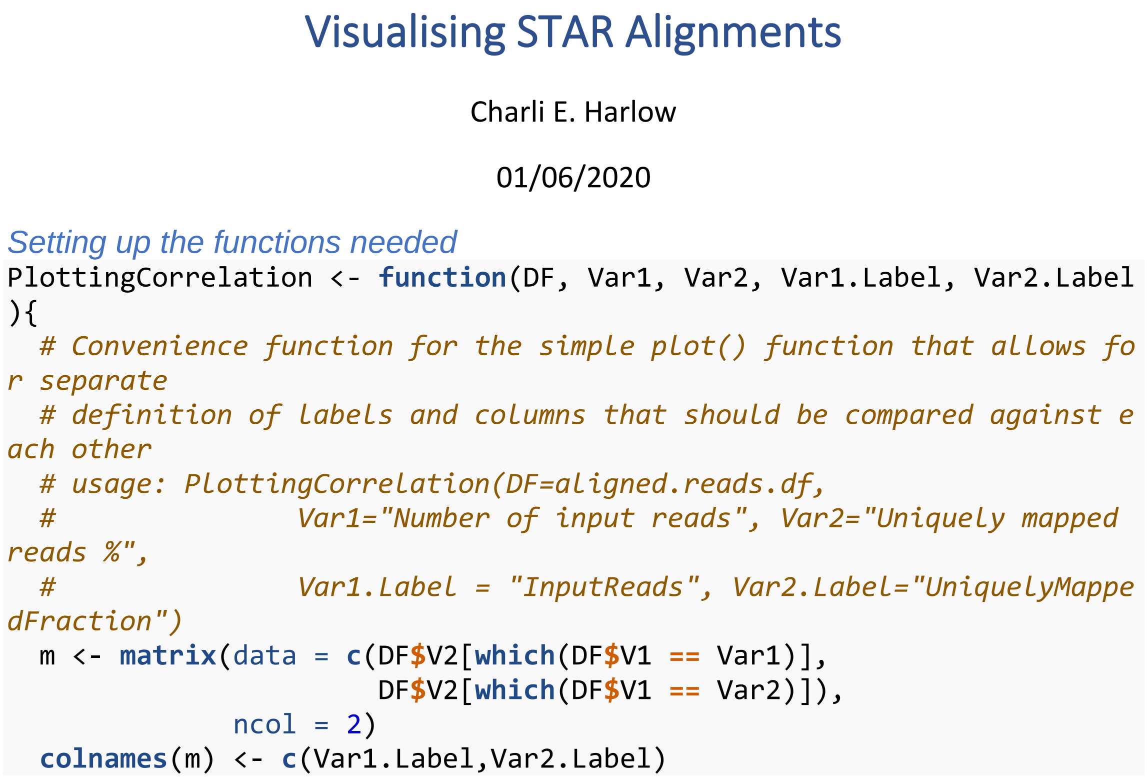
8.1.3 Alignment to the reference genome using STAR

```
#!/bin/bash

## STAR ALIGNMENTS
## CHARLI HARLOW

# 1) Download Reference Genome
# a) Make directory to store reference genome in
mkdir Ref_genome

# b) Download fast files for reference genome
wget ftp://ftp.ensembl.org/pub/release-
98/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.primary_assembly.
fa.gz

# c) Download the gtf file for the reference genome being used:
```

```
wget ftp://ftp.ensembl.org/pub/release-
98/gtf/homo_sapiens/Homo_sapiens.GRCh38.98.gtf.gz

# d) Unzip both these files for use in STAR

gunzip Homo_sapiens.GRCh38.98.gtf.gz
gunzip  Homo_sapiens.GRCh38.98.gtf.gz

# 2) Generate Index file
# only one index file needs to be created per genome. The index
file will contain all the information from the reference genome in
a compressed format that is optimized for efficient access and
comparison with the query read sequences. The main input files for
this step therefore encompass the reference genome sequence and an
annotation file.

# a) Create directory to store the index file in
mkdir /gpfs/mrc0/projects/Research_Project-
MRC158833/cs660/EPO_project/RNA_sequencing_analysis/StarIndex

# b) create alias for the directory with reference genome
Ref_dir="/gpfs/mrc0/projects/Research_Project-
MRC158833/cs660/EPO_project/EPO_RNA_Seq/Ref_genome"

# c) load in STAR module
module load STAR/2.7.1a-foss-2018b

# d) Generate STAR index files

STAR --runMode genomeGenerate \ #specify run mode as generate index
--genomeDir /gpfs/mrc0/projects/Research_Project-
MRC158833/cs660/EPO_project/RNA_sequencing_analysis/StarIndex/ \
#where to store index output
--genomeFastaFiles ${Ref_dir}/GRCh38.p13.genome.fa \ # specify
where fasta reference file is, make sure this file is unzipped
--sjdbGTFfile
${Ref_dir}/gencode.v32.chr_patch_hapl_scaff.annotation.gtf \
#specify where gtf reference file is. Make sure this file is
unzipped
--sjdbOverhang 74 # specifies the length of the genomic sequence
around the annotated junction to be used in constructing the splice
junctions database. Ideally, this length should be equal to the
ReadLength-1, where ReadLength is the length of the reads.

# 3) Run alignment for each sample. If Single End then need to
change options and input names

# a) Make directory to store results
mkdir /gpfs/mrc0/projects/Research_Project-
MRC158833/cs660/EPO_project/RNA_sequencing_analysis/Star_Alignment/

# b) Create alias for index directory
index_dir="/gpfs/mrc0/projects/Research_Project-
MRC158833/cs660/EPO_project/EPO_RNA_Seq/StarIndex”

# run loop to carry out alignment for each sample
# Make sure when reading files in if you are mapping paired end
data then read1 and read2 and space separated NOT separated by
comma. Comma separated list means both files are mapped in one job.
```

```
for SAMPLE in Empty1 Empty2 Empty3 Empty4 KOA1 KOA2 KOA3 KOA5 KOB1
KOB211 KOB3 KOB5
do
STAR --runMode alignReads \ #specify run mode as aligning reads
--genomeDir ${index_dir}/ \ #tell star where to reference file is
--readFilesIn /gpfs/mrc0/projects/Research_Project-
MRC158833/cs660/EPO_project/RNA_sequencing/analysis/ftp1.sequencing
.exeter.ac.uk/H0253/11_trimmed/additional_trimming/${SAMPLE}_additi
onal_trimming_1P.fq.gz /gpfs/mrc0/projects/Research_Project-
MRC158833/cs660/EPO_project/RNA_sequencing/analysis/ftp1.sequencing
.exeter.ac.uk/H0253/11_trimmed/additional_trimming/${SAMPLE}_additi
onal_trimming_2P.fq.gz \ #read in the fast q files for aligning
--readFilesCommand zcat \ #tells STAR that the fastq files are
gzipped
--outFileNamePrefix /gpfs/mrc0/projects/Research_Project-
MRC158833/cs660/EPO_project/RNA_sequencing_analysis/Star_Alignment/
${SAMPLE}_ \ #specifies name of output file
--outSAMtype BAM SortedByCoordinate \ # output in BAM format,
sorted by coordinate. This option will allocate extra memory for
sorting which can be specified by –limitBAMsortRAM
--outReadsUnmapped Fastx \# default is none, fastx will output
unmapped and partially mapped reads in separate files
--runThreadN 4 \ #tells star how many threads to run on
--twopassMode Basic \# STAR will perform mapping, then extract
novel junctions which will be inserted into the genome index which
will then be used to re-map all reads
--outFilterMultimapNmax 1 # only reads with 1 match in the
reference will be returned as aligned
done
```

8.1.4 Visualising STAR alignments



Visualising STAR Alignments

Charli E. Harlow

01/06/2020

Setting up the functions needed

```
PlottingCorrelation <- function(DF, Var1, Var2, Var1.Label, Var2.Label
){
  # Convenience function for the simple plot() function that allows fo
r separate
  # definition of labels and columns that should be compared against e
ach other
  # usage: PlottingCorrelation(DF=aligned.reads.df,
  #              Var1="Number of input reads", Var2="Uniquely mapped
reads %",
  #              Var1.Label = "InputReads", Var2.Label="UniquelyMappe
dFraction")
  m <- matrix(data = c(DF$V2[which(DF$V1 == Var1)],
                      DF$V2[which(DF$V1 == Var2)]),
              ncol = 2)
  colnames(m) <- c(Var1.Label,Var2.Label)
  plot(m)
}
```

Setting up colour for plot

```
library(RColorBrewer)
nb.cols <- 4 # change this to number of coloumns
mycolors <- colorRampPalette(brewer.pal(4, "YlOrRd"))(nb.cols)
```

Setting up the function to produce the plot of 4 side by side

```
PlottingAlignmentResults <- function(Filter, DF, Legend=TRUE, PlotMedian = TRUE){
  # this function extracts those lines that correspond to the value stored in Filter and generates a bar plot where each sample is shown with a different color

  library(ggplot2)
  library(grid) # for unit() function
  filtered.df <- DF[which(DF$V1 == Filter),]
  medians <- as.data.frame(aggregate(V2~sample, data=filtered.df, FUN=median))
  filtered.df <- merge(filtered.df, medians, by.x = "sample", by.y = "sample", all.x=TRUE)

  p <- ggplot(data=filtered.df, aes(fill=replicate, y=V2.x, x=sample)) +
    geom_bar(stat="identity",position=position_dodge()) +
    theme_bw(base_size = 10) +
    scale_fill_manual(values=mycolors) +
    theme(legend.position="bottom",
          legend.text = element_text(size = 8),
          legend.key.size = unit(0.1, "cm"),
          legend.title=element_blank(),
          axis.title.x = element_text(size=10),
          axis.text.x = element_text(size=8)) +
    coord_flip() + ylab("") + ggtitle(Filter)

  if(PlotMedian){
    p <- p + geom_errorbar(aes(y=V2.y, ymax=V2.y, ymin=V2.y), linetype="dashed")
  }

  if(!Legend){
    p <- p +  theme(legend.position="none")
  }

  return(p)
}
```

Set up the Legend function

```
ExtractLegend <- function(Plot){
  library(ggplot2)
  G <- ggplotGrob(Plot)$grobs
  Legend <- G[[which(sapply(G, function(x) x$name) == "guide-box")]]
  Lheight <- sum(Legend$height)
  return(list(legend = Legend, lheight=Lheight))
}
```

```r
Extract.Histo.Info <- function(InList, # output from hist(...,plot=FALSE)
                               Percentage = TRUE
){
  # this function uses the information from hist() that are stored in lists
  # to make a data frame that's suitable for bar plots
  ll <- length(InList$breaks)
  out.df <- data.frame(breaks1 = InList$breaks[c(1:ll-1)],
                       breaks2 = InList$breaks[c(2:ll)],
                       counts = InList$counts)
  if(Percentage){
    out.df <- transform(out.df,
                        breaks = paste(out.df$breaks1*100, "-", out.df$breaks2*100, sep = ""),
                        breaks1 = NULL, breaks2 = NULL)
  }else{
    out.df <- transform(out.df,
                        breaks = paste(out.df$breaks1, "-", out.df$breaks2, sep = ""),
                        breaks1 = NULL, breaks2 = NULL)
  }
  out.df$breaks <- factor(out.df$breaks, levels = unique(as.character(out.df$breaks)), ordered = TRUE)
  return(out.df)
}
```

Load in libraries

```r
library(Cairo)
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.6.2

library(gridExtra) # for composite plotting of ggplots
```

Read in the final.log.out files that were produced from Star Alignment

```r
infiles <- list.files(path="~/Downloads/",pattern="Log.final.out", full.names = TRUE) # listing the files to be read in
```

Check the list looks as expected

```r
head(infiles)

## [1] "/Users/cs660/Downloads//Empty1_10bp_Log.final.out"
## [2] "/Users/cs660/Downloads//Empty2_10bp_Log.final.out"
## [3] "/Users/cs660/Downloads//Empty3_10bp_Log.final.out"
## [4] "/Users/cs660/Downloads//Empty4_10bp_Log.final.out"
## [5] "/Users/cs660/Downloads//KOA1_10bp_Log.final.out"
## [6] "/Users/cs660/Downloads//KOA2_10bp_Log.final.out"
```

Generate a list of data frames from the list of file we read in

```r
align.results <- lapply(infiles, function(x) read.table(x, sep="|", strip.white=TRUE, stringsAsFactor=FALSE, skip=3, fill = TRUE, header = FALSE)) #iterating over the file list to generate a list of data frames
typeof(align.results) #check its a list
```

```
## [1] "list"
```

```
head(align.results[[1]])
```

```
##                                         V1       V2
## 1 Mapping speed, Million of reads per hour   264.01
## 2                    Number of input reads 13493936
## 3                Average input read length       89
## 4                            UNIQUE READS:         
## 5             Uniquely mapped reads number 11782768
## 6                  Uniquely mapped reads %   87.32%
```

Remove the % from the numbers so just a number

```
align.results <- lapply(align.results, function(x)transform(x, V2 = as
.numeric(gsub("%", "", x$V2) ))) #remove the % from some of numbers so
just a number
```

Alter cosmetics of each data frame

```
names(align.results) <- gsub("(Empty|KOA|KOB)*(\\_[0-12]*)*", "\\1\\2"
, infiles) # some cosmetics of each data frames name, specific for the
sample names of the files used here. Instead of 0-12, it will name the
m by the names in the infiles)
```

Check name of each file

```
names(align.results)
```

```
##  [1] "/Users/cs660/Downloads//Empty1_10bp_Log.final.out"
##  [2] "/Users/cs660/Downloads//Empty2_10bp_Log.final.out"
##  [3] "/Users/cs660/Downloads//Empty3_10bp_Log.final.out"
##  [4] "/Users/cs660/Downloads//Empty4_10bp_Log.final.out"
##  [5] "/Users/cs660/Downloads//KOA1_10bp_Log.final.out"
##  [6] "/Users/cs660/Downloads//KOA2_10bp_Log.final.out"
##  [7] "/Users/cs660/Downloads//KOA3_10bp_Log.final.out"
##  [8] "/Users/cs660/Downloads//KOA5_10bp_Log.final.out"
##  [9] "/Users/cs660/Downloads//KOB1_10bp_Log.final.out"
## [10] "/Users/cs660/Downloads//KOB211_10bp_Log.final.out"
## [11] "/Users/cs660/Downloads//KOB3_10bp_Log.final.out"
## [12] "/Users/cs660/Downloads//KOB5_10bp_Log.final.out"
```

Catenate all of the dataframes together to make one dataframe

```
align.results.df <- as.data.frame(do.call(rbind, align.results)) # cat
enating all data frames of align.results together
align.results.df <- align.results.df[complete.cases(align.results.df),
] # remove lines without any values
head(align.results.df)
```

```
##
V1
## /Users/cs660/Downloads//Empty1_10bp_Log.final.out.1 Mapping speed,
Million of reads per hour
## /Users/cs660/Downloads//Empty1_10bp_Log.final.out.2
Number of input reads
## /Users/cs660/Downloads//Empty1_10bp_Log.final.out.3
Average input read length
## /Users/cs660/Downloads//Empty1_10bp_Log.final.out.5             Uni
```

```
quely mapped reads number
## /Users/cs660/Downloads//Empty1_10bp_Log.final.out.6
Uniquely mapped reads %
## /Users/cs660/Downloads//Empty1_10bp_Log.final.out.7
Average mapped length
##                                                               V2
## /Users/cs660/Downloads//Empty1_10bp_Log.final.out.1      264.01
## /Users/cs660/Downloads//Empty1_10bp_Log.final.out.2 13493936.00
## /Users/cs660/Downloads//Empty1_10bp_Log.final.out.3       89.00
## /Users/cs660/Downloads//Empty1_10bp_Log.final.out.5 11782768.00
## /Users/cs660/Downloads//Empty1_10bp_Log.final.out.6       87.32
## /Users/cs660/Downloads//Empty1_10bp_Log.final.out.7       89.56
```

Adding an additional column of sample names Sample name looks like this before: /Users/cs660/Downloads//Empty1_10bp_Log.final.out.1 We want it to look like: Empty, KOA, KOB

```
align.results.df$sample <- gsub("(//*)\\_.*.", "\\1", row.names(align.
results.df)) #create a new column with sample name
align.results.df$sample <- sub("/.*/", "", align.results.df$sample) #r
emove the ./ from each sample name
align.results.df$sample <- sub("_.*.", "", align.results.df$sample) #r
emove the _ from each sample name
align.results.df$sample <- gsub('[[:digit:]]+', '', align.results.df$s
ample) # remove the digits from the end of the sample names

head(align.results.df)

##
V1
## /Users/cs660/Downloads//Empty1_10bp_Log.final.out.1 Mapping speed,
Million of reads per hour
## /Users/cs660/Downloads//Empty1_10bp_Log.final.out.2
Number of input reads
## /Users/cs660/Downloads//Empty1_10bp_Log.final.out.3
Average input read length
## /Users/cs660/Downloads//Empty1_10bp_Log.final.out.5             Uni
quely mapped reads number
## /Users/cs660/Downloads//Empty1_10bp_Log.final.out.6
Uniquely mapped reads %
## /Users/cs660/Downloads//Empty1_10bp_Log.final.out.7
Average mapped length
##                                                               V2 sam
ple
## /Users/cs660/Downloads//Empty1_10bp_Log.final.out.1      264.01  Em
pty
## /Users/cs660/Downloads//Empty1_10bp_Log.final.out.2 13493936.00  Em
pty
## /Users/cs660/Downloads//Empty1_10bp_Log.final.out.3       89.00  Em
pty
## /Users/cs660/Downloads//Empty1_10bp_Log.final.out.5 11782768.00  Em
pty
## /Users/cs660/Downloads//Empty1_10bp_Log.final.out.6       87.32  Em
pty
## /Users/cs660/Downloads//Empty1_10bp_Log.final.out.7       89.56  Em
pty
```

Adding an additional column of replicate id We want each replicate to be number 1-4 So Empty1 = replicate 1, Empty2 = replicate 2, Empty 3 = replicate 3.... KOA1 = replicate 1, KOA2 = replicate 2, KOA3 = replicate 3, KOA5 = replicate 4 KOB1 = replicate 1, KOB211 = replicate 2, KOB3 = replicate 3, KOB5 = replicate 4

```
align.results.df$replicate <- gsub("(//*)\\_.*.", "\\1", row.names(ali
gn.results.df)) #create a new column with sample name
align.results.df$replicate <- sub("/.*/", "", align.results.df$replica
te) #remove the ./ from each sample name
align.results.df$replicate <- sub("_.*.", "", align.results.df$replica
te) #remove the bits after the _ from each sample name
align.results.df$replicate <- gsub("[^0-9.-]+", "", align.results.df$r
eplicate) #remove chracters apart from the number after the sample nam
e e.g. Empty1 = 1
align.results.df$replicate <- ifelse(align.results.df$replicate == 5,
4, ifelse(align.results.df$replicate == 211, 2, align.results.df$repli
cate)) #replace some of the funny sample names to be replicates 1,2,3&
4
align.results.df$replicate <- as.factor(as.numeric(align.results.df$re
plicate)) #change column to a factor variable
```

Plotting

Now that the data frame is set up how we want, we need to pull out the desired categories for plotting Find the unique fields so we know what the categories are called

```
unique(align.results.df$V1)

##  [1] "Mapping speed, Million of reads per hour"
##  [2] "Number of input reads"
##  [3] "Average input read length"
##  [4] "Uniquely mapped reads number"
##  [5] "Uniquely mapped reads %"
##  [6] "Average mapped length"
##  [7] "Number of splices: Total"
##  [8] "Number of splices: Annotated (sjdb)"
##  [9] "Number of splices: GT/AG"
## [10] "Number of splices: GC/AG"
## [11] "Number of splices: AT/AC"
## [12] "Number of splices: Non-canonical"
## [13] "Mismatch rate per base, %"
## [14] "Deletion rate per base"
## [15] "Deletion average length"
## [16] "Insertion rate per base"
## [17] "Insertion average length"
## [18] "Number of reads mapped to multiple loci"
## [19] "% of reads mapped to multiple loci"
## [20] "Number of reads mapped to too many loci"
## [21] "% of reads mapped to too many loci"
## [22] "Number of reads unmapped: too many mismatches"
## [23] "% of reads unmapped: too many mismatches"
## [24] "Number of reads unmapped: too short"
## [25] "% of reads unmapped: too short"
## [26] "Number of reads unmapped: other"
## [27] "% of reads unmapped: other"
## [28] "Number of chimeric reads"
## [29] "% of chimeric reads"
```

Define those entries that we are interested in and want to plot:

```
filters = c("Number of input reads", "Uniquely mapped reads %",
            "% of reads mapped to multiple loci", "% of reads unmapped
: too short")
```

Create plots

```
plots <- lapply(filters, function(x)
  PlottingAlignmentResults(x, align.results.df, Legend = FALSE))
```

Add legend

```
my.legend <- ExtractLegend(PlottingAlignmentResults(align.results.df,
                                                    Filter = filters[1
],
                                                    Legend=TRUE))
```

Combine plots and legend into one final figure

```
grid.arrange(arrangeGrob(plots[[1]], plots[[2]], plots[[3]], plots[[4]
], nrow=2),
             my.legend$legend, nrow=2,
             heights= unit.c(unit(1, "npc") - my.legend$lheight, my.le
gend$lheight)
)
```

Plotting a stacked bar chart of mapped, multiple mapped and unmapped

Define the colours

```
nb.cols <- 4 # change this to number of coloumns (filters in this case
)
mycolors <- colorRampPalette(brewer.pal(4, "RdYlBu"))(nb.cols)
```

Subset to the categories we want to plot

```
filtered.df <- subset(align.results.df, align.results.df$V1 == "Unique
ly mapped reads number" | align.results.df$V1 == "Number of reads unma
pped: too short" | align.results.df$V1 == "Number of reads mapped to m
ultiple loci")
```

Create Stacked Bar Chart

```
ggplot(data=filtered.df, aes(y=V2, x=sample)) +
  geom_bar(aes(fill=V1), stat="identity",position="stack") +
  theme_bw(base_size = 16) +
  ylab("Number of reads") +
  xlab("Sample") +
  theme(axis.text.x = element_text(size=8), axis.title.x = element_tex
t(size=10)) +
  #scale_fill_manual(values=mycolors) +
  theme(legend.position="bottom",
        legend.text = element_text(size = 6),
        legend.key.size = unit(0.4, "cm"),
        legend.title=element_blank())
```

Percentage plots

Filter down to the categories we want to plot

```
filtered.df <- subset(align.results.df, align.results.df$V1 == "Unique
ly mapped reads %" | align.results.df$V1 == "% of reads unmapped: too
short" | align.results.df$V1 ==  "% of reads mapped to multiple loci")
```

Create the Plot

```
ggplot(data=filtered.df, aes(y=V2, x=sample)) +
  geom_bar(aes(fill=V1), stat="identity",position="stack") +
  theme_bw(base_size = 16) +
  theme(axis.text.x = element_text(size=8), axis.title.x = element_tex
t(size=10)) +
  ylab("Percentage of reads (%)") +
  xlab("Sample") +
  #scale_fill_manual(values=mycolors) +
  theme(legend.position="bottom",
        legend.text = element_text(size = 8),
        legend.key.size = unit(0.4, "cm"),
        legend.title=element_blank())
```

8.1.5 Creating a bam index

```
#!/bin/bash

## Create index for bam files
## CHARLI HARLOW

# Load in SAM tools
module load SAMtools

# Create index for each bam file

for ID in Empty1 Empty2 Empty3 Empty4 KOA1 KOA2 KOA3 KOA5 KOB1
KOB211 KOB3 KOB5
do
samtools index /gpfs/mrc0/projects/Research_Project-
MRC158833/cs660/EPO_project/RNA_sequencing_analysis/Star_Alignment_
11_trimmed/${ID}Aligned.sortedByCoord.out.bam
done
```

8.1.6 Gene quantification using FeatureCounts

```
#!/bin/bash

## Gene Quantification using featureCounts
## CHARLI HARLOW
```

```
# 1.Download the feature counts package which is present in the
subread form: <https://sourceforge.net>

wget -r https://sourceforge.net/projects/subread/files/subread-
2.0.0/subread-2.0.0-Linux-x86_64.tar.gz/download

# 2.Uncompress and unpackage the .tar.gz file

tar -zxvf download

# 3.Create directory to put feature count results in

mkdir featureCounts

# 4. If carrying out featureCounts on paired-end data, ensure that
the .bam files are sorted by read name NOT BY coordinate like STAR
output files
# *If not sorted by read name, feature counts will assume that
almost all reads are not properly paired*

# a) Load in samtools
module load SAMtools
# b) sort by read name
samtools sort -n -o
/path/to/directory/output_name_sortByReadName.bam
/path/to/directory/input_name.out.bam


# 5.Run feature counts to count reads per gene

# a) Set up alias for feature counts directory
featureCounts=~/path/to/package/directory/subread-2.0.0-Linux-
x86_64/bin/featureCounts

# b) Run feature counts (if you have not set up alias as above
command states,make sure to supply full path-to-
directory/featureCounts at start of below command to run
featureCounts

$featureCounts -a /path/to/directory/to/reference-
genome/Homo_sapiens.GRCh38.98.gtf \
-T 8 \ # indicates number of threads to use
-p \ #indicates paired-end
-g gene_id \ # indicates how to name the genes. Default is gene_id.
Can change this to gene_name if you want output to contain the gene
names not the accession numbers. Often it is better to use
accession numbers as genes can have more than one gene-name
-F GTF \ # indicates type of reference file
# -f \ #indicates what level to perform the assignment at – default
is to perform assignment at gene-level (meta-feature). If you
specify -f then it will perform quantification at exon-level
(feature level)
-o
/path/to/directory/output/featureCounts/feature_count_results.txt \
#indicates where to put output
/path/to/directory/input_file_name.bam \  # tells feature counts
where input is
2> /path/to/directory/for/log/file/featurecounts_screen_output.log
#indicates to make a log file showing the screen output as
featurecounts is running
```

```
# If you want to run for all bam files then use * for wildcard e.g
/path/to/directory/input_files/*.bam
```

8.1.7 Visualising FeatureCounts

Visualising_FeatureCounts

Charli E. Harlow

01/06/2020

Visualising featureCounts results

It is quite good practice to then plot the statistics produced from featureCounts so you can assess if the quantification worked and can compare the number of alignments that have been assigned vs those that have not.

Alter the results table to contain three columns - one with the sample names, one with the counts and one with the types of counts
e.g. Unassigned_Ambiguity

Once this has been set up, you can then run the below commands to produce a bar chart summarising the featureCount results

```
## Load in required packages
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.6.2

### Read in table
featurecounts <- read.csv("~/Desktop/featureCounts_summary1.csv")

# Make sure table looks like the following
head(featurecounts)

##     Counts     Type  Sample   Name Name1
## 1 11029026 Assigned Empty1    WT-1   WT1
## 2 13534890 Assigned  Empty2   WT-2   WT2
## 3 12080721 Assigned  Empty3   WT-3   WT3
## 4 13127994 Assigned  Empty4   WT-4   WT4
## 5 15917368 Assigned    KOA1  KO1-1  KOA1
## 6 16297446 Assigned    KOA2 KO1-2  KOA2

## Subset the file to the Classes that you are interested in and want
to plot. I focused on assigned, unassigned no features, unassigned amb
iguity and unassigned no features.
featurecounts <- subset(featurecounts, featurecounts$Type=="Assigned"
| featurecounts$Type=="Unassigned_MultiMapping"|featurecounts$Type=="U
nassigned_NoFeatures"|featurecounts$Type=="Unassigned_Ambiguity")

## Plot the data
## Stacked Bar plot - with number ofcounts on the y axis, sample name
on the x axis and fill the bar plot depending on the Filter Class
```

```
ggplot(data=featurecounts, aes(y=Counts, x=Sample)) +
  geom_bar(aes(fill=Type), stat="identity",position="stack") +
  theme_bw(base_size = 12) +
  ylab("Count") +
  xlab("Sample") +
  theme(legend.position="bottom",
        legend.text = element_text(size = 8),
        legend.key.size = unit(0.4, "cm"),
        legend.title=element_blank())
```

8.1.8 Differential Gene expression analysis

Differential Gene Expression Analysis

Charli E. Harlow

01/06/2020

Differential Gene Expression Analysis using DeSeq2

Lots of information about this software and different vignettes can be found online. For example – a recent vignette from the developers of the software; http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#independent-filtering-of-results

Set current working directory where everything will be saved and stored

```
setwd("~/Desktop")
```

Load in packages which will be needed

```
library(magrittr)

## Warning: package 'magrittr' was built under R version 3.6.2
```

1) Load in the count data

```
read.counts <- read.table("~/Desktop/feature_count_results_geneid.txt"
, sep="\t", header = TRUE)
```

The results can now handled as a dataframe in the R environment

```
head(read.counts, n=3)
```

Replace all row names with the names of genes

```
row.names(read.counts) <- read.counts$Geneid
```

Remove the irrelevant columns which contain no count data

```
read.counts <- read.counts[,-c(1:6)]
```

Give meaningful sample names to the columns if your data-frame does not already have clear names - this can be achieved via numerous approaches

```
names(read.counts) <- c("WT1", "WT2","WT3","WT4","KOA1", "KOA2", "KOA3
", "KOA4", "KOB1", "KOB2", "KOB3", "KOB4")

# Check data is what we expect
str(read.counts)
head (read.counts, n = 3)
```

Extract just read count data for Control & KOA or Control & KOB for individual analysis

- Remove KOB samples

```
read.counts.KOA <- read.counts[,-c(9:12)]
str(read.counts.KOA)
head(read.counts.KOA, n = 3)
```

- remove KOA samples

```
read.counts.KOB <- read.counts[,-c(5:8)]
str(read.counts.KOB)
head (read.counts.KOB, n = 3)
```

Read in gtf file with Gene IDs and Gene names

```
gtf <- read.table("~/Downloads/Homo_sapiens.GRCh38.98_gene_annotation_
table.txt", sep="\t", header=T)

gtf$Geneid <- gtf$gene_id
```

2) Create a meta-data dataframe for use in DeSeq2

Create the ColData for DeSeq2 which contains information on the conditions, confounders etc The conditions and sample names should correspond to the column names of read.counts

Create a data-frame with a column called condition which is WT or KO *Use the column names of read.counts dataframe i.e. WT1,WT2 etc but remove the digits so just called WT or KOA or KOB*

```
sample_info <- data.frame(condition = gsub("[[:digit:]]+", "", names(r
ead.counts)),
                          row.names = names(read.counts))
```

Replace WT with Control

```
sample_info$condition <- ifelse(sample_info$condition=="WT", "WT", ife
lse(sample_info$condition=="KOA", "KOA", "KOB"))
```

Change to a factor column not character

```
sample_info$condition <- as.factor(sample_info$condition)
```

Alter the levels so that WT is recognised as ‘level 1’

```
sample_info$condition %<>% relevel("WT")
```

Check we have what we want

```
sample_info$condition
```

Create a column of sample name using the column names of the read.counts dataframe

```
sample_info$sample <- names(read.counts)
sample_info$sample <- factor(sample_info$sample, levels=c("WT1", "WT2"
,"WT3","WT4","KOA1","KOA2","KOA3","KOA4","KOB1","KOB2","KOB3","KOB4"))
sample_info$sample %<>% relevel("WT1")
```

Create a column for biological replicate where WT sample 1 = 1, WT 2=2 etc

```
library(stringr)
regexp <- "[[:digit:]]+"
```

```
sample_info$rep <- data.frame(rep = gsub("(.*)-", "", names(read.count
s)),
                              row.names = names(read.counts))

sample_info$rep <- str_extract(names(read.counts), regexp)
sample_info$rep <- as.factor(sample_info$rep)
```

Create a column for Genotype where WT or KO i.e. combines KOA & KOB to just KO
Use this column to combine all KO samples

```
sample_info$genotype <- ifelse(sample_info$condition=="WT", "WT", ifel
se(sample_info$condition=="KOA", "KO", "KO"))
sample_info$genotype <- as.factor(sample_info$genotype) # change colum
n to factor not character
sample_info$genotype %<>% relevel("WT") # relevel to ensure WT is leve
l 1
sample_info$genotype # check it is what we expect
```

Create a column for cell line where WT is one cell line, KOA is another cell line regardless of replicate and KOB is a cellline regardless of replicate

```
sample_info$cellline <- ifelse(sample_info$condition=="WT", "0", ifels
e(sample_info$condition=="KOA", "1", "2"))
sample_info$cellline <- as.factor(sample_info$cellline) # change to fa
ctor
```

Set up meta-data just for KOA vs Control

```
sample_info_KOA <- data.frame(condition = gsub("[0-9]+", "", names(rea
d.counts.KOA)),
                              row.names = names(read.counts.KOA))

sample_info_KOA$condition <- ifelse(sample_info_KOA$condition=="WT", "
WT", "KOA")
sample_info_KOA$condition <- as.factor(sample_info_KOA$condition)

# alter the levels so that WT is recogised as 'level 1'
sample_info_KOA$condition %<>% relevel("WT")
sample_info_KOA$condition

# create a column of sample name
sample_info_KOA$sample <- names(read.counts.KOA)
sample_info_KOA$sample <- factor(sample_info_KOA$sample, levels=c("WT1
", "WT2","WT3","WT4","KOA1","KOA2","KOA3","KOA4"))
sample_info_KOA$sample %<>% relevel("WT1")

# create a column for biological replicate where WT sample 1 = 1, WT 2
=2 etc
library(stringr)
regexp <- "[[:digit:]]+"
sample_info_KOA$rep <- str_extract(names(read.counts.KOA), regexp)
sample_info_KOA$rep <- as.factor(sample_info_KOA$rep)

# create a column for Genotype where name WT or KO
# use this column to combine all KO samples
sample_info_KOA$genotype <- ifelse(sample_info_KOA$condition=="WT", "W
T", ifelse(sample_info_KOA$condition=="KOA", "KO", "KO"))
```

```
sample_info_KOA$genotype <- as.factor(sample_info_KOA$genotype)
sample_info_KOA$genotype %<>% relevel("WT")
sample_info_KOA$genotype

# create a column for cell line where WT is one cell line, KOA is anot
her cell line regardless of replicate and KOB is a cellline regardless
of replicate
sample_info_KOA$cellline <- ifelse(sample_info_KOA$condition=="WT", "0
", ifelse(sample_info_KOA$condition=="KOA", "1", "2"))
sample_info_KOA$cellline <- as.factor(sample_info_KOA$cellline)
```

Set up meta-data just for KOB vs Control

```
## KOB meta-data
sample_info_KOB <- data.frame(condition = gsub("[0-9]+", "", names(rea
d.counts.KOB)),
                              row.names = names(read.counts.KOB))

sample_info_KOB$condition <- ifelse(sample_info_KOB$condition=="WT", "
WT", "KOB")
sample_info_KOB$condition <- as.factor(sample_info_KOB$condition)

# alter the levels so that WT is recogised as 'level 1'
sample_info_KOB$condition %<>% relevel("WT")
sample_info_KOB$condition

# create a column of sample name
sample_info_KOB$sample <- names(read.counts.KOB)
sample_info_KOB$sample <- factor(sample_info_KOB$sample, levels=c("WT1
", "WT2","WT3","WT4","KOB1","KOB2","KOB3","KOB4"))
sample_info_KOB$sample %<>% relevel("WT1")

# create a column for biological replicate where WT sample 1 = 1, WT 2
=2 etc
library(stringr)
regexp <- "[[:digit:]]+"
sample_info_KOB$rep <- str_extract(names(read.counts.KOB), regexp)
sample_info_KOB$rep <- as.factor(sample_info_KOB$rep)

# create a column for Genotype where name WT or KO
# use this column to combine all KO samples
sample_info_KOB$genotype <- ifelse(sample_info_KOB$condition=="WT", "W
T", ifelse(sample_info_KOB$condition=="KOB", "KO", "KO"))
sample_info_KOB$genotype <- as.factor(sample_info_KOB$genotype)
sample_info_KOB$genotype %<>% relevel("WT")
sample_info_KOB$genotype

# create a column for cell line where WT is one cell line, KOA is anot
her cell line regardless of replicate and KOB is a cellline regardless
of replicate
sample_info_KOB$cellline <- ifelse(sample_info_KOB$condition=="WT", "0
", ifelse(sample_info_KOB$condition=="KOB", "1", "2"))
sample_info_KOB$cellline <- as.factor(sample_info_KOB$cellline)
```

Assign colours to the different groups in the meta-data data-frame # we can use these colours later to colour groups

```
library("RColorBrewer")
col.genotype <- colorRampPalette(c("royalblue", "red3"))(length(unique
(sample_info$genotype)))[factor(sample_info$genotype)]
col.condition <- colorRampPalette(c("lightblue2", "orange", "deeppink"
))(length(unique(sample_info$condition)))[factor(sample_info$condition
)]
col.rep <- colorRampPalette(c("snow2", "azure2","lightblue2","steelblu
e2"))(length(unique(sample_info$rep)))[factor(sample_info$rep)]
col.sample <- colorRampPalette(c("cadetblue1", "cadetblue2","cadetblue
3","cadetblue4", "brown1", "brown2","brown3","brown4","deeppink", "dee
ppink1", "deeppink2", "deeppink3"))(length(unique(rownames(sample_info
))))
```

3) Filter the data now if wanted

Can filter here or filter once data has been correctly read into DeSeq2 (commands for this below after reading data into DeSeq2)

Remove transcripts whose mean raw count across all samples falls below 10

```
#ZeroCountFilterIndices <- which(apply(read.counts, 1, mean)<10)
#print(paste("Total transcripts with mean<10 counts (all samples):", l
ength(ZeroCountFilterIndices), sep=" "))
#if (length(ZeroCountFilterIndices)>0)
#{
# filtered_readcounts <- read.counts[-ZeroCountFilterIndices,]
#}
```

Check if all genes have at least 1 zero (generates error with DESeq2 1. This first converts the entire data frame to TRUE or FALSE (0 or non-zero) 2. It then applies the table function per row, which gives TRUE and FALSE tallies per gene 3. It then checks if any tally equals the total number of samples - as we've already eliminated genes with all 0 values, the only condition that can meet ncol(txi.working$counts) is the FALSE (non-zero) condition This produces a further TRUE or FALSE for each gene and condition 4. Finally, if any TRUE values are present, then we know that at least 1 row has a non-zeros, and therefore we can proceed

```
#if (!any(data.frame(unlist(apply((filtered_readcounts==0), 1, functio
n(x) table(x))))==ncol(filtered_readcounts)))
#{
#  print("All genes contain at least 1 zero.")
#  next()
#}
```

4) Generate the DeSeq2DataSet

Install DeSeq2

```
#BiocManager::install()
#BiocManager::install("DESeq2")
```

Load in DE-Seq2

```
library(DESeq2)

## Loading required package: S4Vectors

## Loading required package: stats4
## Loading required package: BiocGenerics

## Loading required package: parallel

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind, colnames
,
##     dirname, do.call, duplicated, eval, evalq, Filter, Find, get, g
rep,
##     grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget
,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,
##     union, unique, unsplit, which, which.max, which.min

##
## Attaching package: 'S4Vectors'

## The following object is masked from 'package:base':
##
##     expand.grid

## Loading required package: IRanges

## Loading required package: GenomicRanges

## Loading required package: GenomeInfoDb

## Loading required package: SummarizedExperiment

## Loading required package: Biobase

## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.

## Loading required package: DelayedArray

## Loading required package: matrixStats

## Warning: package 'matrixStats' was built under R version 3.6.2

##
## Attaching package: 'matrixStats'
## The following objects are masked from 'package:Biobase':
##
##     anyMissing, rowMedians

## Loading required package: BiocParallel

## Warning: multiple methods tables found for 'type'

##
## Attaching package: 'DelayedArray'

## The following objects are masked from 'package:matrixStats':
##
##     colMaxs, colMins, colRanges, rowMaxs, rowMins, rowRanges

## The following object is masked from 'package:BiocGenerics':
##
##     type

## The following objects are masked from 'package:base':
##
##     aperm, apply, rowsum

## Warning: replacing previous import 'BiocGenerics::type' by 'Delayed
Array::type'
## when loading 'SummarizedExperiment'

library(DESeq2)
```

Convert the data to DESeq format and specify the model We want to test for the effect of genotype (combining all KOs as KO) *Make sure that control/WT is the first level of a factor i.e. the control level*

```
library("magrittr")
sample_info$genotype
# for all KOs combined
dds <- DESeqDataSetFromMatrix(countData=read.counts, #this is the coun
t results data-frame,
                              colData=sample_info, #this is the data-f
rame containing information on samples,
                              design= ~genotype) #specify the design m
odel - here we want to test WT against KO

# for Control vs KOA combined
sample_info_KOA$genotype
dds_KOA <- DESeqDataSetFromMatrix(countData=read.counts.KOA, #this is
the count results data-frame,
                              colData=sample_info_KOA, #this is the da
ta-frame containing information on samples,
                              design= ~genotype) #specify the design m
odel - here we want to test WT against KOA

# for Control vs KOB combined
sample_info_KOB$genotype
dds_KOB <- DESeqDataSetFromMatrix(countData=read.counts.KOB, #this is
the count results data-frame,
                              colData=sample_info_KOB, #this is the da
ta-frame containing information on samples,
                              design= ~genotype) #specify the design m
odel - here we want to test WT against KOB
```

Check the DeSeq2 dataset has been read in correctly

```
colData(dds) %>% head
assay(dds, "counts") %>% head
rowData(dds) %>% head
# test what counts () returns
counts(dds) %>% str
```

Filter (this is very similar to what would have been done above but you can do it the deseq2 dataframe instead)

Remove any genes where the expresison is < 1

```
nrow(dds)
keep <- rowSums(counts(dds)) > 1
dds <- dds[keep,]
nrow(dds)

# For Control vs KOA
nrow(dds_KOA)
keep <- rowSums(counts(dds_KOA)) > 1
dds_KOA <- dds_KOA[keep,]
nrow(dds_KOA)

# For Control vs KOB
nrow(dds_KOB)
keep <- rowSums(counts(dds_KOB)) > 1
dds_KOB <- dds_KOB[keep,]
nrow(dds_KOB)
```

Investigate different library sizes

```
colSums(counts(dds))
colSums(read.counts)
```

5) Normalise the data

DESeq2's default method to normalize read counts to account for differences in sequencing depths is implemented in estimateSizeFactors()

```
dds <- estimateSizeFactors(dds)
sizeFactors(dds)
```

If you check colData () again , you see that this now contains the sizeFactors

```
colData(dds)
```

Counts() allows you to immediately retrieve the _normalized_read counts

```
norm <- counts(dds, normalized=TRUE)
```

Control vs KOA DESeq2's default method to normalize read counts to account for differences in sequencing depths is implemented in estimateSizeFactors()

```
dds_KOA <- estimateSizeFactors(dds_KOA)
sizeFactors(dds_KOA)
```

```
# if you check colData () again , you see that this now contains the s
izeFactors
colData(dds_KOA)

# counts () allows you to immediately retrieve the _normalized_read co
unts
norm_KOA <- counts(dds_KOA, normalized=TRUE)
```

Control vs KOB DESeq2’s default method to normalize read counts to account for differences in sequencing depths is implemented in estimateSizeFactors()

```
dds_KOB <- estimateSizeFactors(dds_KOB)
sizeFactors(dds_KOB)

# if you check colData () again , you see that this now contains the s
izeFactors
colData(dds_KOB)

# counts () allows you to immediately retrieve the _normalized_read co
unts
norm_KOB <- counts(dds_KOB, normalized=TRUE)
```

6) Transformation

Downstream analyses (including clustering) work much better if the read counts are transformed to the log scale following normalization.

Transform size-factor normalized read counts to log2 scale using a pseudocount of 1

```
log.norm.counts <- log2(norm + 1)
```

can also use this command

```
ntd <- normTransform(dds)
```

Control vs KOA

```
# Transform size - factor normalized read counts to log2 scale using a
pseudocount of 1
log.norm.counts.KOA <- log2(norm_KOA + 1)
```

Control vs KOB

```
# Transform size - factor normalized read counts to log2 scale using a
pseudocount of 1
log.norm.counts.KOB <- log2(norm_KOB + 1)
```

Plotting the transformation

```
par(mfrow =c(3 , 1)) # to plot the following two images underneath eac
h other

# first, plot the normalised data: non-transformed
boxplot(norm, notch = TRUE ,
        main = "Untransformed read counts ", ylab = "read counts")
# second, plot the transformed normalised data
boxplot(log.norm.counts, notch = TRUE ,
        main = "Log2 - transformed read counts ",
        ylab = " log2 (read counts)")
# this should give exactly the same as log.norm.counts plot
boxplot(assay(ntd), notch = TRUE ,
        main = "Log2 - transformed read counts ",
        ylab = " log2 (read counts)")
```

7) Visualise the normalised data

Plot the counts in a pairwise manner

```
plot(log.norm.counts[ ,1:2] , cex =.1 , main = " Normalized log2 ( rea
d counts )")
```

Check for heteroscedascity *Many statistical tests and analyses assume that data is homoskedastic, i.e. that all variables have similar variance. However, data with large differences among the sizes of the individual observations often shows heteroskedastic behavior. One way to visually check for heteroskedasticity is to plot the mean vs. the standard deviation*

```
# BiocManager::install("vsn")
library("vsn")
library(ggplot2 )

## Warning: package 'ggplot2' was built under R version 3.6.2

msd_plot <- meanSdPlot(log.norm.counts,
                       ranks =FALSE , # show the data on the original
scale
                       plot = FALSE )
msd_plot$gg +
  ggtitle ("Sequencing depth normalized log2 (read counts )") +
  ylab ("Standard Deviation ")
```

```
msd_plot <- meanSdPlot(assay(ntd),
                       ranks =FALSE , # show the data on the original
scale
                       plot = FALSE )

msd_plot$gg +
  ggtitle ("Sequencing depth normalized log2 (read counts )") +
  ylab ("Standard Deviation ")
```

```
meanSdPlot(assay(ntd))
```

The y-axis shows the variance of the read counts across all samples. Some variability is, in fact, expected, but a clear hump on the left-hand side indicates that for read counts < 32 (2^5 = 32), the variance is higher than for those with greater read counts. That means that there is a dependence of the variance on the mean, which violates the assumption of homoskedasticity.

8) Reduce the heteroskedasticity

Shrink the variance of low read counts For RNA-seq counts, however, the expected variance grows with the mean. A simple and often used strategy to avoid this is to take the logarithm of the normalized count values plus a pseudocount of 1; however, depending on the choice of pseudocount, now the genes with the very lowest counts will contribute a great deal of noise to the resulting plot, because taking the logarithm of small counts actually inflates their variance. As a solution, DESeq2 offers two transformations for count data that stabilize the variance across the mean: the variance stabilizing transformation (VST) for negative binomial data with a dispersion-mean trend (Anders and Huber 2010), implemented in the vst function, and the regularized-logarithm transformation or rlog (Love, Huber, and Anders 2014). For genes with high counts, both the VST and the rlog will give similar result to the ordinary log2 transformation of normalized counts. For genes with lower counts, however, the values are shrunken towards a middle value. The VST or rlog-transformed data then become approximately homoskedastic (more flat trend in the meanSdPlot), and can be used directly for computing distances between samples, making PCA plots, or as input to downstream methods which perform best with homoskedastic data.

R log transformation DESeq2's rlog() function returns values that are both normalized for sequencing depth and transformed to the log2 scale where the values are adjusted to it the experiment-wide trend of the variance-mean relationship *blind = FALSE means that differences between cell lines and treatment (the variables in the design) will not contribute to the expected variance-mean trend of the experiment. The experimental design is not used directly in the transformation, only in estimating the global amount of variability in the counts.* The rlog() function's blind parameter should be set to FALSE if the different conditions lead to strong differences in a large proportion of the genes. If rlog() is applied without incorporating the knowledge of the experimental design (blind = TRUE, the default setting), the dispersion will be greatly overestimated in such cases.

```r
rld <- rlogTransformation(dds, blind = TRUE)
rldMatrix <- data.matrix(assay(rld))
head(assay(rld), 3)

## Control vs KOA
rld_KOA <- rlogTransformation(dds_KOA, blind = TRUE)
rldMatrixKOA <- data.matrix(assay(rld_KOA))

head(assay(rld_KOA), 3)

## Control vs KOB
rld_KOB <- rlogTransformation(dds_KOB, blind = TRUE)
rldMatrixKOB <- data.matrix(assay(rld_KOB))

head(assay(rld_KOB), 3)
```

Plotting the rlog transformation

```r
msd_rlog_plot <- meanSdPlot(assay(rld),
                            ranks =FALSE, # show the data on the origi
nal scale
                            plot = FALSE )
msd_rlog_plot$gg +
  ggtitle ("rlog - transformed read counts") +
  ylab (" standard deviation ")
```

Export the rlog counts

```
#write.table(rldMatrixKOB, "../Control vs KOB/RLogCounts_wtvskob.txt",
row.names=TRUE, col.names=TRUE, sep="\t", quote=FALSE)
```

The variance stabilizing transformation

```
# Control vs KO
vsd <- vst(dds, blind = TRUE)
head(assay(vsd), 3)
colData(vsd)
vsd.norm.counts <- assay(vsd)

vsdMatrix <- data.matrix(assay(vsd))

# Control vs KOA
vsd_KOA <- vst(dds_KOA, blind = TRUE)
head(assay(vsd_KOA), 3)
colData(vsd_KOA)
vsd.norm.counts.KOA <- assay(vsd_KOA)

vsdMatrixKOA <- data.matrix(assay(vsd_KOA))

# Control vs KOB
vsd_KOB <- vst(dds_KOB, blind = TRUE)
head(assay(vsd_KOB), 3)
colData(vsd_KOB)
vsd.norm.counts.KOB <- assay(vsd_KOB)

vsdMatrixKOB <- data.matrix(assay(vsd_KOB))


# Plotting the vsd normalisation
msd_vsd_plot <- meanSdPlot(assay(vsd),
                           ranks =FALSE, # show the data on the origin
al scale
                           plot = FALSE )
msd_vsd_plot$gg +
  ggtitle ("vsd - transformed read counts ") +
  ylab (" standard deviation ")
```

```
meanSdPlot(assay(vsd))
```

```
#write.table(vsdMatrix_KOB, "../Control vs KOB/vsdCounts_wtvskob.txt",
row.names=TRUE, col.names=TRUE, sep="\t", quote=FALSE)
```

9) QC plots to check for normalisation and transformation

Output dispersion plot Need to have estimated the dispersion distance to plot this

```
options(scipen=999)
dds <- estimateDispersions(dds)

## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates

dds_KOA <- estimateDispersions(dds_KOA)

## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates

dds_KOB <- estimateDispersions(dds_KOB)

## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates

par(mfrow =c(1 , 1)) # to plot the following two images underneath eac
h other
options(scipen=999)
plotDispEsts(dds, genecol="black", fitcol="red", finalcol="dodgerblue"
, legend=TRUE, log="xy", cex.axis=0.8, cex=0.3, cex.main=0.8, xlab="Me
an of normalised counts", ylab="Dispersion")
```

```
options(scipen=0)
```

Histograms to check normalisation methods

```
hist(norm, breaks=100, xlab="Counts", col="grey", main="Normalised cou
nts")
```

```
hist(norm, breaks=10000, xlab="Counts", xlim=c(0,2500), col="grey", ma
in="Normalised counts\n(zoomed range 0:2500)")
```

```
hist(log2(norm + 1), breaks=10, xlab="Counts", col="grey", main=bquote
(~Log[2]~normalised~counts))
```

```
hist(rldMatrix, xlab="Counts", breaks=50, col="grey", main="Regularise
d log counts")
```

```
hist(vsdMatrix, xlab="Counts", col="grey", main="Variance Stablised co
unts")
```

Boxplots to check normalisation

```
#par(mar=c(3,3,3,3), mfrow=c(5,1), cex=1, cex.axis=0.8)
boxplot(norm, main="Normalised counts", xlab="", ylab="Normalised coun
ts", names=paste(sample_info$sample), col=col.sample, las=2)
```

```
boxplot(log2(norm+1), main="Log2 + 1 Normalised counts", xlab="", ylab
="Log2 +1 Normalised counts", names=paste(sample_info$sample), col=col
.sample, las=2)
```

```
boxplot(rldMatrix, main="Regularised log counts", xlab="", ylab="Regul
arised log counts", names=paste(sample_info$sample), col=col.sample, l
as=2)
```

```
boxplot(rldMatrix, main="Regularised log counts\n(outlier genes remove
d)", xlab="", ylab="Regularised log counts", names=paste(sample_info$s
ample), col=col.sample, las=2, outline=FALSE)
```

```
boxplot(vsdMatrix, main="Variance Stablised counts", xlab="", ylab="Va
riance Stablised counts", names=paste(sample_info$sample), col=col.sam
ple, las=2)
```

```
boxplot(vsdMatrix, main="Variance Stablised counts\n(outlier genes rem
oved)", xlab="", ylab="Variance Stablised counts", names=paste(sample_
info$sample), col=col.sample, las=2, outline=FALSE)
```

Plotting the normalisation methods on scatter

```
library("dplyr")

## Warning: package 'dplyr' was built under R version 3.6.2

## 
## Attaching package: 'dplyr'

## The following object is masked from 'package:matrixStats':
## 
##     count

## The following object is masked from 'package:Biobase':
## 
##     combine

## The following objects are masked from 'package:GenomicRanges':
## 
##     intersect, setdiff, union
## The following object is masked from 'package:GenomeInfoDb':
##
##     intersect

## The following objects are masked from 'package:IRanges':
##
##     collapse, desc, intersect, setdiff, slice, union

## The following objects are masked from 'package:S4Vectors':
##
##     first, intersect, rename, setdiff, setequal, union

## The following objects are masked from 'package:BiocGenerics':
##
##     combine, intersect, setdiff, union

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library("ggplot2")

norm_methods  <- bind_rows(
  as_tibble(log.norm.counts) %>%
    mutate(transformation = "log2(x + 1)"),
  as_data_frame(assay(vsd)[, 1:2]) %>% mutate(transformation = "vst"),
  as_data_frame(assay(rld)[, 1:2]) %>% mutate(transformation = "rlog")
)

## Warning: `as_data_frame()` was deprecated in tibble 2.0.0.
## Please use `as_tibble()` instead.
## The signature and semantics have changed, see `?as_tibble`.

colnames(norm_methods)[1:2] <- c("x", "y")
ggplot(norm_methods, aes(x = x, y = y)) + geom_hex(bins = 80) +
  coord_fixed() + facet_grid( . ~ transformation)
```

10) Hierarchal Clustering

Calculating distance between samples Using the R log normalisation

Default method for calculating these distances is Euclidean

```
sampleDists_rld <- dist(t(assay(rld)))
sampleDists_rld

## Control vs KOA
sampleDists_rld_KOA <- dist(t(assay(rld_KOA)))
sampleDists_rld_KOA

## Control vs KOB
sampleDists_rld_KOB <- dist(t(assay(rld_KOB)))
sampleDists_rld_KOB
```

```
#Save the distances between samples to table
sampleDistMatrix_rld <- as.matrix(sampleDists_rld)
sampleDistMatrix_rld_KOA <- as.matrix(sampleDists_rld_KOA)
sampleDistMatrix_rld_KOB <- as.matrix(sampleDists_rld_KOB)

#write.table(as.matrix(sampleDists_rld_KOB), "../Control vs KOB/Sample
Distance_rlog_wtvskob.txt", row.names=T, col.names=T, sep="\t", quote=
F)

## Using the vsd normalisation
sampleDists_vsd <- dist(t(assay(vsd)))
sampleDists_vsd
```

Heat map to visualise distance between samples *uses the normalised count matrix*

```
library("pheatmap")
library("RColorBrewer")

# call the row names and the column name
rownames(sampleDistMatrix_rld) <- paste(rld$sample)
colnames(sampleDistMatrix_rld) <- paste(rld$sample)

# change the cluster method to ward.D2
hc <- hclust(sampleDists_rld, method="ward.D2")

# create colours
colors <- colorRampPalette(rev(brewer.pal(9, "Blues")))(255)
mycols <- brewer.pal(3, "Blues")[1:length(unique(sample_info$condition
))]

# change the order to make WT be the first sample
callback = function(hc, mat){
  sv = svd(t(mat))$v[,1]
  dend = reorder(as.dendrogram(hc), wts = sv)
  as.hclust(dend)
}

# plot heat map
pheatmap(sampleDistMatrix_rld,
         clustering_distance_rows = sampleDists_rld,
         clustering_distance_cols = sampleDists_rld,
         col = colors,
        clustering_callback = callback)
```

Heat map using gplots

```
library(gplots)

##
## Attaching package: 'gplots'

## The following object is masked from 'package:IRanges':
##
##     space
## The following object is masked from 'package:S4Vectors':
##
##     space

## The following object is masked from 'package:stats':
##
##     lowess

heatmap.2(as.matrix(sampleDists_rld), key=T, trace="none",
          col=colors,
          #Rowv=F, Colv=F,
          Rowv=as.dendrogram(hc),
          cexRow = 1.5, cexCol=1.5,
          symm=TRUE,
         ColSideColors=mycols[sample_info$condition], RowSideColors=my
cols[sample_info$condition],
          margin=c(5, 5), main="Sample Distance Matrix",  key.title="S
ample similarity", key.xlab="Euclidean distance", key.ylab="")
```

Plot distibution and density with violin plots

```
violinMatrix <- reshape2::melt(rldMatrix, id.vars=NULL)

colnames(violinMatrix) <- c("Gene","Sample","Expression")

library(ggplot2)
ggplot(violinMatrix, aes(x=Sample, y=Expression)) + geom_violin() + th
eme(axis.text.x = element_text(angle=45, hjust=1), axis.line= element_
line(colour = "black"),panel.grid.major = element_blank(), panel.grid.
minor = element_blank(),

panel.background = element_blank())
```

Dendogram cor () calculates the correlation between columns of a matrix

Calculate Pearson’s correlation distance

```
distance.rld <- as.dist(1 - cor(assay(rld), method = "pearson" ))
distance.rld.KOA <- as.dist(1 - cor(assay(rld_KOA), method = "pearson"
))
distance.rld.KOB <- as.dist(1 - cor(assay(rld_KOB), method = "pearson"
))
```

plot () can directly interpret the output of hclust()

```
par(cex=1.0, cex.axis=0.8, cex.main=0.8)
plot(hclust(distance.rld),
     labels = colnames(rld),
     main = "rlog transformed Read Counts Distance: Pearson Correlatio
n")
```

Circular dendrogram and regular dendrogram

```
distmatrix <- dist(t(rldMatrix), method="euclidean")
hclustObject <- hclust(distmatrix, method="ward.D2")
dend <- as.dendrogram(hclustObject)
plot(dend, ylab="Height", main="rlog transformed Read Counts Distance:
Euclidean Distance")
```

Circular dendrogram

```
library(circlize)

## ========================================
## circlize version 0.4.8
## CRAN page: https://cran.r-project.org/package=circlize
## Github page: https://github.com/jokergoo/circlize
## Documentation: http://jokergoo.github.io/circlize_book/book/
##
## If you use it in published research, please cite:
## Gu, Z. circlize implements and enhances circular visualization
##   in R. Bioinformatics 2014.
## ========================================

library(dendextend)

##
## ---------------------
## Welcome to dendextend version 1.13.4
## Type citation('dendextend') for how to cite the package.
##
## Type browseVignettes(package = 'dendextend') for the package vignet
te.
## The github page is: https://github.com/talgalili/dendextend/
##
## Suggestions and bug-reports can be submitted at: https://github.com
/talgalili/dendextend/issues
## Or contact: <tal.galili@gmail.com>
##
##  To suppress this message use:  suppressPackageStartupMessages(libr
ary(dendextend))
## ---------------------

##
## Attaching package: 'dendextend'

## The following object is masked from 'package:stats':
##
##     cutree

#Get the heights for each branch
heights <- round(get_branches_heights(dend, sort=FALSE), 1)

#Get max height
maxHeight= max(heights)

#Set label and dendrogram height for cicular dendrogram
labelHeight=0.1
dendHeight=0.8
```

```
labels(dend) <- gsub("\\.[0-9]$", "", labels(dend))

#Draw the circular dendrogram
circlize_dendrogram(dend, facing="outside", labels=TRUE, labels_track_
height=labelHeight, dend_track_height=dendHeight, cex=0.6)

#Create tick co-ordinates and values for the new axis
#We have to enure that we don't overlap the label plot region (height
specified by labelHeight), nor the central region of the plot (1-(dend
Height+labelHeight))
ticks <- seq(from=(1-(dendHeight+labelHeight)), to=(1-labelHeight), le
ngth.out=5)
values <- round(rev(seq(from=0, to=maxHeight, length.out=5)), 1)

#Add the new axis
library(plotrix)

##
## Attaching package: 'plotrix'

## The following object is masked from 'package:gplots':
##
##     plotCI

ablineclip(h=0, v=ticks, col="black", x1=1-(dendHeight+labelHeight), x
2=1-labelHeight, y1=0, y2=0.04, lwd=1.5)
text(ticks, 0+0.08, values, cex=0.8)
text((1-labelHeight)-(((1-labelHeight)-(1-(dendHeight+labelHeight)))/2
), 0+0.14, "Height")
```

Scatterplots Perform pairwise scatter plots on the samples

```
#library(car)
#scatterplotMatrix(rldMatrix[,c("WT1","WT2","WT3","WT4")], diagonal="b
oxplot", pch=".")
#scatterplotMatrix(rldMatrix[,c("KOA1","KOA2","KOA3","KOA4")], diagona
l="boxplot", pch=".")
#scatterplotMatrix(rldMatrix[,c("WT1","WT2","WT3","WT4","WT4","KOA1","
KOA2","KOA3","KOA4","KOB1","KOB2", "KOB3","KOB4")], diagonal="boxplot"
, pch=".")
```

11) PCA plots

PCA analysis using R function

```
project.pca <- prcomp(t(assay(rld)))

plot(project.pca$x[,1], project.pca$x[,2],
     col = col.sample,
     xlab="PC1",
     ylab="PC2",
     main = "PCA of seq.depth normalized \n and rlog - transformed rea
d counts")
```

```
rownames(rldMatrix) <- rownames(dds)

# Control vs KOA
project.pca.KOA <- prcomp(t(assay(rld_KOA)))
# Control vs KOB
project.pca.KOB <- prcomp(t(assay(rld_KOB)))
```

Accessing the PCA results

```
library(factoextra)

## Warning: package 'factoextra' was built under R version 3.6.2

## Welcome! Want to learn more? See two factoextra-related books at ht
tps://goo.gl/ve3WBa

# Eigenvalues
eig.val <- get_eigenvalue(project.pca)
eig.val

#results for variable
res.var <- get_pca_var(project.pca)
head(res.var$contrib, n=3)      # Contributions to the PCs
head(res.var$coord, n=3)

# results for samples
res.ind <- get_pca_ind(project.pca)
head(res.ind$coord, n=3)          # Coordinates
head(res.ind$contrib, n=3)        # Contributions to the PCs
head(res.ind$cos2, n=3)           # Quality of representation

# results of PCs for samples
project.pca$x
sample_info$projectpca2 <- project.pca$x[,"PC1"]
sample_info$projectpca3 <- project.pca$x[,"PC2"]
```

Determine the proportion of variance of each component Proportion of variance equals (PC stdev^2) / (sum all PCs stdev^2)

```
project.pca.proportionvariances <- ((project.pca$sdev^2) / (sum(projec
t.pca$sdev^2)))*100
# Control vs KOA
project.pca.proportionvariances.KOA <- ((project.pca.KOA$sdev^2) / (su
m(project.pca.KOA$sdev^2)))*100

#Control vs KOB
project.pca.proportionvariances.KOB <- ((project.pca.KOB$sdev^2) / (su
m(project.pca.KOB$sdev^2)))*100
```

Scree plot of PCA

```
barplot(project.pca.proportionvariances, cex.names=1, xlab=paste("Prin
cipal component (PC), 1-", length(project.pca$sdev)), ylab="Proportion
of variation (%)", main="Scree plot", ylim=c(0,100))
```

Scatter Plot of PCA 1-4

```
par(xpd=TRUE)
pairs(project.pca$x[,1:4], col=col.sample, main="Principal components
analysis bi-plot\nPCs 1-4", pch=16, oma=c(2,2,2,13))
legend("bottomright", cex=0.75, fill = unique(col.sample), legend = c(
levels(sample_info$sample)))
```

Scatter Plot of PCA 5-8

```
par(xpd=TRUE)
pairs(project.pca$x[,5:8], col=col.sample, main="Principal components
analysis bi-plot\nPCs 6-10", pch=16, oma=c(2,2,2,13))
legend("bottomright", cex=0.75, fill = unique(col.sample), legend = c(
levels(sample_info$sample)))
```

Scatter Plots

```
par(mar=c(4,4,4,4), mfrow=c(2,3), cex=1.0, cex.main=0.5, cex.axis=0.8)

#Plots scatter plot for PC 1 and 2
plot(project.pca$x, type="n", #main="Principal components analysis bi-
plot",
     xlab=paste("PC1, ", round(project.pca.proportionvariances[1], 2),
"%"), ylab=paste("PC2, ", round(project.pca.proportionvariances[2], 2)
, "%"))
points(project.pca$x, col=col.sample, pch= c(15, 16, 17, 18), cex=1.5)


#Plots scatter plot for PC 1 and 3
plot(project.pca$x[,1], project.pca$x[,3], type="n", main="Principal c
omponents analysis bi-plot", xlab=paste("PC1, ", round(project.pca.pro
portionvariances[1], 2), "%"), ylab=paste("PC3, ", round(project.pca.p
roportionvariances[3], 2), "%"))
points(project.pca$x[,1], project.pca$x[,3], col=col.sample, pch= c(15
, 16, 17, 18), cex=1)

#Plots scatter plot for PC 2 and 3
plot(project.pca$x[,2], project.pca$x[,3], type="n",main="Principal co
mponents analysis bi-plot",  xlab=paste("PC2, ", round(project.pca.pro
portionvariances[2], 2), "%"), ylab=paste("PC3, ", round(project.pca.p
roportionvariances[3], 2), "%"))
points(project.pca$x[,2], project.pca$x[,3], col=col.sample, pch= c(15
, 16, 17, 18), cex=1)

#Plots scatter plot for PC 2 and 4
plot(project.pca$x[,2], project.pca$x[,4], type="n", main="Principal c
omponents analysis bi-plot", xlab=paste("PC2, ", round(project.pca.pro
portionvariances[2], 2), "%"), ylab=paste("PC4, ", round(project.pca.p
roportionvariances[4], 2), "%"))
points(project.pca$x[,2], project.pca$x[,4], col=col.sample, pch= c(15
, 16, 17, 18), cex=1)

#Plots scatter plot for PC 3 and 4
plot(project.pca$x[,3], project.pca$x[,4], type="n",main="Principal co
mponents analysis bi-plot", xlab=paste("PC3, ", round(project.pca.prop
ortionvariances[3], 2), "%"), ylab=paste("PC4, ", round(project.pca.pr
oportionvariances[4], 2), "%"))
points(project.pca$x[,3], project.pca$x[,4], col=col.sample, pch= c(15
, 16, 17, 18), cex=1)

par(xpd=TRUE)
plot.new()
legend("bottomright", bty="n", cex=0.8, title="Condition", legend=c("C
ontrol 1","Control 2","Control 3","Control 4","KOA1","KOA2","KOA3","KO
A4","KOB1","KOB2","KOB3", "KOB4"), col=c(col.sample), pch=c(15, 16, 17
, 18))
```

3D scatter plot of PCAs

```
library(scatterplot3d)
par(mar=c(4,4,4,4), mfrow=c(2,2), cex=1.0, cex.main=0.8, cex.axis=0.8)

scatterplot3d(project.pca$x[,1:3], angle=-40, main="", color=col.genot
ype, pch=17, xlab=paste("PC1, ", round(project.pca.proportionvariances
[1], 2), "%"), ylab=paste("PC2, ", round(project.pca.proportionvarianc
es[2], 2), "%"), zlab=paste("PC3, ", round(project.pca.proportionvaria
nces[3], 2), "%"), grid=FALSE, box=FALSE)
source('http://www.sthda.com/sthda/RDoc/functions/addgrids3d.r')
addgrids3d(project.pca$x[,2:4], grid = c("xy", "xz", "yz"))

par(xpd=TRUE)
plot.new()
legend("left", bty="n", cex=0.8, title="Condition", c("Control 1","Con
trol 2","Control 3","Control 4","KOA1","KOA2","KOA3","KOA4","KOB1","KO
B2","KOB3", "KOB4"), fill=c(col.genotype))

scatterplot3d(project.pca$x[,1:3], angle=40, main="", color=col.condit
ion, pch=17, xlab=paste("PC1, ", round(project.pca.proportionvariances
[1], 2), "%"), ylab=paste("PC2, ", round(project.pca.proportionvarianc
es[2], 2), "%"), zlab=paste("PC3, ", round(project.pca.proportionvaria
nces[3], 2), "%"), grid=FALSE, box=FALSE)
source('http://www.sthda.com/sthda/RDoc/functions/addgrids3d.r')
addgrids3d(project.pca$x[,2:4], grid = c("xy", "xz", "yz"))

par(xpd=TRUE)
plot.new()
legend("left", bty="n", cex=0.8, title="Condition", c("Control 1","Con
trol 2","Control 3","Control 4","KOA1","KOA2","KOA3","KOA4","KOB1","KO
B2","KOB3", "KOB4"), fill=c(col.condition))
```

Write out eigenvector 1 eigenvalues as PC1 seperates the control from KO the best

```
library(plyr); library(dplyr)

## -------------------------------------------------------------------
-----------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems
.
## If you need functions from both plyr and dplyr, please load plyr fi
rst, then dplyr:
## library(plyr); library(dplyr)

## ------------------------------------------------------------------
-----------

##
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following object is masked from 'package:matrixStats':
##
##     count

## The following object is masked from 'package:IRanges':
##
##     desc

## The following object is masked from 'package:S4Vectors':
##
##     rename

project.pca <- prcomp(t(rldMatrix))
summary(project.pca)
project.pca.proportionvariances <- ((project.pca$sdev^2) / (sum(projec
t.pca$sdev^2)))*100

# get the eigenvalues for PC1 for each gene (Control vs KO)

PC1_values <- data.frame(abs(project.pca$rotation[,c("PC1")]))

PC1_values<- data.frame(rownames(PC1_values), PC1_values)
head(PC1_values, n=3)

# Change the column names
names(PC1_values) <- c("Geneid","PC1")

PC1_values <- join(PC1_values, gtf, by="Geneid")
row.names(PC1_values) <- PC1_values$Geneid
head(PC1_values, n=3)

# Control vs KOA
PC1_values_KOA <- data.frame(abs(project.pca.KOA$rotation[,c("PC1")]))

PC1_values_KOA <- data.frame(rownames(PC1_values_KOA), PC1_values_KOA)
head(PC1_values, n=3)

# Change the column names
names(PC1_values_KOA) <- c("Geneid","PC1")

PC1_values_KOA <- join(PC1_values_KOA, gtf, by="Geneid")
```

```
row.names(PC1_values_KOA) <- PC1_values_KOA$Geneid
head(PC1_values, n=3)

#Control vs KOB
PC1_values_KOB <- data.frame(abs(project.pca.KOB$rotation[,c("PC1")]))

PC1_values_KOB <- data.frame(rownames(PC1_values_KOB), PC1_values_KOB)
head(PC1_values, n=3)

# Change the column names
names(PC1_values_KOB) <- c("Geneid","PC1")

PC1_values_KOB <- join(PC1_values_KOB, gtf, by="Geneid")
row.names(PC1_values_KOB) <- PC1_values_KOB$Geneid
head(PC1_values, n=3)

#write.table(PC1_values_KOA, "~/Documents/EPO Project/CRISPR/Whole gen
e knock-out/Confirming EPO KO/RNA Sequencing/DeSeq2/Control vs KOA/PC1
.Eigenvalues.wtvskoa.csv", col.names=TRUE, row.names=FALSE, quote=FALS
E, sep=",")

# If pulling our two eigenvalues
# wObject <- data.frame(abs(project.pca.KOB$rotation[,c("PC1","PC2")])
)

# to calculate the mean if pulling out two eigenvalues
#PC1_PC2_values <- data.frame(abs(project.pca$rotation[,c("PC1","PC2")
]))
#PC1_PC2_values <- data.frame(rownames(PC1_PC2_values), PC1_PC2_values
, apply(PC1_PC2_values, 1, mean))
#PC1_PC2_values$mean <- (PC1_PC2_values$PC1+PC1_PC2_values$PC2)/2
#names(PC1_PC2_values) <- c("GeneID","PC1", "PC2", "Mean", "mean")

#PC1_PC2_values
#write.table(PC1_PC2_values, "PC1.PC2.Eigenvalues.csv", col.names=TRUE
, row.names=FALSE, quote=FALSE, sep=",")
```

Order the results

```
# Default of order is ascending
# We want to order according to the highest eigenvalue so we want desc
ending
sorted_PC1_values <- PC1_values[order(-PC1_values$PC1), ]
head(sorted_PC1_values)

sorted_PC1_values_KOA <- PC1_values_KOA[order(-PC1_values_KOA$PC1), ]
sorted_PC1_values_KOB <- PC1_values_KOB[order(-PC1_values_KOB$PC1), ]
```

Take top 500 genes # wObject <- data.frame(sort(abs(project.pca$rotation[,c(“PC1”,“PC2”)]), decreasing=TRUE)[1:500])

```
top500_PCA <- data.frame(sort(abs(project.pca$rotation[,c("PC1")]), de
creasing=TRUE)[1:500])
top500_PCA_KOA <- PC1_values_KOA[order(-PC1_values_KOA$PC1),][1:500,]
top500_PCA_KOB <- PC1_values_KOB[order(-PC1_values_KOB$PC1),][1:500,]
```

Take Top 20 genes sorted by PC value

```
top20_PCA <- PC1_values[order(-PC1_values$PC1),][1:20,]
```

12) Differential Gene Expression

DeSeq2 uses a negative binomial model to fit the observed read counts to arrive at the estimate for the difference. We need to estimate two parameters from the read counts: the mean as well as the dispersion. The null hypothesis is that there is no systematic difference between the average read count values of the different conditions for a given gene. The p-values are calculated and both tests are some variation of the well-known t-test (How dissimilar are the means of two populations?) or ANOVAs (How well does a reduced model capture the data when compared to the full model with all coefficients?). Once you've obtained a list of p-values for all the genes of your data set, it is important to realize that you just performed the same type of test for thousands and thousands of genes. That means, that even if you decide to focus on genes with a p-value smaller than 0.05, if you've looked at 10,000 genes your final list may contain 0:05/10; 000 = 500 false positive hits. To guard yourself against this, all the tools will offer some sort of correction for the multiple hypotheses you tested, e.g. in the form of the Benjamini-Hochberg formula. You should defnitely rely on the adjusted p-values rather than the original ones to zoom into possible candidates for downstream analyses and follow-up studies.

Running DEG analysis Here we want to look at the effect of the KO's versus the wildtype samples, with the wildtype values used as the denominator for the fold change calculation.

DESeq2 uses the levels of the condition to determine the order of the comparison so it important to set WT as the first-level factor

```
str(colData(dds)$genotype)
colData(dds)$condition <- relevel(colData(dds)$genotype, "WT")
str(colData(dds)$condition)

str(colData(dds_KOA)$genotype)
str(colData(dds_KOA)$condition)
colData(dds_KOA)$condition <- relevel(colData(dds_KOA)$genotype, "WT")
str(colData(dds_KOA)$condition)

str(colData(dds_KOB)$genotype)
str(colData(dds_KOB)$condition)
colData(dds_KOB)$condition <- relevel(colData(dds_KOB)$genotype, "WT")
str(colData(dds_KOB)$condition)
```

Run DeSeq2 The log2 fold change and Wald test p value will be for the last variable in the design formula (in our case just genotype), and if this is a factor, the comparison will be the last level of this variable (KO) over the reference level (control) The order of the variables of the design do not matter so long as the user specifies the comparison to build a results table for, using the name or contrast arguments of results.

The Wald method is the default

```
dds <- DESeq(dds, test="Wald")

## using pre-existing size factors

## estimating dispersions
## found already estimated dispersions, replacing these

## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates

## fitting model and testing

## -- replacing outliers and refitting for 18 genes
## -- DESeq argument 'minReplicatesForReplace' = 7
## -- original counts are preserved in counts(dds)

## estimating dispersions

## fitting model and testing

dds_KOA <- DESeq(dds_KOA, test="Wald")

## using pre-existing size factors

## estimating dispersions

## found already estimated dispersions, replacing these

## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates

## fitting model and testing

dds_KOB <- DESeq(dds_KOB, test="Wald")

## using pre-existing size factors

## estimating dispersions

## found already estimated dispersions, replacing these

## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates

## fitting model and testing
```

Log fold change shrinkage for visualisation and ranking Shrinkage of effect size (LFC estimates) is useful for visualization and ranking of genes.

```
resultsNames(dds)
resLFC <- lfcShrink(dds, coef="genotype_KO_vs_WT", type="apeglm")

## using 'apeglm' for LFC shrinkage. If used in published research, pl
ease cite:
##     Zhu, A., Ibrahim, J.G., Love, M.I. (2018) Heavy-tailed prior di
stributions for
##     sequence count data: removing the noise and preserving large di
```

```
fferences.
##     Bioinformatics. https://doi.org/10.1093/bioinformatics/bty895

head(resLFC, n=3)
```

13) Extract the results

By default the argument alpha is set to 0.1. If the adjusted p value cutoff will be a value other than 0.1, alpha should be set to that value: results function automatically performs independent filtering based on the mean of normalized counts for each gene, optimizing the number of genes which will have an adjusted p value below a given FDR cutoff, alpha.

Run the differential tests on the counts matrix and use FDR correction NB - use 'independentFiltering=FALSE' and 'cooksCutoff=FALSE' to switch off conversion of values to 'NA' if failing FDR conversion or Cook's Distance, respectively

```
unique(sample_info$genotype)
library(DESeq2)
options(scipen=999)

## Simple Extraction of the results
res <- results(dds)
summary(res)

## Modifying extraction for alpha = 0.05
ControlvsKO <- results(dds, pAdjustMethod="BH", independentFiltering =
TRUE, parallel = FALSE, alpha=0.05)
Results_convsko <- as.data.frame(ControlvsKO)
Results_convsko <- data.frame(rownames(Results_convsko), Results_convs
ko, row.names=rownames(Results_convsko))
names(Results_convsko)[1] <- "Geneid"
options(scipen=999)
head(Results_convsko)

# Merge with the file which contains geneid and gene names
Results_convsko <- join(Results_convsko, gtf,by="Geneid")
row.names(Results_convsko) <- Results_convsko$Geneid
head(Results_convsko)

# Control vs KOA
ControlvsKOA <- results(dds_KOA, pAdjustMethod="BH", independentFilter
ing = TRUE, parallel = FALSE, alpha=0.05)
Results_KOA <- as.data.frame(ControlvsKOA)
Results_KOA <- data.frame(rownames(Results_KOA), Results_KOA, row.name
s=rownames(Results_KOA))
names(Results_KOA)[1] <- "Geneid"
options(scipen=999)
head(Results_KOA)

Results_KOA <- join(Results_KOA, gtf,by="Geneid")
row.names(Results_KOA) <- Results_KOA$Geneid
head(Results_KOA)

# Control vs KOB
ControlvsKOB <- results(dds_KOB, pAdjustMethod="BH", independentFilter
ing = TRUE, parallel = FALSE, alpha=0.05)
Results_KOB <- as.data.frame(ControlvsKOB)
Results_KOB <- data.frame(rownames(Results_KOB), Results_KOB, row.name
s=rownames(Results_KOB))
names(Results_KOB)[1] <- "Geneid"
options(scipen=999)
head(Results_KOB)

Results_KOB <- join(Results_KOB, gtf,by="Geneid")
row.names(Results_KOB) <- Results_KOB$Geneid
head(Results_KOB)
```

The DESeqResult object can basically be handled like a data.frame

```
table(ControlvsKO$padj < 0.05)
table(ControlvsKO$padj < 0.05 & abs(ControlvsKO$log2FoldChange)>2)
```

Combine the deseq2 results with pc values

```
merged_pcresults <- merge(Results_convsko, sorted_PC1_values, by="Gene
id")
merged_pcresults_KOA <- merge(Results_KOA, sorted_PC1_values_KOA, by="
Geneid")
merged_pcresults_KOB <- merge(Results_KOB, sorted_PC1_values_KOB, by="
Geneid")
```

Sort results based on PC values

```
merged_pcresults_sorted <- merged_pcresults[order(-merged_pcresults$PC
1),]
row.names(merged_pcresults_sorted) <- merged_pcresults_sorted$Geneid
merged_pcresults_sorted <- merged_pcresults_sorted[,-c(14:18)]
names(merged_pcresults_sorted)[9:12] <- c("GeneSymbol", "Chromosome",
"Class","Strand")

merged_pcresults_sorted_KOA <- merged_pcresults_KOA[order(-merged_pcre
sults_KOA$PC1),]
row.names(merged_pcresults_sorted_KOA) <- merged_pcresults_sorted_KOA$
Geneid
merged_pcresults_sorted_KOA <- merged_pcresults_sorted_KOA[,-c(14:18)]
names(merged_pcresults_sorted_KOA)[9:12] <- c("GeneSymbol", "Chromosom
e", "Class","Strand")

merged_pcresults_sorted_KOB <- merged_pcresults_KOB[order(-merged_pcre
sults_KOB$PC1),]
row.names(merged_pcresults_sorted_KOB) <- merged_pcresults_sorted_KOB$
Geneid
merged_pcresults_sorted_KOB <- merged_pcresults_sorted_KOB[,-c(14:18)]
names(merged_pcresults_sorted_KOB)[9:12] <- c("GeneSymbol", "Chromosom
e", "Class","Strand")
```

Take the top 20 genes according to PCA analysis to obtain their LFC and padj values

```
merged_pcresults_sorted_top20 <- merged_pcresults_sorted[order(-merged
_pcresults_sorted$PC1),][1:20,]
merged_pcresults_sorted_top20_KOA <- merged_pcresults_sorted_KOA[order
(-merged_pcresults_sorted_KOA$PC1),][1:20,]
```

```
merged_pcresults_sorted_top20_KOB <- merged_pcresults_sorted_KOB[order
(-merged_pcresults_sorted_KOB$PC1),][1:20,]
```

Take the top 500 genes according to PCA analysis to obtain their LFC and padj values

```
merged_pcresults_sorted_top500 <- merged_pcresults_sorted[order(-merge
d_pcresults_sorted$PC1),][1:500,]
merged_pcresults_sorted_top500_KOA <- merged_pcresults_sorted_KOA[orde
r(-merged_pcresults_sorted_KOA$PC1),][1:500,]
merged_pcresults_sorted_top500_KOB <- merged_pcresults_sorted_KOB[orde
r(-merged_pcresults_sorted_KOB$PC1),][1:500,]
```

How many genes have a padj < 0.05 Control vs KO

```
sum(merged_pcresults_sorted$padj < 0.05, na.rm=TRUE)
```

Control vs KOA

```
sum(merged_pcresults_sorted_KOA$padj < 0.05, na.rm=TRUE)
```

Control vs KOB

```
sum(merged_pcresults_sorted_KOB$padj < 0.05, na.rm=TRUE)
```

Filter results to those with P<0.05

```
merged_pcresults_sorted_sig <- subset(merged_pcresults_sorted, padj<0.
05)
```

Filter results to those with a log2FC>2

```
merged_pcresults_sorted_sigFC2 <- subset(merged_pcresults_sorted_sig,
abs(log2FoldChange)>=2)
```

Can do this in one command - P<0.05 & logFC>|2|

```
merged_pcresults_sorted_sigFC2 <- subset(merged_pcresults_sorted, padj
<=0.05 & abs(log2FoldChange)>=2)

# Control vs KOA
merged_pcresults_sorted_KOA_sigFC2 <- subset(merged_pcresults_sorted_K
OA, padj<=0.05 & abs(log2FoldChange)>=2)

# Control vs KOB
merged_pcresults_sorted_KOB_sigFC2 <- subset(merged_pcresults_sorted_K
OB, padj<=0.05 & abs(log2FoldChange)>=2)
```

Order the results by the smallest pvalue

```
merged_pcresults_sorted_pvalue <- merged_pcresults_sorted[order(merged
_pcresults_sorted$padj), ]
```

Filter the genes which pass a fold change of > 1.5 and padj < 0.05

```
merged_pcresults_sorted_sigFC15 <- subset(merged_pcresults_sorted_sig,
padj<=0.05 & abs(log2FoldChange)>=1.5)
```

Filter for upregulated genes

```
upregulated_sig_DEGS <- subset(merged_pcresults_sorted_sigFC2, log2Fol
dChange>0)
upregulated_sig_DEGS_15 <- subset(merged_pcresults_sorted_sigFC15, log
2FoldChange>0)

# Control vs KOA
upregulated_sig_DEGs_KOA <- subset(merged_pcresults_sorted_KOA_sigFC2,
log2FoldChange>0)
# Control vs KOB
upregulated_sig_DEGs_KOB <- subset(merged_pcresults_sorted_KOB_sigFC2,
log2FoldChange>0)
```

Filter for downregulated genes

```
downregulated_sig_DEGS <- subset(merged_pcresults_sorted_sigFC2, log2F
oldChange < 0)
downregulated_sig_DEGS_15 <- subset(merged_pcresults_sorted_sigFC15, l
og2FoldChange < 0)

# Control vs KOA
downregulated_sig_DEGs_KOA <- subset(merged_pcresults_sorted_KOA_sigFC
2, log2FoldChange<0)
# Control vs KOB
downregulated_sig_DEGs_KOB <- subset(merged_pcresults_sorted_KOB_sigFC
2, log2FoldChange<0)
```

Obtain gene list

```
DEG_geneids_FC2 <- row.names(merged_pcresults_sorted_sigFC2)
DEG_genenames_FC2 <- data.frame(merged_pcresults_sorted_sigFC2$GeneSym
bol)
names(DEG_genenames_FC2) <- c("GeneSymbol")
head(DEG_genenames_FC2, n=3)
```

14) Plots after Differential Gene Analysis

MA Plots The MA plot provides a global view of the relationship between the expression change between conditions (log ratios, M), the average expression strength of the genes (average mean, A) and the ability of the algorithm to detect differential gene expression: genes that pass the significance threshold (adjusted p-value<0.05) are colored in red

Points will be colored red if the adjusted p value is less than 0.05

```
# Add two lines where logFC > 2 and logFC < -2
drawLines <- function() abline(h=c(-2,2),col="dodgerblue",lwd=2)

par(mar=c(4,4,4,4), mfrow=c(1,1), cex=1.0, cex.main=0.5, cex.axis=0.8)
plot.new()

# MA plot of significant DEGs
plotMA(ControlvsKO, alpha=0.05, main="WT vs. KO")
```

```
# It is more useful visualize the MA-plot for the shrunken log2 fold c
hanges, which remove the noise associated with log2 fold changes from
```

```
low count genes without requiring arbitrary filtering thresholds.
plotMA(resLFC, alpha=0.05, main="WT vs. KO Log Fold Shrinkage")
```

```
## Changing the thresholds to what you want to see
## plotting all those which pass p<0.05 with a folchange >2 or <2 or b
etween -2->2
## lfcThreshold of 1, means log2 fold change is 2
par(mfrow=c(2,2),mar=c(2,2,1,1))

ylim <- c(-10,10)
# Two tailed test where log2 FC is > or < lfc threshold
resGA <- results(dds, lfcThreshold=log2(2), alpha=0.05, altHypothesis=
"greaterAbs")
resGA <- results(dds, lfcThreshold=1, alpha=0.05, altHypothesis="great
erAbs")

# p values are the maximum of the upper and lower tests
resLA <- results(dds, lfcThreshold=log2(2), alpha=0.05, altHypothesis=
"lessAbs")
# log2 FC greater than threshold and P<0.05
resG <- results(dds, lfcThreshold=log2(2), alpha=0.05, altHypothesis="
greater")
# log 2 FC smaller than threshold and P<0.05
resL <- results(dds, lfcThreshold=log2(2), alpha=0.05, altHypothesis="
less")

drawLines <- function() abline(h=c(-1,1),col="dodgerblue",lwd=2)
plotMA(resGA, ylim=ylim, alpha=0.05); drawLines()
plotMA(resLA, ylim=ylim, alpha=0.05); drawLines()
plotMA(resG, ylim=ylim, alpha=0.05); drawLines()
plotMA(resL, ylim=ylim, alpha=0.05); drawLines()
```

```
# Test genes which have fold change more than doubling or less than ha
lving
# lfcThreshold of 1, means log2 fold change is 2 - doubling
#res.thr <- results(dds, lfcThreshold=1, alpha=0.05)
#plotMA(res.thr, ylim=c(-10,10))
```

Histogram of p-values

```
par(mfrow=c(1,1), cex=1.5)
hist(ControlvsKO$padj,
     col= "grey", border = "white", xlab = "P-adjusted value", ylab =
"Frequency",
     main = "Frequencies of p-values for Control vs KO")
```

Bar Plots of Gene expression Plot the gene with the minimum adjusted p-value

```
plotCounts(dds, gene=which.min(ControlvsKO$padj), intgroup="condition"
, normalized=TRUE)
```

More sophisticated plot

```
library(ggplot2)
data <- plotCounts(dds, gene=which.min(ControlvsKO$padj), intgroup=c("
genotype","condition","rep", "sample"), returnData=TRUE)
ggplot(data, aes(x=genotype, y=count, shape=rep, colour=condition)) +
  scale_shape_manual(name="Replicate",
                     labels=c("1","2","3","4"),
                     values = rep(c(15,16,17,18))) +
  scale_colour_manual(name="Condition",
                      labels=c("Control", "KO"),
                      values = rep(c("#B2DFEE","#FFA500"))) +
  geom_point(position=position_jitter(width=.1,height=0)) +
  ggtitle("Expression of RPL22L1") +
  scale_y_log10()
```

Read counts of single genes DESeq2 offers a wrapper function to plot read counts for single genes

```
library (grDevices ) # for italicizing the gene name
# EPO: ENSG00000130427
data <- plotCounts(dds, gene="ENSG00000130427", intgroup=c("genotype",
"condition","rep", "sample"), returnData=TRUE)
ggplot(data, aes(x=genotype, y=count, col=sample_info$condition, shape
=rep)) +
  geom_point(position=position_jitter(width=.1,height=0)) +
  ggtitle("Expression of EPO") +
  scale_y_log10()
```

Plot the expression of the top 50 genes on a bar plot

```
# Load in packages
library(tibble)

## Warning: package 'tibble' was built under R version 3.6.2

library(tidyr)

## Warning: package 'tidyr' was built under R version 3.6.2

##
## Attaching package: 'tidyr'

## The following object is masked from 'package:S4Vectors':
##
##     expand

## The following object is masked from 'package:magrittr':
##
##     extract

library(ggplot2)
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyv
erse 1.3.0 ──
```

```
## ✓ readr   1.4.0     ✓ forcats 0.5.1
## ✓ purrr   0.3.4

## Warning: package 'readr' was built under R version 3.6.2

## Warning: package 'purrr' was built under R version 3.6.2

## Warning: package 'forcats' was built under R version 3.6.2

## ── Conflicts ──────────────────────────────────────────── tidyverse_c
onflicts() ──
## x plyr::arrange()     masks dplyr::arrange()
## x dplyr::collapse()   masks IRanges::collapse()
## x dplyr::combine()    masks Biobase::combine(), BiocGenerics::combi
ne()
## x purrr::compact()    masks plyr::compact()
## x plyr::count()       masks dplyr::count(), matrixStats::count()
## x plyr::desc()        masks dplyr::desc(), IRanges::desc()
## x tidyr::expand()     masks S4Vectors::expand()
## x tidyr::extract()    masks magrittr::extract()
## x plyr::failwith()    masks dplyr::failwith()
## x dplyr::filter()     masks stats::filter()
## x dplyr::first()      masks S4Vectors::first()
## x plyr::id()          masks dplyr::id()
## x dplyr::lag()        masks stats::lag()
## x plyr::mutate()      masks dplyr::mutate()
## x ggplot2::Position() masks BiocGenerics::Position(), base::Positio
n()
## x purrr::reduce()     masks GenomicRanges::reduce(), IRanges::reduc
e()
## x plyr::rename()      masks dplyr::rename(), S4Vectors::rename()
## x purrr::set_names()  masks magrittr::set_names()
## x purrr::simplify()   masks DelayedArray::simplify()
## x dplyr::slice()      masks IRanges::slice()
## x plyr::summarise()   masks dplyr::summarise()
## x plyr::summarize()   masks dplyr::summarize()

# create tibble
merged_pcresults_sorted_tb <- merged_pcresults_sorted  %>%
  data.frame() %>%
  rownames_to_column(var="gene") %>%
  as_tibble()

# subset to significant DEGs
DEGs_ControlvsKO_tb <- subset(merged_pcresults_sorted_tb, abs(log2Fold
Change) >= 2 & padj<0.05)

# pull out the top 50 DEGS based on pvalue
top50_DEGs_genes <- DEGs_ControlvsKO_tb %>%
  arrange(-PC1) %>%     #Arrange rows by padj values (can change this
if you want to any column)
  pull(gene) %>%        #Extract character vector of ordered genes
  head(n=50) # change this to number you want to pull out

top50_DEG_genes <- merged_pcresults_sorted[top50_DEGs_genes,] # pull o
ut from merged pcresults
top50_DEG_genes <- data.frame(top50_DEG_genes) #make into a data frame
```

```
top50_DEG_genes$Geneid <- row.names(top50_DEG_genes) # change row name
s to Gene ID
top50_DEG_genes <- merge(top50_DEG_genes, gtf, by="Geneid") # merge wi
th gtf file so can obtain gene names

#change the normalised counts into a tibble
normalized_counts_tb <- norm %>%
  data.frame() %>%
  rownames_to_column(var="gene") %>%
  as_tibble()

# pull out the top 50 genes from the normalised counts tibble
top50_DEG_norm <- normalized_counts_tb %>%
  filter(gene %in% top50_DEGs_genes)

gathered_top50_DEGs <- top50_DEG_norm %>%
  tidyr::gather(colnames(top50_DEG_norm)[2:13], key = "sample", value
= "normalized_counts")

gathered_top50_DEGs <- inner_join(gathered_top50_DEGs,sample_info, by=
"sample")
names(gathered_top50_DEGs)[1] <- "Geneid"
gathered_top50_DEGs <- inner_join(gtf, gathered_top50_DEGs, by="Geneid
")
gathered_top50_DEGs <- inner_join(merged_pcresults_sorted, gathered_to
p50_DEGs, by="Geneid")

ggplot(gathered_top50_DEGs) +
  geom_point(aes(x = GeneSymbol.x, y = normalized_counts, color = cond
ition)) +
  xlab("Genes") +
  ylab("Regularised Log Counts") +
  ggtitle("Top 50 Significantly Differentially Expressed Genes (P<0.05
& abs(Log2FC) > 2)") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size=6)) +
  theme(plot.title = element_text(hjust = 0.5, size=10))
```

Plotting Heat maps of the count matrix

```
library(pheatmap)
select <- order(rowMeans(counts(dds,normalized=TRUE)),
                decreasing=TRUE)[1:50]

df <- as.data.frame(colData(dds)[,c("genotype", "condition", "rep")])

#Heat map of the Count matrix
pheatmap(data.matrix(norm[select,]), cluster_rows=F, show_rownames=TRU
E,
         cluster_cols=FALSE, annotation_col=df, main="Normalised Count
s", cex=0.8)
```

```
# on the vsd data
pheatmap(assay(vsd)[select,], cluster_rows=T, show_rownames=F,
         cluster_cols=FALSE, annotation_col=df, main="VSD Transformati
on")
```

```
# onthe rlog data
pheatmap(assay(rld)[select,], cluster_rows=T, show_rownames=FALSE,
         cluster_cols=F, annotation_col=df, main="rLog Transformation"
)
```

Heat map of the top most expressed genes based on comparing individual transcripts across all samples Select the most highly expressed genes from the study and perform clustering

```
library("RColorBrewer")
library("gplots")

par(cex=1.0)
select <- order(rowMeans(norm), decreasing=TRUE)[1:50]
hmcol <- colorRampPalette(brewer.pal(9, "BuGn"))(50)
y <- data.matrix(norm[select,])

# now add gene names rather than symbols
ya <- data.frame(y)
ya$Geneid <- row.names(ya)
ya <- join(ya,gtf, by="Geneid")
row.names(ya) <- ya$GeneSymbol
ya <- ya[,-c(13:18)]
ya <- data.matrix(ya)

heatmap.2(ya, main="All samples", cexRow=0.8, cexCol=1.2, offsetCol=1,
labCol=c("Control", NA, NA,NA,"   KOA", NA,NA,NA, "   KOB",NA,NA,NA),
col=hmcol, Rowv=TRUE, Colv=F, scale="none", ColSideColors=c("#B2DFEE",
"#B2DFEE", "#B2DFEE",  "#B2DFEE", "#FFA500","#FFA500","#FFA500","#FFA5
00","#FF1493","#FF1493","#FF1493","#FF1493"), srtCol=0, adjCol=c(-2.2,
-68),dendrogram="none", trace="none", key.title="Counts density", key.
xlab="Counts", key.ylab="")
par(xpd=T)
text(0.05,0.65, "Increasing\nexpression\n(top 50)")
arrows(0.05, 0.6, 0.05, 0.4, lwd=3, xpd=T)
```

Plot the top 20 DEGs sorted by PC1

```
topgenes <- head(rownames(merged_pcresults_sorted_sigFC2),20)
topgenes
mat <- rldMatrix[topgenes,]
mat <- mat - rowMeans(mat)
mat2 <- data.frame(mat)
mat2$Geneid <- row.names(mat2)
mat2 <- join(mat2, gtf,by="Geneid")
row.names(mat2) <- mat2$GeneSymbol
mat2 <- mat2[,-c(13:18)]
```

```
mat2 <- data.matrix(mat2)

hmcol <- colorRampPalette(brewer.pal(11, "RdBu"))(50)
df <- as.data.frame(colData(dds)[,c("genotype", "condition")])
heatmap.2(mat2, main="Top 20 DEGs", cexRow=0.8, cexCol=0.8, col=hmcol,
Rowv=T, Colv=F, labCol=c("Control", NA,NA,NA,"   KOB",NA,NA,NA), adjCo
l=c(-1.6,-62),srtCol=0, scale="none", dendrogram="none", trace="none",
margin=c(1, 5), key.title="Counts density", key.xlab="Counts", key.yla
b="")
```

Plot the top 50 DEGs sorted by PC1

```
topgenes <- head(rownames(merged_pcresults_sorted_sigFC2),50)
mat <- rldMatrix[topgenes,]
mat <- mat - rowMeans(mat)
mat2 <- data.frame(mat)
mat2$Geneid <- row.names(mat2)
mat2 <- join(mat2, gtf,by="Geneid")
row.names(mat2) <- mat2$GeneSymbol
mat2 <- mat2[,-c(13:18)]
mat2 <- data.matrix(mat2)

hmcol <- colorRampPalette(brewer.pal(11, "RdBu"))(50)
df <- as.data.frame(colData(dds)[,c("genotype", "condition")])

library(gplots)
heatmap.2(mat2, main="Top 50 DEGs", cexRow=0.6, cexCol=0.8, col=hmcol,
Rowv=T, Colv=F, labCol=c("Control", NA,NA,NA,"   KOB",NA,NA,NA), adjCo
l=c(-1.6,-62),srtCol=0, scale="none", dendrogram="none", trace="none",
margin=c(1, 5), key.title="Counts density", key.xlab="Counts", key.yla
b="")
```

```
#pheatmap(mat2, annotation_col=df, fontsize_row = 5, fontsize_col = 7,
cluster_cols = F,cluster_rows = T, main="Top 50 DEGs \n P<0.05 & abs|L
og2 Fold Change|>2")
```

Scaled to z-scores

```
# scale = "row" converts to z-score scaling within rows
plot.new()
heatmap.2(mat2, main="Top 50 DEGs", cexRow=0.6, cexCol=0.8, col=hmcol,
Rowv=T, Colv=F, labCol=c("Control", NA,NA,NA,"   KOA",NA,NA,NA,"   KOB
",NA,NA,NA), adjCol=c(-1.6,-62),srtCol=0, scale="row", dendrogram="non
e", trace="none", margin=c(1, 5), key.title="Counts density", key.xlab
="Counts", key.ylab="")
```

Alternative heatmap

```
pheatmap(mat2, annotation_col=df, fontsize_row = 5, fontsize_col = 7,
cluster_cols = F,cluster_rows = T, main="Top 50 DEGs \n P<0.05 & abs|L
og2 Fold Change|>2")
```

Plotting the top 500 genes by PC1

```
topgenes <- head(rownames(merged_pcresults_sorted),500)
mat <- rldMatrix[topgenes,]
mat <- mat - rowMeans(mat)
mat2 <- data.frame(mat)
mat2$Geneid <- row.names(mat2)
mat2 <- join(mat2, gtf,by="Geneid")
row.names(mat2) <- mat2$GeneSymbol
mat2 <- mat2[,-c(13:18)]
mat2 <- data.matrix(mat2)

hmcol <- colorRampPalette(brewer.pal(11, "RdBu"))(50)
df <- as.data.frame(colData(dds)[,c("genotype", "condition")])

# scale = "row" converts to z-score scaling within rows
heatmap.2(mat2, main="Top 50 DEGs", cexRow=0.6, cexCol=0.8, col=hmcol,
Rowv=T, Colv=F, labCol=c("Control", NA,NA,NA,"   KOB",NA,NA,NA), adjCo
l=c(-1.6,-62),srtCol=0, scale="row", dendrogram="none", trace="none",
margin=c(1, 5), key.title="Counts density", key.xlab="Counts", key.yla
b="")
```

Plotting all the significant DEGs (p<0.05 & abs|log2FC|>2)

```
# load the library with the aheatmap () function
library(NMF)

## Loading required package: pkgmaker

## Loading required package: registry

##
## Attaching package: 'pkgmaker'

## The following object is masked from 'package:S4Vectors':
##
##     new2

## Loading required package: rngtools

## Loading required package: cluster

## NMF - BioConductor layer [OK] | Shared memory capabilities [NO: big
memory] | Cores 3/4

##   To enable shared memory capabilities, try: install.extras('
## NMF
## ')

##
## Attaching package: 'NMF'

## The following object is masked from 'package:plotrix':
##
##     dispersion
## The following object is masked from 'package:DelayedArray':
##
##     seed

## The following object is masked from 'package:BiocParallel':
##
##     register

## The following object is masked from 'package:S4Vectors':
##
##     nrun

# aheatmap needs a matrix of values , e.g., a matrix of DE genes with
the transformed read counts for each replicate
# identify gene names with desired cut off and fold change cut off
DEG_genenames_FC2 <- rownames(merged_pcresults_sorted_sigFC2)
head(DEG_genenames_FC2, n=3)
DEGs_up <- rownames(merged_pcresults_sorted_sigFC2[merged_pcresults_so
rted_sigFC2$log2FoldChange>0,])
DEGs_down <- rownames(merged_pcresults_sorted_sigFC2[merged_pcresults_
sorted_sigFC2$log2FoldChange<0,])

# extract the normalized read counts for DE genes into a matrix
hm.mat_DGEgenes <- rldMatrix[DEG_genenames_FC2, ]
hm.mat_DGEgenes_up <- log.norm.counts[DEGs_up, ]
hm.mat_DGEgenes_down <- log.norm.counts[DEGs_down, ]

# plot the normalized read counts of DE genes sorted by the adjusted p
- value
aheatmap(hm.mat_DGEgenes, Rowv = NA , Colv = NA)
```

```
# combine the heatmap with hierarchical clustering
# aheatmap(hm.mat_DGEgenes,
         # Rowv = TRUE, Colv=TRUE, # add dendrograms to rows and colum
ns
         #distfun = "euclidean", hclustfun = "average", annRow=NA, lab
Row=NA)

# scale the read counts per gene to emphasize the sample-type - specif
ic differences
# the read count values are scaled per row so that the colors actually
represent z-scores rather than the underlying read counts.

annotation = data.frame(condition=sample_info$condition)
# aheatmap(hm.mat_DGEgenes,
  #       Rowv = TRUE , Colv = TRUE ,
   #      distfun = "pearson", hclust = "ward",
    #     scale = "row",annCol = annotation, labRow=NA, annRow=NA, mai
n="Heat Map \n DEG Genes with P<0.05 & LogFC > |2|") # values are tran
sformed into distances from the center of the row - specific average :
( actual value - mean of the group ) / standard deviation
```

Complex heat map

```
sigGeneList <- row.names(data.frame(subset(ControlvsKO, abs(log2FoldCh
ange)>=2 & padj<=0.05)))
myCol <- colorRampPalette(c("green", "black", "red"))(10)
myBreaks <- seq(-3, 3, length.out=10)

heat <- t(rldMatrix)[,sigGeneList]
heat <- t(scale(t(heat)))

sampleOrder <- c(
  c("WT1","WT2","WT3","WT4"),
  c("KOA1","KOA2","KOA3","KOA4"), c("KOB1", "KOB2", "KOB3", "KOB4"))

library(ComplexHeatmap)

## Loading required package: grid

## ========================================
## ComplexHeatmap version 2.0.0
## Bioconductor page: http://bioconductor.org/packages/ComplexHeatmap/
## Github page: https://github.com/jokergoo/ComplexHeatmap
## Documentation: http://jokergoo.github.io/ComplexHeatmap-reference
##
## If you use it in published research, please cite:
## Gu, Z. Complex heatmaps reveal patterns and correlations in multidi
mensional
##   genomic data. Bioinformatics 2016.
## ========================================

library(circlize)
library(cluster)
ann <- data.frame(condition=sample_info$condition)
colnames(ann) <- c("Condition")
colours <- list("Control"="lightblue", "KOA"="purple", "KOB"="pink")
rowAnn <- rowAnnotation(df=ann, boxplot=row_anno_boxplot(heat, border=
FALSE, show_annotation_name=FALSE, gp=gpar(fill="#CCCCCC", fontsize=2)
, lim=c(-4,4), pch=".", size=unit(2, "mm"), col=sample_info$condition,
annotation_width=unit(c(1, 7.5), "cm")))
hmap <- Heatmap(heat,
                #split=sampleOrderSplit,
                row_order=sampleOrder,
                name="Gene expression\nZ-score",
                col=colorRamp2(myBreaks, myCol),
                heatmap_legend_param=list(color_bar="continuous", lege
nd_direction="vertical", legend_width=unit(4,"cm"), title_position="to
pcenter", title_gp=gpar(fontsize=6, fontface="bold")),

                cluster_rows=FALSE,
                show_row_dend=FALSE,
                row_title="",
                row_title_side="left",
                row_title_gp=gpar(fontsize=8,  fontface="bold"),
                row_title_rot=0,
                show_row_names=TRUE,
                row_names_gp=gpar(fontsize=8, fontface="bold"),
                row_names_side="left",
                row_dend_width=unit(30,"mm"),
                cluster_columns=TRUE,
                show_column_dend=TRUE,
                column_title="Transcripts",
                column_title_side="bottom",
                column_title_gp=gpar(fontsize=12, fontface="bold"),
                column_title_rot=0,
                show_column_names=FALSE,
                #column_names_gp=gpar(fontsize=termLab, fontface="bold
"),
                #column_names_max_height=unit(15, "cm"),
                column_dend_height=unit(50,"mm"),

                clustering_distance_columns=function(x) as.dist(1-cor(
t(x))),
                clustering_method_columns="ward.D2",
                clustering_distance_rows=function(x) as.dist(1-cor(t(x
))),
                clustering_method_rows="ward.D2")

#top_annotation_height=unit(1.75,"cm"),
#bottom_annotation=sampleBoxplot)
#bottom_annotation_height=unit(4, "cm"))
draw(hmap + rowAnn, heatmap_legend_side="left", annotation_legend_side
="left")
```

Heat map of all significant genes

```
# set the thresholds
padj.cutoff <- 0.05
lfc.cutoff <- 2
sigOE <- merged_pcresults_sorted_tb %>%
  filter(padj < padj.cutoff & abs(log2FoldChange) > lfc.cutoff)

norm_OEsig <- normalized_counts_tb[,c(1,2:13)] %>%
  filter(gene %in% sigOE$gene) %>%
  data.frame() %>%
  column_to_rownames(var = "gene")

annotation <- sample_info %>%
  select(sample, condition) %>%
  data.frame(row.names = "sample")

heat_colors <- brewer.pal(6, "YlOrRd")

pheatmap(norm_OEsig,
         color = heat_colors,
         cluster_rows = T,
         show_rownames = F,
         annotation = annotation,
         border_color = NA,
         fontsize = 10,
         scale = "row",
         fontsize_row = 10,
         height = 20)
```

Volcano Plot

```
# main <- "KO vs WT Volcano Plot"
# merged_pcresults_sorted$genenames <- merged_pcresults_sorted$GeneSymbol
# toptable <- data.frame(merged_pcresults_sorted)
#
# # check how many NA there are under padj
# sum(is.na(merged_pcresults_sorted$padj))

# source("~/Documents/EPO Project/CRISPR/Whole gene knock-out/Confirming EPO KO/RNA Sequencing/DeSeq2/VolcanoPlot.R")
# volcano <- VolcanoPlot(toptable, 0.05, 0.05, 0.05, 2, "Volcano Plot Control vs KO")
# volcano

# Control vs KOA
# main <- "KOA vs WT Volcano Plot"
# merged_pcresults_sorted_KOA$genenames <- merged_pcresults_sorted_KOA$GeneSymbol
# toptable <- data.frame(merged_pcresults_sorted_KOA)

# check how many NA there are under padj
# sum(is.na(merged_pcresults_sorted_KOA$padj))

# source("~/Documents/EPO Project/CRISPR/Whole gene knock-out/Confirming EPO KO/RNA Sequencing/DeSeq2/VolcanoPlot.R")
# volcano <- VolcanoPlot(toptable, 0.05, 0.05, 0.05, 2, "Volcano Plot Control vs KOA")
# volcano

# Control vs KOB
# main <- "KOB vs WT Volcano Plot"
# merged_pcresults_sorted_KOB$genenames <- merged_pcresults_sorted_KOB$GeneSymbol
# toptable <- data.frame(merged_pcresults_sorted_KOB)

# check how many NA there are under padj
#sum(is.na(merged_pcresults_sorted_KOB$padj))

# source("~/Documents/EPO Project/CRISPR/Whole gene knock-out/Confirming EPO KO/RNA Sequencing/DeSeq2/VolcanoPlot.R")
# volcano <- VolcanoPlot(toptable, 0.05, 0.05, 0.05, 2, "Volcano Plot Control vs KOB")
# volcano
```

Focusing on the overlapping significant DEGs

As we have two KO cell lines, we want to refine our list of overlapping DEGs and identify those which as significantly differentially expressed between Control and both KOA and KOB.

First we need to merge the results of Control vs KOA and Control vs KOB together

```
merge_KOA_KOB <- merge(merged_pcresults_sorted_KOA,merged_pcresults_so
rted_KOB,by="Geneid", suffix=c(".WTvsKOA", ".WTvsKOB"), all=T)
```

Calculate the mean PC by average PC1 for WTvsKOA and PC1 for WTvsKOB

```
merge_KOA_KOB$meanPC <- (merge_KOA_KOB$PC1.WTvsKOA + merge_KOA_KOB$PC1
.WTvsKOB)/2
#sort by mean PC value
merge_KOA_KOB_pcsorted <- merge_KOA_KOB[order(-merge_KOA_KOB$meanPC),
]
```

Merge with the overall results when combining all KOs

```
merge_KOA_KOB_pcsorted_all <- merge(merge_KOA_KOB_pcsorted, merged_pcr
esults_sorted, by="Geneid")
# Sort by mean PC
merge_KOA_KOB_pcsorted_all <- merge_KOA_KOB_pcsorted_all[order(-merge_
KOA_KOB_pcsorted_all$meanPC),]

# write.table(merge_KOA_KOB_pcsorted_all, "~/Documents/EPO Project/CRI
SPR/Whole gene knock-out/Confirming EPO KO/RNA Sequencing/DeSeq2/Overl
apping DEG analysis/merged_KOA_KOB_analysis_PC_sorted.txt", sep="/t",
row.names=F, quote=F)
```

Take top 500 genes based on average PC value

```
top500_merge_KOA_KOB_pcsorted_all <- merge_KOA_KOB_pcsorted_all[order(
-merge_KOA_KOB_pcsorted_all$meanPC),][1:500,]

# write.table(top500_merge_KOA_KOB_pcsorted_all, "~/Documents/EPO Proj
ect/CRISPR/Whole gene knock-out/Confirming EPO KO/RNA Sequencing/DeSeq
2/Overlapping DEG analysis/Top_500_DEGs_basedon_meanPC1_KOA_KOB_analys
is.txt", sep="\t", row.names=F, quote=F)
```

Heatmap of the top 500 genes

```
row.names(top500_merge_KOA_KOB_pcsorted_all) <- top500_merge_KOA_KOB_p
csorted_all$Geneid
topgenes <- head(rownames(top500_merge_KOA_KOB_pcsorted_all),500)
mat <- rldMatrix[topgenes,]
mat <- mat - rowMeans(mat)
mat2 <- data.frame(mat)
mat2$Geneid <- row.names(mat2)
mat2 <- join(mat2, gtf,by="Geneid")
row.names(mat2) <- mat2$GeneSymbol
mat2 <- mat2[,-c(13:18)]
mat2 <- data.matrix(mat2)

hmcol <- colorRampPalette(brewer.pal(11, "RdBu"))(50)
df <- as.data.frame(colData(dds)[,c("genotype", "condition")])

# scale = "row" converts to z-score scaling within rows
heatmap.2(mat2, main="Top 50 DEGs", cexRow=0.6, cexCol=0.8, col=hmcol,
Rowv=T, Colv=T, labCol=c("Control", NA,NA,NA,"   KOA",NA,NA,NA,"   KOB
",NA,NA,NA), adjCol=c(-0.8,-62),srtCol=0, scale="row", dendrogram="col
", trace="none", margin=c(1, 5), key.title="Counts density", key.xlab=
"Counts", key.ylab="")
```

```
pheatmap(mat2, annotation_col=df,fontsize_col = 7, clustering_distance
_rows = "euclidean", clustering_method="ward.D2", annotation_row=NA, s
how_rownames=F, cluster_cols = F,cluster_rows = T, main="Top 500 for s
egregating Control from KO along PC1 in PCA")
```

Venn Diagram of WT vs KOA & WT vs KOB

1. Venn diagram of the overall files of for KOA and KOB (how many genes overlap in the beginning)

```
library(VennDiagram)

## Loading required package: futile.logger

##
## Attaching package: 'VennDiagram'

## The following object is masked from 'package:dendextend':
##
##     rotate

y <- list()
y$KOA <- as.character(row.names(merged_pcresults_sorted_KOA))
y$KOB <- as.character(row.names(merged_pcresults_sorted_KOB))
# Generate plot
myCol <- brewer.pal(3, "RdBu")
# Pink: #FF9CEE, purple/blue: #AFCBFF, pastel yellow: #FFF5BA, Light b
lue: #ACE7FF
w <- venn.diagram(y,
                  #circles
                  lwd = 2,
                  lty = 'blank',
                  col="transparent",
                  fill = c(alpha("#FFF5BA",0.5), alpha('#ACE7FF',0.5))
, cex=2,
                  # font inside circles
                  fontface = "plain",
                  fontfamily = "sans",
                  # category names
                  cat.cex=2, cat.fontface="bold", cat.default.pos="out
er",
                  cat.pos=c(-10,10), cat.dist=c(0.055,0.055), cat.font
family="sans",
                  filename=NULL,alpha=0.7, scaled=TRUE)
grid.newpage()
grid.draw(w)
```

2. Create a Venn diagram of the overlapping significant DEGs Those which are significant (p<0.05 & abs(log2FC)>2) in both Control vs KOA and Control vs KOB

```
x <- list()
x$WTvsKOA <- as.character(row.names(merged_pcresults_sorted_KOA_sigFC2
))
```

```
x$WTvsKOB <- as.character(row.names(merged_pcresults_sorted_KOB_sigFC2
))
# Generate plot
# Pink: #FF9CEE, purple/blue: #AFCBFF, pastel yellow: #FFF5BA, Light b
lue: #ACE7FF
v <- venn.diagram(x,
                  #circles
                  lwd = 2,
                  lty = 'blank',
                  col="transparent",
                  fill = c(alpha("#FFF5BA",0.5), alpha('#ACE7FF',0.5))
, cex=2,
                  # font inside circles
                  fontface = "plain",
                  fontfamily = "sans",
                  # category names
                  cat.cex=2, cat.fontface="bold", cat.default.pos="out
er",
                  cat.pos=c(-10,10), cat.dist=c(0.055,0.055), cat.font
family="sans",
                  filename=NULL,alpha=0.7, scaled=TRUE)

# have a look at the default plot
grid.newpage()
grid.draw(v)
```

We now know that 314 genes are overlapping and significant in both KOs.

3. Extract the list of genes in each group

```
# have a look at the names in the plot object v
# We are interested in the labels
lapply(v, function(i) i$label)

# Over-write labels (5 to 7 chosen by manual check of labels)
# in KOA only
v[[5]]$label  <- paste(setdiff(x$WTvsKOA, x$WTvsKOB), collapse="\n")
only_in_KOA  <- data.frame(paste(setdiff(x$WTvsKOA, x$WTvsKOB)))

# in KOB only
v[[6]]$label <- paste(setdiff(x$WTvsKOB, x$WTvsKOA)  , collapse="\n")
only_in_KOB  <- data.frame(paste(setdiff(x$WTvsKOB, x$WTvsKOA)))

# intersection i.e. the genes that are signifcant in both analyses
v[[7]]$label <- paste(intersect(x$WTvsKOA, x$WTvsKOB), collapse="\n")
KOA_KOB  <- data.frame(paste(intersect(x$WTvsKOA, x$WTvsKOB)))

## Obtain gene names and their values for each list
# overlapping list
KOA_KOB$Geneid <- KOA_KOB$paste.intersect.x.WTvsKOA..x.WTvsKOB..
KOA_KOB_values <- merge(KOA_KOB, merge_KOA_KOB_pcsorted_all, by="Genei
d")
sum(abs(KOA_KOB_values$padj.WTvsKOB)>0.05)

# Only significant DEG in KOA
only_in_KOA$Geneid <- only_in_KOA$paste.setdiff.x.WTvsKOA..x.WTvsKOB..
```

```
only_in_KOA_values <- merge(only_in_KOA, merge_KOA_KOB, by="Geneid")

# Only significant DEG in KOB
only_in_KOB$Geneid <- only_in_KOB$paste.setdiff.x.WTvsKOB..x.WTvsKOA..
only_in_KOB_values <- merge(only_in_KOB, merge_KOA_KOB, by="Geneid")
```

4. Create a column with the mean PC values

```
# Create a column of mean PC values
KOA_KOB_values$meanPC <- (KOA_KOB_values$PC1.WTvsKOA + KOA_KOB_values$
PC1.WTvsKOB)/2
```

5. Merge with wt vs all KO file to see if they come up in main analysis

```
KOA_KOB_values_combined <- merge(KOA_KOB_values, merged_pcresults, by=
"Geneid")
sum(KOA_KOB_values_combined$padj>0.05, na.rm=T)
```

6. Sort by mean PC value

```
KOA_KOB_values_sortedPC <- KOA_KOB_values[order(-KOA_KOB_values$meanPC
), ]
# write.table(KOA_KOB_values, file="~/Documents/EPO Project/CRISPR/Who
le gene knock-out/Confirming EPO KO/RNA Sequencing/DeSeq2/Overlapping
DEG analysis/Overlapping_genes_KOA_KOB_analysis_PC1sorted.txt", sep="\
t", row.names=F, quote=F)
```

7. Obtain a list of Upregulated and Downregulated genes

```
#### Upregulated genes
upreg_KOA_KOB_overlap <- KOA_KOB_values_sortedPC[KOA_KOB_values_sorted
PC$log2FoldChange>0,]
# write.table(upreg_KOA_KOB_overlap, file="~/Documents/EPO Project/CRI
SPR/Whole gene knock-out/Confirming EPO KO/RNA Sequencing/DeSeq2/Overl
apping DEG analysis/Overlapping_genes_KOA_KOB_analysis_Upregulated.txt
", sep="\t", row.names=F, quote=F)

#### Downregulated genes
downreg_KOA_KOB_overlap <- KOA_KOB_values_sortedPC[KOA_KOB_values_sort
edPC$log2FoldChange<0,]
# write.table(downreg_KOA_KOB_overlap, file="~/Documents/EPO Project/C
RISPR/Whole gene knock-out/Confirming EPO KO/RNA Sequencing/DeSeq2/Ove
rlapping DEG analysis/Overlapping_genes_KOA_KOB_analysis_Downregulated
.txt", sep="\t", row.names=F, quote=F)
```

8. Plot all of the overlapping genes on a heat map

```
row.names(KOA_KOB_values_sortedPC) <- KOA_KOB_values_sortedPC$Geneid
overlapping_genes <- rownames(KOA_KOB_values_sortedPC)
library(plyr)
library(RColorBrewer)
mat <- rldMatrix[overlapping_genes,]
mat <- mat - rowMeans(mat)
mat2 <- data.frame(mat)
mat2$Geneid <- row.names(mat2)
mat2 <- join(mat2, gtf,by="Geneid")
row.names(mat2) <- mat2$GeneSymbol
mat2 <- mat2[,-c(13:18)]
mat2 <- data.matrix(mat2)
hmcol <- colorRampPalette(brewer.pal(11, "RdBu"))(50)
df <- as.data.frame(colData(dds)[,c("genotype", "condition")])
```

```r
# scale = "row" converts to z-score scaling within rows
library(gplots)

heatmap.2(mat2, main="Heat map of overlapping genes", cexRow=0.6, labR
ow=NA,cexCol=0.8, col=hmcol, Rowv=T, Colv=T, labCol=c("Control", NA,NA
,NA,"   KOA",NA,NA,NA, "   KOB",NA,NA,NA), adjCol=c(-2.5,50),srtCol=0,
scale="row", dendrogram="column", trace="none", margin=c(1, 5), key.ti
tle="Counts density", key.xlab="Counts", key.ylab="")
```

9. Plot the top 50 ordered by the mean of PC1 & using the rldMatrix from WTvsKO study

```r
row.names(KOA_KOB_values_sortedPC) <- KOA_KOB_values_sortedPC$Geneid
topgenes <- head(rownames(KOA_KOB_values_sortedPC),50)
mat <- rldMatrix[topgenes,]
mat <- mat - rowMeans(mat)
mat2 <- data.frame(mat)
mat2$Geneid <- row.names(mat2)
mat2 <- join(mat2, gtf,by="Geneid")
row.names(mat2) <- mat2$GeneSymbol
mat2 <- mat2[,-c(13:18)]
mat2 <- data.matrix(mat2)

hmcol <- colorRampPalette(brewer.pal(11, "RdBu"))(50)
df <- as.data.frame(colData(dds)[,c("genotype", "condition")])

library(gplots)
heatmap.2(mat2, main="Top 50 DEGs", cexRow=0.6, cexCol=0.8, col=hmcol,
Rowv=T, Colv=F, labCol=c("Control", NA,NA,NA,"   KOA",NA,NA,NA,"   KOB
",NA,NA,NA), adjCol=c(-2,-85),srtCol=0, scale="row", dendrogram="none"
, trace="none", margin=c(1, 5), key.title="Counts density", key.xlab="
Counts", key.ylab="")
```

```r
# Alternative plotting method
# pheatmap(mat2, annotation_col=df, fontsize_row = 5, fontsize_col = 7
, cluster_cols = F,cluster_rows = F, main="Top 50 DEGs \n P<0.05 & abs
|Log2 Fold Change|>2")
```

10. Plot the expression of the top 50 genes on a bar plot - sorted by mean PC1 values

```r
library(tibble)
library(tidyr)
library(ggplot2)
library(tidyverse)

KOA_KOB_values_combined_sortedPC_tb <- KOA_KOB_values_sortedPC  %>%
  data.frame() %>%
  rownames_to_column(var="gene") %>%
  as_tibble()

top50_DEGs_genes <- KOA_KOB_values_combined_sortedPC_tb %>%
  arrange(-meanPC) %>%  #Arrange rows by padj values
  pull(gene) %>%        #Extract character vector of ordered genes
  head(n=50)
top50_DEG_genes <- merged_pcresults_sorted[top50_DEGs_genes,]
```

```r
top50_DEG_genes <- data.frame(top50_DEG_genes)
top50_DEG_genes$Geneid <- row.names(top50_DEG_genes)
top50_DEG_genes <- merge(top50_DEG_genes, gtf, by="Geneid")

#change the normalised counts into a tibble
normalized_counts_tb <- norm %>%
  data.frame() %>%
  rownames_to_column(var="gene") %>%
  as_tibble()

top50_DEG_norm <- normalized_counts_tb %>%
  filter(gene %in% top50_DEGs_genes)

gathered_top50_DEGs <- top50_DEG_norm %>%
  tidyr::gather(colnames(top50_DEG_norm)[2:13], key = "sample", value
= "normalized_counts")

gathered_top50_DEGs <- inner_join(gathered_top50_DEGs, sample_info, by
="sample")
names(gathered_top50_DEGs)[1] <- "Geneid"
gathered_top50_DEGs<- inner_join(gtf, gathered_top50_DEGs, by="Geneid"
)
gathered_top50_DEGs<- inner_join(merged_pcresults_sorted, gathered_top
50_DEGs, by="Geneid")
#write.table(gathered_top20_DEGs, "Differential Gene Expression/Top_50
_DEGs.csv", row.names=F, quote=F, sep=",", col.names=T)

#png("BarPlot_Top50_Overlapping_significant_genes.png", width=1200, he
ight=800)
ggplot(gathered_top50_DEGs) +
  geom_point(aes(x = GeneSymbol.x, y = normalized_counts, color = cond
ition)) +
  xlab("Genes") +
  ylab("Regularised Log Counts") +
  ggtitle("Top 50 Overlapping Significant Genes (P<0.05 & abs(Log2FC)
> 2)") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size=6)) +
  theme(plot.title = element_text(hjust = 0.5))

# dev.off()
```

8.1.9 Volcano Plot

Generating_a_volcano_Plot

Charli E. Harlow

11/02/2022

Create a function for a volcano plot

```r
VolcanoPlot <- function(toptable, NominalCutoff, AdjustedCutoff, Label
lingCutoff, FCCutoff, main)
```

```
{
    toptable$Significance <- "NS"
    toptable$Significance[(abs(toptable$log2FoldChange) > FCCutoff)] <- "FC"
    toptable$Significance[(toptable$padj<AdjustedCutoff)] <- "FDR"
    toptable$Significance[(toptable$padj<AdjustedCutoff) & (abs(toptable$log2FoldChange)>FCCutoff)] <- "FC_FDR"
    table(toptable$Significance)
    toptable$Significance <- factor(toptable$Significance, levels=c("NS", "FC", "FDR", "FC_FDR"))

    plot <- ggplot(toptable, aes(x=log2FoldChange, y=-log10(padj))) +

        #Add points:
        #   Colour based on factors set a few lines up
        #   'alpha' provides gradual shading of colour
        #   Set size of points
        geom_point(aes(color=factor(Significance)), alpha=1/2, size=0.8) +

        #Choose which colours to use; otherwise, ggplot2 choose automatically (order depends on how factors are ordered in toptable$Significance)
        scale_color_manual(values=c(NS="grey30", FC="forestgreen", FDR="royalblue", FC_FDR="red2"), labels=c(NS="NS", FC=paste("LogFC>|", FCCutoff, "|", sep=""), FDR=paste("FDR Q<", AdjustedCutoff, sep=""), FC_FDR=paste("FDR Q<", AdjustedCutoff, " & LogFC>|", FCCutoff, "|", sep=""))) +

        #Set the size of the plotting window
        theme_bw(base_size=24) +

        #Modify various aspects of the plot text and legend
        theme(legend.background=element_rect(),
          panel.border = element_blank(),
          axis.line = element_line(colour = "black", size=0.2),
          axis.ticks = element_line(size=0.2),
            plot.title=element_text(angle=0, size=12, face="bold", vjust=1),

            #panel.grid.major=element_blank(),  #Remove gridlines
            #panel.grid.minor=element_blank(),  #Remove gridlines

            axis.text.x=element_text(angle=0, size=12, vjust=1),
            axis.text.y=element_text(angle=0, size=12, vjust=1),
            axis.title=element_text(size=12),

            #Legend
            legend.position="top",          #Moves the legend to the top of the plot
            legend.key=element_blank(),     #removes the border
            legend.key.size=unit(0.5, "cm"),    #Sets overall area/size of the legend
            legend.text=element_text(size=8),   #Text size
            title=element_text(size=8),     #Title text size
            legend.title=element_blank()) +     #Remove the title
        #Change the size of the icons/symbols in the legend
        guides(colour = guide_legend(override.aes=list(size=2.5))) +

        #Set x- and y-axes labels
        xlab(bquote(~Log[2]~ "fold change")) +
        ylab(bquote(~-Log[10]~adjusted~italic(P))) +

        #Set the axis limits
        #xlim(-6.5, 6.5) +
        #ylim(0, 100) +

      # Scale the axis and set the limits
      #scale_x_continuous(breaks = seq(-15, 15, by = 5), lim=c(-15,15)) +
      #scale_y_continuous(breaks=seq(0,250,by=50), lim=c(0,250)) +

        #Set title
        ggtitle(main) +

        #Tidy the text labels for a subset of genes
        geom_text(data=subset(toptable, padj<LabellingCutoff & abs(log2FoldChange)>FCCutoff),
            aes(label=subset(toptable, padj<LabellingCutoff & abs(log2FoldChange)>FCCutoff)$genenames),
            size=2.25,
            #segment.color="black", #This and the next parameter spread out the labels and join them to their points by a line
            #segment.size=0.01,
            check_overlap=TRUE,
            vjust=1.0) +

        #Add a vertical line for fold change cut-offs
        geom_vline(xintercept=c(-FCCutoff, FCCutoff), linetype="longdash", colour="black", size=0.4) +

        #Add a horizontal line for P-value cut-off
        geom_hline(yintercept=-log10(NominalCutoff), linetype="longdash", colour="black", size=0.4)
}

}
```

Appendix 2: Gene ontology analysis

Table 8.1: Gene ontology (GO) analysis of the 3,501 DEGs identified through differential expression analysis of WT cells vs EPO knock-out.*Results passing P-value threshold of 0.05 are shown. Analysis was performed using the online tool Enrichr avaliable at https://maayanlab.cloud/Enrichr/enrich#. BP = Biological Process. MF = Molecular Function. KEGG = KEGG Pathways 2021.*

Ontology	GO Term	P-value
BP	DNA repair (GO:0006281)	1.45E-05
BP	double-strand break repair (GO:0006302)	2.57E-05
BP	mRNA 3'-end processing (GO:0031124)	8.87E-05
BP	DNA metabolic process (GO:0006259)	1.09E-04
BP	negative regulation of ubiquitin-protein transferase activity (GO:0051444)	2.61E-04
BP	negative regulation of intrinsic apoptotic signaling pathway (GO:2001243)	2.91E-04
BP	mRNA processing (GO:0006397)	5.06E-04
BP	mitochondrion organization (GO:0007005)	5.94E-04
BP	regulation of macroautophagy (GO:0016241)	1.10E-03
BP	positive regulation of protein binding (GO:0032092)	1.28E-03
BP	base-excision repair, gap-filling (GO:0006287)	1.33E-03
BP	fatty acid oxidation (GO:0019395)	1.63E-03
BP	regulation of aerobic respiration (GO:1903715)	1.66E-03
BP	positive regulation of fatty acid oxidation (GO:0046321)	1.67E-03
BP	pyrimidine nucleobase metabolic process (GO:0006206)	1.67E-03
BP	mitotic sister chromatid segregation (GO:0000070)	1.86E-03
BP	Notch signaling pathway (GO:0007219)	2.05E-03
BP	positive regulation of binding (GO:0051099)	2.24E-03
BP	ephrin receptor signaling pathway (GO:0048013)	2.51E-03

BP	negative regulation of macroautophagy (GO:0016242)	2.56E-03
BP	regulation of oxidative phosphorylation (GO:0002082)	2.82E-03
BP	regulation of protein binding (GO:0043393)	2.82E-03
BP	fatty acid catabolic process (GO:0009062)	2.92E-03
BP	nucleotide-excision repair (GO:0006289)	3.02E-03
BP	cellular response to leucine (GO:0071233)	3.04E-03
BP	response to leucine (GO:0043201)	3.04E-03
BP	regulation of retrograde protein transport, ER to cytosol (GO:1904152)	3.07E-03
BP	reticulophagy (GO:0061709)	3.07E-03
BP	cellular response to amino acid stimulus (GO:0071230)	3.16E-03
BP	organelle disassembly (GO:1903008)	3.16E-03
BP	recombinational repair (GO:0000725)	3.19E-03
BP	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile (GO:0000377)	3.22E-03
BP	mRNA 3'-end processing by stem-loop binding and cleavage (GO:0006398)	3.92E-03
BP	protein K69-linked ufmylation (GO:1990592)	3.92E-03
BP	protein polyufmylation (GO:1990564)	3.92E-03
BP	negative regulation of proteasomal ubiquitin-dependent protein catabolic process (GO:0032435)	4.01E-03
BP	negative regulation of organelle assembly (GO:1902116)	4.24E-03
BP	regulation of mitotic spindle organization (GO:0060236)	4.24E-03
BP	telomere organization (GO:0032200)	4.38E-03
BP	mRNA splicing, via spliceosome (GO:0000398)	4.43E-03
BP	telomere maintenance (GO:0000723)	5.19E-03
BP	regulation of fatty acid beta-oxidation (GO:0031998)	5.22E-03
BP	fatty acid beta-oxidation (GO:0006635)	5.38E-03
BP	RNA 3'-end processing (GO:0031123)	5.46E-03

BP	protein localization to microtubule cytoskeleton (GO:0072698)	5.56E-03
BP	pyrimidine nucleobase catabolic process (GO:0006208)	5.56E-03
BP	translocation of molecules into host (GO:0044417)	5.56E-03
BP	viral mRNA export from host cell nucleus (GO:0046784)	5.56E-03
BP	anterior/posterior axis specification (GO:0009948)	5.64E-03
BP	endosome transport via multivesicular body sorting pathway (GO:0032509)	6.42E-03
BP	positive regulation of macroautophagy (GO:0016239)	6.66E-03
BP	hydrogen peroxide metabolic process (GO:0042743)	6.91E-03
BP	negative regulation of epithelial to mesenchymal transition (GO:0010719)	6.91E-03
BP	regulation of early endosome to late endosome transport (GO:2000641)	6.92E-03
BP	regulation of membrane depolarization (GO:0003254)	6.92E-03
BP	regulation of spindle organization (GO:0090224)	6.92E-03
BP	cellular response to nutrient levels (GO:0031669)	6.95E-03
BP	macroautophagy (GO:0016236)	7.04E-03
BP	G2/M transition of mitotic cell cycle (GO:0000086)	7.69E-03
BP	negative regulation of endoplasmic reticulum stress-induced intrinsic apoptotic signaling pathway (GO:1902236)	8.34E-03
BP	RNA transport (GO:0050658)	8.62E-03
BP	cell cycle G2/M phase transition (GO:0044839)	8.69E-03
BP	regulation of DNA metabolic process (GO:0051052)	8.88E-03
BP	cellular respiration (GO:0045333)	9.15E-03
BP	microtubule polymerization (GO:0046785)	9.23E-03
BP	sister chromatid segregation (GO:0000819)	9.37E-03
BP	organelle organization (GO:0006996)	9.38E-03
BP	mitotic cell cycle phase transition (GO:0044772)	9.70E-03
BP	regulation of cellular response to stress (GO:0080135)	9.89E-03

BP	regulation of signal transduction by p53 class mediator (GO:1901796)	1.00E-02
BP	cellular response to iron ion (GO:0071281)	1.01E-02
BP	cerebral cortex cell migration (GO:0021795)	1.01E-02
BP	galactose catabolic process (GO:0019388)	1.01E-02
BP	negative regulation of glial cell proliferation (GO:0060253)	1.01E-02
BP	protein ufmylation (GO:0071569)	1.01E-02
BP	regulation of ventricular cardiac muscle cell membrane depolarization (GO:0060373)	1.01E-02
BP	signal complex assembly (GO:0007172)	1.01E-02
BP	telencephalon cell migration (GO:0022029)	1.01E-02
BP	diol metabolic process (GO:0034311)	1.02E-02
BP	telomere capping (GO:0016233)	1.02E-02
BP	amyloid-beta formation (GO:0034205)	1.07E-02
BP	basement membrane assembly (GO:0070831)	1.07E-02
BP	galactose metabolic process (GO:0006012)	1.07E-02
BP	negative regulation of retrograde protein transport, ER to cytosol (GO:1904153)	1.07E-02
BP	negative regulation of ubiquitin protein ligase activity (GO:1904667)	1.07E-02
BP	Notch receptor processing (GO:0007220)	1.07E-02
BP	positive regulation of fatty acid beta-oxidation (GO:0032000)	1.07E-02
BP	membrane protein ectodomain proteolysis (GO:0006509)	1.13E-02
BP	regulation of autophagy (GO:0010506)	1.19E-02
BP	mitotic chromosome condensation (GO:0007076)	1.19E-02
BP	regulation of ATP metabolic process (GO:1903578)	1.20E-02
BP	negative regulation of DNA binding (GO:0043392)	1.21E-02
BP	regulation of cellular senescence (GO:2000772)	1.21E-02
BP	nucleotide catabolic process (GO:0009166)	1.26E-02

BP	regulation of activin receptor signaling pathway (GO:0032925)	1.26E-02
BP	regulation of cell cycle process (GO:0010564)	1.27E-02
BP	mitotic nuclear division (GO:0140014)	1.28E-02
BP	double-strand break repair via homologous recombination (GO:0000724)	1.32E-02
BP	regulation of primary metabolic process (GO:0080090)	1.35E-02
BP	central nervous system development (GO:0007417)	1.40E-02
BP	cardiac ventricle morphogenesis (GO:0003208)	1.42E-02
BP	microtubule nucleation (GO:0007020)	1.44E-02
BP	mRNA export from nucleus (GO:0006406)	1.45E-02
BP	nucleotide-excision repair, DNA gap filling (GO:0006297)	1.53E-02
BP	ribonucleoprotein complex assembly (GO:0022618)	1.54E-02
BP	positive regulation of protein localization to cell surface (GO:2000010)	1.55E-02
BP	pteridine-containing compound metabolic process (GO:0042558)	1.55E-02
BP	response to amino acid (GO:0043200)	1.56E-02
BP	cellular response to acid chemical (GO:0071229)	1.57E-02
BP	negative regulation of oxidative stress-induced cell death (GO:1903202)	1.57E-02
BP	double-strand break repair via nonhomologous end joining (GO:0006303)	1.60E-02
BP	mRNA-containing ribonucleoprotein complex export from nucleus (GO:0071427)	1.72E-02
BP	regulation of epithelial to mesenchymal transition (GO:0010717)	1.73E-02
BP	organelle assembly (GO:0070925)	1.75E-02
BP	mRNA transport (GO:0051028)	1.78E-02
BP	positive regulation of autophagy (GO:0010508)	1.79E-02
BP	regulation of mRNA splicing, via spliceosome (GO:0048024)	1.79E-02
BP	nucleobase-containing compound catabolic process (GO:0034655)	1.83E-02
BP	DNA biosynthetic process (GO:0071897)	1.83E-02

BP	gliogenesis (GO:0042063)	1.83E-02
BP	cellular response to leucine starvation (GO:1990253)	1.84E-02
BP	histone H3-K36 demethylation (GO:0070544)	1.84E-02
BP	myelin maintenance (GO:0043217)	1.84E-02
BP	negative regulation of protein exit from endoplasmic reticulum (GO:0070862)	1.84E-02
BP	neuron projection maintenance (GO:1990535)	1.84E-02
BP	cell morphogenesis involved in differentiation (GO:0000904)	1.88E-02
BP	nucleotide-excision repair, DNA incision, 5'-to lesion (GO:0006296)	1.91E-02
BP	nucleotide-excision repair, DNA incision, 3'-to lesion (GO:0006295)	1.98E-02
BP	nucleotide-excision repair, preincision complex stabilization (GO:0006293)	1.98E-02
BP	actin filament bundle organization (GO:0061572)	1.98E-02
BP	mismatch repair (GO:0006298)	2.03E-02
BP	negative regulation of transferase activity (GO:0051348)	2.03E-02
BP	negative regulation of response to endoplasmic reticulum stress (GO:1903573)	2.03E-02
BP	regulation of spindle assembly (GO:0090169)	2.03E-02
BP	cellular response to prostaglandin E stimulus (GO:0071380)	2.04E-02
BP	detection of muscle stretch (GO:0035995)	2.04E-02
BP	hexose catabolic process (GO:0019320)	2.04E-02
BP	negative regulation of centriole replication (GO:0046600)	2.04E-02
BP	Notch receptor processing, ligand-dependent (GO:0035333)	2.04E-02
BP	nucleobase catabolic process (GO:0046113)	2.04E-02
BP	positive regulation of mitophagy in response to mitochondrial depolarization (GO:0098779)	2.04E-02
BP	protein localization to endoplasmic reticulum exit site (GO:0070973)	2.04E-02
BP	ribonucleoprotein complex disassembly (GO:0032988)	2.04E-02
BP	cellular protein-containing complex assembly (GO:0034622)	2.05E-02

BP	transcription by RNA polymerase II (GO:0006366)	2.17E-02
BP	regulation of G2/M transition of mitotic cell cycle (GO:0010389)	2.18E-02
BP	ubiquitin-dependent ERAD pathway (GO:0030433)	2.18E-02
BP	interstrand cross-link repair (GO:0036297)	2.19E-02
BP	negative regulation of cell cycle process (GO:0010948)	2.21E-02
BP	RNA export from nucleus (GO:0006405)	2.27E-02
BP	negative regulation of multicellular organismal process (GO:0051241)	2.29E-02
BP	protein modification process (GO:0036211)	2.31E-02
BP	regulation of lipid metabolic process (GO:0019216)	2.32E-02
BP	2-oxoglutarate metabolic process (GO:0006103)	2.32E-02
BP	cardiac left ventricle morphogenesis (GO:0003214)	2.32E-02
BP	dicarboxylic acid catabolic process (GO:0043649)	2.32E-02
BP	polarized epithelial cell differentiation (GO:0030859)	2.32E-02
BP	positive regulation of glucose metabolic process (GO:0010907)	2.32E-02
BP	respiratory electron transport chain (GO:0022904)	2.32E-02
BP	sodium-independent organic anion transport (GO:0043252)	2.32E-02
BP	negative regulation of G0 to G1 transition (GO:0070317)	2.36E-02
BP	negative regulation of proteasomal protein catabolic process (GO:1901799)	2.36E-02
BP	regulation of glucose metabolic process (GO:0010906)	2.48E-02
BP	termination of RNA polymerase II transcription (GO:0006369)	2.48E-02
BP	glycosphingolipid metabolic process (GO:0006687)	2.53E-02
BP	embryonic axis specification (GO:0000578)	2.55E-02
BP	regulation of chromosome segregation (GO:0051983)	2.55E-02
BP	negative regulation of metabolic process (GO:0009892)	2.59E-02
BP	DNA geometric change (GO:0032392)	2.64E-02

BP	positive regulation of microtubule polymerization or depolymerization (GO:0031112)	2.64E-02
BP	regulation of cellular response to heat (GO:1900034)	2.64E-02
BP	positive regulation of peptidase activity (GO:0010952)	2.64E-02
BP	negative regulation of cell differentiation (GO:0045596)	2.71E-02
BP	negative regulation of protein-containing complex assembly (GO:0031333)	2.80E-02
BP	membrane protein proteolysis (GO:0033619)	2.89E-02
BP	nucleotide-excision repair, DNA incision (GO:0033683)	2.89E-02
BP	protein acylation (GO:0043543)	2.89E-02
BP	protein monoubiquitination (GO:0006513)	2.89E-02
BP	chromatin-mediated maintenance of transcription (GO:0048096)	2.90E-02
BP	glutathione biosynthetic process (GO:0006750)	2.90E-02
BP	hematopoietic stem cell proliferation (GO:0071425)	2.90E-02
BP	insulin-like growth factor receptor signaling pathway (GO:0048009)	2.90E-02
BP	negative regulation of centrosome duplication (GO:0010826)	2.90E-02
BP	negative regulation of defense response to virus (GO:0050687)	2.90E-02
BP	negative regulation of protein maturation (GO:1903318)	2.90E-02
BP	positive regulation of actin nucleation (GO:0051127)	2.90E-02
BP	positive regulation of cellular respiration (GO:1901857)	2.90E-02
BP	proteasome assembly (GO:0043248)	2.90E-02
BP	tetrahydrobiopterin metabolic process (GO:0046146)	2.90E-02
BP	V(D)J recombination (GO:0033151)	2.90E-02
BP	regulation of fat cell differentiation (GO:0045598)	3.01E-02
BP	base-excision repair (GO:0006284)	3.02E-02
BP	cellular response to interleukin-12 (GO:0071349)	3.02E-02
BP	substrate adhesion-dependent cell spreading (GO:0034446)	3.02E-02

BP	dicarboxylic acid metabolic process (GO:0043648)	3.04E-02
BP	positive regulation of cell cycle G1/S phase transition (GO:1902808)	3.07E-02
BP	myeloid leukocyte differentiation (GO:0002573)	3.15E-02
BP	positive regulation of cellular protein localization (GO:1903829)	3.15E-02
BP	mitochondrial respiratory chain complex assembly (GO:0033108)	3.23E-02
BP	fatty acid beta-oxidation using acyl-CoA oxidase (GO:0033540)	3.31E-02
BP	ganglioside biosynthetic process (GO:0001574)	3.31E-02
BP	glucan catabolic process (GO:0009251)	3.31E-02
BP	glycogen catabolic process (GO:0005980)	3.31E-02
BP	histone H4-K5 acetylation (GO:0043981)	3.31E-02
BP	histone H4-K8 acetylation (GO:0043982)	3.31E-02
BP	negative regulation of androgen receptor signaling pathway (GO:0060766)	3.31E-02
BP	negative regulation of protein processing (GO:0010955)	3.31E-02
BP	nitric oxide biosynthetic process (GO:0006809)	3.31E-02
BP	placenta development (GO:0001890)	3.31E-02
BP	calcium-ion regulated exocytosis (GO:0017156)	3.37E-02
BP	peroxisomal membrane transport (GO:0015919)	3.37E-02
BP	vascular endothelial growth factor receptor signaling pathway (GO:0048010)	3.44E-02
BP	negative regulation of extrinsic apoptotic signaling pathway in absence of ligand (GO:2001240)	3.45E-02
BP	negative regulation of signal transduction in absence of ligand (GO:1901099)	3.45E-02
BP	negative regulation of smoothened signaling pathway (GO:0045879)	3.45E-02
BP	ATP synthesis coupled proton transport (GO:0015986)	3.45E-02
BP	carbohydrate derivative catabolic process (GO:1901136)	3.45E-02
BP	N-glycan processing (GO:0006491)	3.45E-02
BP	negative regulation of oxidative stress-induced intrinsic apoptotic signaling pathway (GO:1902176)	3.45E-02

BP	RNA processing (GO:0006396)	3.47E-02
BP	bone development (GO:0060348)	3.50E-02
BP	glycoside catabolic process (GO:0016139)	3.53E-02
BP	internal protein amino acid acetylation (GO:0006475)	3.53E-02
BP	negative regulation of cAMP-dependent protein kinase activity (GO:2000480)	3.53E-02
BP	negative regulation of neural precursor cell proliferation (GO:2000178)	3.53E-02
BP	negative regulation of response to reactive oxygen species (GO:1901032)	3.53E-02
BP	positive regulation of Arp2/3 complex-mediated actin nucleation (GO:2000601)	3.53E-02
BP	positive regulation of early endosome to late endosome transport (GO:2000643)	3.53E-02
BP	protein localization to microtubule (GO:0035372)	3.53E-02
BP	protein retention in ER lumen (GO:0006621)	3.53E-02
BP	regulation of acetyl-CoA biosynthetic process from pyruvate (GO:0010510)	3.53E-02
BP	regulation of acyl-CoA biosynthetic process (GO:0050812)	3.53E-02
BP	telomeric D-loop disassembly (GO:0061820)	3.53E-02
BP	eye development (GO:0001654)	3.54E-02
BP	positive regulation of mitotic cell cycle phase transition (GO:1901992)	3.54E-02
BP	adherens junction organization (GO:0034332)	3.57E-02
BP	inner mitochondrial membrane organization (GO:0007007)	3.57E-02
BP	response to hydrogen peroxide (GO:0042542)	3.57E-02
BP	negative regulation of fat cell differentiation (GO:0045599)	3.74E-02
BP	regulation of protein import into nucleus (GO:0042306)	3.74E-02
BP	telomere maintenance via telomere lengthening (GO:0010833)	3.74E-02
BP	regulation of DNA repair (GO:0006282)	3.87E-02
BP	cell-cell adhesion mediated by integrin (GO:0033631)	3.97E-02
BP	centriole assembly (GO:0098534)	3.97E-02

BP	DNA protection (GO:0042262)	3.97E-02
BP	dorsal/ventral axis specification (GO:0009950)	3.97E-02
BP	establishment of protein localization to telomere (GO:0070200)	3.97E-02
BP	gamma-aminobutyric acid metabolic process (GO:0009448)	3.97E-02
BP	hematopoietic stem cell migration (GO:0035701)	3.97E-02
BP	immature B cell differentiation (GO:0002327)	3.97E-02
BP	mannose trimming involved in glycoprotein ERAD pathway (GO:1904382)	3.97E-02
BP	positive regulation of protein localization to cell cortex (GO:1904778)	3.97E-02
BP	prostaglandin transport (GO:0015732)	3.97E-02
BP	protein deglycosylation involved in glycoprotein catabolic process (GO:0035977)	3.97E-02
BP	regulation of metaphase plate congression (GO:0090235)	3.97E-02
BP	regulation of receptor catabolic process (GO:2000644)	3.97E-02
BP	spliceosomal conformational changes to generate catalytic conformation (GO:0000393)	3.97E-02
BP	stress granule disassembly (GO:0035617)	3.97E-02
BP	succinyl-CoA metabolic process (GO:0006104)	3.97E-02
BP	trophoblast giant cell differentiation (GO:0060707)	3.97E-02
BP	ubiquitin-dependent glycoprotein ERAD pathway (GO:0097466)	3.97E-02
BP	ERAD pathway (GO:0036503)	3.98E-02
BP	cellular response to topologically incorrect protein (GO:0035967)	3.99E-02
BP	negative regulation of autophagy (GO:0010507)	4.10E-02
BP	carboxylic acid transmembrane transport (GO:1905039)	4.19E-02
BP	regulation of G0 to G1 transition (GO:0070316)	4.19E-02
BP	regulation of mitochondrial membrane potential (GO:0051881)	4.19E-02
BP	cristae formation (GO:0042407)	4.22E-02
BP	positive regulation of microtubule polymerization (GO:0031116)	4.22E-02

BP	positive regulation of protein import into nucleus (GO:0042307)	4.22E-02
BP	ventricular septum morphogenesis (GO:0060412)	4.22E-02
BP	positive regulation of cellular protein metabolic process (GO:0032270)	4.23E-02
BP	regulation of proteasomal ubiquitin-dependent protein catabolic process (GO:0032434)	4.23E-02
BP	actin crosslink formation (GO:0051764)	4.27E-02
BP	energy coupled proton transport, down electrochemical gradient (GO:0015985)	4.27E-02
BP	morphogenesis of a polarized epithelium (GO:0001738)	4.27E-02
BP	nonribosomal peptide biosynthetic process (GO:0019184)	4.27E-02
BP	protein import into peroxisome matrix (GO:0016558)	4.27E-02
BP	regulation of glycogen metabolic process (GO:0070873)	4.27E-02
BP	cilium organization (GO:0044782)	4.43E-02
BP	actin filament organization (GO:0007015)	4.48E-02
BP	basement membrane organization (GO:0071711)	4.53E-02
BP	cellular polysaccharide catabolic process (GO:0044247)	4.53E-02
BP	glycosphingolipid biosynthetic process (GO:0006688)	4.53E-02
BP	histone H2A acetylation (GO:0043968)	4.53E-02
BP	late endosome to vacuole transport via multivesicular body sorting pathway (GO:0032511)	4.53E-02
BP	negative regulation of transcription regulatory region DNA binding (GO:2000678)	4.53E-02
BP	nitric oxide metabolic process (GO:0046209)	4.53E-02
BP	positive regulation by host of viral transcription (GO:0043923)	4.53E-02
BP	positive regulation of small molecule metabolic process (GO:0062013)	4.53E-02
BP	protein localization to endoplasmic reticulum (GO:0070972)	4.53E-02
BP	regulation of ATP biosynthetic process (GO:2001169)	4.53E-02
BP	regulation of interleukin-1 production (GO:0032652)	4.53E-02
BP	regulation of intrinsic apoptotic signaling pathway in response to DNA damage (GO:1902229)	4.53E-02

BP	establishment of epithelial cell polarity (GO:0090162)	4.54E-02
BP	ganglioside metabolic process (GO:0001573)	4.54E-02
BP	miRNA metabolic process (GO:0010586)	4.54E-02
BP	nucleobase-containing compound biosynthetic process (GO:0034654)	4.54E-02
BP	interleukin-12-mediated signaling pathway (GO:0035722)	4.57E-02
BP	positive regulation of nucleocytoplasmic transport (GO:0046824)	4.57E-02
BP	positive regulation of organelle assembly (GO:1902117)	4.57E-02
BP	ribonucleoprotein complex biogenesis (GO:0022613)	4.72E-02
BP	actin filament bundle assembly (GO:0051017)	4.86E-02
BP	regulation of protein localization to cell surface (GO:2000008)	4.86E-02
BP	DNA duplex unwinding (GO:0032508)	4.89E-02
BP	regulation of cellular catabolic process (GO:0031329)	4.92E-02
BP	translesion synthesis (GO:0019985)	4.97E-02
BP	DNA-templated transcription, termination (GO:0006353)	5.10E-02
BP	protein stabilization (GO:0050821)	5.15E-02
MF	ATPase regulator activity (GO:0060590)	3.09E-04
MF	nuclear receptor coactivator activity (GO:0030374)	3.39E-04
MF	ATPase activator activity (GO:0001671)	8.38E-04
MF	RNA binding (GO:0003723)	9.92E-04
MF	antiporter activity (GO:0015297)	3.80E-03
MF	damaged DNA binding (GO:0003684)	5.62E-03
MF	ubiquitin-like protein ligase binding (GO:0044389)	5.77E-03
MF	telomeric DNA binding (GO:0042162)	7.18E-03
MF	ubiquitin protein ligase binding (GO:0031625)	7.33E-03
MF	protein serine/threonine kinase inhibitor activity (GO:0030291)	9.23E-03

MF	exodeoxyribonuclease activity, producing 5'-phosphomonoesters (GO:0016895)	9.75E-03
MF	ATPase binding (GO:0051117)	1.09E-02
MF	protein phosphatase regulator activity (GO:0019888)	1.44E-02
MF	cadherin binding (GO:0045296)	1.76E-02
MF	cAMP-dependent protein kinase regulator activity (GO:0008603)	1.84E-02
MF	endopeptidase activator activity (GO:0061133)	1.84E-02
MF	histone demethylase activity (H3-K36 specific) (GO:0051864)	1.84E-02
MF	mannosyl-oligosaccharide 1,2-alpha-mannosidase activity (GO:0004571)	2.04E-02
MF	mannosyl-oligosaccharide mannosidase activity (GO:0015924)	2.04E-02
MF	DNA-binding transcription factor binding (GO:0140297)	2.11E-02
MF	vitamin D receptor binding (GO:0042809)	2.32E-02
MF	secondary active transmembrane transporter activity (GO:0015291)	2.59E-02
MF	peptidase activator activity (GO:0016504)	2.64E-02
MF	protein kinase inhibitor activity (GO:0004860)	2.71E-02
MF	lysophosphatidic acid acyltransferase activity (GO:0042171)	3.31E-02
MF	Wnt-activated receptor activity (GO:0042813)	3.31E-02
MF	cAMP-dependent protein kinase inhibitor activity (GO:0004862)	3.53E-02
MF	disulfide oxidoreductase activity (GO:0015036)	3.53E-02
MF	ubiquitin ligase inhibitor activity (GO:1990948)	3.53E-02
MF	oxidoreduction-driven active transmembrane transporter activity (GO:0015453)	3.54E-02
MF	nuclear receptor binding (GO:0016922)	3.63E-02
MF	NADP binding (GO:0050661)	3.74E-02
MF	C-acetyltransferase activity (GO:0016453)	3.97E-02
MF	D-loop DNA binding (GO:0062037)	3.97E-02
MF	glutathione transmembrane transporter activity (GO:0034634)	3.97E-02

MF	L-leucine transmembrane transporter activity (GO:0015190)	3.97E-02
MF	phosphatidylserine flippase activity (GO:0140346)	3.97E-02
MF	syndecan binding (GO:0045545)	3.97E-02
MF	thyroid hormone receptor binding (GO:0046966)	4.22E-02
MF	hydro-lyase activity (GO:0016836)	4.48E-02
MF	core promoter sequence-specific DNA binding (GO:0001046)	4.52E-02
MF	CoA hydrolase activity (GO:0016289)	4.53E-02
MF	2-oxoglutarate-dependent dioxygenase activity (GO:0016706)	4.86E-02
MF	active ion transmembrane transporter activity (GO:0022853)	4.97E-02
MF	signal sequence binding (GO:0005048)	4.97E-02
KEGG	Propanoate metabolism	2.77E-04
KEGG	Longevity regulating pathway	4.44E-04
KEGG	AMPK signaling pathway	1.09E-03
KEGG	Peroxisome	2.76E-03
KEGG	Notch signaling pathway	4.38E-03
KEGG	Protein processing in endoplasmic reticulum	7.26E-03
KEGG	Thyroid hormone signaling pathway	8.86E-03
KEGG	Thermogenesis	9.65E-03
KEGG	Pentose phosphate pathway	9.72E-03
KEGG	Lysine degradation	2.01E-02
KEGG	Base excision repair	2.08E-02
KEGG	Small cell lung cancer	2.50E-02
KEGG	Non-alcoholic fatty liver disease	2.66E-02
KEGG	beta-Alanine metabolism	2.69E-02
KEGG	Fatty acid degradation	2.85E-02

KEGG	Glutathione metabolism	3.22E-02
KEGG	Oxidative phosphorylation	3.36E-02
KEGG	Mismatch repair	3.58E-02
KEGG	Purine metabolism	3.65E-02
KEGG	DNA replication	3.91E-02
KEGG	Cysteine and methionine metabolism	4.41E-02
KEGG	N-Glycan biosynthesis	4.41E-02
KEGG	Amoebiasis	4.54E-02
KEGG	mRNA surveillance pathway	4.96E-02

Appendix 3: Selection of 515 Hgb-associated SNPs

Table 8.2: Association statistics for the 515 conditionally independent Hgb-associated variants obtained from Vuckovic et al. (2020) for the analysis performed in Chapter 6.

RSID	Chr	Pos	Effect_allele	Ref_allele	EAF	Beta	SE	Pvalue
rs147804508	19	50035161	T	C	0.010576	0.0655968	0.0109318	1.96652E-09
rs72681869	14	50655357	C	G	0.010983	0.0650354	0.0102392	2.13074E-10
rs181000569	15	67166007	A	C	0.011052	0.0715299	0.0102425	2.87635E-12
rs17850433	21	45746102	T	C	0.987992	0.0619513	0.00977557	2.33736E-10
rs12925612	16	88777073	G	C	0.987987	0.0801445	0.0101211	2.40284E-15
rs114948639	2	46293826	C	T	0.987824	0.122843	0.00992527	3.49017E-35
rs532782761	3	172276502	A	C	0.985836	0.0713986	0.0104224	7.35948E-12
rs115902543	6	25712496	A	C	0.014599	0.0551956	0.00897074	7.61037E-10
rs61750953	19	41306650	C	T	0.984602	0.0955095	0.00865909	2.73957E-28
rs61744929	11	2325427	C	T	0.019006	0.0497109	0.00779756	1.8274E-10
rs192191487	19	41305065	A	G	0.019053	0.0769008	0.00882109	2.83574E-18
rs576750965	19	51003387	C	CAGAG	0.019968	0.0589923	0.00801072	1.78253E-13
rs61835223	1	231562228	G	A	0.02042	0.119324	0.00761638	2.55263E-55
rs77542162	17	67081278	G	A	0.022711	0.0905271	0.00716194	1.27014E-36
rs182552845	7	100306920	G	A	0.976915	0.0480735	0.00737827	7.24229E-11
rs72638978	8	41540828	C	T	0.023161	0.0533022	0.00716916	1.04624E-13
rs184088518	19	41305138	G	T	0.976304	0.118949	0.00727042	3.65041E-60
rs150844304	15	43726625	A	C	0.974707	0.0951662	0.00679515	1.45242E-44
rs74960997	6	163941243	T	G	0.973675	0.0429285	0.00667419	1.25922E-10
rs41278174	16	16259596	A	G	0.02713	0.0524204	0.00655803	1.31358E-15

rs113405739	1	154976518	A	G	0.971172	0.0717367	0.0063737	2.18551E-29
rs142456437	X	129970818	G	C	0.970671	0.0313754	0.00541419	6.83085E-09
rs41295942	7	100218631	T	C	0.030798	0.0462592	0.00615717	5.77602E-14
rs1800961	20	43042364	T	C	0.031089	0.0725911	0.0061426	3.1641E-32
rs9403391	6	142814991	C	T	0.968508	0.0382107	0.00611088	4.02909E-10
rs78415359	18	46207268	G	A	0.968132	0.0488166	0.00606821	8.64952E-16
rs186222325	4	79344554	T	C	0.032868	0.039012	0.00624401	4.15987E-10
rs139273519	12	11803303	G	C	0.961292	0.0359866	0.00555751	9.46175E-11
rs775701197	11	30749171	GTTCCCGCTCTGTAACTTATCAAC	G	0.959211	0.0712129	0.00551689	4.04859E-38
rs11844732	14	24808961	G	T	0.045221	0.0301607	0.00514171	4.46733E-09
rs116971887	16	51170026	G	T	0.954588	0.0375855	0.00518815	4.3407E-13
rs200299716	6	34168059	TA	T	0.046416	0.0365519	0.00507608	5.98548E-13
rs28930677	2	120848049	T	C	0.046661	0.0426472	0.00503343	2.39582E-17
rs16874060	4	23737865	G	A	0.047264	0.0481084	0.00501853	9.143E-22
rs17287978	6	43941137	T	C	0.951909	0.0686304	0.00497296	2.5234E-43
rs34952318	20	11177055	A	G	0.049108	0.0347781	0.00500048	3.52681E-12
rs10410950	19	13128861	G	A	0.950001	0.0300262	0.00490498	9.26552E-10
rs4008347	9	6370088	G	A	0.9478262	0.0400915	0.00478593	5.43166E-17
rs113292219	1	29486128	T	C	0.053762	0.0339747	0.00470016	4.88669E-13
rs34939651	13	110424933	AG	A	0.055329	0.0301039	0.00475521	2.44021E-10
rs60925406	5	52221748	T	C	0.941883	0.0302675	0.00455486	3.03011E-11
rs117900708	15	41190639	T	G	0.939529	0.028021	0.00452146	5.74274E-10
rs12593543	15	95696202	T	C	0.060812	0.0262343	0.00446898	4.3498E-09
rs117793618	11	10279241	A	G	0.938708	0.0605039	0.00446641	8.31808E-42
rs35765401	5	111357751	A	AC	0.937806	0.0323905	0.00453668	9.354E-13

rs8176749	9	136131188	T	C	0.062637	0.0804246	0.00439641	9.36714E-75
rs12881869	14	50923249	C	T	0.929127	0.0298012	0.00416013	7.86279E-13
rs2849017	6	32174048	A	G	0.928676	0.0392043	0.00413104	2.30643E-21
rs4925184	20	61037490	A	G	0.9279162	0.0334781	0.00415373	7.64394E-16
rs3765000	2	44014301	G	A	0.073705	0.0276225	0.00406313	1.05839E-11
rs13107325	4	103188709	T	C	0.074589	0.050046	0.00406064	6.67258E-35
rs58542926	19	19379549	T	C	0.074789	0.0317177	0.00405435	5.15266E-15
rs1265075	6	31113113	A	C	0.07622	0.0353065	0.00401163	1.35559E-18
rs35994626	2	208632817	G	A	0.923751	0.0269083	0.00399096	1.55885E-11
rs17185657	6	29821606	T	A	0.076717	0.0282621	0.00399833	1.56644E-12
rs1800562	6	26093141	A	G	0.076922	0.189783	0.00400401	0
rs78744187	19	33754548	T	C	0.081199	0.0391557	0.00389713	9.44028E-24
rs34651	5	72144005	T	C	0.9185931	0.033691	0.0039204	8.41644E-18
rs73920681	2	25578588	G	C	0.082393	0.0317612	0.00386234	1.9797E-16
rs12481910	21	38168446	T	C	0.915062	0.024986	0.00382577	6.53438E-11
rs34933611	18	12546603	G	C	0.912582	0.0245398	0.00380187	1.08458E-10
rs74414571	7	25420102	G	C	0.912166	0.0273793	0.00378889	4.96674E-13
rs113670117	14	36099366	C	T	0.911796	0.03085	0.00375552	2.12917E-16
rs731749	16	86356582	G	A	0.911476	0.0392431	0.00382712	1.13603E-24
rs1029238	6	30138645	A	G	0.089377	0.0264675	0.00373091	1.30183E-12
rs4240624	8	9184231	G	A	0.0913002	0.0314424	0.00370464	2.11489E-17
rs75224897	12	48496565	C	G	0.092432	0.027608	0.00372617	1.27048E-13
rs56011044	1	48022107	C	T	0.903708	0.0326889	0.00363767	2.55881E-19
rs2612585	18	43085920	G	A	0.9019225	0.0254973	0.00359142	1.25206E-12
rs2228445	1	203667409	T	C	0.0985711	0.0414811	0.00356899	3.16197E-31

rs9816588	3	37987856	A	C	0.098761	0.0357016	0.00357051	1.53913E-23
rs56262900	2	23897725	A	G	0.101602	0.03021	0.00352309	9.92083E-18
rs2229742	21	16339172	G	C	0.896758	0.0260766	0.00350857	1.06759E-13
rs112391082	8	48864284	GGAAA	G	0.103347	0.0226491	0.00350313	1.01038E-10
rs7194649	16	215106	C	A	0.894445	0.033148	0.00356511	1.43261E-20
rs137893155	3	132198815	G	GTTATT	0.894068	0.0255875	0.00350871	3.04106E-13
rs6967414	7	6749758	A	G	0.107017	0.0243489	0.00345115	1.72228E-12
rs17476364	10	71094504	C	T	0.108778	0.149026	0.00341991	0
rs138780918	4	124765262	AAC	A	0.890645	0.022192	0.00341822	8.45537E-11
rs146541367	7	150502468	AT	A	0.889231	0.0277274	0.00340079	3.54351E-16
rs371199618	1	31594199	CAAA	C	0.111684	0.0232723	0.00349061	2.6088E-11
rs61591132	7	150952770	G	A	0.887939	0.0214482	0.00338584	2.3784E-10
rs3847858	12	52318378	T	A	0.113253	0.0222474	0.00336646	3.88117E-11
rs2854528	17	42338248	G	A	0.114247	0.0338113	0.0033507	6.06626E-24
rs12631955	3	66881402	C	G	0.11435	0.0231182	0.00334853	5.05637E-12
rs833805	6	44030011	G	A	0.884699	0.0680859	0.00349099	1.02879E-84
rs55781197	16	67940350	G	A	0.115344	0.0364845	0.00333429	7.24132E-28
rs34130368	10	48411796	G	T	0.884606	0.0258451	0.0033333	8.93181E-15
rs55761633	1	20757820	T	C	0.883547	0.0208652	0.00330339	2.67901E-10
rs55971447	1	161515326	T	C	0.117695	0.0336366	0.00328445	1.29663E-24
rs1354674	5	40623128	C	T	0.117765	0.0237929	0.00332321	8.09047E-13
rs190379045	3	41882697	A	G	0.118236	0.0239721	0.00348204	5.79914E-12
rs12607403	18	46343221	C	T	0.120048	0.0245215	0.00329892	1.06009E-13
rs881144	22	37471290	G	A	0.877321	0.0736692	0.00324719	6.0122E-114
rs12129889	1	79452106	T	C	0.122839	0.0189943	0.00325564	5.40249E-09

rs11543269	18	55313997	C	T	0.876133	0.0234176	0.00324582	5.40552E-13
rs12987661	2	69813458	C	T	0.12528	0.0207983	0.00320821	9.00142E-11
rs6724315	2	46363699	C	T	0.127071	0.0495429	0.00320213	5.37337E-54
rs2928166	5	76479833	C	T	0.128202	0.0276397	0.0031992	5.64269E-18
rs140294987	2	112247747	T	A	0.871148	0.0227731	0.00327299	3.4541E-12
rs11098259	4	115519825	A	G	0.870571	0.0252854	0.00319409	2.44667E-15
rs357282	5	38869035	G	T	0.869474	0.0252743	0.00317019	1.55512E-15
rs13299342	9	136141504	A	G	0.130949	0.0256306	0.00315471	4.49024E-16
rs34914463	17	7366619	T	C	0.868867	0.0285732	0.00315511	1.35116E-19
rs12412959	10	82141337	G	A	0.865923	0.0189318	0.00312769	1.42203E-09
rs11784090	8	23377161	T	C	0.135373	0.0201851	0.00311393	9.0398E-11
rs12119893	1	10483167	A	G	0.135855	0.0214433	0.00325438	4.42625E-11
rs145606348	6	43230103	GGT	G	0.863806	0.0249099	0.00310479	1.03147E-15
rs5955867	X	19959144	C	G	0.137744	0.0181221	0.00267563	1.2612E-11
rs13169	19	808586	G	C	0.138608	0.0358235	0.00307835	2.66498E-31
rs11694902	2	121988884	A	G	0.139399	0.0305995	0.0030635	1.71304E-23
rs556412194	5	154027482	G	GT	0.140487	0.0280995	0.00309732	1.1664E-19
rs76772442	14	65506975	G	A	0.141515	0.0265993	0.00307671	5.36307E-18
rs8133974	21	36450841	T	C	0.143791	0.0233546	0.00303658	1.45891E-14
rs3212018	7	80303700	A	AGCACAAATAAAGCACT	0.14466	0.0208563	0.00317418	5.01094E-11
rs541907913	20	25294041	A	G	0.85531	0.0199732	0.00310329	1.22552E-10
rs2226683	21	39868927	C	T	0.852588	0.0210697	0.00301876	2.96048E-12
rs10981834	9	116349209	G	C	0.851781	0.0280996	0.00299538	6.53623E-21
rs198851	6	26104632	T	G	0.150386	0.108478	0.00297238	1.315E-291
rs12902050	15	51085374	C	T	0.150513	0.0187267	0.00298795	3.67109E-10

rs67037045	12	13050341	ATGAGAAGACC	A	0.848592	0.0318423	0.00297806	1.10598E-26
rs4933432	10	88925529	A	T	0.848427	0.0216818	0.00297154	2.95381E-13
rs1831977	10	5254821	G	C	0.153679	0.0205618	0.00295346	3.35628E-12
rs2834317	21	35356706	G	A	0.846038	0.0372133	0.00297838	8.00209E-36
rs17654742	1	217472265	A	G	0.154143	0.0295032	0.0029617	2.24475E-23
rs111959637	19	33231028	C	G	0.844121	0.0287451	0.00307842	9.85287E-21
rs1086605	1	147287992	T	A	0.156038	0.0263508	0.00292482	2.07234E-19
rs181361	22	21929566	A	T	0.843701	0.0210612	0.00312157	1.50944E-11
rs499974	11	75455021	A	C	0.156567	0.0218025	0.00292873	9.74207E-14
rs62401198	6	43801654	C	T	0.842739	0.0358666	0.00292066	1.1556E-34
rs73047068	19	41297106	G	C	0.83751	0.0183349	0.00289062	2.25487E-10
rs1547651	6	43730644	T	A	0.16472	0.0335167	0.00288668	3.63149E-31
rs12952262	17	46683800	C	T	0.835232	0.0235451	0.00287749	2.77984E-16
rs12889267	14	21542766	A	G	0.832307	0.0242016	0.00285177	2.12919E-17
rs34295474	11	1689409	G	A	0.831521	0.0176224	0.00288172	9.64137E-10
rs9895661	17	59456589	C	T	0.170917	0.0372736	0.00283328	1.57949E-39
rs4837197	9	130622946	T	C	0.171428	0.0384714	0.00282615	3.36791E-42
rs2278243	19	41132022	G	C	0.17315	0.0242426	0.00282245	8.75736E-18
rs34920465	1	22700351	G	A	0.176488	0.0240535	0.00278197	5.32376E-18
rs202049562	2	27779565	A	T	0.178938	0.0242351	0.00284287	1.5298E-17
rs17816693	15	33324280	T	C	0.179382	0.0211103	0.00278035	3.13363E-14
rs515135	2	21286057	T	C	0.180074	0.0170724	0.00276101	6.27397E-10
rs144579321	6	36555803	A	AT	0.819339	0.0182494	0.00278732	5.85894E-11
rs112982980	6	125622737	G	T	0.181257	0.0232425	0.00276746	4.52236E-17
rs9302635	16	72144174	T	C	0.817908	0.0238295	0.00276241	6.33441E-18

rs9327454	5	127249136	A	G	0.817194	0.0200071	0.00277196	5.28848E-13
rs17799476	2	46352326	C	G	0.816452	0.0243846	0.00275006	7.51876E-19
rs8080123	17	59242914	G	T	0.816333	0.0281078	0.002757	2.086E-24
rs2519093	9	136141870	C	T	0.815601	0.0849069	0.00275241	5.9214E-209
rs3859158	16	4676660	A	G	0.815044	0.0184872	0.00274844	1.73872E-11
rs5762813	22	29203314	C	T	0.814784	0.0281293	0.0027426	1.10718E-24
rs4760682	12	48512285	C	A	0.186889	0.0476284	0.00272872	3.18239E-68
rs12352863	9	33116941	G	C	0.812599	0.0177532	0.00274097	9.35784E-11
rs17413015	1	161644811	C	T	0.811738	0.0221227	0.00275961	1.08709E-15
rs8030097	15	66943404	G	C	0.809849	0.0232918	0.00272778	1.35695E-17
rs1399562	8	76493714	A	T	0.807618	0.0185028	0.0027067	8.14772E-12
rs77303550	16	72079657	C	T	0.807502	0.0197289	0.00271001	3.33795E-13
rs2905582	5	137000719	C	T	0.192613	0.0183555	0.0027169	1.41809E-11
rs8090126	18	42811727	C	T	0.193674	0.0203454	0.00269995	4.86568E-14
rs6690625	1	66077590	G	T	0.195771	0.0190782	0.00267574	1.00322E-12
rs34798274	15	63359093	CT	C	0.197262	0.0178608	0.00268889	3.08525E-11
rs10770829	12	21710749	C	T	0.801084	0.0184368	0.00266642	4.69717E-12
rs11030099	11	27677583	A	C	0.198944	0.0167495	0.00267826	4.00395E-10
rs28718978	15	76123209	T	G	0.800378	0.0295666	0.00271921	1.54625E-27
rs7532370	1	217034049	T	C	0.200004	0.020582	0.00265166	8.36464E-15
rs2424	2	42285575	A	T	0.201329	0.0167294	0.00265362	2.8937E-10
rs531302321	16	204126	G	A	0.202266	0.0180474	0.00286642	3.05136E-10
rs7482510	11	2190591	G	C	0.203723	0.0185932	0.00269483	5.21559E-12
rs4953256	2	46005646	G	C	0.204974	0.0174673	0.00264122	3.75731E-11
rs139113145	9	107761370	T	TA	0.205533	0.0173228	0.00268533	1.11187E-10

rs2255293	2	12104445	T	C	0.209067	0.0165568	0.00262204	2.71084E-10
rs752590	2	113972945	G	A	0.210488	0.0314183	0.00260331	1.5475E-33
rs17685306	9	20551460	C	G	0.789458	0.0163305	0.0026145	4.20754E-10
rs117469893	16	52581047	C	G	0.789394	0.0160182	0.00261814	9.46619E-10
rs2306050	16	88800295	T	C	0.211481	0.0246514	0.00263452	8.1975E-21
rs10592192	12	4330068	TTC	T	0.787603	0.0218724	0.00261232	5.62711E-17
rs35240997	3	12379351	A	G	0.786172	0.0313326	0.00259641	1.56517E-33
rs144903947	14	65892694	CA	C	0.78581	0.0189379	0.00263616	6.77523E-13
rs71152258	16	57344314	A	AT	0.78518	0.0167719	0.00278567	1.73581E-09
rs4456287	11	118298252	T	C	0.784682	0.0175607	0.0025918	1.2399E-11
rs10849962	12	112037526	G	A	0.784372	0.0601431	0.00260224	3.503E-118
rs13415550	2	144352423	T	C	0.215801	0.0153445	0.00260333	3.76543E-09
rs738409	22	44324727	G	C	0.216253	0.033724	0.00258242	5.64137E-39
rs116948941	21	38071988	G	A	0.217007	0.0260777	0.00261784	2.24572E-23
rs35062406	X	112157347	CT	C	0.218536	0.01651	0.00216585	2.4809E-14
rs1058	19	44268325	G	C	0.218831	0.0269142	0.00258587	2.2753E-25
rs1155347	6	39146230	T	C	0.780811	0.0175426	0.00260039	1.51828E-11
rs2070895	15	58723939	G	A	0.78052	0.0231314	0.00257697	2.80278E-19
rs6627226	X	149652120	G	T	0.219703	0.0134917	0.00215172	3.60601E-10
rs7849537	9	114897899	T	C	0.780276	0.0209619	0.00257243	3.67929E-16
rs10640460	17	57866165	T	TAA	0.779807	0.0304416	0.00259552	9.10399E-32
rs552708909	19	50090422	CT	C	0.221555	0.0251717	0.00258793	2.32278E-22
rs2835435	21	38055216	C	T	0.777662	0.0316199	0.00260408	6.29169E-34
rs558567978	17	76124810	AG	A	0.222377	0.0213075	0.00258527	1.69517E-16
rs1924930	13	78447373	T	A	0.224612	0.016032	0.00255992	3.78379E-10

rs11072332	15	72108307	G	T	0.773145	0.0212445	0.00254862	7.70491E-17
rs2950015	17	44332351	G	A	0.227758	0.0482432	0.00254317	3.03571E-80
rs34390795	12	133112394	C	CGT	0.229913	0.018972	0.00254564	9.14194E-14
rs34857974	1	45984852	A	ATGT	0.230194	0.0195891	0.00252342	8.29979E-15
rs12423752	12	110294902	T	C	0.769449	0.0196953	0.00253784	8.44921E-15
rs1171615	10	61469090	T	C	0.769072	0.0147984	0.00252751	4.77204E-09
rs8069437	17	44906949	T	C	0.231167	0.0317172	0.00256709	4.55979E-35
rs13133952	4	187977377	C	T	0.768739	0.0148478	0.00254651	5.5213E-09
rs1175550	1	3691528	G	A	0.231502	0.0326381	0.00255302	2.01141E-37
rs11242782	6	2522808	A	G	0.231617	0.0159512	0.00253039	2.90341E-10
rs229932	6	134241797	C	T	0.232317	0.0182203	0.00258719	1.88813E-12
rs143288026	2	145664860	CA	C	0.234183	0.0285235	0.00253132	1.88314E-29
rs12800440	11	95795807	C	T	0.765123	0.0158564	0.00255755	5.65207E-10
rs368417629	11	10195282	G	T	0.764894	0.0467189	0.00253456	7.17475E-76
rs128494	21	37834258	C	T	0.764759	0.0334136	0.00254957	3.05904E-39
rs55639123	15	86160268	G	A	0.235294	0.0156722	0.00251689	4.76023E-10
rs165944	5	88110363	C	T	0.235346	0.0167524	0.0025167	2.80435E-11
rs2184540	9	93801208	A	G	0.235706	0.0174917	0.00250599	2.95272E-12
rs4470295	2	160957824	C	G	0.763718	0.0165058	0.0025031	4.27719E-11
rs543893408	10	77932816	TA	T	0.236877	0.0195677	0.0025348	1.16674E-14
rs3791020	1	173813197	G	A	0.762382	0.0164635	0.00249244	3.96513E-11
rs35790189	3	196176998	G	GGT	0.761941	0.0146975	0.00250986	4.74421E-09
rs5813399	15	66999615	CT	C	0.2402	0.0188309	0.0025147	6.97562E-14
rs12688558	X	57448902	G	T	0.24078	0.0208693	0.00207154	7.17754E-24
rs11708187	3	72396329	A	G	0.241476	0.015815	0.00249456	2.3007E-10

rs76506961	11	8866146	AAAT	A	0.2433	0.0303222	0.00249902	7.00675E-34
rs540730	12	57807114	T	C	0.244861	0.0225578	0.00247353	7.53471E-20
rs79034755	17	80189895	C	T	0.754165	0.0148751	0.00249019	2.32245E-09
rs7168958	15	57719138	G	C	0.753561	0.0165661	0.00247826	2.31605E-11
rs72654647	1	25022314	A	G	0.246726	0.0188636	0.00246901	2.16973E-14
rs7497961	15	66020908	A	G	0.753059	0.0200364	0.00247264	5.35137E-16
rs9270475	6	32559007	A	T	0.752906	0.0491568	0.00252986	4.25026E-84
rs12045893	1	158530416	C	T	0.752769	0.0180726	0.0024578	1.93573E-13
rs2859337	6	12288422	A	G	0.247982	0.0196527	0.00247758	2.15269E-15
rs9614123	22	30371350	A	G	0.750818	0.0199906	0.00245787	4.17751E-16
rs1900920	15	42768198	T	C	0.750641	0.0211169	0.00246537	1.07694E-17
rs218264	4	55408875	A	T	0.750624	0.0443712	0.00248009	1.38689E-71
rs145886884	3	23395310	A	ATT	0.250629	0.0172022	0.00247273	3.48177E-12
rs10864013	1	213039546	T	C	0.747402	0.0175674	0.0024481	7.18161E-13
rs464605	5	55807370	T	C	0.745907	0.0206462	0.00244281	2.86787E-17
rs2069443	7	150755173	G	T	0.25481	0.0189821	0.00243889	7.07743E-15
rs7244849	18	11997350	A	G	0.743018	0.0142129	0.00244543	6.17217E-09
rs7255933	19	45766729	A	G	0.2576	0.0212728	0.0024381	2.65817E-18
rs1016144	17	14822052	A	G	0.741961	0.0169764	0.00246805	6.05001E-12
rs140696582	1	40421617	TG	T	0.739454	0.0198956	0.00242736	2.47705E-16
rs3754140	1	214176380	C	T	0.261392	0.0398752	0.00241319	2.4709E-61
rs7775698	6	135418635	C	T	0.738351	0.0525959	0.00241965	9.1727E-105
rs4075958	5	176784512	A	G	0.2627	0.0279767	0.00241737	5.63431E-31
rs4686671	3	193799549	C	T	0.735495	0.0156546	0.00241506	9.04697E-11
rs5819600	17	17007351	A	AG	0.264722	0.0161183	0.00241773	2.61605E-11

rs2816916	1	200034969	G	A	0.26608	0.0142167	0.00240001	3.14968E-09
rs2836883	21	40466744	A	G	0.266682	0.017799	0.00241796	1.82313E-13
rs5934505	X	8913826	C	T	0.268045	0.03675	0.00199179	5.14357E-76
rs34792580	X	152890468	C	G	0.269759	0.0222047	0.00200014	1.23255E-28
rs10483981	14	86119802	T	C	0.2703	0.0138762	0.00240192	7.59863E-09
rs10494964	1	213966887	C	T	0.271948	0.0260012	0.00243226	1.132E-26
rs9690544	7	129048940	C	A	0.272432	0.018411	0.00239626	1.55135E-14
rs5750644	22	39009135	A	T	0.726258	0.0173242	0.00239962	5.21561E-13
rs1434282	1	199010721	C	T	0.275383	0.0238328	0.00238402	1.57249E-23
rs2983533	6	166074760	C	G	0.275577	0.0157928	0.00239643	4.39423E-11
rs10900027	10	44858681	A	C	0.277205	0.0214638	0.00242021	7.40917E-19
rs9303620	17	27152869	C	T	0.278736	0.020338	0.00237546	1.11205E-17
rs1533495	17	36172155	T	C	0.719963	0.0142406	0.00239677	2.82288E-09
rs2737263	8	116667539	T	G	0.281571	0.0168981	0.00237308	1.07344E-12
rs10819461	9	131841887	T	C	0.718034	0.0156052	0.00236994	4.55991E-11
rs3213545	12	121471337	A	G	0.282906	0.0252117	0.00238877	4.85449E-26
rs10859044	12	91144040	G	T	0.282984	0.0158608	0.00236619	2.04049E-11
rs1495099	17	37784464	C	G	0.283583	0.0206629	0.00236977	2.79755E-18
rs3012053	10	77293109	G	A	0.714844	0.0161399	0.00235659	7.44529E-12
rs549220260	19	4426164	C	CT	0.714625	0.0247229	0.00238343	3.29636E-25
rs10265221	7	151414329	T	C	0.713152	0.0684015	0.00235288	8.2577E-186
rs3814570	10	114708510	T	C	0.287108	0.0141766	0.00235495	1.74508E-09
rs588206	2	45886261	C	G	0.287976	0.0135234	0.00234423	7.98368E-09
rs12863103	13	33723244	T	C	0.288312	0.0177073	0.00235667	5.74813E-14
rs11772705	7	100298904	C	T	0.288452	0.030031	0.00234992	2.13207E-37

rs55777228	16	87869313	C	T	0.28876	0.0161761	0.00239961	1.57152E-11
rs11635174	15	78230929	A	G	0.288788	0.0158236	0.00237385	2.63235E-11
rs2005682	19	35947661	A	T	0.711146	0.0156448	0.00236275	3.55724E-11
rs2379120	20	61030580	A	T	0.28957	0.0178572	0.00238092	6.37565E-14
rs4986080	17	81049741	A	G	0.709804	0.0170248	0.00235859	5.2676E-13
rs10857147	4	81181072	T	A	0.290839	0.0227462	0.00235304	4.17509E-22
rs34160920	14	69264564	CT	C	0.708229	0.0157348	0.00236498	2.86693E-11
rs5906256	X	46557821	T	C	0.293973	0.0152901	0.00194286	3.54945E-15
rs148874756	8	81376290	C	CTGT	0.703858	0.0181814	0.00234001	7.86127E-15
rs12192672	6	7229619	A	G	0.29621	0.0270941	0.00232925	2.83003E-31
rs568754328	19	47623080	C	CT	0.29656	0.0256406	0.00238324	5.39167E-27
rs2823270	21	16794755	A	G	0.700593	0.0280315	0.00233217	2.8069E-33
rs4979080	9	114824708	T	C	0.700433	0.0162957	0.00234005	3.31155E-12
rs13306780	17	42329004	A	C	0.299818	0.0262526	0.00234569	4.47031E-29
rs17278406	9	13976285	G	A	0.698419	0.0350711	0.00233558	5.76739E-51
rs1923032	20	62483993	A	G	0.303062	0.0150565	0.00237999	2.5113E-10
rs35190411	2	25950158	C	CT	0.305191	0.0184688	0.00230687	1.185E-15
rs4803936	19	46878629	G	A	0.69461	0.0153696	0.00231869	3.38953E-11
rs2029466	3	56780003	C	T	0.692401	0.0283006	0.00231697	2.60137E-34
rs9382170	6	13508443	A	T	0.309364	0.0144194	0.00230023	3.64127E-10
rs62435145	7	1286567	G	T	0.309417	0.0373358	0.00237676	1.31864E-55
rs2209098	1	172167226	T	C	0.688915	0.0205705	0.00229334	2.97408E-19
rs1373866	3	143845892	C	T	0.311776	0.0135724	0.00230334	3.80427E-09
rs71450364	2	12930330	T	G	0.311988	0.018579	0.00229107	5.09074E-16
rs34762923	2	204357497	CT	C	0.312384	0.0139664	0.00233512	2.21744E-09

rs11119633	1	211104853	C	T	0.68679	0.0140748	0.00229493	8.62385E-10
rs2396083	6	43804808	C	G	0.686714	0.0379149	0.00230589	9.47151E-61
rs174560	11	61581764	C	T	0.313738	0.0310038	0.00229481	1.35783E-41
rs75792643	22	46372968	T	C	0.315552	0.0326574	0.00229709	7.20154E-46
rs12661188	6	130378833	C	T	0.684157	0.023395	0.0022904	1.70981E-24
rs5995288	22	36762634	T	C	0.317	0.0180151	0.00230611	5.63324E-15
rs768090	2	208003579	A	T	0.317564	0.0190204	0.00229585	1.1842E-16
rs797343	14	34646269	T	C	0.682101	0.0204784	0.00229432	4.4291E-19
rs2576159	10	90296467	T	A	0.319337	0.0142121	0.00228453	4.93943E-10
rs56374617	16	69693035	A	T	0.319345	0.0134	0.00231197	6.79465E-09
rs483180	1	120267505	C	G	0.68042	0.0162737	0.00227859	9.19796E-13
rs6679233	1	151750433	C	T	0.320593	0.0137763	0.00227539	1.40873E-09
rs7910217	10	96999873	A	G	0.321503	0.0138651	0.00227897	1.1727E-09
rs4683294	3	46979013	G	C	0.322768	0.0138302	0.0023868	6.85443E-09
rs6602909	13	114551993	C	T	0.32652	0.0221724	0.00228065	2.43045E-22
rs4886669	15	75448181	C	T	0.32661	0.0190237	0.00227277	5.74861E-17
rs1635474	5	142622669	A	G	0.672814	0.0156694	0.00227917	6.19696E-12
rs12136952	1	12661421	C	G	0.329011	0.0160674	0.00226392	1.27348E-12
rs55995100	7	101231632	G	A	0.670003	0.0154582	0.0022728	1.03614E-11
rs2692533	2	54007171	C	A	0.331385	0.0171644	0.00225719	2.86401E-14
rs2399972	10	13557945	C	G	0.331482	0.0160941	0.00226531	1.2068E-12
rs1544861	11	10679441	C	T	0.666002	0.0196587	0.00226066	3.43788E-18
rs3811444	1	248039451	T	C	0.334557	0.0241785	0.00224738	5.40264E-27
rs10168349	2	46360907	G	C	0.664043	0.0745808	0.00225027	7.1332E-241
rs28606370	10	104670832	C	A	0.664012	0.0154928	0.0022537	6.22581E-12

rs112393215	12	15307775	T	TA	0.663475	0.0153021	0.00227837	1.86462E-11
rs2516470	6	31407331	G	C	0.336773	0.0401324	0.00224548	1.93051E-71
rs565728	2	135283654	C	T	0.662681	0.0149082	0.00228739	7.14606E-11
rs7174574	15	68581390	T	C	0.662653	0.0136674	0.0022757	1.9039E-09
rs2823139	21	16576783	G	A	0.661978	0.0308336	0.00226839	4.42305E-42
rs3803906	19	45715976	G	A	0.341452	0.0229809	0.00225696	2.38027E-24
rs2834318	21	35356814	T	G	0.658359	0.0202799	0.00225028	2.02024E-19
rs9697691	22	46309893	G	C	0.342784	0.0247777	0.00224368	2.36116E-28
rs2720659	8	129060804	A	G	0.343424	0.016659	0.00224099	1.05547E-13
rs9649959	8	128972721	A	G	0.655124	0.0204683	0.0022418	6.83116E-20
rs5846851	3	14929509	C	CA	0.344917	0.0185642	0.00224962	1.55571E-16
rs13028787	2	28941262	C	T	0.348046	0.0142753	0.00224951	2.21043E-10
rs4280597	3	25053482	G	A	0.348493	0.0134875	0.00223506	1.59442E-09
rs1868274	2	46309262	C	G	0.650636	0.0275725	0.00224532	1.16017E-34
rs9791312	6	143142313	C	A	0.349802	0.013291	0.00225546	3.7968E-09
rs34678368	16	51187951	GA	G	0.35265	0.0189993	0.00229889	1.40218E-16
rs7259119	19	32593682	T	C	0.353451	0.0157387	0.00224314	2.27707E-12
rs12644772	4	55335501	C	T	0.354235	0.0231339	0.00222827	2.99434E-25
rs1340817	13	29230581	A	G	0.645242	0.0234	0.0022345	1.15989E-25
rs140581697	2	227111435	T	TTA	0.645129	0.01677	0.00221838	4.04402E-14
rs12894354	14	102987884	C	T	0.644532	0.014257	0.0022351	1.78614E-10
rs10909942	1	3318769	C	G	0.356359	0.0170791	0.00221991	1.43052E-14
rs261291	15	58680178	T	C	0.643317	0.0180782	0.00222905	5.05069E-16
rs865483	17	35851177	A	C	0.357526	0.0146876	0.0022226	3.88767E-11
rs863678	2	176974104	T	G	0.642472	0.025021	0.00227349	3.5933E-28

	6	50791584	C	CACAA	0.358079	0.0183149	0.00232345	3.20558E-15
rs2718146	7	134659326	A	G	0.359459	0.0155163	0.00222242	2.91601E-12
rs3780474	9	32425676	G	T	0.359913	0.0206119	0.00221743	1.46653E-20
rs2871974	15	99284074	T	C	0.639079	0.0133583	0.00222252	1.85032E-09
rs2309753	2	100775920	C	A	0.638762	0.0143686	0.00220935	7.84506E-11
rs6679817	1	89363264	T	C	0.363739	0.0136485	0.00221317	6.96215E-10
rs9380559	6	36207598	A	G	0.634234	0.01388	0.00221121	3.44932E-10
rs4776806	15	66921963	C	G	0.366072	0.0154798	0.0022167	2.88401E-12
rs2934849	6	166162335	T	C	0.368684	0.0146562	0.00220349	2.90402E-11
rs553725343	15	60945980	A	AT	0.631309	0.0174817	0.0022275	4.22301E-15
rs999010	1	231495316	G	A	0.629001	0.0277804	0.00219669	1.16998E-36
rs35124400	12	2522077	T	C	0.371132	0.0284353	0.00220722	5.62153E-38
rs12718730	7	50436828	T	A	0.627098	0.0131337	0.00220554	2.60282E-09
rs66520518	3	58405636	TTAAG	T	0.627052	0.0217412	0.00220537	6.31085E-23
rs34881325	9	2622134	T	C	0.375621	0.0237125	0.00225365	6.85366E-26
rs218476	20	57237670	A	G	0.377838	0.0138841	0.00220433	3.00428E-10
rs28432336	4	87984331	G	A	0.37842	0.0325823	0.00220826	2.86799E-49
rs60992881	16	157592	CAA	C	0.381745	0.0220035	0.00227119	3.38795E-22
rs78374304	12	48403839	A	T	0.617865	0.0338903	0.00219047	5.38635E-54
rs575138	1	53328394	G	C	0.617618	0.0174932	0.00218322	1.12337E-15
rs4660253	1	43761651	C	T	0.382407	0.0203162	0.00222984	8.1583E-20
rs11022762	11	13335926	T	C	0.383346	0.0173931	0.00218987	1.98104E-15
rs937851	11	6667353	A	G	0.383568	0.0139143	0.00220442	2.75451E-10
rs71535075	6	16289908	G	GTC	0.616262	0.020488	0.00219272	9.30724E-21
rs12897414	14	34724550	C	T	0.384387	0.0146433	0.00221875	4.11721E-11

rs6665323	1	47953054	T	C	0.386022	0.0138137	0.00218318	2.4944E-10
rs11122450	1	230301811	T	G	0.388306	0.0151786	0.00217973	3.31867E-12
rs11556924	7	129663496	C	T	0.610994	0.0171029	0.00218432	4.88449E-15
rs1558801	12	109036359	C	A	0.390936	0.0129219	0.00219598	3.99638E-09
rs13389219	2	165528876	C	T	0.607385	0.0193156	0.00217216	5.98316E-19
rs5798300	12	54749111	AGT	A	0.39295	0.0136749	0.0022842	2.14095E-09
rs4233937	2	66752251	G	A	0.606598	0.0150982	0.00217496	3.87056E-12
rs3053061	15	78537238	C	CGGGGTGCGG	0.605686	0.0232103	0.00218463	2.29577E-26
rs123698	19	807442	C	G	0.604603	0.029195	0.00217808	5.72847E-41
rs142351272	X	109887299	T	C	0.603933	0.0148976	0.00180844	1.75379E-16
rs36068213	1	48116939	T	TAA	0.396194	0.0160025	0.00221277	4.76333E-13
rs6415788	9	4118111	G	T	0.396543	0.02412	0.00220626	8.05717E-28
rs9969563	8	12897602	T	C	0.603286	0.0131874	0.00219498	1.87847E-09
rs760077	1	155178782	T	A	0.602237	0.033005	0.00216713	2.24208E-52
rs11122174	1	231114595	C	T	0.601757	0.0145879	0.00216511	1.60894E-11
rs7224610	17	53364788	C	A	0.398504	0.0140132	0.00219566	1.74531E-10
rs13103534	4	148976981	A	G	0.398846	0.0175928	0.00218333	7.76904E-16
rs140307022	1	16370178	AGCTCT	A	0.399727	0.0237954	0.00221501	6.40741E-27
rs12985346	19	2164351	A	T	0.400983	0.0319754	0.00218059	1.10193E-48
rs3809627	16	30103160	A	C	0.402081	0.0218503	0.00217603	1.00282E-23
rs56388355	3	30181399	AG	A	0.40279	0.015903	0.00227244	2.59263E-12
rs139384	22	39527935	C	T	0.403155	0.01609	0.00218118	1.62211E-13
rs78839561	20	4142839	A	G	0.403289	0.0135037	0.00220605	9.28662E-10
rs8000803	13	71714717	T	G	0.595352	0.01367	0.00217458	3.25181E-10
rs11405520	21	44139647	A	AG	0.405197	0.0153039	0.00218953	2.75706E-12

rs7309382	12	115366182	C	T	0.405319	0.0160515	0.00217464	1.56844E-13
rs72717436	5	448291	A	C	0.405452	0.0181457	0.00217194	6.56522E-17
rs1472226	1	212402580	G	A	0.594062	0.0146935	0.00216562	1.16173E-11
rs3809770	17	47047596	A	G	0.593638	0.0136028	0.00217402	3.92503E-10
rs1535099	14	104194278	C	T	0.407766	0.0136751	0.00217414	3.17692E-10
rs35538465	4	157836033	G	GGGAAAGTCT	0.408116	0.0133262	0.002187	1.10558E-09
rs7703616	5	1115115	C	T	0.408174	0.0151988	0.00217716	2.93042E-12
rs668459	6	139835689	C	T	0.41002	0.0170228	0.00215699	2.97564E-15
rs56267269	8	42399667	T	C	0.58915	0.0267479	0.00216652	5.11812E-35
rs35750745	7	671593	CA	C	0.588338	0.0242886	0.00221422	5.36444E-28
rs4683603	3	141092050	T	G	0.411702	0.0155617	0.00218034	9.51949E-13
rs3740689	11	47380593	A	G	0.586218	0.0165885	0.00216153	1.66162E-14
rs714195	4	146445680	C	T	0.583483	0.0136561	0.002164	2.77988E-10
rs1110542	16	79199414	A	T	0.582578	0.0155597	0.00216446	6.54009E-13
rs4969145	17	76406170	T	C	0.417873	0.0134549	0.00219758	9.20615E-10
rs60325980	17	61880270	C	CA	0.417876	0.019725	0.0022245	7.50343E-19
rs2647187	1	17409382	C	T	0.581413	0.0170685	0.00216293	2.98866E-15
rs2281841	10	45406608	T	C	0.419001	0.0151051	0.00223804	1.48607E-11
rs17006441	3	69841880	A	C	0.419336	0.0236806	0.00216394	7.15636E-28
rs57828851	17	7732351	T	A	0.419452	0.0230896	0.00217466	2.46911E-26
rs5785906	10	71007973	C	CT	0.57989	0.0274961	0.00218791	3.19488E-36
rs4554203	5	147886011	G	A	0.579675	0.0131515	0.0021614	1.167E-09
rs551118	16	88856084	C	G	0.420693	0.0466922	0.00219417	1.735E-100
rs9823829	3	194506427	A	G	0.578133	0.0189208	0.0021908	5.79716E-18
rs6428637	1	88641063	G	A	0.422003	0.0138517	0.00215267	1.2375E-10

rs743417	21	35347960	C	T	0.423447	0.0237126	0.00216544	6.60989E-28
rs4927705	3	195751397	A	T	0.575788	0.0150169	0.00215371	3.11196E-12
rs11345369	8	61175880	T	TA	0.42456	0.0140097	0.00217528	1.19149E-10
rs2894802	6	52656169	G	T	0.572889	0.0239644	0.00215748	1.15229E-28
rs228917	22	37506410	T	C	0.572476	0.041752	0.00215078	6.05193E-84
rs78968605	12	117588433	TG	T	0.567132	0.0155344	0.00215305	5.39074E-13
rs1841677	13	51386889	T	C	0.5671	0.0141471	0.00216088	5.87374E-11
rs6998007	8	40045119	C	T	0.435111	0.0171805	0.00215796	1.70029E-15
rs9429088	1	46497500	T	A	0.564551	0.0281245	0.00214046	1.95604E-39
rs2250598	16	89698070	T	C	0.435744	0.0253551	0.00216493	1.10977E-31
rs34099733	1	182997286	C	CCT	0.436663	0.0133804	0.00215223	5.06771E-10
rs7916396	10	3245548	C	T	0.436898	0.0133303	0.00214624	5.26479E-10
rs919798	19	4498154	A	G	0.562361	0.0289401	0.00215526	4.16357E-41
rs988397	3	169098791	C	T	0.561784	0.0192779	0.00214661	2.69232E-19
rs855791	22	37462936	G	A	0.561407	0.10122	0.00215389	0
rs9398804	6	126703390	T	A	0.560897	0.0204196	0.00216175	3.52479E-21
rs7956653	12	64901246	G	A	0.559444	0.0135227	0.00216051	3.87376E-10
rs6927173	6	7222093	T	G	0.443209	0.0207656	0.00215737	6.24491E-22
rs13007705	2	239069196	T	C	0.444478	0.0153928	0.00214679	7.49172E-13
rs560176773	6	1740921	AT	A	0.444556	0.0155624	0.00216599	6.72613E-13
rs6595714	5	125843369	T	C	0.444611	0.0127823	0.00214189	2.40523E-09
rs2122113	2	46225778	T	C	0.553377	0.0149907	0.00214259	2.62375E-12
rs34723331	11	8742865	TC	T	0.553102	0.0285938	0.00220113	1.38502E-38
rs1427445	2	219555573	C	A	0.553075	0.0207235	0.00213489	2.81342E-22
rs3842397	18	43845880	CTT	C	0.446981	0.017215	0.00214967	1.16394E-15

rs12056626	8	116557955	G	C	0.552966	0.0196153	0.00214754	6.61079E-20
rs4823309	22	46004023	C	T	0.552354	0.0287116	0.00219213	3.3992E-39
rs10930763	2	177779131	A	T	0.449825	0.0123388	0.00213591	7.61188E-09
rs6016505	20	39678289	T	C	0.548564	0.0145559	0.00214338	1.11285E-11
rs1539508	6	43868986	A	G	0.547901	0.0205501	0.00216093	1.90955E-21
rs559406	18	12857002	G	T	0.452299	0.0166508	0.002145	8.31987E-15
rs11356211	19	46219958	CA	C	0.547466	0.0150505	0.00216772	3.83834E-12
rs963837	11	30749090	C	T	0.452886	0.0240441	0.00213799	2.41995E-29
rs6126019	20	49101590	C	T	0.546847	0.0214311	0.00214967	2.07304E-23
rs2466076	8	32432796	T	G	0.545661	0.0144568	0.00214593	1.61862E-11
rs8027685	15	76277092	T	C	0.454913	0.0262767	0.00214014	1.18859E-34
rs6911827	6	22130601	T	C	0.455242	0.019636	0.0021358	3.79398E-20
rs115986297	6	2050791	A	G	0.456599	0.025845	0.00213961	1.35865E-33
rs112597538	10	60272708	C	T	0.457513	0.0170114	0.0021384	1.78842E-15
rs35736498	2	242644728	T	C	0.45804	0.0138188	0.00213504	9.64682E-11
rs464499	5	34660677	A	G	0.541748	0.0166549	0.00213987	7.07483E-15
rs11258533	10	13737768	T	C	0.458288	0.0163895	0.0021671	3.94265E-14
rs142720800	11	3944329	C	CA	0.540174	0.0131067	0.00219362	2.30206E-09
rs4953348	2	46558432	G	A	0.460644	0.0162481	0.00213115	2.45743E-14
rs1949481	11	16259405	T	C	0.538446	0.0126251	0.00214228	3.78557E-09
rs28418580	4	89742244	T	C	0.461651	0.0153326	0.00214981	9.8869E-13
rs550369338	1	26183738	A	AT	0.53762	0.0150981	0.00214311	1.8553E-12
rs139551207	22	38603571	C	CCAGTAGCTGGGACTA	0.537579	0.0133619	0.00214445	4.63699E-10
rs66782572	3	52567617	A	G	0.463361	0.0223702	0.00233631	1.01844E-21
rs9849756	3	187641417	C	T	0.535332	0.0136114	0.00227146	2.06821E-09

rs1239682	13	51165494	C	T	0.465395	0.0162294	0.00214073	3.42253E-14
rs972762	5	34507508	T	G	0.533841	0.0154586	0.00215829	7.92596E-13
rs7040995	9	92226172	C	G	0.53272	0.0128548	0.00215219	2.33083E-09
rs10900568	1	204220232	C	T	0.532405	0.0135896	0.00212858	1.72127E-10
rs1076230	19	1244900	G	A	0.468938	0.0128987	0.00213375	1.49331E-09
rs867606797	3	12737066	GGTT	G	0.53051	0.0156746	0.00222085	1.69003E-12
rs8020977	14	76624547	G	T	0.529856	0.0160255	0.00213795	6.59294E-14
rs8138197	22	43114551	A	G	0.472068	0.0278201	0.00213508	8.26038E-39
rs6667862	1	15814186	T	C	0.527518	0.0159325	0.00212532	6.55353E-14
rs10171620	2	43429058	G	T	0.527464	0.0199781	0.00213623	8.59764E-21
rs1860302	17	19914399	A	T	0.527018	0.0127991	0.00213572	2.06165E-09
rs1256061	14	64703593	G	T	0.524517	0.0250164	0.00214058	1.48998E-31
rs36012880	14	24881986	C	T	0.476547	0.0135486	0.00213857	2.36796E-10
rs35943712	3	121628658	AT	A	0.523255	0.0168966	0.00213848	2.76188E-15
rs4897160	6	126223944	A	G	0.477008	0.0146574	0.00212615	5.4295E-12
rs6465351	7	91773213	T	C	0.522987	0.0187511	0.00213129	1.3927E-18
rs6108789	20	10974785	G	C	0.518873	0.0127125	0.00214323	3.00238E-09
rs3184504	12	111884608	T	C	0.481371	0.0679936	0.00213004	1.3531E-223
rs10142359	14	73884540	G	A	0.481701	0.0155303	0.00213632	3.60418E-13
rs998584	6	43757896	A	C	0.482625	0.0200236	0.00213124	5.7065E-21
rs17773190	2	47030363	G	A	0.482717	0.0213011	0.00214481	3.03767E-23
rs2337106	18	46460903	C	G	0.483325	0.0205874	0.00214904	9.72214E-22
rs9296668	6	51838263	G	A	0.483801	0.0145156	0.00215398	1.59517E-11
rs10956934	8	95992473	C	A	0.515658	0.0248437	0.00213529	2.7419E-31
rs8073217	17	28164263	G	C	0.51339	0.0124998	0.00214586	5.70862E-09

rs1893989	11	111202478	G	A	0.513307	0.0169737	0.00213062	1.6316E-15
rs4886755	15	76298132	A	G	0.48761	0.0357024	0.0021328	6.73437E-63
rs2970876	4	23888619	A	G	0.509836	0.012925	0.00213705	1.46603E-09
rs9366927	6	37027232	C	T	0.490845	0.022003	0.00213584	6.91538E-25
rs10898979	11	74039147	T	C	0.508753	0.01256	0.00213371	3.94554E-09
rs1863127	12	46199589	C	T	0.508463	0.015054	0.00213246	1.6716E-12
rs9482771	6	127446610	C	G	0.493905	0.0265257	0.00212896	1.24203E-35
rs1062601	20	56137834	G	A	0.505726	0.0231614	0.00213988	2.6585E-27
rs734663	2	39743876	C	T	0.505569	0.013874	0.00212708	6.91189E-11
rs1257415	14	99698931	A	C	0.505518	0.0152783	0.00214257	9.97663E-13
rs13127730	4	48590606	C	T	0.495795	0.0144425	0.00212912	1.17446E-11
rs833061	6	43737486	C	T	0.504045	0.0335787	0.00212603	3.41712E-56
rs1331308	6	135405122	A	C	0.503785	0.0138464	0.00213483	8.81746E-11
rs10904089	10	3791224	T	G	0.496234	0.0159163	0.00216652	2.03492E-13
rs3760994	19	1435771	A	G	0.496267	0.015308	0.00212957	6.56003E-13
rs66533066	1	172415351	GA	G	0.503075	0.0134659	0.00213048	2.60559E-10
rs871841	17	8216468	C	T	0.498004	0.0210849	0.00213298	4.82694E-23
rs1075871	3	194681297	G	A	0.498266	0.0199046	0.00213387	1.07954E-20
rs2870238	4	77373079	T	C	0.501562	0.026815	0.00213036	2.48722E-36
rs66468814	10	36461955	C	CT	0.501517	0.0145919	0.0021396	9.10915E-12

Bibliography

Adli, M. (2018). The CRISPR tool kit for genome editing and beyond. *Nature Communications*, *9*(1), 1911. https://doi.org/10.1038/s41467-018-04252-2

Agrotis, A., & Ketteler, R. (2015). A new age in functional genomics using CRISPR/Cas9 in arrayed library screening . In *Frontiers in Genetics* (Vol. 6, p. 300). https://www.frontiersin.org/article/10.3389/fgene.2015.00300

Akizawa, T., Iwasaki, M., Otsuka, T., Yamaguchi, Y., & Reusch, M. (2021). Phase 3 Study of Roxadustat to Treat Anemia in Non–Dialysis-Dependant CKD. *Kidney International Reports*, *6*(7), 1810–1828. https://doi.org/https://doi.org/10.1016/j.ekir.2021.04.003

Akizawa, T., Iwasaki, M., Yamaguchi, Y., Majikawa, Y., & Reusch, M. (2020). Phase 3, Randomized, Double-Blind, Active-Comparator (Darbepoetin Alfa) Study of Oral Roxadustat in CKD Patients with Anemia on Hemodialysis in Japan. *Journal of the American Society of Nephrology*, *31*(7), 1628–1639. https://doi.org/10.1681/ASN.2019060623

Akizawa, T., Nangaku, M., Yonekawa, T., Okuda, N., Kawamatsu, S., Onoue, T., Endo, Y., Hara, K., & Cobitz, A. R. (2020). Efficacy and Safety of Daprodustat Compared with Darbepoetin Alfa in Japanese Hemodialysis Patients with Anemia: A Randomized, Double-Blind, Phase 3 Trial. *Clinical Journal of the American Society of Nephrology : CJASN*, *15*(8), 1155–1165. https://doi.org/10.2215/CJN.16011219

Albert, F. W., & Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, *16*(4), 197–212. https://doi.org/10.1038/nrg3891

Amanzada, A., Goralczyk, A. D., Reinhardt, L., Moriconi, F., Cameron, S., & Mihm, S. (2014). Erythropoietin rs1617640 G allele associates with an attenuated rise of serum erythropoietin and a marked decline of hemoglobin in hepatitis C patients undergoing antiviral therapy. *BMC Infectious Diseases*, *14*, 503. https://doi.org/10.1186/1471-2334-14-503

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106. https://doi.org/10.1186/gb-2010-11-10-r106

Au, V., Li-Leger, E., Raymant, G., Flibotte, S., Chen, G., Martin, K., Fernando, L., Doell, C., Rosell, F. I., Wang, S., Edgley, M. L., Rougvie, A. E., Hutter, H., & Moerman, D. G. (2019). CRISPR/Cas9 Methodology for the Generation of Knockout Deletions in Caenorhabditis elegans. *G3 (Bethesda, Md.)*, *9*(1), 135–144. https://doi.org/10.1534/g3.118.200778

Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Aviner, R. (2020). The science of puromycin: From studies of ribosome function to applications in biotechnology. *Computational and Structural Biotechnology Journal*, *18*, 1074–1083. https://doi.org/10.1016/j.csbj.2020.04.014

Babitt, J. L., & Lin, H. Y. (2012). Mechanisms of anemia in CKD. *Journal of the American Society of Nephrology : JASN*, *23*(10), 1631–1634. https://doi.org/10.1681/ASN.2011111078

Backman, J. D., Li, A. H., Marcketta, A., Sun, D., Mbatchou, J., Kessler, M. D., Benner, C., Liu, D., Locke, A. E., Balasubramanian, S., Yadav, A., Banerjee, N., Gillies, C. E., Damask, A., Liu, S., Bai, X., Hawes, A., Maxwell, E., Gurski, L., … DiscovEHR. (2021). Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature*, *599*(7886), 628–634. https://doi.org/10.1038/s41586-021-04103-z

Baird-Gunning, J., & Bromley, J. (2016). Correcting iron deficiency. *Australian Prescriber*, *39*(6), 193–199. https://doi.org/10.18773/austprescr.2016.069

Batchelor, E. K., Kapitsinou, P., Pergola, P. E., Kovesdy, C. P., & Jalal, D. I. (2020). Iron Deficiency in Chronic Kidney Disease: Updates on Pathophysiology, Diagnosis, and Treatment. *Journal of the American Society of Nephrology*, *31*(3), 456–468. https://doi.org/10.1681/ASN.2019020213

Becker, S., & Boch, J. (2021). TALE and TALEN genome editing technologies. *Gene and Genome Editing*, *2*, 100007. https://doi.org/https://doi.org/10.1016/j.ggedit.2021.100007

Benchling [Biology Software]. (2021). *Benchling [Biology Software]*. https://benchling.com

Benner, C., Spencer, C. C. A., Havulinna, A. S., Salomaa, V., Ripatti, S., & Pirinen, M. (2016). FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics (Oxford, England)*, *32*(10), 1493–1501. https://doi.org/10.1093/bioinformatics/btw018

Bennett, D. A., & Holmes, M. V. (2017). Mendelian randomisation in cardiovascular research: an introduction for clinicians. *Heart*, *103*(18), 1400 LP – 1407. https://doi.org/10.1136/heartjnl-2016-310605

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59. https://doi.org/10.1038/nature07517

Benyamin, B., Esko, T., Ried, J. S., Radhakrishnan, A., Vermeulen, S. H., Traglia, M., Gögele, M., Anderson, D., Broer, L., Podmore, C., Luan, J., Kutalik, Z., Sanna, S., van der Meer, P., Tanaka, T., Wang, F., Westra, H.-J., Franke, L., Mihailov, E., … Whitfield, J. B. (2014). Novel loci affecting iron homeostasis and their effects in individuals at risk for hemochromatosis. *Nature Communications*, *5*, 4926. https://doi.org/10.1038/ncomms5926

Berlian, G., Tandrasasmita, O. M., & Tjandrawinata, R. R. (2019). Upregulation of endogenous erythropoietin expression by DLBS6747, a bioactive fraction of Ipomoea batatas L. leaves, via increasing HIF1α transcription factor in HEK293 kidney cells. *Journal of Ethnopharmacology*, *235*, 190–198. https://doi.org/10.1016/j.jep.2019.01.016

Bialk, P., Rivera-Torres, N., Strouse, B., & Kmiec, E. B. (2015). Regulation of Gene Editing Activity Directed by Single-Stranded Oligonucleotides and CRISPR/Cas9 Systems. *PloS One*, *10*(6), e0129308–e0129308. https://doi.org/10.1371/journal.pone.0129308

Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstadt, A., & Bonas, U. (2009). Breaking the code of DNA binding specificity of TAL-type III effectors. *Science (New York, N.Y.)*, *326*(5959), 1509–1512. https://doi.org/10.1126/science.1178811

Bodapati, S., Daley, T. P., Lin, X., Zou, J., & Qi, L. S. (2020). A benchmark of algorithms for the analysis of pooled CRISPR screens. *Genome Biology*, *21*(1), 62. https://doi.org/10.1186/s13059-020-01972-x

Boel, A., De Saffel, H., Steyaert, W., Callewaert, B., De Paepe, A., Coucke, P. J., & Willaert, A. (2018). CRISPR/Cas9-mediated homology-directed repair by ssODNs in zebrafish induces complex mutational patterns resulting from genomic integration of repair-template fragments. *Disease Models & Mechanisms*, *11*(10), dmm035352. https://doi.org/10.1242/dmm.035352

Bollen, Y., Post, J., Koo, B.-K., & Snippert, H. J. G. (2018). How to create state-of-the-art genetic model systems: strategies for optimal CRISPR-mediated genome editing. *Nucleic Acids Research*, *46*(13), 6435–6454. https://doi.org/10.1093/nar/gky571

Bonjoch, L., Mur, P., Arnau-Collell, C., Vargas-Parra, G., Shamloo, B., Franch-Expósito, S., Pineda, M., Capellà, G., Erman, B., & Castellví-Bel, S. (2019). Approaches to functionally validate candidate genetic variants involved in colorectal cancer predisposition. *Molecular Aspects of Medicine*, *69*, 27–40. https://doi.org/https://doi.org/10.1016/j.mam.2019.03.004

Bonomini, M., Del Vecchio, L., Sirolli, V., & Locatelli, F. (2016). New Treatment Approaches for the Anemia of CKD. *American Journal of Kidney Diseases*, *67*(1), 133–142. https://doi.org/10.1053/j.ajkd.2015.06.030

Bonora, M., Patergnani, S., Rimessi, A., De Marchi, E., Suski, J. M., Bononi, A., Giorgi, C., Marchi, S., Missiroli, S., Poletti, F., Wieckowski, M. R., & Pinton, P. (2012). ATP synthesis and storage. *Purinergic Signalling*, *8*(3), 343–357. https://doi.org/10.1007/s11302-012-9305-8

Boratyn, G. M., Camacho, C., Cooper, P. S., Coulouris, G., Fong, A., Ma, N., Madden, T. L., Matten, W. T., McGinnis, S. D., Merezhuk, Y., Raytselis, Y., Sayers, E. W., Tao, T., Ye, J., & Zaretskaya, I. (2013). BLAST: a more efficient report with usability improvements. *Nucleic Acids Research*, *41*(W1), W29–W33. https://doi.org/10.1093/nar/gkt282

Bovijn, J., Lindgren, C. M., & Holmes, M. V. (2020). Genetic variants mimicking therapeutic inhibition of IL-6 receptor signaling and risk of COVID-19. *The Lancet. Rheumatology*, *2*(11), e658–e659. https://doi.org/10.1016/S2665-9913(20)30345-3

Bowden, J., Davey Smith, G., & Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*, *44*(2), 512–525. https://doi.org/10.1093/ije/dyv080

Bowden, J., Davey Smith, G., Haycock, P. C., & Burgess, S. (2016). Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*, *40*(4), 304–314.

Bowden, J., Del Greco M, F., Minelli, C., Zhao, Q., Lawlor, D. A., Sheehan, N. A., Thompson, J., & Davey Smith, G. (2019). Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the NOME assumption. *International Journal of Epidemiology*, *48*(3), 728–742. https://doi.org/10.1093/ije/dyy258

Brandt, M., Gokden, A., Ziosi, M., & Lappalainen, T. (2020). A polyclonal allelic expression assay for detecting regulatory effects of transcript variants. *Genome Medicine*, *12*(1), 79. https://doi.org/10.1186/s13073-020-00777-8

Brigandi, R. A., Johnson, B., Oei, C., Westerman, M., Olbina, G., de Zoysa, J., Roger, S. D., Sahay, M., Cross, N., McMahon, L., Guptha, V., Smolyarchuk, E. A., Singh, N., Russ, S. F., & Kumar, S. (2016). A Novel Hypoxia-Inducible Factor-Prolyl Hydroxylase Inhibitor (GSK1278863) for Anemia in CKD: A 28-Day, Phase 2A Randomized Trial. *American Journal of Kidney Diseases : The Official Journal of the National Kidney Foundation*, *67*(6), 861–871. https://doi.org/10.1053/j.ajkd.2015.11.021

Broekema, R. V, Bakker, O. B., & Jonkers, I. H. (2020). A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. *Open Biology*, *10*(1), 190221. https://doi.org/10.1098/rsob.190221

Broxmeyer, H. E. (2013). Erythropoietin: multiple targets, actions, and modifying influences for biological and clinical consideration. *The Journal of Experimental Medicine*, *210*(2), 205–208. https://doi.org/10.1084/jem.20122760

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., … Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(D1), D1005–D1012. https://doi.org/10.1093/nar/gky1120

Burgess, S., Bowden, J., Fall, T., Ingelsson, E., & Thompson, S. G. (2017). Sensitivity analyses for robust causal inference from Mendelian randomization analyses with multiple genetic variants. *Epidemiology (Cambridge, Mass.)*, *28*(1), 30.

Burgess, S., Butterworth, A., Malarstig, A., & Thompson, S. G. (2012). Use of Mendelian randomisation to assess potential benefit of clinical intervention. *BMJ*, *345*. https://doi.org/10.1136/bmj.e7325

Burgess, S., Davies, N. M., & Thompson, S. G. (2016). Bias due to participant overlap in two-sample Mendelian randomization. *Genetic Epidemiology*, *40*(7), 597–608. https://doi.org/10.1002/gepi.21998

Burgess, S., Small, D. S., & Thompson, S. G. (2017). A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research*, *26*(5), 2333–2355. https://doi.org/10.1177/0962280215597579

Burgess, S., & Thompson, S. G. (2010). *Avoiding bias from weak instruments in Mendelian randomization studies*.

Burgess, S., & Thompson, S. G. (2017). Interpreting findings from Mendelian randomization using the MR-Egger method. *European Journal of Epidemiology*, *32*(5), 377–389. https://doi.org/10.1007/s10654-017-0255-x

Burgess, S., Timpson, N. J., Ebrahim, S., & Davey Smith, G. (2015). Mendelian randomization: where are we now and where are we going? *International Journal of Epidemiology*, *44*(2), 379–388. https://doi.org/10.1093/ije/dyv108

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS Computational Biology*, *8*(12), e1002822–e1002822. https://doi.org/10.1371/journal.pcbi.1002822

Bush, W. S., Oetjens, M. T., & Crawford, D. C. (2016). Unravelling the human genome–phenome relationship using phenome-wide association studies. *Nature Reviews Genetics*, *17*(3), 129–145. https://doi.org/10.1038/nrg.2015.36

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203–209. https://doi.org/10.1038/s41586-018-0579-z

Cano-Gamez, E., & Trynka, G. (2020). From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases . In *Frontiers in Genetics* (Vol. 11, p. 424). https://www.frontiersin.org/article/10.3389/fgene.2020.00424

Carey, J. L., Nader, N., Chai, P. R., Carreiro, S., Griswold, M. K., & Boyle, K. L. (2017). Drugs and Medical Devices: Adverse Events and the Impact on Women's Health. *Clinical Therapeutics*, *39*(1), 10–22. https://doi.org/https://doi.org/10.1016/j.clinthera.2016.12.009

Carroll, D. (2011). Genome engineering with zinc-finger nucleases. *Genetics*, *188*(4), 773–782. https://doi.org/10.1534/genetics.111.131433

Cases, A., Egocheaga, M. I., Tranche, S., Pallarés, V., Ojeda, R., Górriz, J. L., & Portolés, J. M. (2018). Anemia of chronic kidney disease: Protocol of study, management and referral to Nephrology. *Nefrologia*, *38*(1), 8–12. https://doi.org/10.1016/j.nefro.2017.09.004

Cavadas, M. A. S., Mesnieres, M., Crifo, B., Manresa, M. C., Selfridge, A. C., Keogh, C. E., Fabian, Z., Scholz, C. C., Nolan, K. A., Rocha, L. M. A., Tambuwala, M. M., Brown, S., Wdowicz, A., Corbett, D., Murphy, K. J., Godson, C., Cummins, E. P., Taylor, C. T., & Cheong, A. (2016). REST is a hypoxia-responsive transcriptional repressor. *Scientific Reports*, *6*(1), 31355. https://doi.org/10.1038/srep31355

Chamorro, M. E., Wenker, S. D., Vota, D. M., Vittori, D. C., & Nesse, A. B. (2013). Signaling pathways of cell proliferation are involved in the differential effect of erythropoietin and its carbamylated derivative. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, *1833*(8), 1960–1968. https://doi.org/https://doi.org/10.1016/j.bbamcr.2013.04.006

Chaparro, C. M., & Suchdev, P. S. (2019). Anemia epidemiology, pathophysiology, and etiology in low- and middle-income countries. *Annals of the New York Academy of Sciences*, *1450*(1), 15–31. https://doi.org/10.1111/nyas.14092

Chapman, J. R., Taylor, M. R. G., & Boulton, S. J. (2012). Playing the end game: DNA double-strand break repair pathway choice. *Molecular Cell*, *47*(4), 497–510. https://doi.org/10.1016/j.molcel.2012.07.029

Chen, C.-L., Rodiger, J., Chung, V., Viswanatha, R., Mohr, S. E., Hu, Y., & Perrimon, N. (2020). SNP-CRISPR: A Web Tool for SNP-Specific Genome Editing. *G3 (Bethesda, Md.)*, *10*(2), 489–494. https://doi.org/10.1534/g3.119.400904

Chen, C.-Y. (2014). DNA polymerases drive DNA sequencing-by-synthesis technologies: both past and present. *Frontiers in Microbiology*, *5*, 305. https://doi.org/10.3389/fmicb.2014.00305

Chen, C., Hou, J., Shi, X., Yang, H., Birchler, J. A., & Cheng, J. (2021). DeepGRN: prediction of transcription factor binding site across cell-types using attention-based deep neural networks. *BMC Bioinformatics*, *22*(1), 38. https://doi.org/10.1186/s12859-020-03952-1

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., & Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, *14*, 128. https://doi.org/10.1186/1471-2105-14-128

Chen, M.-H., Raffield, L. M., Mousas, A., Sakaue, S., Huffman, J. E., Moscati, A., Trivedi, B., Jiang, T., Akbari, P., Vuckovic, D., Bao, E. L., Zhong, X., Manansala, R., Laplante, V., Chen, M., Lo, K. S., Qian, H., Lareau, C. A., Beaudoin, M., … Lettre, G. (2020). Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell*, *182*(5), 1198-1213.e14. https://doi.org/10.1016/j.cell.2020.06.045

Chen, N., Hao, C., Liu, B.-C., Lin, H., Wang, C., Xing, C., Liang, X., Jiang, G., Liu, Z., Li, X., Zuo, L., Luo, L., Wang, J., Zhao, M., Liu, Z., Cai, G.-Y., Hao, L., Leong, R., Wang, C., … Yu, K.-H. P. (2019). Roxadustat Treatment for Anemia in Patients Undergoing Long-Term Dialysis. *New England Journal of Medicine*, *381*(11), 1011–1022. https://doi.org/10.1056/NEJMoa1901713

Chen, N., Hao, C., Peng, X., Lin, H., Yin, A., Hao, L., Tao, Y., Liang, X., Liu, Z., Xing, C., Chen, J., Luo, L., Zuo, L., Liao, Y., Liu, B.-C., Leong, R., Wang, C., Liu, C., Neff, T., … Yu, K.-H. P. (2019). Roxadustat for Anemia in Patients with Kidney Disease Not Receiving Dialysis. *New England Journal of Medicine*, *381*(11), 1001–1010.

https://doi.org/10.1056/NEJMoa1813599

Chen, Q., Luo, W., Veach, R. A., Hickman, A. B., Wilson, M. H., & Dyda, F. (2020). Structural basis of seamless excision and specific targeting by piggyBac transposase. *Nature Communications*, *11*(1), 3446. https://doi.org/10.1038/s41467-020-17128-1

Chen, T. K., Knicely, D. H., & Grams, M. E. (2019). Chronic Kidney Disease Diagnosis and Management: A Review. *JAMA*, *322*(13), 1294–1304. https://doi.org/10.1001/jama.2019.14745

Chertow, G. M., Pergola, P. E., Farag, Y. M. K., Agarwal, R., Arnold, S., Bako, G., Block, G. A., Burke, S., Castillo, F. P., Jardine, A. G., Khawaja, Z., Koury, M. J., Lewis, E. F., Lin, T., Luo, W., Maroni, B. J., Matsushita, K., McCullough, P. A., Parfrey, P. S., … Eckardt, K.-U. (2021). Vadadustat in Patients with Anemia and Non–Dialysis-Dependent CKD. *New England Journal of Medicine*, *384*(17), 1589–1600. https://doi.org/10.1056/NEJMoa2035938

Childebayeva, A., Harman, T., Weinstein, J., Goodrich, J. M., Dolinoy, D. C., Day, T. A., Bigham, A. W., & Brutsaert, T. D. (2019). DNA Methylation Changes Are Associated With an Incremental Ascent to High Altitude . In *Frontiers in Genetics* (Vol. 10). https://www.frontiersin.org/article/10.3389/fgene.2019.01062

Chin, C. L., Goh, J. B., Srinivasan, H., Liu, K. I., Gowher, A., Shanmugam, R., Lim, H. L., Choo, M., Tang, W. Q., Tan, A. H.-M., Nguyen-Khuong, T., Tan, M. H., & Ng, S. K. (2019). A human expression system based on HEK293 for the stable production of recombinant erythropoietin. *Scientific Reports*, *9*(1), 16768. https://doi.org/10.1038/s41598-019-53391-z

Choi, J., Zhang, T., Vu, A., Ablain, J., Makowski, M. M., Colli, L. M., Xu, M., Hennessey, R. C., Yin, J., Rothschild, H., Gräwe, C., Kovacs, M. A., Funderburk, K. M., Brossard, M., Taylor, J., Pasaniuc, B., Chari, R., Chanock, S. J., Hoggart, C. J., … Brown, K. M. (2020). Massively parallel reporter assays of melanoma risk variants identify MX2 as a gene promoting melanoma. *Nature Communications*, *11*(1), 2718. https://doi.org/10.1038/s41467-020-16590-1

Chu, V. T., Weber, T., Wefers, B., Wurst, W., Sander, S., Rajewsky, K., & Kühn, R. (2015). Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. *Nature Biotechnology*, *33*(5), 543–548. https://doi.org/10.1038/nbt.3198

Cimmino, F., Avitabile, M., Lasorsa, V. A., Montella, A., Pezone, L., Cantalupo, S., Visconte, F., Corrias, M. V., Iolascon, A., & Capasso, M. (2019). HIF-1 transcription activity: HIF1A driven response in normoxia and in hypoxia. *BMC Medical Genetics*, *20*(1), 37. https://doi.org/10.1186/s12881-019-0767-1

Clement, F. M., Klarenbach, S., Tonelli, M., Wiebe, N., Hemmelgarn, B., & Manns, B. J. (2010). An Economic Evaluation of Erythropoiesis-Stimulating Agents in CKD. *American Journal of Kidney Diseases*, *56*(6), 1050–1061. https://doi.org/https://doi.org/10.1053/j.ajkd.2010.07.015

Cline, M. G., Meredith, K. E., Boyer, J. T., & Burrows, B. (1989). Decline of height with age in adults in a general population sample: estimating maximum height and distinguishing birth cohort effects from actual loss of stature with aging. *Human Biology*, *61*(3), 415–425.

Cohen, J. C., Boerwinkle, E., Mosley, T. H., & Hobbs, H. H. (2006). Sequence Variations in PCSK9, Low LDL, and Protection against Coronary Heart Disease. *New England Journal of Medicine*, *354*(12), 1264–1272. https://doi.org/10.1056/NEJMoa054013

Collins, R. (2012). What makes UK Biobank special? *Lancet (London, England)*, *379*(9822), 1173–1174. https://doi.org/10.1016/S0140-6736(12)60404-8

Concordet, J.-P., & Haeussler, M. (2018). CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Research*, *46*(W1), W242–W245. https://doi.org/10.1093/nar/gky354

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*(1), 13. https://doi.org/10.1186/s13059-016-0881-8

Cong, L., Ran, A. F., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Xuebing, W., Wenyan, J., Luciano, M. A., & Zhang, F. (2013). Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science*, *339*(6121), 819–823. https://doi.org/10.1126/science.1231143

Conroy, M., Sellors, J., Effingham, M., Littlejohns, T. J., Boultwood, C., Gillions, L., Sudlow, C. L. M., Collins, R., & Allen, N. E. (2019). The advantages of UK Biobank's open-access strategy for health research. *Journal of Internal Medicine*, *286*(4), 389–397. https://doi.org/10.1111/joim.12955

Console, L., Scalise, M., Giangregorio, N., Tonazzi, A., Barile, M., & Indiveri, C. (2020). The Link Between the Mitochondrial Fatty Acid Oxidation Derangement and Kidney Injury . In *Frontiers in Physiology* (Vol. 11, p. 794). https://www.frontiersin.org/article/10.3389/fphys.2020.00794

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., & Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nature Reviews Genetics*, *10*(3), 184–194. https://doi.org/10.1038/nrg2537

Courtney, D. G., Moore, J. E., Atkinson, S. D., Maurizi, E., Allen, E. H. A., Pedrioli, D. M. L., McLean, W. H. I., Nesbit, M. A., & Moore, C. B. T. (2016). CRISPR/Cas9 DNA cleavage at SNP-derived PAM enables both in vitro and in vivo KRT12 mutation-specific targeting. *Gene Therapy*, *23*(1), 108–112. https://doi.org/10.1038/gt.2015.82

Couser, W. G., Remuzzi, G., Mendis, S., & Tonelli, M. (2011). The contribution of chronic kidney disease to the global burden of major noncommunicable diseases. *Kidney International*, *80*(12), 1258–1270. https://doi.org/https://doi.org/10.1038/ki.2011.368

Damman, J., Bloks, V. W., Daha, M. R., J. Van Der Most, P., Sanjabi, B., Van Der Vlies, P., Snieder, H., Ploeg, R. J., Krikke, C., Leuvenink, H. G. D., & Seelen, M. A. (2015). Hypoxia and complement-and-coagulation pathways in the deceased organ donor as the major target for intervention to improve renal allograft outcome. *Transplantation*, *99*(6), 1293–1300. https://doi.org/10.1097/TP.0000000000000500

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P.-R., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., … Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*(10), 1284–1287. https://doi.org/10.1038/ng.3656

Davey Smith, G., & Ebrahim, S. (2003). 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?*. *International Journal of Epidemiology*, *32*(1), 1–22. https://doi.org/10.1093/ije/dyg070

Davey Smith, G., & Hemani, G. (2014). Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics*, *23*(R1), R89–R98. https://doi.org/10.1093/hmg/ddu328

Davies, N. M., Holmes, M. V., & Davey Smith, G. (2018). Reading Mendelian randomisation studies: A guide, glossary, and checklist for clinicians. *BMJ (Online)*, *362*. https://doi.org/10.1136/bmj.k601

Deaton, A. M., Parker, M. M., Ward, L. D., Flynn-Carroll, A. O., BonDurant, L., Hinkle, G., Akbari, P., Lotta, L. A., Abecasis, G., Baras, A., Cantor, M., Coppola, G., Economides, A., Lotta, L. A., Overton, J. D., Reid, J. G., Shuldiner, A., Karalis, K., Deubler, A., … Personnel, G. (2021). Gene-level analysis of rare variants in 379,066 whole exome sequences identifies an association of GIGYF1 loss of function with type 2 diabetes. *Scientific Reports*, *11*(1), 21565. https://doi.org/10.1038/s41598-021-99091-5

Denny, J. C., Bastarache, L., & Roden, D. M. (2016). Phenome-Wide Association Studies as a Tool to Advance Precision Medicine. *Annual Review of Genomics and Human Genetics*, *17*(1), 353–373. https://doi.org/10.1146/annurev-genom-090314-024956

Dhillon, S. (2020). Daprodustat: First Approval. *Drugs*, *80*(14), 1491–1497. https://doi.org/10.1007/s40265-020-01384-y

Di Lullo, L., House, A., Gorini, A., Santoboni, A., Russo, D., & Ronco, C. (2015). Chronic kidney disease and cardiovascular complications. *Heart Failure Reviews*, *20*(3), 259–272. https://doi.org/10.1007/s10741-014-9460-9

DiMasi, J. A., Grabowski, H. G., & Hansen, R. W. (2016). Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics*, *47*, 20–33. https://doi.org/https://doi.org/10.1016/j.jhealeco.2016.01.012

Ding, S., Wu, X., Li, G., Han, M., Zhuang, Y., & Xu, T. (2005). Efficient Transposition of the piggyBac (PB) Transposon in Mammalian Cells and Mice. *Cell*, *122*(3), 473–483. https://doi.org/https://doi.org/10.1016/j.cell.2005.07.013

Diogo, D., Tian, C., Franklin, C. S., Alanne-Kinnunen, M., March, M., Spencer, C. C. A., Vangjeli, C., Weale, M. E., Mattsson, H., Kilpeläinen, E., Sleiman, P. M. A., Reilly, D. F., McElwee, J., Maranville, J. C., Chatterjee, A. K., Bhandari, A., Nguyen, K.-D. H., Estrada, K., Reeve, M.-P., … Runz, H. (2018). Phenome-wide association studies across large population cohorts support drug target validation. *Nature Communications*, *9*(1), 4285. https://doi.org/10.1038/s41467-018-06540-3

Dixon, A. L., Liang, L., Moffatt, M. F., Chen, W., Heath, S., Wong, K. C. C., Taylor, J., Burnett, E., Gut, I., Farrall, M., Lathrop, G. M., Abecasis, G. R., & Cookson, W. O. C. (2007). A genome-wide association study of global gene expression. *Nature Genetics*, *39*(10), 1202–1207. https://doi.org/10.1038/ng2109

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., Virgin, H. W., Listgarten, J., & Root, D. E. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology*, *34*(2), 184–191. https://doi.org/10.1038/nbt.3437

Doench, J. G., Hartenian, E., Graham, D. B., Tothova, Z., Hegde, M., Smith, I., Sullender, M., Ebert, B. L., Xavier, R. J., & Root, D. E. (2014). Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nature Biotechnology*, *32*(12), 1262–1267. https://doi.org/10.1038/nbt.3026

Doudna, J. A., & Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. *Science*, *346*(6213), 1258096. https://doi.org/10.1126/science.1258096

Dowden, H., & Munro, J. (2019). Trends in clinical success rates and therapeutic focus. In

Nature reviews. Drug discovery (Vol. 18, Issue 7, pp. 495–496). https://doi.org/10.1038/d41573-019-00074-z

Duarte, S., Woll, P. S., Buza-Vidas, N., Chin, D. W. L., Boukarabila, H., Luís, T. C., Stenson, L., Bouriez-Jones, T., Ferry, H., Mead, A. J., Atkinson, D., Jin, S., Clark, S.-A., Wu, B., Repapi, E., Gray, N., Taylor, S., Mutvei, A. P., Tsoi, Y. L., … Jacobsen, S. E. W. (2018). Canonical Notch signaling is dispensable for adult steady-state and stress myelo-erythropoiesis. *Blood*, *131*(15), 1712–1719. https://doi.org/10.1182/blood-2017-06-788505

Dudbridge, F., & Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology*, *32*(3), 227–234. https://doi.org/10.1002/gepi.20297

Eckardt, K.-U., Agarwal, R., Aswad, A., Awad, A., Block, G. A., Bacci, M. R., Farag, Y. M. K., Fishbane, S., Hubert, H., Jardine, A., Khawaja, Z., Koury, M. J., Maroni, B. J., Matsushita, K., McCullough, P. A., Lewis, E. F., Luo, W., Parfrey, P. S., Pergola, P., … Chertow, G. M. (2021). Safety and Efficacy of Vadadustat for Anemia in Patients Undergoing Dialysis. *New England Journal of Medicine*, *384*(17), 1601–1612. https://doi.org/10.1056/NEJMoa2025956

Eckardt, K. U. (1996). Erythropoietin production in liver and kidneys. *Current Opinion in Nephrology and Hypertension*, *5*(1), 28–34. https://doi.org/10.1097/00041552-199601000-00007

Eichler, E. E. (2019). Genetic Variation, Comparative Genomics, and the Diagnosis of Disease. *The New England Journal of Medicine*, *381*(1), 64–74. https://doi.org/10.1056/NEJMra1809315

Eknoyan, G., Lameire, N., Eckardt, K., Kasiske, B., Wheeler, D., Levin, A., Stevens, P. E., Bilous, R. W., Lamb, E. J., & Coresh, J. (2013). KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney Int*, *3*(1), 5–14.

Etheridge, A. S., Gallins, P. J., Jima, D., Broadaway, K. A., Ratain, M. J., Schuetz, E., Schadt, E., Schroder, A., Molony, C., Zhou, Y., Mohlke, K. L., Wright, F. A., & Innocenti, F. (2020). A New Liver Expression Quantitative Trait Locus Map From 1,183 Individuals Provides Evidence for Novel Expression Quantitative Trait Loci of Drug Response, Metabolic, and Sex-Biased Phenotypes. *Clinical Pharmacology and Therapeutics*, *107*(6), 1383–1393. https://doi.org/10.1002/cpt.1751

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics (Oxford, England)*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Fadista, J., Manning, A. K., Florez, J. C., & Groop, L. (2016). The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *European Journal of Human Genetics*, *24*(8), 1202–1205. https://doi.org/10.1038/ejhg.2015.269

Fan, Y., Fu, Y.-Y., Chen, Z., Hu, Y.-Y., & Shen, J. (2016). Gene-gene interaction of erythropoietin gene polymorphisms and diabetic retinopathy in Chinese Han. *Experimental Biology and Medicine (Maywood, N.J.)*, *241*(14), 1524–1530. https://doi.org/10.1177/1535370216645210

Ference, B. A., Holmes, M. V, & Smith, G. D. (2021). Using Mendelian Randomization to Improve the Design of Randomized Trials. *Cold Spring Harbor Perspectives in Medicine* , *11*(7). https://doi.org/10.1101/cshperspect.a040980

Ference, B. A., Robinson, J. G., Brook, R. D., Catapano, A. L., Chapman, M. J., Neff, D. R., Voros, S., Giugliano, R. P., Davey Smith, G., Fazio, S., & Sabatine, M. S. (2016). Variation in PCSK9 and HMGCR and Risk of Cardiovascular Disease and Diabetes. *New England Journal of Medicine*, *375*(22), 2144–2153. https://doi.org/10.1056/NEJMoa1604304

Ferrucci, L., Bandinelli, S., Benvenuti, E., Di Iorio, A., Macchi, C., Harris, T. B., & Guralnik, J. M. (2000). Subsystems contributing to the decline in ability to walk: bridging the gap between epidemiology and geriatric practice in the InCHIANTI study. *Journal of the American Geriatrics Society*, *48*(12), 1618–1625. https://doi.org/10.1111/j.1532-5415.2000.tb03873.x

Fishbane, S., El-Shahawy, M. A., Pecoits-Filho, R., Van, B. P., Houser, M. T., Frison, L., Little, D. J., Guzman, N. J., & Pergola, P. E. (2021). Roxadustat for Treating Anemia in Patients with CKD Not on Dialysis: Results from a Randomized Phase 3 Study. *Journal of the American Society of Nephrology*, *32*(3), 737–755. https://doi.org/10.1681/ASN.2020081150

Fishbane, S., & Spinowitz, B. (2018). Update on Anemia in ESRD and Earlier Stages of CKD: Core Curriculum 2018. *American Journal of Kidney Diseases : The Official Journal of the National Kidney Foundation*, *71*(3), 423–435. https://doi.org/10.1053/j.ajkd.2017.09.026

Fisher, J. W., Koury, S., Ducey, T., & Mendel, S. (1996). Erythropoietin production by interstitial cells of hypoxic monkey kidneys. *British Journal of Haematology*, *95*(1), 27–32. https://doi.org/10.1046/j.1365-2141.1996.d01-1864.x

Flister, M. J., Tsaih, S.-W., O'Meara, C. C., Endres, B., Hoffman, M. J., Geurts, A. M., Dwinell, M. R., Lazar, J., Jacob, H. J., & Moreno, C. (2013). Identifying multiple causative genes at a single GWAS locus. *Genome Research*, *23*(12), 1996–2002. https://doi.org/10.1101/gr.160283.113

Forejtnikovà, H., Vieillevoye, M., Zermati, Y., Lambert, M., Pellegrino, R. M., Guihard, S., Gaudry, M., Camaschella, C., Lacombe, C., Roetto, A., Mayeux, P., & Verdier, F. (2010). Transferrin receptor 2 is a component of the erythropoietin receptor complex and is required for efficient erythropoiesis. *Blood*, *116*(24), 5357–5367. https://doi.org/10.1182/blood-2010-04-281360

Franklin, B. H. (2013). Erythropoietin. *Cold Spring Harbor Perspectives in Medicine*, *3*(3), a011619–a011619. https://doi.org/10.1101/cshperspect.a011619

Frayling, T.M., & Stoneman, C. E. (2018). Mendelian randomisation in type 2 diabetes and coronary artery disease. *Current Opinion in Genetics and Development*, *50*. https://doi.org/10.1016/j.gde.2018.05.010

Frayling, Timothy M, Beaumont, R. N., Jones, S. E., Yaghootkar, H., Tuke, M. A., Ruth, K. S., Casanova, F., West, B., Locke, J., Sharp, S., Ji, Y., Thompson, W., Harrison, J., Etheridge, A. S., Gallins, P. J., Jima, D., Wright, F., Zhou, Y., Innocenti, F., … Wood, A. R. (2018). A Common Allele in FGF21 Associated with Sugar Intake Is Associated with Body Shape, Lower Total Body-Fat Percentage, and Higher Blood Pressure. *Cell Reports*, *23*(2), 327–336. https://doi.org/10.1016/j.celrep.2018.03.070

Frayling, Timothy M, Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., Perry, J. R. B., Elliott, K. S., Lango, H., Rayner, N. W., Shields, B., Harries, L. W., Barrett, J. C., Ellard, S., Groves, C. J., Knight, B., Patch, A.-M., Ness, A. R., Ebrahim, S., … McCarthy, M. I. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science (New York, N.Y.)*, *316*(5826), 889–894. https://doi.org/10.1126/science.1141634

Frede, S., Freitag, P., Geuting, L., Konietzny, R., & Fandrey, J. (2011). Oxygen-regulated expression of the erythropoietin gene in the human renal cell line REPC. *Blood*, *117*(18), 4905–4914. https://doi.org/https://doi.org/10.1182/blood-2010-07-298083

Friedmann, B., Frese, F., Menold, E., Kauper, F., Jost, J., & Bärtsch, P. (2005). Individual variation in the erythropoietic response to altitude training in elite junior swimmers. *British Journal of Sports Medicine*, *39*(3), 148–153. https://doi.org/10.1136/bjsm.2003.011387

Fu, Y., Foden, J. A., Khayter, C., Maeder, M. L., Reyon, D., Joung, J. K., & Sander, J. D. (2013). High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nature Biotechnology*, *31*(9), 822–826. https://doi.org/10.1038/nbt.2623

Fuchs, F. D., & Whelton, P. K. (2020). High Blood Pressure and Cardiovascular Disease. *Hypertension*, *75*(2), 285–292. https://doi.org/10.1161/HYPERTENSIONAHA.119.14240

Fuller, C. W., Middendorf, L. R., Benner, S. A., Church, G. M., Harris, T., Huang, X., Jovanovich, S. B., Nelson, J. R., Schloss, J. A., Schwartz, D. C., & Vezenov, D. V. (2009). The challenges of sequencing by synthesis. *Nature Biotechnology*, *27*(11), 1013–1023. https://doi.org/10.1038/nbt.1585

Gabriel, R., Lombardo, A., Arens, A., Miller, J. C., Genovese, P., Kaeppel, C., Nowrouzi, A., Bartholomae, C. C., Wang, J., Friedman, G., Holmes, M. C., Gregory, P. D., Glimm, H., Schmidt, M., Naldini, L., & von Kalle, C. (2011). An unbiased genome-wide analysis of zinc-finger nuclease specificity. *Nature Biotechnology*, *29*(9), 816–823. https://doi.org/10.1038/nbt.1948

Gaj, T., Gersbach, C. A., & Barbas 3rd, C. F. (2013). ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in Biotechnology*, *31*(7), 397–405. https://doi.org/10.1016/j.tibtech.2013.04.004

Gaj, T., Sirk, S. J., Shui, S.-L., & Liu, J. (2016). Genome-Editing Technologies: Principles and Applications. *Cold Spring Harbor Perspectives in Biology*, *8*(12), a023754. https://doi.org/10.1101/cshperspect.a023754

Gallagher, M. D., & Chen-Plotkin, A. S. (2018). The Post-GWAS Era: From Association to Function. *The American Journal of Human Genetics*, *102*(5), 717–730. https://doi.org/https://doi.org/10.1016/j.ajhg.2018.04.002

Gardie, B., Percy, M. J., Hoogewijs, D., Chowdhury, R., Bento, C., Arsenault, P. R., Richard, S., Almeida, H., Ewing, J., Lambert, F., McMullin, M. F., Schofield, C. J., & Lee, F. S. (2014). The role of PHD2 mutations in the pathogenesis of erythrocytosis. *Hypoxia (Auckland, N.Z.)*, *2*, 71–90. https://doi.org/10.2147/HP.S54455

Garimella, P. S., Katz, R., Patel, K. V, Kritchevsky, S. B., Parikh, C. R., Ix, J. H., Fried, L. F., Newman, A. B., Shlipak, M. G., Harris, T. B., & Sarnak, M. J. (2016). Association of Serum Erythropoietin With Cardiovascular Events, Kidney Function Decline, and Mortality. *Circulation: Heart Failure*, *9*(1), e002124. https://doi.org/10.1161/CIRCHEARTFAILURE.115.002124

Gazzin, S., Vitek, L., Watchko, J., Shapiro, S. M., & Tiribelli, C. (2016). A Novel Perspective on the Biology of Bilirubin in Health and Disease. *Trends in Molecular Medicine*, *22*(9), 758–768. https://doi.org/10.1016/j.molmed.2016.07.004

Giacomini, K. M., Yee, S. W., Mushiroda, T., Weinshilboum, R. M., Ratain, M. J., & Kubo, M. (2017). Genome-wide association studies of drug response and toxicity: an opportunity for genome medicine. *Nature Reviews. Drug Discovery*, *16*(1), 1. https://doi.org/10.1038/nrd.2016.234

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., & Plagnol, V. (2013). *Bayesian Test for Colocalisation Between Pairs of Genetic Association Studies Using Summary Statistics*. https://doi.org/10.1371/journal.pgen.1004383

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., &

Plagnol, V. (2014). Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLOS Genetics*, *10*(5), e1004383. https://doi.org/10.1371/journal.pgen.1004383

Gilbert, L. A., Larson, M. H., Morsut, L., Liu, Z., Brar, G. A., Torres, S. E., Stern-Ginossar, N., Brandman, O., Whitehead, E. H., Doudna, J. A., Lim, W. A., Weissman, J. S., & Qi, L. S. (2013). CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell*, *154*(2), 442–451. https://doi.org/10.1016/j.cell.2013.06.044

Gill, D., Georgakis, M. K., Koskeridis, F., Jiang, L., Feng, Q., Wei, W.-Q., Theodoratou, E., Elliott, P., Denny, J. C., Malik, R., Evangelou, E., Dehghan, A., Dichgans, M., & Tzoulaki, I. (2019). Use of Genetic Variants Related to Antihypertensive Drugs to Inform on Efficacy and Side Effects. *Circulation*, *140*(4), 270–279. https://doi.org/10.1161/CIRCULATIONAHA.118.038814

Goh, J. B., Wallace, D. F., Hong, W., & Subramaniam, V. N. (2015). Endofin, a novel BMP-SMAD regulator of the iron-regulatory hormone, hepcidin. *Scientific Reports*, *5*(1), 13986. https://doi.org/10.1038/srep13986

Graham, D. B., & Root, D. E. (2015). Resources for the design of CRISPR gene editing experiments. *Genome Biology*, *16*(1), 260. https://doi.org/10.1186/s13059-015-0823-x

Greenawalt, D. M., Dobrin, R., Chudin, E., Hatoum, I. J., Suver, C., Beaulaurier, J., Zhang, B., Castro, V., Zhu, J., Sieberts, S. K., Wang, S., Molony, C., Heymsfield, S. B., Kemp, D. M., Reitman, M. L., Lum, P. Y., Schadt, E. E., & Kaplan, L. M. (2011). A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort. *Genome Research*, *21*(7), 1008–1016. https://doi.org/10.1101/gr.112821.110

Grote Beverborg, N., Verweij, N., Klip, Ij. T., van der Wal, H. H., Voors, A. A., van Veldhuisen, D. J., Gansevoort, R. T., Bakker, S. J. L., van der Harst, P., & van der Meer, P. (2015). Erythropoietin in the General Population: Reference Ranges and Clinical, Biochemical and Genetic Correlates. *PLOS ONE*, *10*(4), e0125215. https://doi.org/10.1371/journal.pone.0125215

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics (Oxford, England)*, *32*(18), 2847–2849. https://doi.org/10.1093/bioinformatics/btw313

Guo, Y., Ma, J., Xiao, L., Fang, J., Li, G., Zhang, L., Xu, L., Lai, X., Pan, G., & Chen, Z. (2019). Identification of key pathways and genes in different types of chronic kidney disease based on WGCNA. *Molecular Medicine Reports*, *20*(3), 2245–2257. https://doi.org/10.3892/mmr.2019.10443

Gupta, N., & Wish, J. B. (2017). Hypoxia-Inducible Factor Prolyl Hydroxylase Inhibitors: A Potential New Treatment for Anemia in Patients With CKD. *American Journal of Kidney Diseases : The Official Journal of the National Kidney Foundation*, *69*(6), 815–826. https://doi.org/10.1053/j.ajkd.2016.12.011

Gustafsson, M. V, Zheng, X., Pereira, T., Gradin, K., Jin, S., Lundkvist, J., Ruas, J. L., Poellinger, L., Lendahl, U., & Bondesson, M. (2005). Hypoxia Requires Notch Signaling to Maintain the Undifferentiated Cell State. *Developmental Cell*, *9*(5), 617–628. https://doi.org/10.1016/j.devcel.2005.09.010

Haase, V. H. (2010). Hypoxic regulation of erythropoiesis and iron metabolism. *American Journal of Physiology. Renal Physiology*, *299*(1), F1–F13. https://doi.org/10.1152/ajprenal.00174.2010

Haase, V. H. (2013). Regulation of erythropoiesis by hypoxia-inducible factors. *Blood Reviews*, *27*(1), 41–53. https://doi.org/10.1016/j.blre.2012.12.003

Halldorsson, B. V, Eggertsson, H. P., Moore, K. H. S., Hauswedell, H., Eiriksson, O., Ulfarsson, M. O., Palsson, G., Hardarson, M. T., Oddsson, A., Jensson, B. O., Kristmundsdottir, S., Sigurpalsdottir, B. D., Stefansson, O. A., Beyter, D., Holley, G., Tragante, V., Gylfason, A., Olason, P. I., Zink, F., … Stefansson, K. (2021). The sequences of 150,119 genomes in the UK biobank. *BioRxiv*. https://doi.org/10.1101/2021.11.16.468246

Hannon, E., Weedon, M., Bray, N., O'Donovan, M., & Mill, J. (2017). Pleiotropic Effects of Trait-Associated Genetic Variation on DNA Methylation: Utility for Refining GWAS Loci. *American Journal of Human Genetics*, *100*(6), 954–959. https://doi.org/10.1016/j.ajhg.2017.04.013

Harris, R., Bradburn, M., Deeks, J., Harbord, R., Altman, D., & Sterne, J. (2008). metan: fixed- and random-effects meta-analysis. *Stata Journal*, *8*(1), 3–28. http://www.stata-journal.com/article.html?article=sbe24_2

Hartwig, F. P., Davey Smith, G., & Bowden, J. (2017). Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *International Journal of Epidemiology*, *46*(6), 1985–1998. https://doi.org/10.1093/ije/dyx102

Hebbring, S. J. (2014). The challenges, advantages and future of phenome-wide association studies. *Immunology*, *141*(2), 157–165.

Heigwer, F., Kerr, G., & Boutros, M. (2014). E-CRISP: fast CRISPR target site identification. *Nature Methods*, *11*(2), 122–123. https://doi.org/10.1038/nmeth.2812

Hemani, G., Bowden, J., Haycock, P., Zheng, J., Davis, O., Flach, P., Gaunt, T., & Smith, G. D. (2017). Automating Mendelian randomization through machine learning to construct a putative causal map of the human phenome. *BioRxiv*. https://doi.org/10.1101/173682

Hemani, G., Tilling, K., & Davey Smith, G. (2017). Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLOS Genetics*, *13*(11), e1007081. https://doi.org/10.1371/journal.pgen.1007081

Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., Langdon, R., Tan, V. Y., Yarmolinsky, J., Shihab, H. A., Timpson, N. J., Evans, D. M., Relton, C., Martin, R. M., Davey Smith, G., Gaunt, T. R., & Haycock, P. C. (2018). The MR-Base platform supports systematic causal inference across the human phenome. *ELife*, *7*, e34408. https://doi.org/10.7554/eLife.34408

Hendriks, W. T., Jiang, X., Daheron, L., & Cowan, C. A. (2015). TALEN- and CRISPR/Cas9-Mediated Gene Editing in Human Pluripotent Stem Cells Using Lipid-Based Transfection. *Current Protocols in Stem Cell Biology*, *34*, 5B.3.1-5B.3.25. https://doi.org/10.1002/9780470151808.sc05b03s34

Hernández, C. C., Burgos, C. F., Gajardo, A. H., Silva-Grecchi, T., Gavilan, J., Toledo, J. R., & Fuentealba, J. (2017). Neuroprotective effects of erythropoietin on neurodegenerative and ischemic brain diseases: the role of erythropoietin receptor. *Neural Regeneration Research*, *12*(9), 1381–1389. https://doi.org/10.4103/1673-5374.215240

Hernandez, D. G., Nalls, M. A., Moore, M., Chong, S., Dillman, A., Trabzuni, D., Gibbs, J. R., Ryten, M., Arepalli, S., Weale, M. E., Zonderman, A. B., Troncoso, J., O'Brien, R., Walker, R., Smith, C., Bandinelli, S., Traynor, B. J., Hardy, J., Singleton, A. B., & Cookson, M. R. (2012). Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain. *Neurobiology of Disease*, *47*(1), 20–28. https://doi.org/https://doi.org/10.1016/j.nbd.2012.03.020

Heuberger, J. A. A. C., Cohen Tervaert, J. M., Schepers, F. M. L., Vliegenthart, A. D. B., Rotmans, J. I., Daniels, J. M. A., Burggraaf, J., & Cohen, A. F. (2013). Erythropoietin doping in cycling: lack of evidence for efficacy and a negative risk-benefit. *British Journal of Clinical Pharmacology*, *75*(6), 1406–1421. https://doi.org/10.1111/bcp.12034

Hill, N. R., Fatoba, S. T., Oke, J. L., Hirst, J. A., O'Callaghan, C. A., Lasserson, D. S., & Hobbs, F. D. R. (2016). Global Prevalence of Chronic Kidney Disease – A Systematic Review and Meta-Analysis. *PLOS ONE*, *11*(7), e0158765. https://doi.org/10.1371/journal.pone.0158765

Hirschhorn, J. N. (2009). Genomewide association studies--illuminating biologic pathways. *The New England Journal of Medicine*, *360*(17), 1699–1701. https://doi.org/10.1056/NEJMp0808934

Hockemeyer, D., Soldner, F., Beard, C., Gao, Q., Mitalipova, M., DeKelver, R. C., Katibah, G. E., Amora, R., Boydston, E. A., Zeitler, B., Meng, X., Miller, J. C., Zhang, L., Rebar, E. J., Gregory, P. D., Urnov, F. D., & Jaenisch, R. (2009). Efficient targeting of expressed and silent genes in human ESCs and iPSCs using zinc-finger nucleases. *Nature Biotechnology*, *27*(9), 851–857. https://doi.org/10.1038/nbt.1562

Holdstock, L., Cizman, B., Meadowcroft, A. M., Biswas, N., Johnson, B. M., Jones, D., Kim, S. G., Zeig, S., Lepore, J. J., & Cobitz, A. R. (2019). Daprodustat for anemia: A 24-week, open-label, randomized controlled trial in participants with chronic kidney disease. *Clinical Kidney Journal*, *12*(1), 129–138. https://doi.org/10.1093/ckj/sfy013

Holzner, L. M. W., & Murray, A. J. (2021). Hypoxia-Inducible Factors as Key Players in the Pathogenesis of Non-alcoholic Fatty Liver Disease and Non-alcoholic Steatohepatitis . In *Frontiers in Medicine* (Vol. 8). https://www.frontiersin.org/article/10.3389/fmed.2021.753268

Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B., & Eskin, E. (2014). Identifying Causal Variants at Loci with Multiple Signals of Association. *Genetics*, *198*(2), 497–508. https://doi.org/10.1534/genetics.114.167908

Hormozdiari, F., van de Bunt, M., Segrè, A. V., Li, X., Joo, J. W. J., Bilow, M., Sul, J. H., Sankararaman, S., Pasaniuc, B., & Eskin, E. (2016). Colocalization of GWAS and eQTL Signals Detects Target Genes. *The American Journal of Human Genetics*, *99*(6), 1245–1260. https://doi.org/https://doi.org/10.1016/j.ajhg.2016.10.003

Horscroft, J. A., Kotwica, A. O., Laner, V., West, J. A., Hennis, P. J., Levett, D. Z. H., Howard, D. J., Fernandez, B. O., Burgess, S. L., Ament, Z., Gilbert-Kawai, E. T., Vercueil, A., Landis, B. D., Mitchell, K., Mythen, M. G., Branco, C., Johnson, R. S., Feelisch, M., Montgomery, H. E., … Murray, A. J. (2017). Metabolic basis to Sherpa altitude adaptation. *Proceedings of the National Academy of Sciences*, *114*(24), 6382 LP – 6387. https://doi.org/10.1073/pnas.1700527114

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., Davidson, C., Dodiya, K., El Houdaigui, B., Fatima, R., … Flicek, P. (2021). Ensembl 2021. *Nucleic Acids Research*, *49*(D1), D884–D891. https://doi.org/10.1093/nar/gkaa942

Howie, B., Marchini, J., & Stephens, M. (2011). Genotype Imputation with Thousands of Genomes. *G3: Genes, Genomes, Genetics*, *1*(6), 457–470. https://doi.org/10.1534/g3.111.001198

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLOS Genetics*, *5*(6), e1000529. https://doi.org/10.1371/journal.pgen.1000529

Hsu, C., McCulloch, C. E., & Curhan, G. C. (2002). Iron Status and Hemoglobin Level in Chronic Renal Insufficiency. *Journal of the American Society of Nephrology*, *13*(11), 2783–2786. https://doi.org/10.1097/01.ASN.0000034200.82278.DC

Hsu, P. D., Lander, E. S., & Zhang, F. (2014). Development and applications of CRISPR-Cas9 for genome engineering. *Cell*, *157*(6), 1262–1278. https://doi.org/10.1016/j.cell.2014.05.010

Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., Li, Y., Fine, E. J., Wu, X., Shalem, O., Cradick, T. J., Marraffini, L. A., Bao, G., & Zhang, F. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature Biotechnology*, *31*(9), 827–832. https://doi.org/10.1038/nbt.2647

Huang, C., Yang, D., Ye, G. W., Powell, C. A., & Guo, P. (2021). Vascular Notch Signaling in Stress Hematopoiesis . In *Frontiers in Cell and Developmental Biology* (Vol. 8, p. 1693). https://www.frontiersin.org/article/10.3389/fcell.2020.606448

Huang, Y., Pettitt, S. J., Guo, G., Liu, G., Li, M. A., Yang, F., & Bradley, A. (2012). Isolation of homozygous mutant mouse embryonic stem cells using a dual selection system. *Nucleic Acids Research*, *40*(3), e21–e21. https://doi.org/10.1093/nar/gkr908

Hurle, M. R., Nelson, M. R., Agarwal, P., & Cardon, L. R. (2016). Impact of genetically supported target selection on R&D productivity. *Nature Reviews Drug Discovery*, *15*(9), 596–597. https://doi.org/10.1038/nrd.2016.164

Hutchinson, A., Asimit, J., & Wallace, C. (2020). Fine-mapping genetic associations. *Human Molecular Genetics*, *29*(R1), R81–R88. https://doi.org/10.1093/hmg/ddaa148

Imeri, F., Nolan, K. A., Bapst, A. M., Santambrogio, S., Abreu-Rodríguez, I., Spielmann, P., Pfundstein, S., Libertini, S., Crowther, L., Orlando, I. M. C., Dahl, S. L., Keodara, A., Kuo, W., Kurtcuoglu, V., Scholz, C. C., Qi, W., Hummler, E., Hoogewijs, D., & Wenger, R. H. (2019). Generation of renal Epo-producing cell lines by conditional gene tagging reveals rapid HIF-2 driven Epo kinetics, cell autonomous feedback regulation, and a telocyte phenotype. *Kidney International*, *95*(2), 375–387. https://doi.org/10.1016/j.kint.2018.08.043

Innocenti, F., Cooper, G. M., Stanaway, I. B., Gamazon, E. R., Smith, J. D., Mirkov, S., Ramirez, J., Liu, W., Lin, Y. S., Moloney, C., Aldred, S. F., Trinklein, N. D., Schuetz, E., Nickerson, D. A., Thummel, K. E., Rieder, M. J., Rettie, A. E., Ratain, M. J., Cox, N. J., & Brown, C. D. (2011). Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS Genetics*, *7*(5), e1002078. https://doi.org/10.1371/journal.pgen.1002078

Inoue, F., & Ahituv, N. (2015). Decoding enhancers using massively parallel reporter assays. *Genomics*, *106*(3), 159–164.

Ishida, K., Xu, H., Sasakawa, N., Lung, M. S. Y., Kudryashev, J. A., Gee, P., & Hotta, A. (2018). Site-specific randomization of the endogenous genome by a regulatable CRISPR-Cas9 piggyBac system in human cells. *Scientific Reports*, *8*(1), 310. https://doi.org/10.1038/s41598-017-18568-4

Ivics, Z., Li, M. A., Mátés, L., Boeke, J. D., Nagy, A., Bradley, A., & Izsvák, Z. (2009). Transposon-mediated genome manipulation in vertebrates. *Nature Methods*, *6*(6), 415–422. https://doi.org/10.1038/nmeth.1332

Iwaki, T., & Umemura, K. (2011). A single plasmid transfection that offers a significant advantage associated with puromycin selection, fluorescence-assisted cell sorting, and doxycycline-inducible protein expression in mammalian cells. *Cytotechnology*, *63*(4), 337–343. https://doi.org/10.1007/s10616-011-9357-6

Jankowski, J., Floege, J., Fliser, D., Böhm, M., & Marx, N. (2021). Cardiovascular Disease in Chronic Kidney Disease. *Circulation*, *143*(11), 1157–1172. https://doi.org/10.1161/CIRCULATIONAHA.120.050686

Jayaram, N., Usvyat, D., & R. Martin, A. C. (2016). Evaluating tools for transcription factor binding site prediction. *BMC Bioinformatics*, *17*(1), 547. https://doi.org/10.1186/s12859-016-1298-9

Jelkmann, W. (2011). Regulation of erythropoietin production. *The Journal of Physiology*, *589*(Pt 6), 1251–1258. https://doi.org/10.1113/jphysiol.2010.195057

Jelkmann, W. (2013). Physiology and pharmacology of erythropoietin. *Transfusion Medicine and Hemotherapy : Offizielles Organ Der Deutschen Gesellschaft Fur Transfusionsmedizin Und Immunhamatologie*, *40*(5), 302–309. https://doi.org/10.1159/000356193

Jha, V., Garcia-Garcia, G., Iseki, K., Li, Z., Naicker, S., Plattner, B., Saran, R., Wang, A. Y.-M., & Yang, C.-W. (2013). Chronic kidney disease: global dimension and perspectives. *Lancet (London, England)*, *382*(9888), 260–272. https://doi.org/10.1016/S0140-6736(13)60687-X

Jiang, F., & Doudna, J. A. (2017). CRISPR–Cas9 Structures and Mechanisms. *Annual Review of Biophysics*, *46*(1), 505–529. https://doi.org/10.1146/annurev-biophys-062215-010822

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (New York, N.Y.)*, *337*(6096), 816–821. https://doi.org/10.1126/science.1225829

Jørgensen, A. B., Frikke-Schmidt, R., Nordestgaard, B. G., & Tybjærg-Hansen, A. (2014). Loss-of-Function Mutations in APOC3 and Risk of Ischemic Vascular Disease. *New England Journal of Medicine*, *371*(1), 32–41. https://doi.org/10.1056/NEJMoa1308027

Joung, J. K., & Sander, J. D. (2013). TALENs: a widely applicable technology for targeted genome editing. *Nature Reviews. Molecular Cell Biology*, *14*(1), 49–55. https://doi.org/10.1038/nrm3486

Junk, A. K., Mammis, A., Savitz, S. I., Singh, M., Roth, S., Malhotra, S., Rosenbaum, P. S., Cerami, A., Brines, M., & Rosenbaum, D. M. (2002). Erythropoietin administration protects retinal neurons from acute ischemia-reperfusion injury. *Proceedings of the National Academy of Sciences*, *99*(16), 10659–10664. https://doi.org/10.1073/pnas.152321399

Kaitin, K. I. (2010). Deconstructing the drug development process: the new face of innovation. *Clinical Pharmacology and Therapeutics*, *87*(3), 356–361. https://doi.org/10.1038/clpt.2009.293

Kanai, H., Nangaku, M., Nagai, R., Okuda, N., Kurata, K., Nagakubo, T., Endo, Y., & Cobitz, A. (2021). Efficacy and safety of daprodustat in Japanese peritoneal dialysis patients. *Therapeutic Apheresis and Dialysis : Official Peer-Reviewed Journal of the International Society for Apheresis, the Japanese Society for Apheresis, the Japanese Society for Dialysis Therapy*. https://doi.org/10.1111/1744-9987.13686

Kang, S.-H., Lee, W., An, J.-H., Lee, J.-H., Kim, Y.-H., Kim, H., Oh, Y., Park, Y.-H., Jin, Y. B., Jun, B.-H., Hur, J. K., Kim, S.-U., & Lee, S. H. (2020). Prediction-based highly sensitive CRISPR off-target validation using target-specific DNA enrichment. *Nature Communications*, *11*(1), 3596. https://doi.org/10.1038/s41467-020-17418-8

Kaplan, J. M., Sharma, N., & Dikdan, S. (2018). Hypoxia-Inducible Factor and Its Role in the Management of Anemia in Chronic Kidney Disease. *International Journal of Molecular Sciences*, *19*(2). https://doi.org/10.3390/ijms19020389

Kästner, A., Grube, S., El-Kordi, A., Stepniak, B., Friedrichs, H., Sargin, D., Schwitulla, J., Begemann, M., Giegling, I., Miskowiak, K. W., Sperling, S., Hannke, K., Ramin, A., Heinrich, R., Gefeller, O., Nave, K.-A., Rujescu, D., & Ehrenreich, H. (2012). Common variants of the genes encoding erythropoietin and its receptor modulate cognitive performance in schizophrenia. *Molecular Medicine (Cambridge, Mass.)*, *18*(1), 1029–1040. https://doi.org/10.2119/molmed.2012.00190

KDOQI. (2006). KDOQI Clinical Practice Guidelines and Clinical Practice Recommendations for Anemia in Chronic Kidney Disease. *American Journal of Kidney Diseases : The Official Journal of the National Kidney Foundation*, *47*(5 Suppl 3), S11-145. https://doi.org/10.1053/j.ajkd.2006.03.010

Khabour, O. F., Bani-Ahmad, M. A., & Hammash, N. M. (2012). Association between polymorphisms in erythropoietin gene and upper limit haematocrit levels among regular blood donors. *Transfusion Clinique et Biologique*, *19*(6), 353–357. https://doi.org/https://doi.org/10.1016/j.tracli.2012.10.001

Kim, D., Bae, S., Park, J., Kim, E., Kim, S., Yu, H. R., Hwang, J., Kim, J.-I., & Kim, J.-S. (2015). Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nature Methods*, *12*(3), 237–243, 1 p following 243. https://doi.org/10.1038/nmeth.3284

Kim, H., & Kim, J.-S. (2014). A guide to genome engineering with programmable nucleases. *Nature Reviews Genetics*, *15*(5), 321–334. https://doi.org/10.1038/nrg3686

King, E. A., Davis, J. W., & Degner, J. F. (2019). Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLOS Genetics*, *15*(12), e1008489. https://doi.org/10.1371/journal.pgen.1008489

Kondrashov, A., Duc Hoang, M., Smith, J. G. W., Bhagwan, J. R., Duncan, G., Mosqueira, D., Munoz, M. B., Vo, N. T. N., & Denning, C. (2018). Simplified Footprint-Free Cas9/CRISPR Editing of Cardiac-Associated Genes in Human Pluripotent Stem Cells. *Stem Cells and Development*, *27*(6), 391–404. https://doi.org/10.1089/scd.2017.0268

Kopan, R. (2012). Notch signaling. *Cold Spring Harbor Perspectives in Biology*, *4*(10), a011213. https://doi.org/10.1101/cshperspect.a011213

Koury, M. J., & Haase, V. H. (2015). Anaemia in kidney disease: harnessing hypoxia responses for therapy. *Nature Reviews. Nephrology*, *11*(7), 394–410. https://doi.org/10.1038/nrneph.2015.82

Krapf, R., & Hulter, H. N. (2009). Arterial hypertension induced by erythropoietin and erythropoiesis-stimulating agents (ESA). *Clinical Journal of the American Society of Nephrology : CJASN*, *4*(2), 470–480. https://doi.org/10.2215/CJN.05040908

Kukurba, K. R., & Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harbor Protocols*, *2015*(11), 951–969. https://doi.org/10.1101/pdb.top084970

Kuleshov, M. V, Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W., & Ma'ayan, A. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, *44*(W1), W90-7. https://doi.org/10.1093/nar/gkw377

La Gerche, A., & Brosnan, M. J. (2017). Cardiovascular Effects of Performance-Enhancing Drugs. *Circulation*, *135*(1), 89–99. https://doi.org/10.1161/CIRCULATIONAHA.116.022535

Lamb, B. M., Mercer, A. C., & Barbas, C. F. 3rd. (2013). Directed evolution of the TALE N-terminal domain for recognition of all 5' bases. *Nucleic Acids Research*, *41*(21), 9779–9785. https://doi.org/10.1093/nar/gkt754

Lamon, S., & Russell, A. (2013). The role and regulation of erythropoietin (EPO) and its receptor in skeletal muscle: how much do we really know?   . In *Frontiers in Physiology* (Vol. 4). https://www.frontiersin.org/article/10.3389/fphys.2013.00176

Lauridsen, B. K., Stender, S., Frikke-Schmidt, R., Nordestgaard, B. G., & Tybjærg-Hansen, A. (2015). Genetic variation in the cholesterol transporter NPC1L1, ischaemic vascular disease, and gallstone disease. *European Heart Journal*, *36*(25), 1601–1608. https://doi.org/10.1093/eurheartj/ehv108

Lawlor, D. A. (2016). Commentary: Two-sample Mendelian randomization: opportunities and challenges. *International Journal of Epidemiology*, *45*(3), 908–915. https://doi.org/10.1093/ije/dyw127

Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N., & Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, *27*(8), 1133–1163. https://doi.org/10.1002/sim.3034

Lawrenson, K., Li, Q., Kar, S., Seo, J.-H., Tyrer, J., Spindler, T. J., Lee, J., Chen, Y., Karst, A., Drapkin, R., Aben, K. K. H., Anton-Culver, H., Antonenkova, N., Bowtell, D., Webb, P. M., deFazio, A., Baker, H., Bandera, E. V, Bean, Y., … Group, A. O. C. S. (2015). Cis-eQTL analysis and functional validation of candidate susceptibility genes for high-grade serous ovarian cancer. *Nature Communications*, *6*(1), 8234. https://doi.org/10.1038/ncomms9234

Lee, F. S., & Percy, M. J. (2011). The HIF Pathway and Erythrocytosis. *Annual Review of Pathology: Mechanisms of Disease*, *6*(1), 165–192. https://doi.org/10.1146/annurev-pathol-011110-130321

Lee, J. W., Ko, J., Ju, C., & Eltzschig, H. K. (2019). Hypoxia signaling in human diseases and therapeutic targets. *Experimental & Molecular Medicine*, *51*(6), 1–13. https://doi.org/10.1038/s12276-019-0235-1

Leenaars, C. H. C., Kouwenaar, C., Stafleu, F. R., Bleich, A., Ritskes-Hoitinga, M., De Vries, R. B. M., & Meijboom, F. L. B. (2019). Animal to human translation: a systematic scoping review of reported concordance rates. *Journal of Translational Medicine*, *17*(1), 223. https://doi.org/10.1186/s12967-019-1976-2

Levey, A S. (2021). Defining AKD: The Spectrum of AKI, AKD, and CKD. *Nephron*. https://doi.org/10.1159/000516647

Levey, Andrew S, Coresh, J., Greene, T., Marsh, J., Stevens, L. A., Kusek, J. W., & Van Lente, F. (2007). Expressing the Modification of Diet in Renal Disease Study equation for estimating glomerular filtration rate with standardized serum creatinine values. *Clinical Chemistry*, *53*(4), 766–772. https://doi.org/10.1373/clinchem.2006.077180

Levey, Andrew S, Eckardt, K.-U., Tsukamoto, Y., Levin, A., Coresh, J., Rossert, J., De Zeeuw, D., Hostetter, T. H., Lameire, N., & Eknoyan, G. (2005). Definition and classification of chronic kidney disease: a position statement from Kidney Disease: Improving Global Outcomes (KDIGO). *Kidney International*, *67*(6), 2089–2100. https://doi.org/10.1111/j.1523-1755.2005.00365.x

Levin, A., Stevens, P. E., Bilous, R. W., Coresh, J., De Francisco, A. L. M., De Jong, P. E., Griffith, K. E., Hemmelgarn, B. R., Iseki, K., & Lamb, E. J. (2013). Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney International Supplements*, *3*(1), 1–150.

Li, H., Yang, Y., Hong, W., Huang, M., Wu, M., & Zhao, X. (2020). Applications of genome editing technology in the targeted therapy of human diseases: mechanisms, advances and prospects. *Signal Transduction and Targeted Therapy*, *5*(1), 1. https://doi.org/10.1038/s41392-019-0089-y

Li, J., Hou, R., Niu, X., Liu, R., Wang, Q., Wang, C., Li, X., Hao, Z., Yin, G., & Zhang, K. (2016). Comparison of microarray and RNA-Seq analysis of mRNA expression in dermal mesenchymal stem cells. *Biotechnology Letters*, *38*(1), 33–41. https://doi.org/10.1007/s10529-015-1963-5

Li, M. A., Pettitt, S. J., Eckert, S., Ning, Z., Rice, S., Cadiñanos, J., Yusa, K., Conte, N., & Bradley, A. (2013). The piggyBac transposon displays local and distant reintegration preferences and can cause mutations at noncanonical integration sites. *Molecular and Cellular Biology*, *33*(7), 1317–1330. https://doi.org/10.1128/MCB.00670-12

Li, X., Burnight, E. R., Cooney, A. L., Malani, N., Brady, T., Sander, J. D., Staber, J., Wheelan, S. J., Joung, J. K., McCray, P. B., Bushman, F. D., Sinn, P. L., & Craig, N. L. (2013). piggyBac transposase tools for genome engineering. *Proceedings of the National Academy of Sciences*, *110*(25), E2279--E2287. https://doi.org/10.1073/pnas.1305987110

Li, Z., Michael, I. P., Zhou, D., Nagy, A., & Rini, J. M. (2013). Simple piggyBac transposon-based mammalian cell expression system for inducible protein production. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(13), 5004–5009. https://doi.org/10.1073/pnas.1218620110

Liang, Q., Kong, J., Stalker, J., & Bradley, A. (2009). Chromosomal mobilization and reintegration of Sleeping Beauty and PiggyBac transposons. In *Genesis (New York, N.Y. : 2000)* (Vol. 47, Issue 6, pp. 404–408). https://doi.org/10.1002/dvg.20508

Liao, Y., Smyth, G. K., & Shi, W. (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research*, *41*(10), e108. https://doi.org/10.1093/nar/gkt214

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)*, *30*(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

Lichou, F., & Trynka, G. (2020). Functional studies of GWAS variants are gaining momentum. *Nature Communications*, *11*(1), 6283. https://doi.org/10.1038/s41467-020-20188-y

Lin, C. S., Lim, S. K., D'Agati, V., & Costantini, F. (1996). Differential effects of an erythropoietin receptor gene disruption on primitive and definitive erythropoiesis. *Genes & Development*, *10*(2), 154–164. https://doi.org/10.1101/gad.10.2.154

Lin, J., & Musunuru, K. (2018). From Genotype to Phenotype: A Primer on the Functional Follow-up of Genome-Wide Association Studies in Cardiovascular Disease. *Circulation. Genomic and Precision Medicine*, *11*(2), e001946. https://pubmed.ncbi.nlm.nih.gov/29915816

Lino, C. A., Harper, J. C., Carney, J. P., & Timlin, J. A. (2018). Delivering CRISPR: a review of the challenges and approaches. *Drug Delivery*, *25*(1), 1234–1257. https://doi.org/10.1080/10717544.2018.1474964

Liu, Guanqing, Zhang, Y., & Zhang, T. (2020). Computational approaches for effective CRISPR guide RNA design and evaluation. *Computational and Structural Biotechnology Journal*, *18*, 35–44. https://doi.org/https://doi.org/10.1016/j.csbj.2019.11.006

Liu, Guowen, Roy, J., & Johnson, E. A. (2006). Identification and function of hypoxia-response genes in Drosophila melanogaster. *Physiological Genomics*, *25*(1), 134–141. https://doi.org/10.1152/physiolgenomics.00262.2005

Liu, L. Y., Schaub, M. A., Sirota, M., & Butte, A. J. (2012). Sex differences in disease risk from reported genome-wide association study findings. *Human Genetics*, *131*(3), 353–364. https://doi.org/10.1007/s00439-011-1081-y

Liu, Q., Fan, D., Adah, D., Wu, Z., Liu, R., Yan, Q., Zhang, Y., Du, Z., Wang, D., Li, Y., Bao, S., & Liu, L. (2018). CRISPR/Cas9-mediated hypoxia inducible factor-1α knockout enhances the antitumor effect of transarterial embolization in hepatocellular carcinoma. *Oncol Rep*, *40*(5), 2547–2557. https://doi.org/10.3892/or.2018.6667

Liu, S., Wang, Q., Yu, X., Li, Y., Guo, Y., Liu, Z., Sun, F., Hou, W., Li, C., Wu, L., Guo, D., & Chen, S. (2018). HIV-1 inhibition in cells with CXCR4 mutant genome created by CRISPR-Cas9 and piggyBac recombinant technologies. *Scientific Reports*, *8*(1), 8573. https://doi.org/10.1038/s41598-018-26894-4

Liu, T., Jia, P., Ma, H., Reed, S. A., Luo, X., Larman, H. B., & Schultz, P. G. (2017). Construction and Screening of a Lentiviral Secretome Library. *Cell Chemical Biology*, *24*(6), 767-771.e3. https://doi.org/10.1016/j.chembiol.2017.05.017

Liu, X.-H., Kirschenbaum, A., Yao, S., Stearns, M. E., Holland, J. F., Claffey, K., & Levine, A. C. (1999). Upregulation of vascular endothelial growth factor by cobalt chloride-simulated hypoxia is mediated by persistent induction of cyclooxygenase-2 in a metastatic human prostate cancer cell line. *Clinical & Experimental Metastasis*, *17*(8), 687–694. https://doi.org/10.1023/A:1006728119549

Liu, Y., Morley, M., Brandimarto, J., Hannenhalli, S., Hu, Y., Ashley, E. A., Tang, W. H. W., Moravec, C. S., Margulies, K. B., Cappola, T. P., & Li, M. (2015). RNA-Seq identifies novel myocardial gene expression signatures of heart failure. *Genomics*, *105*(2), 83–89. https://doi.org/10.1016/j.ygeno.2014.12.002

Livak, K. J., & Schmittgen, T. D. (2001). Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method. *Methods*, *25*(4), 402–408. https://doi.org/https://doi.org/10.1006/meth.2001.1262

Locatelli, F., & Del Vecchio, L. (2003). Pure red cell aplasia secondary to treatment with erythropoietin. *Journal of Nephrology*, *16*(4), 461–466.

Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., Chasman, D. I., Ridker, P. M., Neale, B. M., Berger, B., Patterson, N., & Price, A. L. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, *47*(3), 284–290. https://doi.org/10.1038/ng.3190

Lotta, L. A., Mokrosiński, J., Mendes de Oliveira, E., Li, C., Sharp, S. J., Luan, J., Brouwers, B., Ayinampudi, V., Bowker, N., Kerrison, N., Kaimakis, V., Hoult, D., Stewart, I. D., Wheeler, E., Day, F. R., Perry, J. R. B., Langenberg, C., Wareham, N. J., & Farooqi, I. S. (2019). Human Gain-of-Function MC4R Variants Show Signaling Bias and Protect against Obesity. *Cell*, *177*(3), 597-607.e9. https://doi.org/https://doi.org/10.1016/j.cell.2019.03.044

Lotta, L. A., Sharp, S. J., Burgess, S., Perry, J. R. B., Stewart, I. D., Willems, S. M., Luan, J., Ardanaz, E., Arriola, L., Balkau, B., Boeing, H., Deloukas, P., Forouhi, N. G., Franks, P. W., Grioni, S., Kaaks, R., Key, T. J., Navarro, C., Nilsson, P. M., … Wareham, N. J. (2016). Association Between Low-Density Lipoprotein Cholesterol-Lowering Genetic Variants and Risk of Type 2 Diabetes: A Meta-analysis. *JAMA*, *316*(13), 1383–1391. https://doi.org/10.1001/jama.2016.14568

Love, M. I., Anders, S., Kim, V., & Huber, W. (2015). RNA-Seq workflow: gene-level exploratory analysis and differential expression. *F1000Research*, *4*, 1070. https://doi.org/10.12688/f1000research.7035.1

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017). Transcriptomics technologies. *PLoS Computational Biology*, *13*(5), e1005457–e1005457. https://doi.org/10.1371/journal.pcbi.1005457

Luk, C. T., Shi, S. Y., Choi, D., Cai, E. P., Schroer, S. A., & Woo, M. (2013). In Vivo Knockdown of Adipocyte Erythropoietin Receptor Does Not Alter Glucose or Energy Homeostasis. *Endocrinology*, *154*(10), 3652–3659. https://doi.org/10.1210/en.2013-1113

Lundby, C., Thomsen, J. J., Boushel, R., Koskolou, M., Warberg, J., Calbet, J. A. L., & Robach, P. (2007). Erythropoietin treatment elevates haemoglobin concentration by increasing red cell volume and depressing plasma volume. *The Journal of Physiology*, *578*(Pt 1), 309–314. https://doi.org/10.1113/jphysiol.2006.122689

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., Norville, J. E., & Church, G. M. (2013). RNA-guided human genome engineering via Cas9. *Science (New York, N.Y.)*, *339*(6121), 823–826. https://doi.org/10.1126/science.1232033

Malik, R., Chauhan, G., Traylor, M., Sargurupremraj, M., Okada, Y., Mishra, A., Rutten-Jacobs, L., Giese, A.-K., van der Laan, S. W., Gretarsdottir, S., Anderson, C. D., Chong, M., Adams, H. H. H., Ago, T., Almgren, P., Amouyel, P., Ay, H., Bartz, T. M., Benavente, O. R., … Consortium, M. (2018). Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nature Genetics*, *50*(4), 524–537. https://doi.org/10.1038/s41588-018-0058-3

Mandelboum, S., Manber, Z., Elroy-Stein, O., & Elkon, R. (2019). Recurrent functional misinterpretation of RNA-seq data caused by sample-specific gene length bias. *PLOS Biology*, *17*(11), e3000481. https://doi.org/10.1371/journal.pbio.3000481

Marignol, L., Rivera-Figueroa, K., Lynch, T., & Hollywood, D. (2013). Hypoxia, notch signalling, and prostate cancer. *Nature Reviews Urology*, *10*(7), 405–413. https://doi.org/10.1038/nrurol.2013.110

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal; Vol 17, No 1: Next Generation Sequencing Data Analysis*. https://doi.org/10.14806/ej.17.1.200

Maruyama, T., Dougan, S. K., Truttmann, M. C., Bilate, A. M., Ingram, J. R., & Ploegh, H. L. (2015). Increasing the efficiency of precise genome editing with CRISPR-Cas9 by inhibition of nonhomologous end joining. *Nature Biotechnology*, *33*(5), 538–542. https://doi.org/10.1038/nbt.3190

Masoud, G. N., & Li, W. (2015). HIF-1α pathway: role, regulation and intervention for cancer therapy. *Acta Pharmaceutica Sinica. B*, *5*(5), 378–389. https://doi.org/10.1016/j.apsb.2015.05.007

Matías-García, P. R., Wilson, R., Guo, Q., Zaghlool, S. B., Eales, J. M., Xu, X., Charchar, F. J., Dormer, J., Maalmi, H., Schlosser, P., Elhadad, M. A., Nano, J., Sharma, S., Peters, A., Fornoni, A., Mook-Kanamori, D. O., Winkelmann, J., Danesh, J., Di Angelantonio, E., … Resource, H. K. T. (2021). Plasma Proteomics of Renal Function: A Transethnic Meta-Analysis and Mendelian Randomization Study. *Journal of the American Society of Nephrology*, *32*(7), 1747 LP – 1763. https://doi.org/10.1681/ASN.2020071070

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M., Fuchsberger, C., Danecek, P., Sharp, K., Luo, Y., Sidore, C., Kwong, A., Timpson, N., Koskinen, S., Vrieze, S., Scott, L. J., Zhang, H., Mahajan, A., … Consortium, the H. R. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, *48*(10), 1279–1283. https://doi.org/10.1038/ng.3643

McGonigle, P., & Ruggeri, B. (2014). Animal models of human disease: Challenges in enabling translation. *Biochemical Pharmacology*, *87*(1), 162–171. https://doi.org/https://doi.org/10.1016/j.bcp.2013.08.006

McGuire, A. L., Gabriel, S., Tishkoff, S. A., Wonkam, A., Chakravarti, A., Furlong, E. E. M., Treutlein, B., Meissner, A., Chang, H. Y., López-Bigas, N., Segal, E., & Kim, J.-S. (2020). The road ahead in genetics and genomics. *Nature Reviews Genetics*, *21*(10), 581–596. https://doi.org/10.1038/s41576-020-0272-6

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 122. https://doi.org/10.1186/s13059-016-0974-4

Meadowcroft, A. M., Cizman, B., Holdstock, L., Biswas, N., Johnson, B. M., Jones, D., Nossuli, A. K., Lepore, J. J., Aarup, M., & Cobitz, A. R. (2019). Daprodustat for anemia: A 24-week, open-label, randomized controlled trial in participants on hemodialysis. *Clinical Kidney Journal*, *12*(1), 139–148. https://doi.org/10.1093/ckj/sfy014

Meier, I. D., Bernreuther, C., Tilling, T., Neidhardt, J., Wong, Y. W., Schulze, C., Streichert, T., & Schachner, M. (2010). Short DNA sequences inserted for gene targeting can accidentally interfere with off-target gene expression. *FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology*, *24*(6), 1714–1724. https://doi.org/10.1096/fj.09-140749

Melzer, D., Perry, J. R. B., Hernandez, D., Corsi, A.-M., Stevens, K., Rafferty, I., Lauretani, F., Murray, A., Gibbs, J. R., Paolisso, G., Rafiq, S., Simon-Sanchez, J., Lango, H., Scholz, S., Weedon, M. N., Arepalli, S., Rice, N., Washecka, N., Hurst, A., … Ferrucci, L. (2008). A Genome-Wide Association Study Identifies Protein Quantitative Trait Loci (pQTLs). *PLOS Genetics*, *4*(5), e1000072. https://doi.org/10.1371/journal.pgen.1000072

Mercher, T., Cornejo, M. G., Sears, C., Kindler, T., Moore, S. A., Maillard, I., Pear, W. S., Aster, J. C., & Gilliland, D. G. (2008). Notch signaling specifies megakaryocyte development from hematopoietic stem cells. *Cell Stem Cell*, *3*(3), 314–326. https://doi.org/10.1016/j.stem.2008.07.010

Mikhail, A., Brown, C., Williams, J. A., Mathrani, V., Shrivastava, R., Evans, J., Isaac, H., & Bhandari, S. (2017). Renal association clinical practice guideline on Anaemia of Chronic Kidney Disease. *BMC Nephrology*, *18*(1), 345. https://doi.org/10.1186/s12882-017-0688-1

Minamishima, Y. A., Moslehi, J., Bardeesy, N., Cullen, D., Bronson, R. T., & Kaelin Jr, W. G. (2008). Somatic inactivation of the PHD2 prolyl hydroxylase causes polycythemia and congestive heart failure. *Blood*, *111*(6), 3236–3244. https://doi.org/10.1182/blood-2007-10-117812

Minikel, E. V., Karczewski, K. J., Martin, H. C., Cummings, B. B., Whiffin, N., Rhodes, D., Alföldi, J., Trembath, R. C., van Heel, D. A., Daly, M. J., Team, G. A. D. P., Consortium, G. A. D., Schreiber, S. L., & MacArthur, D. G. (2020). Evaluating drug targets through human loss-of-function genetic variation. *Nature*, *581*(7809), 459–464. https://doi.org/10.1038/s41586-020-2267-z

Mo, Y., Lim, C., Watson, J. A., White, N. J., & Cooper, B. S. (2020). Non-adherence in non-inferiority trials: pitfalls and recommendations. *BMJ*, *370*. https://doi.org/10.1136/bmj.m2215

Mokry, L. E., Ahmad, O., Forgetta, V., Thanassoulis, G., & Richards, J. B. (2015). Mendelian randomisation applied to drug development in cardiovascular disease: a review. *Journal of Medical Genetics*, *52*(2), 71–79. https://doi.org/10.1136/jmedgenet-2014-102438

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, *5*(7), 621–628. https://doi.org/10.1038/nmeth.1226

Muckenthaler, M. U., Rivella, S., Hentze, M. W., & Galy, B. (2017). A Red Carpet for Iron Metabolism. *Cell*, *168*(3), 344–361. https://doi.org/10.1016/j.cell.2016.12.034

Muir, A. J., Gong, L., Johnson, S. G., Lee, M. T. M., Williams, M. S., Klein, T. E., Caudle, K. E., & Nelson, D. R. (2014). Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines for IFNL3 (IL28B) genotype and PEG interferon-α-based regimens. *Clinical Pharmacology and Therapeutics*, *95*(2), 141–146. https://doi.org/10.1038/clpt.2013.203

Muñoz-López, M., & García-Pérez, J. L. (2010). DNA transposons: nature and applications in genomics. *Current Genomics*, *11*(2), 115–128. https://doi.org/10.2174/138920210790886871

Murphy, W. G. (2014). The sex difference in haemoglobin levels in adults - mechanisms, causes, and consequences. *Blood Reviews*, *28*(2), 41–47. https://doi.org/10.1016/j.blre.2013.12.003

Naeem, M., Majeed, S., Hoque, M. Z., & Ahmad, I. (2020). Latest Developed Strategies to Minimize the Off-Target Effects in CRISPR-Cas-Mediated Genome Editing. *Cells*, *9*(7), 1608. https://doi.org/10.3390/cells9071608

Nagai, T., Yasuoka, Y., Izumi, Y., Horikawa, K., Kimura, M., Nakayama, Y., Uematsu, T., Fukuyama, T., Yamazaki, T., Kohda, Y., Hasuike, Y., Nanami, M., Kuragano, T., Kobayashi, N., Obinata, M., Tomita, K., Tanoue, A., Nakanishi, T., Kawahara, K., & Nonoguchi, H. (2014). Reevaluation of erythropoietin production by the nephron. *Biochemical and Biophysical Research Communications*, *449*(2), 222–228. https://doi.org/10.1016/j.bbrc.2014.05.014

Nangaku, M., Hamano, T., Akizawa, T., Tsubakihara, Y., Nagai, R., Okuda, N., Kurata, K., Nagakubo, T., Jones, N. P., Endo, Y., & Cobitz, A. R. (2021). Daprodustat Compared with Epoetin Beta Pegol for Anemia in Japanese Patients Not on Dialysis: A 52-Week Randomized Open-Label Phase 3 Trial. *American Journal of Nephrology*, *52*(1), 26–35. https://doi.org/10.1159/000513103

Nelson, M. R., Tipney, H., Painter, J. L., Shen, J., Nicoletti, P., Shen, Y., Floratos, A., Sham, P. C., Li, M. J., Wang, J., Cardon, L. R., Whittaker, J. C., & Sanseau, P. (2015). The support of human genetic evidence for approved drug indications. *Nature Genetics*, *47*(8), 856–860. https://doi.org/10.1038/ng.3314

Nemudryi, A. A., Valetdinova, K. R., Medvedev, S. P., & Zakian, S. M. (2014). TALEN and CRISPR/Cas Genome Editing Systems: Tools of Discovery. *Acta Naturae*, *6*(3), 19–40. https://pubmed.ncbi.nlm.nih.gov/25349712

Neupane, P., Bhuju, S., Thapa, N., & Bhattarai, H. K. (2019). ATP Synthase: Structure, Function and Inhibition. *Biomolecular Concepts*, *10*(1), 1–10. https://doi.org/doi:10.1515/bmc-2019-0001

Newcombe, P. J., Conti, D. V, & Richardson, S. (2016). JAM: A Scalable Bayesian Framework for Joint Analysis of Marginal SNP Effects. *Genetic Epidemiology*, *40*(3), 188–201. https://doi.org/10.1002/gepi.21953

Ng, S. R., Rideout, W. M., Akama-Garren, E. H., Bhutkar, A., Mercer, K. L., Schenkel, J. M., Bronson, R. T., & Jacks, T. (2020). CRISPR-mediated modeling and functional validation of candidate tumor suppressor genes in small cell lung cancer. *Proceedings of the National Academy of Sciences*, *117*(1), 513–521. https://doi.org/10.1073/pnas.1821893117

Nguyen, P. A., Born, D. A., Deaton, A. M., Nioi, P., & Ward, L. D. (2019). Phenotypes associated with genes encoding drug targets are predictive of clinical trial side effects. *Nature Communications*, *10*(1), 1579. https://doi.org/10.1038/s41467-019-09407-3

Nica, A. C., & Dermitzakis, E. T. (2013). Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *368*(1620), 20120362. https://doi.org/10.1098/rstb.2012.0362

Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I., & Dermitzakis, E. T. (2010). Candidate Causal Regulatory Effects by Integration of Expression QTLs with Complex Trait Genetic Associations. *PLOS Genetics*, *6*(4), e1000895. https://doi.org/10.1371/journal.pgen.1000895

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., & Cox, N. J. (2010). Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *PLOS Genetics*, *6*(4), e1000888. https://doi.org/10.1371/journal.pgen.1000888

Nikpay, M., Goel, A., Won, H.-H., Hall, L. M., Willenborg, C., Kanoni, S., Saleheen, D., Kyriakou, T., Nelson, C. P., Hopewell, J. C., Webb, T. R., Zeng, L., Dehghan, A., Alver, M., Armasu, S. M., Auro, K., Bjonnes, A., Chasman, D. I., Chen, S., … Farrall, M. (2015). A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*, *47*(10), 1121–1130. https://doi.org/10.1038/ng.3396

Noguchi, C. T., Wang, L., Rogers, H. M., Teng, R., & Jia, Y. (2008). Survival and proliferative roles of erythropoietin beyond the erythroid lineage. *Expert Reviews in Molecular Medicine*, *10*, e36–e36. https://doi.org/10.1017/S1462399408000860

O'Brien, A. R., Wilson, L. O. W., Burgio, G., & Bauer, D. C. (2019). Unlocking HDR-mediated nucleotide editing by identifying high-efficiency target sites using machine learning. *Scientific Reports*, *9*(1), 2788. https://doi.org/10.1038/s41598-019-39142-0

O'Seaghdha, C. M., Parekh, R. S., Hwang, S.-J., Li, M., Köttgen, A., Coresh, J., Yang, Q., Fox, C. S., & Kao, W. H. L. (2011). The MYH9/APOL1 region and chronic kidney disease in European-Americans. *Human Molecular Genetics*, *20*(12), 2450–2456. https://doi.org/10.1093/hmg/ddr118

Obara, N., Suzuki, N., Kim, K., Nagasawa, T., Imagawa, S., & Yamamoto, M. (2008). Repression via the GATA box is essential for tissue-specific erythropoietin gene expression. *Blood*, *111*(10), 5223–5232. https://doi.org/10.1182/blood-2007-10-115857

Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., Graham, R. R., Manoharan, A., Ortmann, W., Bhangale, T., Denny, J. C., Carroll, R. J., Eyler, A. E., Greenberg, J. D., Kremer, J. M., … Plenge, R. M. (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, *506*(7488), 376–381. https://doi.org/10.1038/nature12873

Okamoto, S., Amaishi, Y., Maki, I., Enoki, T., & Mineno, J. (2019). Highly efficient genome editing for single-base substitutions using optimized ssODNs with Cas9-RNPs. *Scientific Reports*, *9*(1), 4811. https://doi.org/10.1038/s41598-019-41121-4

Orlić, L., Mikolasevic, I., Bagic, Z., Racki, S., Stimac, D., & Milic, S. (2014). Chronic Kidney Disease and Nonalcoholic Fatty Liver Disease—Is There a Link? *Gastroenterology*

Research and Practice, *2014*, 847539. https://doi.org/10.1155/2014/847539

Owen, K. R., Thanabalasingham, G., James, T. J., Karpe, F., Farmer, A. J., McCarthy, M. I., & Gloyn, A. L. (2010). Assessment of high-sensitivity C-reactive protein levels as diagnostic discriminator of maturity-onset diabetes of the young due to HNF1A mutations. *Diabetes Care*, *33*(9), 1919–1924. https://doi.org/10.2337/dc10-0288

Paananen, J., & Fortino, V. (2020). An omics perspective on drug target discovery platforms. *Briefings in Bioinformatics*, *21*(6), 1937–1953. https://doi.org/10.1093/bib/bbz122

Page, P. (2014). Beyond statistical significance: clinical interpretation of rehabilitation research literature. *International Journal of Sports Physical Therapy*, *9*(5), 726–736. https://pubmed.ncbi.nlm.nih.gov/25328834

Paliege, A., Rosenberger, C., Bondke, A., Sciesielski, L., Shina, A., Heyman, S. N., Flippin, L. A., Arend, M., Klaus, S. J., & Bachmann, S. (2010). Hypoxia-inducible factor-2alpha-expressing interstitial fibroblasts are the only renal cells that express erythropoietin under hypoxia-inducible factor stabilization. *Kidney International*, *77*(4), 312–318. https://doi.org/10.1038/ki.2009.460

Panjeta, M., Tahirović, I., Sofić, E., Ćorić, J., & Dervišević, A. (2017). Interpretation of Erythropoietin and Haemoglobin Levels in Patients with Various Stages of Chronic Kidney Disease. *Journal of Medical Biochemistry*, *36*(2), 145–152. https://doi.org/10.1515/jomb-2017-0014

Paquet, D., Kwart, D., Chen, A., Sproul, A., Jacob, S., Teo, S., Olsen, K. M., Gregg, A., Noggle, S., & Tessier-Lavigne, M. (2016). Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9. *Nature*, *533*(7601), 125–129. https://doi.org/10.1038/nature17664

Parfrey, P. (2021). Hypoxia-Inducible Factor Prolyl Hydroxylase Inhibitors for Anemia in CKD. *New England Journal of Medicine*. https://doi.org/10.1056/NEJMe2117100

Park, S. L., Won, S. Y., Song, J.-H., Kambe, T., Nagao, M., Kim, W.-J., & Moon, S.-K. (2015). EPO gene expression promotes proliferation, migration and invasion via the p38MAPK/AP-1/MMP-9 pathway by p21WAF1 expression in vascular smooth muscle cells. *Cellular Signalling*, *27*(3), 470–478. https://doi.org/10.1016/j.cellsig.2014.12.001

Pattanayak, V., Ramirez, C. L., Joung, J. K., & Liu, D. R. (2011). Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection. *Nature Methods*, *8*(9), 765–770. https://doi.org/10.1038/nmeth.1670

Pendergrass, S. A., & Ritchie, M. D. (2015). Phenome-wide association studies: leveraging comprehensive phenotypic and genotypic data for discovery. *Current Genetic Medicine Reports*, *3*(2), 92–100.

Perez-Pinera, P., Kocak, D. D., Vockley, C. M., Adler, A. F., Kabadi, A. M., Polstein, L. R., Thakore, P. I., Glass, K. A., Ousterout, D. G., Leong, K. W., Guilak, F., Crawford, G. E., Reddy, T. E., & Gersbach, C. A. (2013). RNA-guided gene activation by CRISPR-Cas9-based transcription factors. *Nature Methods*, *10*(10), 973–976. https://doi.org/10.1038/nmeth.2600

Perret-Guillaume, C., Joly, L., & Benetos, A. (2009). Heart rate as a risk factor for cardiovascular disease. *Progress in Cardiovascular Diseases*, *52*(1), 6–10. https://doi.org/10.1016/j.pcad.2009.05.003

Pers, T. H., Karjalainen, J. M., Chan, Y., Westra, H.-J., Wood, A. R., Yang, J., Lui, J. C., Vedantam, S., Gustafsson, S., Esko, T., Frayling, T., Speliotes, E. K., Boehnke, M., Raychaudhuri, S., Fehrmann, R. S. N., Hirschhorn, J. N., Franke, L., & Consortium, G. I. of

An. T. (GIANT). (2015). Biological interpretation of genome-wide association studies using predicted gene functions. *Nature Communications*, *6*(1), 5890. https://doi.org/10.1038/ncomms6890

Pfeffer, M. A., Burdmann, E. A., Chen, C.-Y., Cooper, M. E., de Zeeuw, D., Eckardt, K.-U., Feyzi, J. M., Ivanovich, P., Kewalramani, R., Levey, A. S., Lewis, E. F., McGill, J. B., McMurray, J. J. V, Parfrey, P., Parving, H.-H., Remuzzi, G., Singh, A. K., Solomon, S. D., & Toto, R. (2009). A Trial of Darbepoetin Alfa in Type 2 Diabetes and Chronic Kidney Disease. *New England Journal of Medicine*, *361*(21), 2019–2032. https://doi.org/10.1056/NEJMoa0907845

Pham, T.-N. D., Ma, W., Miller, D., Kazakova, L., & Benchimol, S. (2019). Erythropoietin inhibits chemotherapy-induced cell death and promotes a senescence-like state in leukemia cells. *Cell Death & Disease*, *10*(1), 22. https://doi.org/10.1038/s41419-018-1274-6

Phillips, T. M., Kim, K., Vlashi, E., McBride, W. H., & Pajonk, F. (2007). Effects of recombinant erythropoietin on breast cancer-initiating cells. *Neoplasia (New York, N.Y.)*, *9*(12), 1122–1129. https://doi.org/10.1593/neo.07694

Pickar-Oliver, A., & Gersbach, C. A. (2019). The next generation of CRISPR–Cas technologies and applications. *Nature Reviews Molecular Cell Biology*, *20*(8), 490–507. https://doi.org/10.1038/s41580-019-0131-5

PINTO-SIETSMA, S.-J., JANSSEN, W. M. T., HILLEGE, H. L., NAVIS, G., ZEEUW, D. D. E., & JONG, P. E. D. E. (2000). Urinary Albumin Excretion Is Associated with Renal Functional Abnormalities in a Nondiabetic Population. *Journal of the American Society of Nephrology*, *11*(10), 1882 LP – 1888. https://doi.org/10.1681/ASN.V11101882

Plenge, R. M. (2016). Disciplined approach to drug discovery and early development. *Science Translational Medicine*, *8*(349), 349ps15-349ps15. https://doi.org/10.1126/scitranslmed.aaf2608

Plenge, R. M., Scolnick, E. M., & Altshuler, D. (2013). Validating therapeutic targets through human genetics. *Nature Reviews Drug Discovery*, *12*(8), 581–594. https://doi.org/10.1038/nrd4051

Porcu, E., Rüeger, S., Lepik, K., Agbessi, M., Ahsan, H., Alves, I., Andiappan, A., Arindrarto, W., Awadalla, P., Battle, A., Beutner, F., Jan Bonder, M., Boomsma, D., Christiansen, M., Claringbould, A., Deelen, P., Esko, T., Favé, M.-J., Franke, L., … Consortium, B. (2019). Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nature Communications*, *10*(1), 3300. https://doi.org/10.1038/s41467-019-10936-0

Portolés, J., Martín, L., Broseta, J. J., & Cases, A. (2021). Anemia in Chronic Kidney Disease: From Pathophysiology and Current Treatments, to Future Agents . In *Frontiers in Medicine* (Vol. 8). https://www.frontiersin.org/article/10.3389/fmed.2021.642296

Prestigiacomo, V., & Suter-Dick, L. (2018). Nrf2 protects stellate cells from Smad-dependent cell activation. *PLOS ONE*, *13*(7), e0201044. https://doi.org/10.1371/journal.pone.0201044

Provenzano, R., Fishbane, S., Szczech, L., Leong, R., Saikali, K. G., Zhong, M., Lee, T. T., Houser, M. T., Frison, L., Houghton, J., Little, D. J., Peony Yu, K.-H., & Neff, T. B. (2021). Pooled Analysis of Roxadustat for Anemia in Patients With Kidney Failure Incident to Dialysis. *Kidney International Reports*, *6*(3), 613–623. https://doi.org/https://doi.org/10.1016/j.ekir.2020.12.018

Pulit, S. L., Stoneman, C., Morris, A. P., Wood, A. R., Glastonbury, C. A., Tyrrell, J., Yengo, L., Ferreira, T., Marouli, E., Ji, Y., Yang, J., Jones, S., Beaumont, R., Croteau-Chonka, D. C., Winkler, T. W., GIANT, C., Hattersley, A. T., Loos, R. J. F., Hirschhorn, J. N., … Lindgren, C. M. (2018). Meta-analysis of genome-wide association studies for body fat distribution in 694,649 individuals of European ancestry. *Human Molecular Genetics*. https://doi.org/10.1093/hmg/ddy327

Pulley, J. M., Shirey-Rice, J. K., Lavieri, R. R., Jerome, R. N., Zaleski, N. M., Aronoff, D. M., Bastarache, L., Niu, X., Holroyd, K. J., Roden, D. M., Skaar, E. P., Niswender, C. M., Marnett, L. J., Lindsley, C. W., Ekstrom, L. B., Bentley, A. R., Bernard, G. R., Hong, C. C., & Denny, J. C. (2017). Accelerating Precision Drug Development and Drug Repurposing by Leveraging Human Genetics. *ASSAY and Drug Development Technologies*, *15*(3), 113–119. https://doi.org/10.1089/adt.2016.772

Qi, L. S., Larson, M. H., Gilbert, L. A., Doudna, J. A., Weissman, J. S., Arkin, A. P., & Lim, W. A. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*, *152*(5), 1173–1183. https://doi.org/10.1016/j.cell.2013.02.022

Raj, T., Rothamel, K., Mostafavi, S., Ye, C., Lee, M. N., Replogle, J. M., Feng, T., Lee, M., Asinovski, N., Frohlich, I., Imboywa, S., Von Korff, A., Okada, Y., Patsopoulos, N. A., Davis, S., McCabe, C., Paik, H., Gyan, S. P., Raychaudhuri, S., … De Jager, P. L. (2014). Polarization of the Effects of Autoimmune and Neurodegenerative Risk Alleles in Leukocytes. *Science*, *344*(6183), 519–523. https://doi.org/10.1126/science.1249547

Ramakrishnan, S., Anand, V., & Roy, S. (2014). Vascular endothelial growth factor signaling in hypoxia and inflammation. *Journal of Neuroimmune Pharmacology : The Official Journal of the Society on NeuroImmune Pharmacology*, *9*(2), 142–160. https://doi.org/10.1007/s11481-014-9531-7

Ramakrishnan, S. K., & Shah, Y. M. (2017). A central role for hypoxia-inducible factor (HIF)-2α in hepatic glucose homeostasis. *Nutrition and Healthy Aging*, *4*(3), 207–216. https://doi.org/10.3233/NHA-170022

Ran, F. A., Hsu, P. D., Lin, C.-Y., Gootenberg, J. S., Konermann, S., Trevino, A. E., Scott, D. A., Inoue, A., Matoba, S., Zhang, Y., & Zhang, F. (2013). Double Nicking by RNA-Guided CRISPR Cas9 for Enhanced Genome Editing Specificity. *Cell*, *154*(6), 1380–1389. https://doi.org/https://doi.org/10.1016/j.cell.2013.08.021

Ran, F. A., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A., & Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. *Nature Protocols*, *8*(11), 2281–2308. https://doi.org/10.1038/nprot.2013.143

Rana, N. K., Singh, P., & Koch, B. (2019). CoCl2 simulated hypoxia induce cell proliferation and alter the expression pattern of hypoxia associated genes involved in angiogenesis and apoptosis. *Biological Research*, *52*(1), 12. https://doi.org/10.1186/s40659-019-0221-z

Randall, J. C., Winkler, T. W., Kutalik, Z., Berndt, S. I., Jackson, A. U., Monda, K. L., Kilpeläinen, T. O., Esko, T., Mägi, R., Li, S., Workalemahu, T., Feitosa, M. F., Croteau-Chonka, D. C., Day, F. R., Fall, T., Ferreira, T., Gustafsson, S., Locke, A. E., Mathieson, I., … Heid, I. M. (2013). Sex-stratified Genome-wide Association Studies Including 270,000 Individuals Show Sexual Dimorphism in Genetic Loci for Anthropometric Traits. *PLOS Genetics*, *9*(6), e1003500. https://doi.org/10.1371/journal.pgen.1003500

Rao, M. S., Van Vleet, T. R., Ciurlionis, R., Buck, W. R., Mittelstadt, S. W., Blomme, E. A. G., & Liguori, M. J. (2019). Comparison of RNA-Seq and Microarray Gene Expression Platforms for the Toxicogenomic Evaluation of Liver From Short-Term Rat Toxicity Studies . In *Frontiers in Genetics* (Vol. 9). https://www.frontiersin.org/article/10.3389/fgene.2018.00636

Rao, Shuquan, Yao, Y., & Bauer, D. E. (2021). Editing GWAS: experimental approaches to dissect and exploit disease-associated genetic variation. *Genome Medicine*, *13*(1), 41. https://doi.org/10.1186/s13073-021-00857-3

Rao, Shuyun, Lee, S.-Y., Gutierrez, A., Perrigoue, J., Thapa, R. J., Tu, Z., Jeffers, J. R., Rhodes, M., Anderson, S., Oravecz, T., Hunger, S. P., Timakhov, R. A., Zhang, R., Balachandran, S., Zambetti, G. P., Testa, J. R., Look, A. T., & Wiest, D. L. (2012). Inactivation of ribosomal protein L22 promotes transformation by induction of the stemness factor, Lin28B. *Blood*, *120*(18), 3764–3773. https://doi.org/10.1182/blood-2012-03-415349

Rath, D., Amlinger, L., Rath, A., & Lundgren, M. (2015). The CRISPR-Cas immune system: Biology, mechanisms and applications. *Biochimie*, *117*, 119–128. https://doi.org/https://doi.org/10.1016/j.biochi.2015.03.025

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Göttgens, B., … Yosef, N. (2017). The Human Cell Atlas. *ELife*, *6*. https://doi.org/10.7554/eLife.27041

Renner, W., Kaiser, M., Khuen, S., Trummer, O., Mangge, H., & Langsenlehner, T. (2020). The Erythropoetin rs1617640 Gene Polymorphism Associates with Hemoglobin Levels, Hematocrit and Red Blood Cell Count in Patients with Peripheral Arterial Disease. *Genes*, *11*(11), 1305. https://doi.org/10.3390/genes11111305

Richmond, R. C., & Davey Smith, G. (2021). Mendelian Randomization: Concepts and Scope. *Cold Spring Harbor Perspectives in Medicine*, *12*(1). http://perspectivesinmedicine.cshlp.org/content/12/1/a040501.abstract

Robert-Moreno, À., Espinosa, L., Sanchez, M. J., de la Pompa, J. L., & Bigas, A. (2007). The notch pathway positively regulates programmed cell death during erythroid differentiation. *Leukemia*, *21*(7), 1496–1503. https://doi.org/10.1038/sj.leu.2404705

Robert, C., & Watson, M. (2015). Errors in RNA-Seq quantification affect genes of relevance to human disease. *Genome Biology*, *16*(1), 177. https://doi.org/10.1186/s13059-015-0734-x

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L., & Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology*, *12*(3), R22. https://doi.org/10.1186/gb-2011-12-3-r22

Robinson, J. R., Denny, J. C., Roden, D. M., & Van Driest, S. L. (2018). Genome-wide and Phenome-wide Approaches to Understand Variable Drug Actions in Electronic Health Records. *Clinical and Translational Science*, *11*(2), 112–122. https://doi.org/10.1111/cts.12522

Rodriguez, D., Watts, D., Gaete, D., Sormendi, S., & Wielockx, B. (2021). Hypoxia Pathway Proteins and Their Impact on the Blood Vasculature. *International Journal of Molecular Sciences*, *22*(17), 9191. https://doi.org/10.3390/ijms22179191

Rohde, P. D., Østergaard, S., Kristensen, T. N., Sørensen, P., Loeschcke, V., Mackay, T. F. C., & Sarup, P. (2018). Functional Validation of Candidate Genes Detected by Genomic Feature Models. *G3 (Bethesda, Md.)*, *8*(5), 1659–1668. https://doi.org/10.1534/g3.118.200082

RStudio Team. (2018). *RStudio: Integrated Development for R.* RStudio, Inc., Boston, MA. http://www.rstudio.com/

Sakaue, S., & Okada, Y. (2019). GREP: genome for REPositioning drugs. *Bioinformatics (Oxford, England)*, *35*(19), 3821–3823. https://doi.org/10.1093/bioinformatics/btz166

Sanson, K. R., Hanna, R. E., Hegde, M., Donovan, K. F., Strand, C., Sullender, M. E., Vaimberg, E. W., Goodale, A., Root, D. E., Piccioni, F., & Doench, J. G. (2018). Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. *Nature Communications*, *9*(1), 5416. https://doi.org/10.1038/s41467-018-07901-8

Santhanam, A. V. R., d'Uscio, L. V, & Katusic, Z. S. (2010). Cardiovascular effects of erythropoietin an update. *Advances in Pharmacology (San Diego, Calif.)*, *60*, 257–285. https://doi.org/10.1016/B978-0-12-385061-4.00009-X

Saran, R., Robinson, B., Abbott, K. C., Agodoa, L. Y. C., Bragg-Gresham, J., Balkrishnan, R., Bhave, N., Dietrich, X., Ding, Z., Eggers, P. W., Gaipov, A., Gillen, D., Gipson, D., Gu, H., Guro, P., Haggerty, D., Han, Y., He, K., Herman, W., … Shahinian, V. (2019). US Renal Data System 2018 Annual Data Report: Epidemiology of Kidney Disease in the United States. *American Journal of Kidney Diseases : The Official Journal of the National Kidney Foundation*, *73*(3 Suppl 1), A7–A8. https://doi.org/10.1053/j.ajkd.2019.01.001

Schadt, E. E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P. Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C., Zhu, J., Millstein, J., Sieberts, S., Lamb, J., GuhaThakurta, D., Derry, J., Storey, J. D., Avila-Campillo, I., Kruger, M. J., … Ulrich, R. (2008). Mapping the genetic architecture of gene expression in human liver. *PLoS Biology*, *6*(5), e107. https://doi.org/10.1371/journal.pbio.0060107

Schaid, D. J., Chen, W., & Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews. Genetics*, *19*(8), 491–504. https://doi.org/10.1038/s41576-018-0016-z

Schertzer, M. D., Thulson, E., Braceros, K. C. A., Lee, D. M., Hinkle, E. R., Murphy, R. M., Kim, S. O., Vitucci, E. C. M., & Calabrese, J. M. (2019). A piggyBac-based toolkit for inducible genome editing in mammalian cells. *RNA (New York, N.Y.)*, *25*(8), 1047–1058. https://doi.org/10.1261/rna.068932.118

Schmidt, A. F., Finan, C., Gordillo-Marañón, M., Asselbergs, F. W., Freitag, D. F., Patel, R. S., Tyl, B., Chopade, S., Faraway, R., Zwierzyna, M., & Hingorani, A. D. (2020). Genetic drug target validation using Mendelian randomisation. *Nature Communications*, *11*(1), 3255. https://doi.org/10.1038/s41467-020-16969-0

Schmidt, A. F., Swerdlow, D. I., Holmes, M. V, Patel, R. S., Fairhurst-Hunter, Z., Lyall, D. M., Hartwig, F. P., Horta, B. L., Hyppönen, E., & Power, C. (2017). PCSK9 genetic variants and risk of type 2 diabetes: a mendelian randomisation study. *The Lancet Diabetes & Endocrinology*, *5*(2), 97–105.

Schönenberger, M., & Kovacs, W. (2015). Hypoxia signaling pathways: modulators of oxygen-related organelles . In *Frontiers in Cell and Developmental Biology* (Vol. 3). https://www.frontiersin.org/article/10.3389/fcell.2015.00042

Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., Lightfoot, S., Menzel, W., Granzow, M., & Ragg, T. (2006). The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Molecular Biology*, *7*(1), 3. https://doi.org/10.1186/1471-2199-7-3

Scott, R. A., Freitag, D. F., Li, L., Chu, A. Y., Surendran, P., Young, R., Grarup, N., Stancáková, A., Chen, Y., Varga, T. V, Yaghootkar, H., Luan, J., Zhao, J. H., Willems, S. M., Wessel, J., Wang, S., Maruthur, N., Michailidou, K., Pirie, A., … Waterworth, D. M. (2016). A genomic approach to therapeutic target validation identifies a glucose-lowering GLP1R variant protective for coronary heart disease. *Science Translational Medicine*, *8*(341), 341ra76. https://doi.org/10.1126/scitranslmed.aad3744

Shafi, T., & Coresh, J. (2019). 1 - Chronic Kidney Disease: Definition, Epidemiology, Cost, and

Outcomes. In J. Himmelfarb & T. A. Ikizler (Eds.), *Chronic Kidney Disease, Dialysis, and Transplantation (Fourth Edition)* (Fourth Edi, pp. 2-22.e3). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-323-52978-5.00001-X

Sharan, S. K., Thomason, L. C., Kuznetsov, S. G., & Court, D. L. (2009). Recombineering: a homologous recombination-based method of genetic engineering. *Nature Protocols*, *4*(2), 206–223. https://doi.org/10.1038/nprot.2008.227

Sheehan, N. A., & Didelez, V. (2019). Epidemiology, genetic epidemiology and Mendelian randomisation: more need than ever to attend to detail. *Human Genetics*, *139*(1), 121–136. https://doi.org/10.1007/s00439-019-02027-3

Sheehan, N. A., Didelez, V., Burton, P. R., & Tobin, M. D. (2008). Mendelian Randomisation and Causal Inference in Observational Epidemiology. *PLOS Medicine*, *5*(8), e177. https://doi.org/10.1371/journal.pmed.0050177

Shepshelovich, D., Rozen-Zvi, B., Avni, T., Gafter, U., & Gafter-Gvili, A. (2016). Intravenous Versus Oral Iron Supplementation for the Treatment of Anemia in CKD: An Updated Systematic Review and Meta-analysis. *American Journal of Kidney Diseases : The Official Journal of the National Kidney Foundation*, *68*(5), 677–690. https://doi.org/10.1053/j.ajkd.2016.04.018

Shi, H., Zhou, Y., Jia, E., Pan, M., Bai, Y., & Ge, Q. (2021). Bias in RNA-seq Library Preparation: Current Challenges and Solutions. *BioMed Research International*, *2021*, 6647597. https://doi.org/10.1155/2021/6647597

Shih, H.-M., Wu, C.-J., & Lin, S.-L. (2018). Physiology and pathophysiology of renal erythropoietin-producing cells. *Journal of the Formosan Medical Association*, *117*(11), 955–963. https://doi.org/https://doi.org/10.1016/j.jfma.2018.03.017

Shock, N. W. (1984). *Normal human aging: The Baltimore longitudinal study of aging*.

Shu, L., Blencowe, M., & Yang, X. (2018). Translating GWAS Findings to Novel Therapeutic Targets for Coronary Artery Disease . In *Frontiers in Cardiovascular Medicine* (Vol. 5, p. 56). https://www.frontiersin.org/article/10.3389/fcvm.2018.00056

Simonsick, E. M., Newman, A. B., Nevitt, M. C., Kritchevsky, S. B., Ferrucci, L., Guralnik, J. M., Harris, T., & Group, for the H. A. B. C. S. (2001). Measuring Higher Level Physical Function in Well-Functioning Older Adults: Expanding Familiar Approaches in the Health ABC Study. *The Journals of Gerontology: Series A*, *56*(10), M644–M649. https://doi.org/10.1093/gerona/56.10.M644

Singh, A. K. (2018). *Chapter 12 - Erythropoiesis: The Roles of Erythropoietin and Iron* (A. K. Singh & G. H. B. T.-T. of N.-E. (Second E. Williams (eds.); pp. 207–215). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-803247-3.00012-X

Singh, A. K., Carroll, K., McMurray, J. J. V, Solomon, S., Jha, V., Johansen, K. L., Lopes, R. D., Macdougall, I. C., Obrador, G. T., Waikar, S. S., Wanner, C., Wheeler, D. C., Więcek, A., Blackorby, A., Cizman, B., Cobitz, A. R., Davies, R., DiMino, T. L., Kler, L., … Perkovic, V. (2021). Daprodustat for the Treatment of Anemia in Patients Not Undergoing Dialysis. *New England Journal of Medicine*. https://doi.org/10.1056/NEJMoa2113380

Singh, A. K., Carroll, K., Perkovic, V., Solomon, S., Jha, V., Johansen, K. L., Lopes, R. D., Macdougall, I. C., Obrador, G. T., Waikar, S. S., Wanner, C., Wheeler, D., Więcek, A., Blackorby, A., Cizman, B., Cobitz, A. R., Davies, R., Dole, J., Kler, L., … ASCEND-D Study Group. (2021). Daprodustat for the Treatment of Anemia in Pateints Undergoing Dialysis. *New England Journal of Medicine*, *385*, 2325–2335. https://doi.org/10.1056/NEJMoa2113379

Singh, A. K., Szczech, L., Tang, K. L., Barnhart, H., Sapp, S., Wolfson, M., & Reddan, D. (2006). Correction of anemia with epoetin alfa in chronic kidney disease. *The New England Journal of Medicine*, *355*(20), 2085–2098. https://doi.org/10.1056/NEJMoa065485

Singh, A. M., Adjan Steffey, V. V, Yeshi, T., & Allison, D. W. (2015). Gene Editing in Human Pluripotent Stem Cells: Choosing the Correct Path. *Journal of Stem Cell and Regenerative Biology*, *1*(1), http://www.ommegaonline.org/article-details/Gene-E. https://pubmed.ncbi.nlm.nih.gov/26702451

Singh, A. M., Perry, D. W., Steffey, V. V. A., Miller, K., & Allison, D. W. (2016). *Decoding the Epigenetic Heterogeneity of Human Pluripotent Stem Cells with Seamless Gene Editing BT - Stem Cell Heterogeneity: Methods and Protocols* (K. Turksen (ed.); pp. 153–169). Springer New York. https://doi.org/10.1007/7651_2016_324

Smith, G. D., & Ebrahim, S. (2005). What can mendelian randomisation tell us about modifiable behavioural and environmental exposures? *BMJ*, *330*(7499), 1076–1079. https://doi.org/10.1136/bmj.330.7499.1076

Smith, G. D., Ebrahim, S., Lewis, S., Hansell, A. L., Palmer, L. J., & Burton, P. R. (2005). Genetic epidemiology and public health: hope, hype, and future prospects. *The Lancet*, *366*(9495), 1484–1498. https://doi.org/https://doi.org/10.1016/S0140-6736(05)67601-5

Sofianopoulou, E., Kaptoge, S. K., Afzal, S., Jiang, T., Gill, D., Gundersen, T. E., Bolton, T. R., Allara, E., Arnold, M. G., Mason, A. M., Chung, R., Pennells, L. A. M., Shi, F., Sun, L., Willeit, P., Forouhi, N. G., Langenberg, C., Sharp, S. J., Panico, S., … Burgess, S. (2021). Estimating dose-response relationships for vitamin D with coronary heart disease, stroke, and all-cause mortality: observational and Mendelian randomisation analyses. *The Lancet Diabetes & Endocrinology*, *9*(12), 837–846. https://doi.org/10.1016/S2213-8587(21)00263-1

Solaini, G., Baracca, A., Lenaz, G., & Sgarbi, G. (2010). Hypoxia and mitochondrial oxidative metabolism. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, *1797*(6), 1171–1177. https://doi.org/https://doi.org/10.1016/j.bbabio.2010.02.011

Sorkin, J. D., Muller, D. C., & Andres, R. (1999). Longitudinal change in height of men and women: implications for interpretation of the body mass index: the Baltimore Longitudinal Study of Aging. *American Journal of Epidemiology*, *150*(9), 969–977. https://doi.org/10.1093/oxfordjournals.aje.a010106

Sormendi, S., & Wielockx, B. (2018). Hypoxia Pathway Proteins As Central Mediators of Metabolism in the Tumor Cells and Their Microenvironment . In *Frontiers in Immunology* (Vol. 9). https://www.frontiersin.org/article/10.3389/fimmu.2018.00040

Souma, T., Suzuki, N., & Yamamoto, M. (2015). Renal erythropoietin-producing cells in health and disease . In *Frontiers in Physiology* (Vol. 6). https://www.frontiersin.org/article/10.3389/fphys.2015.00167

Spencer, C. C. A., Su, Z., Donnelly, P., & Marchini, J. (2009). Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genetics*, *5*(5), e1000477. https://doi.org/10.1371/journal.pgen.1000477

St Peter, W. L., Guo, H., Kabadi, S., Gilbertson, D. T., Peng, Y., Pendergraft, T., & Li, S. (2018). Prevalence, treatment patterns, and healthcare resource utilization in Medicare and commercially insured non-dialysis-dependent chronic kidney disease patients with and without anemia in the United States. *BMC Nephrology*, *19*(1), 67. https://doi.org/10.1186/s12882-018-0861-1

Stauffer, M. E., & Fan, T. (2014). Prevalence of anemia in chronic kidney disease in the United States. *PloS One*, *9*(1), e84943. https://doi.org/10.1371/journal.pone.0084943

Stepanenko, A. A., & Dmitrenko, V. V. (2015). HEK293 in cell biology and cancer research: phenotype, karyotype, tumorigenicity, and stress-induced genome-phenotype evolution. *Gene*, *569*(2), 182–190. https://doi.org/https://doi.org/10.1016/j.gene.2015.05.065

Steyer, B., Bu, Q., Cory, E., Jiang, K., Duong, S., Sinha, D., Steltzer, S., Gamm, D., Chang, Q., & Saha, K. (2018). Scarless Genome Editing of Human Pluripotent Stem Cells via Transient Puromycin Selection. *Stem Cell Reports*, *10*(2), 642–654. https://doi.org/10.1016/j.stemcr.2017.12.004

Stitziel, N. O., Won, H.-H., Morrison, A. C., Peloso, G. M., Do, R., Lange, L. A., Fontanillas, P., Gupta, N., Duga, S., Goel, A., Farrall, M., Saleheen, D., Ferrario, P., König, I., Asselta, R., Merlini, P. A., Marziliano, N., Notarangelo, M. F., Schick, U., … Kathiresan, S. (2014). Inactivating mutations in NPC1L1 and protection from coronary heart disease. *The New England Journal of Medicine*, *371*(22), 2072–2082. https://doi.org/10.1056/NEJMoa1405386

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*, *12*(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Paolo Casale, F., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., … Consortium, T. 1000 G. P. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*(7571), 75–81. https://doi.org/10.1038/nature15394

Sugahara, M., Tanaka, T., & Nangaku, M. (2017). Prolyl hydroxylase domain inhibitors as a novel therapeutic approach against anemia in chronic kidney disease. *Kidney International*, *92*(2), 306–312. https://doi.org/10.1016/j.kint.2017.02.035

Sun, B. B., Kurki, M. I., Foley, C. N., Mechakra, A., Chen, C.-Y., Marshall, E., Wilk, J. B., Team, B. B., Chahine, M., Chevalier, P., Christé, G., FinnGen, Palotie, A., Daly, M. J., & Runz, H. (2021). Genetic associations of protein-coding variants in human disease. *MedRxiv*. https://doi.org/10.1101/2021.10.14.21265023

Suresh, S., Rajvanshi, P. K., & Noguchi, C. T. (2020). The Many Facets of Erythropoietin Physiologic and Metabolic Response   . In *Frontiers in Physiology*   (Vol. 10, p. 1534). https://www.frontiersin.org/article/10.3389/fphys.2019.01534

Swerdlow, D. I., Kuchenbaecker, K. B., Shah, S., Sofat, R., Holmes, M. V, White, J., Mindell, J. S., Kivimaki, M., Brunner, E. J., Whittaker, J. C., Casas, J. P., & Hingorani, A. D. (2016). Selecting instruments for Mendelian randomization in the wake of genome-wide association studies. *International Journal of Epidemiology*, *45*(5), 1600–1616. https://doi.org/10.1093/ije/dyw088

Swerdlow, D. I., Preiss, D., Kuchenbaecker, K. B., Holmes, M. V, Engmann, J. E. L., Shah, T., Sofat, R., Stender, S., Johnson, P. C. D., Scott, R. A., Leusink, M., Verweij, N., Sharp, S. J., Guo, Y., Giambartolomei, C., Chung, C., Peasey, A., Amuzu, A., Li, K., … Sattar, N. (2015). HMG-coenzyme A reductase inhibition, type 2 diabetes, and bodyweight: evidence from genetic analysis and randomised trials. *Lancet (London, England)*, *385*(9965), 351–361. https://doi.org/10.1016/S0140-6736(14)61183-1

Takeda, K., Ho, V. C., Takeda, H., Duan, L.-J., Nagy, A., & Fong, G.-H. (2006). Placental but not heart defects are associated with elevated hypoxia-inducible factor alpha levels in mice lacking prolyl hydroxylase domain protein 2. *Molecular and Cellular Biology*, *26*(22), 8336–8346. https://doi.org/10.1128/MCB.00425-06

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, *20*(8), 467–484. https://doi.org/10.1038/s41576-019-0127-1

Tanaka, Y., Nishida, N., Sugiyama, M., Kurosaki, M., Matsuura, K., Sakamoto, N., Nakagawa, M., Korenaga, M., Hino, K., Hige, S., Ito, Y., Mita, E., Tanaka, E., Mochida, S., Murawaki, Y., Honda, M., Sakai, A., Hiasa, Y., Nishiguchi, S., … Mizokami, M. (2009). Genome-wide association of IL28B with response to pegylated interferon-alpha and ribavirin therapy for chronic hepatitis C. *Nature Genetics*, *41*(10), 1105–1109. https://doi.org/10.1038/ng.449

Tani, J., Ito, Y., Tatemichi, S., Yamakami, M., Fukui, T., Hatano, Y., Kakimoto, S., Kotani, A., Sugimura, A., Mihara, K., Yamamoto, R., Tanaka, N., Minami, K., Takahashi, K., & Hirato, T. (2020). Physicochemical and biological evaluation of JR-131 as a biosimilar to a long-acting erythropoiesis-stimulating agent darbepoetin alfa. *PloS One*, *15*(4), e0231830–e0231830. https://doi.org/10.1371/journal.pone.0231830

Thanabalasingham, G., Shah, N., Vaxillaire, M., Hansen, T., Tuomi, T., Gašperíková, D., Szopa, M., Tjora, E., James, T. J., Kokko, P., Loiseleur, F., Andersson, E., Gaget, S., Isomaa, B., Nowak, N., Raeder, H., Stanik, J., Njolstad, P. R., Malecki, M. T., … Owen, K. R. (2011). A large multi-centre European study validates high-sensitivity C-reactive protein (hsCRP) as a clinical biomarker for the diagnosis of diabetes subtypes. *Diabetologia*, *54*(11), 2801–2810. https://doi.org/10.1007/s00125-011-2261-y

Thomas, P., & Smart, T. G. (2005). HEK293 cell line: A vehicle for the expression of recombinant proteins. *Journal of Pharmacological and Toxicological Methods*, *51*(3), 187–200. https://doi.org/https://doi.org/10.1016/j.vascn.2004.08.014

Thomas, R., Kanso, A., & Sedor, J. R. (2008). Chronic kidney disease and its complications. *Primary Care*, *35*(2), 329–vii. https://doi.org/10.1016/j.pop.2008.01.008

Tipanee, J., VandenDriessche, T., & Chuah, M. K. (2017). Transposons: Moving Forward from Preclinical Studies to Clinical Trials. *Human Gene Therapy*, *28*(11), 1087–1104. https://doi.org/10.1089/hum.2017.128

Tokish, J. M., Kocher, M. S., & Hawkins, R. J. (2004). Ergogenic Aids: A Review of Basic Science, Performance, Side Effects, and Status in Sports. *The American Journal of Sports Medicine*, *32*(6), 1543–1553. https://doi.org/10.1177/0363546504268041

Tong, Z., Yang, Z., Patel, S., Chen, H., Gibbs, D., Yang, X., Hau, V. S., Kaminoh, Y., Harmon, J., Pearson, E., Buehler, J., Chen, Y., Yu, B., Tinkham, N. H., Zabriskie, N. A., Zeng, J., Luo, L., Sun, J. K., Prakash, M., … Zhang, K. (2008). Promoter polymorphism of the erythropoietin gene in severe diabetic eye and kidney complications. *Proceedings of the National Academy of Sciences*, *105*(19), 6998–7003. https://doi.org/10.1073/pnas.0800454105

Tycko, J., Wainberg, M., Marinov, G. K., Ursu, O., Hess, G. T., Ego, B. K., Aradhana, Li, A., Truong, A., Trevino, A. E., Spees, K., Yao, D., Kaplow, I. M., Greenside, P. G., Morgens, D. W., Phanstiel, D. H., Snyder, M. P., Bintu, L., Greenleaf, W. J., … Bassik, M. C. (2019). Mitigation of off-target toxicity in CRISPR-Cas9 screens for essential non-coding elements. *Nature Communications*, *10*(1), 4063. https://doi.org/10.1038/s41467-019-11955-7

Tzur, S., Rosset, S., Shemer, R., Yudkovsky, G., Selig, S., Tarekegn, A., Bekele, E., Bradman, N., Wasser, W. G., Behar, D. M., & Skorecki, K. (2010). Missense mutations in the APOL1 gene are highly associated with end stage kidney disease risk previously attributed to the

MYH9 gene. *Human Genetics*, *128*(3), 345–350. https://doi.org/10.1007/s00439-010-0861-0

Udupa, K.-B. (2006). Functional significance of erythropoietin receptor on tumor cells. *World Journal of Gastroenterology*, *12*(46), 7460–7462. https://doi.org/10.3748/wjg.v12.i46.7460

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), 59. https://doi.org/10.1038/s43586-021-00056-9

Urnov, F. D., Miller, J. C., Lee, Y.-L., Beausejour, C. M., Rock, J. M., Augustus, S., Jamieson, A. C., Porteus, M. H., Gregory, P. D., & Holmes, M. C. (2005). Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature*, *435*(7042), 646–651. https://doi.org/10.1038/nature03556

Urnov, F. D., Rebar, E. J., Holmes, M. C., Zhang, H. S., & Gregory, P. D. (2010). Genome editing with engineered zinc finger nucleases. *Nature Reviews. Genetics*, *11*(9), 636–646. https://doi.org/10.1038/nrg2842

van der Weyden, L., Adams, D. J., & Bradley, A. (2002). Tools for targeted manipulation of the mouse genome. *Physiological Genomics*, *11*(3), 133–164. https://doi.org/10.1152/physiolgenomics.00074.2002

VanderWeele, T. J., Tchetgen Tchetgen, E. J., Cornelis, M., & Kraft, P. (2014). Methodological challenges in mendelian randomization. *Epidemiology (Cambridge, Mass.)*, *25*(3), 427–435. https://doi.org/10.1097/EDE.0000000000000081

Vasan, R. S., Larson, M. G., Leip, E. P., Evans, J. C., O'Donnell, C. J., Kannel, W. B., & Levy, D. (2001). Impact of High-Normal Blood Pressure on the Risk of Cardiovascular Disease. *New England Journal of Medicine*, *345*(18), 1291–1297. https://doi.org/10.1056/NEJMoa003417

Våtsveen, T. K., Sponaas, A.-M., Tian, E., Zhang, Q., Misund, K., Sundan, A., Børset, M., Waage, A., & Brede, G. (2016). Erythropoietin (EPO)-receptor signaling induces cell death of primary myeloma cells in vitro. *Journal of Hematology & Oncology*, *9*(1), 75. https://doi.org/10.1186/s13045-016-0306-x

Verbanck, M., Chen, C.-Y., Neale, B., & Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature Genetics*, *50*(5), 693–698. https://doi.org/10.1038/s41588-018-0099-7

Verma, A., & Ritchie, M. D. (2017). Current Scope and Challenges in Phenome-Wide Association Studies. *Current Epidemiology Reports*, *4*(4), 321–329. https://doi.org/10.1007/s40471-017-0127-7

Vis, M. A. M., Ito, K., & Hofmann, S. (2020). Impact of Culture Medium on Cellular Interactions in in vitro Co-culture Systems . In *Frontiers in Bioengineering and Biotechnology* (Vol. 8). https://www.frontiersin.org/article/10.3389/fbioe.2020.00911

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics*, *101*(1), 5–22. https://doi.org/https://doi.org/10.1016/j.ajhg.2017.06.005

Vora, S., Tuttle, M., Cheng, J., & Church, G. (2016). Next stop for the CRISPR revolution: RNA-guided epigenetic regulators. *The FEBS Journal*, *283*(17), 3181–3193. https://doi.org/10.1111/febs.13768

Vuckovic, D., Bao, E. L., Akbari, P., Lareau, C. A., Mousas, A., Jiang, T., Chen, M.-H., Raffield, L. M., Tardaguila, M., Huffman, J. E., Ritchie, S. C., Megy, K., Ponstingl, H., Penkett, C. J., Albers, P. K., Wigdor, E. M., Sakaue, S., Moscati, A., Manansala, R., … Soranzo, N. (2020). The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell*, *182*(5), 1214-1231.e11. https://doi.org/10.1016/j.cell.2020.08.008

Walker, V. M., Davey Smith, G., Davies, N. M., & Martin, R. M. (2017). Mendelian randomization: a novel approach for the prediction of adverse drug events and drug repurposing opportunities. *International Journal of Epidemiology*, *46*(6), 2078–2089. https://doi.org/10.1093/ije/dyx207

Wallace, C. (2020). Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLOS Genetics*, *16*(4), e1008720. https://doi.org/10.1371/journal.pgen.1008720

Wallace, C. (2021). A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLOS Genetics*, *17*(9), e1009440. https://doi.org/10.1371/journal.pgen.1009440

Wang, C., Gong, B., Bushel, P. R., Thierry-Mieg, J., Thierry-Mieg, D., Xu, J., Fang, H., Hong, H., Shen, J., Su, Z., Meehan, J., Li, X., Yang, L., Li, H., Łabaj, P. P., Kreil, D. P., Megherbi, D., Gaj, S., Caiment, F., … Tong, W. (2014). The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nature Biotechnology*, *32*(9), 926–932. https://doi.org/10.1038/nbt.3001

Wang, D., Zhang, C., Wang, B., Li, B., Wang, Q., Liu, D., Wang, H., Zhou, Y., Shi, L., Lan, F., & Wang, Y. (2019). Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. *Nature Communications*, *10*(1), 4284. https://doi.org/10.1038/s41467-019-12281-8

Wang, Gang, Yang, L., Grishin, D., Rios, X., Ye, L. Y., Hu, Y., Li, K., Zhang, D., Church, G. M., & Pu, W. T. (2017). Efficient, footprint-free human iPSC genome editing by consolidation of Cas9/CRISPR and piggyBac technologies. *Nature Protocols*, *12*(1), 88–103. https://doi.org/10.1038/nprot.2016.152

Wang, Gao, Sarkar, A., Carbonetto, P., & Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*(5), 1273–1300.

Wang, H., La Russa, M., & Qi, L. S. (2016). CRISPR/Cas9 in Genome Editing and Beyond. *Annual Review of Biochemistry*, *85*(1), 227–264. https://doi.org/10.1146/annurev-biochem-060815-014607

Wang, L., Di, L., & Noguchi, C. T. (2014). Erythropoietin, a novel versatile player regulating energy metabolism beyond the erythroid system. *International Journal of Biological Sciences*, *10*(8), 921–939. https://doi.org/10.7150/ijbs.9518

Wang, T., Wei, J. J., Sabatini, D. M., & Lander, E. S. (2014). Genetic Screens in Human Cells Using the CRISPR-Cas9 System. *Science*, *343*(6166), 80–84. https://doi.org/10.1126/science.1246981

Wang, W., Koka, V., & Lan, H. Y. (2005). Transforming growth factor-beta and Smad signalling in kidney diseases. *Nephrology (Carlton, Vic.)*, *10*(1), 48–56. https://doi.org/10.1111/j.1440-1797.2005.00334.x

Wang, X., Wang, H., Lu, J., Feng, Z., Liu, Z., Song, H., Wang, H., Zhou, Y., & Xu, J. (2020). Erythropoietin-Modified Mesenchymal Stem Cells Enhance Anti-fibrosis Efficacy in Mouse Liver Fibrosis Model. *Tissue Engineering and Regenerative Medicine*, *17*(5), 683–693.

https://doi.org/10.1007/s13770-020-00276-2

Wang, Y., Nudel, R., Benros, M. E., Skogstrand, K., Fishilevich, S., iPSYCH-BROAD, Lancet, D., Sun, J., Hougaard, D. M., Andreassen, O. A., Mortensen, P. B., Buil, A., Hansen, T. F., Thompson, W. K., & Werge, T. (2020). Genome-wide association study identifies 16 genomic regions associated with circulating cytokines at birth. *PLOS Genetics*, *16*(11), e1009163. https://doi.org/10.1371/journal.pgen.1009163

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

Watts, D., Gaete, D., Rodriguez, D., Hoogewijs, D., Rauner, M., Sormendi, S., & Wielockx, B. (2020). Hypoxia Pathway Proteins are Master Regulators of Erythropoiesis. *International Journal of Molecular Sciences*, *21*(21). https://doi.org/10.3390/ijms21218131

Weidemann, A., & Johnson, R. S. (2009). Nonrenal regulation of EPO synthesis. *Kidney International*, *75*(7), 682–688. https://doi.org/https://doi.org/10.1038/ki.2008.687

Weischenfeldt, J., Symmons, O., Spitz, F., & Korbel, J. O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature Reviews Genetics*, *14*(2), 125–138. https://doi.org/10.1038/nrg3373

Weiss, G., Ganz, T., & Goodnough, L. T. (2019). Anemia of inflammation. *Blood*, *133*(1), 40–50. https://doi.org/10.1182/blood-2018-06-856500

Wen, Y., Liao, G., Pritchard, T., Zhao, T.-T., Connelly, J. P., Pruett-Miller, S. M., Blanc, V., Davidson, N. O., & Madison, B. B. (2017). A stable but reversible integrated surrogate reporter for assaying CRISPR/Cas9-stimulated homology-directed repair. *Journal of Biological Chemistry*, *292*(15), 6148–6162. https://doi.org/https://doi.org/10.1074/jbc.M117.777722

Wetzels, J. F. M., Kiemeney, L. A. L. M., Swinkels, D. W., Willems, H. L., & Heijer, M. de. (2007). Age- and gender-specific reference values of estimated GFR in Caucasians: The Nijmegen Biomedical Study. *Kidney International*, *72*(5), 632–637. https://doi.org/10.1038/sj.ki.5002374

Wiedenheft, B., Sternberg, S. H., & Doudna, J. A. (2012). RNA-guided genetic silencing systems in bacteria and archaea. *Nature*, *482*(7385), 331–338. https://doi.org/10.1038/nature10886

Wilhelm, B. T., & Landry, J.-R. (2009). RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods (San Diego, Calif.)*, *48*(3), 249–257. https://doi.org/10.1016/j.ymeth.2009.03.016

Willer, C. J., Li, Y., & Abecasis, G. R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics (Oxford, England)*, *26*(17), 2190–2191. https://doi.org/10.1093/bioinformatics/btq340

Woodard, L. E., & Wilson, M. H. (2015). piggyBac-ing models and new therapeutic strategies. *Trends in Biotechnology*, *33*(9), 525–533. https://doi.org/10.1016/j.tibtech.2015.06.009

Wu, H., Liu, X., Jaenisch, R., & Lodish, H. F. (1995). Generation of committed erythroid BFU-E and CFU-E progenitors does not require erythropoietin or the erythropoietin receptor. *Cell*, *83*(1), 59–67. https://doi.org/10.1016/0092-8674(95)90234-1

Xie, Fei, Ye, L., Chang, J. C., Beyer, A. I., Wang, J., Muench, M. O., & Kan, Y. W. (2014). Seamless gene correction of β-thalassemia mutations in patient-specific iPSCs using CRISPR/Cas9 and piggyBac. *Genome Research*, *24*(9), 1526–1533. https://doi.org/10.1101/gr.173427.114

Xie, Fuliang, Xiao, P., Chen, D., Xu, L., & Zhang, B. (2012). miRDeepFinder: a miRNA analysis tool for deep sequencing of plant small RNAs. *Plant Molecular Biology*. https://doi.org/10.1007/s11103-012-9885-2

Xie, Z., Bailey, A., Kuleshov, M. V, Clarke, D. J. B., Evangelista, J. E., Jenkins, S. L., Lachmann, A., Wojciechowicz, M. L., Kropiwnicki, E., Jagodnik, K. M., Jeon, M., & Ma'ayan, A. (2021). Gene Set Knowledge Discovery with Enrichr. *Current Protocols*, *1*(3), e90. https://doi.org/10.1002/cpz1.90

Yamazaki, S., Hirano, I., Kato, K., Yamamoto, M., & Suzuki, N. (2021). Defining the functionally sufficient regulatory region and liver-specific roles of the erythropoietin gene by transgene complementation. *Life Sciences*, *269*, 119075. https://doi.org/https://doi.org/10.1016/j.lfs.2021.119075

Yang, H.-C., Zuo, Y., & Fogo, A. B. (2010). Models of chronic kidney disease. *Drug Discovery Today. Disease Models*, *7*(1–2), 13–19. https://doi.org/10.1016/j.ddmod.2010.08.002

Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Madden, P. A. F., Heath, A. C., Martin, N. G., Montgomery, G. W., Weedon, M. N., Loos, R. J., Frayling, T. M., McCarthy, M. I., Hirschhorn, J. N., Goddard, M. E., Visscher, P. M., Consortium, G. I. of An. T. (GIANT), & Consortium, Dia. G. R. A. M. (DIAGRAM). (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics*, *44*(4), 369–375. https://doi.org/10.1038/ng.2213

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, *88*(1), 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011

Yang, L., Guell, M., Byrne, S., Yang, J. L., De Los Angeles, A., Mali, P., Aach, J., Kim-Kiselak, C., Briggs, A. W., Rios, X., Huang, P.-Y., Daley, G., & Church, G. (2013). Optimization of scarless human stem cell genome editing. *Nucleic Acids Research*, *41*(19), 9049–9061. https://doi.org/10.1093/nar/gkt555

Yang, M., Su, H., Soga, T., Kranc, K. R., & Pollard, P. J. (2014). Prolyl hydroxylase domain enzymes: important regulators of cancer metabolism. *Hypoxia (Auckland, N.Z.)*, *2*, 127–142. https://doi.org/10.2147/HP.S47968

Yarmolinsky, J., Díez-Obrero, V., Richardson, T. G., Pigeyre, M., Sjaarda, J., Paré, G., Walker, V. M., Vincent, E. E., Tan, V. Y., Obón-Santacana, M., Albanes, D., Hampe, J., Gsur, A., Hampel, H., Pai, R. K., Jenkins, M., Gallinger, S., Casey, G., Zheng, W., … Moreno, V. (2022). Genetically proxied therapeutic inhibition of antihypertensive drug targets and risk of common cancers: A mendelian randomization analysis. *PLOS Medicine*, *19*(2), e1003897. https://doi.org/10.1371/journal.pmed.1003897

Ye, Z., Mayer, J., Ivacic, L., Zhou, Z., He, M., Schrodi, S. J., Page, D., Brilliant, M. H., & Hebbring, S. J. (2015). Phenome-wide association studies (PheWASs) for functional variants. *European Journal of Human Genetics : EJHG*, *23*(4), 523–529. https://doi.org/10.1038/ejhg.2014.123

Yeh, T.-L., Leissing, T. M., Abboud, M. I., Thinnes, C. C., Atasoylu, O., Holt-Martyn, J. P., Zhang, D., Tumber, A., Lippl, K., Lohans, C. T., Leung, I. K. H., Morcrette, H., Clifton, I. J., Claridge, T. D. W., Kawamura, A., Flashman, E., Lu, X., Ratcliffe, P. J., Chowdhury, R., … Schofield, C. J. (2017). Molecular and cellular mechanisms of HIF prolyl hydroxylase inhibitors in clinical trials. *Chemical Science*, *8*(11), 7651–7668. https://doi.org/10.1039/c7sc02103h

Young, M. D., Wakefield, M. J., Smyth, G. K., & Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*, *11*(2), R14.

https://doi.org/10.1186/gb-2010-11-2-r14

Yusa, K. (2013). Seamless genome editing in human pluripotent stem cells using custom endonuclease–based gene targeting and the piggyBac transposon. *Nature Protocols*, *8*(10), 2061–2078. https://doi.org/10.1038/nprot.2013.126

Yusa, K., Mick, C., & Craig, N. (2021). piggyBac Transposon. *Microbiology Spectrum*, *3*(2), 3.2.04. https://doi.org/10.1128/microbiolspec.MDNA3-0028-2014

Yusa, K., Rad, R., Takeda, J., & Bradley, A. (2009). Generation of transgene-free induced pluripotent mouse stem cells by the piggyBac transposon. *Nature Methods*, *6*(5), 363–369. https://doi.org/10.1038/nmeth.1323

Yusa, K., Rashid, S. T., Strick-Marchand, H., Varela, I., Liu, P.-Q., Paschon, D. E., Miranda, E., Ordóñez, A., Hannan, N. R. F., Rouhani, F. J., Darche, S., Alexander, G., Marciniak, S. J., Fusaki, N., Hasegawa, M., Holmes, M. C., Di Santo, J. P., Lomas, D. A., Bradley, A., & Vallier, L. (2011). Targeted gene correction of α1-antitrypsin deficiency in induced pluripotent stem cells. *Nature*, *478*(7369), 391–394. https://doi.org/10.1038/nature10424

Zeggini, E., & Ioannidis, J. P. A. (2009). Meta-analysis in genome-wide association studies. *Pharmacogenomics*, *10*(2), 191–201. https://doi.org/10.2217/14622416.10.2.191

Zeisberg, M., & Kalluri, R. (2015). Physiology of the Renal Interstitium. *Clinical Journal of the American Society of Nephrology : CJASN*, *10*(10), 1831–1840. https://doi.org/10.2215/CJN.00640114

Zeng, Y., Gong, M., Lin, M., Gao, D., & Zhang, Y. (2020). A Review About Transcription Factor Binding Sites Prediction Based on Deep Learning. *IEEE Access*, *8*, 219256–219274. https://doi.org/10.1109/ACCESS.2020.3042903

Zhang, J.-P., Li, X.-L., Li, G.-H., Chen, W., Arakaki, C., Botimer, G. D., Baylink, D., Zhang, L., Wen, W., Fu, Y.-W., Xu, J., Chun, N., Yuan, W., Cheng, T., & Zhang, X.-B. (2017). Efficient precise knockin with a double cut HDR donor after CRISPR/Cas9-mediated double-stranded DNA cleavage. *Genome Biology*, *18*(1), 35. https://doi.org/10.1186/s13059-017-1164-8

Zhang, X.-H., Tee, L. Y., Wang, X.-G., Huang, Q.-S., & Yang, S.-H. (2015). Off-target Effects in CRISPR/Cas9-mediated Genome Engineering. *Molecular Therapy - Nucleic Acids*, *4*, e264. https://doi.org/https://doi.org/10.1038/mtna.2015.37

Zhang, Yi, Thamer, M., Stefanik, K., Kaufman, J., & Cotter, D. J. (2004). Epoetin requirements predict mortality in hemodialysis patients. *American Journal of Kidney Diseases*, *44*(5), 866–876. https://doi.org/https://doi.org/10.1053/j.ajkd.2004.08.002

Zhang, Yong, Duc, A.-C. E., Rao, S., Sun, X.-L., Bilbee, A. N., Rhodes, M., Li, Q., Kappes, D. J., Rhodes, J., & Wiest, D. L. (2013). Control of hematopoietic stem cell emergence by antagonistic functions of ribosomal protein paralogs. *Developmental Cell*, *24*(4), 411–425. https://doi.org/10.1016/j.devcel.2013.01.018

Zhang, Yuanyuan, Wang, L., Dey, S., Alnaeeli, M., Suresh, S., Rogers, H., Teng, R., & Noguchi, C. T. (2014). Erythropoietin action in stress response, tissue maintenance and metabolism. *International Journal of Molecular Sciences*, *15*(6), 10296–10333. https://doi.org/10.3390/ijms150610296

Zhao, G., Yang, W., Wu, J., Chen, B., Yang, X., Chen, J., McVey, D. G., Andreadi, C., Gong, P., Webb, T. R., Samani, N. J., & Ye, S. (2018). Influence of a Coronary Artery Disease-Associated Genetic Variant on FURIN Expression and Effect of Furin on Macrophage Behavior. *Arteriosclerosis, Thrombosis, and Vascular Biology*, *38*(8), 1837–1844. https://doi.org/10.1161/ATVBAHA.118.311030

Zhao, Q., Wang, J., Hemani, G., Bowden, J., & Small, D. S. (2018). *Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score*. https://arxiv.org/abs/1801.09652

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., & Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PloS One*, *9*(1), e78644. https://doi.org/10.1371/journal.pone.0078644

Zheng, J., Haberland, V., Baird, D., Walker, V., Haycock, P. C., Hurle, M. R., Gutteridge, A., Erola, P., Liu, Y., Luo, S., Robinson, J., Richardson, T. G., Staley, J. R., Elsworth, B., Burgess, S., Sun, B. B., Danesh, J., Runz, H., Maranville, J. C., … Gaunt, T. R. (2020). Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nature Genetics*, *52*(10), 1122–1131. https://doi.org/10.1038/s41588-020-0682-6

Zheng, Q., Wang, Y., Yang, H., Sun, L., Fu, X., Wei, R., Liu, Y. N., & Liu, W. J. (2021). Efficacy and Safety of Daprodustat for Anemia Therapy in Chronic Kidney Disease Patients: A Systematic Review and Meta-Analysis. *Frontiers in Pharmacology*, *11*, 2208. https://doi.org/10.3389/fphar.2020.573645

Zheng, W., Chung, L. M., & Zhao, H. (2011). Bias detection and correction in RNA-Sequencing data. *BMC Bioinformatics*, *12*(1), 290. https://doi.org/10.1186/1471-2105-12-290

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, *44*(7), 821–824. https://doi.org/10.1038/ng.2310

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., Montgomery, G. W., Goddard, M. E., Wray, N. R., Visscher, P. M., & Yang, J. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics*, *48*(5), 481–487. https://doi.org/10.1038/ng.3538

Zhu, Z., Zheng, Z., Zhang, F., Wu, Y., Trzaskowski, M., Maier, R., Robinson, M. R., McGrath, J. J., Visscher, P. M., Wray, N. R., & Yang, J. (2018). Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nature Communications*, *9*(1), 224. https://doi.org/10.1038/s41467-017-02317-2

Ziello, J. E., Jovin, I. S., & Huang, Y. (2007). Hypoxia-Inducible Factor (HIF)-1 regulatory pathway and its potential for therapeutic intervention in malignancy and ischemia. *The Yale Journal of Biology and Medicine*, *80*(2), 51–60. https://pubmed.ncbi.nlm.nih.gov/18160990

Zmajkovic, J., Lundberg, P., Nienhold, R., Torgersen, M. L., Sundan, A., Waage, A., & Skoda, R. C. (2018). A Gain-of-Function Mutation in EPO in Familial Erythrocytosis. *New England Journal of Medicine*, *378*(10), 924–930. https://doi.org/10.1056/NEJMoa1709064

Zumbrennen-Bullough, K., & Babitt, J. L. (2013). The iron cycle in chronic kidney disease (CKD): from genetics and experimental models to CKD patients. *Nephrology Dialysis Transplantation*, *29*(2), 263–273. https://doi.org/10.1093/ndt/gft443

Zwaka, T. P., & Thomson, J. A. (2009). *Chapter 46 - Homologous Recombination in Human Embryonic Stem Cells* (R. Lanza, J. Gearhart, B. Hogan, D. Melton, R. Pedersen, E. D. Thomas, J. Thomson, & I. B. T.-E. of S. C. B. (Second E. Wilmut (eds.); pp. 417–422). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-374729-7.00046-9