THE UNIVERSITY
of EDINBURGH

Cortisol Responsive Gene Networks in Cardiovascular Disease

THE UNIVERSITY *of* EDINBURGH

Sean Alexander Bankier

The University of Edinburgh

Doctor of Philosophy with Integrated Study

2021

Declaration

I declare that this thesis has been composed by myself and that it has not been submitted for any other degree or professional qualification. I confirm that the work presented here is my own, with the exception of joint-author publications, where my own contribution and the contributions of co-authors has been explicitly defined. Any inclusion of work conducted by others has been appropriately referenced and credited.

21st June 2021

Abstract

Increased plasma cortisol levels are associated with Cardiovascular Disease (CVD) and CVD risk factors, however the tissue specific mechanisms underpinning this process are poorly understood. A genome wide meta-analysis by the CORtisol NETwork (CORNET) consortium identified genetic variants, spanning the *SERPINA6*/*SERPINA1* locus on chromosome 14, associated with morning plasma cortisol and shown to be causal for ischaemic heart disease. *SERPINA6* encodes Corticosteroid Binding Globulin (CBG), responsible for binding most cortisol in blood and putatively mediating delivery of cortisol to target tissues. This thesis addresses the hypothesis that genetic variants in *SERPINA6* influence CBG expression in liver and cortisol delivery to extra-hepatic tissues, responsible for mediating cortisol-regulated gene expression.

The Stockholm Tartu Atherosclerosis Reverse Networks Engineering Task study (STARNET) provides RNA sequencing data for 7 vascular and metabolic tissues from 600 genotyped individuals (mean age 65.8, 70.3% male) undergoing coronary artery bypass grafting. We have identified 21 Single Nucleotide Polymorphisms (SNPs) associated with both variation for plasma cortisol at genome wide significance in CORNET ($p \leq 5\times10^{-8}$) and with *SERPINA6* gene expression in STARNET-liver ($q \leq 0.05$) as cis-expression Quantitative Trait Loci (eQTLs). We go on to describe the extra-hepatic consequences of genetic variation for plasma cortisol by linking SNPs associated with plasma cortisol to genes expressed in trans across STARNET tissues, finding the highest representation of trans-genes in liver, subcutaneous and visceral abdominal adipose tissue (FDR = 15%). Through the use of published evidence, we

then identify a sub-set of cortisol associated trans-genes that are putatively regulated by the glucocorticoid receptor, the primary transcription factor activated by cortisol.

Using causal methods, we have identified glucocorticoid regulated trans-genes that are responsible for the regulation of tissue specific gene networks. Cis-eQTLs were used as genetic instruments for the identification of pairwise causal relationships, from which cortisol associated gene networks could be reconstructed. Gene networks were identified in liver, subcutaneous fat and visceral abdominal fat, including a high confidence gene network specific to subcutaneous adipose (FDR < 10%) under the regulation of the interferon regulatory transcription factor, *IRF2*. Targets in this network include *LDB2* and *LIPA*, both associated with coronary artery disease. Finally, we identify coordinated patterns of gene expression, representative of the STARNET gene networks, in gene expression data from the Metabolic Syndrome in Men (METSIM) and the Stockholm Atherosclerosis Gene Expression (STAGE) Study, both independent datasets from STARNET.

This thesis describes genetic variation at the *SERPINA6*/ *SERPINA1* locus that is associated with changes in both morning plasma cortisol and *SERPINA6* expression in liver, the gene that encodes CBG. Altered CBG levels in turn impact gene expression in extra-hepatic tissues through modulation of cortisol delivery. This supports a dynamic role for CBG in modulating cortisol delivery to tissues. The cortisol-responsive gene networks identified here represent candidate pathways to mediate cardiovascular risk attributable to elevated cortisol.

Lay summary

Cortisol is a steroid hormone that is produced by the adrenal glands and released in response to stress, thereby regulating a variety of biological functions. However, if cortisol levels are not properly controlled this can result in the development of disease. Individuals with high levels of cortisol over a prolonged period may go on to develop Cushing's syndrome which can result in high blood pressure, high blood glucose, and obesity. Moreover, subtle elevations in cortisol levels in otherwise healthy individuals has been linked to high blood pressure, type II diabetes and increased risk of heart disease. Variation in cortisol levels in the population has been attributed, in part, to genetic changes in a gene known as *SERPINA6*.

This PhD thesis describes how genetic variation in *SERPINA6* can be linked to molecular changes within networks of genes. Using large genetic datasets, variants in the genetic code for *SERPINA6* were identified that are associated with changes in levels of *SERPINA6* in liver. These genetic variants are also associated with changes in the activity of networks of different genes in fat, where cortisol influences body weight and metabolism. By understanding how these networks work, this provides insight into how our genes influence cortisol levels in the body and how this may lead to disease.

Acknowledgements

I am greatly appreciative of the many people who have aided and supported me both academically and personally, making this PhD a truly enriching and enjoyable experience.

I would like to thank Johan Björkegren for his generous access to the STARNET cohort, without which this project would not have been possible, as well as other members of the Björkegren lab and the scientific computing team at Mount Sinai School of Medicine.

I am indebted to the developers of the many open access computational tools used throughout this project, in particular to Lingfei Wang and Tom Michoel for the development of the causal inference package Findr. Also to the teams behind the many Python libraries, including SciPy, Pandas and Seaborn.

I have benefited greatly from the work of Andrew Crawford and all other analysts and principal investigators of the CORNET consortium whose painstaking work has made it possible to dissect the role of genetic variation for cortisol. I am grateful to the Medical Research Council and The University of Edinburgh for the funding of my PhD studentship. I also wish to thank all of the patients who generously contributed to the many studies that were used over the course of this project.

I am greatly appreciative of members of the human steroid signalling group at the University of Edinburgh Centre for Cardiovascular Science; Ruth Morgan for kindly providing access to murine RNA-seq data, Elisa Villalobos for her time and patient guidance conducting cell culture experiments and Marisa Magennis for her tireless

help with the various administrative and technical issues encountered. My thanks also to Allende Miguelez-Crespo, Mark Nixon, Rebecca Reynolds, Kerri Devine, Lisa Ivatt, Natalie Homer, Clare MacLeod, Jo Simpson, Roland Stimson, Scott Denham for contributing to the lively discussion during our lab meetings. I also wish to thank the former Michoel lab members Pau Erola, Siddharth Jayaraman (and Lingfei again) for their advice and guidance while introducing me into the world of bioinformatics.

I particularly want to thank all of my supervisors: Ruth Andrew for her patient and invaluable feedback for written work as well as sound advice throughout my PhD; Brian Walker for his guidance, both academic and career orientated, and for pushing me toward new collaborations and opportunities; Tom Michoel for his unwavering mentorship and support that has allowed me to develop as a scientist. I would also like to thank Tom and Brian together, for their work in conceptualising an exciting cross-disciplinary project that has been a privilege to be a part of.

A huge debt of gratitude is owed my parents for continued support and understanding throughout my lengthy academic journey and to my friends who have helped keep me sane throughout. My final thanks goes out to Erin for her encouragement, patience and for our late night science discussions while working through the problems of the day.

Table of Abbreviations

Abbreviation	Definition
ABC	ATP-binding cassette
ACTH	Adrenocorticotropic Hormone
CAD	Coronary Artery Disease
CBG	Corticosteroid Binding Globulin
ChIP-seq	Chromatin Immunoprecipitation sequencing
CIT	Causal Inference Test
CMD	Cardiometabolic Disease
CORNET	Cortisol Network
CVD	Cardiovascular Disease
DAVID	Database for Annotation, Visualization and Integrated Discovery
dbGaP	Database of Genotypes and Phenotypes
ENCODE	The Encyclopedia of DNA Elements
ER	Endoplasmic Reticulum
FDR	False Discovery Rate
FFL	Feed-Forward Loop
Findr	Fast Inference of Networks from Directed Regulations
GO	Gene Ontology
GR	Glucocorticoid Receptor
GRE	Glucocorticoid Response Element
GTEx	Genotype-Tissue Expression project
GWAMA	Genome Wide Meta Analysis
GWAS	Genome Wide Association Studies
HMDP	Hybrid Mouse Diversity Panel
HPA	Hypothalamic–Pituitary–Adrenal (axis)
IRF	Interferon Regulatory Factor
IV	Instrumental Variable
LCMS	likelihood-based causality model selection
LD	Linkage Disequilibrium
LLR	Log Likelihood Ratio
MAF	Minor Allele Frequency
METSIM	Metabolic Syndrome in Men study
MR	Mineralocorticoid Receptor
PCA	Principal Component Analysis
QTL	Quantitative Trait Loci
RMA	Robust Microarray Average
SERPIN	Serine Protease Inhibitor
SGBS	Simpson–Golabi–Behmel Syndrome
SNP	Single Nucleotide Polymorphism
STAGE	Stockholm Atherosclerosis Gene Expression study
STARNET	Stockholm Tartu Atherosclerosis Reverse Networks Engineering Task study
TPM	Transcripts Per Million

Contents

Chapter 1

Introduction

1.1 Thesis scope

Since the inception of Genome Wide Association Studies (GWAS), nearly two decades ago[1], there has been a steady expansion in the number of studies conducted as well as the sample size, yielding a wealth of new genetic associations with complex traits and disease. This approach has offered many new opportunities in endocrinology, where hormonal networks are well understood and hence lend themselves to informed mapping approaches. However, loci identified by GWAS alone are insufficient to elucidate the mechanisms by which these traits emerge[2] and efforts to understand the biology underpinning these associations has proved to be a significant challenge.

Much of the genetic research related to hormones has focused on monogenic endocrine disorders with scope for clinical intervention through genetic testing schemes. Examples include Autoimmune Polyglandular Syndrome Type 1[3] and IPEX syndrome[4] involving germline mutations within the *AIRE* and *FOXP3* genes respectively. However many genetic variants that contribute to different endocrine disorders and risk factors arise from common variants identified from GWAS. Natural genetic variation is a fundamental component of the development of endocrine phenotypes and dis-

orders, however it has been a challenge to obtain mechanistic understanding from genetic associations alone. Steroid profiles have been exploited as GWAS traits to identify regions of the genome associated with changes in hormone levels (Table 1.1), but the downstream consequences of these variants are still poorly understood.

Trait	Discovery sample	Ancestry	Publication year
Plasma cortisol[5]	12597	European	2014
Testosterone levels in polycystic ovary syndrome[6]	957	European	2015
Saliva cortisol[7]	7703	European	2017
Plasma cortisol[7]	7667	European	2019
DHEAS[7]	7667	European	2019
Estradiol[7]	7667	European	2019
Testosterone[7]	7667	European	2019
Plasma cortisol[8]	25,314	European	2021

Table 1.1: Notable publicly available GWAS and Genome Wide Meta Analyses (GWAMA) for endocrine factors.

For medical conditions that arise via a complex disease aetiology, a single mutation is often insufficient to translate the disease genotype to an observable phenotype. Unlike Mendelian disorders such as Cystic Fibrosis[9] and Huntington's disease[10], it is the cumulative effect of causal risk variants and their interactions with environmental factors that is required to cross the threshold of complex disease phenotype. These effects are often not as simple as an additive model and it is important to consider the epistatic effects of genetic variation. This is where a complex trait may be reliant upon the interaction of multiple gene loci[11], as in the case with certain metabolic disorders such as obesity[12] and type II diabetes[13], with known endocrine involvement.

There is an unmet need to dissect the influence of complex genetic variation within highly dynamic systems through the incorporation of multi-omic datasets. This is particularly prevalent within endocrine systems, where hormone and related metabolite levels vary in response to environmental and genetic perturbations[14,15]. Cortisol is the most prominent hormone in the glucocorticoid class of steroid hormones and is involved in mediating the stress response, as well as influencing biological functions relating to metabolism and the immune response[16].

However, variation in plasma cortisol levels has been linked to Cardiovascular disease (CVD)[7,8] and CVD risk factors such as hypertension[17] and type II diabetes[18].

A Genome Wide Meta Analysis (GWAMA) carried out by the CORtsiol NETwork (CORNET) consortium has identified genetic variation associated with morning levels of cortisol in human plasma[5,8]. This thesis aims to go beyond genome level observations, by examining how genetic variation at the GWAMA locus may have functional implications for cortisol signalling. Through the integration of multi-tissue gene expression data, we present a systems level framework for the consequences of plasma cortisol variation. Steroid signalling is well suited to such an approach, as steroid hormone networks have been well characterised and are mediated through transcriptional changes, hence lending themselves to a systems genetics approach. As well as linking gene expression to plasma cortisol associated genetic variation, we identify regulatory gene networks being driven by key glucocorticoid responsive genes. These gene networks were reconstructed using causal inference methodologies and provide a nuanced understanding of trait linked genetic variation.

In this introduction, we set out to describe how GWAS have been used to identify genetic variants that are associated with complex traits, as well as discussing the limitations of association based approaches. The role of Quantitative Trait Loci (QTLs) in systems genetics is then discussed, including how these have been used to model the impact of genetic variation upon molecular phenotypes. We then introduce how causal inference methods have been used to overcome some of the limitations of GWAS and how these can be integrated for the reconstruction of causal molecular networks. Finally, we will then review the aspects of glucocorticoid biology that are most pertinent to this thesis, before stating the aims and hypotheses that this thesis will address.

1.2 Moving beyond GWAS

GWAS have been highly successful in identifying trait associated loci, however they are limited in their ability to ascribe function to genetic variation. Additionally, GWAS variants, mainly Single Nucleotide Polymorphisms (SNPs), that do cross the threshold of genome wide significance, have been shown to suffer from the "winner's curse"[19], where associations fail to replicate in independent cohorts. This can result from the overestimation of effect sizes[20], meaning follow up studies are underpowered as effect size calculations are based on these inflated effect sizes, highlighting the need for robust validation of loci.

The genetic drivers behind many common disease phenotypes present through complex multi-factorial models of inheritance[21], as is the case in instances of obesity[12], cardiovascular disease[22] and type II diabetes[23]. The identification of causal SNPs is further complicated by the presence of pleiotropy, the phenomenon whereby genetic variation can be seen to influence multiple phenotypic traits[24]. Pleiotropy has been shown to be highly prevalent across the human genome[25], with studies showing that up to 90% of trait associated loci are associated with multiple traits[26].

These limitations have encouraged the wider integration of multi-omic data to provide functional context for GWAS results, linking SNPs to intermediate molecular phenotypes using systems genetics approaches that consider the global response to genetic variation[27] (Figure 1.1). This is particularly relevant in the case of hormone associated genetic variation, as many hormones mediate signalling across tissues, through transcriptional changes and can be modelled using these systems based approaches.

As transcriptomic data has become more readily available from highly powered studies, there has been a drive to link SNPs to variation in gene expression as expression Quantitative Trait Loci (eQTLs). High throughput sequencing technologies such as RNA-seq have facilitated the analysis of gene expression on a genome wide scale, replacing SNP microarrays as the leading method for gene expression

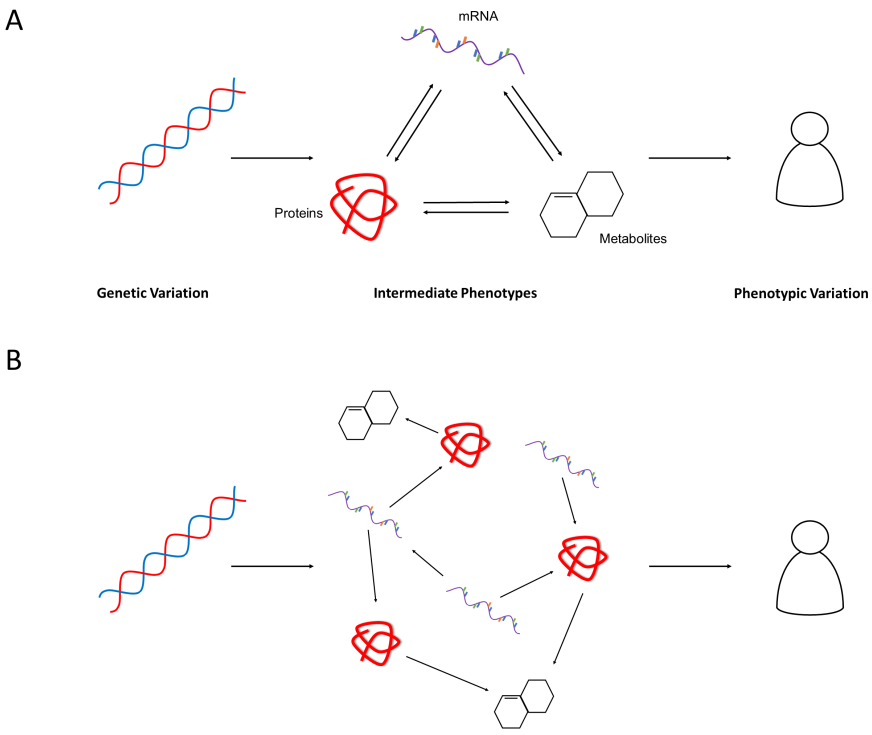


Figure 1.1: (A) Genetic variation (Left) influences complex traits (Right) through quantitative changes in intermediate phenotypes (Middle). Molecular interactions are shown as arrows, where the direction of the arrow indicates the direction of the flow of biological information. **(B)** Intermediate phenotypes can be modelled as biological networks using causal inference to uncover directed relationships between the molecular determinants that mediate the effect of genetic changes on complex traits.

analysis[28]. In conjunction with the emergence of large deeply genotyped cohorts, this has allowed for the mapping of eQTLs on an unprecedented scale. If SNPs associated with complex traits are also shown to be eQTLs, this may be interpreted as indicative of functional variation. However, causal relationships are required to translate statistical associations into biologically meaningful models[29]. Such approaches are crucial for deciphering gene-trait relationships, which is something that GWAS alone are unable to do.

During meiosis, alleles are randomly segregated within chromosomes during gamete production. This independent assortment ensures that alleles are randomly distributed across a given population, much in the same way that treatments are allocated during randomised controlled trials. Therefore, perturbing biological systems using the natural genetic variation present in a representative population[30,31]

makes it possible to examine the impact of genotypic variation given biological context. This provides an avenue to distinguish between direct and indirect consequences of common genetic variation[32], something that is not possible using traditional transgenic models[33–35] where the idea of experimentally replicating complex genotypes within a study population quickly becomes unfeasible at scale.

eQTL genotypes have the potential to establish causal relationships within biological data[36]. Systems genetics focuses on the flow of genetic information, from the gene level to downstream molecular changes. Biological information flows from eQTLs at the level of the genome to intermediate phenotypes in a unidirectional manner, allowing for the identification of causal relationships between genetic variation and downstream traits[37]. Variation in hormone levels influence a wide range of traits across multiple tissues[38], with a high risk of unobserved confounding affecting putative relationships obtained from association analysis. Understanding the direction of causal relationships is a necessary step to combat complex disease by distinguishing potential therapeutic targets from their downstream effects[39].

Biological networks reconstructed from multi-omic data that respond to both genetic and environmental perturbations, provide a more holistic view of how genetics may influence phenotype[27]. Causal inference methodologies are necessary to distinguish between causes, consequences and confounding factors of associations with genetic variation[40]. Such approaches can be used to transform a group of associations to a well-organised network of directed causal relationships[41].

It has been challenging to identify causal variants from GWAS results alone as a result of Linkage Disequilibrium (LD), resulting in the observation of the non-random inheritance of alleles at a given loci with SNPs that are in LD[42]. Therefore, if a true causal SNP for a trait is present and detectable at a given locus, the causal SNP and all other SNPs in LD will be identified as being associated with the trait in question, leading to an increase in type I error rate[43].

Efforts to understand the role genetic variation plays in regulating hormone levels, including cortisol, have been greatly enhanced by the emergence of GWAS. How-

ever, the tissue specific mechanisms underlying these associations cannot be truly uncovered by association based methodology alone. Likewise, traditional molecular based approaches are ill-suited to examine the impact of common SNPs, often with very small effect sizes. The type of approaches outlined in this chapter allow for integration of multi-omic data sources to develop directed causal relationships between intermediate phenotypes. These methods can be applied to uncover the tissue-specific mechanisms underpinning genetic variation associated with endocrine traits, including plasma cortisol variation.

1.3 Retrospective of eQTL studies

GWAS has exploded in popularity over the course of the last decade. As of 2018, the NHGRI-EBI GWAS Catalog lists 5687 studies for more than 71,000 traits[44]. However most of the genome-wide significant loci identified are of low to moderate penetrance, exacerbating the issue of missing heritability that has been predicted for complex traits[45–48]. This includes traits such as type II diabetes where only 10% of heritability is explained by the GWAS variants that have currently been identified[23], although twin and population studies estimate heritability to be between 20-80%[49]. Missing heritability, combined with the uniform distribution of GWAS hits across the genome, has even led to speculation of an "omnigenic" model of inheritance in which all genes in trait-related cells play a functional role in the resultant phenotype[45].

To overcome these limitations and garner greater benefit from GWAS datasets, there has been an increased focus to identify the functional and mechanistic consequences that are brought on by complex genetic variation. Linking GWAS loci to gene expression provides some indication of a functional relationship, and indeed data have demonstrated that trait-associated SNPs are more likely than non-trait-associated SNPs to also be associated with changes in gene expression (eSNPs)[50]. eSNPs describe the association between a single SNP with changes in gene expression whereas eQTLs are reflective of the association between a genetic locus and

gene expression.

Although association with changes in gene expression is not a direct proxy for function, this does help to better characterise systemic changes that are elicited in response to trait-associated genetic variation. This approach has been been used to expand understanding of conditions such as obesity. In 2015 researchers used a porcine model to first identify obesity related genes using a linear model, and then integrated gene expression data to find eQTLs for these genes within subcutaneous adipose tissue. Many of these eQTL linked genes identified were associated with lipid pathways, which highlights the capability of eQTL based approaches to generate functional findings that can be targeted by experimentalists[51]. It is important in such studies to consider external variables when sampling, as unlike genotypes, gene expression is highly variable. Factors such as time of day among other covariates, can introduce bias into any association study that can lead to findings that are not reflective of genotype related changes in gene expression.

eQTLs are categorised on the basis of their proximity to the gene locus with which they are associated (Figure 1.2). This distinction is important as it provides insight into the mechanisms by which an eQTL mediates an effect on gene expression. Cis-eQTLs, located close to their associated gene, are more likely to be acting locally than those located further away. Typically this distance is defined as being within 1 Mb of the associated transcription start site[52]. Outside of this threshold, eQTLs are said to be acting distally with associated genes in trans. Trans-eQTLs can be associated with genes located several megabases away including those on other chromosomes.

If an eQTL is cis-acting, this is likely to suggests a physical interaction between the eQTL and the associated gene. For example, a cis-eQTL sitting in an enhancer region may facilitate either an increased or decreased affinity for binding with a transcription factor[53,54]. Trans-eQTLs, on the other hand, associated with a distal gene, may influence transcription indirectly through an intermediary gene product or working in conjunction with local cis-eQTLs[55]. Most gene regulation takes place

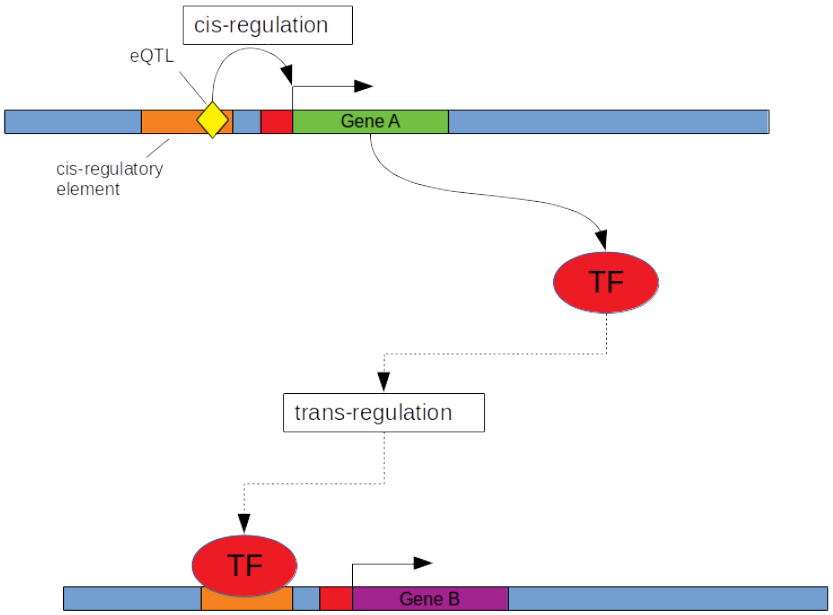


Figure 1.2: Cis and trans gene regulation. Gene A (green) encodes a transcription factor (TF) which regulates the expression of gene B (purple). The eQTL (yellow), acts as a cis-eQTL for gene A by causing a change in the sequence of gene A's cis-regulatory element (orange) which may either increase or decrease the binding affinity of any corresponding TFs. The same eQTL is a trans-eQTL for Gene B as by changing the expression of the TF encoded by gene A, this in turn influences the expression of gene B.

in cis, within regulatory regions and this is reflected by cis signals appearing more strongly than trans effects from eQTL mapping studies[56].

Multiple reproducible trans-eQTLs have been identified, often located in non-coding regions of the genome. Some of these trans-eQTL signals influence the expression of multiple genes, leading to the identification of regions of the genome termed as eQTL hot spots[57]. As these mechanisms are indirect, this presents a mechanism for genetic variation influencing hormone mediated transcription factors, which will then in turn influence the expression of trans-associated genes. Examples of this include a 2020 study where researchers interested in genomic loci associated with intramuscular fat, identified nine eQTL hot spot regions associated with intramuscular fat and harbouring transcription factors involved in lipid metabolism[53].

The study of eQTLs has been greatly aided by the advancement in high through-

put sequencing technologies such as RNA-seq. As one of the limiting factors for large scale eQTL analysis has been sample size, reduced sequencing costs have alleviated this limitation in conjunction with reduced costs for high performance computing to analyse millions of SNP-gene combinations simultaneously[58]. This also facilitates the processing of a wider variety of biological samples in terms of tissue or cell types. In turn this has led to the establishment of large database projects aiming to produce catalogues of eQTLs across multiple tissues (Table 1.2). Understanding the tissue specific context of eQTLs is a crucial step in understanding how these SNPs may influence phenotypic variation.

Study name	Number of samples	Description
GTEx	15,201	Multi-tissue dataset from post mortem donors[59].
HipSci	322	Induced pluripotent stem cells derived from healthy donors[60].
Geuvadis	445	Lymphoblastoid cell lines from five populations[61].
BLUEPRINT	554	Monocyte, neutrophils and T-cells from healthy donors[62].
TwinsUK	1,364	Fat, skin, blood and lympoblastic cell lines from twin pairs[63].
STARNET	3,786	Multi-tissue dataset from individuals undergoing surgery for coronary artery disease[64].
eQTLGen	31,684	Meta-analysis of cis and trans eQTL analysis in blood[65].

Table 1.2: Examples of major publicly available eQTL datasets.

One of the most comprehensive projects to produce an atlas of transcriptome wide genetic effects has been conducted by the GTEx consortium[59], who have generated tissue specific eQTL data from an impressive number of post-mortem samples. With the latest release of data from GTEx v8[66], the authors present both trans and cis associations from 49 different tissues. Genomes are consistent between cell types, but the way in which these genes are expressed varies drastically between tissues, and much of the regulation that mediates this disparity takes place at the transcriptome. Therefore, to effectively characterise the genetic architecture that regulates the transcriptome it is important to sample from as many different tissues as possible.

Due to the increase in available eQTL datasets, there has been a need for the compilation and curation of existing data. Differences in analysis and quantification methods, impedes certain downstream analyses such as colocalisation[67] which requires comparable summary statistics. Colocalisation refers to the colocalisa-

tion of trait-associated SNPs with changes in gene expression to ascribe function to GWAS hits and prioritise the identification of causal SNPs[68,69].

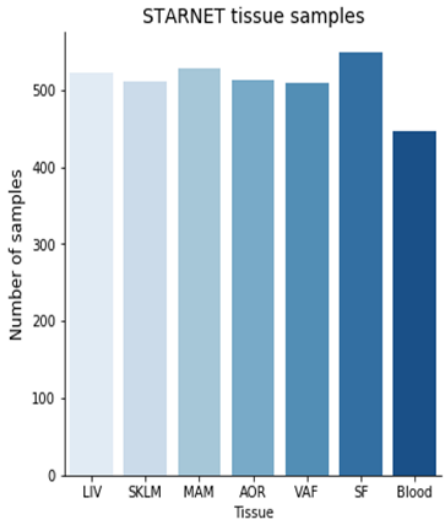
Although eQTL studies provide useful functional insight to GWAS hits, some important caveats remain. One question focuses upon the diversity of donor populations in both GWAS and eQTL studies. In GTEx the majority (85.3%) of donors have a European American ancestry which is common among many such studies[70], but may raise issues where findings are inferred to be generalisable without being fully representative of different human populations.

Another concern is that as transcriptional output does not perfectly correlate with proteomic output, the influence of many eQTLs at the protein level has been uncharacterised. More studies now aim to identify protein Quantitative Trait Loci[71] (pQTLs), genetic variants associated with protein expression, however proteomic analysis is more costly than RNA-seq and requires considerable sample sizes to remain robust. A 2018 study by Yao and colleagues[72] demonstrated colocalisation between eQTL and pQTL signals for 190 variants out of a total of 372 pQTLs. This shows that although there are signals consistent between transcript and protein levels, confounding factors such as protein turnover and protein: protein interactions means that changes in the transcriptome are not always directly reflected at the proteome.

In this thesis, we examined the transcriptomic impact of cortisol associated genetic variation using the Stockholm Tartu Atherosclerosis Reverse Networks Engineering Task (STARNET) study. Individuals in this study were genotyped and RNA sequencing was conducted for seven different vascular and metabolic tissue samples (Figure 1.3A), for 600 patients undergoing bypass surgery for Coronary Artery Disease (CAD). This resource has been used to identify tissue specific and cross tissue cis (proximal) and trans (distal) gene regulation associated with CAD[64].

This cohort was composed of Caucasian individuals from Eastern European origin (30% female), with a confirmed diagnosis of CAD (Figure 1.3B). Of these individuals 27% had diabetes, 77% had hypertension and 37% had suffered a myocardial

A



B

	variable	mean	SD	median	percent
A. Basic clinical characteristics	Age (years)	65.8	8.7	67	
	Gender (male, %)				70.3
	Body mass index (kg/m^2)	29.9	4.4	28.9	
	Waist/hip ratio (WHR)	0.98	0.07	0.9	
	Systolic blood pressure (mm Hg)	130.1	18.3	130	
	Diastolic blood pressure (mm Hg)	74.6	11.3	75	
	Pulse rate (beats/min)	74.4	12	74	
	Current smoking (%)				18.4
B. Basic blood chemistry	Hemoglobin (g/l)	137	16	138	
	White blood cell count (10^9/l)	7.15	1.9	7	
	Platelets (10^9/l)	236.1	75.3	224	
	Creatinine (µmol/l)	84	23.3	80	
	C-reactive protein (mg/l)	6.5*	11.8*	2	
	TSH (mU/l)	2.1*	5.3*	1.3	
	HbA1c (% units)	6.1	0.8	5.8	
	Blood glucose (mmol/l)	6.5	2	5.9	
	ASAT (U/l)	26.3	14	22	
	ALAT (U/l)	25.8	17.3	21	
	GGT (U/l)	36.8	31.6	27	
	Plasma cholesterol (mmol/l)	4.9	1.1	4.9	
	LDL (mmol/l)	2.9	1.2	2.8	
	HDL (mmol/l)	1.5	0.9	1.3	
	LDL/HDL ratio	2.3	1.3	2.2	
	Triglycerides (mmol/l)	1.6	0.1	1.3	
C. Current diagnoses	Hyperlipidemia (%)				68.3
	Hypertension (%)				76.7
	Type 2 diabetes (%)				27.2
D. Current drug therapies	Beta-blockers (%)				74
	Lipid lowering (%)				65.9
	Oral anti-diabetics (%)				19.4
	Insulin (%)				7.7
	Aspirin (%)				81.4
E. Heart function classification	Canadian Cardiovascular Society grading of angina pectoris (arbitrary unit)	2.2*	0.7*	2	
	New York Heart Association Functional Classification (arbitrary unit)	1.6*	0.7*	2	
F. Pre-operative angiographic assessment of CAD	Syntax score	44.3	16.9	44.5	
	Duke CAD Severity Score	65.7	13.8	67	
	Number of diseased coronary vessels	2.8*	2.0*	3	
	Number of coronary stenotic lesions	4.4*	1.6*	4	
H. Prior Evidence of CAD or stroke	Myocardial infarction (%)				37.1
	Prior Stroke (%)				7.6
	TIA				7.8
	Claudication				17.1
	Family history of heart attack (%)				33.5

* Data not normally distributed.

Figure 1.3: (A) Number of samples collected per tissue from STARNET. Tissue samples include; Subcutaneous Fat (SF), Internal Mammary Artery (MAM), Liver (LIV), Atherosclerotic Aortic Root (AOR), Skeletal Muscle (SKLM), Visceral Abdominal Fat (VAF) and Whole Blood (Blood). **(B)** Summary statistics of the STARNET cohort. Adapted from Franzen *et al* 2016[64]

infarction before 60. The whole blood samples were taken pre-operatively and the remaining tissue biopsies were obtained during open-heart surgery. This cohort presents a unique opportunity to systematically link common genetic variation for plasma cortisol to changes in gene expression in a representative cohort of individuals with a confirmed diagnosis of CAD.

There are caveats relating to the study design, that need to be considered when considering studies such as STARNET. Although the initial aim of the study was to identify changes in gene expression that could be associated with CAD, specially atherosclerosis, other comorbidities such as hypertension and diabetes may act as confounding factors for any associations. The fact that this is not a healthy population, also may lead to findings which are not generalisable to the broader population. Another issue to be considered is the application of such a study when examining the impact of genetic variation for plasma cortisol, as both the anticipatory and surgical stress response means that cortisol levels will be higher for samples

obtained during surgery compared to non-surgical conditions.

1.4 Causal inference in genetic epidemiology

A fundamental aim in genetic analysis has been to identify functional relationships between genetic variation and complex phenotypes, however this has not always been possible in a laboratory setting. Randomised controlled trials are often, rightly, held up as the gold standard for studying causal effects. However there are many instances when a randomised controlled trial is either unfeasible or unethical to carry out[73]. Fortunately, the random segregation of alleles during meiosis[36] offers a solution. Given a large enough sample size, it is possible to predict causal relationships between different phenotypes, using associated genotypes as genetic instruments for causal inference analysis.

Instrumental Variable (IV) analysis is a causal inference framework that has been applied to obtain causal relationships between biological traits (Figure 1.4). For this methodology, the IV is used to infer a causal relationship between an exposure and an outcome variable. IV analysis requires the following assumptions: 1) The IV should be robustly associated with the exposure. 2) The IV should only be causal for the outcome through the exposure[73]. 3) The IV should be independent from any confounding factors that are causal for the exposure or the outcome[74,75]. Given these assumptions, it is possible to use the IV as a proxy for the exposure to infer a directed relationship between the exposure and outcome.

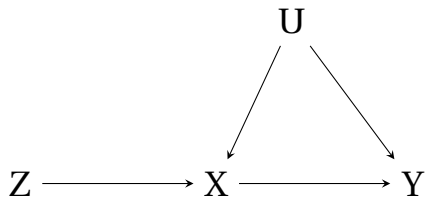


Figure 1.4: Instrumental Variable paradigm. The instrumental variable (Z) is causally associated with the exposure (X) which in turn is causally associated with the outcome (Y). The IV will account for any confounding (U) that affects the exposure or outcome, assuming independence of U.

Mendelian randomisation is a causal inference based approach that has become

increasingly popular in recent years as an extension of IV analysis. The randomisation refers to the way in which alleles randomly segregate from parent to offspring[73]. Mendelian Randomisation aims to address the missing functionality of GWAS, while using GWAS data, often now with just summary statistics, to infer causality between traits given the exposure-outcome paradigm. Given the flexibility of this model, it is possible to infer causality between gene expression and GWAS traits using eQTLs as IVs[37].

The use of genetic variants as instruments in IV analysis has become an important method for establishing causal relationships in biological systems where gene expression acts as an exposure. eQTLs have been shown to satisfy the IV assumptions through a robust association with gene expression, given the same eQTL is not also directly associated with the outcome. This also overcomes issues related to confounding as genetic variation is fixed at conception of the foetus and is therefore highly unlikely to be confounded by the same causal factors influencing downstream phenotypes[76]. Issues can arise when an eQTL is also directly associated with the outcome, however this can be overcome through careful instrument selection and testing for pleiotropy[77].

IV analysis overcomes some of the pitfalls that are present in other statistical techniques such as mediation analysis. Mediation analysis establishes a relationship between an exposure and an outcome variable through a third variable, the mediator. This assesses the direct and indirect relationship between the exposure and the outcome by conditioning on the mediator to calculate the probability of both the direct and indirect relationship[78]. A drawback is that mediation is susceptible to collider bias, resulting in confounding between the mediator and outcome. This occurs when the researcher conditions on the mediator, opening a backdoor path between the exposure and the outcome suggesting a causal relationship where one does not exist[79].

Other causal methods have focused on the identification of causal variants, as opposed to obtaining directed causal relationships. Statistical fine-mapping in-

tegrates data related to LD and the structural context of trait-associated SNPs to identify variants that are causal for a given trait[80]. Fine-mapping methods such as RASQUAL, aim to identify causal variants that result in changes in allelic imbalance in regulatory regions by examining the prevalence of chromatin accessibility QTLs (caQTLs) using ATAC-seq data[81].

Other causal approaches include colocalisation analysis. Most colocalisation methods have aimed to identify a shared causal variant in a comparison between traits, while taking into account LD. One of the most commonly used tools for colocalisation analysis, is the Bayesian method, Coloc, developed by Giambartolomei *et al*[82]. This method uses five hypothesis tests to estimate the probability of a shared causal variant against a null hypothesis and has become one of the foremost methods for colocalisation analysis. Colocalisation analysis is desirable as it allows for the identification of SNP peaks shared by multiple traits, including association with gene expression, and allows for the prioritisation of causal SNPs.

1.5 Reconstruction of causal gene networks

Jansen and Nap first proposed the integration of genomic information to identify changes in continuous molecular traits associated with the segregation of genotypes within a population in 2001[83]. What the authors originally describe as "genetical genomics", outlines a strategy to link genetic variation within a population to gene expression data, at the time obtained from microarray assays, and to other sources of expression data relating to proteins and metabolites. This has provided the foundation for modern day systems genetics[27], which allows for the integration of genetic and quantitative data with the ultimate aim of generating biological networks that can be linked to complex traits.

Most network based approaches to date have focused on correlation, through the development of co-expression networks using transcriptomic data[84]. Co-expression networks were first proposed in the 1990s[85] and have been used to identify novel

pathways in complex traits and disease as wide ranging as depression[86], muscular disease,[87] and CVD[88]. They have also been used to identify clusters of genes that are linked to different phenotypic characteristics for conditions such as endometriosis[89]. These methods are capable of reconstructing edges (connections between nodes) between co-expressed genes but are limited due to their inability to distinguish between different causal models (Figure 1.5).

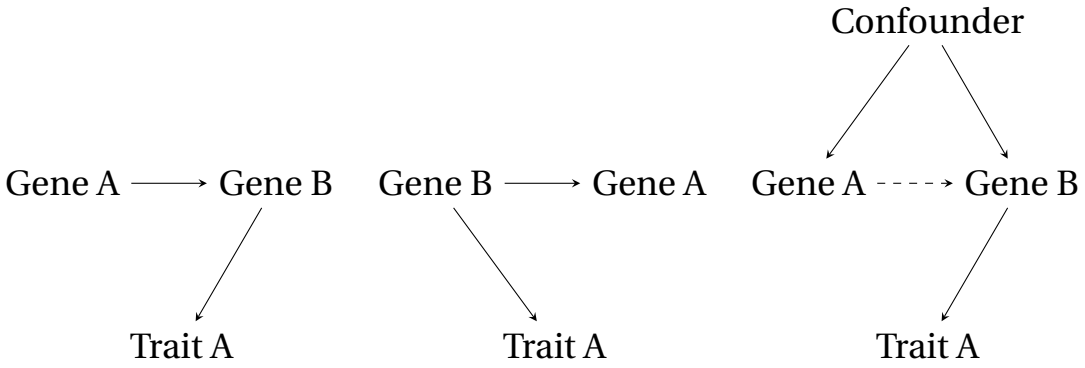


Figure 1.5: Causal modelling of pairwise gene-gene relationships. (Left) Simple causal model where Trait A is influenced by Gene A, through Gene B. (Middle) Reactive model where Gene B influences both Gene A and Trait A, therefore any association between Gene A and Trait A is a non-causal relationship. (Right) Association between Genes A and B is a result of unobserved confounding, therefore there is no causal relationship between Gene A and Trait A.

Pairwise gene-gene relationships are capable of providing a foundation for gene network reconstruction using sufficiently large transcriptomic datasets[90–92]. IV based methods can be used to obtain probability estimates for causal relationships between genes when provided with robust genetic instruments. A method that facilities this approach is the tool Findr, which incorporates eQTLs within an IV framework to obtain the Bayesian posterior probability of a causal relationship between a pair of genes, using a combination of likelihood ratio tests to account for any unobserved confounding[92].

Bayesian networks are acyclic graphs that have been used for modelling gene networks as they allow for the incorporation of prior knowledge and are capable of resolving issues of conditional independence in data[93,94]. Bayesian networks are developed using frequency tables from discrete data, however in cases of continuous data such as transcriptomic datasets, posterior probabilities can be calculated from density functions[95]. By obtaining posterior probabilities for pairwise relation-

ships between genes with tools such as Findr, it is possible to reconstruct networks of genes (nodes) that are connected by posterior probabilities (edges) at a given threshold (Figure 1.6).

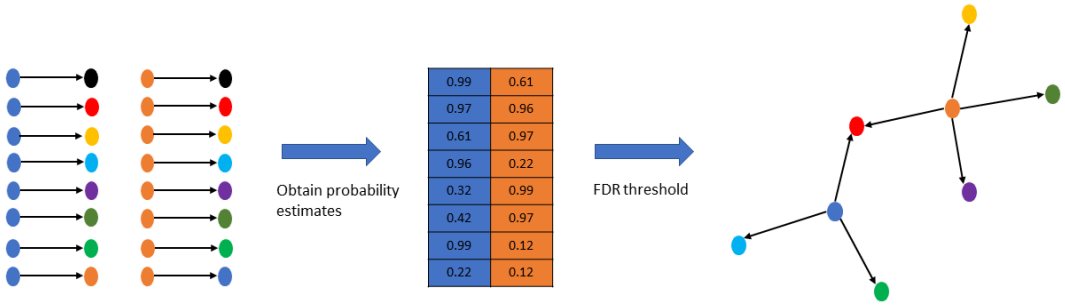


Figure 1.6: Reconstructing gene networks from pairwise relationships. (Left) Prospective pairwise relationships between genes with a robust eQTL (blue and orange) and other genes within a dataset. (Middle) Causal inference approaches are employed to obtain a probability matrix for the likelihood of a causal relationship between gene pairs. (Right) A filtering step is imposed e.g. a FDR cut-off, which will return relationships that cross this threshold to be assembled as directed networks.

There is a wealth of data relating to the role of gene regulation, including available cis-regulatory elements[96] and transcription factor binding sites[97]. The incorporation of these data allows for the construction of robust priors for Bayesian causal inference. eQTLs are also particularly well suited to filling this role and have been used to identify genes driving cardiovascular disease[88], type II diabetes[98] and Acute Myeloid Leukaemia[99] when combined with gene expression data.

An issue encountered within Bayesian network analysis, is that as the number of networks nodes increases so does the number of potential network edges. Given the high dimensional datasets commonly generated from next generation sequencing, standard Bayesian network methods are often computationally prohibitive[100]. Novel methods to overcome the computational burden include the use of eQTL and transcriptomic data within a node ordering approach which prioritises given relationships, reducing the number of possible networks[101].

Other types of omics data are also also amenable to this approach. Steroid profiling has its roots in the 1960s with the use of gas chromatography for separation[102] and advances in mass spectrometry have allowed for high resolution analysis of

ionised modules[103]. It is possible to link genotypes to changes in steroid levels, as evidenced by the number of GWAS conducted in this area (Table 1.1). Causal inference has the potential to integrate relationships between gene expression and metabolites, using eQTLs as instruments in the IV paradigm previously described. These methods have the potential to uncover upstream regulatory targets responsible for variation at the level of the metabolome by fully exploiting multi-omic datasets, with the potential to identify therapeutic targets that correspond to disease.

Analysis by Gallois *et al* in 158 serum metabolites, demonstrates that by integrating traditional association studies with genetic-metabolite networks, a small number of highly pleiotropic genes can be linked to causal variants responsible for mediating variation in metabolite profiles[104]. This is an example of how integration of multi-omic datasets can uncover regulatory targets responsible for downstream variation. However any analysis of SNP-metabolite associations will be subject to confounding due to the high number of variables that influence changes in metabolite levels, highlighting the need for causal inference.

There is further scope to elucidate the mechanisms of disease through the dissection of GWAS hits for complex disease, as demonstrated by Small and colleagues who were able to reconstruct networks of genes associated with SNPs linked to type II diabetes and mediated through the gene *KLF14*[105]. The researchers were able to show that cis-eQTLs for *KLF14* regulated a larger adipose specific gene network that was significantly enriched for metabolic pathways. This highlights the role of a network approach when combined with traditional genetic association and linkage studies.

There have also been attempts towards the incorporation of *in vivo* data within a systems based framework. The Hybrid Mouse Diversity Panel (HMDP) uses a mouse reference population to examine the cross-tissue impact of gene expression variation[106]. The researchers in this study used natural transcript variation to identify and functionally annotate endocrine circuits. Seldin *et al* began by screening for

any correlations between transcripts in their mouse cohort to identify endocrine factors that correspond to changes in target gene expression while filtering for tissue. They then used pathway enrichment to formulate hypotheses based on their statistical analysis, which were followed up by experimental validation.

Other approaches have aimed to integrate data from different sources to identify genes causal for cholesterol metabolism. In this 2020 study[107], Li *et al* used human lipid GWAS with murine liver co-expression networks to identify causal genes for cholesterol metabolism. The researchers began with the identification of cholesterol associated modules within liver networks in mice, followed by cross-referencing against human GWAS datasets. Such approaches help to prioritise results that can be taken forward for validation that would be unsuited to a high throughput experimental approach.

In this thesis, we have used a strategy that is similar to those that been outlined in this section, while taking causal relationships into account. By identifying pairwise relationships between genes, this has allowed for the reconstruction of causal gene regulatory networks. These networks aim to characterise the downstream transcriptomic consequences of genetic variation for plasma cortisol.

1.6 Cortisol and the glucocorticoid receptor

The biosynthesis of cortisol takes place in the adrenal glands[108], through activation of the hypothalamic–pituitary–adrenal (HPA) axis[109] in response to stress. The HPA axis is stimulated by Adrenocorticotropic hormone (ACTH) secretion, which results in the release of cortisol, which then inhibits the HPA axis as part of a negative feedback loop[110]. Outside of the stress response, cortisol levels follow a diurnal rhythm, peaking in the morning and then declining throughout the day[111,112]. However, cortisol levels have also been demonstrated to respond to ultradian rhythms within a 24-hour period, composed of an intrinsic pulsatility, resulting from delayed ACTH release and negative feedback regulation[113,114].

In addition to variation in plasma cortisol levels, cortisol is also modulated intracellularly by metabolising enzymes. 11β-HSD1/2 influence tissue availability of cortisol, with 11β-HSD1 involved in the conversion of cortisone to cortisol[115] and 11β-HSD2 inactivating cortisol to cortisone[116]. Additionally, transmembrane transporters are also involved in the modulation of intracellular cortisol, in particular the ATP-binding cassette (ABC) family of transporters. An example is the transporter *ABCB1*, which is involved in the export of cortisol, but not corticosterone across the blood brain barrier[117].

The primary mediator of glucocorticoid action in humans is the Glucocorticoid Receptor (GR), which is present in the vast majority of human cells and is highly conserved across vertebrates[118]. GR is a transcription factor and is involved in the regulation of different genes involved with metabolism, immune response and development[16]. GR modulates anti-inflammatory functions through the regulation of pro-inflammatory genes[119]. It also plays a crucial role in the metabolism of glucose[120] and fatty acids[121], as well as modulating the cytotoxic[122] and oxidative stress response[123].

When not bound to a ligand, GR is present in the cytoplasm, as part of a heat shock protein complex[124]. Once bound to a ligand such as cortisol, GR will translocate to the nucleus to modulate gene expression by binding to Glucocorticoid Response Elements (GREs) as a dimer[125], although GR can also act as a monomer which plays an important role in mediating tissue specific transcription by interacting with other DNA bound transcription factors[126]. Glucocorticoid binding to GR mostly takes place intracellularly, however there is evidence of membrane bound receptors with a lower affinity for glucocorticoids as a mechanism of non-genomic glucocorticoid action[127].

GR belongs to the nuclear receptor superfamily which includes the Mineralocorticoid Receptor (MR)[128]. MR has 10-fold greater binding affinity for cortisol than GR[129], which means that MR becomes occupied at low basal levels of cortisol, with GR being activated as glucocorticoid levels increase[130]. MR-GR heterodimers have

also been shown to form, which can bind to GREs and have different transactivation properties than their respective homodimers[131].

Although GR is expressed ubiquitously, there is significant variation in glucocorticoid sensitivity between tissues and among individuals[132], which can play a role in the development of disease. In metabolic syndrome, there are reports of altered GR mRNA observed in skeletal muscle[133] and in adipose[134], however these reports have been inconsistent, highlighting the role of complex regulatory pathways mediating transcriptional differences between tissues. Cushing's syndrome occurs in response to chronic activation of the HPA axis by increased ACTH secretion or by autonomous adrenal cortex cortisol release[135], resulting in insulin resistance, obesity and hypertension among other symptoms. Conversely, primary adrenal insufficiency or Addison's disease presents in response to a lack of cortisol as result of damage or destruction of the adrenal glands[136], resulting in fatigue, hypotension and weight loss.

1.7 Plasma cortisol and cardiovascular disease

Both Cushing's and Addison's are extreme examples of impaired HPA axis function. However, small but sustained changes in cortisol levels have been shown to be linked to complex disease phenotypes at a population level, including Cardiovascular Disease (CVD). CVD is a group of diseases, the presentation and severity of which are influenced by interactions between environmental, genetic and metabolic risk factors. The global impact of CVD across the world is striking: in 2012 approximately 17.5 million deaths were as a result of CVD which is the equivalent of 33% of all deaths worldwide[64]. Increased plasma cortisol levels have been demonstrated to be associated with the development of CVD risk factors such as hypertension[17] and type II diabetes[18]. However, the tissue specific mechanisms and relative contribution of cortisol to these risk factors are poorly understood.

CVD risk is elevated in patients with Cushing's and is a major cause of morbidity

and mortality for these individuals. Glucocorticoid excess increases the risk of CVD, which can be observed in patients who are prescribed synthetic glucocorticoids[137]. Placental environment and low birth weight has also been shown to be mediated through changes in HPA axis reactivity. Men born with a lower birth weight show increased responsiveness to ACTH as measured by urinary cortisol metabolite excretion, resulting in increased HPA axis activation which leads to raised blood pressure, glucose intolerance and hypertriglyceridemia[138].

A prospective cohort of 2512 men was studied to understand the contribution of CVD risk factors, including cortisol and testosterone, to CVD outcomes[139]. This study was able to identify links between cortisol: testosterone ratio and insulin resistance, in addition to incidents of ischaemic heart disease. A more recent study of 798 individuals, identified a positive association between morning plasma cortisol and incident of CVD, this finding was backed up by both a meta-analysis of published studies and two-sample Mendelian randomisation analysis[140].

1.8 Genetic variation for plasma cortisol

As discussed previously, cortisol levels are influenced by many factors including stress, temporal variation in response to circadian rhythms and tissue specific variation. However, there is inter-individual variation in basal levels of cortisol that can be attributed to genetic variation. The heritability of plasma cortisol levels has been estimated to be between 30-60%, although the genetic mechanisms underpinning this heritability are poorly understood[141]. Associations between plasma cortisol and CVD have likely been underestimated, due to the challenges in obtaining representative epidemiological samples. It is very difficult to obtain samples at a set time point, an issue that is further complicated by inter-individual variation in wake times.

This thesis is concerned with the genetic factors that mediate changes in plasma cortisol levels within a population and stems from the GWAMA by the CORNET con-

sortium which first identified genetic variants that are associated with changes in morning levels of plasma cortisol (n=12,597)[5]. The CORNET GWAMA was expanded to 25,314 individuals of European ancestry[8], however a single locus was identified that was associated with variation in plasma cortisol in both GWAMAs. This locus contains the genes *SERPINA6* and *SERPINA1*, which is notable as both genes are involved in cortisol biology.

There is previously reported evidence of cis-regulation at the *SERPINA1*/ *SERPINA6* locus in the form of a locus control region (LCR). LCRs are transcriptional regulators in mammalian systems, first discovered in the human β-globin locus[142], and responsible for the control of sets of co-regulated genes. In this case the SERPIN LCR has been shown to interact with multiple liver specific transcription factors[143] and deletion of key binding sites has been shown to impact both α-1 antitrypsin and CBG expression. The SERPIN LCR is located within the region ~8Kb upstream of *SERPINA1*[144] and contains 55 SNPs that were identified as being at genome wide significance in the CORNET GWAMA.

As part of the expanded GWAMA, SNPs associated with plasma cortisol were used as instruments for Mendelian randomisation to infer causal relationships between plasma cortisol and CVD risk factors. Causal relationships were established between plasma cortisol and increased risk of both ischaemic heart disease and myocardial infarction. In another GWAMA, causal relationships were identified between cortisol and Coronary Artery Disease (CAD)[7]. However, despite the presence of causal relationships between cortisol and CVD outcomes, little is known about the tissue specific mechanisms linking these factors.

1.9 Cortisol binding Globulin

Cortisol is released into the bloodstream following secretion from the adrenal cortex, where it can diffuse into tissues and bind to GR. However, most cortisol is not transported freely, as more than 80-90%[145,146] of cortisol in the blood is bound to

corticosteroid binding globulin (CBG) and unable to diffuse into tissues. This leads to the formation of different cortisol pools in the blood; with up to 90% of cortisol bound to CBG, 4-5% bound to albumin and a remaining 5% that is free and able to act in tissues[147] (Figure 1.7).

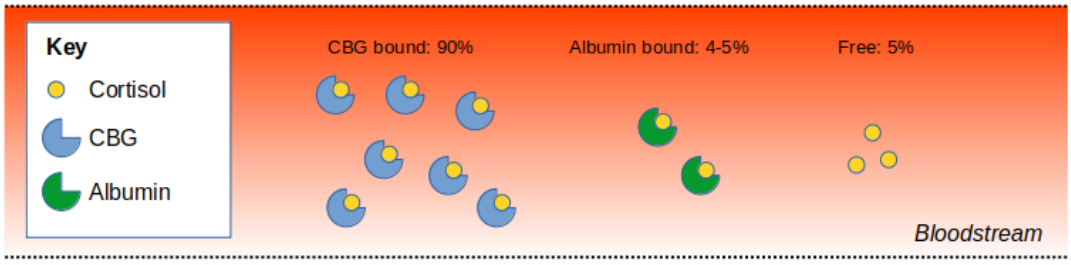


Figure 1.7: Depiction of different cortisol pools in the bloodstream.

Although the identification of SNPs associated with plasma cortisol does not reveal a mechanism linking genetic variation to changes in cortisol levels, the identification of the *SERPINA6*/ *SERPINA1* locus does indicate directions for investigation. *SERPINA6* is the gene that is responsible for encoding CBG. Additionally, *SERPINA1* encodes α-1 antitrypsin, which is responsible for the inhibition of neutrophil elastase, a serine protease responsible for cleavage of the reactive centre loop of CBG, resulting in a 9-10 fold reduction in binding affinity to cortisol[148,149].

CBG is primarily expressed in the liver, where it was first isolated revealing its structure[150]. CBG is a member of the serine protease inhibitor (SERPIN) superfamily of proteins, but has lost its inhibitory function while retaining the SERPIN protein structure[151]. CBG binds cortisol in the bloodstream, preventing cortisol diffusion to tissues, although the precise function of CBG in terms of acting as either a transporter of reservoir for cortisol is disputed[146,152].

Mutations affecting CBG have been implicated in CBG deficiency. CBG deficiency was first described in familial studies[153] and since then has been linked to both heterozygous and homozygous mutations in *SERPINA6*. A null mutation was first described in an Italian-Australian family, resulting in a complete loss of function of CBG resulting in chronic fatigue and relative hypotension[154]. Another study identified a patient who was heterozygous for a *de novo* mutation in *SERPINA6*,

which led to severe muscle fatigue and abnormally high salivary cortisol levels[155]. Simard and colleagues, examined the impact of 32 uncharacterised polymorphisms in *SERPINA6* and identified 8 naturally occurring CBG mutants, which result in changes to cortisol binding affinity, CBG production/ secretion and sensitivity to proteolytic cleavage[156].

Interactions between CBG and glucocorticoids have been studied in an experimental setting through the use of different animal models. Researchers have demonstrated that CBG-null mice, have a 10 fold increase in free corticosterone levels (corticosterone being the primary glucocorticoid in mice)[157]. This suggests that CBG plays a role in influencing negative feedback of the HPA axis, which in turn would influence tissue CORT levels. Additionally, BioBreeding rats which are used as a model for diabetes, have been shown to contain a variant of CBG which has a reduced binding affinity for cortisol[158].

Although model organisms are useful for examining the tissue specific impact of the type of mutations identified from familial studies, they are impractical to use for studying the impact of common genetic variants, as described by the CORNET GWAMA. The objective of this thesis has been to expand upon the identification of the *SERPINA6/ SERPINA1* locus, associated with variation in cortisol levels. By unpicking the impact of common genetic variation on gene expression across different tissues, it is possible to better understand the role of CBG expression in mediating GR responsive gene expression. This also helps to address the question of whether CBG influences tissue specific delivery of cortisol. As most genetic variation for cortisol is mediated through GR and therefore through altered transcription, this lends the study of genetic variation for plasma cortisol to a systems genetics approach.

1.10 Hypothesis and aims

We hypothesise that cortisol associated genetic variants in the SERPINA6/ SERPINA1 locus influence the levels of CBG in liver and cortisol delivery to extra-hepatic tissues, influencing GR mediated gene expression.

This hypothesis has been addressed through the following aims:

1. To determine if genetic variation associated with plasma cortisol influences *SERPINA6* expression in liver.
2. To investigate the role of any other genes trans-associated with genetic variation for plasma cortisol across different vascular and metabolic tissues in the STARNET cohort.
3. To identify GR responsive genes associated with genetic variation for plasma cortisol and to utilise causal inference methods to identify key regulator genes responsible for mediating the effects of genetic variation on downstream transcriptional networks.
4. To replicate any cortisol associated trans-genes or cortisol responsive gene networks in independent data sources.

1.11 Workflow

A complete workflow of the analyses presented in this thesis can be found at figure 1.8.

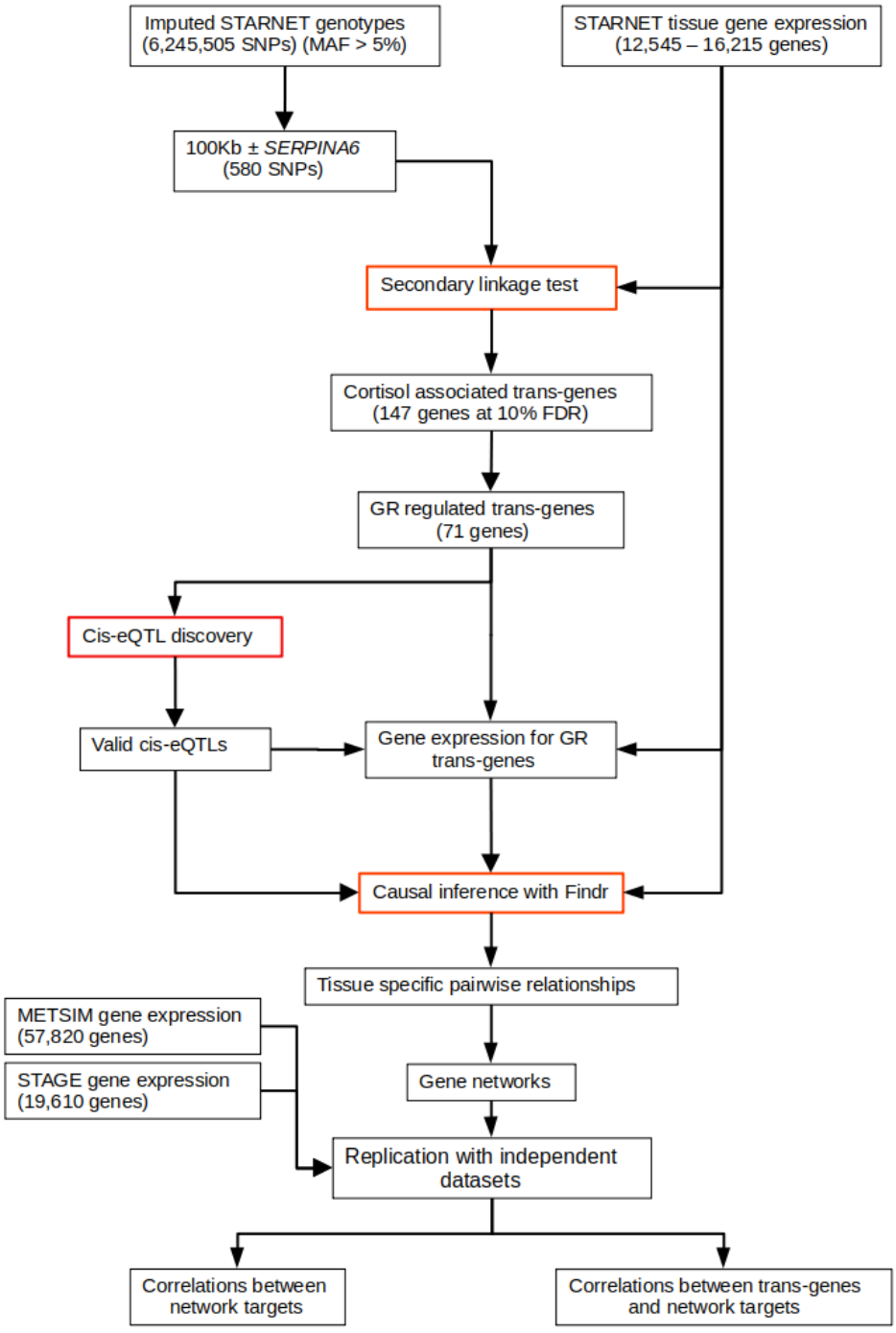


Figure 1.8: Complete workflow of analyses undertaken in this thesis. Workflow describes data inputs, intermediate results and methods. Red boxes indicate specific analyses.

Chapter 2

Cis-eQTL discovery and global tissue specific influence of CORNET SNPs

2.1 Introduction

2.1.1 Genetic variation and plasma cortisol

In their 2014 GWAMA[5], the CORNET consortium describes a single SNP peak on chromosome 14 spanning the *SERPINA6/ SERPINA1* locus associated with morning plasma cortisol. This was followed up with an expanded GWAMA, where the number of subjects was increased from 12,597 to 25,314 and the number of SNPs from ~2.2 M to ~7 M, across 17 population-based cohorts of European ancestries[8].

This expanded GWAMA confirmed the previously identified genetic association spanning the *SERPINA6/ SERPINA1* locus on chromosome 14 (Figure 2.1a) without identifying any new loci. In an additive genetic model, the top SNP, rs9989237, reported a per minor allele effect of 0.11 cortisol z-score ($p = 2.2\times10^{-19}$). The minor allele frequency of this SNP was 0.22, explaining 0.13% of the morning cortisol variance. The locus contained 73 SNPs that crossed the threshold of genome wide significance ($p \leq 5\times10^{-8}$) and contains 4 blocks of SNPs in low LD ($r^2 < 0.3$) (Figure

2.1b-c).

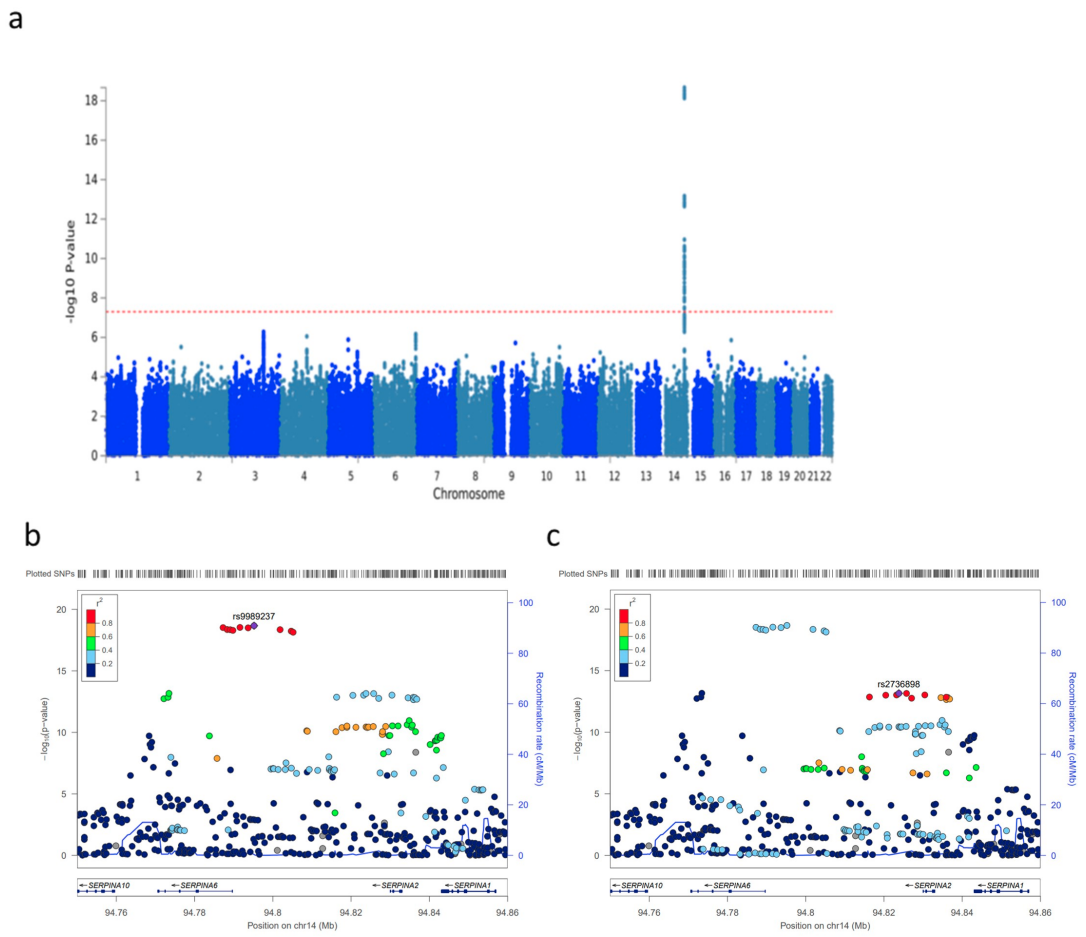


Figure 2.1: (a) Manhattan plot of -log10 P values of the SNP-based association analysis of morning plasma cortisol (n = 25,314). The locus on chr14 spans *SERPINA6* and *SERPINA1* genes; no other loci reached genome-wide significance. (b, c) Zoomed in Manhattan plot (LocusZoom plot) of -log10 P values of the SNP-based association analysis of morning plasma cortisol (n = 25,314). These show two (of the four) LD blocks ($r^2 > 0.3$) in this locus. Adapted from (Crawford, Bankier *et al*, 2021[8]).

Additionally, Mendelian Randomisation analysis was undertaken to identity causal relationships between plasma cortisol and CVD phenotypes. LD clumping yielded 4 SNPs, representative of each LD block, that were used as instruments for the cortisol associated loci. This analysis demonstrated that each standard deviation increase in morning plasma cortisol was associated with increased risk of chronic ischaemic heart disease (0.32, 95% CI 0.06–0.59) and myocardial infarction (0.21, 95% CI 0.00–0.43).

This study showed the *SERPINA6*/ *SERPINA1* locus to be associated with as plasma cortisol and provided evidence of a causal relationship between plasma cortisol and

CVD phenotypes. However, without understanding the impact of genetic variation on intermediate phenotypes such as gene expression, it is not possible to describe mechanistic links between cortisol linked SNPs and CVD. Additionally, is important to note that while the *SERPINA6*/ *SERPINA1* finding was retained, this remains a surprising result as cortisol influences the expression of many genes outside of this single locus.

In this chapter, we provide evidence of associations between cortisol linked SNPs at the *SERPINA6*/ *SERPINA1* locus and gene expression in both cis and trans. This includes findings that were presented in the publication of the 2021 CORNET GWAMA[8].

2.1.2 Plasma cortisol linked gene expression

The influence of cortisol, as mediated by GR, can be observed across many different tissue types. The primary goal of this chapter is to describe the impact of cortisol linked genetic variation upon tissue specific gene expression. The STARNET cohort is well suited to address the impact of genetic variation across tissue types, given the availability of multi-tissue gene expression data from ~600 genotyped individuals.

STARNET itself is an expansion of a prior study, the Stockholm Atherosclerosis Gene Expression (STAGE) Study, which examined the role of tissue specific gene expression variation across 114 individuals[159]. The STAGE authors initially carried out two way clustering across the transcriptional profiles in STAGE, with the aim of identifying clusters related to CAD development. This resulted in the identification of a module of genes represented by LIM domain binding 2 (*LDB2*), a transcription co-factor associated with CAD and atherosclerosis[159].

In later work, the STAGE authors carried out a global eQTL discovery analysis, identifying 8156 cis-eQTLs across the seven CAD relevant tissues present in the STAGE cohort[160]. Since then, STAGE has been used to identify cross-tissue co-expression networks, validated with data from the Hybrid Mouse Diversity Panel[88].

STARNET was established with the aim to link SNPs associated with Cardiometabolic

Disease (CMD) to variation in gene expression within a representative cohort of individuals with a confirmed diagnosis of CAD. In their initial analysis, the STARNET authors identified 2047 disease associated SNP that overlapped with a STARNET eQTL, moreover it was demonstrated that the cis-eQTLs identified in STARNET were enriched for GWAS associations with CAD and Alzheimer's disease.

The authors were able to identify 562 risk SNPs for CAD and Alzheimer's disease that also had cis-eQTL in a STARNET tissue. They also identified correlations between cis and trans-genes, revealing 37 cis-genes and 994 trans-genes connected in a cross tissue regulatory network for CAD. The authors followed up by demonstrating that the trans-genes in this network were enriched for CAD and atherosclerosis. Additionally, they identified tissue specific regulatory regions, driven by GWAS SNPs, including linking variation in coronary artery disease risk gene *PCSK9* in visceral abdominal fat to plasma LDL levels[64].

The work described here, has established STARNET and STAGE as important resources for linking tissue specific gene expression to both phenotypic and genetic variation. To better characterise the downstream consequences of genetic variation for plasma cortisol, we used STARNET to link SNPs identified from the CORNET GWAMA to variation in gene expression across the available tissues in both cis and trans.

Initially we worked to characterise the impact of genetic variation at the *SERPINA6*/ *SERPINA1* locus through a cis-eQTL discovery approach. We then examined to what extent cis-eQTLs in this region overlapped with cortisol linked genetic variants from the CORNET GWAMA. Finally we looked to examine the impact of genes that were trans-associated with these genetic variants across the different tissues in STARNET, using gene set enrichment to identify any functional similarities within trans-gene sets.

2.1.3 Chapter objectives

1. To determine if genetic variation for plasma cortisol is linked to variation in gene expression at the *SERPINA6*/ *SERPINA1* locus.

2. To identify genes that are trans-associated with genetic variation for plasma cortisol across STARNET tissues, and to characterise the global transcriptomic response to cortisol linked SNPs.

3. To investigate the functional impact of cortisol associated trans-genes, using clustering and gene set enrichment methods.

2.2 Materials and methods

2.2.1 Datasets

Three major datasets were used over the course of this chapter as described in table 2.1. The summary statistics from the CORNET GWAMA were used in this chapter in their processed form (available at `https://datashare.ed.ac.uk/handle/10283/3836`)[8].

Dataset	Full name	Number of participants	Study type	Number of imputed SNPs
STAGE	Stockholm Tartu Atherosclerosis Gene Expression study	114	Genotype and Microarray data	909,622
STARNET	Stockholm Tartu Atherosclerosis Reverse Networks Engineering Task study	600	Genotype and RNA sequencing	14,098,064
CORNET	CORtisol NETwork (2021)	25,000	GWAMA	8,452,427

Table 2.1: Summary of datasets used throughout this project.

2.2.2 QC and normalisation of STARNET gene expression

All STARNET genotype and gene expression data obtained for this project had previously undergone both Quality Control (QC) and normalisation and is described in full at (Franzen *et al*, 2016)[64].

Gene expression for STARNET tissue samples was measured using RNA-sequencing. RNA samples with less than 1M uniquely mapped reads were excluded, which removed 12 samples with extremely low read counts. The distribution of read counts can be found at figure S2.9 and the samples used in the final analysis had between 15-30 million reads. Normalisation was performed in each tissue separately and counts were adjusted for GC bias and library size using EDAseq (version 1.8.0). Sex was also confirmed using Y chromosome genes and *XIST*. Linear regression was used to correct for age, sex and library protocols. Cross tissue gene expression was

compared using principal component analysis and hierarchical clustering.

The numbers of samples and genes retained can be seen in Table 2.2. Having obtained gene expression matrices from Franzen *et al*, we conducted PCA analysis to confirm that there were no outliers within the samples (Figure S2.10). Ensembl Biomart (GRCh37) was used to label transcripts (provided as Ensembl IDs) with; gene name, chromosome location, gene start and gene end.

Tissue	**Number of samples**		**Number of genes**	
	STARNET	**STAGE**	**STARNET**	**STAGE**
Liver	523	77	13875	19610
Skeletal muscle	512	78	12544	19610
Internal mammary artery	529	79	15458	19610
Atherosclerotic aortic root	514	N/A	16214	N/A
Subcutaneous fat	549	63	14120	19610
Visceral abdominal fat	509	88	14965	19610
Whole blood	448	103	12843	19610
Atherosclerotic arterial wall	N/A	68	N/A	19610

Table 2.2: Summary of number of samples and genes from both STARNET and STAGE datasets.

2.2.3 eQTL discovery

Genotype Pre-processing

STRANET genotypes were obtained using blood DNA genotyping with the Illumina Infinium assay. Again quality control was carried out by Franzen *et al.* The Human OmniExpressExome-8v1 bead chip was used with GRCh37 and contains 951,117 genomic markers. QC was performed using PLINK, with a confirmation of self reported sex via heterozygosity rates on X chromosome. Genome wide heterozygosity per sample was also computed and Hardy-Weinberg equilibrium was computed per SNP. IMPUTE2 was used to increase number of variants via imputation, using the 1000 Genomes Project phase 1 SNPs.

Following QC, STARNET genotypes required further pre-processing before they

could be utilised for eQTL discovery (Figure 2.2A). For both STAGE and STARNET datasets, genotype and gene expression data were contained within matrices. Genotype data for each SNP would be presented as either 2, 1 or 0 representing whether an individual was homozygous, heterozygous or alternative homozygous for a given SNP. All steps were conducted using Python (version 3.7.4).

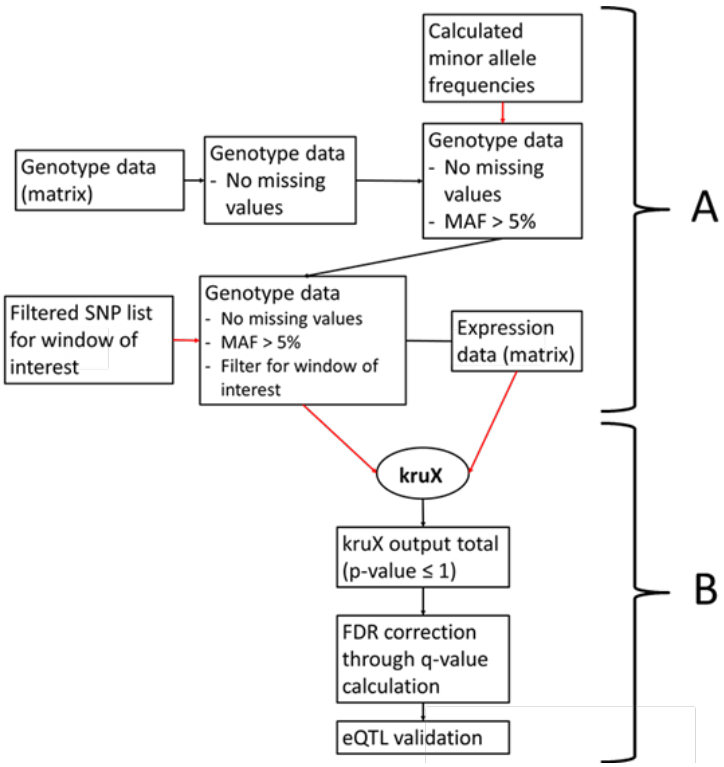


Figure 2.2: Pipeline depicting pre-processing steps (A) and analysis (B) for cis-eQTL discovery using kruX[161] with multiple testing correction.

Genotype data were filtered to remove any SNPs with missing data for any individuals. A further filtering step was undertaken to exclude SNPs that had a Minor Allele Frequency (MAF) < 5% to avoid biases introduced by the inclusion of singletons or rare variants[162]. An additional filtering step was included to only include only SNPs within a given window of interest e.g. for cis associations this is defined as ± 1 Mb of the transcription start/ end point. All of these steps were performed in Python using the Pandas dataframe library (version 1.2.4).

Cis-eQTL Discovery and multiple testing correction

Cis-eQTLs were identified using the kruX algorithm[161]. kruX calculates the Kruskal-Wallis test statistic for millions of SNP gene pairs simultaneously. The Kruskal-Wallis test is a non-parametric ANOVA used to determine whether sample groups originate from the same distribution that has been used previously in genetic association studies when applied to genotype groups[98].

Processed genotype and gene expression data were inputted to kruX as matrices, using the kruX Python package (Figure 2.2B). This calculated the Kruskal-Wallis test statistic for each SNP-gene pair in addition to a corresponding p-value. Multiple testing correction was obtained through the calculation of q-values using the Storey and Tibshirani method for a given FDR threshold (α)[163] using the qvalue Python package.

When effect size (β-values) are required, the package MatrixeQTL has been used which identifies eQTLs using linear regression[164]. This was implemented in R (version 3.6.3) and uses the same input as previous described with kruX. A comparison between kruX and MatrixeQTL output shows a strong correlation between p-values from both methods (R^2 = 0.97) (Figure S2.11).

eQTL discovery results were visualised using the web app of the GWAS visualisation tool LocusZoom[165]. This plots the genomic location of each SNP at its given significance level (-$\log_{10}$ p-value).

Global transcriptomic analysis

The global tissue-specific effect on gene expression for SNPs associated with plasma cortisol at a genome-wide level of significance was depicted using Q-Q plots showing the observed transcriptome-wide SNP-gene associations against the expected uniform distribution. Deviation from the uniform distribution was tested using the Kolmogorov Smirnov test statistic and p-value for each SNP.

Colocalisation

Bayes factor colocalisation analysis was performed in R using the package Coloc[69]. For visualisation of the colocalisation event, Linkage Disequilibrium with the lead SNP was calculated using the package LDlinkR[166].

2.2.4 Identification of trans-associated genes

Trans-gene identification Using secondary linkage test

The secondary linkage test (P2) is a likelihood ratio test in the Findr package[92] (version 1.0.8) that is used to identify associations between a given SNP (E) and a gene (B) without requiring a linear function. P2 proposes a null hypothesis where E and B are independent and and alternative hypothesis where E is causal for B (E $\rightarrow$B). Maximum likelihood estimators are then used to obtain a log likelihood ratio (LLR) between the alternative and null hypothesis (Equation 2.1).

$$P(E \rightarrow B) = P(\mathscr{H}_{alt}^{(P2)} \mid LLR^{(P2)}). \tag{2.1}$$

The LLR is then converted to the posterior probability of $\mathscr{H}_{alt}^{(P2)}$ with empirical estimation of the local FDR as a score from 0-1. The input required for this is composed of three matrices: (E) Genotype information for SNPs of interest. (A) Randomly permutated data in the same shape as E. (B) Expression data for all genes in the given dataset. This can be used to identify genes that are trans-associated with any given SNP as Findr will return a score for the association for all genes from the B matrix with build in FDR correction. To obtain the probability of a false positive within a set of E $\rightarrow$B interactions, this was calculated as 1 minus the mean of all posterior probabilities. A P2 score cut off was set to reflect the desired global FDR[167].

The output of kruX and Findr P2 were compared (Figure S2.12) and plotted using the Python plotting package Seaborn, with Findr P2 score against 1-q-value for

the kruX output. The Spearman's correlation coefficient was then calculated using the Python package scipy.stats (version 1.3.0) showing the different methods to be comparable (R^2 = 0.68).

Hierarchical Clustering

Hierarchical clustering was performed using correlation matrices generated from expression data. The *corr()* function in Pandas was used to construct correlation matrices for all genes in a given gene set. These were generated using Pearson's correlation coefficient for pairwise gene correlations and then plotted using the clustermap function in Seaborn to produce clusters represented by heatmap dendrograms.

K-means Clustering

Gene-sets were also analysed using a K-means clustering approach through use of the Python machine learning package SciKit Learn[168] (version 0.21.2). Gene expression data was used to generate clusters based on a given value of K. The data were then subjected to Principal Component Analysis (PCA) across two components, again using SciKit Learn. The output of the PCA was then presented using the scatterplot function from Seaborn.

Functional annotation and clustering with DAVID

Gene sets were functionally annotated using the Database for Annotation, Visualization and Integrated Discovery (DAVID)[169]. This web-based application allows for the generation gene clusters that have been grouped in relation to an enrichment of functional terms, including but not limited to Gene Ontology (GO) terms. The strength of the gene-term interactions is measured by EASE scores, a modified Fisher's exact test. An enrichment score for a given cluster is generated as the geometric mean of all the EASE scores within a cluster that has undergone -log transfor-

mation. For all analyses Ensembl Gene IDs were used as the input format for DAVID as opposed to universal gene symbols.

For the analyses conducted, all of the default annotation options were selected in addition to: GAD DISEASE, GO TERM BP FAT, GO TERM CC FAT, GO TERM MF FAT, PUBMED ID, REACTOME PATHWAY, BIOGRID INTERACTIONS and UP TISSUE. Gene sets were then run using DAVID and functionally enriched clusters generated using high classification stringency. Tissue specific genes from STARNET RNA-seq datasets were used as background for enrichment (Table 2.2).

2.3 Results

2.3.1 Identification of *SERPINA6* cis-eQTLs in STARNET-liver associated with plasma cortisol

Cis-eQTL Discovery at the *SERPINA6/ SERPINA1* locus

The CORNET GWAMA identified a peak of 73 SNPs that were associated with plasma cortisol ($p < 5 \times 10^{-8}$) at the *SERPINA6/ SERPINA1* locus on chromosome 14[8] (Figure 2.1). Due to the location of the CORNET peak and the role of *SERPINA6* as the gene which encodes CBG, we aimed to identify SNPs associated with plasma cortisol that also influence *SERPINA6* expression.

As this analysis was focused upon SNPs identified in the CORNET GWAMA peak as opposed to all potential *SERPINA6* cis-eQTLs, we restricted the analysis window to only include SNPs within a 100 Kb of *SERPINA6*. The Kruskal Wallis test was performed using genotype and gene expression data from STARNET and STAGE for all SNPs within this window, to infer associations between SNP and gene pairs across all tissues.

The Kruskal Wallis test statistic was calculated for each SNP gene combination of all 580 SNPs in STARNET and 72 SNPs in STAGE. A cis-eQTL peak was identified for *SERPINA6* in liver, the predominant tissue for *SERPINA6* expression[150]. All SNP-*SERPINA6* associations for both STARNET (Figure 2.3A) and STAGE (Figure 2.3B) were plotted in relation to their genomic location and the strength of the association and compared to the original CORNET peak for plasma cortisol variation.

Identification of high confidence *SERPINA6* cis-eQTLs in STARNET-liver

Following correction for multiple testing, 32 cis-eQTLs for *SERPINA6* were identified in STARNET-liver at a 5% FDR threshold ($q \leq 0.05$). These *SERPINA6* cis-eQTLs were then filtered to include only SNPs which were also at genome wide signifi-

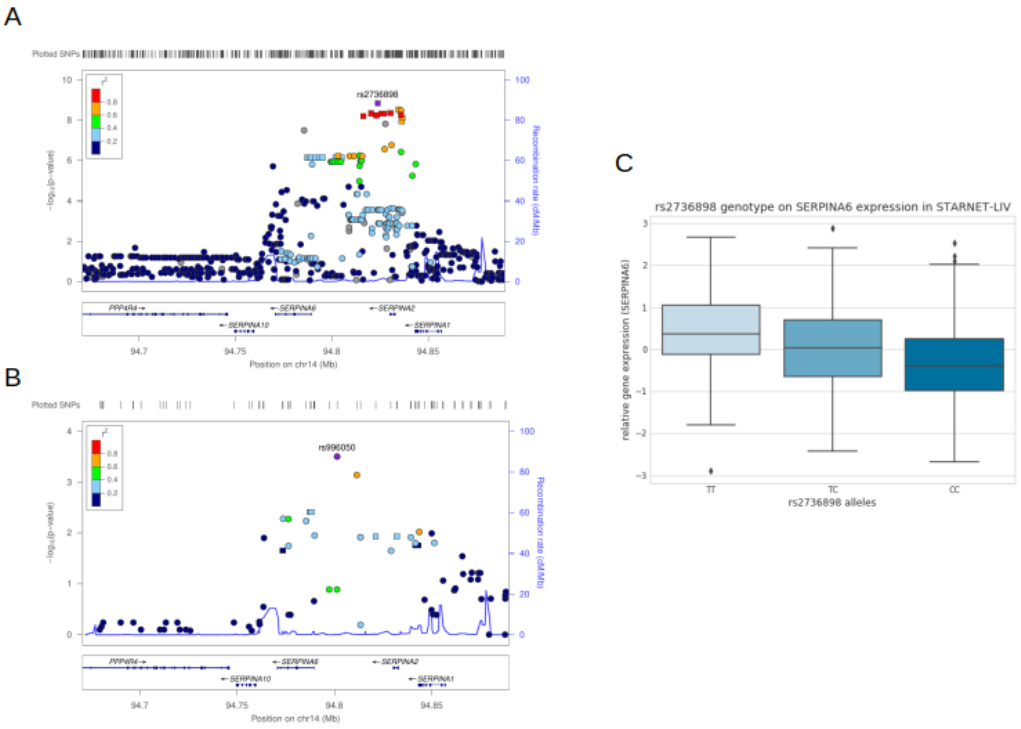


Figure 2.3: Cis-eQTL discovery for *SERPINA6* cis-eQTLs in STARNET-liver. (A) LocusZoom plot showing genomic loci of given SNPs against Kruskal Wallis p-value (-log10 p-value) for an eQTL analysis in liver for all SNPs within 100 Kb of *SERPINA6* in STARNET-liver. Squares represent SNPs ($q \leq 0.05$) that are also at genome wide significance in CORNET ($p \leq 5 \times 10^{-8}$). (B) LocusZoom plot showing cis-eQTLs in STAGE-liver. (C) Genotypic effect of lead *SERPINA6* cis-eQTL in STARNET-liver, rs2736898, on *SERPINA6* gene expression in liver.

cance in CORNET ($p \leq 5 \times 10^{-8}$). This resulted in the identification of 21 cis-eQTLs for *SERPINA6* which were also associated with variation for plasma cortisol (Table 2.3). No cis-eQTLs for *SERPINA6* were identified in any other STARNET tissue. The lead *SERPINA6* cis-eQTL identified was the SNP rs2736898, with the alternate allele C which exerted a negative effect on *SERPINA6* expression in liver ($q = 0.00015$) (Figure 2.3C). This SNP was strongly associated with plasma cortisol ($p = 7.03 \times 10^{-14}$). No cis-eQTLs for *SERPINA6* in liver were identified in GTEx, the most comprehensive public eQTL dataset. However the lead cis-eQTL, rs2736898, was present in GTEx as a cis-eQTL for *SERPINA6* in transverse colon ($p = 0.0000033$)[66].

It is also notable that of the 21 SNPs identified as being associated with *SERPINA6* expression and plasma cortisol, 16 of these were found within the SERPIN LCR, located within the region ~8Kb upstream of *SERPINA1*[144]. These LCR located

SNP	Kruskal Wallis	p-value	q-value	CORNET p-value	Findr P2 score	LCR
rs2736898	40.682	1.465E-09	0.000153	7.026E-14	0.99	TRUE
rs3762132	39.246	3.00E-09	0.000304	1.566E-13	0.99	TRUE
rs59036614	38.457	4.46E-09	0.000441	9.494E-14	0.99	TRUE
rs2749529	38.414	4.56E-09	0.000445	9.917E-14	0.99	TRUE
rs2013150	38.345	4.72E-09	0.000445	7.124E-14	0.99	TRUE
rs2749527	38.340	4.73E-09	0.000445	1.747E-13	0.99	TRUE
rs941594	38.049	5.47E-09	0.000508	1.454E-13	0.99	TRUE
rs2736899	37.939	5.78E-09	0.000522	9.513E-14	0.99	TRUE
rs2749530	37.749	6.35E-09	0.000551	1.399E-13	0.99	TRUE
rs1243171	37.441	7.41E-09	0.0006350	2.022E-13	0.99	TRUE
rs1243173	36.559	1.15E-08	0.000947	1.528E-13	0.99	TRUE
rs2749539	28.667	5.96E-07	0.0342	3.043E-08	0.95	TRUE
rs4491436	28.297	7.17E-07	0.0381	5.97E-19	0.98	TRUE
rs718187	28.297	7.17E-07	0.0381	4.52E-19	0.98	TRUE
rs9989237	28.297	7.17E-07	0.0381	2.157E-19	0.98	TRUE
rs12589136	28.297	7.17E-07	0.0381	3.226E-19	0.98	FALSE
rs6575415	28.297	7.17E-07	0.0381	2.97E-19	0.98	FALSE
rs2281518	28.297	7.17E-07	0.0381	4.579E-19	0.98	FALSE
rs941599	28.297	7.17E-07	0.0381	4.406E-19	0.98	FALSE
rs4905187	28.297	7.17E-07	0.0381	7.338E-19	0.98	TRUE
rs7161521	28.297	7.17E-07	0.0381	3.073E-19	0.98	FALSE

Table 2.3: Cis-eQTLs (q-value ≤ 0.05) identified for *SERPINA6* that were at genome wide significance in CORNET (p-value ≤ 5×10^{-8}) and corresponding P2 score.

SNPs are highlighted in table 2.3. Additionally the *SERPINA6* cis-eQTL rs9989237 (q = 0.0381) is associated with multiple SERPIN genes and other cis-genes across multiple tissues in the FIVEx database, a browser that combines data from multiple eQTL datasets[170]. This may highlight the presence of a LCR within this region providing a biological basis underpinning the observed associations with this SNP.

Two cis-eQTLs for *SERPINA6*, rs941598 and rs996050, were identified in STAGE-liver at the same 5% FDR threshold. However, neither demonstrated an association with plasma cortisol at the level of genome wide significance (p=1.23×10^{-07} and 1.05×10^{-07} respectively). Although, given that individuals were less densely genotyped in STAGE compared to STARNET, many SNPs were unavailable for analysis, including rs2736898.

Cis associations with other genes were also examined in the surrounding region, to confirm that the variation for plasma cortisol was mediated through *SERPINA6*

and not another gene. Cis-eQTLs were identified for *SERPINA10* in the liver and *SERPINA1* in the blood (Figure S2.16), however these cis-eQTLs were not present in the CORNET GWAMA at genome wide significance, providing no evidence to link these genes to a role in mediating genetic variation for plasma cortisol.

Identification of cis-eQTLs for *SERPINA6* using Findr linkage test

The secondary linkage test (P2) from the Findr package was used in order to test for consistency between methods for cis-eQTL discovery. This likelihood ratio test calculates the Bayesian posterior probability for a given SNP-gene interaction. Using P2 it was possible to validate all of the *SERPINA6* cis-eQTL interactions identified using the Kruskal Wallis test within a 5% local precision FDR threshold (P2 $\geq$ 0.95) (Table 2.3).

Colocalisation

Colocalisation is used to determine if separate signals from two association studies, contain a shared causal variant. This is important while taking into account factors such as LD, to identify any variants that colocalise through the integration of eQTL and GWAS signals[69].

To determine if the signals identified for *SERPINA6* cis-eQTLs in liver and SNPs associated with plasma cortisol are driven by the same causal variant, Bayes factor colocalisation analysis was performed while accounting for allelic heterogeneity (Figure 2.4). The probability of both traits sharing a causal variant was low (40.6%) when examining all SNPs within 100Kb of *SERPINA6*. However, when examining each LD block individually, the block represented by rs2736898 returns a 99.2% probability of shared causal variant in this region (Table S2.6).

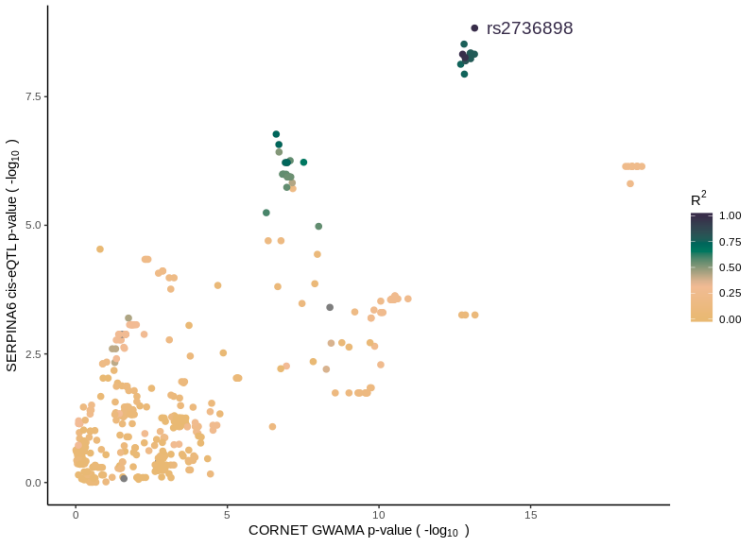


Figure 2.4: Scatterplot showing colocalisation of joint signal from CORNET GWAMA and *SERPINA6* cis-eQTLs from STARNET-liver. Includes all SNPs within 100 Kb of *SERPINA6* that were present in both datasets (n=535). Colour bar indicates degree of LD with rs2736898. Formal colocalisation analysis with Coloc indicates 99.2% probability of the presence of a shared causal variant within LD block 2 mediating GWAMA and *SERPINA6* cis-eQTL signal.

2.3.2 Cortisol associated SNPs mediate global tissue specific effects

We assessed the global transcriptional impact of genetic variation for plasma cortisol across all STARNET tissues. In particular, we aimed to address the role of each LD block that composed the CORNET peak. The CORNET GWAMA peak spanning the *SERPINA6*/ *SERPINA1* locus was composed of 4 LD blocks, each represented by a lead SNP. Of the 21 cis-eQTLs for *SERPINA6*, that were also associated with plasma cortisol for the CORNET GWAMA ($p < 5\times10^{-8}$), 12 were present in LD block 2, and 9 were present in LD block 4, including the representative SNPs for these blocks. The lead *SERPINA6* cis-eQTL in STARNET liver, rs2736898, was identified as the representative SNP for LD block 2 ($p = 7.026\times10^{-14}$).

The global effect on tissue-specific gene expression of representative SNPs from each LD block was assessed using the distribution of transcriptome-wide SNP-gene associations. For these SNP-gene associations we used Findr P2, however with the p-values rather than posterior probabilities which could be plotted logarithmically as Quantile-Quantile plots (Figure S2.13). For all SNP-gene associations per tissue, deviation from the expected uniform distribution of p-values was calculated using

SNP	eQTL p-value	eQTL q-value	CORNET_p-value	Representative SNP	LD block
rs2736898	1.465E-09	0.00015	7.026E-14	rs2736898	2
rs3762132	3.01E-09	0.00030	1.566E-13	rs2736898	2
rs59036614	4.459E-09	0.00044	9.494E-14	rs2736898	2
rs2749529	4.555E-09	0.00044	9.917E-14	rs2736898	2
rs2749527	4.726E-09	0.00044	1.747E-13	rs2736898	2
rs2013150	4.716E-09	0.00044	7.124E-14	rs2736898	2
rs941594	5.466E-09	0.00051	1.454E-13	rs2736898	2
rs2736899	5.778E-09	0.00052	9.513E-14	rs2736898	2
rs2749530	6.352E-09	0.00055	1.399E-13	rs2736898	2
rs1243171	7.408E-09	0.00064	2.022E-13	rs2736898	2
rs1243173	1.152E-08	0.00095	1.528E-13	rs2736898	2
rs2749539	5.958E-07	0.03418	3.043E-08	rs2736898	2
rs4491436	7.169E-07	0.03810	5.97E-19	rs9989237	4
rs718187	7.169E-07	0.03810	4.52E-19	rs9989237	4
rs9989237	7.169E-07	0.03810	2.157E-19	rs9989237	4
rs12589136	7.169-07	0.03810	3.226E-19	rs9989237	4
rs6575415	7.169E-07	0.03810	2.97E-19	rs9989237	4
rs2281518	7.169E-07	0.03810	4.579E-19	rs9989237	4
rs941599	7.169E-07	0.03810	4.406E-19	rs9989237	4
rs4905187	7.169E-07	0.03810	7.338E-19	rs9989237	4
rs7161521	7.169E-07	0.03810	3.073E-19	rs9989237	4

Table 2.4: Distribution of *SERPINA6* cis-eQTL across CORNET GWAMA LD blocks.

the Kolmogorov-Smirnov test(Figure S2.14).

LD block 2 was of particular interest, due to its role in harbouring the strongest cis-eQTL for *SERPINA6* in STARNET liver. When analysing the distribution of P2 p-values, the strongest tissue-specific effects were observed in visceral abdominal fat, subcutaneous fat and liver (Figure 2.5). In all cases the allele associated with higher plasma cortisol in the GWAMA was the allele associated with higher *SERPINA6* expression in STARNET (Figure S2.15). Inflation of p-values is noticeable, particularly in STARNET-liver. This may have arisen due to unaccounted covariates within the population structure[171].

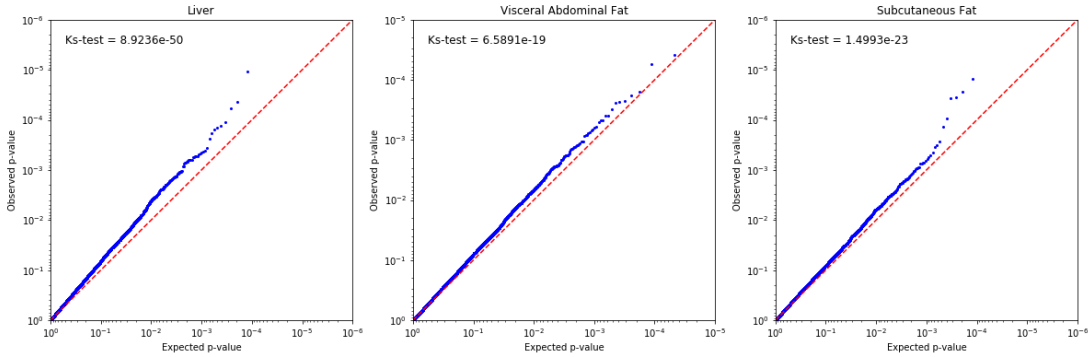


Figure 2.5: Global tissue specific effects on gene expression for rs2736898 represented as Q-Q plots for genes in liver, subcutaneous fat and visceral abdominal fat describing observed p-values vs those expected by chance. Deviation from expected uniform distribution described by Kolmogorov-Smirnov test p-value (Ks-test).

2.3.3 Identification of genes trans-associated to CORNET SNPs

Cortisol trans-genes identified in multiple STARNET tissues

Having identified 21 SNPs that were strongly associated with changes in *SERPINA6* expression in liver and variation for plasma cortisol levels, we next aimed to identify specific genes that were trans-associated with cortisol associated SNPs outside of the *SERPINA6*/ *SERPINA1* locus. Findr P2 was used to identify sets of trans-genes by testing SNP-gene pairs between all cortisol associated SNPs at genome wide significance ($p \leq 5\times10^{-8}$) and every gene in the STARNET dataset across all tissues. We then imposed a FDR threshold to yield tissue specific gene sets.

As trans-gene associations are expected to be weaker than their cis counterparts, a lower threshold of 15% FDR was used to compensate for an increased false negative rate. This approach was used to identify distal genes, including those on different chromosomes, that were associated with the SNPs linked to plasma cortisol. This approach yielded 263 trans-associated genes across seven STARNET tissues (Figure 2.6, Table S2.8), composed of 704 individual trans-associations as many trans-genes were associated with more than one SNP (Figure S2.17).

The tissues with the greatest number of unique trans-genes were subcutaneous fat, visceral abdominal fat and liver, with a combined total of 157 unique trans-genes

and total of 422 total associations (FDR = 15%). FDR was calculated per tissue as the mean of the local precision FDR (P2 scores) (Table S2.7).

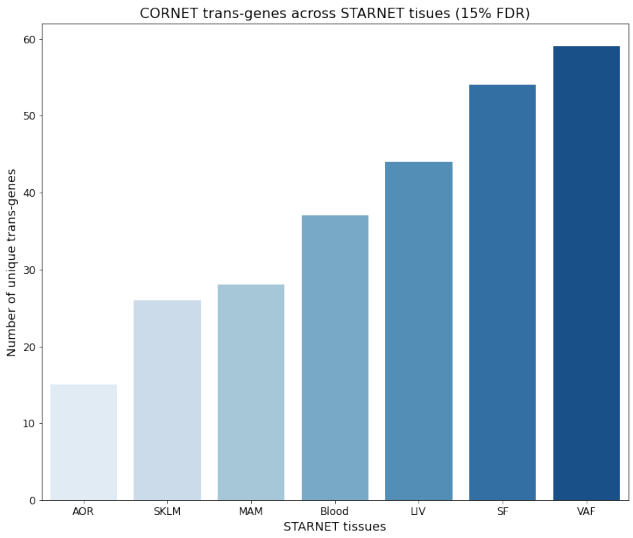


Figure 2.6: Number trans-genes identified across different STARNET tissues. Associations with 73 SNPs that are at genome wide significance in CORNET ($p \leq 5 \times 10\text{-}8$) for a given trans threshold (FDR = 15%)

An examination of the trans-associations for visceral adipose at this threshold revealed that the majority of trans-genes were associated with the same SNP, rs2005945. Although not the strongest SNP identified in the CORNET GWAMA, rs2005945 is at genome wide significance ($p = 9.76 \times 10^{-9}$) and has a small but positive effect on plasma cortisol (effect size = 0.0602).

Functional enrichment of tissue specific trans-gene sets

Functional enrichment methods were used to determine biological functions of cortisol associated tissue specific trans-gene sets. The Database for Annotation, Visualization and Integrated Discovery (DAVID)[169] allows for the annotation of genes and generation of gene-sets based upon common functional enrichment terms.

Gene set enrichment was carried out for trans-genes in liver, visceral abdominal fat and subcutaneous fat, as these tissues contained the greatest number of trans-

associations with cortisol associated SNPs at a 15% FDR threshold. Each tissue gene set was tested in turn, against the background of the total number of genes per tissue that were originally tested for trans-associations. Enriched clusters were filtered with an enrichment score ≥ 1 for each gene set and summarised by a common term (Table 2.5). Full cluster breakdown can be found at Table S2.9.

Tissue	Enrichment score	Cluster
liver	1.24	Regulation of cell signalling
	1.19	GTPase regulation
	1.00	Nucleotide binding
subcutaneous fat	1.99	Desmosome
	1.29	Nucleotide metabolic process
visceral abdominal fat	1.51	Cytoplasmic side of plasma membrane
	1.34	Endoplasmic reticulum
	1.11	Organelle organisation
	1.05	Phospholipid metabolic process

Table 2.5: Function enrichment of CORNET trans-genes using DAVID for top tissues. Clusters have been summarised by a single term.

In liver, the strongest cluster identified was related to regulation of cell signalling, including GO terms for regulation of cell communication (p = 0.05) and regulation of signal transduction, however following multiple testing correction these associations are no longer retained. The strongest cluster across all tissues was a desmosome cluster in subcutaneous fat (p = 0.001). Desmosomes are composed of intermediate filaments and link to the cytoskeleton, they are involved in cell to cell adhesion. Several clusters were identified in visceral abdominal fat, including clusters involving the cell membrane (p = 0.01) and organelle organisation (p = 0.01), however again following multiple testing correction none of these associations are retained.

Hierarchical clustering of CORNET associated trans-genes

To further investigate the role of these trans-genes, a clustering approach was utilised for each trans-gene set using gene expression data from STARNET. For each gene set, a correlation matrix was generated for each gene pair across all samples. Hi-

erarchical clustering was then performed and presented as heatmap dendrograms (Figure 2.7).

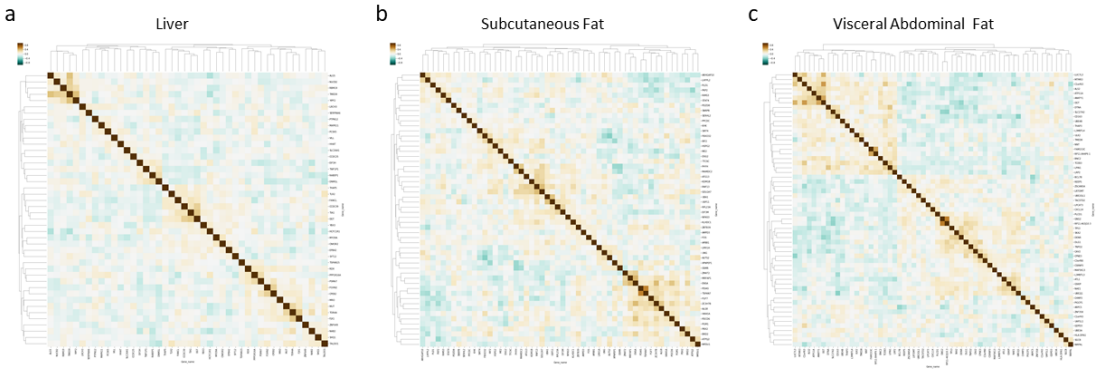


Figure 2.7: Hierarchical clustering of cortisol associated trans-genes in (a) liver (b) subcutaneous fat and (c) visceral abdominal fat.

From visualising the heatmap for visceral abdominal fat trans-genes, the tissue with the largest number of trans-genes, it was possible to observe two large clusters of trans-genes identified based on a positive expression signal. Clustering for both liver and subcutaneous fat trans-genes was poorly defined, however using this approach it was not possible to quantify differences between clusters.

K-means clustering of CORNET associated trans-genes

Having identified visible clusters using hierarchical clustering methods, we decided to use K-means to extract groups in a more quantitative way. K-means clustering is a machine learning method used to partition a set of observations into a given number of clusters, denoted by K. As with hierarchical clustering, expression data for each trans gene set were used and each gene was assigned to the K cluster with the closest mean to its own gene expression. As before, K-means clustering was performed for liver, subcutaneous and visceral adipose fat as these had the greatest number or trans-genes and strongest relationship with glucocorticoid biology.

Following clustering of genes, principal component analyses (PCA) were performed to transform the high dimensional expression data for all patients across two principal components. For each gene set, K was initially set to the lowest pos-

sible value (K=2) and gradually increased to increase the specificity of the classification until discrete clusters observed from the PCA were no longer being retained. Clustering was not observed for either subcutaneous fat or liver trans-genes.

Visceral abdominal fat again stood out as the strongest candidate in the strength of the K-means clusters, where two clusters of genes could be observed. Gene set enrichment was performed in each individual cluster to identify any functional difference between clusters (enrichment score ≥ 1). Out of the two clusters, no functional enrichment was observed for K0 however enrichment was observed in K1 at higher levels than when looking at all trans-genes combined. The functional cluster with the highest enrichment score was related to the mitotic cell cycle (p = 0.003), other clusters include terms related to regulation of organelle organisation (p = 0.007) and cytoplasmic side of cell membrane (p = 0.02) (Figure S2.10), however these enriched terms are no longer retained when correcting for multiple testing.

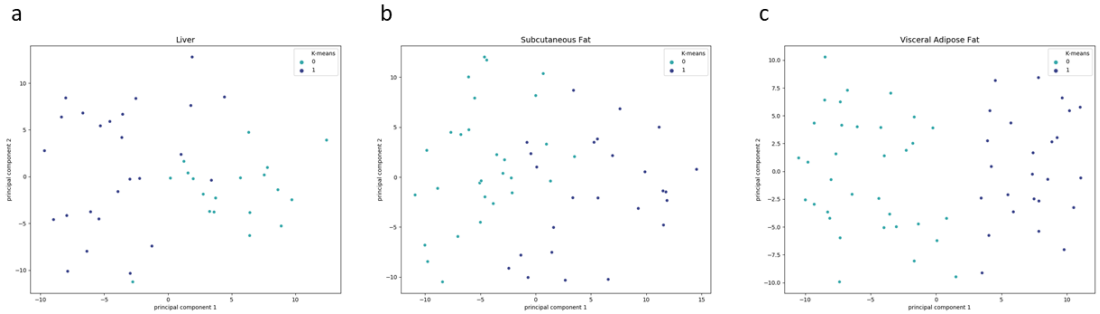


Figure 2.8: K-means clustering of cortisol associated trans-genes in (a) liver (b) subcutaneous fat and (c) visceral abdominal fat

Following the identification and appraisal of differential clustered genes within each tissue set, visceral adipose not only contained the created number of trans-associated genes but also showed the greatest level of structure within the gene set.

2.4 Discussion

2.4.1 Genetic variation for plasma cortisol is mediated through CBG

The identification of 21 high confidence cis-eQTLs for *SERPINA6* that were also at genome wide significance in CORNET suggests that CBG biology may play a role in mediating genetic variation for plasma cortisol. As *SERPINA6* encodes CBG this could shed light on a functional mechanism for how genetic variation may influence plasma cortisol levels.

As CBG binds cortisol in the blood for transport to other tissues, this could account for any cross-tissue effects of plasma cortisol variation. Although *SERPINA6* is only expressed in liver, its influence is global as CBG is exported to the bloodstream and therefore exerts an effect upon all tissues. As GR is ubiquitously expressed and interacts with a vast range of targets, this can account for any pleiotropic effects that may be associated with plasma cortisol variation.

It is important to account for other potential interactions with the significant CORNET SNPs that could act as vehicles for the plasma cortisol variation to the expression level. *SERPINA6* was initially considered due to the proximity of its genomic loci to the CORNET peak on Chromosome 14 and as it already had a defined role within glucocorticoid biology. In addition, cis-eQTL interactions are typically stronger than trans interactions. The nominal threshold of 1 Mb that is set for cis-interaction is done so as this is what is assumed to be the upper distance limit for chromatin interactions where an eQTL could have a physical interaction with its target gene[172]. It is also notable that 16 of the 21 cis-eQTLs for *SERPINA6* were located in the SERPIN LCR, a previously described cis-regulatory region for control of SERPIN genes. Genetic variation in this region may lead to the disruption of previously described liver specific transcription factors binding sites[143], leading to the observed changes in *SERPINA6* expression in STARNET liver.

Although cis-eQTL interactions were identified for other genes within the 100

Kb window containing the CORNET peak, these were able to be excluded, as these could not account for plasma cortisol variation. High confidence cis-eQTLs ($q \leq 0.05$) for both *SERPINA10* in Liver and *SERPINA1* in the blood were identified within this region, however of those cis-eQTLs that were also in CORNET, none crossed the threshold for genome wide significance. As there were no cis-eQTLs for *SERPINA1* that were associated with genetic variation for plasma cortisol, this suggests that it is indeed CBG as opposed to α-1 antitrypsin that is responsible for mediating the genetic effect on cortisol.

The identification of cis-eQTLs for *SERPINA6* in liver has not been previously observed in other large eQTL datasets such as GTEx. Although, GTEx does contain cis-eQTLs for *SERPINA6*, these are in tissues other than liver, the tissue where *SERPINA6* is most highly expressed. The eQTLs identified from GTEx are from from post-mortem donors, which may also have influenced gene expression levels, missing key associations. This highlights the importance of selecting a representative dataset when conducting such analyses and the selection of appropriate, especially considering that liver was the only STARNET tissue where *SERPINA6* cis-eQTLs were identified. There are a number of healthy controls in the STARNET dataset, however this number was too small to conduct eQTL discovery and compare the differences between case and control populations.

Outside of the previously stated advantages, the selection of STARNET as a cohort for measuring the genetic influence of plasma cortisol may have introduced some biases. Although genotypes are fixed, gene expression varies in response to both endogenous and exogenous factors. Whole blood samples were taken preoperatively, all other tissues including liver were taken during a coronary artery bypass grafting procedure. The human stress response to surgery has been well characterised and results in stimulation of the HPA axis leading to high levels of cortisol in the blood both during and post-surgery[173]. In addition to surgical stress there is also anticipatory stress which may result in elevated cortisol levels. As the individuals in the STARNET cohort will have higher cortisol levels as a result of coronary artery bypass grafting, it is unclear if gene expression levels would be representative

of a healthy population, given evidence of reduced cortisol clearance in patients suffering from critical illness[174].

An additional issue is that although GR mediates changes through transcription, it is unclear what affect these variants are having upon CBG protein levels. This is due to a lack of proteomic data in the STARNET cohort. Although *SERPINA6* expression can be thought of as a proxy for CBG levels, not all *SERPINA6* mRNA will be translated, so it is unclear how much of an effect these variants will have on the protein level.

It is important to highlight some of the shortcomings of the CORNET GWAMA, which underpins this thesis, as any issues in study design may have impacted any downstream analyses. Cortisol measurements obtained in the CORNET GWAMA are defined as morning cortisol measurements, however these cover a time period between wake and 11am[8]. The impact of inter-individual variation in diurnal rhythms and wake times may lead to comparisons between individuals that are not reflective of variation in basal levels of plasma cortisol between individuals. An improved study design could involve cortisol measurements taken across multiple time points throughout the period between wake and 11am to account for inter-individual variation.

2.4.2 Trans-effects of cortisol associated SNPs predominantly observed in liver and fat

The identification of genes that are trans associated with cortisol associated SNPs is a crucial step in understanding the downstream effects of genetic variation for plasma cortisol. If CBG is the vehicle through which plasma cortisol variation exerts its influence, it is through the changes in expression of these downstream genes that the tissue specific consequences for this variation are mediated.

The tissues with the greatest number of trans-genes identified were liver and both subcutaneous and visceral abdominal fat. It is interesting to note that these

tissues all play a crucial role in glucocorticoid biology. The stimulation of gluconeogenesis occurs mainly within the liver and this is stimulated by glucocorticoids, including cortisol[175]. Cortisol plays a role in fat through the stimulation of the expression of genes involved in adipocyte metabolism[176]. In extreme cases this can lead to clinical consequences. Cushing's Syndrome results in high blood pressure and obesity and occurs in response to chronically increased levels of plasma cortisol[177].

Functional annotation clustering across these tissues has identified trends in the genes that are associated with variation for plasma cortisol. In liver, this resulted in an enrichment of genes associated with the regulation of cell signalling and GTPase regulation. As the effects of GR are ubiquitous these effects are often mediated throughout and between cells. Cell signalling is important in this process.

Visceral abdominal fat contained the highest number of trans-genes out of all of the tissues studied. When performing functional clustering on all visceral adipose trans-genes together, functional clusters related to the cell membrane, organelle organisation and the endoplasmic reticulum were identified. K-means clustering revealed a difference in functional clustering between k-clusters, where one cluster was enriched with terms related to organelle organisation, cell membrane and new clusters such as cell cycle and vesicle trafficking.

A 15% FDR threshold was used to call trans-gene sets, as trans-associations tend to be weaker than their cis counterparts. Therefore setting a more lenient threshold protects against a high false negative rate, however given the high number of trans-association, with a 15% FDR threshold, a number of false positive findings will inevitably be introduced. This is counteracted in subsequent chapters, by scrutinising key genes within these trans-gene sets that are suspected drivers of genetic variation for cortisol.

As the primary transcriptional signal at the CORNET GWAMA locus appears to be driven by changes in CBG expression, the possibility exists that the trans-effects observed in extra-hepatic tissues are being predominantly driven by changes in

CBG levels as opposed to variation in cortisol acting through CBG. HPA adaptation is crucical for regulating the genetic impact of plasma cortisol, however a recent rodent model found only the transporter *ABCB1* to be responsible for mediating HPA adaptation[178].

As it is a relatively small proportion of the heritability of plasma cortisol that is explained by the *SERPINA6*/ *SERPINA1* locus, it is worth considering that the transgene effect observed is in fact mediated by CBG acting independently of plasma cortisol. This could be indicative of a new role of CBG that is not related to the transport of plasma cortisol, however future functional experiments would be required to either substantiate or refute this claim.

2.5 Conclusion

By using a systems genetics approach to examine the impact of genetic variation for plasma cortisol upon downstream gene expression, it has been possible to investigate how the effects of cortisol linked genetic variation are mediated. The identification of SNPs that are associated with both variation for plasma cortisol and CBG expression highlights a mechanism by which genetic variation at the genomic level can have tissue specific influences through the machinery of glucocorticoid biology that is ubiquitous across tissues. By expanding these associations across the genome, it has been possible to identify trans effects of cortisol variation, with a specific emphasis in hepatic and adipose tissues. This chapter provides a foundation for understanding the role of cortisol associated genetic variation in mediating tissue specific changes in gene expression.

2.6 Supplementary data

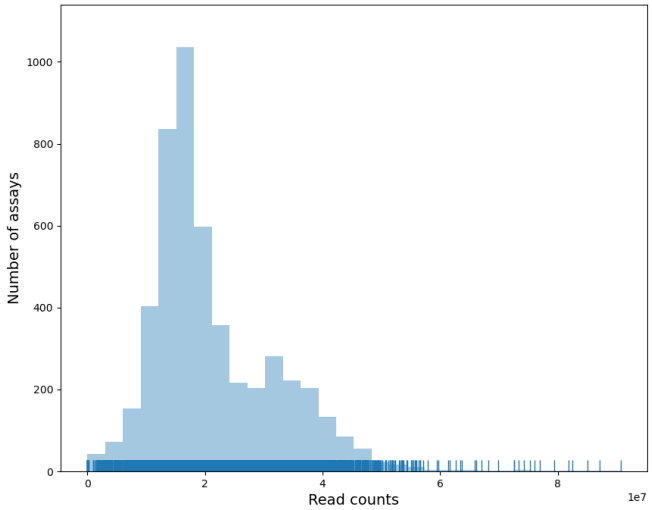


Figure 2.9: Distribution of RNA-seq read counts across all STARNET tissues.

LD_block	nsnps	PP.H0.abf	PP.H1.abf	PP.H2.abf	PP.H3.abf	PP.H4.abf
LD1	4	4.16E-05	7.82E-07	0.259387	0.004138	0.736433
LD2	13	4.81E-14	1.55E-10	2.92E-06	0.008404	0.991593
LD3	2	5.14E-05	1.34E-06	0.047832	0.000298	0.951817
LD4	52	2.24E-17	5.87E-16	0.00038	0.008972	0.990648
All_SNPs	535	9.76E-18	3.48E-14	0.000167	0.594	0.406

Table 2.6: Analysis using Coloc, a Bayesian test for colocalisation. Coloc Approximate Bayes Factor Colocalisation Analysis (ABF) return 5 hypothesis tests to determine if two genetic association signals share the same causal variant. H0.abf: neither trait has a genetic association in the region, H1.abf: only cis-eQTL has a genetic association in the region. H2.abf: only GWAMA has a genetic association in the region, H3: both traits are associated, but with different causal variants. H4: both traits are associated and share a single causal variant. This test assumes a single shared causal variant for both signals. Posterior probability (PP) of a shared causal variant is low when examining all SNPs together (40.6%) but increases in certain LD blocks when examined individually. Strongest signal is present in LD blocks 2 (99.2%) and 4 (99.1%).

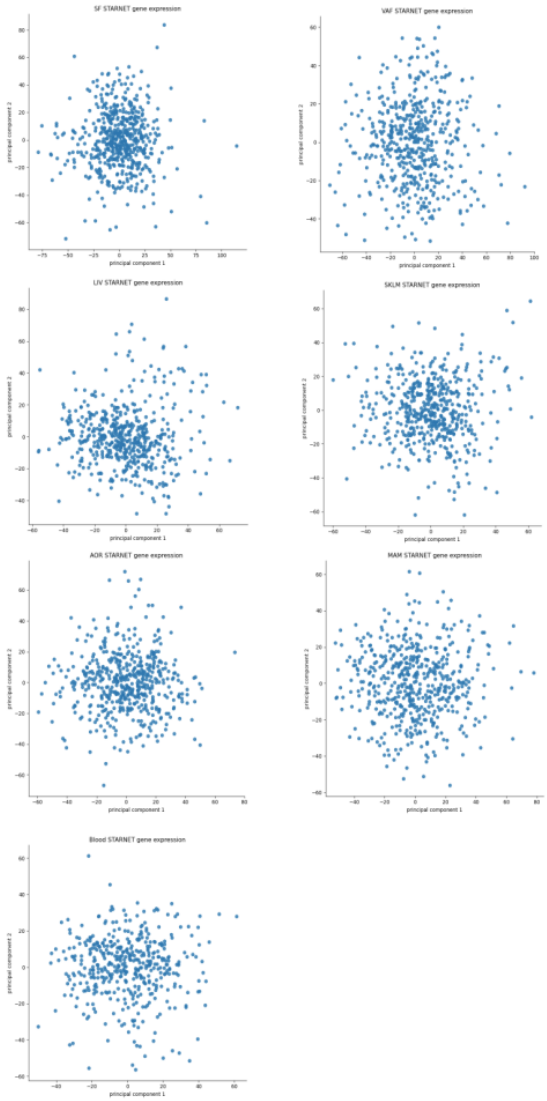


Figure 2.10: Principal component analysis of gene expression samples across all STARNET tissues.

Tissue	10% FDR	15% FDR
LIV	0.81	0.72
SKLM	0.87	0.72
AOR	0.85	0.79
MAM	0.82	0.72
VAF	0.8	0.87
SF	0.83	0.73
BLOOD	0.81	0.71

Table 2.7: Tissue specific local precision FDR (Findr P2 scores) used to establish FDR thresholds for trans-gene sets.

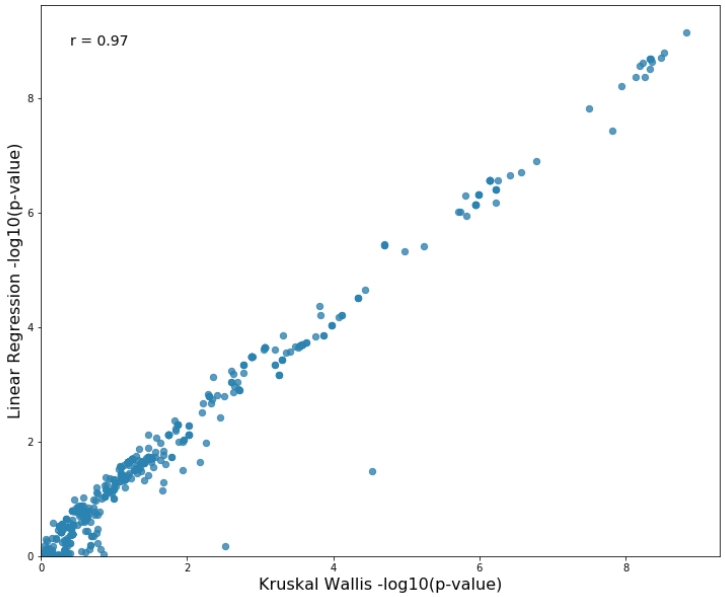


Figure 2.11: Comparison of -log10 p-values for *SERPINA6* cis-eQTLs calculated using Linear regression vs Kruskal Wallis. Spearman's R = 0.97.

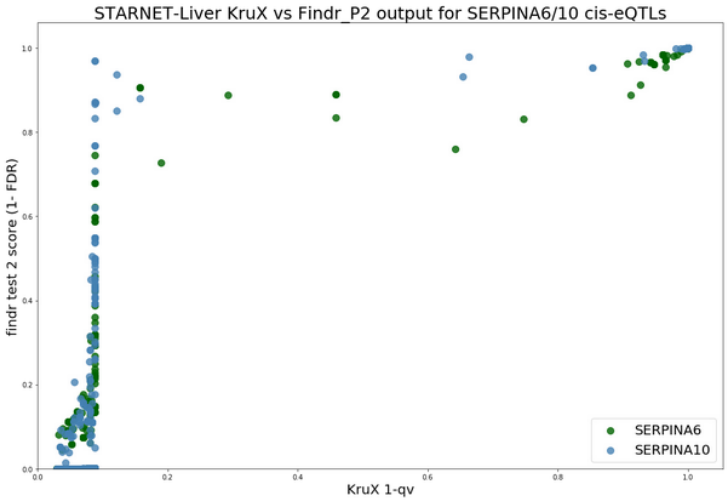


Figure 2.12: Comparison of eQTL disocvery using KruX vs Findr secondary linkage test (P2) method for *SERPINA6* and *SERPINA10* in STARNET-liver. Spearman's R = 0.68

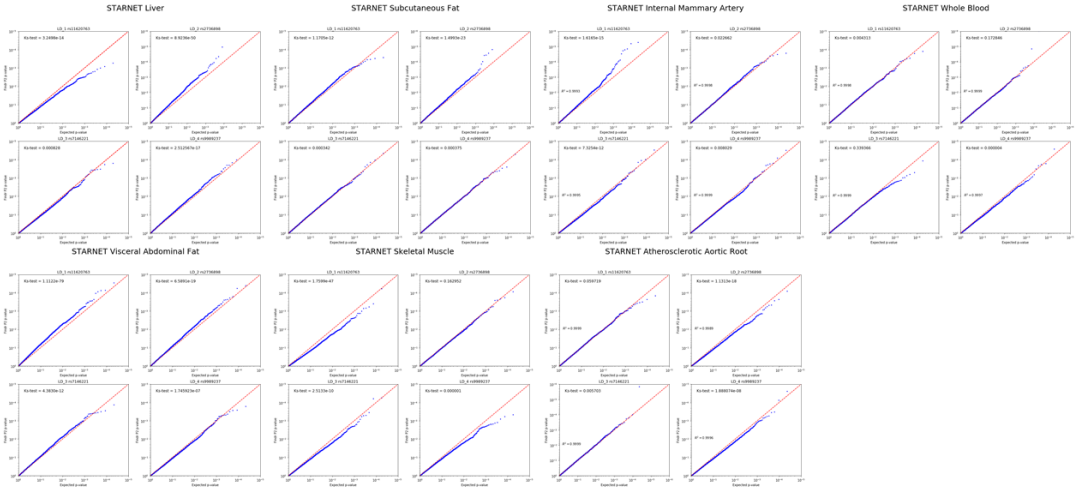


Figure 2.13: Quantile-quantile plot of tissue-specific *SERPINA6* expression according to each LD block

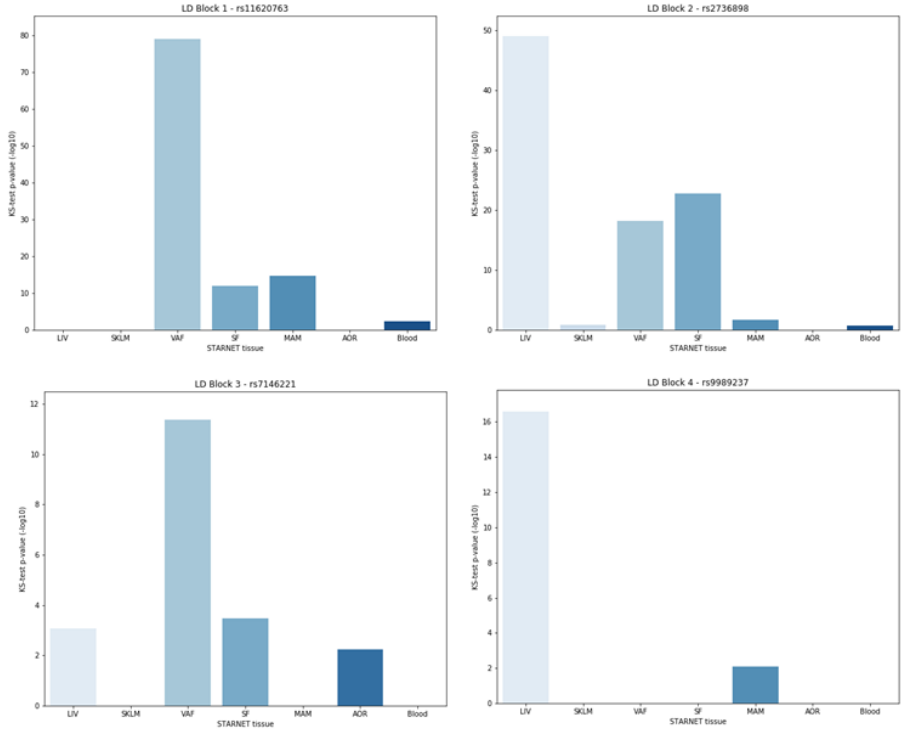


Figure 2.14: Kolmogorov-Smirnov test for *SERPINA6* expression of the lead SNP in each LD block in multiple tissues

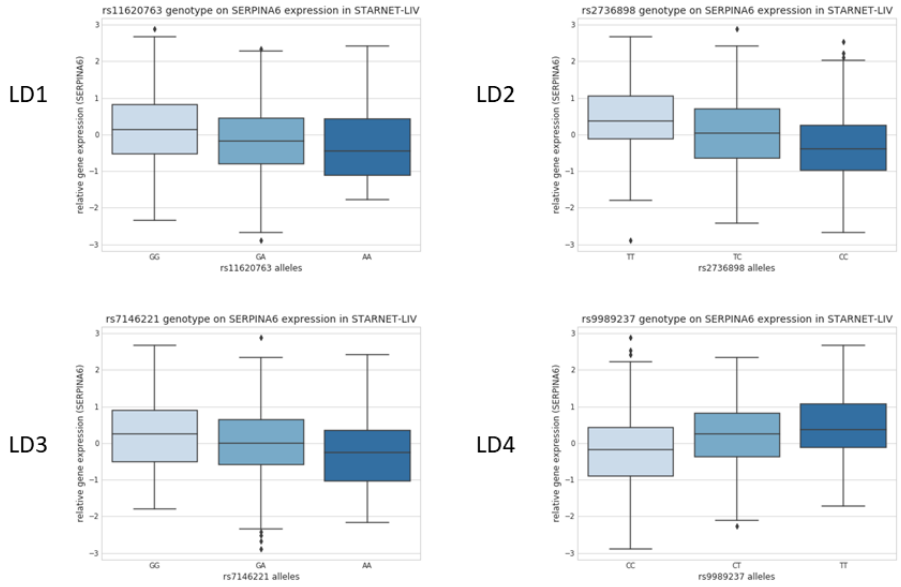


Figure 2.15: Magnitude and allele-specific effects of lead SNP in each LD block on *SERPINA6* expression in STARNET-liver

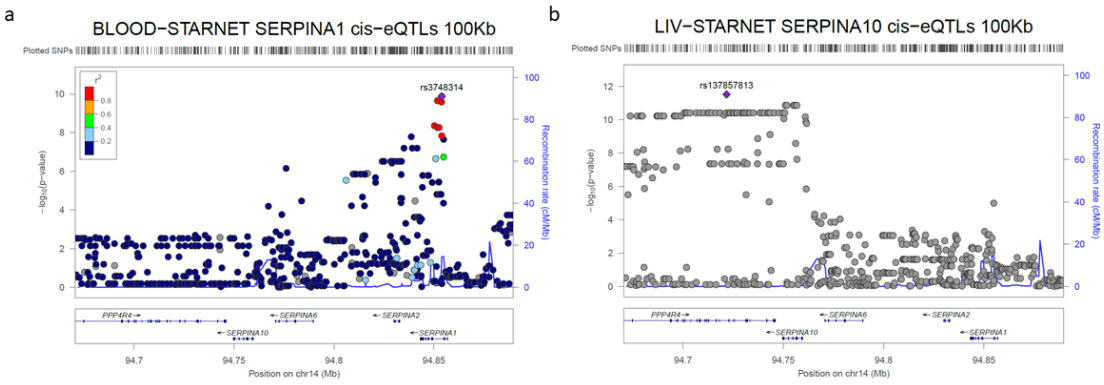


Figure 2.16: LocusZoom plots showing genomic loci of given SNPs against measure of significance (-log10 (p-value)). (a) eQTL analysis for STAGE-Blood samples for *SERPINA1* ($p \leq 1$) (n=580). (B) eQTL analysis for STARNET-Liver samples for *SERPINA10* ($p \leq 1$) (n=580)

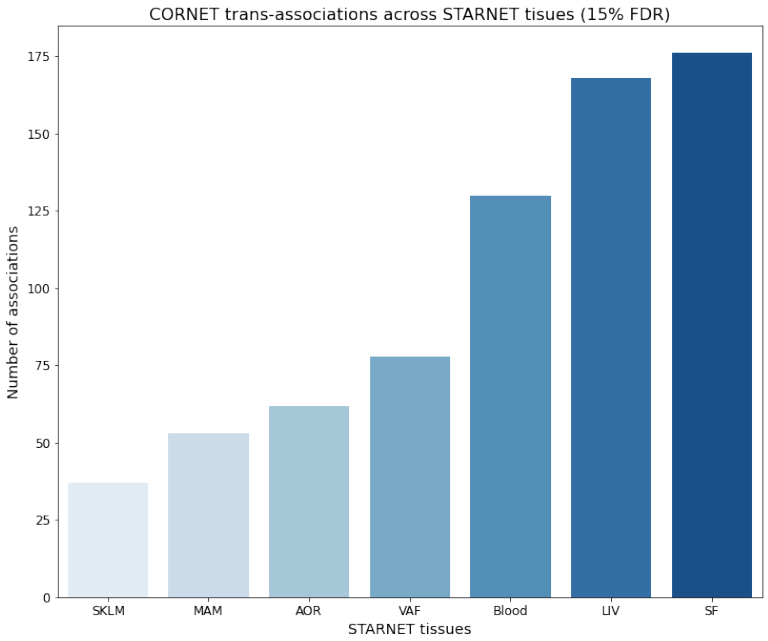


Figure 2.17: Number trans-associations identified across different STARNET tissues. Associations with 73 SNPs that are at genome wide significance in CORNET ($p \leq 5 \times 10\text{-}8$) for a given trans threshold (P2 ≥ 15% FDR)

Gene name	Ensembl Gene ID	Top SNP	P2 score	Tissue	Number of associations
TNC	ENSG00000041982	rs7146221	0.980	AOR	1
AL139147.1	ENSG00000248458	rs1950652	0.952	AOR	11
RP11-844P9.3	ENSG00000251667	rs877081	0.880	AOR	10
ZSCAN2	ENSG00000176371	rs3748319	0.863	AOR	4
LINC01021	ENSG00000250337	rs3790035	0.860	AOR	3
DBF4	ENSG00000006634	rs12590834	0.853	AOR	10
SLC25A34	ENSG00000162461	rs2749529	0.849	AOR	1
ACAA2	ENSG00000167315	rs1950652	0.846	AOR	4
C1QC	ENSG00000159189	rs877081	0.835	AOR	1
CD6	ENSG00000013725	rs12590834	0.833	AOR	1
FAM229B	ENSG00000203778	rs9989237	0.832	AOR	10
GLA	ENSG00000102393	rs12590834	0.817	AOR	1
C8orf59	ENSG00000176731	rs8015996	0.812	AOR	3
ALS2CL	ENSG00000178038	rs4283161	0.801	AOR	1
TAPT1	ENSG00000169762	rs2749529	0.792	AOR	1
RBM17	ENSG00000134453	rs1243171	0.996	Blood	10
CEP164	ENSG00000110274	rs1243171	0.996	Blood	11
DONSON	ENSG00000159147	rs2281517	0.969	Blood	12
CENPBD1	ENSG00000177946	rs3790035	0.957	Blood	3
ZC3H14	ENSG00000100722	rs3819333	0.947	Blood	8
ARHGAP32	ENSG00000134909	rs4283161	0.941	Blood	1
PRPF6	ENSG00000101161	rs4905194	0.937	Blood	8
ZNF555	ENSG00000186300	rs3819333	0.926	Blood	7
GTF2F1	ENSG00000125651	rs3762130	0.911	Blood	7
CHAC2	ENSG00000143942	rs2749539	0.897	Blood	1
PNPO	ENSG00000108439	rs12590834	0.878	Blood	2
TUBG1	ENSG00000131462	rs4990242	0.865	Blood	2
RPL23AP7	ENSG00000240356	rs67994395	0.863	Blood	1
SBDS	ENSG00000126524	rs4283161	0.851	Blood	1
ARHGAP12	ENSG00000165322	rs6575415	0.850	Blood	9
IPO7	ENSG00000205339	rs718187	0.840	Blood	9
ADAMTS6	ENSG00000049192	rs4283161	0.837	Blood	1
ASAH1	ENSG00000104763	rs4990242	0.836	Blood	1
TLR5	ENSG00000187554	rs941594	0.826	Blood	1
KANK2	ENSG00000197256	rs4283161	0.816	Blood	1
TTC1	ENSG00000113312	rs67994395	0.806	Blood	10
SYF2	ENSG00000117614	rs113375097	0.789	Blood	1
CD27-AS1	ENSG00000215039	rs877081	0.786	Blood	1
USP35	ENSG00000118369	rs877081	0.786	Blood	1
BRICD5	ENSG00000182685	rs877081	0.786	Blood	1
CAMK2N1	ENSG00000162545	rs877081	0.786	Blood	1
TRAP1	ENSG00000126602	rs35854995	0.784	Blood	1
NPC2	ENSG00000119655	rs877081	0.784	Blood	1
ARL4A	ENSG00000122644	rs877081	0.780	Blood	1
TTF2	ENSG00000116830	rs877081	0.779	Blood	1
NDRG3	ENSG00000101079	rs877081	0.772	Blood	1
SERPINA1	ENSG00000197249	rs877081	0.770	Blood	3

RSAD1	ENSG00000136444	rs877081	0.766	Blood	1
DGKZP1	ENSG00000179611	rs877081	0.765	Blood	1
GUF1	ENSG00000151806	rs877081	0.759	Blood	1
LINC00152	ENSG00000222041	rs11620777	0.742	Blood	1
TMEM260	ENSG00000070269	rs4905194	0.731	Blood	7
SERPINA6	ENSG00000170099	rs3762132	0.999	LIV	24
YBX3	ENSG00000060138	rs56045385	0.976	LIV	3
HCFC1R1	ENSG00000103145	rs56045385	0.965	LIV	14
HHAT	ENSG00000054392	rs3762132	0.955	LIV	7
SYT12	ENSG00000173227	rs11622665	0.952	LIV	1
VILL	ENSG00000136059	rs67994395	0.952	LIV	1
MRO	ENSG00000134042	rs3762132	0.944	LIV	3
SHQ1	ENSG00000144736	rs1950652	0.908	LIV	8
SLC26A1	ENSG00000145217	rs2736898	0.906	LIV	3
CCDC39	ENSG00000145075	rs56045385	0.905	LIV	14
CPEB2	ENSG00000137449	rs4905194	0.893	LIV	1
TMED9	ENSG00000184840	rs4905194	0.893	LIV	25
THAP5	ENSG00000177683	rs2749539	0.891	LIV	1
NAB2	ENSG00000166886	rs3762127	0.889	LIV	9
FOXN2	ENSG00000170802	rs2013150	0.888	LIV	1
TWF1P1	ENSG00000178082	rs2013150	0.888	LIV	1
FANCL	ENSG00000115392	rs35854995	0.883	LIV	4
CNKSR2	ENSG00000149970	rs8022616	0.873	LIV	1
LRCH3	ENSG00000186001	rs8022616	0.873	LIV	1
RBM19	ENSG00000122965	rs4283161	0.872	LIV	2
FJX1	ENSG00000179431	rs4905194	0.871	LIV	5
TOR4A	ENSG00000198113	rs3762127	0.860	LIV	5
PPP1R16A	ENSG00000160972	rs4990242	0.857	LIV	1
MUT	ENSG00000146085	rs12590834	0.835	LIV	11
TIA1	ENSG00000116001	rs67994395	0.834	LIV	3
YIPF2	ENSG00000130733	rs4283161	0.833	LIV	1
RDX	ENSG00000137710	rs2736898	0.814	LIV	1
TSPAN15	ENSG00000099282	rs8022616	0.805	LIV	1
CCDC25	ENSG00000147419	rs2281517	0.803	LIV	1
ALG5	ENSG00000120697	rs4283161	0.787	LIV	1
EIF3H	ENSG00000147677	rs4990242	0.786	LIV	1
PCSK5	ENSG00000099139	rs2736898	0.777	LIV	1
TLR2	ENSG00000137462	rs2749527	0.763	LIV	1
DNM1L	ENSG00000087470	rs2749527	0.762	LIV	1
NUCB2	ENSG00000070081	rs2749527	0.762	LIV	1
PTPN12	ENSG00000127947	rs2749527	0.762	LIV	1
MAPK11	ENSG00000185386	rs35854995	0.761	LIV	1
ZNF649	ENSG00000198093	rs2749527	0.754	LIV	1
RABEP1	ENSG00000029725	rs2749527	0.748	LIV	1
EPB42	ENSG00000166947	rs1956179	0.728	LIV	1
MYO9A	ENSG00000066933	rs3762132	0.726	LIV	1
PSMA7	ENSG00000101182	rs1956179	0.723	LIV	1
TALDO1	ENSG00000177156	rs2749527	0.721	LIV	1

FAM127B	ENSG00000203950	rs11620763	0.961	MAM	3
MTND5P12	ENSG00000251544	rs11620763	0.954	MAM	3
RP11-846F4.1	ENSG00000266529	rs11620777	0.953	MAM	3
ZSCAN16-AS1	ENSG00000269293	rs113375097	0.944	MAM	1
GTPBP3	ENSG00000130299	rs11620777	0.929	MAM	3
RARRES1	ENSG00000118849	rs11620777	0.925	MAM	1
SSH1	ENSG00000084112	rs11620777	0.925	MAM	3
CACYBP	ENSG00000116161	rs7141205	0.908	MAM	3
C1orf52	ENSG00000162642	rs2749527	0.906	MAM	1
RP11-488L18.10	ENSG00000259865	rs7146221	0.896	MAM	1
MAP6D1	ENSG00000180834	rs2005945	0.891	MAM	1
GRIPAP1	ENSG00000068400	rs2005945	0.891	MAM	1
PIPOX	ENSG00000179761	rs2749527	0.878	MAM	1
CNOT3	ENSG00000088038	rs2749527	0.874	MAM	2
SIX4	ENSG00000100625	rs12892767	0.862	MAM	1
RP11-56B16.2	ENSG00000259556	rs7146221	0.844	MAM	1
KIF3B	ENSG00000101350	rs2005945	0.843	MAM	1
TTL	ENSG00000114999	rs2005945	0.841	MAM	1
LZIC	ENSG00000162441	rs11629326	0.841	MAM	5
RP11-736K20.6	ENSG00000246523	rs7146221	0.835	MAM	1
RP11-323I15.5	ENSG00000259336	rs2005945	0.822	MAM	1
HNRNPCP6	ENSG00000213305	rs4283161	0.810	MAM	1
PI4KB	ENSG00000143393	rs4283161	0.810	MAM	1
SLC37A2	ENSG00000134955	rs2281517	0.798	MAM	1
TRMU	ENSG00000100416	rs11622665	0.779	MAM	1
RP11-5C23.2	ENSG00000273183	rs877081	0.767	MAM	1
IFI27L2	ENSG00000119632	rs2749539	0.752	MAM	1
NCKAP5L	ENSG00000167566	rs7161521	0.736	MAM	9
ALG8	ENSG00000159063	rs2749530	0.999	SF	11
NPAS3	ENSG00000151322	rs8022616	0.966	SF	1
RNF13	ENSG00000082996	rs11622665	0.963	SF	1
OSMR	ENSG00000145623	rs8022616	0.954	SF	1
ATP5J2	ENSG00000241468	rs2013150	0.950	SF	7
SIRT4	ENSG00000089163	rs7145181	0.948	SF	18
PPCDC	ENSG00000138621	rs4900229	0.948	SF	18
ZBTB39	ENSG00000166860	rs2005945	0.873	SF	1
LMX1A	ENSG00000162761	rs11629171	0.873	SF	3
VWA5A	ENSG00000110002	rs11629171	0.873	SF	3
AUTS2	ENSG00000158321	rs8015996	0.873	SF	3
TTC9C	ENSG00000162222	rs2013150	0.872	SF	5
DSG2	ENSG00000046604	rs3790035	0.871	SF	3
UNG	ENSG00000076248	rs3748319	0.864	SF	7
RPL23A	ENSG00000198242	rs4900229	0.859	SF	1
ATG13	ENSG00000175224	rs4990242	0.857	SF	2
PLD1	ENSG00000075651	rs4990242	0.857	SF	1
FGF7	ENSG00000140285	rs4990242	0.857	SF	1
RIMS3	ENSG00000117016	rs1243171	0.856	SF	1
ZC3H7B	ENSG00000100403	rs941594	0.856	SF	3

PBX2	ENSG00000204304	rs1243171	0.856	SF	3
PGM1	ENSG00000079739	rs1243171	0.856	SF	1
PKP2	ENSG00000057294	rs1243171	0.856	SF	1
STAT4	ENSG00000138378	rs941594	0.856	SF	2
EIF3M	ENSG00000149100	rs941594	0.856	SF	4
MPDU1	ENSG00000129255	rs941594	0.852	SF	1
BRD2	ENSG00000204256	rs1243171	0.850	SF	2
AMPD3	ENSG00000133805	rs1950652	0.848	SF	15
USP11	ENSG00000102226	rs1950652	0.837	SF	6
PDZD8	ENSG00000165650	rs7146221	0.834	SF	1
KLHDC1	ENSG00000197776	rs17090691	0.829	SF	9
PDCD6	ENSG00000249915	rs7146221	0.827	SF	1
PHYH	ENSG00000107537	rs8022616	0.822	SF	1
MAMDC2	ENSG00000165072	rs8022616	0.822	SF	1
IRF2	ENSG00000168310	rs8022616	0.822	SF	1
GOLGA7	ENSG00000147533	rs8022616	0.822	SF	1
TSPAN7	ENSG00000156298	rs1243171	0.821	SF	1
FOS	ENSG00000170345	rs58622098	0.819	SF	10
LHFPL2	ENSG00000145685	rs2005945	0.814	SF	1
UBA1	ENSG00000130985	rs67994395	0.805	SF	1
APBB1	ENSG00000166313	rs67994395	0.805	SF	1
ME2	ENSG00000082212	rs8022616	0.785	SF	1
ZMAT2	ENSG00000146007	rs2749529	0.782	SF	5
HSPG2	ENSG00000142798	rs7146221	0.772	SF	1
MRFAP1	ENSG00000179010	rs2749527	0.772	SF	1
PDIA5	ENSG00000065485	rs1243171	0.769	SF	2
KDM1B	ENSG00000165097	rs8022616	0.768	SF	1
XPNPEP1	ENSG00000108039	rs8022616	0.764	SF	1
ARHGAP10	ENSG00000071205	rs2005945	0.742	SF	1
ENSA	ENSG00000143420	rs1243171	0.738	SF	2
SNRPB	ENSG00000125835	rs67994395	0.736	SF	1
KHK	ENSG00000138030	rs8022616	0.733	SF	1
SERHL2	ENSG00000183569	rs67994395	0.732	SF	1
FBXO32	ENSG00000156804	rs8015996	0.731	SF	3
ATHL1	ENSG00000142102	rs8022616	0.971	SKLM	1
ZNF566	ENSG00000186017	rs11629171	0.910	SKLM	3
ESRRG	ENSG00000196482	rs7146221	0.908	SKLM	1
KLHDC3	ENSG00000124702	rs7146221	0.906	SKLM	1
PLEKHH2	ENSG00000152527	rs2749539	0.898	SKLM	1
PPP1R35	ENSG00000160813	rs2749539	0.894	SKLM	5
WDR55	ENSG00000120314	rs2749539	0.892	SKLM	1
PPP1R21	ENSG00000162869	rs11620777	0.891	SKLM	3
ZNF503	ENSG00000165655	rs4283161	0.884	SKLM	1
OGG1	ENSG00000114026	rs2749539	0.884	SKLM	1
PPP1R9B	ENSG00000108819	rs2749539	0.882	SKLM	1
CCND3	ENSG00000112576	rs2749539	0.880	SKLM	1
SULT1A1	ENSG00000196502	rs2749539	0.879	SKLM	1
MTA2	ENSG00000149480	rs2749539	0.879	SKLM	1

TMEM117	ENSG00000139173	rs2749539	0.879	SKLM	1
EFNB1	ENSG00000090776	rs2749539	0.879	SKLM	1
UTP14A	ENSG00000156697	rs11622665	0.867	SKLM	1
RAPSN	ENSG00000165917	rs12892767	0.858	SKLM	1
ALG1L	ENSG00000189366	rs56045385	0.845	SKLM	2
ZNF35	ENSG00000169981	rs2749529	0.826	SKLM	2
ARHGEF28	ENSG00000214944	rs1956179	0.825	SKLM	1
NME7	ENSG00000143156	rs2736899	0.816	SKLM	1
CEP76	ENSG00000101624	rs877081	0.798	SKLM	1
MAPK12	ENSG00000188130	rs4283161	0.783	SKLM	1
KIAA0391	ENSG00000258790	rs2749539	0.765	SKLM	2
DNASE2	ENSG00000105612	rs877081	0.730	SKLM	1
SKA2	ENSG00000182628	rs2005945	0.992	VAF	10
FAM153C	ENSG00000204677	rs2749539	0.989	VAF	2
RP11-844P9.1	ENSG00000251458	rs2749539	0.967	VAF	2
UBE2QL1	ENSG00000215218	rs8022616	0.959	VAF	1
DLG1	ENSG00000075711	rs2005945	0.956	VAF	1
REEP5	ENSG00000129625	rs2749539	0.946	VAF	1
C3orf80	ENSG00000180044	rs2005945	0.943	VAF	1
GNG2	ENSG00000186469	rs2736899	0.934	VAF	2
LUC7L3	ENSG00000108848	rs2005945	0.926	VAF	1
ARPC5	ENSG00000162704	rs2749539	0.903	VAF	2
SLC27A2	ENSG00000140284	rs35854995	0.895	VAF	1
TMED8	ENSG00000100580	rs35854995	0.888	VAF	1
L3MBTL4	ENSG00000154655	rs2749539	0.887	VAF	2
DENR	ENSG00000139726	rs12590834	0.887	VAF	1
GSKIP	ENSG00000100744	rs35854995	0.883	VAF	1
RP11-463J10.3	ENSG00000259007	rs2749539	0.868	VAF	1
UBE3A	ENSG00000114062	rs35854995	0.864	VAF	1
OGT	ENSG00000147162	rs35854995	0.864	VAF	2
PIGCP1	ENSG00000213713	rs35854995	0.864	VAF	2
ZNF358	ENSG00000198816	rs11620763	0.835	VAF	2
SGCB	ENSG00000163069	rs113375097	0.828	VAF	1
NAE1	ENSG00000159593	rs4900229	0.823	VAF	2
C1orf43	ENSG00000143612	rs113375097	0.820	VAF	1
CXCL14	ENSG00000145824	rs1956179	0.820	VAF	1
BNC2	ENSG00000173068	rs113375097	0.819	VAF	1
LPCAT3	ENSG00000111684	rs1956179	0.802	VAF	1
PLCD1	ENSG00000187091	rs1956179	0.795	VAF	1
TACSTD2	ENSG00000184292	rs1956179	0.795	VAF	2
CHMP3	ENSG00000115561	rs1956179	0.795	VAF	1
ALS2	ENSG00000003393	rs1956179	0.795	VAF	1
ULK2	ENSG00000083290	rs1956179	0.793	VAF	1
THAP3	ENSG00000041988	rs1956179	0.792	VAF	1
MAP3K13	ENSG00000073803	rs2005945	0.788	VAF	1
BCL7B	ENSG00000106635	rs2005945	0.788	VAF	1
LRP2	ENSG00000081479	rs2005945	0.788	VAF	1
ZSCAN5A	ENSG00000131848	rs2005945	0.788	VAF	1

NNT	ENSG00000112992	rs2005945	0.788	VAF	1
LPIN1	ENSG00000134324	rs2005945	0.788	VAF	1
TP53	ENSG00000141510	rs2005945	0.788	VAF	1
ATL1	ENSG00000198513	rs2005945	0.788	VAF	1
CAV2	ENSG00000105971	rs2005945	0.788	VAF	1
UBE2J1	ENSG00000198833	rs2005945	0.788	VAF	1
TNPO2	ENSG00000105576	rs2005945	0.788	VAF	1
CD163	ENSG00000177575	rs2005945	0.788	VAF	1
L3MBTL3	ENSG00000198945	rs2005945	0.788	VAF	1
MTMR3	ENSG00000100330	rs2005945	0.788	VAF	1
TCEB3	ENSG00000011007	rs2005945	0.788	VAF	1
HLA-DPA1	ENSG00000231389	rs2005945	0.788	VAF	1
UBE4B	ENSG00000130939	rs2005945	0.788	VAF	1
ANKFY1	ENSG00000185722	rs2005945	0.788	VAF	1
CPNE3	ENSG00000085719	rs2005945	0.788	VAF	1
DTNA	ENSG00000134769	rs1956179	0.785	VAF	1
LRTOMT	ENSG00000184154	rs1956179	0.784	VAF	1
GDPD3	ENSG00000102886	rs11620777	0.783	VAF	1
CSRNP3	ENSG00000178662	rs11620777	0.783	VAF	1
WAPAL	ENSG00000062650	rs1956179	0.783	VAF	1
UAP1L1	ENSG00000197355	rs1956179	0.782	VAF	1
C1orf63	ENSG00000117616	rs1956179	0.781	VAF	2
ATP11A	ENSG00000068650	rs1956179	0.781	VAF	1

Table 2.8: All genes associated with variation for plasma cortisol across all STARNET tissues (FDR = 15%). Only unique associations are included with the top SNP-gene pair. Number of associations refers to the total number of cortisol associated SNPs associated with a given gene.

Tissue	Cluster enrichment score	term
LIV	1.24	GO:0010646~regulation of cell communication
		GO:0023051~regulation of signaling
		GO:0009966~regulation of signal transduction
	1.19	GO:0030695~GTPase regulator activity
		GO:0060589~nucleoside-triphosphatase regulator activity
		GO:0043087~regulation of GTPase activity
	1	domain:RRM 2
		omain:RRM 1
		IPR000504:RNA recognition motif domain
		IPR012677:Nucleotide-binding, alpha-beta plait
		GO:0000166~nucleotide binding
		RNA-binding
SF	1.99	GO:0030057~desmosome
		GO:0030057~desmosome
		GO:0005911~cell-cell junction
	1.29	GO:0009117~nucleotide metabolic process
		GO:0006753~nucleoside phosphate metabolic process
		GO:0055086~nucleobase-containing small molecule metabolic process
VAF	1.51	GO:0009898~cytoplasmic side of plasma membrane
		GO:0098562~cytoplasmic side of membrane
		GO:0031234~extrinsic component of cytoplasmic side of plasma membrane
		GO:0019898~extrinsic component of membrane
		GO:0019897~extrinsic component of plasma membrane
		GO:0098552~side of membrane
	1.34	Endoplasmic reticulum
		GO:0005783~endoplasmic reticulum
		GO:0044432~endoplasmic reticulum part
	1.11	GO:0090174~organelle membrane fusion
		GO:0016050~vesicle organization
		GO:0044801~single-organism membrane fusion
		GO:0048284~organelle fusion
		GO:0061025~membrane fusion
	1.05	GO:0006644~phospholipid metabolic process
		GO:0008654~phospholipid biosynthetic process
		GO:0006650~glycerophospholipid metabolic process
		GO:0046486~glycerolipid metabolic process
		GO:0046474~glycerophospholipid biosynthetic process
		GO:0008610~lipid biosynthetic process
		GO:0045017~glycerolipid biosynthetic process
		GO:0090407~organophosphate biosynthetic process

Table 2.9: Function enrichment of CORNET trans-genes using DAVID for top tissues.

Tissue	Cluster enrichment score	Term
VAF (K=1)	1.89	GO:0000278~mitotic cell cycle
		GO:1903047~mitotic cell cycle process
		GO:0051726~regulation of cell cycle
		GO:0022402~cell cycle process
		GO:0007049~cell cycle
	1.66	GO:0033043~regulation of organelle organization
		GO:1902589~single-organism organelle organization
		GO:0051128~regulation of cellular component organization
	1.41	GO:0007093~mitotic cell cycle checkpoint
		GO:0045930~negative regulation of mitotic cell cycle
		GO:0000075~cell cycle checkpoint
	1.4	GO:0009898~cytoplasmic side of plasma membrane
		GO:0098562~cytoplasmic side of membrane
		GO:0098552~side of membrane
	1.37	Endoplasmic reticulum
		GO:0005783~endoplasmic reticulum
		GO:0044432~endoplasmic reticulum part
	1.23	GO:0051726~regulation of cell cycle
		GO:0016032~viral process
		GO:0044764~multi-organism cellular process
		GO:0044403~symbiosis, encompassing mutualism through parasitism
		O:0044419~interspecies interaction between organisms
		GO:0044802~single-organism membrane organization
	1.16	GO:0031625~ubiquitin protein ligase binding
		GO:0044389~ubiquitin-like protein ligase binding
		GO:0031625~ubiquitin protein ligase binding
	1.15	GO:0065003~macromolecular complex assembly
		GO:0006461~protein complex assembly
		GO:0070271~protein complex biogenesis
		GO:0071822~protein complex subunit organization
	1.15	GO:0070062~extracellular exosome
		GO:1903561~extracellular vesicle
		GO:0043230~extracellular organelle
		GO:0031988~membrane-bounded vesicle
		GO:0070062~extracellular exosome
		GO:0044421~extracellular region part
		GO:0005576~extracellular region
	1.05	GO:0065003~macromolecular complex assembly
		GO:0022607~cellular component assembly
		GO:0044085~cellular component biogenesis

Table 2.10: Function enrichment of CORNET trans-genes using DAVID visceral adipose following K-means clustering.

Chapter 3

Glucocorticoid regulated causal gene networks

3.1 Introduction

3.1.1 Glucocorticoid regulated trans-genes

The previous chapter described the identification of genes trans-associated with genetic variation for plasma cortisol at the *SERPINA6*/ *SERPINA1* locus on chromosome 14 (Chapter 2.3.3). These trans-genes were primarily expressed in subcutaneous fat, visceral abdominal fat and liver, all tissues in which glucocorticoids play an important role. Glucocorticoid signalling in liver regulates genes required for metabolising lipids and glucose[179]. Glucocorticoids also play a crucial role in adipogenesis and are involved in lipolysis during fasting, yielding glycerol which is required for gluconeogenesis in the liver[180].

Glucocorticoids primarily influence gene expression through activation of GR. As a dimer, GR is capable of activating transcription of target genes through the binding of GREs[125] (Chapter 1.6). This chapter investigates whether the influence of genetic variation for plasma cortisol is mediated by trans-genes that are specifically

regulated by GR. We identify a subset of cortisol associated trans-genes where there is prior evidence of glucocorticoid regulation from the literature.

Chromatin Immunoprecipitation sequencing (ChIP-seq) has been established as a tool for studying the role of protein-DNA interactions. This involves using a cross linking agent, such as formaldehyde, to capture the protein-DNA interaction within cells to preserve the chromatin conformation. Cells are then lysed, fragmented and the protein-DNA complexes are precipitated, usually through the use of antibodies specific to the protein. When coupled with next generation DNA sequencing, it becomes possible to analyse millions of protein bound DNA fragments[181].

ChIP-seq has been shown to be an important method for identifying transcription factor binding sites, such as GREs. By using antibodies specific to the transcription factor under investigation, it is possible to isolate and sequence the bound DNA, which can then be mapped back to the target gene. Projects such as The Encyclopedia of DNA Elements (ENCODE) have utilised ChIP-seq, in addition to other techniques such as DNase footprinting, to identify transcription factor binding interactions across the genome[182]. Other projects such as ChEA have integrated different ChIP experiments (ChIP-X) to produce databases of transcription factor binding targets, including targets for *NR3C1*, the gene that encodes GR[183].

In addition to global binding approaches for identifying transcription factor binding, perturbation based experiments can identify transcription factor targets under prescribed biological conditions. As glucocorticoids have a highly specified response, it is important to consider tissue context of transcription factor binding. Dexamethasone is a synthetic GR agonist and is commonly used to activate GR in a laboratory setting in place of endogenous glucocorticoids[184]. Experiments by Yu *et al*[185], have considered the role of GR binding using ChIP-seq to identify GR binding sites in 3T3-L1 adipocytes treated with dexamethasone, identifying 274 glucocorticoid responsive genes. Furthermore, adrenalectomised mice treated with the GR agonist dexamethasone, have been used to identify GR responsive genes with RNA-seq in subcutaneous adipose tissue (Morgan *et al*, Unpublished). It should however

be noted that as dexamethasone is a far stronger GR agonist than endogenous cortisol[186], activation of GR by dexamethasone may not be as clinically comparable to cortisol activated genes.

The aforementioned datasets from both large scale, transcription factor binding experiments and tissue specific perturbation experiments were used to identify a subset cortisol associated trans-genes with evidence of regulation by glucocorticoids. Assigning targets to transcription factors is a non-trivial problem, as there is limited overlap between different methods, as many transcription factor interactions with chromatin do not result in changes in gene expression[187]. Therefore it is important to consider the global impact of transcription factor binding in addition to experiments targeting activation of specific transcription factors within the biological context e.g. GR activation in response to dexamethasone treatment. This approach aims to address the hypothesis that genetic variation for plasma cortisol impacts phenotypic variation through modulation of GR action.

3.1.2 Glucocorticoid regulated gene networks

To examine the transcriptional impact of genetic variation for cortisol, as mediated by glucocorticoid regulated trans-genes, causal inference approaches were utilised to develop causal gene networks. Not all trans-genes will directly impact phenotypic change, but causal networks have the capability to identify cortisol associated genes that are key regulators of networks of genes that may impact phenotypic variation. Here we use eQTLs to establish causal relationships between pairs of genes, from which networks of causal gene-gene relationships can be reconstructed.

In this chapter we used causal inference to identify pairwise causal relationships between genes, using transcriptomic and genomic data from STARNET. This involved the use of the software package Findr[92], which implements a series of statistical tests to infer causality between a pair of genes using eQTLs as genetic instruments (Chapter 1.4). Findr introduces six likelihood ratio tests (Figure 3.1A) relating to a given gene pair (genes A and B) and an eQTL of gene A (E). For each test, Findr

calculates the Bayesian posterior probability for the selected hypothesis being true, allowing for the development of composite tests to infer the probability of an A →B relationship while using E as an instrumental variable (Figure 3.1B).

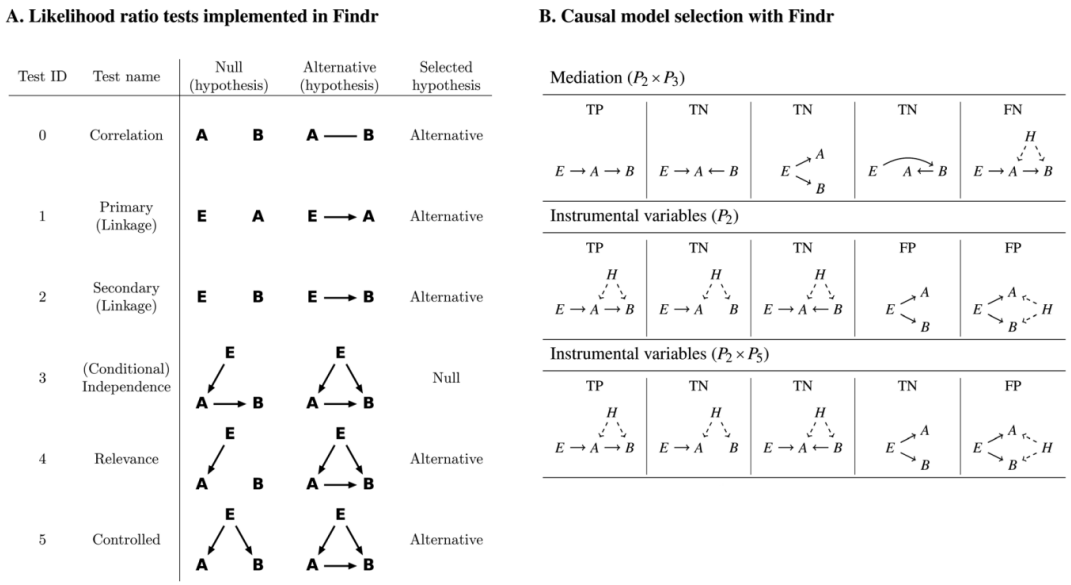


Figure 3.1: (A) Likelihood ratio tests from Findr package used to test for a causal relationship between genes A and B using eQTL E. (B) Test combination in Findr. Composite tests can be constructed from the base Findr tests to infer an A →B relationship. Row 1 shows mediation analysis as inferred by P2*P3. A false negative (missing interaction) is returned in instances where a true interaction is subject to unobserved or unknown confounding (H). Row 2 shows the secondary linkage test (P2) which is capable of resolving causal models with unobserved confounding, although it is unable to distinguish causal models from models where E is a common cause of A and B independently, which raises a false positive (wrong prediction). Row 3 shows instrumental variable analysis where P2*P5 is capable of resolving causal models *with* unobserved confounding, as well as models where A and B are associated independently with E *without* unobserved confounding (i.e. E explains all the correlation between A and B), and will only return a false positive if E affects A and B independently but does not explain all of the correlation between A and B due to unobserved confounding. Figure adapted from Ludl and Michoel, 2021[188].

eQTLs have been used in mediation analysis to distinguish between different causal models and to separate traits that are causal for complex traits from those which are reactive. In mediation analysis, causality is established between two traits (A and B) by testing if the association between the eQTL and one of the traits is mediated by the other trait and exhibits conditional independence with the eQTL[30]. An A →B relationship is inferred if, an E →B relationship can be established which is dependent on A i.e. the relationship collapses following the blocking of (conditioning on) A. If this is the case, a causal A →B relationship can be established as the

flow of information will travel from the mediating trait to the other.

Examples of mediation analysis with QTLs include likelihood-based causality model selection (LCMS)[189], which selects the best fitting causal model using conditional correlations between QTLs and two traits. This approach has also been applied specifically to examine relationships between genes with methods such as Trigger, using the same concept of combining posterior probabilities as calculated in Findr to test for mediated associations between eQTLs and RNA-expression levels[167]. Millstein and colleagues used mediation analysis to develop the Causal Inference Test (CIT) which is a method similar to LCMS that uses conditional independence to determine if the mediator is causal for the trait[90].

The mediation method can be used in Findr as a composite of the secondary linkage and conditional independence tests (P2*P3). However, a significant drawback of mediation analysis is that it preforms poorly in instances where common regulators of A and B act as a confounder, where it will reject the hypothesis of a causal interaction, leading to a false negative prediction. The original Findr manuscript demonstrated that P2*P3 fails in the presence of hidden confounders and is unable to detect weak causal interactions[92]. When using large datasets such as STARNET, increased power obtained from large sample sizes increases the probability of detecting weak correlations due to unknown confounders (common upstream regulators), adding to the false negative rate. Thus, paradoxically, the performance of mediation was shown to *decrease* with increasing sample size[92].

The controlled test (P5) is a causal inference test, which can be used as a composite test with secondary linkage (P2*P5) to infer a causal A $\rightarrow$B relationship while using a cis-eQTL, E, as an instrumental variable. P5 examines whether A and B are associated independently with E, while P2 tests for a direct association between E $\rightarrow$B. Previous work has demonstrated that most cis-eQTLs are only associated with a single gene[190], therefore selecting cis-eQTLs specifically as an instrument, allows E $\rightarrow$B to be used as a proxy for estimating causal effects between A $\rightarrow$B. When combined with P2, P5 can then be used to account for the comparatively few instances

where E is a cis-eQTL for more than one gene, although in such cases a false positive may still occur when A and B are confounded by a common regulator. Therefore it is important when using P2*P5 to examine manually all cis-associations for selected E of interest, to account for any sources of pleiotropy that may have been missed by P5.

Systematic analysis by the Findr authors demonstrates that the increased false positive rate of P2*P5 when compared to P2*P3 is outweighed by the benefit of a greatly decreased false negative rate, and P2*P5 performs significantly better in both synthetic and real world data[92]. Based on this evidence, P2*P5 was taken forward as the primary approach for estimating causal effects between glucocorticoid regulated trans-genes and other genes in STARNET. This required cis-eQTL discovery to be undertaken for all GR-regulated trans-genes, where only genes with a valid cis-eQTL could be included in the analysis.

3.1.3 Chapter objectives

1. To identify cortisol linked trans-genes that have evidence of regulation by glucocorticoids.

2. To estimate causal effects between glucocorticoid regulated trans-genes and other genes in respective tissue.

3. To reconstruct glucocorticoid responsive gene networks.

3.2 Materials and methods

3.2.1 Identification of glucocorticoid regulated trans-genes

GR target datasets

Multiple datasets were used to identify genes that had prior evidence of regulation by GR (Table 3.1). These datasets have been filtered to include targets for *NR3C1*.

Dataset	Study type	Target definition
Yu et al, 2010[185]	ChiP-seq and microarray analysis in adipocyte specific (3TS-L1).	Use of PinkThing to identify gene closest to peak.
Morgan, unpublished*	RNA-seq for subcutaneous and epididymal adipose from dex treated mice.	Differential expression based on fold change.
ENCODE[182]	Large scale high throughput ChiP-seq.	Map ChIP-seq peak to closest gene. If multiple peaks are identified then it is closest to the gene.
TRANSFAC[191]	Curated Transcription Factor Targets dataset based on Positional Weight Matrices (PWMs).	PWMs are used to uncharacterised bindings sequences to closet gene using MATCH.
CHEA[183]	ChIP-X Enrichment Analysis from published ChIP-X experiments.	Target calling based on 400 bp sliding window to detect overlap between peak and closest gene.

Table 3.1: Datasets for identifying genes regulated by glucocorticoids. *Murine RNA-seq data is currently unpublished and was kindly provided by Dr Ruth Morgan from The University of Edinburgh Centre for Cardiovascular Science.

Prioritisation of GR regulated trans-genes

Genes were labelled according to evidence of GR regulation from datasets shown in Table 3.1. This approach was taken for trans-genes from STARNET liver, subcutaneous fat and visceral abdominal fat. The critera included: 1) appearing in a transcription factor database (ENCODE, TRANSFAC, CHEA). 2) Identified as GR target from ChIP-seq experiment in adipocytes from Yu *et al*[185] as according to authors

criteria for calling GR targets. 3) Gene is differentially expressed in response to dex treatment in adipocytes from Yu *et al*[185] as according to authors criteria for calling GR targets. 4) Murine homolog of human gene is differentially expressed in murine RNA-seq experiments (FC > 1; p-value < 0.05). Genes were then given a score and ranked according how well they met the criteria for GR regulation (+1 for each item matched from criteria 1-4).

For each GR trans-gene, the direction of the cortisol associated effect was estimated from the Pearson correlation coefficient of the gene expression level within the associated tissue and the associated cortisol SNP genotype (-012 format). A positive correlation was reported as a gene up-regulated in response to the effect of the alternate allele, and a negative correlation was reported as down-regulation.

Transcription factor target enrichment

Enrichment of transcription factors targets within gene sets was performed using Fisher's exact test using the Python module Scipy Stats. This involved the creation of a 2x2 contingency table based on a tissue specific background consisting of all genes tested in the trans-gene analysis (Table S3.11).

3.2.2 Gene network reconstruction

cis-eQTL discovery for instrument selection

Cis-eQTL discovery was carried out to identify genetic instruments to be used for causal inference analysis with Findr (Figure 3.2). An automated pipeline was established to use the secondary linkage test (P2) to calculate SNP-gene associations when supplied with a list of genes. This was completed using the same process as described in methods section 2.2.3 where SNP-gene associations were obtained between all SNPs within 1 Mb of the trans-gene and all other genes for the tissue specific to the trans-gene.

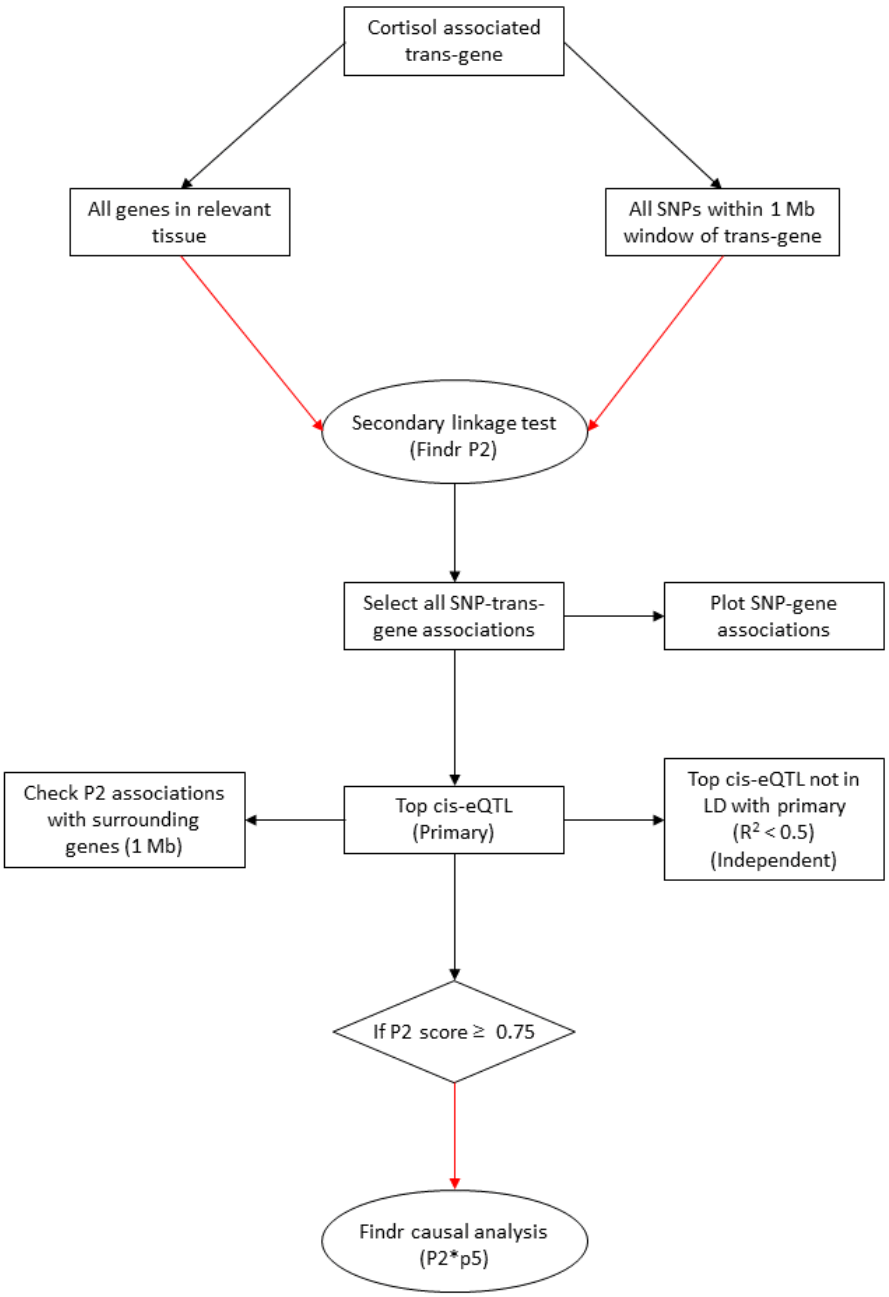


Figure 3.2: Instrument selection for causal analysis with Findr. Flowchart depicts identification of cis-eQTLs for use as genetic instruments for causal analysis with Findr.

Associations between all SNPs and the trans-gene were extracted from the output. and a primary cis-eQTL was selected for each gene, defined as the SNP-gene association with the highest Findr P2 score for the trans-gene. An alternate, independent, cis-eQTL was selected as the second strongest cis-association, not in LD with the primary cis-eQTL. LD between SNPs was calculated as the Pearson corre-

lation coefficient between the primary cis-eQTL genotype with all other SNP genotypes. The alternate cis-eQTL was defined as the top cis-association, which was not in LD with the primary cis-eQTL ($R^2 < 0.5$).

To test for pleiotropy between the selected instrument and other cis-genes, cis associations between all cis genes (± 1 Mb of the trans-gene) and the primary instrument were obtained. These associations were then extracted and compared to the cis-association between the instrument and trans-gene in question.

Pairwise causal inference with Findr

All genes with a valid cis-eQTL (P2 ≥ 0.75) were taken forward for causal analysis with Findr[92] (version 1.0.8). Causal relationships were inferred between these cis-eQTL genes (A-genes) and all other genes in the same tissue (B-genes). The input was as follows: (dg) array of eQTL genotypes A-gene in 012 format, (d) array of normalised A-gene expression levels, (dt) array of expression levels for all B-genes in the relevant tissue sorted with d appearing on top.

The output of all tests in Findr was calculated using the *pijs_gassist* function from the Findr Python package. The posterior probability of a causal interaction (P(A →B)) was calculated from the product of the alternative hypotheses from the secondary linkage test (P2) and the controlled test (P5) (Figure 3.1). This approach was undertaken for each A-gene in a given tissue in a iterative fashion. Following completion of all A-genes in a tissue, the output was converted from the default matrix format to a Pandas DataFrame.

Implementation of global FDR threshold

Each tissue specific gene set of A →B pairwise interactions was filtered according to a local precision FDR threshold (Findr score) for each interaction, to correspond to a global FDR for all interactions in the tissue set. The Findr score for a given A →B pairwise interaction, is calculated as 1 - the probability of that interaction being a

false positive.

To obtain the probability of a false positive across all interactions in a gene set, this was calculated as 1 minus the mean of all local precision FDR scores for a given tissue. A Findr score cut off was set to reflect the desired global FDR[167].

Network visualisation using Cytoscape

Networks were assembled using the network visualisation tool, Cytoscape (version 3.8.0). Networks were assembled from FDR thresholded pairwise gene interactions previously described. These were assembled as directed networks where the A-gene acts as the parent node and the B-gene as the child node, with the posterior probability of an A $\rightarrow$B interaction forming the network edge.

3.3 Results

3.3.1 Identification of glucocorticoid responsive trans-genes in liver and adipose tissue

GR regulated trans-genes identified from transcription factor datasets and experimental data

Having previously identified genes trans-associated with genetic variation for plasma cortisol (Chapter 2.3.3), we next aimed to define a subset of these genes that were regulated by GR. These genes were cross-referenced with GR targets identified from a variety of different sources as described in Table 3.1. This included large projects such as ENCODE, TRANSFAC and CHEA which predict transcription factor binding targets from high throughput transcription factor binding assays. We also included predicted GR targets from perturbation based experiments, in specific tissues. ChIP-seq and microarray analysis has been used to identify 274 glucocorticoid regulated genes in 3TS-L1 adipocytes, a murine derived cell line. In addition RNA-seq data in subcutaneous fat from adrenalectomised mice treated with dexamethasone, a GR agonist, has been used to identify genes that are differentially expressed in response to dex treatment.

The greatest number of unique cortisol associated trans-genes were identified in liver (n=43), subcutaneous fat (n=54) and visceral abdominal fat (n=59) at a 15% FDR threshold. The involvement of these tissues in glucocorticoid signalling has been well documented in the literature and as a result the identification of GR regulated trans-genes was restricted to these tissues. As the 3TS-L1 adipocyte study was carried out in adipose specifically, comparisons of these genes were only made with subcutaneous and visceral adipose trans-genes. Likewise, as the murine RNA-seq experiments were restricted to subcutaneous adipose, only subcutaneous adipose trans-genes were compared to these differentially expressed genes.

Liver

In the liver trans-gene set, 19/43 genes were identified that were present in either the ENCODE, TRANSFAC or CHEA datasets (FDR = 15%). This includes *SERPINA6* which is cis-associated with the genetic variation for plasma cortisol, as described previously (Chapter 2.3.1). One gene, *CPEB2*, was identified in more than one dataset and was present in both ENCODE and CHEA. *CPEB2* (P2 score = 0.89) is a regulator of translation, splice variants of which have been linked to cancer metastasis[192].

Gene name	Transcription factor db	GR count	Pearson R^2	Direction
CPEB2	CHEA, ENCODE	2	0.173	+
SERPINA6	ENCODE	1	0.260	+
RDX	ENCODE	1	-0.130	-
YBX3	ENCODE	1	-0.178	-
TOR4A	ENCODE	1	0.065	+
OGT	ENCODE	1	-0.167	-
ALG5	ENCODE	1	0.169	+
EIF3H	CHEA	1	-0.194	-
CNKSR2	CHEA	1	-0.111	-
PCSK5	ENCODE	1	-0.100	-
FOXN2	ENCODE	1	-0.088	-
NAB2	ENCODE	1	0.153	+
DNM1L	ENCODE	1	-0.029	-
NUCB2	CHEA	1	-0.036	-
PTPN12	ENCODE	1	-0.120	-
SLC26A1	ENCODE	1	-0.145	-
MAPK11	ENCODE	1	-0.177	-
ZNF649	ENCODE	1	0.073	+
RABEP1	ENCODE	1	0.089	+

Table 3.2: Cortisol associated trans-genes from STARNET-liver (FDR = 15%) with evidence of GR regulation. Transcription factor db includes ENCODE, TRANSFAC and CHEA transcription factor datasets. Direction of effect is estimated from the Pearson correlation coefficient of the gene expression level and cortisol associated genotype.

Subcutaneous fat

Out of the cortisol associated trans-genes identified in subcutaneous adipose (FDR = 15%), 28/54 genes were present in either a transcription factor dataset or identified

from the adipose specific perturbation datasets. There were 13 genes that had been identified as GR targets from both high throughput transcription factor binding assays and adipose specific experiments. Evidence from both types of assay indicates that these are the strongest candidates for GR regulation.

The subcutaneous fat trans-gene with the strongest evidence of GR regulation is *PKP2* (P2 score = 0.86) which was identified as a GR target in both CHEA and ENCODE as well as being differentially expressed in dex treated mice and identified as a GR target in adipocytes[185]. *PKP2* is a poorly characterised gene that encodes desmosomal protein plakophilin-2. Deletions in this gene have however been associated with arrhythmogenic right ventricular cardiomyopathy. Interestingly, *PKP2* has also been shown to respond to dex treatment in a reporter system, however the authors were unable to detect a traditional GRE consensus sequence for *PKP2*[193].

Visceral adipose fat

Visceral adipose tissue had the largest number of cortisol associated trans-genes. 21/59 of these genes had some evidence of being targets of GR, which is less than for subcutaneous adipose. There were 5 genes that had been identified as GR targets from both high throughput transcription factor binding assays and adipose specific experiments. The top gene in this case was *BNC2* which was identified as a GR target in CHEA, TRANSFAC and ENCODE as well as appearing as a target from ChIP-seq in adipocytes. *BNC2* encodes Basonuclin 2 which is a zinc finger protein. Mutations in this gene have been linked to congenital lower urinary-tract obstruction[194].

No enrichment of GR regulated trans-genes with Fisher's exact test

Fisher's exact test was used to identify if there are more GR genes in the cortisol associated trans-gene sets than what would be expected by chance. Enrichment analysis was carried out for each tissue set of trans-genes at a 15% FDR threshold across the different GR datasets presented in the previous section. This was performed using

Gene name	ChIP-seq	Microarray	Murine Dex 5% FDR	Transcription factor db	GR count	Pearson R^2	Direction
*PKP2**	True	False	True	CHEA, ENCODE	4	0.088	+
*RNF13**	True	False	False	TRANSFAC, ENCODE	3	-0.171	-
*OSMR**	True	False	True	ENCODE	3	-0.106	-
*PHYH**	True	False	True	CHEA	3	0.076	+
*ZC3H7B**	True	False	True	CHEA	3	-0.052	-
*AMPD3**	True	False	False	CHEA	2	-0.165	-
*PLD1**	True	False	False	ENCODE	2	0.078	+
*IRF2**	True	False	False	ENCODE	2	0.187	+
*LHFPL2**	False	True	False	ENCODE	2	0.048	+
*ME2**	False	False	True	ENCODE	2	0.174	+
*FOS**	True	False	False	ENCODE	2	-0.081	-
*PPCDC**	True	False	False	ENCODE	2	-0.002	-
*AUTS2**	True	False	False	TRANSFAC	2	-0.129	-
STAT4	False	False	True	False	1	0.066	+
XPNPEP1	True	False	False	False	1	-0.161	-
ATP5J2	False	False	False	ENCODE	1	0.051	+
MAMDC2	False	False	True	False	1	0.056	+
KLHDC1	True	False	False	False	1	-0.144	-
PDZD8	False	False	False	ENCODE	1	-0.115	-
USP11	False	False	False	ENCODE	1	-0.086	-
KHK	False	False	False	ENCODE	1	0.172	+
HSPG2	False	False	False	ENCODE	1	0.097	+
PGM1	False	False	True	False	1	0.043	+
PBX2	False	False	True	False	1	-0.046	-
ENSA	False	False	True	False	1	-0.108	-
TSPAN7	True	False	False	False	1	-0.067	-
ATG13	False	False	False	ENCODE	1	0.013	+
PDIA5	False	False	False	ENCODE	1	-0.058	-

Table 3.3: Cortisol associated trans-genes from STARNET-subcutaneous fat (FDR = 15%) with evidence of GR regulation. Transcription factor db includes ENCODE, TRANSFAC and CHEA transcription factor datasets. ChIP-seq and Microarray fields are from Yu et al experiments in adipocytes[185]. Murine dex is from unpublished dexamethasone treated adrenalectomised mice. * Indicates genes that have been identified as GR targets from both global TF binding and perturbation experiments. Direction of effect is estimated from the Pearson correlation coefficient of the gene expression level and cortisol associated genotype.

all expressed genes in the corresponding tissue as a background.

Following analysis across different data sources of GR regulation, there is no indication that there is an enrichment of GR-regulated genes in any of the trans-gene sets examined (Table 3.5)

Gene name	ChIP-seq	Microarray	Transcription factor db	GR count	Pearson R^2	Direction
*BNC2**	True	False	CHEA, TRANSFAC, ENCODE	4	-0.157	-
*DTNA**	True	False	CHEA, TRANSFAC	3	-0.158	-
*LPIN1**	True	False	ENCODE	2	0.125	+
*ULK2**	True	False	ENCODE	2	-0.104	-
*CAV2**	True	False	ENCODE	2	-0.048	-
LUC7L3	False	False	TRANSFAC, ENCODE	2	0.147	+
BCL7B	False	False	ENCODE	1	0.027	+
ATL1	False	False	ENCODE	1	-0.173	-
CD163	False	False	CHEA	1	0.153	+
CXCL14	False	False	TRANSFAC	1	0.205	+
ANKFY1	False	False	ENCODE	1	0.083	+
NNT	True	False	False	1	0.121	+
UBE3A	False	False	TRANSFAC,	1	-0.169	-
OGT	False	False	ENCODE	1	-0.100	-
DENR	False	False	ENCODE	1	0.071	+
CSRNP3	False	False	CHEA	1	-0.168	-
DLG1	True	False	False	1	-0.093	-
REEP5	False	False	ENCODE	1	-0.022	-
WAPAL	True	False	False	1	0.146	+
GNG2	True	False	False	1	-0.098	-
SLC27A2	True	False	False	1	-0.180	-

Table 3.4: Cortisol associated trans-genes from STARNET-visceral adipose fat (FDR = 15%) with evidence of GR regulation. Transcription factor db includes ENCODE, TRANSFAC and CHEA transcription factor datasets. ChIP-seq and Microarray fields are from Yu et al experiments in adipocytes[185]. * Indicates genes that have been identified as GR targets from both global TF binding and perturbation experiments. Direction of effect is estimated from the Pearson correlation coefficient of the gene expression level and cortisol associated genotype.

3.3.2 Identification of cortisol responsive gene networks in hepatic and adipose tissues

Cis-eQTL discovery for prioritised trans-genes

Having identified cortisol associated trans-genes that are regulated by GR, causal analysis was used to identify genes that may regulate transcriptional networks. eQTLs were used as genetic instruments to infer causality between GR trans-genes and other genes in STARNET using the instrumental variable composite test from Findr (P2*P5). This test requires cis-eQTLs as instruments specifically, therefore cis-eQTL mapping was undertaken for all GR regulated trans-genes.

As in Chapter 2.3.1, the linkage test (Findr P2) was used to identify cis-eQTLs for

GR dataset	Tissue	Odds ratio	p-value
TF databse	liver	1.71	0.10
	subcutaneous fat	1.35	0.30
	visceral abdominal fat	0.87	0.67
GR ChiP-seq	subcutaneous fat	1.54	0.16
	visceral abdominal fat	0.93	1.00
GR Microarray	subcutaneous fat	1.73	0.45
	visceral abdominal fat	0.00	1.00

Table 3.5: Enrichment of GR regulated genes within cortisol associated trans-genes sets. Odds ratio and p-value obtained from Fisher's exact test with tissue specific background.

each GR target gene in turn, within the tissue where it is associated with cortisol. For each gene, all SNPs within a 1 Mb window of the target gene were tested against all genes from the relevant tissue, to obtain the null distributions required for Findr P2. The top SNP associations with the cis gene was then extracted. Only genes with a validated cis-eQTL (Findr P2 ≥ 0.75) were taken forward for causal analysis using the eQTL instruments as a genetic instrument in Findr (Table 3.6). Out of a total of 68 trans-genes with evidence of GR regulation 38 had a cis-eQTL that could be taken forward for causal analysis, broken down as 12 from liver, 19 from subcutaneous fat and 7 from visceral abdominal fat (Findr P2 ≥ 0.75).

Reconstruction of gene networks in liver, subcutaneous fat and visceral abdominal fat

Using the alternative test combination in Findr (P2*P5), causal estimates were obtained for pairwise relationships between GR regulated trans-genes and all other genes within the given tissue. This was carried out for all GR regulated trans-genes in liver, subcutaneous fat and visceral abdominal fat with a valid cis-eQTL instrument (Findr P2 ≥ 0.75). Both 10% and 15% global FDR thresholds were then imposed from the mean of the local precision FDR (Findr score) of each gene-gene interaction (Table 3.7). Primary networks were obtained by filtering to include only GR trans-genes with a minimum of 4 targets genes at the global FDR threshold.

Tissue	Gene name	Ensembl gene ID	SNP	P2 Findr score
liver	*ALG5*	ENSG00000120697	rs9576151	1.00
liver	*NUCB2*	ENSG00000070081	rs214080	1.00
liver	*FOXN2*	ENSG00000170802	rs79073127	1.00
liver	*YBX3*	ENSG00000060138	rs11053915	1.00
liver	*PTPN12*	ENSG00000127947	rs7783866	1.00
liver	*SERPINA6*	ENSG00000170099	rs2736898	1.00
liver	*TOR4A*	ENSG00000198113	rs28567631	1.00
liver	*RABEP1*	ENSG00000029725	rs56176579	0.99
liver	*RDX*	ENSG00000137710	rs7107823	0.97
liver	*MAPK11*	ENSG00000185386	rs742186	0.96
liver	*CPEB2*	ENSG00000137449	rs62410848	0.90
liver	*DNM1L*	ENSG00000087470	rs11052028	0.86
liver	*SLC26A1*	ENSG00000145217	rs3733346	0.84
liver	*PCSK5*	ENSG00000099139	rs11145221	0.74
subcutaneous fat	*OSMR*	ENSG00000145623	rs13165709	1.00
subcutaneous fat	*AUTS2*	ENSG00000158321	rs2141205	1.00
subcutaneous fat	*PDZD8*	ENSG00000165650	rs149832558	1.00
subcutaneous fat	*KHK*	ENSG00000138030	rs7560144	1.00
subcutaneous fat	*ATP5J2*	ENSG00000241468	rs138229375	1.00
subcutaneous fat	*PPCDC*	ENSG00000138621	rs3812943	1.00
subcutaneous fat	*PGM1*	ENSG00000079739	rs139945547	1.00
subcutaneous fat	*PHYH*	ENSG00000107537	rs6602646	1.00
subcutaneous fat	*AMPD3*	ENSG00000133805	rs11042759	1.00
subcutaneous fat	*PKP2*	ENSG00000057294	rs12825217	1.00
subcutaneous fat	*STAT4*	ENSG00000138378	rs4341966	1.00
subcutaneous fat	*ATG13*	ENSG00000175224	rs61882678	0.99
subcutaneous fat	*PLD1*	ENSG00000075651	rs10936700	0.97
subcutaneous fat	*IRF2*	ENSG00000168310	rs34985265	0.94
subcutaneous fat	*PBX2*	ENSG00000204304	6:32853219	0.93
subcutaneous fat	*XPNPEP1*	ENSG00000108039	rs3780953	0.90
subcutaneous fat	*RNF13*	ENSG00000082996	rs9853321	0.81
subcutaneous fat	*ZC3H7B*	ENSG00000100403	rs9611739	0.77
visceral abdominal fat	*NNT*	ENSG00000112992	rs6451720	1.00
visceral abdominal fat	*DTNA*	ENSG00000134769	rs71363449	1.00
visceral abdominal fat	*ULK2*	ENSG00000083290	rs79506397	0.88
visceral abdominal fat	*CD163*	ENSG00000177575	rs73059776	0.86
visceral abdominal fat	*LUC7L3*	ENSG00000108848	rs6504682	0.80
visceral abdominal fat	*DENR*	ENSG00000139726	rs73230017	0.76
visceral abdominal fat	*ATL1*	ENSG00000198513	rs61543335	0.76
visceral abdominal fat	*BNC2*	ENSG00000173068	rs10810646	0.75

Table 3.6: GR regulated cortisol linked trans-genes (FDR = 15%) with a valid cis-eQTL for causal analysis (Findr P2 ≥ 0.75).

Liver

In STARNET-liver 48 causal interactions were obtained at a global 10% FDR threshold (Findr score ≥ 0.855) (Table S3.12). When filtering to a minimum of 4 targets, the

Tissue	FDR threshold	Total targets	Network regulator	Regulator targets
liver	15%	197	*CPEB2*	190
	10%	48	*CPEB2*	44
subcutaneous fat	15%	1701	*RNF13*	416
			IRF2	247
			PBX2	883
	10%	486	*RNF13*	215
			IRF2	128
			PBX2	138
visceral abdominal fat	15%	396	*CD163*	378
			LUC7L3	15
	10%	17	*CD163*	4
			LUC7L3	11

Table 3.7: Network targets following FDR filtering. Total targets includes all pairwise interactions at given threshold and network regulators correspond to trans-genes with at least 4 network targets at the given FDR threshold. Inclusive of network regulators present at both 10% and 15% thresholds.

only GR regulated trans-gene that remained was *CPEB2* (Figure 3.3). Notably, this was also the trans-gene that appeared in the most GR target datasets forming a network with 44 target genes. Functional enrichment was performed using DAVID for all *CPEB2* target genes (Table 3.8). The strongest cluster was related to fatty acid beta oxidation and lipid metabolism, including 5 genes related to GO:0006635 - fatty acid beta-oxidation (adj p-value = 0.002). Other enrichments stem from 8 genes related to acquired immunodeficiency syndrome and disease progression (adj p-value = 0.003).

The strongest causal relationship within this network was between *CPEB2* and the gene *HADHA*. (Findr score = 0.99), responsible for encoding the alpha subunit of the mitochondrial trifunctional protein[195]. Mutations affecting this protein have been linked to long-chain 3-hydroxyacyl-CoA dehydrogenase (LCHAD) deficiency, which affects the ability to metabolise fatty acids in the liver[196]. These mutations have also been linked to maternal acute fatty liver during pregnancy[197].

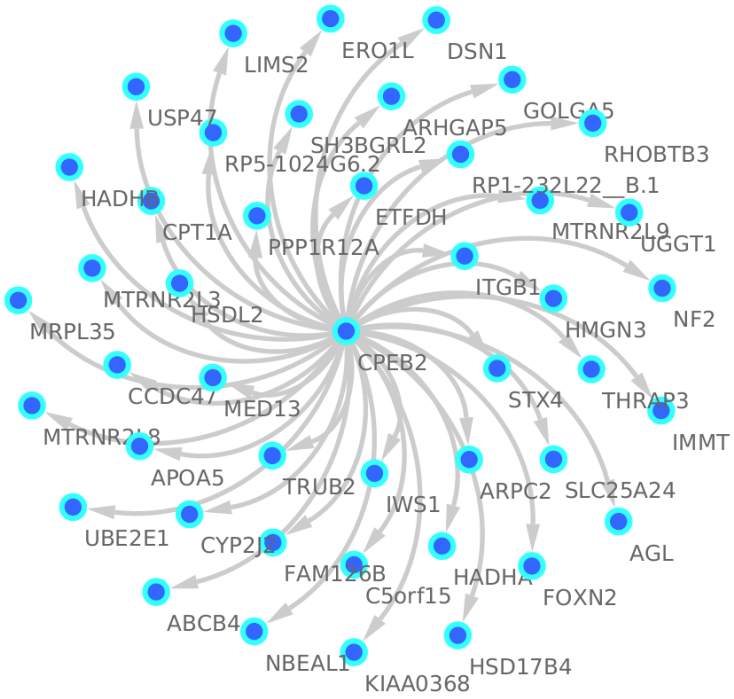


Figure 3.3: 10% FDR gene network for STRANET-liver. Reconstructed from pairwise interactions (A →B) from GR regulated trans-genes (A) with a valid cis-eQTL (Findr P2 ≥ 0.75) against all other STARNET-liver genes. Edges represent Bayesian posterior probabilities of pairwise interaction between genes (nodes). Arrow indicates direction of regulation and interactions were only retained where parent node (A gene) had at least 4 targets.

Subcutaneous fat

In STARNET-subcutaneous fat, 486 causal relationships were detected at a 10% FDR threshold (Findr score = 0.87), which is the most out of all tissues examined (Table S3.13). When filtering to exclude trans-genes with less than 4 targets at this threshold, three major sub-networks are represented under the regulation of the genes *RNF13*, *PBX2* and *IRF2* (Figure 3.4). This includes a total of 481 causal relationships across all three sub-networks, including 8 genes that are co-regulated by at least two sub-networks.

RNF13 is the largest subcutaneous fat sub-network with 215 gene targets at a 10% FDR threshold. *RNF13* encodes IRE1α-interacting protein which plays an important role in the endoplasmic reticulum (ER) stress response, through regulation of IRE1α which is a critical sensor of unfolded proteins[198]. *ERN1*, the gene that encodes IRE1α does not appear as a target of *RNF13* in this sub-network. A detailed

Tissue	Network regulator	Cluster enrichment score	Cluster
Liver	*CPEB2*	3.54	Fatty acid beta oxidation and lipid metabolism
		3.07	Mitochondria and acquired immunodeficiency syndrome
		2.45	3-hydroxyacyl-CoA dehydrogenase activity
		2.1	Mitochondrial inner membrane
		1.58	Fatty acid metabolism
		1.31	Focal adhesion
Subcutaneous fat	*IRF2*	1.94	Poly(A) RNA binding
		1.85	Ubiquitin-associated/translation elongation factor EF1B, N-terminal, eukaryote
		1.76	Viral translational termination-reinitiation
		1.74	Transcription factor activity/ binding
		1.74	Cellular protein catabolic process
	PBX2	2.12	Endoplasmic reticulum unfolded protein response
		1.95	Poly(A) RNA binding
		1.85	Nucleotide-binding, alpha-beta plait
		1.84	transcription factor activity/ binding
		1.63	mRNA splicing, via spliceosome
	RNF13	3.6	Poly(A) RNA binding
		3.18	Centriole
		2.92	Zinc finger, C2H2
		2.92	Transcription regulation
		2.3	Mitotic cell cycle process

Table 3.8: Functional enrichment of causal network targets using DAVID. Filtered to enrichment score > 1

view of the *RNF13* sub-network can be seen in Figure S3.7.

The strongest functional enrichment term for *RNF13* targets is related to Poly(A) RNA binding, where 33 targets are included for this term, GO:0044822 poly(A) RNA binding (adj p-value = 0.01), and 39 targets are included for RNA binding, GO:0003723 RNA binding (adj p-value = 0.04). Other notable terms include 44 genes related to transcriptional regulation (adj p-value) and 23 genes related to Zinc finger motifs (adj p-value = 0.05).

The *IRF2* sub-network contains 128 targets at a 10% FDR threshold. *IRF2* is a member of the Interferon Regulatory Factor (IRF) family and encodes the transcription factor Interferon Regulatory Factor 2. *IRF2* has been shown to play an important role as a repressor of the transcription factor from the same family, *IRF1* which

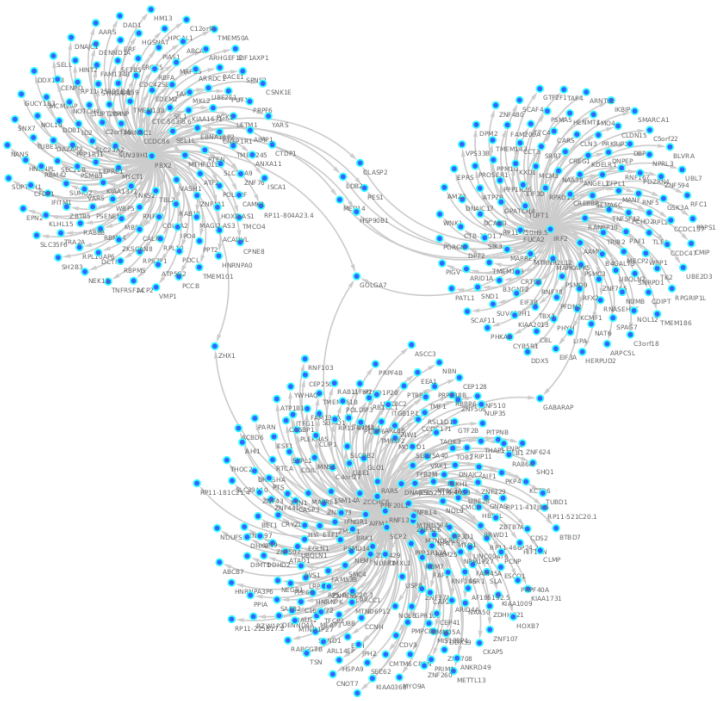


Figure 3.4: 10% FDR networks for STARNET-subcutaneous fat. Reconstructed from pairwise interactions (A →B) from GR regulated trans-genes (A) with a valid cis-eQTL (Findr P2 ≥ 0.75) against all other STARNET-subcutaneous fat genes. Edges represent Bayesian posterior probabilities of pairwise interaction between genes (nodes). Arrow indicates direction of regulation and interactions were only retained where parent node (A gene) had at least 4 targets.

responds to interferons and plays an important role in immune response[199]. Furthermore, *IRF1* has previously been identified as a marker for glucocorticoid sensitivity in peripheral blood[200], although it does not appear as a target of *IRF2* in this network (Findr score = 0.13). A detailed view of the *IRF2* sub-network can be seen in Figure S3.8.

Following functional enrichment of *IRF2* targets, the strongest enrichment term includes 19 genes related to Poly(A) RNA binding (p-value = 0.009), however this association is no longer retained following multiple testing correction. Some notable targets of *IRF2* include *LDB2* (Findr score = 0.94) and *LIPA* (Findr score = 0.91). GWAS suggests functions for *LIPA* related to CAD and ischaemic cardiomyopathy and *LDB2* has been demonstrated to be involved in the development of atherosclerosis[201]. Additionally, cortisol has been shown to induce a 5 fold reduction in *LDB2* expression in adipocytes[202].

The *PBX2* sub-network contains 138 targets at a 10% FDR threshold. The strongest

enrichment in this region includes 6 target genes for endoplasmic reticulum unfolded protein response (p-value = 0.004), however this association is not retained following multiple testing. *PBX2* is another transcription factor and a member of the TALE/PBX homeobox family. It activates transcription through binding to the TLX1 promoter. In this network, it shares five targets with *IRF2* and one with *RNF13*. A detailed view of the *PBX2* sub-network can be seen in Figure S3.9.

Visceral adipose fat

Although STARNET-visceral abdominal fat contained the largest number of trans-associations with cortisol SNPs, the fewest causal relationships were detected in this tissue at 10% FDR (Table 3.14). Two small sub-networks were detected, regulated by the genes *LUC7L3* and *CD163* composed of eleven and four targets specifically. No functional enrichment was observed for any of the targets of *CD163* or *LUC7L3*. Interestingly, when the FDR threshold is reduced to 15% the sub-network for *CD163* is expanded to include 378 targets, a much more dramatic expansion compared to reducing the threshold to 15% FDR with other regulators.

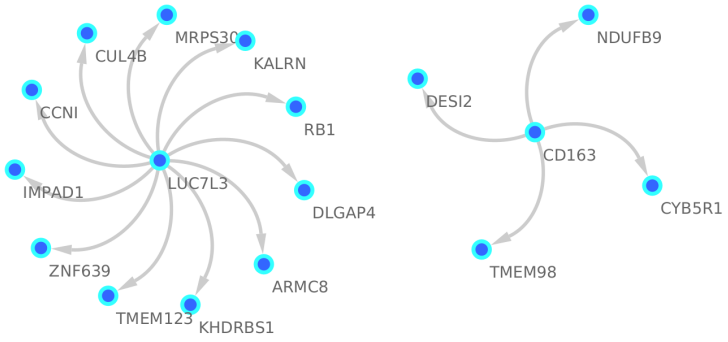


Figure 3.5: 10% FDR networks for visceral abdominal fat Reconstructed from pairwise interactions (A →B) from GR regulated trans-genes (A) with a valid cis-eQTL (Findr P2 ≥ 0.75) against all other STARNET-visceral abdominal fat genes. Edges represent Bayesian posterior probabilities of pairwise interaction between genes (nodes). Arrow indicates direction of regulation and interactions were only retained where parent node (A gene) had at least 4 targets.

CD163 is a haemoglobin scavenger protein that is expressed in macrophages, the expression of which has likely been captured from macrophages within visceral

adipose. It is involved in the clearance of hemoglobin/haptoglobin complexes and may play a role in protection from oxdidative damage from haemoglobin. It has also been shown to play a role in activating macrophages as part of the inflammatory response[203]. *LUC7L3*, also known as CROP, encodes a protein that is involved in alternative splicing and is associated with human heart failure[204]. It has also been shown to play a role in the inhibition of hepatitis B replication[205].

3.3.3 *IRF2* targets overrepresented within *IRF2* network

Predicted *IRF2* transcription factor targets have been previously described as part of the TRANSFAC dataset. We examined the overlap between predicted *IRF2* targets in TRANSFAC and gene targets within the previously described *IRF2* causal networks in subcutaneous fat. A true network of *IRF2* targets would be expected to show an enrichment of predicted *IRF2* targets, compared to what would be expected by chance.

Fisher's exact test was used to examine if there is an enrichment of *IRF2* targets within the causal network identified in subcutaneous fat, across different thresholds. At a 10% FDR threshold, the *IRF2* network had 128 target genes, 35 of which were also predicted *IRF2* targets (p = 0.08). The strongest enrichment was observed at the 15% FDR cut off where 104/247 causal targets were also predicted targets of *IRF2* in TRANSFAC (p = 0.005). Decreasing the global FDR beyond this threshold increases the number of TRANSFAC targets within the pool of causal targets, however at a lower enrichment (p = 0.046).

FDR threshold	Findr threshold	Network targets	*IRF2* targets (TRANSFAC)	odds ratio	P-value
10% FDR	0.87	128	35	1.733	0.080
15% FDR	0.8	247	104	1.903	0.005
20% FDR	0.7	545	229	1.412	0.046

Table 3.9: Fisher's exact test for *IRF2* targets from TRANSFAC predicted targets. STARNET subcutaneous fat genes used as background for enrichment.

In addition to examining the prevalence of *IRF2* targets within the *IRF2* causal

network, we also investigated the overlap between network genes that are also regulated by GR. Interestingly, we observe that there is an enrichment of ENCODE GR targets at the 15-20% FDR threshold ($p \leq 0.05$). The 15% and 20% networks include 68 and 138 GR targets respectively. No GR enrichment was observed in either CHEA or TRANSFAC datasets for *IRF2* networks.

FDR threshold	Findr threshold	Network targets	GR targets (ENCODE)	odds ratio	p-value
10% FDR	0.87	128	25	0.894	0.666
15% FDR	0.8	247	68	1.410	0.019
20% FDR	0.7	545	138	1.263	0.022

Table 3.10: Fisher's exact test for GR targets from ENCODE transcription factor dataset. STARNET subcutaneous fat genes used as background for enrichment.

3.3.4 Application of independent genetic instruments for gene network reconstruction

To study the impact of instrument selection on the reconstruction of causal networks we examined the distribution of local cis-eQTLs for each of the GR regulated trans-genes that was found to regulate a network. Primary instruments were selected as the strongest cis-eQTL within a 1 Mb window of the associated gene, as determined by Findr P2 score. However the landscape of gene expression linked genetic variation can involve several loci associated with the expression of the same gene to differing degrees. In addition to selecting a primary cis-eQTL as an instrument, alternate, independent, instruments were also identified. These were defined as as the second strongest cis SNP-gene association which was not in LD with the primary instrument ($R^2 < 0.5$) (Figure S3.15).

Having previously identified trans-genes responsible for the regulation of transcriptional networks, we next aimed to investigate the prevalence of these networks when using an independent cis-instrument. In each case, an independent instrument associated with the gene in question was identified (Figure 3.6), however causal analysis was only carried out with the independent instrument if it was valid (Findr

P2 ≥ 0.75).

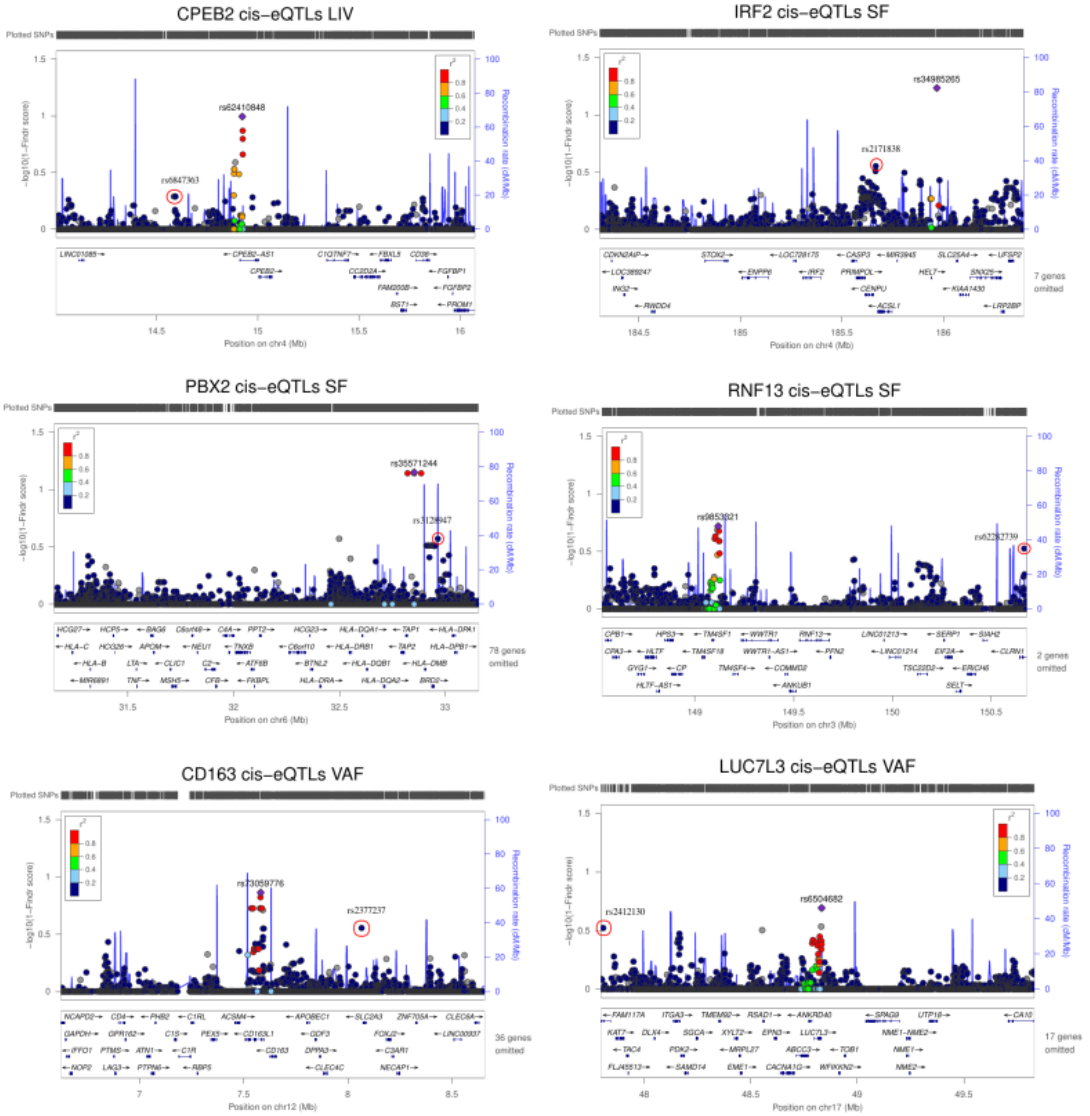


Figure 3.6: Cis-eQTL discovery for network drivers. SNP-gene associations within a 1 Mb window of the associated gene calculated using the Findr secondary linkage test (P2) and presented as 1-findr score (-$\log_{10}$) with LocusZoom. Lead cis-eQTL is primary instrument used for causal analysis. Red circle indicates independent ($R^2 \leq 0.5$) alternate instrument.

Causal relationships in STARNET liver were defined by a GR regulated network under the regulation of *CPEB2* (FDR = 10%) (Figure 3.3). The genetic instrument used to construct this network, rs62410848 (Findr P2 = 0.90), is the strongest cis-eQTL for *CPEB2*, located less than 100 Kb upstream of the *CPEB2* locus. An independent peak was identified 400 Kb upstream of *CPEB2*, represented by rs6847363 as the top cis-association in this region (Findr P2 = 0.48). As this independent in-

strument fell below the required threshold, causal analysis was not carried out using rs6847363 as an instrument.

To determine the robustness of the primary instrument, we examined cis-associations with other genes within this locus (± 1 Mb). While *CPEB2* was the strongest cis-eQTL association in this region, rs62410848 was also seen to be associated with *CD38* (Findr P2 = 0.85), a gene ~800 Kb downstream of *CPEB2*. Although *CD38* is not associated with any cortisol variants at the *SERPINA6*/ *SERPINA1* locus, it has been identified as being regulated by glucocorticoids in smooth muscle cells[206] and has been identified as a GR target in ENCODE. However, *CD38* does not appear as a target of *CPEB2*, which suggests a low P5 score. This suggests that *CPEB2* and *CD38* are independently associated with rs62410848 and that *CPEB2* is the true network regulator in this cis region.

In STARNET-subcutaneous fat, the *IRF2* sub-network was generated using the SNP rs34985265 (Findr P2 = 0.94), located ~500 Kb upstream of *IRF2*. The strongest independent cis-eQTL for *IRF2*, rs2171838 (Findr P2 = 0.72), is located closer to *IRF2*, ~300 of the *IRF2* transcription start site. This association did not reach the association threshold for use as a causal instrument (Findr P2 = 0.72). Examining cis-associations between rs34985265 and all genes within 1 Mb of *IRF2*, *IRF2* is the only gene to show an association with this SNP.

For *RNF13* the primary instrument, rs9853321 (Findr P2 = 0.81), was located in a peak 400 Kb upstream of the *RNF13* transcription start site. The strongest independent cis-eQTL, rs62282739, is located nearly 1 Mb downstream of *RNF13* and was too weak to be taken froward for causal analysis (Findr P2 = 0.70). Cis-associations for rs9853321 in this region, include an association with the gene *TM4SF1* (Findr P2 = 0.93) at a higher level than the association with *RNF13*. There is some indication of a causal relationship between *RNF13* and *TM4S1* (Findr score = 0.73), however *TM4S1* is not a target of *RNF13* at either a 10% or 15% global FDR threshold, suggesting that *TM4S1* is independently associated with rs9853321.

The third subcutaneous fat sub-network was predicted using the SNP rs35571244

as a cis-eQTL for *PBX2* (Findr P2 = 0.93). This SNP is located ~800 Kb downstream of the *PBX2* transcription start site and is the strongest cis-eQTL for a peak of SNPs in this region. An alternate cis-eQTL, rs3128947 (Findr P2 = 0.73), is located 500 Kb upstream of the *PBX2* transcription start site. Again, this cis-eQTL was too weak to be taken forward for causal analysis. There are 31 cis-associations between rs35571244 and genes within a 1 Mb window of *PBX2* at a 15% FDR threshold (Findr P2 ≥ 0.8), of which PBX2 is the 7th strongest association. Of these cis associations, 10 are causal targets of *PBX2* when using rs35571244 as a genetic instrument at a 15% FDR threshold and 4 are targets at a 10% FDR threshold. This suggests that these genes are not independently linked to rs35571244, which raises the possibility that these targets would be predicted from other cis-genes and not just *PBX2* specifically.

The primary instrument used to reconstruct the *CD163* sub-network in visceral abdominal fat, rs7954905 (Findr P2 = 0.86), is located less than 100 Kb downstream of *CD163*. The strongest independent cis-eQTL, rs2377237 (Findr P2 = 0.72), is located ~500 Kb upstream of the *CD163* transcription start site, however this SNP was below the threshold for use a causal instrument. There were 6 cis-associations at a 15% FDR threshold (Findr P2 ≥ 0.78). One of these cis-genes is a target of *CD163* (Findr score = 0.86) at a 15% FDR threshold, but no genes are targets at a 10% FDR threshold. This target gene is *CD163L1*, which is a paralog of *CD163* located downstream of of *CD163*. The peak represented by rs7954905 is located in the *CD163L1* gene body. *CD163L1* arose as a gene duplication of *CD163* and colocalises with *CD163*[207].

The primary instrument used to reconstruct the *LUC7L3* sub-network, rs6504682 (Findr P2 = 0.8), is located within the *LUC7L3* gene body. An independent cis-eQTL, rs2412130 (Findr P2 = 0.7) is located in a peak ~1000 Kb upstream of *LUC7L3*. Again, this alternate cis-eQTL did not meet the threshold for use as an instrument. There is only one other cis-gene associated with rs6504682, *ANKRD40* (Findr score = 0.81), however this gene is not a target of *LUC7L3* in either the 15% or 10% FDR causal networks in visceral adipose.

3.4 Discussion

3.4.1 Cortisol associated trans-genes include genes regulated by GR

Having previously identified genes associated with genetic variation for plasma cortisol, we aimed to prioritise those where there is prior evidence of regulation by glucocorticoids for causal analysis. Previously published studies have used a variety of different methods to identify targets of GR, using both high throughput binding assays and tissue specific perturbation experiments. Using the GR targets identified from these studies, we were able to identify tissue specific subsets of cortisol associated trans-genes that have prior evidence of being regulated by GR.

We focused our examination upon the STARNET tissues with the greatest number of trans-genes identified, specifically liver, subcutaneous fat and visceral abdominal fat. GR associations were identified in all tissues, however most were identified in subcutaneous fat where 28/54 trans-genes were either present in a transcription factor database for GR, or were shown to be GR regulated in a perturbation experiment. However, it is worth noting that as most of the perturbation based experiments were specific to subcutaneous fat, increasing the likelihood of identifying GR targets in this tissue. Likewise, as no hepatic specific datasets for GR regulation were employed, it is possible potential GR targets in STARNET liver have been missed. Trans-genes that have been confirmed as GR targets from multiple sources of evidence were considered more likely true targets and less likely false positive findings. This is particularly true for trans-genes that were confirmed as GR targets from both transcription factor binding studies and perturbation experiments.

In addition to searching systematically for GR targets within the cortisol associated trans-genes, a look-up approach was used to identify associations with canonical GR targets genes that have been well characterised in the literature, such as *PEPCK*[208], *GILZ*[209], *LPL*[210] and *PER1*[211]. None of these genes were identified as trans-genes targets in any STARNET tissue (FDR = 15%), even though several are

expressed in liver and adipose. Additionally, none of these genes appear as targets for any networks reconstructed at either 10% or 15% FDR thresholds. The strongest causal relationship for these genes was detected between *RNF13* and *TSC22D3* (the gene that encodes *GILZ*) (Findr score = 0.76), however following network reconstruction this relationship was not retained.

We also examined whether in addition to containing genes that are glucocorticoid regulated, these trans-gene sets were statistically enriched for GR targets, above what would be expected by chance. Within the defined trans-gene sets (15% FDR), enrichment was not observed for GR targets for any tissue. This may be due to a lack of high quality tissue specific GR sets; as previously mentioned this hampered the analysis of GR targets in liver due to lack of a hepatic GR target set. As GR is ubiquitously expressed and regulates a large number of genes, a number of GR targets would be expected in any random gene set, which raises the threshold of GR-targets needed to observe enrichment. However, by prioritising those trans-genes with multiple sources of evidence, it is possible to identify the strongest candidates for GR regulation within a trans-gene set and reduce the likelihood of identifying false positive targets.

3.4.2 GR regulated trans-genes mediate transcriptional networks

Having prioritised trans-genes that are regulated by GR, causal inference was used to identify key regulators of gene networks. Cis-eQTL discovery was carried out in STARNET to identify cis-eQTLs, that would be suitable for use as genetic instruments. A drawback of this approach is that only genes with a genetic instrument (e.g. a cis-eQTL) could be included in this analysis. Therefore, any regulatory effects mediated by GR regulated trans-genes that do not have a cis-eQTL instrument would be missed. Cis-eQTLs were identified using the Findr P2 method, which was demonstrated to be robust and comparable to methods such as kruX (Figure S2.12). Findr P2 measures the strength of association relative to a null distribution of associations with the genome-wide background, which would correct for SNPs that

systematically have high associations to all genes. This is useful, as it is important when selecting cis-instruments that the instrument is not strongly associated with multiple cis-genes.

Using the instrumental variable test combination in Findr (P2*P5) we identified pairwise causal relationships between GR regulated trans-genes (A) and target genes within the corresponding STARNET tissue (B), while using the best eQTL of A as an instrumental variable. As stated in the chapter introduction (Chapter 3.1.2), mediation based causal inference performs poorly in response to weak interactions and hidden confounders, both of which are common features in transcriptomic networks. As the instrumental variable test combination requires cis-eQTLs for A to be used as an instrument, this allows for any E $\rightarrow$B interactions to be used as a proxy for A $\rightarrow$B. Although this does result in an increased false positive rate, this has been previously demonstrated to be a substantial improvement over the large false negative rate observed with mediation analysis[92].

Having used this test combination, networks were reconstructed from the pairwise interactions identified from Findr. Both 10% and 15% FDR networks were generated to examine the difference in number of associations retained across threshold. We filtered these interactions to include only GR trans-genes (A-genes) which had at least 4 causal targets (B-genes) at a given FDR threshold. These trans-genes with at least 4 targets corresponded to sub-networks of target genes within the larger tissue network.

In liver, *CPEB2* was the only trans-gene to regulate a causal network at a 10% FDR threshold. This is also the only trans-gene in liver that appeared in two transcription factor databases (CHEA and ENCODE), and therefore has the strongest evidence of being GR regulated out of all of the liver trans-genes. *CPEB2* has been annotated as an RNA binding protein and regulator of transcription, which is suggestive of a potential biological mechanism whereby it could mediate transcriptional changes. Additionally, functional enrichment of *CPEB2* targets show an enrichment of genes related to fatty acid oxidation and lipid metabolism. A recent study has also high-

lighted the role of *CPEB2* in the regulation of hormone sensing in mammary gland and plays a role in the development of ER-positive breast tumors[212]. The key *CPEB2* targets from this manuscript do not appear as targets in the causal *CPEB2* networks with *TNFSF11* and *CCND1*, however this study was carried out in breast cancer cell lines rather than liver which account for transcriptional differences.

The largest network at a 10% FDR threshold was identified in subcutaneous fat, composed of three sub-networks driven by the genes *RNF13*, *IRF2* and *PBX2*. As *IRF2* and *PBX2* are both transcription factors, this suggests a biological mechanism explaining how these genes could be involved in mediating transcription. *RNF13* encodes a ubiquitin protein ligase that is linked to the endoplasmic reticulum stress response[198]. Protein ubiqutination has been demonstrated to play a role in the modulation of gene expression[213] which could be a possible link for how *RNF13* may influence gene expression by mediating changes in target proteins. Interestingly, *PKP2* which had the strongest evidence of GR regulation did not appear as a regulator of a transcriptional network or as a network target.

Interferon regulatory factors are a family of transcription factors that play important roles in regulating the immune response and signalling. While *IRF2* does not mediate the production of interferons, as with *IRF3*, *IRF5* and *IRF7*, it has been shown to play a role in regulatory the inflammatory response in macrophages[214] and is involved in the regulation of human keratinocyte stem cell fate[215]. Chromatin analysis has also identified a role for IRF2 in the activation of transcription through binding to gene promoters. In particular, the researchers examined adult and fetal proerythroblast, demonstrating a preference for binding in adult specific genes related to type I interferon and interferon gamma signalling[216]. There were no strong functional terms, enriched among *IRF2* targets, although some key genes include *LDB2* and *LIPA* which have been shown to play a role in the development of CAD and atherosclerosis[201], with *LDB2* appearing as a key regulator of atherosclerosis from the STAGE study[159]. Moreover, evidence demonstrating that *LDB2* expression is modulated by cortisol[202], may be explained by regulation by *IRF2*, although further research would be required to confirm a mechanism.

Previously identified *IRF2* targets, as predicted from ENCODE, appear enriched within the *IRF2* networks, however at a 15% FDR threshold rather than a 10% threshold. This effect is diminished as the threshold increased from 15% to 20%. Interestingly, in addition to seeing an enrichment of *IRF2* predicted targets, at a 15% FDR threshold there is also an enrichment of GR targets among the targets of the *IRF2* network. This is initially somewhat counter-intuitive, considering that *IRF2* itself is GR regulated, the targets of *IRF2* would not necessarily be expected to also be GR regulated. However, this finding may be indicative of a feed-forward loop (FFL) which is a type of transcriptional motif that is often seen enriched within transcriptional networks[217]. Feed-forward loops involve two transcription factors (X and Y) with their own inducers, where X regulates Y, in addition to the targets of Y as well. Such loops have been noted previously in glucocorticoid biology as a regulatory model that modulates GR activity[218], specifically researchers were able to identify a FFL driven by GR and the transcription factor *KLF15* in murine lungs that is linked to the regulation of amino acid metabolism and mitochondrial function[219]. There is a possibility that *IRF2* may be involved in a similar relationship with GR, however this would require further experiments to test for the presence of such a motif.

An examination of the cis-region of *IRF2*, reveals a lack of associations between rs34985265 and any genes other than *IRF2*. This suggests a lack of pleiotropy interfering with any causal estimates, ensuring that *IRF2* is responsible as the regulator of this network and not another gene in this region. This is not as clear in the case of *PBX2*, as the instrument used to predict this network is associated with 31 other genes in the cis region. Of these, 10 are causal targets of *PBX2* at a 15% FDR threshold, which may lead to issues of pleiotropy. In these instances, as the instrument is associated with both the A and B gene, it is unclear which gene is responsible for the signal. Likewise, for *RNF13* there is an additional cis-association between the instrument used to construct the *RNF13* sub-network outside of *RNF13* itself. However, in this case the other cis gene is not a target of *RNF13*, suggesting that this is is not responsible for the *RNF13* network.

When considering the impact of transcription factor interactions on gene ex-

pression data, there are methodological issues related to these analyses. It has been noted by Castadldi *et al*[220] that issues relating to data normalisation of gene expression data have a scale dependent effect on the number of transcription factor interactions identified. Additionally the authors comment on the presence of heteroscedasticity when comparing eQTL effect sizes to GWAS effect sizes for common traits and disease. In this analysis the pre-processing of gene expression data was consistent across STARNET tissues, which would limit the variability of interactions across tissues, however this highlights the importance of validation with independent datasets. Effect sizes for eQTLs were not used in the causal network analysis, therefore any issues relating to heteroscedasticity between eQTL and GWAS effect sizes would have minimal impact on these analyses.

It is worth noting, that although we identify the presence of novel glucocorticoid regulated gene networks, critical regulators of glucocorticoid and circadian mediated transcription have been previously described in the literature. The nuclear receptor hepatocyte nuclear factor 4A (*HNF4A*) plays a pivotal role in the liver transcriptome through the regulation of BMAL1::CLOCK[221,222],. resulting in changes in clock genes. Although *HNF4A* was not identified as GR regulated trans-gene, it would be an interesting experiment to see if it would be possible to reconstruct a network of clock genes using an instrument for *HNF4A* and Findr. Unfortunately, although gene expression data for *HNF4A* are present in STARNET-liver, it does not have a valid cis-eQTL to use as an instrument. This does highlight a drawback of instrumental variable analysis, as biological signals may be missed if an appropriate instrument is not available.

3.5 Conclusion

In this chapter we have demonstrated that there are a sub-set of genes trans-associated with genetic variation for plasma cortisol in liver, that are targets of GR. Moreover, we have identified key genes that are regulators of GR regulated gene networks in liver, subcutaneous and visceral abdominal adipose. These gene networks are regulated by genes such as *CPEB2* which links to transcriptional regulation and *IRF2*, a transcription factor whose targets are enriched within the *IRF2* sub-network. In summary, these networks help to characterise the downstream consequences of genetic variation for plasma cortisol through a mechanism that is mediated by GR.

3.6 Supplementary data

		Trans-genes	
		yes	no
GR-regulated	yes		
	no		

Table 3.11: Example of 2x2 contingency table used to measure GR enrichment within a trans-gene set using Fisher's exact test.

A-genes	B-genes	Findr score	GR count
CPEB2	*HADHA*	0.994	2
CPEB2	*C5orf15*	0.975	2
CPEB2	*CCDC47*	0.954	2
CPEB2	*HADHB*	0.943	2
CPEB2	*TRUB2*	0.939	2
CPEB2	*IMMT*	0.931	2
CPEB2	*SH3BGRL2*	0.920	2
CPEB2	*ARHGAP5*	0.919	2
CPEB2	*ITGB1*	0.916	2
CPEB2	*AGL*	0.910	2
CPEB2	*MTRNR2L3*	0.907	2
CPEB2	*PPP1R12A*	0.904	2
CPEB2	*GOLGA5*	0.904	2
CPEB2	*MRPL35*	0.903	2
CPEB2	*STX4*	0.902	2
CPEB2	*THRAP3*	0.898	2
CPEB2	*UGGT1*	0.897	2
CPEB2	*ARPC2*	0.892	2
CPEB2	*HSDL2*	0.888	2
CPEB2	*CPT1A*	0.886	2
CPEB2	*DSN1*	0.884	2
CPEB2	*RP1-232L22__B.1*	0.882	2
CPEB2	*HSD17B4*	0.881	2
CPEB2	*APOA5*	0.879	2
CPEB2	*NF2*	0.877	2
CPEB2	*LIMS2*	0.876	2
CPEB2	*UBE2E1*	0.875	2
CPEB2	*KIAA0368*	0.872	2
CPEB2	*HMGN3*	0.871	2
CPEB2	*ETFDH*	0.870	2
CPEB2	*FOXN2*	0.866	2
CPEB2	*FAM126B*	0.866	2
CPEB2	*RP5-1024G6.2*	0.864	2
CPEB2	*RHOBTB3*	0.863	2
CPEB2	*CYP2J2*	0.863	2
CPEB2	*SLC25A24*	0.863	2
CPEB2	*IWS1*	0.861	2
CPEB2	*MED13*	0.859	2
CPEB2	*USP47*	0.857	2
CPEB2	*MTRNR2L8*	0.857	2
CPEB2	*NBEAL1*	0.857	2
CPEB2	*ERO1L*	0.856	2
CPEB2	*ABCB4*	0.856	2
CPEB2	*MTRNR2L9*	0.855	2

Table 3.12: All pairwise interactions from Findr (P2*P5) in liver at a 10% FDR threshold.

A-genes	B-genes	Findr score	GR count
IRF2	*TLE4*	0.982	2
IRF2	*BLVRA*	0.982	2
IRF2	*KDELR1*	0.958	2
IRF2	*MAPRE2*	0.956	2
IRF2	*B4GALT3*	0.954	2
IRF2	*SNRPD1*	0.953	2
IRF2	*TRIB2*	0.950	2
IRF2	*FAM206A*	0.949	2
IRF2	*SPAG7*	0.944	2
IRF2	*PSMD9*	0.944	2
IRF2	*ATP7B*	0.944	2
IRF2	*MTRNR2L12*	0.943	2
PBX2	*SUPT4H1*	0.943	1
IRF2	*SMARCA1*	0.943	2
IRF2	*EIF3B*	0.940	2
IRF2	*LDB2*	0.937	2
IRF2	*HERPUD2*	0.937	2
IRF2	*KIAA2013*	0.935	2
IRF2	*CCDC47*	0.935	2
IRF2	*DPM2*	0.933	2
IRF2	*RNF5*	0.932	2
RNF13	*SMC4*	0.931	3
IRF2	*RPL29*	0.931	2
RNF13	*ARL14EP*	0.930	3
RNF13	*PSKH1*	0.930	3
IRF2	*WNK1*	0.930	2
IRF2	*ARPC5L*	0.930	2
RNF13	*SNW1*	0.930	3
PBX2	*PGK1*	0.930	1
IRF2	*CCT3*	0.930	2
RNF13	*KIAA1009*	0.930	3
RNF13	*ZCCHC6*	0.928	3
IRF2	*CRTC2*	0.928	2
IRF2	*CDIPT*	0.928	2
IRF2	*CMIP*	0.926	2
IRF2	*PORCN*	0.924	2
IRF2	*RFX2*	0.924	2
IRF2	*RNF167*	0.924	2
RNF13	*ZHX1*	0.923	3
RNF13	*ATAD1*	0.922	3
RNF13	*NEMF*	0.922	3
RNF13	*ZNF441*	0.922	3
RNF13	*PRPF40A*	0.922	3
RNF13	*RAB6A*	0.922	3
RNF13	*IFNGR1*	0.922	3
RNF13	*ZEB1*	0.922	3
RNF13	*ARL8B*	0.922	3

RNF13	*CDS2*	0.922	3
RNF13	*RP11-521C20.1*	0.922	3
RNF13	*CENPC*	0.922	3
RNF13	*PRPF38B*	0.921	3
RNF13	*LRRCC1*	0.921	3
RNF13	*CRYZL1*	0.921	3
RNF13	*ANKRD49*	0.921	3
RNF13	*ZNF624*	0.921	3
RNF13	*MNS1*	0.921	3
RNF13	*GABARAP*	0.921	3
IRF2	*MANF*	0.920	2
RNF13	*ZNF510*	0.920	3
RNF13	*PLEKHA5*	0.920	3
RNF13	*TOB2*	0.920	3
IRF2	*GSK3A*	0.919	2
RNF13	*ZNF43*	0.919	3
RNF13	*PITPNB*	0.919	3
RNF13	*TUBD1*	0.919	3
RNF13	*NT5C3A*	0.919	3
RNF13	*ING3*	0.919	3
RNF13	*POLDIP3*	0.919	3
RNF13	*DENND4A*	0.919	3
RNF13	*THAP1*	0.919	3
RNF13	*RP11-466P24.2*	0.919	3
RNF13	*ZNF429*	0.918	3
RNF13	*ZNF708*	0.918	3
RNF13	*BZW1P2*	0.918	3
RNF13	*TAOK3*	0.918	3
RNF13	*PHF20L1*	0.917	3
RNF13	*CAPZB*	0.917	3
IRF2	*RFC1*	0.917	2
IRF2	*KXD1*	0.917	2
RNF13	*RTCA*	0.917	3
RNF13	*SLA*	0.916	3
RNF13	*CLMP*	0.916	3
RNF13	*DIMT1*	0.916	3
RNF13	*DROSHA*	0.916	3
RNF13	*ZNF260*	0.916	3
RNF13	*HNRNPK*	0.916	3
RNF13	*ZNF197*	0.915	3
IRF2	*EPRS*	0.914	2
RNF13	*C4orf27*	0.914	3
RNF13	*LRP12*	0.914	3
RNF13	*SEC62*	0.914	3
IRF2	*CYB5R1*	0.913	2
RNF13	*CLIP1*	0.913	3
RNF13	*TMF1*	0.912	3
IRF2	*DNPEP*	0.912	2

RNF13	*PPIA*	0.912	3
RNF13	*VRK1*	0.912	3
RNF13	*SUMO1*	0.912	3
IRF2	*CLDN15*	0.912	2
RNF13	*MTND6P12*	0.912	3
IRF2	*FMO4*	0.911	2
IRF2	*RP11-750H9.5*	0.911	2
IRF2	*TMEM109*	0.911	2
IRF2	*PDZRN4*	0.911	2
IRF2	*PRKRIP1*	0.911	2
IRF2	*SCAF4*	0.911	2
RNF13	*FAF1*	0.911	3
IRF2	*UBE2D3*	0.911	2
RNF13	*TMEM161B*	0.911	3
IRF2	*PSMC3*	0.911	2
IRF2	*NAA38*	0.911	2
IRF2	*ZFPL1*	0.911	2
RNF13	*ZNF506*	0.910	3
IRF2	*ZNF765*	0.910	2
RNF13	*ZNF37A*	0.910	3
IRF2	*TUFT1*	0.910	2
IRF2	*PATL1*	0.910	2
IRF2	*LIPA*	0.910	2
IRF2	*EIF3D*	0.910	2
RNF13	*PTBP2*	0.909	3
IRF2	*TAF4*	0.909	2
IRF2	*CTB-36O1.7*	0.909	2
IRF2	*AMZ2*	0.909	2
RNF13	*NARS2*	0.908	3
RNF13	*PCGF6*	0.908	3
RNF13	*ATP1B1*	0.908	3
RNF13	*EEA1*	0.908	3
RNF13	*ARID4A*	0.907	3
RNF13	*NAA50*	0.907	3
RNF13	*SCP2*	0.906	3
RNF13	*PRIM1*	0.906	3
IRF2	*GTF2F1*	0.905	2
RNF13	*DNAJC15*	0.905	3
RNF13	*COIL*	0.905	3
RNF13	*JPH2*	0.905	3
RNF13	*GTF2B*	0.905	3
RNF13	*MOSPD1*	0.905	3
IRF2	*GOLGA7*	0.905	2
IRF2	*NOL12*	0.905	2
RNF13	*USP8*	0.904	3
RNF13	*LINC00476*	0.904	3
RNF13	*SAFB2*	0.904	3
RNF13	*KTN1*	0.904	3

RNF13	*MTO1*	0.904	3
IRF2	*SUV420H1*	0.904	2
RNF13	*CRBN*	0.904	3
IRF2	*RPRD1B*	0.903	2
RNF13	*HIF1AN*	0.903	3
RNF13	*NBN*	0.903	3
IRF2	*WBP1*	0.902	2
IRF2	*PPM1G*	0.902	2
IRF2	*KCMF1*	0.902	2
IRF2	*PFDN2*	0.902	2
RNF13	*CEP41*	0.902	3
RNF13	*MIS18BP1*	0.902	3
IRF2	*NUMB*	0.901	2
RNF13	*HOXB7*	0.901	3
IRF2	*NPRL3*	0.901	2
RNF13	*RP11-181C21.4*	0.901	3
RNF13	*RARS*	0.901	3
IRF2	*SCAF11*	0.901	2
RNF13	*GLO1*	0.901	3
RNF13	*SLC9B2*	0.901	3
IRF2	*UBL7*	0.901	2
IRF2	*PAF1*	0.901	2
RNF13	*RBBP6*	0.900	3
RNF13	*NOL8*	0.900	3
RNF13	*ZMAT1*	0.900	3
RNF13	*GS1-251I9.4*	0.900	3
RNF13	*FAM45A*	0.900	3
RNF13	*MTND5P4*	0.900	3
PBX2	*PPP1R11*	0.900	1
RNF13	*ITFG1*	0.900	3
RNF13	*YWHAQ*	0.900	3
RNF13	*CMTM6*	0.900	3
RNF13	*ZNF626*	0.900	3
RNF13	*CEP250*	0.900	3
RNF13	*KIAA1731*	0.900	3
RNF13	*PRPF4B*	0.900	3
RNF13	*ESCO1*	0.900	3
RNF13	*RP11-417J8.6*	0.900	3
RNF13	*RNF103*	0.900	3
RNF13	*MTND5P16*	0.900	3
RNF13	*AP3D1*	0.899	3
RNF13	*BTBD7*	0.899	3
RNF13	*ZNF655*	0.899	3
RNF13	*ZNF273*	0.899	3
RNF13	*CASP3*	0.899	3
RNF13	*NUP35*	0.899	3
RNF13	*ZNF814*	0.899	3
RNF13	*HYI*	0.899	3

RNF13	*ABCB7*	0.899	3
RNF13	*ZNF329*	0.899	3
RNF13	*TFB2M*	0.899	3
RNF13	*PPP6C*	0.899	3
RNF13	*CEP128*	0.899	3
RNF13	*UQCRC2*	0.899	3
IRF2	*AAMP*	0.898	2
RNF13	*MFAP3*	0.898	3
RNF13	*BRK1*	0.898	3
RNF13	*MAPRE1*	0.898	3
RNF13	*CNOT7*	0.898	3
IRF2	*HSP90B1*	0.898	2
RNF13	*ACBD6*	0.898	3
IRF2	*GPATCH4*	0.898	2
RNF13	*AIFM1*	0.898	3
IRF2	*EIF3A*	0.898	2
RNF13	*ZDHHC21*	0.898	3
RNF13	*GPR107*	0.897	3
IRF2	*TK2*	0.897	2
IRF2	*CREBBP*	0.896	2
RNF13	*DDX59*	0.896	3
RNF13	*PMPCB*	0.896	3
RNF13	*FAM53B*	0.896	3
RNF13	*TM9SF2*	0.896	3
IRF2	*CBL*	0.896	2
RNF13	*PCNP*	0.895	3
RNF13	*RP11-225B17.2*	0.895	3
RNF13	*RBM7*	0.895	3
IRF2	*UBQLN2*	0.895	2
IRF2	*TMEM186*	0.895	2
IRF2	*TTC4*	0.895	2
RNF13	*BRWD1*	0.894	3
RNF13	*RB1CC1*	0.894	3
RNF13	*SLC39A10*	0.894	3
RNF13	*TSN*	0.894	3
RNF13	*MTND1P27*	0.893	3
RNF13	*UBB*	0.893	3
RNF13	*PKP4*	0.893	3
RNF13	*KCTD6*	0.893	3
IRF2	*ANGEL1*	0.893	2
RNF13	*IWS1*	0.892	3
RNF13	*GOLGA7*	0.892	3
RNF13	*LSM14A*	0.892	3
RNF13	*KIAA0368*	0.892	3
RNF13	*ASCC3*	0.892	3
RNF13	*THOC2*	0.891	3
RNF13	*EGLN1*	0.891	3
IRF2	*MECP2*	0.891	2

IRF2	*PSMA5*	0.891	2
RNF13	*NEGR1*	0.891	3
RNF13	*CA5BP1*	0.891	3
RNF13	*GBE1*	0.890	3
RNF13	*TRIP11*	0.890	3
IRF2	*PRPS1*	0.890	2
IRF2	*DPF2*	0.890	2
RNF13	*ITGB1P1*	0.890	3
RNF13	*RSL1D1*	0.890	3
IRF2	*MCM3*	0.890	2
RNF13	*SBNO1*	0.890	3
RNF13	*UBQLN1*	0.890	3
IRF2	*TBX3*	0.890	2
RNF13	*SLC25A40*	0.890	3
RNF13	*NPAT*	0.889	3
RNF13	*DDHD2*	0.889	3
IRF2	*RNF38*	0.889	2
RNF13	*AIF1*	0.889	3
IRF2	*CLASP2*	0.889	2
RNF13	*PARN*	0.889	3
RNF13	*ZNF507*	0.889	3
IRF2	*TNFSF12*	0.889	2
RNF13	*NPM1P27*	0.889	3
RNF13	*MTND1P20*	0.889	3
RNF13	*RBM25*	0.889	3
RNF13	*ZNF107*	0.888	3
RNF13	*C16orf72*	0.888	3
RNF13	*CCDC171*	0.888	3
IRF2	*DNAJC17*	0.888	2
RNF13	*RP11-97I14.1*	0.888	3
IRF2	*RANBP10*	0.887	2
RNF13	*CKAP5*	0.887	3
IRF2	*PIGV*	0.887	2
IRF2	*VPS33B*	0.887	2
IRF2	*PHYH*	0.887	2
RNF13	*RMND5A*	0.886	3
RNF13	*SYPL1*	0.886	3
RNF13	*CDV3*	0.886	3
IRF2	*FUCA2*	0.886	2
IRF2	*PHKA2*	0.886	2
IRF2	*PROSER1*	0.886	2
IRF2	*DDX5*	0.885	2
RNF13	*PTS*	0.885	3
IRF2	*NAT6*	0.885	2
IRF2	*PES1*	0.885	2
RNF13	*CCNH*	0.885	3
RNF13	*HSPA9*	0.885	3
RNF13	*ESF1*	0.885	3

RNF13	*ZBTB7A*	0.885	3
RNF13	*DHX8*	0.885	3
IRF2	*CARS*	0.884	2
RNF13	*UBE2B*	0.884	3
RNF13	*ERH*	0.883	3
RNF13	*SHQ1*	0.883	3
RNF13	*HAUS2*	0.883	3
IRF2	*SIK3*	0.883	2
IRF2	*MED14*	0.883	2
IRF2	*CCDC137*	0.883	2
IRF2	*SRRT*	0.883	2
RNF13	*PSMD14*	0.883	3
RNF13	*AHI1*	0.882	3
IRF2	*HENMT1*	0.882	2
RNF13	*HBS1L*	0.882	3
RNF13	*GNAS*	0.882	3
RNF13	*RAB11FIP2*	0.882	3
RNF13	*RP5-874C20.3*	0.882	3
IRF2	*IKBIP*	0.882	2
IRF2	*FCHO2*	0.881	2
RNF13	*SSR1*	0.881	3
IRF2	*DCAKD*	0.881	2
RNF13	*ETF1*	0.881	3
RNF13	*TENM1*	0.880	3
RNF13	*NDUFS3*	0.880	3
IRF2	*C3orf18*	0.880	2
IRF2	*RPGRIP1L*	0.879	2
RNF13	*FAM134A*	0.879	3
RNF13	*NUBP1*	0.879	3
RNF13	*METTL13*	0.878	3
RNF13	*DMXL1*	0.878	3
IRF2	*MAPKAPK5*	0.878	2
IRF2	*CLN3*	0.877	2
IRF2	*CREG1*	0.877	2
RNF13	*NOL9*	0.877	3
IRF2	*ARNTL2*	0.877	2
RNF13	*BET1*	0.877	3
RNF13	*TFCP2*	0.877	3
IRF2	*SEMA6C*	0.876	2
RNF13	*MYO9A*	0.876	3
IRF2	*B3GNT2*	0.875	2
IRF2	*PPP1R26*	0.875	2
RNF13	*PPP1R12A*	0.875	3
RNF13	*DNAJC2*	0.875	3
RNF13	*HNRNPA3P6*	0.874	3
IRF2	*ZNF480*	0.874	2
IRF2	*ZNF594*	0.874	2
IRF2	*RNASEH2C*	0.874	2

PBX2	*PPT2*	0.874	1
PBX2	*SYNGAP1*	0.874	1
PBX2	*TBL2*	0.873	1
PBX2	*VARS*	0.873	1
PBX2	*MANSC1*	0.873	1
PBX2	*RBM42*	0.873	1
PBX2	*LDB2*	0.873	1
PBX2	*CLASP2*	0.873	1
PBX2	*HGSNAT*	0.873	1
IRF2	*TMEM143*	0.873	2
PBX2	*EPN2*	0.873	1
PBX2	*CTC-503J8.6*	0.873	1
PBX2	*LEPRE1*	0.873	1
PBX2	*SLC38A9*	0.873	1
PBX2	*TNKS2*	0.873	1
PBX2	*MCM3AP*	0.873	1
PBX2	*DNAJC1*	0.873	1
PBX2	*MBD6*	0.873	1
PBX2	*RAB1B*	0.873	1
PBX2	*CTDP1*	0.873	1
PBX2	*SH2B3*	0.873	1
PBX2	*ERCC5*	0.873	1
PBX2	*RPL10*	0.873	1
PBX2	*DCTN1*	0.873	1
PBX2	*ZNF76*	0.873	1
PBX2	*TRA2A*	0.873	1
PBX2	*MED14*	0.873	1
PBX2	*IFITM1*	0.873	1
PBX2	*HNRNPA0*	0.873	1
PBX2	*GUCY1B3*	0.873	1
PBX2	*PSMB9*	0.873	1
RNF13	*RNF146*	0.873	3
PBX2	*CCDC86*	0.873	1
PBX2	*HOXD-AS1*	0.873	1
PBX2	*PIAS1*	0.873	1
PBX2	*PTEN*	0.873	1
PBX2	*HSP90B1*	0.873	1
PBX2	*ANXA11*	0.873	1
PBX2	*SLC35F6*	0.873	1
PBX2	*SEC22B*	0.873	1
PBX2	*DDX19B*	0.873	1
PBX2	*FUT11*	0.873	1
PBX2	*CPNE8*	0.873	1
PBX2	*C12orf4*	0.873	1
RNF13	*AF186192.5*	0.873	3
PBX2	*KIAA1377*	0.873	1
PBX2	*FAM134B*	0.873	1
PBX2	*CAMK1*	0.873	1

PBX2	*RP11-804A23.4*	0.873	1
PBX2	*SPNS2*	0.873	1
PBX2	*ZNF101*	0.873	1
PBX2	*ATP5G2*	0.873	1
PBX2	*DENND1A*	0.873	1
PBX2	*SUV39H1*	0.873	1
PBX2	*IPO4*	0.873	1
PBX2	*DAZAP2*	0.873	1
PBX2	*RBPMS*	0.873	1
PBX2	*SUMO2*	0.873	1
RNF13	*CMC1*	0.873	3
PBX2	*MKL2*	0.873	1
PBX2	*ERF*	0.872	1
PBX2	*BACE1*	0.872	1
PBX2	*ARHGEF12*	0.872	1
PBX2	*EDEM2*	0.872	1
PBX2	*HINT2*	0.872	1
PBX2	*SELL*	0.872	1
IRF2	*ARID1A*	0.872	2
PBX2	*TMEM245*	0.872	1
PBX2	*SF3B5*	0.872	1
RNF13	*RABGGTB*	0.872	3
PBX2	*NOL10*	0.872	1
PBX2	*SUPT20H*	0.872	1
PBX2	*SEL1L*	0.872	1
PBX2	*C2orf16*	0.872	1
PBX2	*UBE2E1*	0.872	1
PBX2	*COL1A2*	0.872	1
PBX2	*KIAA1671*	0.872	1
PBX2	*YARS*	0.872	1
PBX2	*CFDP1*	0.872	1
PBX2	*ARRDC1*	0.872	1
PBX2	*NEK11*	0.872	1
PBX2	*ATF2*	0.872	1
PBX2	*RAB8B*	0.872	1
PBX2	*AARS*	0.872	1
PBX2	*AIMP1*	0.872	1
PBX2	*TNFRSF14*	0.872	1
PBX2	*TARS*	0.872	1
PBX2	*TMCO4*	0.872	1
PBX2	*SLC23A2*	0.872	1
PBX2	*VASH1*	0.872	1
PBX2	*WBP5*	0.872	1
PBX2	*PDCL*	0.872	1
PBX2	*PRNP*	0.872	1
PBX2	*CNEP1R1*	0.872	1
PBX2	*NANS*	0.872	1
IRF2	*GABARAP*	0.872	2

PBX2	*KLHL15*	0.872	1
PBX2	*PSENEN*	0.872	1
PBX2	*MTHFD1L*	0.872	1
PBX2	*ZKSCAN8*	0.872	1
PBX2	*PRPF6*	0.872	1
PBX2	*ID2*	0.872	1
PBX2	*TMEM138*	0.872	1
PBX2	*PES1*	0.872	1
PBX2	*EBNA1BP2*	0.872	1
PBX2	*HM13*	0.872	1
PBX2	*HNRNPL*	0.872	1
IRF2	*DBP*	0.872	2
PBX2	*GOLGA7*	0.872	1
PBX2	*NOTCH4*	0.872	1
PBX2	*MYCT1*	0.872	1
PBX2	*EIF1AXP1*	0.871	1
PBX2	*ACP2*	0.871	1
PBX2	*TUBE1*	0.871	1
PBX2	*CSNK1E*	0.871	1
PBX2	*PCCB*	0.871	1
PBX2	*HPCAL1*	0.871	1
PBX2	*ACADVL*	0.871	1
PBX2	*RBM14*	0.871	1
PBX2	*CD59*	0.871	1
PBX2	*ABCA3*	0.871	1
PBX2	*SIL1*	0.871	1
IRF2	*C5orf22*	0.871	2
PBX2	*DDB1*	0.871	1
PBX2	*ISCA1*	0.871	1
PBX2	*ZBTB5*	0.871	1
PBX2	*RPS7P1*	0.871	1
PBX2	*MAGI2-AS3*	0.871	1
PBX2	*LETM1*	0.871	1
PBX2	*TMEM101*	0.871	1
PBX2	*RNF40*	0.871	1
PBX2	*POLR2F*	0.871	1
PBX2	*CALR*	0.871	1
RNF13	*LIN9*	0.871	3
IRF2	*SND1*	0.871	2
PBX2	*TMEM50A*	0.871	1
PBX2	*VMP1*	0.871	1
PBX2	*ZHX1*	0.870	1
PBX2	*MRPS5*	0.870	1
PBX2	*RBFA*	0.870	1
PBX2	*SNX7*	0.870	1
PBX2	*RP11-750B16.1*	0.870	1
PBX2	*CENPO*	0.870	1
PBX2	*DAD1*	0.870	1

PBX2	*RPL10AP6*	0.870	1
PBX2	*CDC42SE1*	0.870	1

Table 3.13: All pairwise interactions from Findr (P2*P5) in subcutaneous fat at a 10% FDR threshold.

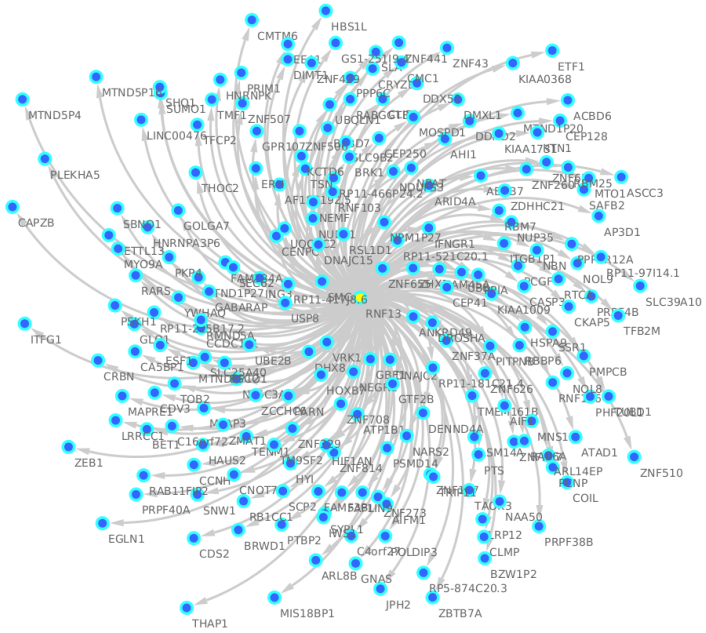


Figure 3.7: Subcutaneous fat network network reconstructed from STARNET-subcutaneous fat at 10% FDR and filtered to only include interactions for *RNF13*. Arrows indicate direction of regulation.

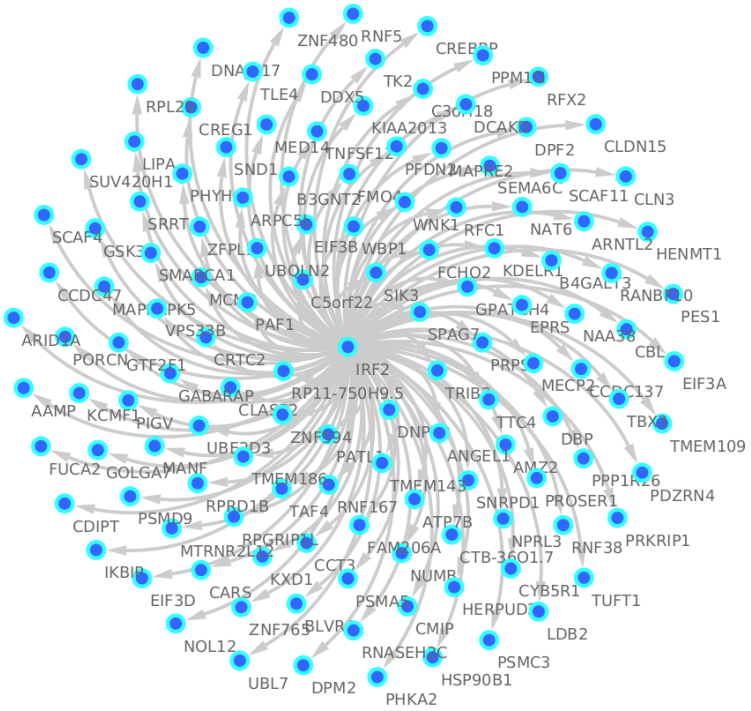


Figure 3.8: Subcutaneous fat network network reconstructed from STARNET-subcutaneous fat at 10% FDR and filtered to only include interactions for *IRF2*. Arrows indicate direction of regulation.

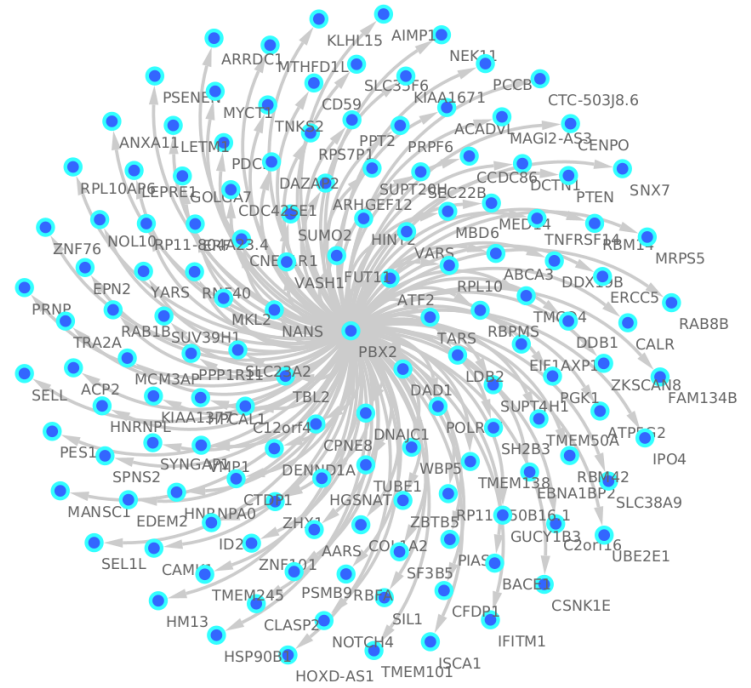


Figure 3.9: Subcutaneous fat network network reconstructed from STARNET-subcutaneous fat at 10% FDR and filtered to only include interactions for *PBX2*. Arrows indicate direction of regulation.

A-genes	B-genes	Findr score	GR count
LUC7L3	*TMEM123*	0.994	2
CD163	*NDUFB9*	0.977	1
LUC7L3	*CCNI*	0.950	2
LUC7L3	*KHDRBS1*	0.933	2
LUC7L3	*KALRN*	0.915	2
LUC7L3	*MRPS30*	0.915	2
LUC7L3	*RB1*	0.913	2
LUC7L3	*DLGAP4*	0.912	2
CD163	*CYB5R1*	0.904	1
LUC7L3	*ZNF639*	0.901	2
LUC7L3	*IMPAD1*	0.886	2
LUC7L3	*ARMC8*	0.876	2
LUC7L3	*CUL4B*	0.872	2
CD163	*DESI2*	0.861	1
CD163	*TMEM98*	0.861	1

Table 3.14: All pairwise interactions from Findr (P2*P5) in visceral abdominal fat at a 10% FDR threshold.

Tissue	Gene name	Ensembl gene ID	Primary instrument	Primary Findr score	Independent instrument	Independent Findr score
liver	*ALG5*	ENSG00000120697	rs9576151	1.00	rs731877	0.97
liver	*NUCB2*	ENSG00000070081	rs214080	1.00	rs199782153	1.00
liver	*FOXN2*	ENSG00000170802	rs79073127	1.00	rs7566996	1.00
liver	*YBX3*	ENSG00000060138	rs11053915	1.00	rs35166079	0.98
liver	*PTPN12*	ENSG00000127947	rs7783866	1.00	rs56111978	1.00
liver	*SERPINA6*	ENSG00000170099	rs2736898	1.00	rs7161231	0.94
liver	*TOR4A*	ENSG00000198113	rs28567631	1.00	rs12004799	0.61
liver	*RABEP1*	ENSG00000029725	rs56176579	0.99	rs199634929	0.78
liver	*RDX*	ENSG00000137710	rs7107823	0.97	rs55969611	0.78
liver	*MAPK11*	ENSG00000185386	rs742186	0.96	rs4838867	0.95
liver	*CPEB2*	ENSG00000137449	rs62410848	0.90	rs6847363	0.48
liver	*DNM1L*	ENSG00000087470	rs11052028	0.86	rs4931017	0.75
liver	*SLC26A1*	ENSG00000145217	rs3733346	0.84	rs33965806	0.61
liver	*PCSK5*	ENSG00000099139	rs11145221	0.74	rs5898407	0.56
liver	*ZNF649*	ENSG00000198093	rs6509633	0.73	rs34110879	0.72
liver	*NAB2*	ENSG00000166886	rs34577247	0.61	rs4759254	0.55
liver	*EIF3H*	ENSG00000147677	rs1569365	0.44	rs11775371	0.43
subcutaneous fat	*OSMR*	ENSG00000145623	rs13165709	1.00	rs11949864	1.00
subcutaneous fat	*AUTS2*	ENSG00000158321	rs2141205	1.00	rs36018096	0.93
subcutaneous fat	*PDZD8*	ENSG00000165650	rs149832558	1.00	rs363237	1.00
subcutaneous fat	*KHK*	ENSG00000138030	rs7560144	1.00	rs12714026	1.00
subcutaneous fat	*ATP5J2*	ENSG00000241468	rs138229375	1.00	rs74774557	1.00
subcutaneous fat	*PPCDC*	ENSG00000138621	rs3812943	1.00	rs8042558	1.00
subcutaneous fat	*PGM1*	ENSG00000079739	rs139945547	1.00	rs2269267	1.00
subcutaneous fat	*PHYH*	ENSG00000107537	rs6602646	1.00	rs2484528	1.00
subcutaneous fat	*AMPD3*	ENSG00000133805	rs11042759	1.00	rs2957658	1.00
subcutaneous fat	*PKP2*	ENSG00000057294	rs12825217	1.00	rs11052277	0.99
subcutaneous fat	*STAT4*	ENSG00000138378	rs4341966	1.00	rs11691372	1.00
subcutaneous fat	*ATG13*	ENSG00000175224	rs61882678	0.99	rs112018914	0.63
subcutaneous fat	*PLD1*	ENSG00000075651	rs10936700	0.97	rs13068741	0.74
subcutaneous fat	*IRF2*	ENSG00000168310	rs34985265	0.94	rs2171838	0.72
subcutaneous fat	*PBX2*	ENSG00000204304	6:32853219	0.93	rs3128947	0.73
subcutaneous fat	*XPNPEP1*	ENSG00000108039	rs3780953	0.90	rs143664187	0.71
subcutaneous fat	*RNF13*	ENSG00000082996	rs9853321	0.81	rs62282739	0.70
subcutaneous fat	*ZC3H7B*	ENSG00000100403	rs9611739	0.77	rs2267429	0.71
subcutaneous fat	*KLHDC1*	ENSG00000197776	rs61984263	0.71	rs2883893	0.70
subcutaneous fat	*ME2*	ENSG00000082212	rs8092419	0.71	rs2586777	0.59
subcutaneous fat	*ENSA*	ENSG00000143420	rs111906083	0.70	rs191222557	0.43
subcutaneous fat	*MAMDC2*	ENSG00000165072	rs78116045	0.60	rs149208553	0.56
subcutaneous fat	*HSPG2*	ENSG00000142798	rs76643224	0.58	rs7513223	0.55
subcutaneous fat	*LHFPL2*	ENSG00000145685	rs74347692	0.58	rs4704459	0.56
subcutaneous fat	*PDIA5*	ENSG00000065485	rs72285796	0.49	rs73186456	0.42
subcutaneous fat	*FOS*	ENSG00000170345	rs11627282	0.44	rs2240624	0.39
visceral abdominal fat	*NNT*	ENSG00000112992	rs6451720	1.00	rs11951515	1.00
visceral abdominal fat	*DTNA*	ENSG00000134769	rs71363449	1.00	rs9950794	0.77
visceral abdominal fat	*ULK2*	ENSG00000083290	rs79506397	0.88	rs1634418	0.47
visceral abdominal fat	*CD163*	ENSG00000177575	rs73059776	0.86	rs2377237	0.72
visceral abdominal fat	*LUC7L3*	ENSG00000108848	rs6504682	0.80	rs2412130	0.70
visceral abdominal fat	*DENR*	ENSG00000139726	rs73230017	0.76	rs201556706	0.47
visceral abdominal fat	*ATL1*	ENSG00000198513	rs61543335	0.76	rs11849026	0.59
visceral abdominal fat	*BNC2*	ENSG00000173068	rs10810646	0.75	rs4961707	0.69
visceral abdominal fat	*REEP5*	ENSG00000129625	rs79334785	0.74	rs153562	0.73
visceral abdominal fat	*CXCL14*	ENSG00000145824	rs72802346	0.74	rs62366015	0.56
visceral abdominal fat	*ANKFY1*	ENSG00000185722	rs9891529	0.70	rs9912501	0.68
visceral abdominal fat	*GNG2*	ENSG00000186469	rs59825463	0.68	rs17124893	0.54
visceral abdominal fat	*SLC27A2*	ENSG00000140284	rs17509944	0.68	rs934633	0.54
visceral abdominal fat	*LPIN1*	ENSG00000134324	rs79911731	0.64	rs11365598	0.64
visceral abdominal fat	*CSRNP3*	ENSG00000178662	rs61553904	0.62	rs10203713	0.61
visceral abdominal fat	*UBE3A*	ENSG00000114062	rs112605074	0.59	rs185740871	0.57
visceral abdominal fat	*DLG1*	ENSG00000075711	rs9325375	0.59	rs9820797	0.53
visceral abdominal fat	*WAPAL*	ENSG00000062650	rs72404633	0.57	rs7086163	0.50
visceral abdominal fat	*BCL7B*	ENSG00000106635	rs73134935	0.50	rs78703841	0.49
visceral abdominal fat	*CAV2*	ENSG00000105971	rs201679967	0.43	rs62476996	0.43

Table 3.15: All primary and independent instruments identified for GR regulated trans-genes (FDR = 15%) in liver, subcutaneous fat and visceral abdominal fat.

Chapter 4

Replication of cortisol associated trans-genes and networks

4.1 Introduction

4.1.1 Replication of cortisol associated trans-genes

In Chapter 2, we described the identification of genetic variants associated with plasma cortisol and how these can be linked to tissue specific gene expression within the STARNET dataset. STARNET was selected for studying the impact of genetic variation for plasma cortisol, as it contains a sufficiently high number of individuals to detect small effect sizes commonly associated with complex traits[223]. Furthermore, it presented an opportunity to examine the role of cortisol linked gene expression in different tissues, taken from the same individuals and sampled at the same time. The STARNET cohort is also composed of individuals who have been diagnosed with CAD, making this an appropriate cohort for examining how genetic variation for cortisol may impact CVD linked phenotypes.

In this chapter we aimed to see if the trans-associations identified in STARNET were robust and were represented in independent cohorts, outside of STARNET.

Trans-gene associations tend to be weaker than their cis counterparts, however have been predicted to contribute to high levels of complex trait heritability[223]. Cis-eQTLs can often be mapped to cis-regulatory elements, which can be involved in direct regulation of the associated gene[54]. However, the biology underpinning trans-eQTL associations is often more challenging to uncover and can involve indirect mechanisms of regulation, sometimes acting in conjunction with cis-eQTLs[55].

Trans-associations are often challenging to replicate in independent datasets, whereby SNPs can be associated with multiple transcripts, either alluding to intermediate genes in a causal pathway or as the result of an increased false positive rate[224]. Pierce and colleagues sought to describe the interplay between different cis and trans action. Following the identification of 21 trans-associations that were shown to be mediated by a cis-eQTL, the researchers were able to replicate 7 of these associations in an independent cohort[225]. Although trans-eQTLs have been linked to heritability for complex traits[223], Yap *et al* report a lack of enrichment of trans-eQTLs and GWAS associated SNPs in blood[226], highlighting the difficulty in parsing functional variants from large sets of trans-associations.

Results from chapter 2 have shown trans-associations to be highly tissue specific. The most high profile example of the role of tissue specificity and gene expression can be seen in the latest release from the GTEx consortium[66]. This dataset has characterised the gene expression landscape across 49 tissues from 838 post-mortem donors and shows that tissue specificity and cell composition are key factors in linking genetic variation to mechanism[66]. Therefore, it is important when attempting to replicate trans-genes that replication is attempted in a comparable tissue set.

The first replication cohort considered for this project, was from the Metabolic Syndrome in Men study (METSIM), a population based study examining Finnish men that was conducted between 2005-2010[227]. Within this study, there are 982 individuals who underwent genotyping and whole exome sequencing and there are expression data for 434 individuals in this cohort obtained from subcutaneous fat

biopsies. The METSIM authors initially aimed to use the study to identify factors associated with the development of type II diabetes and CVD and since then METSIM has been used to identify inflammatory markers associated with insulin secretion and sensitivity[228].

METSIM was identified as an avenue for replicating cortisol associated trans-genes that were identified in STARNET-subcutaneous fat and access to the dataset was obtained through the Database of Genotypes and Phenotypes (dbGaP). The availability of individuals genotypes meant that it was possible to link cortisol associated SNPs to gene transcripts in subcutaneous fat. Furthermore, where cortisol linked SNPs are unavailable it is possible to compare patterns of gene expression in METSIM to those in STARNET.

A second replication cohort was considered from a recent study by the BIOCORT consortium[229], using a perturbation based approach to examine variation in gene expression in response to glucocorticoid treatment. This was a randomised, crossover, doubleblind study of 10 individuals with primary adrenal insufficiency. Individuals in the study were randomised to either intravenous saline or hydrocortisone, and microarray analysis was used to measure gene expression in both peripheral blood mononuclear cells (PBMCs) and abdominal subcutaneous adipose tissue.

Both METSIM and BIOCORT represent an approach to validate genes associated with plasma cortisol within STARNET tissues. By comparing genes that are differentially expressed in response to cortisol treatment, to genes that are associated with genetic variation for plasma cortisol this is a useful approach to identify genes where there is strong evidence for a cortisol response.

4.1.2 Replication of glucocorticoid regulated gene networks

Having first identified genes that are associated with plasma cortisol, in Chapter 3 we describe a subset of these trans-genes that are both *regulated* by glucocorti-

coids and in turn *regulate* a network of target genes in the same tissue where they are associated with cortisol. As with the trans-gene associations, tissue specificity is important when making a comparison of gene networks in an independent dataset. In addition, target specificity is also important to see if the same regulators are targeting the same genes when using different data.

An example of targeted replication of gene networks includes work by Small *et al*, in follow-up experiments of a type II diabetes associated gene network in adipose tissue. Having identified a key cis-eQTL for *KLF14*, the researchers then followed up by examining the role of a murine *KLF14* knockout model and replicating some of the network target genes in murine adipose[105].

Talukdar and colleagues also used data from an animal model to validate gene regulatory networks identified from the STAGE dataset. Using data from the hybrid mouse diversity panel, consisting of different strains of mice with broad phenotypic variation. the researchers sought to replicate 26 CAD related gene networks by examining how well the networks segregated among mice with CAD related traits. Overall the researchers were able to show that 12 of their 26 networks segregated according to phenotype[88].

Cohain and colleagues, examined the reproducibility of Bayesian networks using both STARNET and whole blood data from GTEx. The researchers found that individual network edges are challenging to replicate and are highly dependent on sample size, however that the key drivers (regulators) of these networks are robust across different datasets and sample sizes[230]. Therefore, it can be taken that although there may be variability in the network targets, the regulators of these networks are conserved across datasets.

These examples highlight a diversity of different approaches that can be used to replicate gene networks. Replication of gene networks is challenging, as the networks themselves describe subtle changes in gene expression over multiple genes, which respond in a coordinated way to an associated exposure. Therefore, although it is unlikely that all network targets will be replicated, the broad changes in gene

expression should be consistent between datasets and able to be captured.

4.1.3 Chapter objectives

1. To provide evidence of the presence of cortisol associated trans-genes identified from STARNET in an independent dataset.

2. To provide evidence of glucocorticoid regulated gene networks identified from STARNET in an independent dataset.

4.2 Materials and methods

4.2.1 Datasets

Three major datasets were used over the course of this chapter as described in Table 4.1. Gene expression data for METSIM is available publicly at GEO (assession no GSE70353) and METSIM genotype data was accessed through an application to dbGaP (assession no. phs000743.v1.p1).

Dataset	Full name	Number of participants	Study type	Number of SNPs
STAGE	Stockholm Atherosclerosis Gene Expression study	114	Genotype and Microarray data	909,622
METSIM	Metabolic Syndrome in Man study	982	Whole exome sequencing and RNA-seq	444,342
BIOCORT	Chantzichristos *et al*, 2021	10	Randomised crossover trail	N/A

Table 4.1: Dataset used for replication of cortisol associated trans-genes and networks

4.2.2 trans-gene replication

Processing of METSIM genotypes

METSIM genotypes were obtained through an application to dbGaP. Genotypes were obtained using an Exome Chip assay (Illumina). VCF files were processed using vcftools (version 0.1.13) to convert genotypes to -012 format, as a flat text file. Genotype files were then filtered to align with individuals where gene expression data is also available. A further filtering step was used to identify SNPs that were also present in the CORNET GWAMA.

Processing of METSIM gene expression data

Subcutaneous fat biopsies underwent RNA isolation and gene expression levels were measured as described in Civelek *et al, 2017*[231]. To summarise; expression profiling was performed using the Affymetrix U219 microarray. Microarray processing was carried out using Affymetrix GCOS algorithm via the robust multiarray average (RMA) method. PCA was conducted on downloaded subcutaneous fat gene expression, across sample (Figure S4.10). Ensembl Biomart (GRCh37) was used to label transcripts (provided as Ensembl IDs) with; gene name, chromosome location, gene start and gene end.

Processing of STAGE gene expression data

Gene expression analysis was carried out as described in Foroughi *et al*, 2015[160], using a custom Affymetrix array (HuRSTA-2a520709). Normalisation was carried out using the RMA method and a custom annotation file was used to match probes to 19,610 probe sets for unique genes using the hg19 human genome assembly. PCA was conducted on all STAGE tissues, across sample (Figure S4.9). Ensembl Biomart (GRCh37) was used to label transcripts (provided as Ensembl IDs) with; gene name, chromosome location, gene start and gene end.

Differential gene expression in BIOCORT

Raw Affymetrix microarray files (HuGene-2_0-st) were obtained from BIOCORT authors[229]. Cel files were pre-processed using the oligo package[232] in R and expression was normalised using the gcrma method. Following this, differential expression for treated vs untreated samples was obtained using limma[233] with the application of empirical Bayes smoothing to the standard errors. All steps were carried out in R.

Trans-gene discovery

Association between SNPs and trans-genes were obtained using the secondary linkage test in Findr[92] (P2). This was carried out as described in Chapter 2.2.3, by testing all SNPs within the prescribed window against all genes in the corresponding dataset.

4.2.3 Gene network replication

Correlations between network targets

Correlations between between gene network targets were calculated using gene expression data from STARNET, STAGE and METSIM. Gene expression matrices were filtered to only include the target genes under investigation and the *.corr* function in Pandas was used to construct correlation matrices of corresponding Pearson correlation coefficients as an absolute value.

A background gene-set was constructed from the overlapping genes between the STARNET gene expression set that was used for network discovery and the corresponding gene expression set that was being used for replication. The previously described correlation analysis, was then repeated using a random set of genes (the same size as the target set) selected from the background gene-set using the *.sample* function in Pandas. The Kruskal Wallis test was used in Scipy Stats to test if the targeted and randomly sampled correlations follow the same distribution. Both the targeted and random correlations were then plotted as both a distribution plot and a box plot using the Python plotting package Seaborn.

Regulator-target network correlations

Correlations were also calculated between predicted network regulators and the corresponding targets. Instead of constructing a matrix, this was performed iter-

atively to obtain the Pearson correlation coefficients between regulators and target expression profiles as absolute values.

Random gene sets the same size of the target gene set were constructed from a background set, and correlation coefficients were obtained between the network regulator and random sets, using the same iterative method. The Kruskal Wallis test, from Scipy Stats, was calculated using the random and target sets for each regulator and this was again presented as distribution plots and box plots.

4.3 Results

4.3.1 Replication of cortisol associated trans-genes

Cortisol associated SNPs not present within METSIM genotypes

Our initial approach to address the replication of STARNET trans-genes described in Chapter 2.3.3, was to repeat the trans-eQTL discovery with SNPs associated with plasma cortisol from the CORNET GWAMA ($p < 5 \times 10^{-8}$), using METSIM genotypes and gene expression data. Following genotype processing, METSIM SNPs were filtered to include those present at the CORNET peak on Chromosome 14 (± 100 Kb of *SERPINA6*). We then selected samples from individuals where where both genotype and expression data was present (n = 189).

METSIM genotypes were limited to exonic regions, therefore there were only 17 SNPs identified in the locus identified from the CORNET GWAMA which was mainly intronic. Of these SNPs, the SNP with the strongest association with plasma cortisol was rs2273399 ($p = 2.1 \times 10^{-5}$). No SNPs were identified that surpassed the threshold of genome wide significance. For this reason, no SNPs were taken forward for trans-eQTL analysis with METSIM gene expression data.

Cortisol associated genes in STAGE tissues

As STAGE contains many of the same tissues measured in STARNET, this presented as an alternate avenue to replicate cortisol linked trans-associations identified in STARNET. Trans-eQTL discovery was carried out in five STAGE tissues that were also present in STARNET, namely; liver, skeletal muscle, subcutaneous fat, visceral abdominal fat and whole blood. Following genotype processing, SNPs in STAGE were filtered for those present at the CORNET peak on chromosome 14 (± 100 Kb of *SERPINA6*), and both genotype and gene expression data was filtered to only include individuals where there was both genotype and gene expression data available (Ta-

ble 2.1).

72 SNPs were identified within a 100 Kb window surrounding *SERPINA6*, 8 of which were present in the CORNET GWAMA at genome wide significance ($p \leq 5\times10^{-8}$). Trans-eQTL analysis was carried out between all genes from the five STAGE tissue and all 72 SNPs at the *SERPINA6* loci, after which we selected for SNP-gene associations between genes and SNPs associated with plasma cortisol at genome wide significance at a 15% FDR threshold. A total of 44 SNP-gene associations between cortisol linked SNPs and genes in STAGE tissues were identified, with 21 unique genes across all tissues. Of these 21 genes; 20 were associated with cortisol SNPs in trans (Table 4.2). One gene, *SERPINA9*, was cis-associated with rs7161521 ($p = 3.07\times10^{-19}$) in blood (Findr P2 = 0.89). Cis-associations between cortisol SNPs and *SERPINA6* in STAGE are described in Chapter 2.3.1.

The strongest trans-associations were identified in liver and subcutaneous fat for the genes *HMSD* and *UBXN2B* respectively (Findr P2 = 0.97). *HMSD* is associated with rs11629171 (CORNET $p = 6.83\times10^{-14}$) but was not measured in STARNET liver so could not be compared. *UBXN2B* was associated with 4 cortisol linked SNPs, the strongest of which was rs1884548 (CORNET $p = 4.54\times10^{-10}$). This association was not retained in STARNET-subcutaneous fat (Findr P2 = 0).

The only gene to appear as cortisol linked trans-gene in both STARNET and in STAGE (FDR = 15%) was *OGG1*, which was expressed in skeletal muscle. *OGG1* is associated in STAGE with rs1884549 (CORNET $p = 4.94\times10^{-10}$) and STARNET with rs2749539 (CORNET $p = 3.04\times10^{-8}$), interestingly the two SNPs are in different LD block, as described in Chapter 2.3.2.

Cortisol associated trans-genes are differentially expressed in response to hydrocortisone treatment

The results from experiments by the BIOCORT consortium reveals genes which are differentially expressed in Addison's disease patients (n = 10) in response to hydro-

Tissue	Gene name	SNP	P2 Findr score	Number of associations
Blood	*OR5B2*	rs1884549	0.83	3
	C8orf22	rs1884549	0.91	5
Liver	*HADH*	rs4905188	0.86	2
	PSG2	rs11629171	0.92	1
	HMSD	rs11629171	0.97	1
	C21orf15	rs941599	0.76	2
Subcutaneous fat	*OR4D10*	rs11629171	0.91	1
	UBQLN3	rs11629171	0.90	1
	SLC37A3	rs11629171	0.90	1
	OR8K1	rs11629171	0.79	1
	UBXN2B	rs1884548	0.97	4
	ZKSCAN2	rs11629171	0.93	1
Skeletal Muscle	*TTC33*	rs941599	0.87	2
	PPP4R1	rs4905188	0.86	5
	FCER2	rs941595	0.96	2
	EIF2S1	rs941599	0.83	2
	HTR2A	rs11629171	0.81	1
	DUOX1	rs11629171	0.81	1
	*OGG1**	rs1884549	0.80	3
Visceral abdominal fat	*MYF6*	rs11629171	0.94	3

Table 4.2: Trans-associations between SNPs associated with plasma cortisol ($p \leq 5\times10^{-8}$) and genes in STAGE (FDR = 15%). SNP column indicates strongest eQTL and number of associations is specific to SNPs that are associated with plasma cortisol. *Indicates genes that are also trans-associated with cortisol linked SNPs in the same tissue in STARNET (FDR = 15%).

cortisone treatment. This has been obtained from gene expression data from subcutaneous fat and PBMCs. We compared the overlap between genes which were differentially expressed in BIOCORT ($p \leq 0.05$) to genes that were associated with plasma cortisol in subcutaneous fat (FDR = 15%) (Table 4.3).

Five unique genes were differentially expressed in BIOCORT subcutaneous fat, out of the 54 genes that were trans-associated with genetic variation for plasma cortisol (FDR = 15%). None of these genes were identified as major network regulators, however *STAT4* is shown to have three causal targets at a 15% FDR threshold. Additionally *BRD2* is a target of the *PBX2* sub-network at a 15% FDR cut off (Findr score

	Discovery		Replication		
Gene name	**SNP**	**P2 Findr score**	**p-value**	**Fold change**	**direction of regulation**
DSG2	rs3790035	0.871	0.024	1.235	+
DSG2	rs8015996	0.871	0.024	1.235	+
DSG2	rs11629171	0.871	0.024	1.235	+
PKP2	rs1243171	0.856	0.003	1.398	+
STAT4	rs941594	0.856	0.031	0.805	+
STAT4	rs2749527	0.779	0.031	0.805	+
BRD2	rs1243171	0.850	0.022	1.083	+
BRD2	rs59036614	0.736	0.022	1.083	+
MAMDC2	rs8022616	0.822	0.013	1.450	+

Table 4.3: BIOCORT - CORNET trans-gene crossover in subcutaneous fat. Table contains genes that are both differentially expressed in BIOCORT ($p \leq 0.05$) and trans-associated with SNPs associated with plasma cortisol (FDR = 15%).

= 0.85). Using Fisher's exact test there is no evidence of an enrichment of cortisol associated trans-genes within the genes differentially expressed in BIOCORT subcutaneous fat ($p = 0.801$). However, these findings do not take into account any differences in gene expression that may arise as a result of a cortisol free system.

4.3.2 Correlations in gene expression between network targets

Correlations of network targets in METSIM subcutaneous fat

The causal gene networks identified in Chapter 3, represent coordinated changes in gene expression in response to genetic perturbations. Therefore, it is possible to examine if these changes in gene expression are present in other datasets, from gene expression data alone. We used gene expression data from subcutaneous fat from METSIM to compare patterns in gene expression within causal networks predicted from STARNET subcutaneous fat. As METSIM only contains gene expression data for subcutaneous fat, analysis was restricted to the causal networks identified in STARNET subcutaneous fat.

Correlation matrices were constructed from the targets of the subcutaneous fat sub-network regulators; *IRF2*, *RNF13* and *PBX2* at both 10% and 15% FDR thresholds. This allowed for correlations between these network targets to be calculated

and for the distribution of these correlations to be examined (Figure 4.1). Absolute correlation distributions were compared to distributions of a random set of genes selected from METSIM subcutaneous fat, the same size as the corresponding target gene set. The difference between targeted and random distributions was formalised using the Kruskal Wallis test for each sub-network (Table 4.4).

Network regulator	FDR threshold	KW test statistic	p-value	No. Target Genes
IRF2	10% FDR	72.35	1.80E-17	128
IRF2	15% FDR	543.29	3.63E-120	247
RNF13	10% FDR	10125.33	< 1.0E-300	215
RNF13	15% FDR	33660.99	< 1.0E-300	416
PBX2	10% FDR	83.87	5.30E-20	138
PBX2	15% FDR	22620.51	< 1.0E-300	883

Table 4.4: Correlations between network targets in METSIM subcutaneous fat. Kruskal Wallis test calculated for distribution of correlations between network targets compared to correlations within random gene set of same size.

Target gene correlations show a distribution skewed towards higher correlation values compared to the random distribution at both 10% and 15% FDR network thresholds, suggesting that the expression of these genes is more highly correlated than what would be expected by chance. For each gene-set, the disparity between target and random sets is strongest at the 15% threshold which was also the same threshold where GR target enrichment was highest in *IRF2* targets (Chapter 3.3.3). The largest divergence between the targeted and random distribution is from the *RNF13* targets, in particular within the 15% network ($p < 1.0 \times 10^{-300}$). The differences between *IRF2* and *PBX2* targets are more modest, however both show a distribution skewed towards higher correlation values compared to the random distribution at both 10% and 15% FDR network thresholds.

As the gene networks had originally been identified using causal methods in STARNET rather than by co-expression, we also examined the co-expression of network targets in STARNET as well as METSIM (Table S4.8, Figure S4.11). Interestingly, the correlations within both targeted and random gene sets appear to be systematically weaker in STARNET when compared to METSIM. However, the disparity between random and targets correlations is more pronounced than correlations

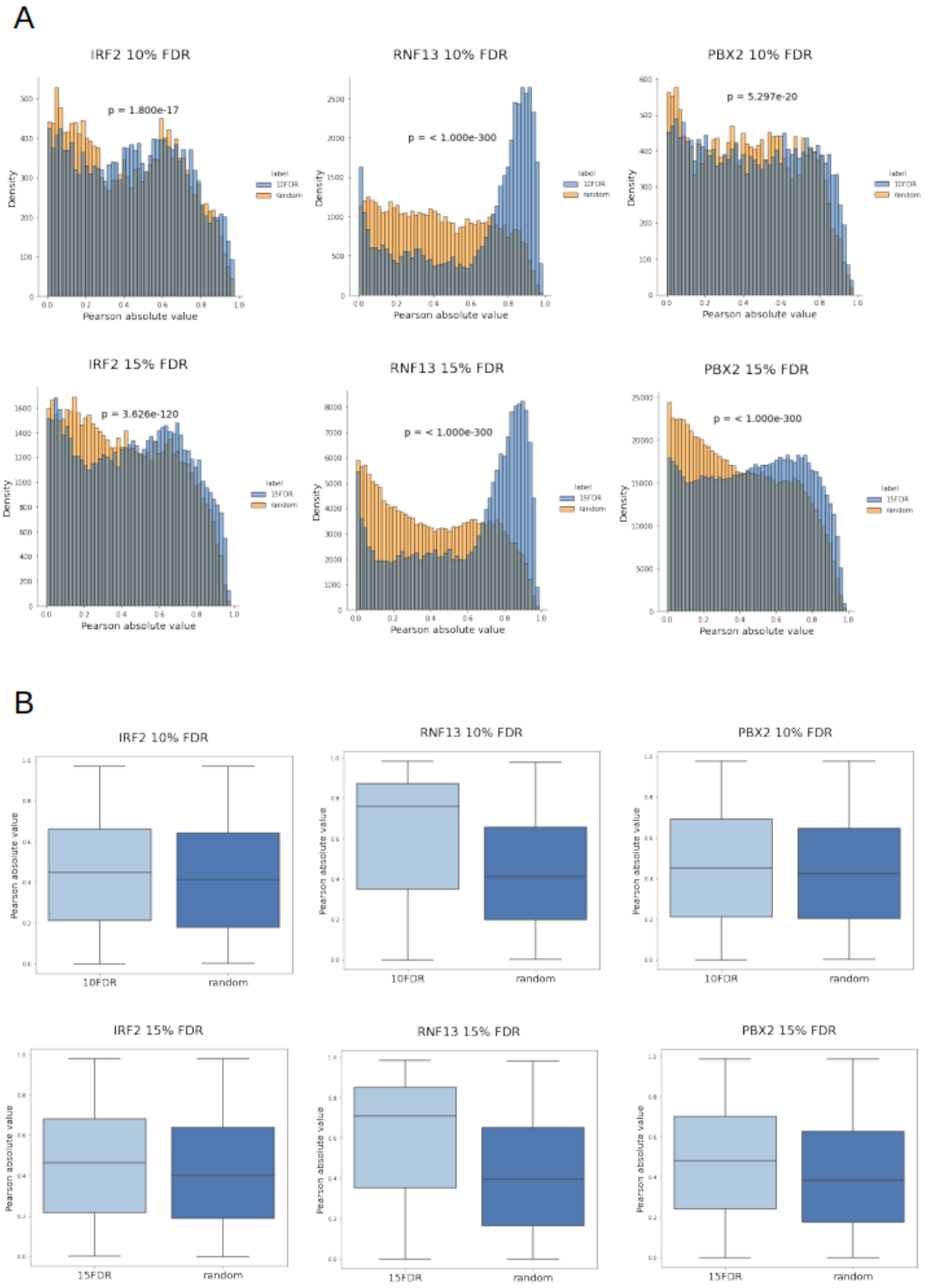


Figure 4.1: Correlations in gene expression between network targets in METSIM subcutaneous fat as absolute values. (A) Distribution plot of subcutaneous fat sub-network correlations compared to random gene set of same size (orange). Comparison between targeted and random distribution formalised by Kruskal Wallis p-value. (B) Boxplots comparing distribution between targets and random gene set.

within METSIM for all gene sets at both 10% and 15% FDR threshold ($p < 1.0 \times 10^{-300}$).

Correlations of network targets in STAGE liver and adipose

Outside of identifying trans-gene associations in STAGE, it is also possible to examine correlations in gene expression between network targets that have been predicted from STARNET. Although gene expression in STAGE represents a smaller sample size than METSIM, it has the benefit of the inclusion of a broader range of tissues, including all 3 STARNET tissues where gene networks were identified (Table 2.1).

We calculated correlations in gene expression from STAGE for network targets from liver, subcutaneous fat and visceral abdominal fat at both 10% and 15% FDR thresholds. As with METSIM, we then calculated correlations in gene expression for random gene-sets of the same size as the network targets, from a tissue specific background constructed from overlap between STAGE and STARNET genes. We then calculated the Kruskal Wallis test to formalise the difference between the random and target distributions (Table 4.5).

Tissue	Network regulator	FDR threshold	KW test statistic	p-value	No. Target Genes
Subcutaneous fat	*IRF2*	10% FDR	435.48	1.04E-96	128
	IRF2	15% FDR	2094.16	< 1.0E-300	247
	RNF13	10% FDR	1397.30	< 1.0E-300	215
	RNF13	15% FDR	5946.61	< 1.0E-300	416
	PBX2	10% FDR	235.21	4.34E-53	138
	PBX2	15% FDR	18524.43	< 1.0E-300	883
Visceral abdominal fat	*LUC7L3*	10% FDR	8.03	0.0046	11
	LUC7L3	15% FDR	9.26	0.0023	15
	CD163	10% FDR	2.62	0.1	4
	CD163	15% FDR	1272.63	9.99-279	378
Liver	*CPEB2*	10% FDR	133.06	8.76E-31	44
	CPEB2	15% FDR	82.95	8.43E-20	190

Table 4.5: Correlations between network targets in STAGE. Kruskal Wallis test calculated for distribution of correlations between network targets compared to correlations within random gene set of same size.

In STAGE subcutaneous fat, correlations in gene expression were stronger be-

tween target genes for each sub-network compared to random distributions (Figure 4.2). As with METSIM, the disparity between random and target correlations was stronger at the 15% FDR threshold compared to the 10% cut off for all gene sets ($p < 1.0\times10^{-300}$). However, differences were still observed at the 10% level for all gene sets ($p \leq 1.6\times10^{-30}$).

Two causal gene sub-networks had been previously identified in STARNET-visceral abdominal fat, regulated by the genes *CD163* and *LUC7L3*. As with previous correlations, the disparity between target and random distributions (Figure 4.3) was observed at the 15% FDR threshold, however in this instance this may be influenced by the fact that *CD163* and *LUC7L3* have relatively few targets at the 10% threshold (4 and 12 respectively). *CD163* shows a strong difference between target and random distributions at the 15% threshold ($p = 2.73\times10^{-273}$), however this is far weaker for *LUC7L3* ($p = 0.0021$). This was repeated using STARNET gene expression data, where all target sets show a skewed distribution in comparison to the random distribution (Figure S4.12). However, as with METSIM there is a large difference between the targets of *CD163* and *LUC7L3* at 15% FDR ($p < 1.00\times10^{-300}$; 1.62×10^{-21} respectively).

A single network was identified in STARNET-liver, under the regulation of *CPEB2*. Correlations in STAGE-liver were obtained between *CPEB2* networks targets and a random gene set, at both 10% and 15% network thresholds (Figure 4.4). The difference between target and random distributions is present at both 10% ($p = 3.34\times10^{-11}$) and 15% ($p = 3.01\times10^{-26}$) threshold, however this is weaker than has been previously observed for network regulators of other tissues. Again, this was repeated using gene expression data from STARNET liver, which showed strong contact between target and random distributions for both 10% and 15% targets for *CPEB2* (Figure S4.13).

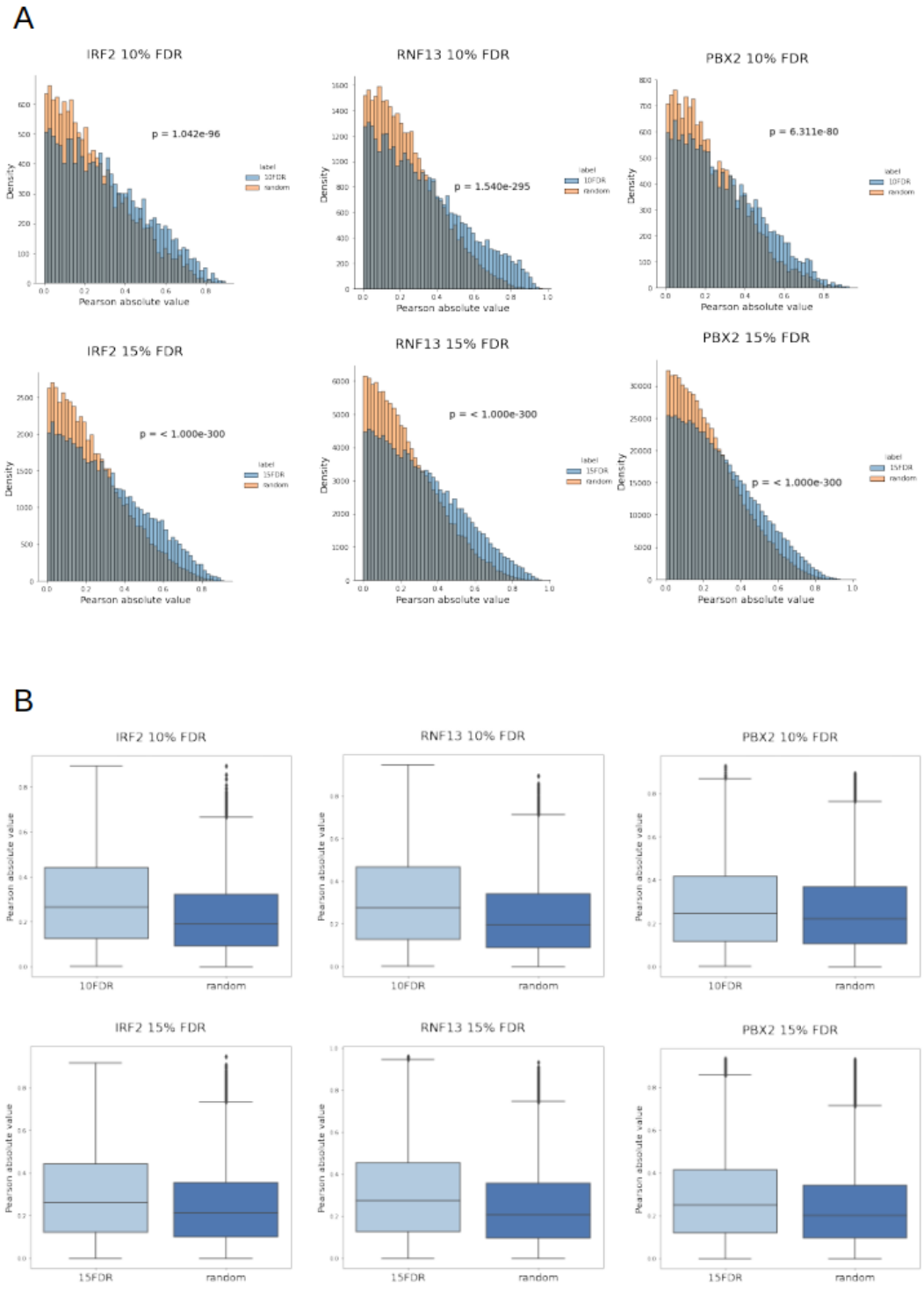


Figure 4.2: Correlations in gene expression between network targets in STAGE subcutaneous fat as absolute values. (A) Distribution plot of subcutaneous fat sub-network correlations compared to correlations within random gene set of same size (orange). Comparison between targeted and random distribution formalised by Kruskal Wallis p-value. (B) Boxplots comparing distribution between targets and random gene set.

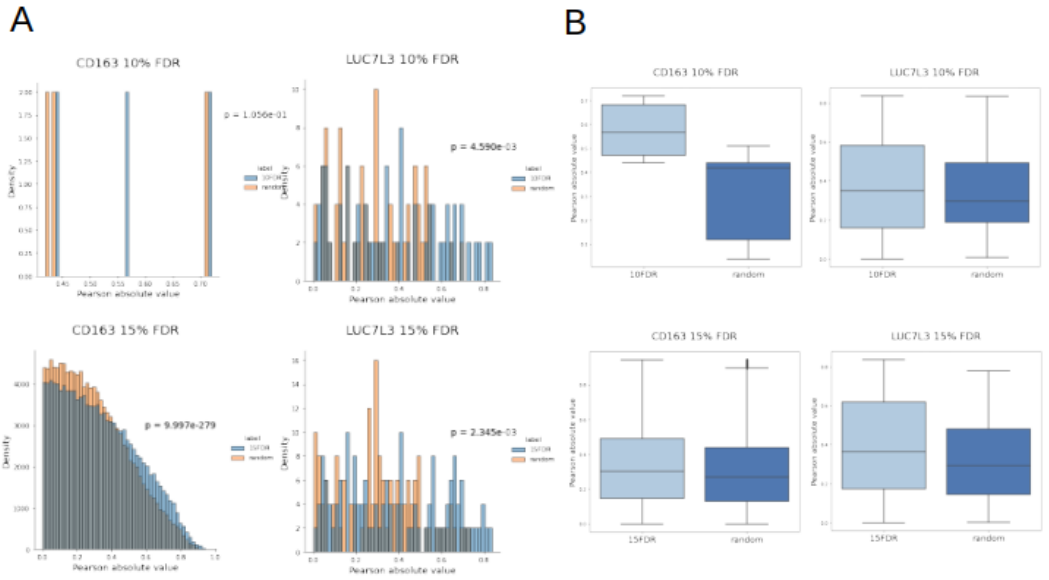


Figure 4.3: Correlations in gene expression between network targets in STAGE visceral adipose fat as absolute values. (A) Distribution plot of visceral adipose fat sub-networks compared to correlations within random gene set of same size (orange). Comparison between targeted and random distribution formalised by Kruskal Wallis p-value. (B) Boxplots comparing distribution between targets and random gene set.

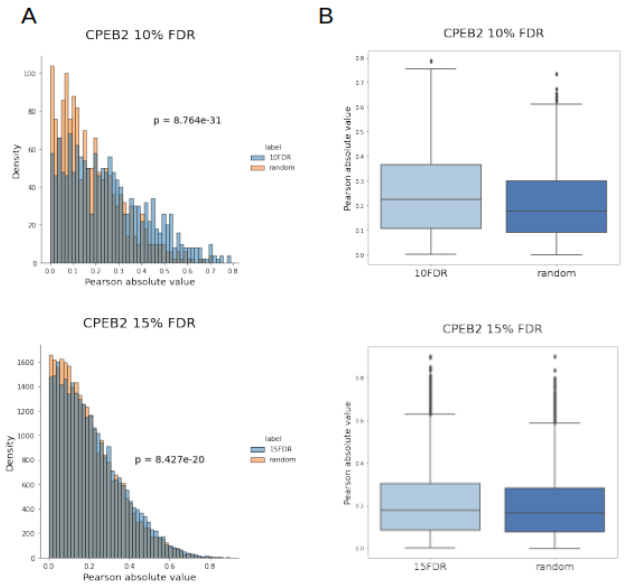


Figure 4.4: Correlations in gene expression between network targets in STAGE liver as absolute values. (A) Distribution plot of liver sub-network compared to correlations within random gene set of same size (orange). Comparison between targeted and random distribution formalised by Kruskal Wallis p-value. (B) Boxplots comparing distribution between targets and random gene set.

4.3.3 Correlations in gene expression between network regulators and targets

Regulator-target correlations in METSIM subcutaneous fat

In addition to examining correlations between network targets, we also examined the variability in correlation distributions between network regulators and their predicted targets. Correlations were calculated between sub-network regulators and their predicted targets for both the true predicted targets and a random set of genes of the same size in METSIM subcutaneous fat (Figure 4.5). Again the Kruskal Wallis test was used to formalise the difference between the target correlation distribution and that of the random selection (Table 4.6).

Network regulator	FDR threshold	KW test statistic	p-value	No. Target genes
IRF2	10% FDR	0.82	0.36	128
IRF2	15% FDR	4.30	0.038	247
RNF13	10% FDR	48.68	3.01E-12	215
RNF13	15% FDR	68.78	1.10E-16	416
PBX2	10% FDR	0.31	0.58	138
PBX2	15% FDR	7.75	0.0054	883

Table 4.6: Correlations between network regulators and targets in METSIM subcutaneous fat. Kruskal Wallis test calculated for distribution of correlations between network regulators and targets compared to correlations within random gene set of same size.

For all sub-networks, regulator-target correlations were skewed towards higher correlations, however the same was also true for the regulator-random correlations meaning that for most sub-networks there was little difference between the two distributions. The greatest difference could be seen between the random and target distributions of *RNF13* at both the 10% ($p = 3.011 \times 10^{-12}$) and 15% ($p = 1.01 \times 10^{-16}$) FDR thresholds. As had been seen with the correlations between targets, for all sub-networks the difference in distributions was strongest at the 15% threshold including for *IRF2* ($p = 0.038$) and *PBX2* ($p = 0.0054$).

Again, this approach was repeated using STARNET expression data (Table S4.9, Figure S4.14). Correlations again appeared lower in STARNET compared to MET-

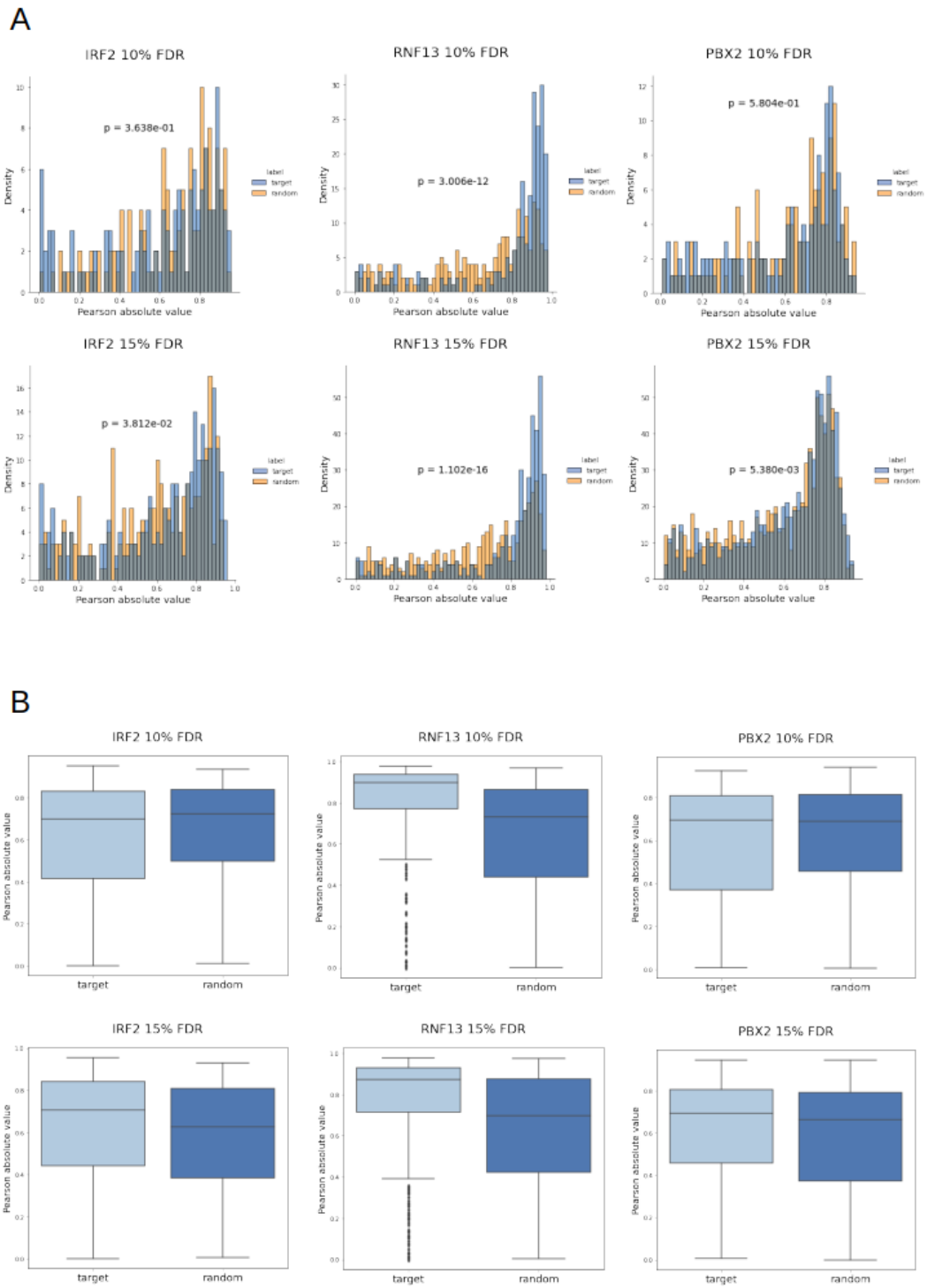


Figure 4.5: Correlations in gene expression between network regulators and predicted targets in METSIM subcutaneous fat as absolute values. (A) Distribution plot of subcutaneous fat regulator-target compared to correlations between regulator and random gene set of same size (orange). Comparison between targeted and random distribution formalised by Kruskal Wallis p-value. (B) Boxplots comparing distribution between regulator-targets and regulator-random gene set.

SIM, however there was greater distinction between the target vs random correlations when compared to the expression of the regulator compared to the distribution of correlations between target genes. All networks showed a distinction between the distribution of regulator-target correlations compared to correlations between the regulator and random genes.

Regulator-target correlations in STAGE liver and adipose

As with the correlations between network targets, STAGE was used to independently examine patterns in gene expression between network regulators and their corresponding targets in liver, subcutaneous fat and visceral abdominal fat. Regulator-target correlations were calculated for each sub-network and compared to a regulator-random distribution, selected from a tissue specific background. Again, this was formalised using the Kruskal Wallis test (Table 4.7).

Tissue	Network regulator	FDR threshold	KW test statistic	p-value	No. Target Genes
Subcutaneous fat	*IRF2*	10% FDR	3.25	0.07	128
	IRF2	15% FDR	1.28	0.26	247
	RNF13	10% FDR	1.65	0.20	215
	RNF13	15% FDR	4.63	0.031	416
	PBX2	10% FDR	4.95	0.026	138
	PBX2	15% FDR	23.26	1.41E-06	883
Visceral abdominal fat	*LUC7L3*	10% FDR	3.13	0.077	11
	LUC7L3	15% FDR	3.22	0.073	15
	CD163	10% FDR	3.03	0.082	4
	CD163	15% FDR	0.05	0.83	378
Liver	*CPEB2*	10% FDR	0.02	0.88	44
	CPEB2	15% FDR	0.09	0.77	190

Table 4.7: Correlations between network regulators and targets in STAGE. Kruskal Wallis test calculated for distribution of correlations between network regulators and targets compared to correlations within random gene set of same size.

In STAGE-subcutaneous fat, there was minimal difference between the regulator-random and regulator-target correlations across all all sub-networks. The strongest difference was observed in *PBX2* at the 15% threshold ($p = 1.41 \times 10^{-6}$). For both *IRF2* and *PBX2*, the differences between random and targeted correlations at both

10% and 15% cut offs are minimal, however interestingly for *IRF2* they are stronger at the 10% FDR threshold compared to 15% (p = 0.07).

In STAGE-visceral abdominal fat, there was minimal difference between regulator-target and regulator-random distributions for either *LUC7L3* or *CD163*. There was some variability in distributions at the 10% thresholds, however this is likely due to the relatively low number of genes at these thresholds. In STARNET-visceral abdominal fat, the difference is more stark between random and targets, although the overall correlation values are low (Figure S4.15).

There was no difference observed between regulator-target and regulator-random distributions in STAGE-liver. For *CPEB2*, there was no distinction between the target and random correlation distributions at either the 10% or 15% FDR thresholds. In STARNET-liver, although correlation values remained low, the distinction between targeted and random correlations was strong at both the 10% ($p = 1.83 \times 10^{-13}$) and 15% ($p = 1.68 \times 10^{-45}$) thresholds (Figure S4.15).

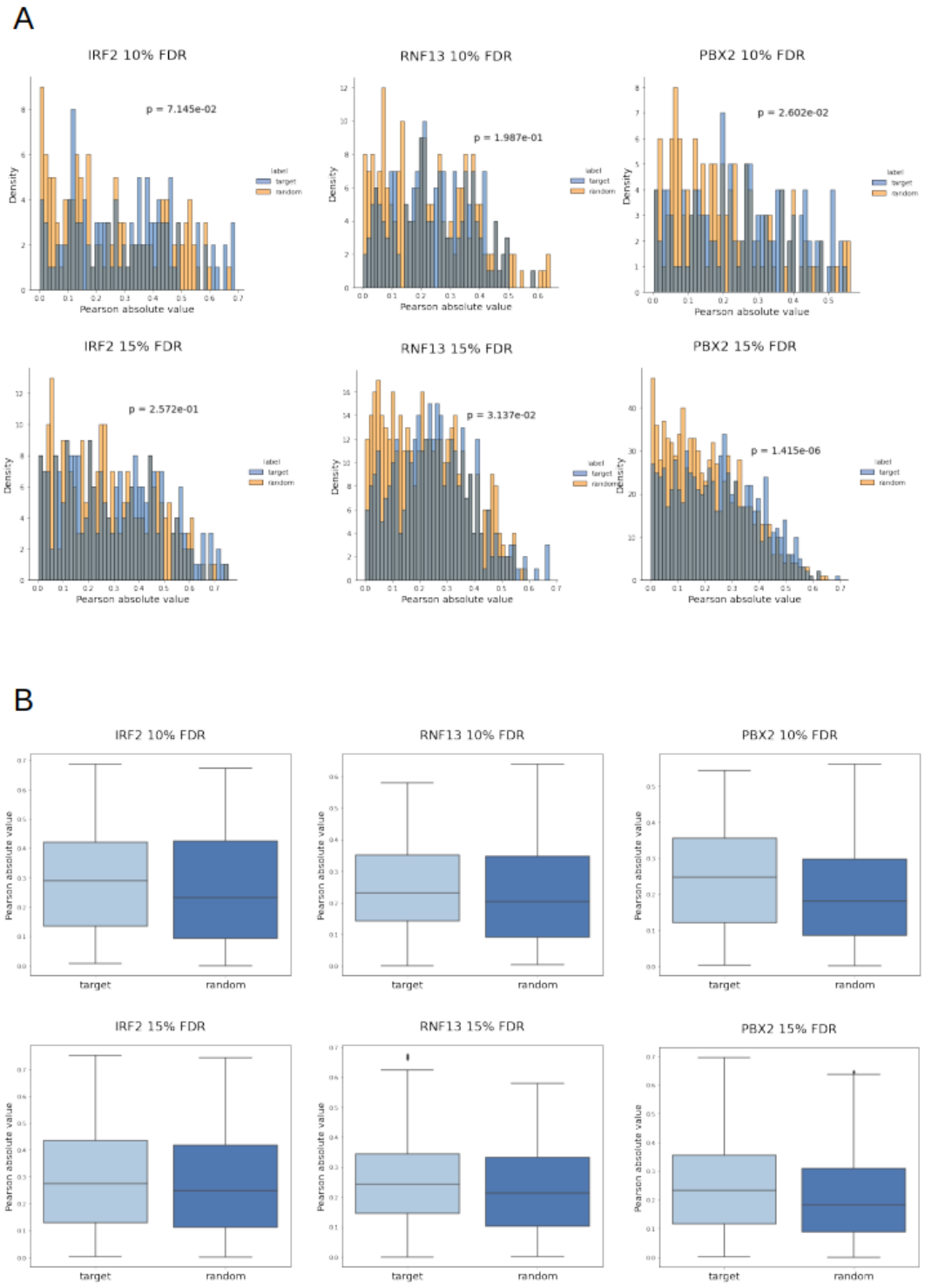


Figure 4.6: Correlations in gene expression between network regulators and predicted targets in STAGE subcutaneous fat as absolute values. (A) Distribution plot of subcutaneous fat regulator-target compared to correlations between regulator and random gene set of same size (orange). Comparison between targeted and random distribution formalised by Kruskal Wallis p-value. (B) Boxplots comparing distribution between regulator-targets and regulator-random gene set.

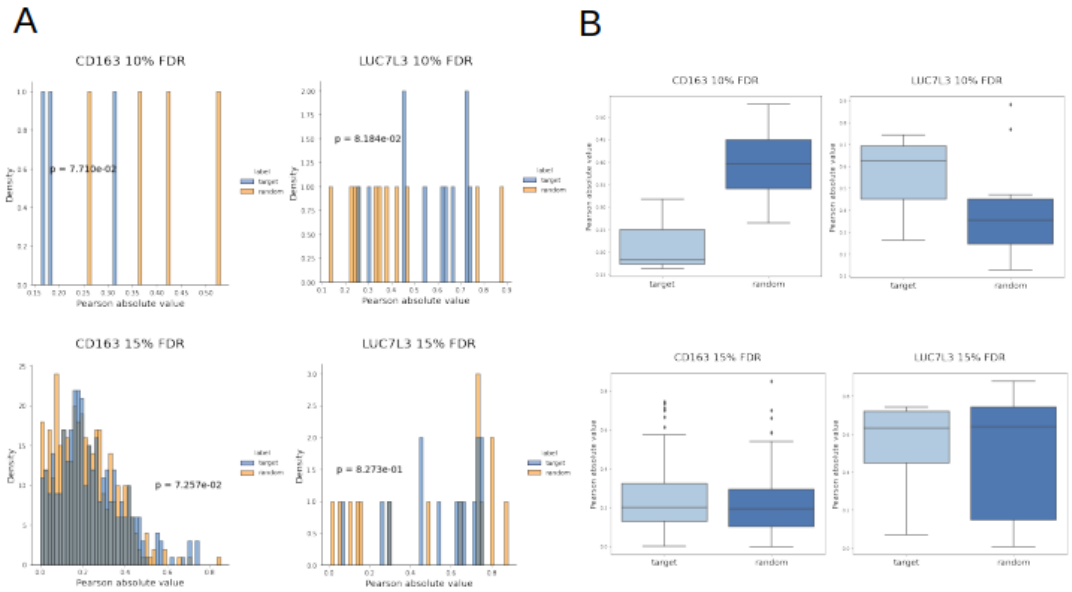


Figure 4.7: Correlations in gene expression between network regulators and predicted targets in STAGE visceral adipose fat as absolute values. (A) Distribution plot of visceral adipose fat regulator-target compared to correlations between regulator and random gene set of same size (orange). Comparison between targeted and random distribution formalised by Kruskal Wallis p-value. (B) Boxplots comparing distribution between regulator-targets and regulator-random gene set.

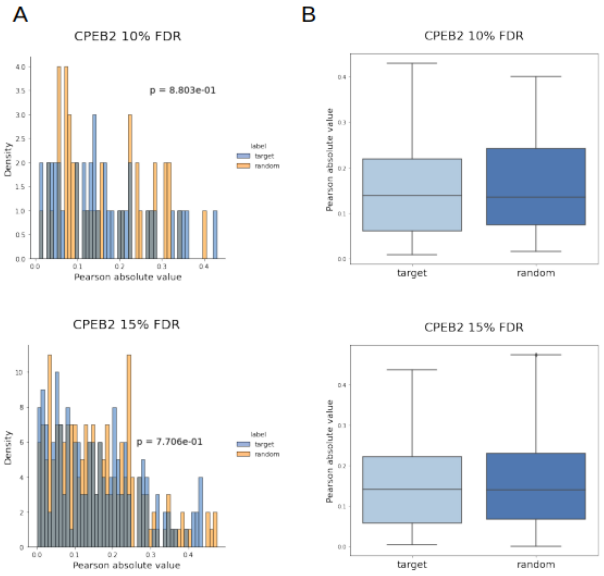


Figure 4.8: Correlations in gene expression between network regulators and predicted targets in STAGE liver as absolute values. (A) Distribution plot of liver regulator-target compared to correlations between regulator and random gene set of same size (orange). Comparison between targeted and random distribution formalised by Kruskal Wallis p-value. (B) Boxplots comparing distribution between regulator-targets and regulator-random gene set.

4.4 Discussion

4.4.1 Independent trans-associations within *SERPINA6/SERPINA1* locus

Trans-eQTL associations identified from a single dataset often prove challenging to reproduce in independent data sources, a problem which is compounded by marginal effect sizes. There is also the added difficulty of sourcing comparable independent datasets that fulfil the criteria necessary for studying the impact of genetic variation on gene expression. This can often lead to trans-associations being missed, resulting in an increased false negative rate.

Although STAGE follows a similar study design to STARNET, it is a unique cohort and is independent from STARNET. As well as being independent, STAGE is composed of individuals with comparable (European) ancestry, and contains gene expression data for many of the same tissues found in STARNET. These factors were taken in account when selecting STAGE as a source of replication for the cortisol associated trans-genes identified in STARNET. A significant drawback however is the difference in sample sizes between STAGE (n = 114) and STARNET (n = 600), which makes it more challenging to detect some of the small effect size associations identified in STARNET.

Although, cortisol associated trans-genes were identified across all 5 STAGE tissues examined, there was minimal overlap with the trans-associations previously identified using STARNET. One trans-gene, *OGG1*, was identified in both STARNET and STAGE for skeletal muscle. *OGG1* encodes the enzyme 8-Oxoguanine glycosylase which is involved in DNA repair[234]. Murine models have also demonstrated that *OGG1* deficiency, result in increased lipid deposition in skeletal muscle[235]. However, *OGG1* is associated with a different SNP in STAGE compared to STARNET, although both SNPs are associated with plasma cortisol.

METSIM is a large cohort (n = 982), with genotype and subcutaneous fat gene

expression data. METSIM was used to identify cortisol associated trans-genes that could be compared to those identified in STARNET-subcutaneous fat. However, the number of individuals available for eQTL discovery was dramatically reduced when selecting for individuals where where both subcutaneous fat gene expression and genotype data was available (n = 189). Moreover, as METSIM genotypes had been obtained using an exome chip array, this excluded the majority of intronic SNPs that were associated with plasma cortisol at the *SERPINA6*/ *SERPINA1* locus. No plasma cortisol associated SNPs were identified, meaning that we were unable to to identify any cortisol associated trans-genes. Although this resource is currently lacking the capability to measure these cortisol associated genotypes, this could potentially be overcome through the application of imputation methods[236] however issues relating to sample size will still remain.

In addition to using genotype data to link cortisol associated genetic variants to gene expression, we used data obtained from perturbation experiments to identify trans-genes, whose expression varies in response to treatment with cortisol. Individuals with primary adrenal insufficiency, or Addison's disease, are lacking in their production of adrenocortical hormones, including cortisol[237]. These individuals represent a unique opportunity to measure changes in gene expression in response to cortisol, in what constitutes a human glucocorticoid "knock-out". The BIOCORT authors identified subcutaneous fat genes in Addison's disease patients that were differentially expressed in response to treatment with hydrocortisone[229]. This presented an opportunity to identify genes associated with genetic variation for cortisol, that have also been demonstrated to respond to hydrocortisone treatment in individuals with low basal levels of glucocorticoids.

Five unique genes were identified that were both associated with genetic variation for plasma cortisol in STARNET and differentially expressed in response to hydrocortisone treatment in the BIOCORT, within the same tissue. Although none of these genes were key network regulators, it is notable that there was one gene, *BRD2*, which was both a cortisol associated trans-gene and a target of the *PBX2* subnetwork in subcutaneous fat. Additionally another trans-gene, *STAT4* was shown to

be associated with plasma cortisol, differentially expressed in response to hydrocortisone treatment and regulated a small sub-network of 3 target genes. These findings represent sustained changes in gene expression, across independent datasets, identified using different methodological approaches.

A limitation of identifying cortisol associated trans-genes with BIOCORT, is that gene expression was not measured in many of the tissues where trans-genes had been identified in STARNET. As a result, this approach could only be applied to trans-genes identified in subcutaneous fat and not any of the other STARNET tissues. Another consideration is that gene expression has been shown to be highly variable when sampling from different subcutaneous adipose depots[238]. Although the subcutaneous fat measured in both STARNET and BIOCORT has both been sampled from the abdominal region, depot specific variation may occur. These regional differences contribute added difficulty in comparing patterns in gene expression across datasets, where not all variation can be attributed to the studied perturbation, be it genetic or experimental.

An additional limitation of using patients with Addison's disease, is that in a cortisol free system there are systemic changes that may alter gene expression which are not reflective of response to hydrocortisone treatment. Therefore, it is unclear if differentially expressed genes in Addison's patients can be used as a direct proxy for cortisol responsive genes. There is evidence suggesting that gene expression levels of cortisol responsive gene respond to exogenous glucocorticoid treatment[239], however it is unclear if this is the case in a cortisol free system.

4.4.2 Network targets are highly correlated in independent cohorts

Gene networks were reconstructed in STARNET using causal methods, through the use of eQTLs as genetic instruments to infer pairwise gene-gene relationships. We aimed to investigate if these coordinated changes in gene expression are present in other datasets outside of STARNET. It is possible to compare correlations in gene expression within predicted networks to random sets of genes under the hypothesis

that correlations within true gene networks would be expected to be stronger than a random set. We examined gene network correlations using two approaches; calculating correlation distributions between network targets and calculating correlation distributions between network targets and associated regulator.

The gene networks identified in subcutaneous fat were the primary targets of replication for several reasons; they represented the most sub-network regulators, with the most targets at the strictest threshold (10% FDR) and there were multiple datasets available for replication (METSIM and STAGE). For all subcutaneous fat sub-networks, a shift was observed between the targeted and random distributions within the network targets, with the *RNF13* network appearing most strongly correlated. This was particularly prevalent at the 15% FDR threshold leading to a stronger Kruskal Wallis test statistic, although the bulk of this difference may be driven by the larger numbers of genes in the 15% FDR sub-networks.

When examining correlations between regulators and targets, the distinction between targeted and random distributions becomes less clear even when correlation distributions are heavily shifted towards being positively correlated. However, as with the correlations between targets, the shift between regulator-target and regulator-random distributions is strongest at the 15% threshold. Again, the strongest shift can be observed for the *RNF13* sub-network targets, particularly in the METSIM dataset. One reason that the *IRF2* and *PBX2* sub-network shifts are not as prominent as those observed in *RNF13* could be explained by the fact that both *IRF2* and *PBX2* are transcription factor, and therefore likely to regulate a number of genes outside of these putative cortisol regulated networks, leading to an inflation in the correlations of the random distributions.

Replication of gene networks in visceral abdominal adipose and liver, was limited by the fact that correlations could only be examined in the STAGE dataset. There was only one sub-network identified in STARNET-liver, which was under the regulation of *CPEB2*. As with subcutaneous fat, there is a greater disparity observed between random and targeted correlations at the 15% threshold. The same is true

for the sub-networks of visceral abdominal fat, particularly for *CD163* where there is a dramatic increase in the number of targets from 4 to 378. There is a more modest increase for *LUC7L3* from 10% to 15%, however this corresponds to a small increase from 11 to 15 targets. As with subcutaneous fat, there is little to no difference between the regulator-target and the regulator-random correlations.

It is worth noting that there are observable differences in the range of correlation values between different datasets, giving the impression that network correlation values are higher in METSIM than they are in STARNET. This is likely to be due to numerical differences in the pre-processing of gene expression data from these respective datasets where METSIM is presented as the transcripts per million (TMP) and STARNET has undergone $\log_2$ transformation. For this reason, it is important to consider the divergence between random and targeted correlations within the respective datasets as opposed to directly comparing correlation values.

4.5 Conclusion

We have identified genes trans-associated with genetic variation for plasma cortisol, a subset of which show evidence of regulation by GR and in turn act as regulators for glucocorticoid responsive transcriptional networks. This work has been characterised within a single dataset, STARNET, therefore it is critical to provide evidence of reproducibility in independent data sources. We have demonstrated that associations with genetic variation for cortisol are present within the STAGE dataset, although there is limited overlap with trans-genes identified in STARNET, an issue likely to be exacerbated by small sample sizes in STAGE. However, we describe patterns of gene expression that correspond to predicted cortisol network targets in both METSIM and STAGE datasets, with higher correlations between network targets over what would be expected by chance. These data highlight the impact that genetic variation for plasma cortisol has upon robust and reproducible tissue specific transcriptional variation, providing insight into the wider impact of glucocorticoid biology, as mediated by GR.

4.6 Supplementary data

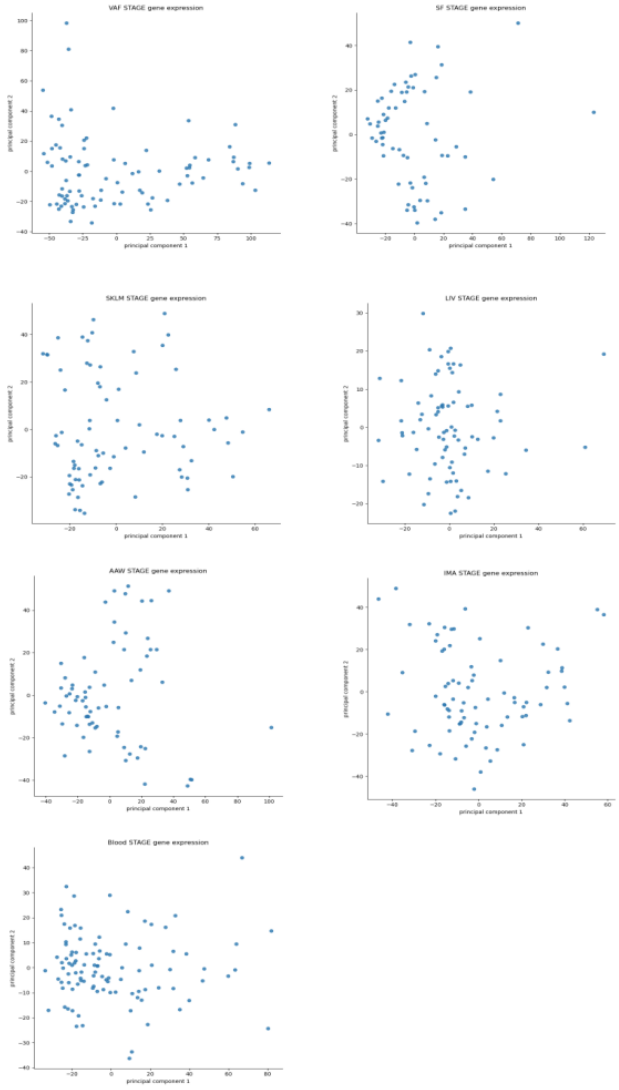


Figure 4.9: Principal component analysis of gene expression samples across all STAGE tissues.

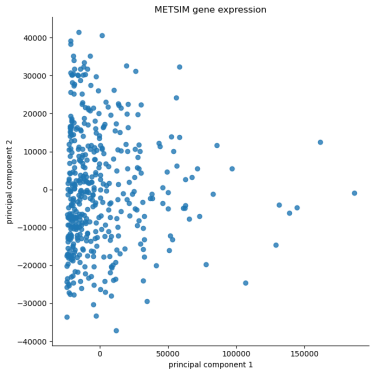


Figure 4.10: Principal component analysis of gene expression samples in MESTIM subcutaneous fat.

Tissue	Network regulator	FDR threshold	KW test statistic	p-value	No. Target Genes
Subcutaneous fat	*IRF2*	10% FDR	5217.64	1.00E-300	128
	IRF2	15% FDR	14452.98	1.00E-300	247
	RNF13	10% FDR	15327.41	1.00E-300	125
	RNF13	15% FDR	26976.73	1.00E-300	416
	PBX2	10% FDR	3974.48	1.00E-300	138
	PBX2	15% FDR	60503.15	1.00E-300	883
Visceral abdominal fat	*LUC7L3*	10% FDR	86.96	1.11E-20	11
	LUC7L3	15% FDR	85.45	2.38E-20	15
	CD163	10% FDR	5.36	0.021	4
	CD163	15% FDR	13135.27	1.00E-300	378
Liver	*CPEB2*	10% FDR	1047.46	8.67E-230	44
	CPEB2	15% FDR	6088.62	1.00E-300	190

Table 4.8: Correlations between network targets in STARNET. Kruskal Wallis test calculated for distribution of correlations between network targets compared to correlations within random gene set of same size.

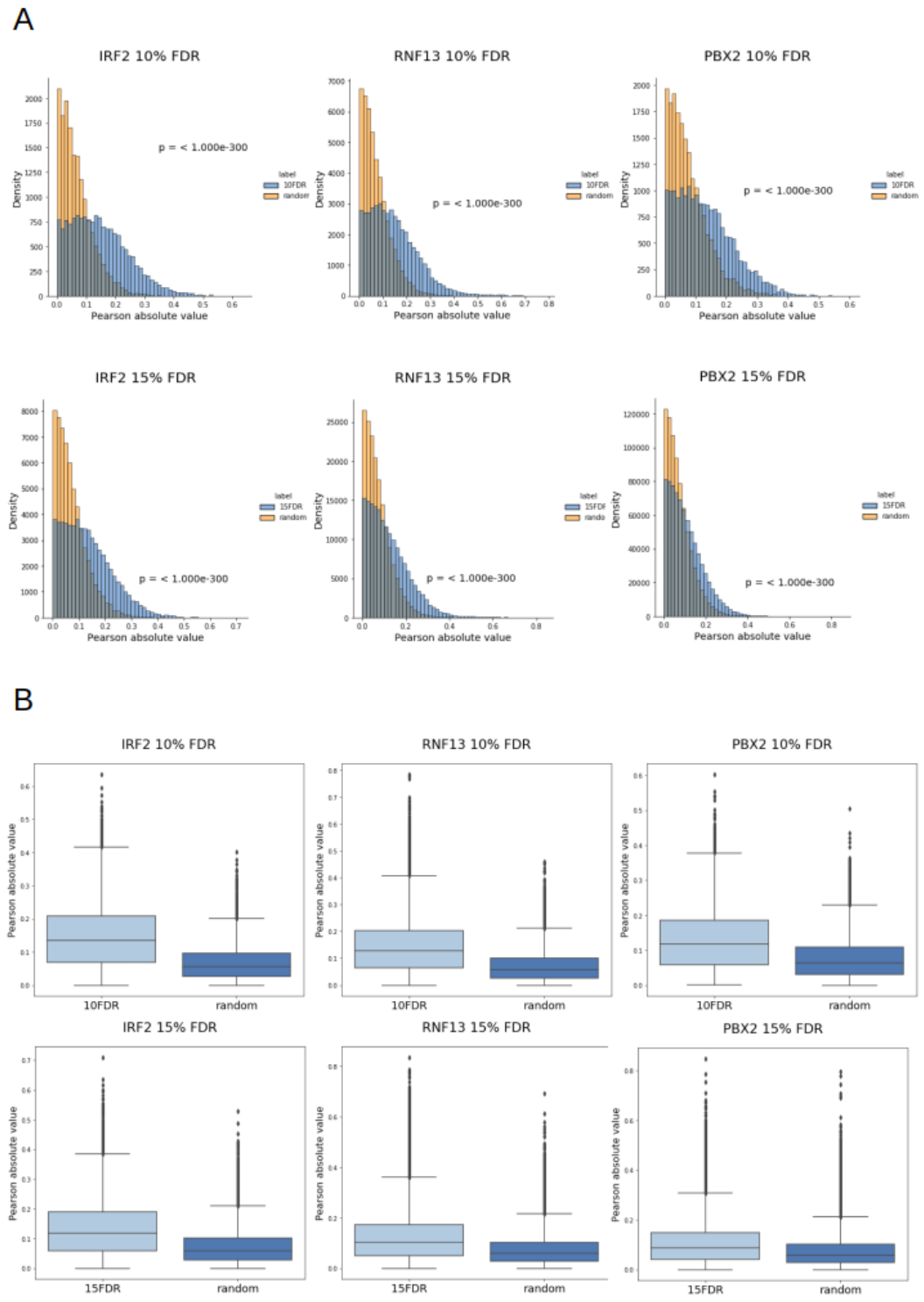


Figure 4.11: Correlations in gene expression between network targets in STARNET subcutaneous fat as absolute values. (A) Distribution plot of subcutaneous fat sub-network correlations compared to random gene set of same size (orange). Comparison between targeted and random distribution formalised by Kruskal Wallis p-value. (B) Boxplots comparing distribution between targets and random gene set.

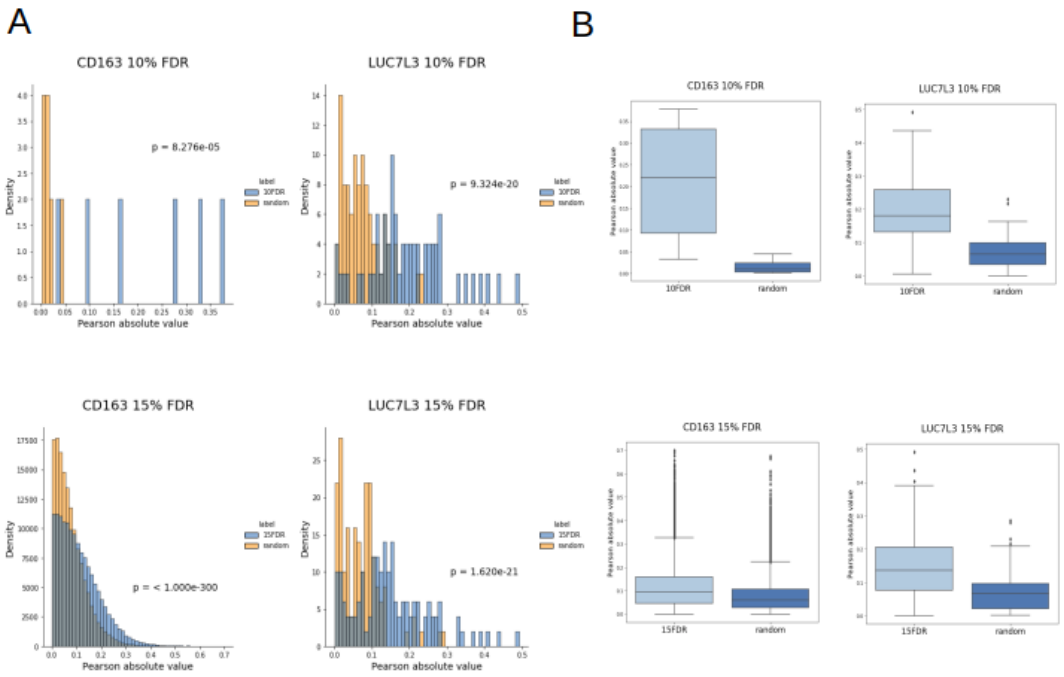


Figure 4.12: Correlations in gene expression between network targets in STARNET visceral adipose fat as absolute values. (A) Distribution plot of visceral adipose fat sub-network correlations compared to random gene set of same size (orange). Comparison between targeted and random distribution formalised by Kruskal Wallis p-value. (B) Boxplots comparing distribution between targets and random gene set.

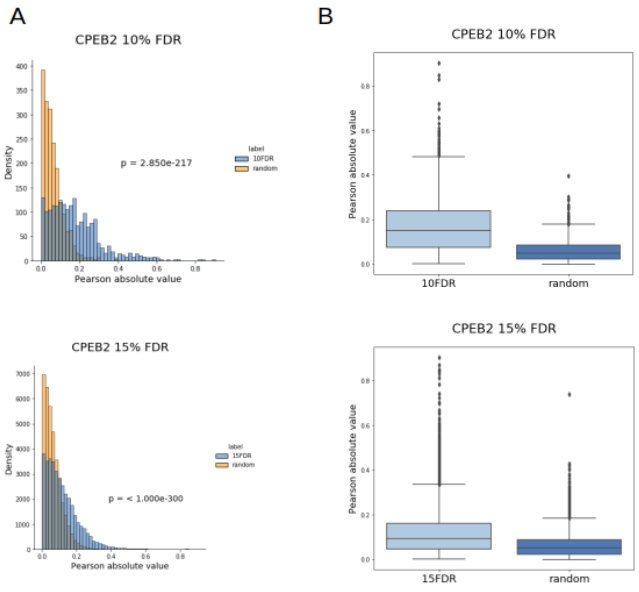


Figure 4.13: Correlations in gene expression between network targets in STARNET liver as absolute values. (A) Distribution plot of liver sub-network correlations compared to correlations between random gene set of same size (orange). Comparison between targeted and random distribution formalised by Kruskal Wallis p-value. (B) Boxplots comparing distribution between targets and random gene set.

Tissue	Network regulator	FDR threshold	KW test statistic	p-value	No. Target Genes
Subcutaneous fat	*IRF2*	10% FDR	116.39	3.91E-27	128
	IRF2	15% FDR	220.40	7.38E-50	247
	RNF13	10% FDR	202.88	4.91E-46	215
	RNF13	15% FDR	280.28	6.53E-63	416
	PBX2	10% FDR	174.76	6.75E-40	138
	PBX2	15% FDR	804.20	6.58E-177	883
Visceral abdominal fat	*CD163*	10% FDR	2.08	0.15	11
	CD163	15% FDR	309.55	2.74E-69	15
	LUC7L3	10% FDR	11.88	0.00057	4
	LUC7L3	15% FDR	12.87	0.00033	378
Liver	*CPEB2*	10% FDR	49.38	2.11E-12	44
	CPEB2	15% FDR	200.12	1.97E-45	198

Table 4.9: Correlations between network regulators and targets in STARNET. Kruskal Wallis test calculated for distribution of correlations between network regulators and targets compared to correlations within random gene set of same size.

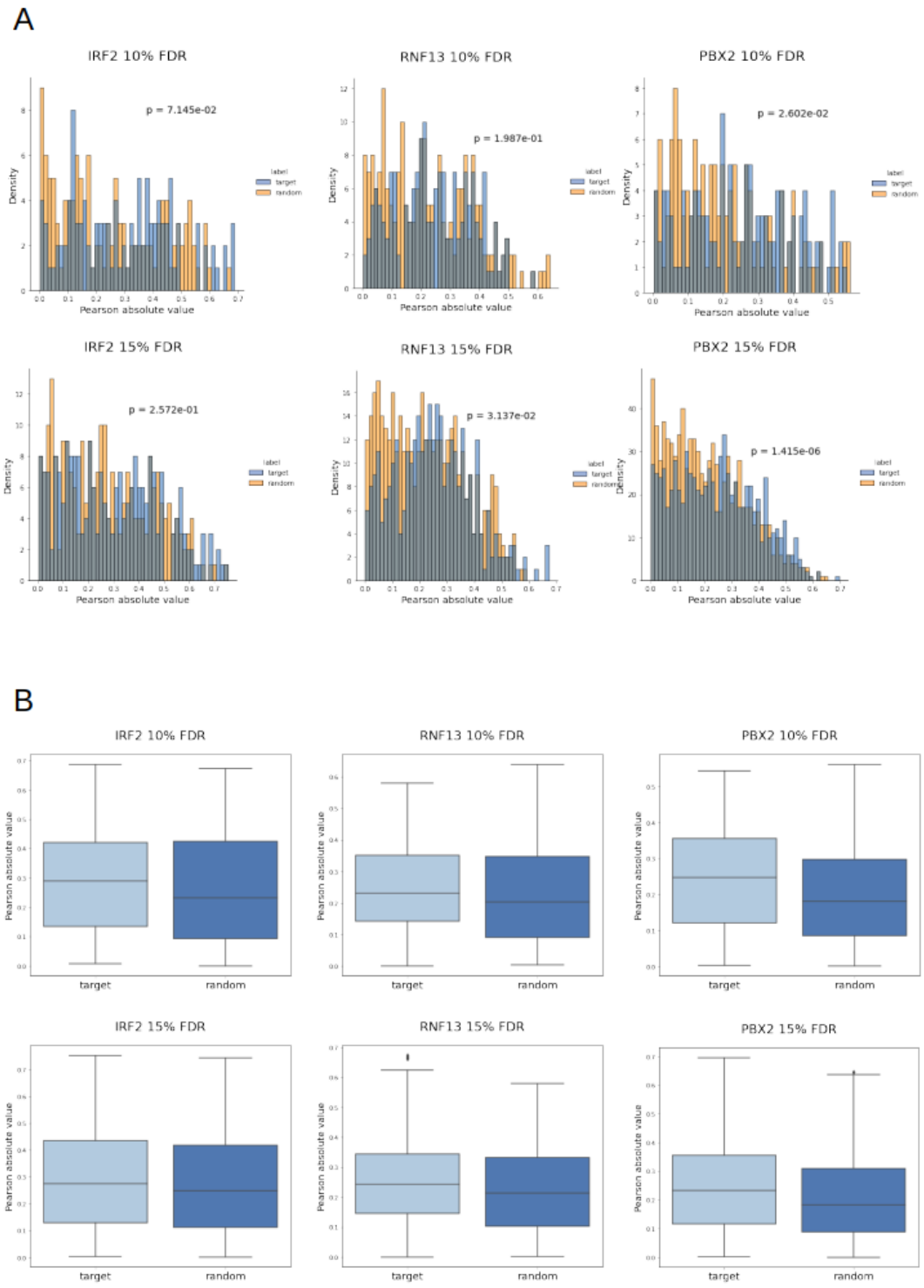


Figure 4.14: Correlations in gene expression between network regulators and predicted targets in STARNET subcutaneous fat as absolute values. (A) Distribution plot of subcutaneous fat regulator-target compared to correlations between regulator and random gene set of same size (orange). Comparison between targeted and random distribution formalised by Kruskal Wallis p-value. (B) Boxplots comparing distribution between regulator-targets and regulator-random gene set.

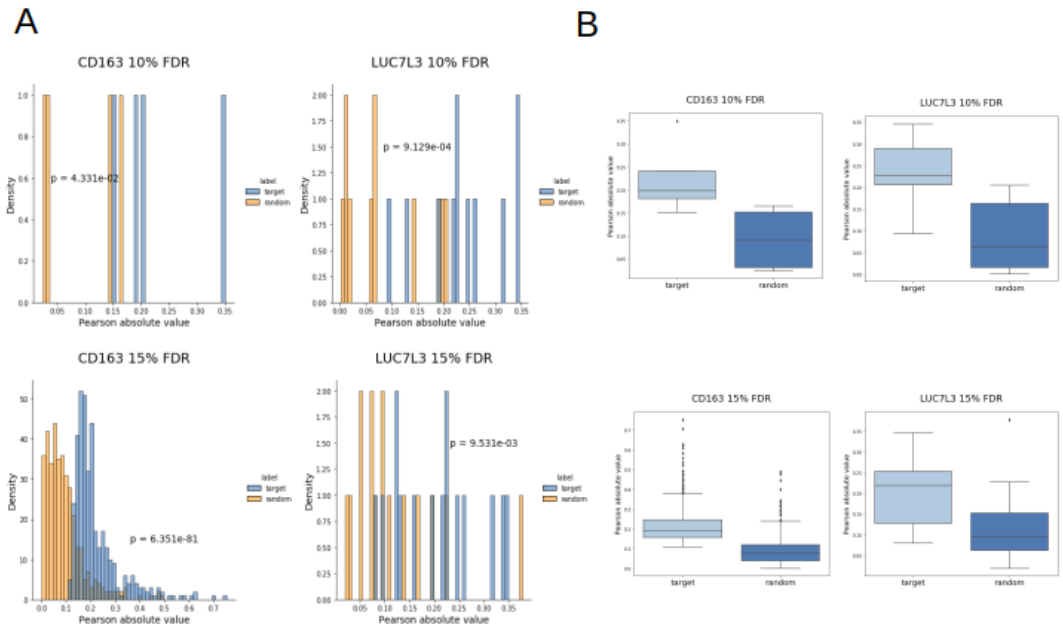


Figure 4.15: Correlations in gene expression between network regulators and predicted targets in STARNET visceral adipose fat as absolute values. (A) Distribution plot of visceral adipose fat regulator-target compared to correlations between regulator and random gene set of same size (orange). Comparison between targeted and random distribution formalised by Kruskal Wallis p-value. (B) Boxplots comparing distribution between regulator-targets and regulator-random gene set.

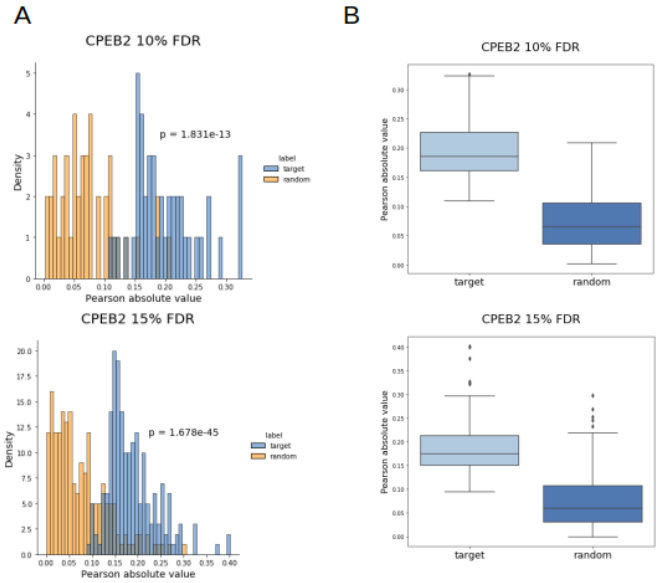


Figure 4.16: Correlations in gene expression between network regulators and predicted targets in STARNET liver as absolute values. (A) Distribution plot of liver regulator-target compared to correlations between regulator and random gene set of same size (orange). Comparison between targeted and random distribution formalised by Kruskal Wallis p-value. (B) Boxplots comparing distribution between regulator-targets and regulator-random gene set.

Chapter 5

Conclusions and future work

This thesis has aimed to characterise the impact that genetic variation for plasma cortisol has upon changes in gene expression throughout the transcriptome. This work has led to the identification of cortisol linked genetic variants, which mediate changes in gene expression in both cis and trans. We have scrutinised these trans-associations, to identify a subset of genes that are regulated by glucocorticoids, and in turn regulate downstream transcriptional networks, thus providing a deeper understanding of the transcriptional landscape driven by cortisol linked genetic variation.

5.1 Cortisol associated genetic variation is driven by changes in *SERPINA6* expression in liver

The sole SNP peak identified by the CORNET consortium as associated with plasma cortisol, was mapped to the *SERPINA6*/*SERPINA1* locus, containing genes with known cortisol involvement. For this reasons we chose to start our investigation by examining cis-associations within this locus. Both *SERPINA6* and *SERPINA1* play a role in glucocorticoid regulation through the protein CBG, where the former is the gene responsible for encoding CBG[151] and the latter is involved in CBG cleavage[148]. We

identified a peak of cis-eQTLs for *SERPINA6* in liver, the tissue in which *SERPINA6* is most highly expressed[145]. A subset of these cis-eQTLs was demonstrated to overlap and colocalise with cortisol associated SNPs, suggesting the presence of common SNPs responsible for both signals.

These findings are suggestive of a functional role for cortisol associated variation, as mediated by CBG. We were careful to examine other cis-associations within this region, including *SERPINA1*, where we were able to identify a cis-eQTL peak in blood. However, this peak did not include any cortisol associated SNPs, and was therefore unlikely to be contributing to the cortisol associated variation identified from the CORNET GWAMA. We were able to further refine the cortisol associated peak within the GWAMA, as *SERPINA6* cis-eQTLs only represented two out of the four LD blocks (LD blocks 2 and 4) that were shown to be associated with plasma cortisol at a level of genome wide significance, suggesting that only these blocks are responsible for mediating cortisol associated variation through CBG.

The lead cis-eQTL for *SERPINA6*, rs2736898, was also identified as the representative SNP for the second cortisol associated LD block following LD clumping. The alternate allele for this SNP exerts a negative effect on *SERPINA6*, however the lead SNP for the fourth LD block, rs9989237, increases *SERPINA6* expression. This is notable, as it suggests that although both blocks are associated with cortisol at the same locus, they have different functional implications. Further research would be required to uncover the impact of cross-talk within individuals with mutations for SNPs from more than one LD block, to determine if these alleles work synergistically or antagonistically, depending on the haplotype structure of the relative effect and reference alleles.

It is possible that a reduction in CBG expression would correspond to decreased levels of bound cortisol, which could lead to more free cortisol able to diffuse into tissues to bind to GR[240]. However, it is important to consider that a rise in free cortisol as a result of a decrease in CBG, would also be expected to be adjusted by negative feedback of the HPA axis[110]. This is a key finding, as we do in fact observe

downstream transcriptomic consequences in extra-hepatic tissues that appear to be mediated by CBG, which might be suggestive of a role for CBG beyond binding cortisol in the blood. Conversely, if CBG expression was increased this would lead to more bound cortisol and potentially a decrease in the gene expression of cortisol associated trans-genes.

It is important to note that while variation at the *SERPINA6/ SERPINA1* locus has an important impact on *SERPINA6* expression in liver, this impact is observed in other tissues as well. This impact is particularly prominent for rs2736898, which in addition to an association with plasma cortisol and with *SERPINA6* expression in liver, also mediates global transcriptional changes in liver and adipose. This suggests that variation affecting CBG, also mediates changes in gene expression in extra-hepatic tissues. Such a finding would be consistent with animal models, where CBG knock out mice show increased levels of free corticosterone[157] and how familial mutation in CBG have been linked to disease[154,155].

A missing feature from the description of cortisol associated cis-eQTLs for *SERPINA6*, is the functional role that this has upon the CBG protein, beyond changes in *SERPINA6* gene expression. Although *SERPINA6* is translated to CBG, the levels at the transcriptome are not directly reflected at the proteome[241]. Additionally, there are further modifications that can influence CBG protein level, including but not limited to cleavage of the reactive centre loop by α-1 antitrypsin, which significantly reduces the binding affinity of CBG[148,149]. This could be overcome by conducting a cis-pQTL analysis in the same region to identify genotypes associated with CBG protein levels. It would then be possible to identify cis-eQTLs for *SERPINA6*, which correspond to pQTLs for CBG to establish a through line from genetic changes to CBG expression, as mediated by modulation of *SERPINA6* gene expression.

CBG has been identified as the predominant mechanism for mediating genetic variation for cortisol, however this is only true in respect to the genetic variation identified from the CORNET GWAMA specifically. The issue of missing heritability is common among GWAS, and has still to be addressed by the field as a whole[46].

Estimates for the heritability of cortisol range between 40-60%[141], however the lead SNP identified from the CORNET GWAMA contributes only 0.13% of plasma cortisol variance[8]. Despite increasing the sample size from 12,597 in the original GWAMA[5] to 25,314 individuals no new loci outside of the *SERPINA6/ SERPINA1* locus were identified. It is likely that additional variation for cortisol exists outside of this locus, however this is mediated at a level of sub genome wide significance in the CORNET GWAMA. Therefore it is important to point out that although these findings suggests CBG is an important mediator of genetic variation for plasma cortisol, it is unlikely to be the only mechanism.

5.2 Genes trans-associated with plasma cortisol are regulated by GR in different tissues

In addition to identifying cis-associations at the *SERPINA6* locus, we describe how genetic variation for plasma cortisol is trans-associated with genes in multiple tissues within the STARNET dataset. We identified genes that were trans-associated with cortisol linked genetic variation across all STARNET tissues. The tissues with the highest number of unique trans-genes were liver, subcutaneous fat and visceral abdominal fat. By identifying genes that are trans-associated with cortisol, this helps to better describe the transcriptional landscape influenced by genetic variation at the *SERPINA6/ SERPINA1*.

Having identified a strong signal associated with *SERPINA6* expression in liver, this suggests a potential avenue for modulation of trans-associated genes. As CBG directly influences levels of free cortisol in the blood, where 80-90%[147] of cortisol is bound to CBG, changes in CBG concentration would result in altered levels of free cortisol able to activate GR. We made the decision to look for trans-genes across all cortisol associated SNPs at genome wide significance, including but not limited to those which are also associated with *SERPINA6* expression in liver. This was to capture the full extent of variation as mediated by this locus, as opposed to that

which is mediated by *SERPINA6* gene expression. As there are more factors that influence CBG expression and cortisol binding efficacy outside of *SERPINA6* expression, it made sense to include all cortisol associated SNPs.

Under our hypothesis, we predict that genetic effects of cortisol upon trans-associated genes will be mediated by glucocorticoids. Therefore, we aimed to identify a subset of cortisol-associated trans-genes where there is evidence of GR regulation. By combining evidence of GR regulation from a variety of experimental evidence, we were able to refine a subset of genes that were both associated with genetic variation for plasma cortisol and regulated by GR. This was done to identify true GR targets and justified the use of a modest (15%) FDR threshold, as selecting for GR responsive genes would reduce the probability of false positive findings.

We identified a subset of GR responsive genes in liver, subcutaneous fat and visceral adipose fat. However, we failed to observe a statistical enrichment of GR regulated genes in any of these trans-gene sets. This does not negate the identification of strong GR targets that are associated with plasma cortisol, but it may imply that there are some effects of cortisol linked genetic variation that are mediated outside of GR. It is also worth noting that some of the genes with higher levels of evidence for GR regulation have gone on to demonstrate regulation of transcription networks e.g. *CPEB2*, *IRF2*, *RNF13*. Therefore this instills confidence in our strategy of casting a wider net in selecting genes associated with plasma cortisol with a more lenient FDR threshold and then refining our trans-genes sets to include genes that are relevant to glucocorticoid biology.

Our genes sets in liver, subcutaneous and visceral abdominal adipose were filtered to select for GR regulated trans-genes. These tissues contained the greatest number of trans-genes and there is well documented GR involvement in these tissues[179,180]. However glucocorticoids are also major targets in skeletal muscle where they modulate protein and glucose metabolism[242]. A lack of resources for identifying tissue specific GR targets in other tissues, means that potential GR targets may have been missed in tissues outside of liver and adipose. With the availability of

skeletal muscle specific GR targets, future work could include an examination of GR regulated trans-genes in skeletal muscle or other GR tissues where cortisol associated trans-genes were identified.

It should also be noted that published evidence does not provide direct insight into the mechanism of GR regulation for GR-regulated trans-genes. GRE activation occurs in response to the binding of GR dimers[125], however ChIP-seq does not make a distinction between GR dimer or monomer action. GR monomers have been shown to mediate an effect through interacting with other DNA-bound transcription factors[126], therefore in the case of transcription factors such as *IRF2*, it is unclear if regulation by glucocorticoids is mediated by a GR dimer binding to a GRE, or a GR momomer interacting with *IRF2*. Additionally, for genes identified as differentially expressed in response to dexamethasone, we do not know the mechanism of this GR regulation. Functional experiments could work to resolve this, however these are ill-suited to a high throughput approach such as those used in this thesis to identify potential GR targets. However, this would be amenable for the smaller subset of network regulators identified in this thesis, as part of a future research strategy.

5.3 GR regulated trans-genes drive transcriptional networks

Although we identified cortisol associated trans-genes across multiple tissues, the phenotypic impact of this variation may not always be exerted by these trans-genes directly, but can be mediated through transcriptional targets under the regulation of GR responsive trans-genes. Having identified genes that were trans-associated with plasma cortisol and putatively regulated by GR, we used causal inference to identify regulatory networks driven by key trans-genes. This approach aimed to prioritise key GR responsive trans-genes that are responsible for regulating these downstream gene networks.

We identified causal gene networks in liver, subcutaneous fat and visceral abdominal fat where select GR regulated trans-genes act as regulators of sub-networks within overarching tissue specific networks. Pairwise causal relationships were established between network regulators and downstream targets using cis-eQTLs as genetic instruments. This approach has the benefit of generating directed relationships, where we have established that the regulator is influencing expression of the target and not the other way around. Additionally, this accounts for any unobserved confounding, where an outside factor could be responsible for the co-expression of the regulator and target, suggesting the presence of an interaction where one does not exist. However, a drawback of this approach is that we are limited by only being able to examine GR regulated trans-genes with valid cis-eQTLs. This means that there could be valid cortisol responsive networks regulated by GR trans-genes which we were unable to predict due to lack of a corresponding instrument.

The presence of pleiotropy can invalidate one of the fundamental assumptions of instrumental variable analysis, that the instrument should only be associated with the outcome through the exposure[243]. As we were careful to only select cis instruments, this issue would only occur if the selected instrument is associated with an additional gene within the cis region, outside of the exposure. Therefore, we calculated cis-association with the genes surrounding our network regulators to ensure that pleiotropy was not influencing our causal estimates. All sub-network regulators were free of any interfering cis associations with the exception of *PBX2* in subcutaneous fat. Multiple cis-associations for the *PBX2* instrument were identified, with 10 being targets of *PBX2* which raises the possibility that the transcriptional targets attributed to *PBX2* are being regulated by a genes other than *PBX2*. This does not invalidate the targets, but may mean that this network is not being regulated by a GR target.

All other sub-networks identified were determined to be robustly regulated by their targets. *IRF2* stands out as a network regulator of particular interest. There is strong evidence of GR regulation, where *IRF2* has been identified as a GR target from published dexamethasone treated adipocyte ChIP-seq experiments[185] and as

a putative GR target within ENCODE. It is robustly associated with its corresponding cis-eQTL instrument and there is an enrichment of *IRF2* targets within our predicted *IRF2* regulated causal network. Additionally, we show evidence of regulation by glucocorticoids within the targets of *IRF2*, potentially suggesting evidence of a feed-forward loop motif[217]. Interestingly the genotype for rs8022616, the cortisol associated SNP linked to *IRF2* expression in subcutaneous fat, is associated with a decrease in *IRF2* expression. Previous evidence suggests that interferon signalling is inhibited by glucocorticoids which would potentially align with this finding[244,245].

It is notable that enrichment of both GR and *IRF2* targets appears at the 15% FDR threshold. This is the same threshold where *IRF2* predicted targets are most strongly correlated, compared to what would be expected by chance, in both METSIM and STAGE datasets, suggesting that this is the optimal threshold for balancing both false positive and false negative findings. It should be noted that although we have demonstrated replication in independent datasets, all participants were of European ancestry, therefore not all findings will be generalisable to other populations and ethnicities.

The *RNF13* network is also of note, as the targets for this sub-network identified in STARNET subcutaneous fat appear to replicate better than any other sub-network in both METSIM and STAGE datasets. This is also one of the few network regulators, where correlations between the regulator and its corresponding targets appear more strongly than what would be expected by chance. *RNF13* encodes a protein that has been shown to play an important role in the unfolded protein stress response in the endoplasmic reticulum[198]. Apart from this, *RNF13* has not been well characterised so it is not immediately obvious by which mechanism *RNF13* may be modulating gene expression in subcutaneous fat. It is however notable that *RNF13* is a RING domain family member, as these proteins are known to possess E3 ubiquitin ligase activity, which are known to regulate a wide variety of cellular functions[246].

We have worked to effectively characterise the transcriptomic consequences of genetic variation for plasma cortisol, but for a comprehensive systems levels un-

derstanding of these networks, expanded integration of additional omics data is required. Following the same argument for the integration of proteomic data to understand links between genetic variation and CBG, future work should focus on the impact of trans-associations downstream of translation. Although gene expression plays a crucial role in the regulation of biological mechanisms, proteins form the functional foundation for biological activity, therefore improved characterisation of this omic level would be beneficial in better understanding how these networks may influence phenotypic variation.

5.4 Future work

It is important to consider that while we have replicated patterns of gene expression consistent with cortisol associated gene networks, we have yet to link gene expression of key network regulators to cortisol associated genetic variation in an independent dataset. Outside of the avenues for future work, that have already been outlined in this chapter, this is a key area that additional research should address.

Additional research is required to confirm the extent to which genes are influenced by cortisol linked genetic variation, including how this may be mediated by CBG. However, partially due to the uniqueness of STARNET as a dataset, it has been challenging to identify corresponding replication cohort, with appropriately genotypes samples and the necessary gene expression data required to conduct a replication trans-gene analysis.

Initially our replication strategy involved a two-pronged approach, whereby in addition to pursuing replication via statistical analyses using independent datasets, perturbation based experiments in a representative cell line could act as a source of validation. Work towards this approach was initiated, but was suspended during the COVID-19 pandemic in 2020. However this remains a viable strategy for future work.

SGBS cells are obtained from stromal vascular fraction from subcutaneous adi-

pose tissue of male infant with Simpson–Golabi–Behmel Syndrome[247]. These have been used to study human adipose biology, including through perturbation with glucocorticoids to study the role of leptin[248] and adipose steroidogenic activity[249], making SGBS a promising model for investigating how predictions of cortisol responsive genes may influence gene expression within an *in vitro* environment. SGBS cells require differentiation from pre-adipocytes to adipocytes (Figure 5.1), at which point they would be able to perturbed experimentally.

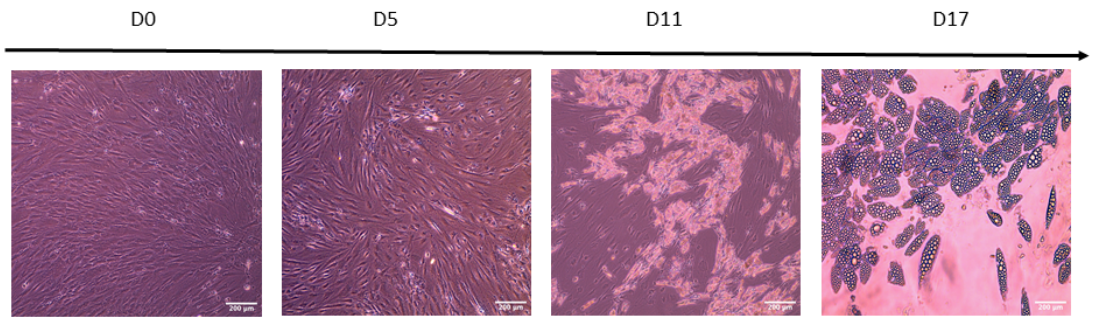


Figure 5.1: Differentiation of SBGS pre-adipocytes to mature adipocytes. 17 day differentiation process from day 0 (D0) to day 17 (D17). Maturation of adipocytes observable through the accumulation of lipid droplets. Scalebar depicts 200 µm.

A potential strategy would have been to first measure gene expression of predicted cortisol associated trans-genes within SGBS cells. Following this, it would be possible to measure how gene expression of these trans-genes vary in response to perturbation with a GR agonist such a dexamethasone. To validate potential gene networks, it would be possible to target subcutaneous fat sub-network regulators such as *IRF2* using an knock-down approach such as siRNA inhibition (Figure 5.2).

Following knock-down of the predicted network regulator it would be possible to measure the corresponding change in gene expression in the SGBS transcriptome, in the presence of and absence of a GR agonist. Technologies such as RNA-seq, could measure changes in gene expression which could be filtered for predicted network targets to see if expression varies in response to knock-down of their predicted regulator.

Experimental validation in a representative model, could be extremely valuable in measuring the impact of perturbations within cortisol responsive gene networks.

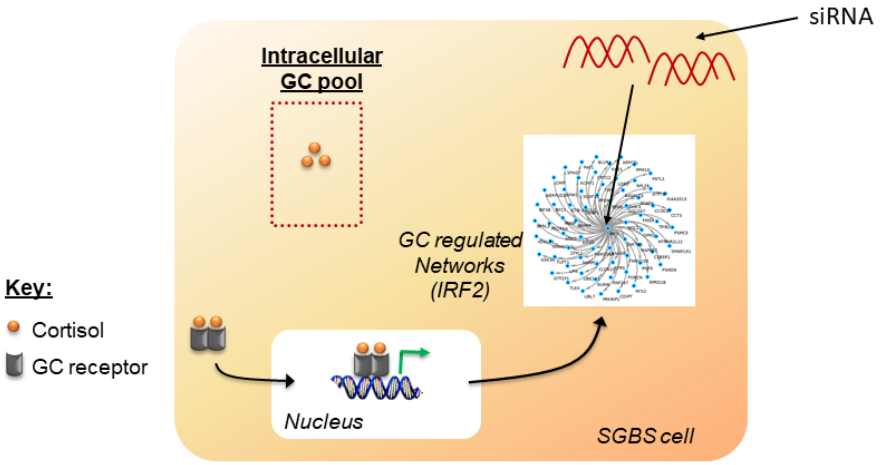


Figure 5.2: Analysis plan for perturbation of glucocorticoid responsive trans-genes. SGBS cells to me used as model for studying the effects of cortisol associated trans-genes and networks identified in STARNET-subcutaneous fat.

Although experimental models are ill-suited to study the direct effect of complex genetic variation, putative gene networks downstream of such variation, present as targets for experimental perturbation, with regulators acting as nexus points for mediating the effects of genetic variation. This opens the door for a variety of experiments, in addition to those already outlined in this section.

5.5 Conclusion

In this thesis, we have characterised the transcriptomic consequences of genetic variation for plasma cortisol. These variants result in changes in basal plasma cortisol levels that are modest but sustained at a population level. However, we describe how these small effect size variants can be linked to profound molecular changes across different tissues. We highlight CBG as the source these genetic changes that lead to coordinated variation in gene expression across the transcriptome, as mediated by GR.

We have categorised transcriptional networks using causal methods, which provide far higher resolution than other non-causal forms of network analysis, and ultimately provide greater scope for translation benefit. We identify key network regulators, while taking into account tissue specificity, expanding upon current understanding of the transcriptional consequences of GR signalling. Additionally, we highlight future directions for understanding the impact of complex genetic variation for cortisol on a systems wide level. These networks represent robust and sustained genetic changes at a population level and highlight key points of regulation with applications for both precision medicine and improved understanding of glucocorticoid biology.

Bibliography

[1] Ozaki, Kouichi et al. “Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction”. en. *Nature Genetics* 32.4 (2002), pp. 650–654.

[2] Hunter, David J. and Kraft, Peter. “Drinking from the fire hose–statistical issues in genomewide association studies”. eng. *The New England Journal of Medicine* 357.5 (2007), pp. 436–439.

[3] Husebye, E. S., Perheentupa, J., Rautemaa, R., and Kämpe, O. “Clinical manifestations and management of patients with autoimmune polyendocrine syndrome type I”. en. *Journal of Internal Medicine* 265.5 (2009), pp. 514–529.

[4] d’Hennezel, Eva, Dhuban, Khalid Bin, Torgerson, Troy, and Piccirillo, Ciriaco. “The immunogenetics of immune dysregulation, polyendocrinopathy, enteropathy, X linked (IPEX) syndrome”. en. *Journal of Medical Genetics* 49.5 (2012), pp. 291–302.

[5] Bolton, Jennifer L. et al. “Genome Wide Association Identifies Common Variants at the SERPINA6/SERPINA1 Locus Influencing Plasma Cortisol and Corticosteroid Binding Globulin”. *PLOS Genetics* 10.7 (2014), e1004474.

[6] Hayes, M. Geoffrey et al. “Genome-wide association of polycystic ovary syndrome implicates alterations in gonadotropin secretion in European ancestry populations”. *Nature Communications* 6 (2015).

[7] Pott, Janne et al. “Genetic Association Study of Eight Steroid Hormones and Implications for Sexual Dimorphism of Coronary Artery Disease”. en. *The Journal of Clinical Endocrinology & Metabolism* 104.11 (2019), pp. 5008–5023.

[8] Crawford, Andrew A. et al. “Variation in the SERPINA6/SERPINA1 locus alters morning plasma cortisol, hepatic corticosteroid binding globulin expression, gene expression in peripheral tissues, and risk of cardiovascular disease”. en. *Journal of Human Genetics* (2021), pp. 1–12.

[9] Knowles, Michael R. and Drumm, Mitchell. “The Influence of Genetics on Cystic Fibrosis Phenotypes”. *Cold Spring Harbor Perspectives in Medicine* 2.12 (2012).

[10] Holmans, Peter A., Massey, Thomas H., and Jones, Lesley. “Genetic modifiers of Mendelian disease: Huntington’s disease and the trinucleotide repeat disorders”. en. *Human Molecular Genetics* 26.R2 (2017), R83–R90.

[11] Mackay, Trudy FC and Moore, Jason H. “Why epistasis is important for tackling complex human disease genetics”. *Genome Medicine* 6.6 (2014), p. 42.

[12] Nagrani, Rajini et al. “Common genetic variation in obesity, lipid transfer genes and risk of Metabolic Syndrome: Results from IDEFICS/I.Family study and meta-analysis”. en. *Scientific Reports* 10.1 (2020), p. 7189.

[13] Smushkin, Galina and Vella, Adrian. “Genetics of Type 2 Diabetes”. *Current opinion in clinical nutrition and metabolic care* 13.4 (2010), pp. 471–477.

[14] Purifoy, Frances E. “Endocrine-Environment Interaction in Human Variability”. *Annual Review of Anthropology* 10 (1981), pp. 141–162.

[15] Goodarzi, Mark O. “Genetics of Common Endocrine Disease: The Present and the Future”. en. *The Journal of Clinical Endocrinology & Metabolism* 101.3 (2016), pp. 787–794.

[16] Le, Phillip Phuc et al. “Glucocorticoid Receptor-Dependent Gene Regulatory Networks”. *PLOS Genetics* 1.2 (2005), e16.

[17] Whitworth, Judith A., Brown, Mark A., Kelly, John J., and Williamson, Paula M. “Mechanisms of cortisol-induced hypertension in humans”. *Steroids*. Aldosterone and Hypertension 60.1 (1995), pp. 76–80.

[18] Chiodini, Iacopo et al. “Cortisol Secretion in Patients With Type 2 Diabetes: Relationship with chronic complications”. en. *Diabetes Care* 30.1 (2007), pp. 83–88.

[19] Xiao, Rui and Boehnke, Michael. “Quantifying and correcting for the winner’s curse in genetic association studies”. *Genetic epidemiology* 33.5 (2009), pp. 453–462.

[20] Palmer, Cameron and Pe’er, Itsik. “Statistical correction of the Winner’s Curse explains replication variability in quantitative trait genome-wide association studies”. *PLoS Genetics* 13.7 (2017).

[21] Schadt, Eric E. “Molecular networks as sensors and drivers of common human diseases”. en. *Nature* 461.7261 (2009), nature08454.

[22] Humphries Steve E. “Common Variants for Cardiovascular Disease”. *Circulation* 135.22 (2017), pp. 2102–2105.

[23] Xue, Angli et al. “Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes”. en. *Nature Communications* 9.1 (2018), p. 2941.

[24] Solovieff, Nadia, Cotsapas, Chris, Lee, Phil H., Purcell, Shaun M., and Smoller, Jordan W. “Pleiotropy in complex traits: challenges and strategies”. *Nature reviews. Genetics* 14.7 (2013), pp. 483–495.

[25] Chesmore, Kevin, Bartlett, Jacquelaine, and Williams, Scott M. “The ubiquity of pleiotropy in human disease”. en. *Human Genetics* 137.1 (2018), pp. 39–44.

[26] Watanabe, Kyoko et al. “A global overview of pleiotropy and genetic architecture in complex traits”. en. *Nature Genetics* 51.9 (2019), pp. 1339–1348.

[27] Civelek, Mete and Lusis, Aldons J. “Systems genetics approaches to understand complex traits”. *Nature reviews. Genetics* 15.1 (2014), pp. 34–48.

[28] Sun, Wei. “A Statistical Framework for eQTL Mapping Using RNA-seq Data”. en. *Biometrics* 68.1 (2012), pp. 1–11.

[29] Battle, A.J. and Montgomery, S.B. “Determining causality and consequence of expression quantitative trait loci”. *Human genetics* 133.6 (2014), pp. 727–735.

[30] Rockman, Matthew V. “Reverse engineering the genotype–phenotype map with natural genetic variation”. en. *Nature* 456.7223 (2008), nature07633.

[31] Jansen, Ritsert C. “Studying complex biological systems using multifactorial perturbation”. en. *Nature Reviews Genetics* 4.2 (2003), pp. 145–151.

[32] Ferkingstad, Egil, Frigessi, Arnoldo, and Lyng, Heidi. “Indirect genomic effects on survival from gene expression data”. *Genome Biology* 9.3 (2008), R58.

[33] Petters, R. M. and Sommer, J. R. “Transgenic animals as models for human disease”. eng. *Transgenic Research* 9.4-5 (2000), 347–351, discussion 345–346.

[34] Moore, Bethany B. and Hogaboam, Cory M. “Murine models of pulmonary fibrosis”. *American Journal of Physiology-Lung Cellular and Molecular Physiology* 294.2 (2008), pp. L152–L160.

[35] Adamson, Kathryn Isabel, Sheridan, Eamonn, and Grierson, Andrew James. “Use of zebrafish models to investigate rare human disease”. en. *Journal of Medical Genetics* 55.10 (2018), pp. 641–649.

[36] Latvala, Antti and Ollikainen, Miina. “Mendelian randomization in (epi)genetic epidemiology: an effective tool to be handled with care”. *Genome Biology* 17 (2016).

[37] Porcu, Eleonora et al. “Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits”. en. *Nature Communications* 10.1 (2019), pp. 1–12.

[38] Harden, K. Paige and Klump, Kelly L. “Introduction to the Special Issue on Gene-Hormone Interplay”. *Behavior genetics* 45.3 (2015), pp. 263–267.

[39] Li, Yang, Tesson, Bruno M., Churchill, Gary A., and Jansen, Ritsert C. “Critical reasoning on causal inference in genome-wide linkage and association studies”. en. *Trends in Genetics* 26.12 (2010), pp. 493–498.

[40] Pingault, Jean-Baptiste et al. “Using genetic data to strengthen causal inference in observational research”. en. *Nature Reviews Genetics* 19.9 (2018), pp. 566–580.

[41] Chindelevitch, Leonid et al. “Causal reasoning on biological networks: interpreting transcriptional changes”. en. *Bioinformatics* 28.8 (2012), pp. 1114–1121.

[42] Slatkin, Montgomery. “Linkage disequilibrium — understanding the evolutionary past and mapping the medical future”. en. *Nature Reviews Genetics* 9.6 (2008), pp. 477–485.

[43] Joiret, Marc, Mahachie John, Jestinah M., Gusareva, Elena S., and Van Steen, Kristel. “Confounding of linkage disequilibrium patterns in large scale DNA based gene-gene interaction studies”. *BioData Mining* 12.1 (2019), p. 11.

[44] Buniello, Annalisa et al. “The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019”. en. *Nucleic Acids Research* 47.D1 (2019), pp. D1005–D1012.

[45] Boyle, Evan A., Li, Yang I., and Pritchard, Jonathan K. “An Expanded View of Complex Traits: From Polygenic to Omnigenic”. en. *Cell* 169.7 (2017), pp. 1177–1186.

[46] Bogardus, Clifton. “Missing Heritability and GWAS Utility”. en. *Obesity* 17.2 (2009), pp. 209–210.

[47] Clarke, Angus J. and Cooper, David N. “GWAS: heritability missing in action?” en. *European Journal of Human Genetics* 18.8 (2010), pp. 859–861.

[48] Makowsky, Robert et al. “Beyond Missing Heritability: Prediction of Complex Traits”. *PLoS Genetics* 7.4 (2011).

[49] Manolio, Teri A. et al. “Finding the missing heritability of complex diseases”. en. *Nature* 461.7265 (2009), pp. 747–753.

[50] Nicolae, Dan L. et al. “Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS”. en. *PLOS Genetics* 6.4 (2010), e1000888.

[51] Kogelman, Lisette J. A. et al. “An integrative systems genetics approach reveals potential causal genes and pathways related to obesity”. *Genome Medicine* 7.1 (2015), p. 105.

[52] Kirsten, Holger et al. “Dissecting the genetics of the human transcriptome identifies novel trait-related trans-eQTLs and corroborates the regulatory relevance of non-protein coding loci”. *Human Molecular Genetics* 24.16 (2015), pp. 4746–4763.

[53] Cesar, Aline S. M. et al. “Identification of putative regulatory regions and transcription factors associated with intramuscular fat content traits”. *BMC Genomics* 19 (2018).

[54] Brown, Christopher D., Mangravite, Lara M., and Engelhardt, Barbara E. “Integrative Modeling of eQTLs and Cis-Regulatory Elements Suggests Mechanisms Underlying Cell Type Specificity of eQTLs”. en. *PLOS Genetics* 9.8 (2013), e1003649.

[55] Albert, Frank Wolfgang, Bloom, Joshua S, Siegel, Jake, Day, Laura, and Kruglyak, Leonid. "Genetics of trans-regulatory variation in gene expression". *eLife* 7 ().

[56] Dixon, Anna L. et al. "A genome-wide association study of global gene expression". en. *Nature Genetics* 39.10 (2007), pp. 1202–1207.

[57] Yao, Chen et al. "Dynamic Role of trans Regulation of Gene Expression in Relation to Complex Traits". en. *The American Journal of Human Genetics* 100.4 (2017), pp. 571–580.

[58] Dijk, Erwin L. van, Auger, Hélène, Jaszczyszyn, Yan, and Thermes, Claude. "Ten years of next-generation sequencing technology". en. *Trends in Genetics* 30.9 (2014), pp. 418–426.

[59] Lonsdale, John et al. "The Genotype-Tissue Expression (GTEx) project". en. *Nature Genetics* 45.6 (2013), pp. 580–585.

[60] Kilpinen, Helena et al. "Common genetic variation drives molecular heterogeneity in human iPSCs". eng. *Nature* 546.7658 (2017), pp. 370–375.

[61] Lappalainen, Tuuli et al. "Transcriptome and genome sequencing uncovers functional variation in humans". en. *Nature* 501.7468 (2013), pp. 506–511.

[62] Chen, Lu et al. "Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells". en. *Cell* 167.5 (2016), 1398–1414.e24.

[63] Buil, Alfonso et al. "Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins". en. *Nature Genetics* 47.1 (2015), pp. 88–91.

[64] Franzén, Oscar et al. "Cardiometabolic risk loci share downstream cis- and trans-gene regulation across tissues and diseases". en. *Science* 353.6301 (2016), pp. 827–830.

[65] Võsa, Urmo et al. "Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis". en. *bioRxiv* (2018), p. 447367.

[66] Consortium, The GTEx. "The GTEx Consortium atlas of genetic regulatory effects across human tissues". en. *Science* 369.6509 (2020), pp. 1318–1330.

[67] Hormozdiari, Farhad et al. "Colocalization of GWAS and eQTL Signals Detects Target Genes". en. *The American Journal of Human Genetics* 99.6 (2016), pp. 1245–1260.

[68] Nica, Alexandra C. et al. "Candidate Causal Regulatory Effects by Integration of Expression QTLs with Complex Trait Genetic Associations". en. *PLOS Genetics* 6.4 (2010), e1000895.

[69] Wu, Ying et al. "Colocalization of GWAS and eQTL signals at loci with multiple signals identifies additional candidate genes for body fat distribution". en. *Human Molecular Genetics* 28.24 (2019), pp. 4161–4172.

[70] Mills, Melinda C. and Rahal, Charles. "A scientometric review of genome-wide association studies". en. *Communications Biology* 2.1 (2019), pp. 1–11.

[71] Suhre, Karsten, McCarthy, Mark I., and Schwenk, Jochen M. "Genetics meets proteomics: perspectives for large population-based studies". en. *Nature Reviews Genetics* 22.1 (2021), pp. 19–37.

[72] Yao, Chen et al. "Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease". en. *Nature Communications* 9.1 (2018), pp. 1–11.

[73] Hinke, Stephanie von, Davey Smith, George, Lawlor, Debbie A., Propper, Carol, and Windmeijer, Frank. "Genetic markers as instrumental variables". *Journal of Health Economics* 45 (2016), pp. 131–148.

[74] Glymour, M. Maria, Tchetgen Tchetgen, Eric J., and Robins, James M. "Credible Mendelian Randomization Studies: Approaches for Evaluating the Instrumental Variable Assumptions". *American Journal of Epidemiology* 175.4 (2012), pp. 332–339.

[75] Lousdal, Mette Lise. “An introduction to instrumental variable assumptions, validation and estimation”. *Emerging Themes in Epidemiology* 15 (2018).

[76] Neto, Elias Chaibub, Keller, Mark P., Attie, Alan D., and Yandell, Brian S. “CAUSAL GRAPHICAL MODELS IN SYSTEMS GENETICS: A UNIFIED FRAMEWORK FOR JOINT INFERENCE OF CAUSAL NETWORK AND GENETIC ARCHITECTURE FOR CORRELATED PHENOTYPES”. *The annals of applied statistics* 4.1 (2010), pp. 320–339.

[77] Neumeyer, Sonja, Hemani, Gibran, and Zeggini, Eleftheria. “Strengthening Causal Inference for Complex Disease Using Molecular Quantitative Trait Loci”. English. *Trends in Molecular Medicine* 0.0 (2019).

[78] Hutton, Joseph et al. “Mediation analysis to understand genetic relationships between habitual coffee intake and gout”. *Arthritis Research & Therapy* 20.1 (2018), p. 135.

[79] Howe, Laura D et al. “Relationship between mediation analysis and the structured life course approach”. *International Journal of Epidemiology* 45.4 (2016), pp. 1280–1294.

[80] Broekema, R. V., Bakker, O. B., and Jonkers, I. H. “A practical view of fine-mapping and gene prioritization in the post-genome-wide association era”. *Open Biology* 10.1 (2020), p. 190221.

[81] Kumasaka, Natsuhiko, Knights, Andrew J., and Gaffney, Daniel J. “Fine-mapping cellular QTLs with RASQUAL and ATAC-seq”. en. *Nature Genetics* 48.2 (2016), pp. 206–213.

[82] Giambartolomei, Claudia et al. “Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics”. en. *PLOS Genetics* 10.5 (2014), e1004383.

[83] Jansen, Ritsert C. and Nap, Jan-Peter. “Genetical genomics: the added value from segregation”. en. *Trends in Genetics* 17.7 (2001), pp. 388–391.

[84] Dam, Sipko van, Võsa, Urmo, Graaf, Adriaan van der, Franke, Lude, and Magalhães, João Pedro de. “Gene co-expression analysis for functional classification and gene–disease predictions”. en. *Briefings in Bioinformatics* 19.4 (2018), pp. 575–592.

[85] Butte, Atul J. and Kohane, Isaac S. “Unsupervised knowledge discovery in medical databases using relevance networks”. *Proceedings. AMIA Symposium* (1999), pp. 711–715.

[86] Gerring, Zachary F., Gamazon, Eric R., Derks, Eske M., and Consortium, for the Major Depressive Disorder Working Group of the Psychiatric Genomics. “A gene co-expression network-based analysis of multiple brain tissues reveals novel genes and molecular pathways underlying major depression”. en. *PLOS Genetics* 15.7 (2019), e1008245.

[87] Mukund, Kavitha and Subramaniam, Shankar. “Co-expression Network Approach Reveals Functional Similarities among Diseases Affecting Human Skeletal Muscle”. English. *Frontiers in Physiology* 8 (2017).

[88] Talukdar, Husain A. et al. “Cross-Tissue Regulatory Gene Networks in Coronary Artery Disease”. *Cell Systems* 2.3 (2016), pp. 196–208.

[89] Bakhtiarizadeh, Mohammad Reza, Hosseinpour, Batool, Shahhoseini, Maryam, Korte, Arthur, and Gifani, Peyman. “Weighted Gene Co-expression Network Analysis of Endometriosis and Identification of Functional Modules Associated With Its Main Hallmarks”. English. *Frontiers in Genetics* 9 (2018).

[90] Millstein, Joshua, Zhang, Bin, Zhu, Jun, and Schadt, Eric E. “Disentangling molecular relationships with a causal inference test”. *BMC Genetics* 10.1 (2009), p. 23.

[91] Meinshausen, Nicolai et al. “Methods for causal inference from gene perturbation experiments and validation”. en. *Proceedings of the National Academy of Sciences* 113.27 (2016), pp. 7361–7368.

[92] Wang, Lingfei and Michoel, Tom. “Efficient and accurate causal inference with hidden confounders from genome-transcriptome variation data”. *PLOS Computational Biology* 13.8 (2017), e1005703.

[93] Friedman, Nir, Linial, Michal, Nachman, Iftach, and Pe’er, Dana. “Using Bayesian Networks to Analyze Expression Data”. *Journal of Computational Biology* 7.3-4 (2000), pp. 601–620.

[94] Werhli, Adriano V. and Husmeier, Dirk. “Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge”. eng. *Statistical Applications in Genetics and Molecular Biology* 6 (2007), Article15.

[95] Pearl, Judea. *Causality*. Cambridge: Cambridge University Press, 2009.

[96] Wittkopp, Patricia J. and Kalay, Gizem. “Cis -regulatory elements: molecular mechanisms and evolutionary processes underlying divergence”. en. *Nature Reviews Genetics* 13.1 (2012), pp. 59–69.

[97] Prestridge, Dan S. “Predicting Pol II Promoter Sequences using Transcription Factor Binding Sites”. en. *Journal of Molecular Biology* 249.5 (1995), pp. 923–932.

[98] Schadt, Eric E. et al. “Mapping the Genetic Architecture of Gene Expression in Human Liver”. en. *PLOS Biology* 6.5 (2008), e107.

[99] Agrahari, Rupesh et al. “Applications of Bayesian network models in predicting types of hematological malignancies”. en. *Scientific Reports* 8.1 (2018), p. 6951.

[100] Wilkinson, Darren J. “Bayesian methods in bioinformatics and computational systems biology”. en. *Briefings in Bioinformatics* 8.2 (2007), pp. 109–116.

[101] Wang, Lingfei, Audenaert, Pieter, and Michoel, Tom. “High-Dimensional Bayesian Network Inference From Systems Genetics Data Using Genetic Node Ordering”. English. *Frontiers in Genetics* 10 (2019).

[102] Horning, E. C. “Gas Phase Analytical Methods for the Study of Steroid Hormones and their Metabolites”. en. *Gas Phase Chromatography of Steroids*. Ed. by Eik-Nes, K. B. and Horning, E. C. Monographs on Endocrinology. Berlin, Heidelberg: Springer, 1968, pp. 1–71.

[103] Andrew, Ruth and Homer, Natalie. “Mass spectrometry: future opportunities for profiling and imaging steroids and steroid metabolites”. en. *Current Opinion in Endocrine and Metabolic Research* (2020).

[104] Gallois, Apolline et al. “A comprehensive study of metabolite genetics reveals strong pleiotropy and heterogeneity across time and context”. en. *Nature Communications* 10.1 (2019), p. 4788.

[105] Small, Kerrin S. et al. “Regulatory variants at KLF14 influence type 2 diabetes risk via a female-specific effect on adipocyte size and body composition”. en. *Nature Genetics* 50.4 (2018), pp. 572–580.

[106] Seldin, Marcus M. et al. “A Strategy for Discovery of Endocrine Interactions with Application to Whole-Body Metabolism”. en. *Cell Metabolism* 27.5 (2018), 1138–1155.e6.

[107] Li, Zhonggang et al. “Integrating Mouse and Human Genetic Data to Move beyond GWAS and Identify Causal Genes in Cholesterol Metabolism”. en. *Cell Metabolism* 31.4 (2020), 741–754.e5.

[108] Whitehouse, Barbara J. and Vinson, G. P. “Pathway for Cortisol Biosynthesis in the Foetal Adrenal Cortex”. en. *Nature* 221.5185 (1969), pp. 1051–1052.

[109] Pariante, Carmine M. and Lightman, Stafford L. “The HPA axis in major depression: classical theories and new developments”. *Trends in Neurosciences* 31.9 (2008), pp. 464–468.

[110] Lightman, Stafford L, Birnie, Matthew T, and Conway-Campbell, Becky L. “Dynamics of ACTH and Cortisol Secretion and Implications for Disease”. *Endocrine Reviews* 41.3 (2020), pp. 470–490.

[111] WEITZMAN, ELLIOT D. et al. “Twenty-four Hour Pattern of the Episodic Secretion of Cortisol in Normal Subjects”. *The Journal of Clinical Endocrinology & Metabolism* 33.1 (1971), pp. 14–22.

[112] Krieger, Dorothy T., Allen, William, Rizzo, Frank, and Krieger, Howard P. “Characterization of the Normal Temporal Pattern of Plasma Corticosteroid Levels”. en. *The Journal of Clinical Endocrinology & Metabolism* 32.2 (1971), pp. 266–284.

[113] Russell, G M, Kalafatakis, K, and Lightman, S L. “The Importance of Biological Oscillators for Hypothalamic-Pituitary-Adrenal Activity and Tissue Glucocorticoid Response: Coordinating Stress and Neurobehavioural Adaptation”. *Journal of Neuroendocrinology* 27.6 (2015), pp. 378–388.

[114] Kalafatakis, K. et al. “Ultradian rhythmicity of plasma cortisol is necessary for normal emotional and cognitive responses in man”. en. *Proceedings of the National Academy of Sciences* 115.17 (2018), E4091–E4100.

[115] Walker, Brian R., Campbell, Jill C., Fraser, R., Stewart, Paul M., and Edwards, Christopher R. W. “Mineralocorticoid excess and inhibition of 11 β-hydroxysteroid dehydrogenase in patients with ectopic ACTH syndrome”. en. *Clinical Endocrinology* 37.6 (1992), pp. 483–492.

[116] Tomlinson, Jeremy W. and Stewart, Paul M. “Cortisol metabolism and the role of 11β-hydroxysteroid dehydrogenase”. en. *Best Practice & Research Clinical Endocrinology & Metabolism* 15.1 (2001), pp. 61–78.

[117] Nixon, Mark et al. “ABCC1 confers tissue-specific sensitivity to cortisol versus corticosterone: A rationale for safer glucocorticoid replacement therapy”. en. *Science Translational Medicine* 8.352 (2016), 352ra109–352ra109.

[118] Denver, Robert John. “Structural and Functional Evolution of Vertebrate Neuroendocrine Stress Systems”. en. *Annals of the New York Academy of Sciences* 1163.1 (2009), pp. 1–16.

[119] Coutinho, Agnes E. and Chapman, Karen E. “The anti-inflammatory and immunosuppressive effects of glucocorticoids, recent developments and mechanistic insights”. *Molecular and Cellular Endocrinology* 335.1 (2011), pp. 2–13.

[120] Opherk, Christian et al. “Inactivation of the Glucocorticoid Receptor in Hepatocytes Leads to Fasting Hypoglycemia and Ameliorates Hyperglycemia in Streptozotocin-Induced Diabetes Mellitus”. *Molecular Endocrinology* 18.6 (2004), pp. 1346–1353.

[121] Guia, Roldan M. de and Herzig, Stephan. “How Do Glucocorticoids Regulate Lipid Metabolism?” en. *Glucocorticoid Signaling: From Molecules to Mice to Man*. Ed. by Wang, Jen-Chywan and Harris, Charles. Advances in Experimental Medicine and Biology. New York, NY: Springer, 2015, pp. 127–144.

[122] Zhou, Jie et al. “Glucocorticoid Regulation of Natural Cytotoxicity: Effects of Cortisol on the Phenotype and Function of a Cloned Human Natural Killer Cell Line”. en. *Cellular Immunology* 178.2 (1997), pp. 108–116.

[123] Bjelaković, G. et al. “Glucocorticoids and oxidative stress”. eng. *Journal of Basic and Clinical Physiology and Pharmacology* 18.2 (2007), pp. 115–127.

[124] Vandevyver, Sofie, Dejager, Lien, and Libert, Claude. “Comprehensive Overview of the Structure and Regulation of the Glucocorticoid Receptor”. *Endocrine Reviews* 35.4 (2014), pp. 671–693.

[125] Xavier, Andre Machado, Anunciato, Aparecida Kataryna Olimpio, Rosenstock, Tatiana Rosado, and Glezer, Isaias. “Gene Expression Control by Glucocorticoid Receptors during Innate Immune Responses”. English. *Frontiers in Endocrinology* 7 (2016).

[126] Lim, Hee-Woong et al. “Genomic redistribution of GR monomers and dimers mediates transcriptional response to exogenous glucocorticoid in vivo”. *Genome Research* 25.6 (2015), pp. 836–844.

[127] Orchinik, M, Murray, T F, Franklin, P H, and Moore, F L. “Guanyl nucleotides modulate binding to steroid receptors in neuronal membranes.” *Proceedings of the National Academy of Sciences of the United States of America* 89.9 (1992), pp. 3830–3834.

[128] Whitfield, G. Kerr, Jurutka, Peter W., Haussler, Carol A., and Haussler, Mark R. “Steroid hormone receptors: Evolution, ligands, and molecular basis of biologic function”. en. *Journal of Cellular Biochemistry* 75.S32 (1999), pp. 110–122.

[129] Gomez-Sanchez, Elise and Gomez-Sanchez, Celso E. “The Multifaceted Mineralocorticoid Receptor”. *Comprehensive Physiology* 4.3 (2014), pp. 965–994.

[130] Medina, Adriana et al. “Glucocorticoid and Mineralocorticoid Receptor Expression in the Human Hippocampus in Major Depressive Disorder”. *Journal of psychiatric research* 47.3 (2013), pp. 307–314.

[131] Trapp, T., Rupprecht, R., Castrén, M., Reul, J. M., and Holsboer, F. “Heterodimerization between mineralocorticoid and glucocorticoid receptor: a new principle of glucocorticoid action in the CNS”. eng. *Neuron* 13.6 (1994), pp. 1457–1462.

[132] Walker, Brian R. “Glucocorticoids and cardiovascular disease”. eng. *European Journal of Endocrinology* 157.5 (2007), pp. 545–559.

[133] Reynolds, Rebecca M. et al. “Skeletal Muscle Glucocorticoid Receptor Density and Insulin Resistance”. *JAMA* 287.19 (2002), pp. 2505–2506.

[134] Wake, Deborah J. et al. “Local and Systemic Impact of Transcriptional Up-Regulation of 11beta-Hydroxysteroid Dehydrogenase Type 1 in Adipose Tissue in Human Obesity”. *The Journal of Clinical Endocrinology & Metabolism* 88.8 (2003), pp. 3983–3988.

[135] Raff, Hershel and Carroll, Ty. “Cushing’s syndrome: from physiological principles to diagnosis and clinical care”. *The Journal of Physiology* 593.Pt 3 (2015), pp. 493–506.

[136] Hahner, Stefanie et al. “Adrenal insufficiency”. en. *Nature Reviews Disease Primers* 7.1 (2021), pp. 1–24.

[137] Souverein, P C et al. “Use of oral glucocorticoids and risk of cardiovascular and cerebrovascular disease in a population based case–control study”. *Heart* 90.8 (2004), pp. 859–865.

[138] Reynolds, Rebecca M. et al. “Altered Control of Cortisol Secretion in Adult Men with Low Birth Weight and Cardiovascular Risk Factors”. en. *The Journal of Clinical Endocrinology & Metabolism* 86.1 (2001), pp. 245–250.

[139] Smith, George Davey et al. “Cortisol, Testosterone, and Coronary Heart Disease”. *Circulation* 112.3 (2005), pp. 332–340.

[140] Crawford, Andrew A et al. “Morning plasma cortisol as a cardiovascular risk factor: findings from prospective cohort and Mendelian randomization studies”. *European Journal of Endocrinology* 181.4 (2019), pp. 429–438.

[141] Bartels, Meike, Geus, Eco J. C. de, Kirschbaum, Clemens, Sluyter, Frans, and Boomsma, Dorret I. “Heritability of Daytime Cortisol Levels in Children”. en. *Behavior Genetics* 33.4 (2003), pp. 421–433.

[142] Li, Qiliang, Peterson, Kenneth R., Fang, Xiangdong, and Stamatoyannopoulos, George. "Locus control regions". *Blood* 100.9 (2002), pp. 3077–3086.

[143] Zhao, Hui, Friedman, Richard D., and Fournier, R. E. K. "The Locus Control Region Activates Serpin Gene Expression through Recruitment of Liver-Specific Transcription Factors and RNA Polymerase II". *Molecular and Cellular Biology* 27.15 (2007), pp. 5286–5295.

[144] Baxter, Euan W., Cummings, W. Jason, and Fournier, R. E. K. "Formation of a large, complex domain of histone hyperacetylation at human 14q32.1 requires the serpin locus control region". *Nucleic Acids Research* 33.10 (2005), pp. 3313–3322.

[145] Hammond, Geoffrey L. "Molecular Properties of Corticosteroid Binding Globulin and the Sex-Steroid Binding Proteins*". *Endocrine Reviews* 11.1 (1990), pp. 65–79.

[146] Perogamvros, Ilias, Ray, David W., and Trainer, Peter J. "Regulation of cortisol bioavailability—effects on hormone measurement and action". en. *Nature Reviews Endocrinology* 8.12 (2012), pp. 717–727.

[147] Lewis, John G., Bagley, Christopher J., Elder, Peter A., Bachmann, Anthony W., and Torpy, David J. "Plasma free cortisol fraction reflects levels of functioning corticosteroid-binding globulin". en. *Clinica Chimica Acta* 359.1 (2005), pp. 189–194.

[148] Chan, Wee Lee, Carrell, Robin W., Zhou, Aiwu, and Read, Randy J. "How Changes in Affinity of Corticosteroid-binding Globulin Modulate Free Cortisol Concentration". *The Journal of Clinical Endocrinology and Metabolism* 98.8 (2013), pp. 3315–3322.

[149] Nenke, Marni A., Holmes, Mark, Rankin, Wayne, Lewis, John G., and Torpy, David J. "Corticosteroid-binding globulin cleavage is paradoxically reduced in alpha-1 antitrypsin deficiency: Implications for cortisol homeostasis". en. *Clinica Chimica Acta* 452 (2016), pp. 27–31.

[150] Hammond, G L et al. "Primary structure of human corticosteroid binding globulin, deduced from hepatic and pulmonary cDNAs, exhibits homology with serine protease inhibitors." *Proceedings of the National Academy of Sciences of the United States of America* 84.15 (1987), pp. 5153–5157.

[151] Pemberton, P. A., Stein, P. E., Pepys, M. B., Potter, J. M., and Carrell, R. W. "Hormone binding globulins undergo serpin conformational change in inflammation". en. *Nature* 336.6196 (1988), pp. 257–258.

[152] Moisan, Marie-Pierre. "CBG: a cortisol reservoir rather than a transporter". en. *Nature Reviews Endocrinology* 9.2 (2013), pp. 78–78.

[153] Doe, Richard P., Lohrenz, Francis N., and Seal, Ulysses S. "Familial decrease in corticosteroid-binding globulin". en. *Metabolism* 14.8 (1965), pp. 940–943.

[154] Torpy, David J. et al. "Familial Corticosteroid-Binding Globulin Deficiency Due to a Novel Null Mutation: Association with Fatigue and Relative Hypotension". en. *The Journal of Clinical Endocrinology & Metabolism* 86.8 (2001), pp. 3692–3700.

[155] Buss, C. et al. "Haploinsufficiency of the SERPINA6 gene is associated with severe muscle fatigue: A de novo mutation in corticosteroid-binding globulin deficiency". en. *Journal of Neural Transmission* 114.5 (2007), pp. 563–569.

[156] Simard, Marc, Hill, Lesley A., Lewis, John G., and Hammond, Geoffrey L. "Naturally Occurring Mutations of Human Corticosteroid-Binding Globulin". *The Journal of Clinical Endocrinology & Metabolism* 100.1 (2015), E129–E139.

[157] Petersen, Helle Heibroch et al. "Hyporesponsiveness to Glucocorticoids in Mice Genetically Deficient for the Corticosteroid Binding Globulin". *Molecular and Cellular Biology* 26.19 (2006), pp. 7236–7245.

[158] Smith, C. L. and Hammond, G. L. “An amino acid substitution in biobreeding rat corticosteroid binding globulin results in reduced steroid binding affinity.” en. *Journal of Biological Chemistry* 266.28 (1991), pp. 18555–18559.

[159] Hägg, Sara et al. “Multi-Organ Expression Profiling Uncovers a Gene Module in Coronary Artery Disease Involving Transendothelial Migration of Leukocytes and LIM Domain Binding 2: The Stockholm Atherosclerosis Gene Expression (STAGE) Study”. *PLOS Genetics* 5.12 (2009), e1000754.

[160] Foroughi Asl Hassan et al. “Expression Quantitative Trait Loci Acting Across Multiple Tissues Are Enriched in Inherited Risk for Coronary Artery Disease”. *Circulation: Cardiovascular Genetics* 8.2 (2015), pp. 305–315.

[161] Qi, Jianlong, Asl, Hassan Foroughi, Björkegren, Johan, and Michoel, Tom. “kruX: matrix-based non-parametric eQTL discovery”. *BMC Bioinformatics* 15 (2014), p. 11.

[162] Linck, Ethan B. and Battey, C. J. “Minor allele frequency thresholds strongly affect population structure inference with genomic datasets”. en. *bioRxiv* (2017), p. 188623.

[163] Storey, John D. and Tibshirani, Robert. “Statistical significance for genomewide studies”. en. *Proceedings of the National Academy of Sciences* 100.16 (2003), pp. 9440–9445.

[164] Shabalin, Andrey A. “Matrix eQTL: ultra fast eQTL analysis via large matrix operations”. *Bioinformatics* 28.10 (2012), pp. 1353–1358.

[165] Pruim, Randall J. et al. “LocusZoom: regional visualization of genome-wide association scan results”. en. *Bioinformatics* 26.18 (2010), pp. 2336–2337.

[166] Machiela, Mitchell J. and Chanock, Stephen J. “LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants”. en. *Bioinformatics* 31.21 (2015), pp. 3555–3557.

[167] Chen, Lin S., Emmert-Streib, Frank, and Storey, John D. “Harnessing naturally randomized transcription to infer regulatory relationships among genes”. *Genome Biology* 8.10 (2007), R219.

[168] Pedregosa, Fabian et al. “Scikit-learn: Machine Learning in Python”. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[169] Huang, Da Wei, Sherman, Brad T., and Lempicki, Richard A. “Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources”. en. *Nature Protocols* 4.1 (2009), pp. 44–57.

[170] Kwong, Alan et al. “FIVEx: an interactive eQTL browser across public datasets”. *Bioinformatics* btab614 (2021).

[171] Balding, David J. “A tutorial on statistical methods for population association studies”. en. *Nature Reviews Genetics* 7.10 (2006), pp. 781–791.

[172] Guo, Xingyi et al. “A Comprehensive cis-eQTL Analysis Revealed Target Genes in Breast Cancer Susceptibility Loci Identified in Genome-wide Association Studies”. *The American Journal of Human Genetics* 102.5 (2018), pp. 890–903.

[173] Finnerty, Celeste C., Mabvuure, Nigel Tapiwa, Ali, Arham, Kozar, Rosemary A., and Herndon, David N. “The Surgically Induced Stress Response”. *JPEN. Journal of parenteral and enteral nutrition* 37.5 0 (2013), 21S–29S.

[174] Boonen, Eva et al. “Reduced Cortisol Metabolism during Critical Illness”. *New England Journal of Medicine* 368.16 (2013), pp. 1477–1488.

[175] Monder, Carl and Coufalik, Alena. “Influence of Cortisol on Glycogen Synthesis and Gluconeogenesis in Fetal Rat Liver in Organ Culture”. en. *Journal of Biological Chemistry* 247.11 (1972), pp. 3608–3617.

[176] Pavlatou, Maria G. et al. “Circulating cortisol-associated signature of glucocorticoid-related gene expression in subcutaneous fat of obese subjects”. en. *Obesity* 21.5 (2013), pp. 960–967.

[177] Nieman, Lynnette K. and Ilias, Ioannis. “Evaluation and treatment of Cushing’s syndrome”. *The American Journal of Medicine* 118.12 (2005), pp. 1340–1346.

[178] Lopez, Juan Pablo et al. “Single-cell molecular profiling of all three components of the HPA axis reveals adrenal ABCB1 as a regulator of stress adaptation”. eng. *Science Advances* 7.5 (2021), eabe4497.

[179] Præstholm, Stine M., Correia, Catarina M., and Grøntved, Lars. “Multifaceted Control of GR Signaling and Its Impact on Hepatic Transcriptional Networks and Metabolism”. English. *Frontiers in Endocrinology* 11 (2020).

[180] Lee, Rebecca A., Harris, Charles A., and Wang, Jen-Chywan. “Glucocorticoid Receptor and Adipocyte Biology”. *Nuclear receptor research* 5 (2018).

[181] Mundade, Rasika, Ozer, Hatice Gulcin, Wei, Han, Prabhu, Lakshmi, and Lu, Tao. “Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond”. *Cell Cycle* 13.18 (2014), pp. 2847–2852.

[182] Djebali, Sarah et al. “Landscape of transcription in human cells”. en. *Nature* 489.7414 (2012), pp. 101–108.

[183] Lachmann, Alexander et al. “ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments”. *Bioinformatics* 26.19 (2010), pp. 2438–2444.

[184] Scheschowitsch, Karin, Leite, Jacqueline Alves, and Assreuy, Jamil. “New insights in glucocorticoid receptor signaling – more than just a ligand binding receptor”. English. *Frontiers in Endocrinology* 8 (2017).

[185] Yu, Chi-Yi et al. “Genome-Wide Analysis of Glucocorticoid Receptor Binding Regions in Adipocytes Reveal Gene Network Involved in Triglyceride Homeostasis”. en. *PLOS ONE* 5.12 (2010), e15188.

[186] Koning, Anne-Sophie C A M, Buurstede, Jacobus C, Weert, Lisa T C M van, and Meijer, Onno C. “Glucocorticoid and Mineralocorticoid Receptors in the Brain: A Transcriptional Perspective”. *Journal of the Endocrine Society* 3.10 (2019), pp. 1917–1930.

[187] Cusanovich, Darren A., Pavlovic, Bryan, Pritchard, Jonathan K., and Gilad, Yoav. “The Functional Consequences of Variation in Transcription Factor Binding”. en. *PLOS Genetics* 10.3 (2014), e1004226.

[188] Ludl, Adriaan-Alexander and Michoel, Tom. “Comparison between instrumental variable and mediation-based methods for reconstructing causal gene networks in yeast”. en. *Molecular Omics* 17.2 (2021), pp. 241–251.

[189] Schadt, Eric E. et al. “An integrative genomics approach to infer causal associations between gene expression and disease”. eng. *Nature Genetics* 37.7 (2005), pp. 710–717.

[190] Tong, Pin, Monahan, Jack, and Prendergast, James G. D. “Shared regulatory sites are abundant in the human genome and shed light on genome evolution and disease pleiotropy”. en. *PLOS Genetics* 13.3 (2017), e1006673.

[191] Matys, V. et al. “TRANSFAC ® : transcriptional regulation, from patterns to profiles”. *Nucleic Acids Research* 31.1 (2003), pp. 374–378.

[192] DeLigio, James T., Lin, Grace, Chalfant, Charles E., and Park, Margaret A. “Splice variants of cytosolic polyadenylation element–binding protein 2 (CPEB2) differentially regulate pathways linked to cancer metastasis”. *The Journal of Biological Chemistry* 292.43 (2017), pp. 17909–17918.

[193] Ishikawa, Kosuke, Tamamura, Sakura, Semba, Kentaro, and Watanabe, Shinya. "Establishment of reporter cells that respond to glucocorticoids by a transposon-mediated promoter-trapping system". en. *European Journal of Pharmaceutical Sciences* (2021), p. 105819.

[194] Kolvenbach, Caroline M. et al. "Rare Variants in BNC2 Are Implicated in Autosomal-Dominant Congenital Lower Urinary-Tract Obstruction". en. *The American Journal of Human Genetics* 104.5 (2019), pp. 994–1006.

[195] Kamijo, T., Aoyama, T., Komiyama, A., and Hashimoto, T. "Structural Analysis of cDNAs for Subunits of Human Mitochondrial Fatty Acid β-Oxidation Trifunctional Protein". en. *Biochemical and Biophysical Research Communications* 199.2 (1994), pp. 818–825.

[196] IJlst, Lodewijk, Wanders, Ronald J. A., Ushikubo, Sciichi, Kamijo, Takehiko, and Hashimoto, Takashi. "Molecular basis of long-chain 3-hydroxyacyl-CoA dehydrogenase deficiency: identification of the major disease-causing mutation in the α-subunit of the mitochondrial trifunctional protein". en. *Biochimica et Biophysica Acta (BBA) - Lipids and Lipid Metabolism* 1215.3 (1994), pp. 347–350.

[197] Sims, H. F. et al. "The molecular basis of pediatric long chain 3-hydroxyacyl-CoA dehydrogenase deficiency associated with maternal acute fatty liver of pregnancy". en. *Proceedings of the National Academy of Sciences* 92.3 (1995), pp. 841–845.

[198] Arshad, Muhammad et al. "RNF13, a RING Finger Protein, Mediates Endoplasmic Reticulum Stress-induced Apoptosis through the Inositol-requiring Enzyme (IRE1α)/c-Jun NH2-terminal Kinase Pathway". *The Journal of Biological Chemistry* 288.12 (2013), pp. 8726–8736.

[199] Harada, H. et al. "Anti-oncogenic and oncogenic potentials of interferon regulatory factors-1 and -2". en. *Science* 259.5097 (1993), pp. 971–974.

[200] Chapin, William J. et al. "Peripheral blood IRF1 expression as a marker for glucocorticoid sensitivity". *Pharmacogenetics and genomics* 25.3 (2015), pp. 126–133.

[201] Shang Ming-Mei et al. "Lim Domain Binding 2". *Arteriosclerosis, Thrombosis, and Vascular Biology* 34.9 (2014), pp. 2068–2077.

[202] Bujalska, I. J. et al. "Expression profiling of 11beta-hydroxysteroid dehydrogenase type-1 and glucocorticoid-target genes in subcutaneous and omental human preadipocytes". eng. *Journal of Molecular Endocrinology* 37.2 (2006), pp. 327–340.

[203] Etzerodt, Anders and Moestrup, Søren K. "CD163 and Inflammation: Biological, Diagnostic, and Therapeutic Aspects". *Antioxidants & Redox Signaling* 18.17 (2013), pp. 2352–2363.

[204] Gao, Ge et al. "The Role of RBM25/LUC7L3 in Abnormal Cardiac Sodium Channel Splicing Regulation in Human Heart Failure". *Circulation* 124.10 (2011), pp. 1124–1131.

[205] Li, Yuan et al. "LUC7L3/CROP inhibits replication of hepatitis B virus via suppressing enhancer II/basal core promoter activity". en. *Scientific Reports* 6.1 (2016), p. 36741.

[206] Kang, Bit Na, Jude, Joseph A., Panettieri, Reynold A., Walseth, Timothy F., and Kannan, Mathur S. "Glucocorticoid regulation of CD38 expression in human airway smooth muscle cells: role of dual specificity phosphatase 1". *American Journal of Physiology-Lung Cellular and Molecular Physiology* 295.1 (2008), pp. L186–L193.

[207] Moeller, Jesper B. et al. "CD163-L1 Is an Endocytic Macrophage Protein Strongly Regulated by Mediators in the Inflammatory Response". en. *The Journal of Immunology* 188.5 (2012), pp. 2399–2409.

[208] Friedman, J. E. “Role of glucocorticoids in activation of hepatic PEPCK gene transcription during exercise”. *American Journal of Physiology-Endocrinology and Metabolism* 266.4 (1994), E560–E566.

[209] Krzysiek, R. “Role of Glucocorticoid-Induced Leucine Zipper (GILZ) Expression by Dendritic Cells in Tolerance Induction”. en. *Transplantation Proceedings* 42.8 (2010), pp. 3331–3332.

[210] Fried, S K, Russell, C D, Grauso, N L, and Brolin, R E. “Lipoprotein lipase regulation by insulin and glucocorticoid in subcutaneous and omental adipose tissues of obese women and men.” *Journal of Clinical Investigation* 92.5 (1993), pp. 2191–2198.

[211] Reddy, Timothy E., Gertz, Jason, Crawford, Gregory E., Garabedian, Michael J., and Myers, Richard M. “The Hypersensitive Glucocorticoid Response Specifically Regulates Period 1 and Expression of Circadian Genes”. *Molecular and Cellular Biology* 32.18 (2012), pp. 3756–3767.

[212] Pascual, Rosa et al. “The RNA binding protein CPEB2 regulates hormone sensing in mammary gland development and luminal breast cancer”. en. *Science Advances* 6.20 (2020), eaax3868.

[213] Ouni, Ikram, Flick, Karin, and Kaiser, Peter. “Ubiquitin and transcription”. *Transcription* 2.3 (2011), pp. 135–139.

[214] Cui, Huachun, Banerjee, Sami, Guo, Sijia, Xie, Na, and Liu, Gang. “Interferon regulatory factor 2 inhibits expression of glycolytic genes and lipopolysaccharide induced pro-inflammatory responses in macrophages”. *Journal of immunology (Baltimore, Md. : 1950)* 200.9 (2018), pp. 3218–3230.

[215] Mercado, Nicolas et al. “IRF2 is a master regulator of human keratinocyte stem cell fate”. en. *Nature Communications* 10.1 (2019), p. 4676.

[216] Sun, Hongduo et al. “Quantitative integration of epigenomic variation and transcription factor binding using MAmotif toolkit identifies an important role of IRF2 as transcription activator at gene promoters”. en. *Cell Discovery* 4.1 (2018), pp. 1–4.

[217] Mangan, S. and Alon, U. “Structure and function of the feed-forward loop network motif”. en. *Proceedings of the National Academy of Sciences* 100.21 (2003), pp. 11980–11985.

[218] Chen, Sheng-hong, Masuno, Kiriko, Cooper, Samantha B., and Yamamoto, Keith R. “Incoherent feed-forward regulatory logic underpinning glucocorticoid receptor action”. en. *Proceedings of the National Academy of Sciences* 110.5 (2013), pp. 1964–1969.

[219] Sasse, Sarah K. et al. “The Glucocorticoid Receptor and KLF15 Regulate Gene Expression Dynamics and Integrate Signals through Feed-Forward Circuitry”. en. *Molecular and Cellular Biology* 33.11 (2013), pp. 2104–2115.

[220] Castaldi, Peter J. et al. “Screening for interaction effects in gene expression data”. *PLoS ONE* 12.3 (2017), e0173847.

[221] Qu, Meng, Duffy, Tomas, Hirota, Tsuyoshi, and Kay, Steve A. “Nuclear receptor HNF4A transrepresses CLOCK:BMAL1 and modulates tissue-specific circadian networks”. en. *Proceedings of the National Academy of Sciences* 115.52 (2018), E12305–E12312.

[222] Qu, Meng, Qu, Han, Jia, Zhenyu, and Kay, Steve A. “HNF4A defines tissue-specific circadian rhythms by beaconing BMAL1::CLOCK chromatin binding and shaping the rhythmic chromatin landscape”. en. *Nature Communications* 12.1 (2021), p. 6350.

[223] Liu, Xuanyao, Li, Yang I., and Pritchard, Jonathan K. “Trans Effects on Gene Expression Can Drive Omnigenic Inheritance”. en. *Cell* 177.4 (2019), 1022–1034.e6.

[224] Fehrmann, Rudolf S. N. et al. “Trans-eQTLs Reveal That Independent Genetic Variants Associated with a Complex Phenotype Converge on Intermediate Genes, with a Major Role for the HLA”. en. *PLOS Genetics* 7.8 (2011), e1002197.

[225] Pierce, Brandon L. et al. “Mediation Analysis Demonstrates That Trans-eQTLs Are Often Explained by Cis-Mediation: A Genome-Wide Analysis among 1,800 South Asians”. en. *PLOS Genetics* 10.12 (2014), e1004818.

[226] Yap, Chloe X. et al. “Trans -eQTLs identified in whole blood have limited influence on complex disease biology”. en. *European Journal of Human Genetics* 26.9 (2018), pp. 1361–1368.

[227] Laakso, Markku et al. “The Metabolic Syndrome in Men study: a resource for studies of metabolic and cardiovascular diseases”. *Journal of Lipid Research* 58.3 (2017), pp. 481–493.

[228] Fizelova, Maria et al. “Differential Associations of Inflammatory Markers With Insulin Sensitivity and Secretion: The Prospective METSIM Study”. *The Journal of Clinical Endocrinology & Metabolism* 102.9 (2017), pp. 3600–3609.

[229] Chantzichristos, Dimitrios et al. “Identification of human glucocorticoid response markers using integrated multi-omic analysis from a randomized crossover trial”. *eLife* 10 (2021). Ed. by Mangoni, Arduino A, Rosen, Cliff J, and Bordag, Natalie, e62236.

[230] Cohain, Ariella et al. “Exploring the reproducibility of probabilistic causal molecular network models”. *Biocomputing 2017*. WORLD SCIENTIFIC, 2016, pp. 120–131.

[231] Civelek, Mete et al. “Genetic Regulation of Adipose Gene Expression and Cardio-Metabolic Traits”. en. *The American Journal of Human Genetics* 100.3 (2017), pp. 428–443.

[232] Carvalho, Benilton S. and Irizarry, Rafael A. “A framework for oligonucleotide microarray preprocessing”. eng. *Bioinformatics (Oxford, England)* 26.19 (2010), pp. 2363–2367.

[233] Ritchie, Matthew E. et al. “limma powers differential expression analyses for RNA-sequencing and microarray studies”. eng. *Nucleic Acids Research* 43.7 (2015), e47.

[234] Wagatsuma, Akira and Sakuma, Kunihiro. “Molecular Mechanisms for Age-Associated Mitochondrial Deficiency in Skeletal Muscle”. en. *Journal of Aging Research* 2012 (2012), e768304.

[235] Vartanian, Vladimir et al. “8-oxoguanine DNA glycosylase (OGG1) deficiency elicits coordinated changes in lipid and mitochondrial metabolism in muscle”. en. *PLOS ONE* 12.7 (2017), e0181687.

[236] Marchini, Jonathan and Howie, Bryan. “Genotype imputation for genome-wide association studies”. en. *Nature Reviews Genetics* 11.7 (2010), pp. 499–511.

[237] Johannsson, Gudmundur et al. “Adrenal insufficiency: review of clinical outcomes with current glucocorticoid replacement therapy”. en. *Clinical Endocrinology* 82.1 (2015), pp. 2–11.

[238] Passaro, Angelina et al. “Gene expression regional differences in human subcutaneous adipose tissue”. *BMC Genomics* 18.1 (2017), p. 202.

[239] Hu, Yanhua et al. “Development of a Molecular Signature to Monitor Pharmacodynamic Responses Mediated by In Vivo Administration of Glucocorticoids”. en. *Arthritis & Rheumatology* 70.8 (2018), pp. 1331–1342.

[240] Oakley, Robert H. and Cidlowski, John A. “The Biology of the Glucocorticoid Receptor: New Signaling Mechanisms in Health and Disease”. *The Journal of allergy and clinical immunology* 132.5 (2013), pp. 1033–1044.

[241] Greenbaum, Dov, Colangelo, Christopher, Williams, Kenneth, and Gerstein, Mark. "Comparing protein abundance and mRNA expression levels on a genomic scale". *Genome Biology* 4.9 (2003), p. 117.

[242] Bodine, Sue C. and Furlow, J. David. "Glucocorticoids and Skeletal Muscle". en. *Glucocorticoid Signaling: From Molecules to Mice to Man*. Ed. by Wang, Jen-Chywan and Harris, Charles. Advances in Experimental Medicine and Biology. New York, NY: Springer, 2015, pp. 145–176.

[243] Jiang, Lai et al. "Constrained instruments and their application to Mendelian randomization with pleiotropy". *Genetic Epidemiology* 43.4 (2019), pp. 373–401.

[244] Hu, Xiaoyu, Li, Wai-Ping, Meng, Charis, and Ivashkiv, Lionel B. "Inhibition of IFN-γ Signaling by Glucocorticoids". en. *The Journal of Immunology* 170.9 (2003), pp. 4833–4839.

[245] Flammer, Jamie R. et al. "The Type I Interferon Signaling Pathway Is a Target for Glucocorticoid Inhibition". *Molecular and Cellular Biology* 30.19 (2010), pp. 4564–4574.

[246] Zheng, Ning and Shabek, Nitzan. "Ubiquitin Ligases: Structure, Function, and Regulation". *Annual Review of Biochemistry* 86.1 (2017), pp. 129–157.

[247] Fischer-Posovszky, Pamela, Newell, Felicity S., Wabitsch, Martin, and Tornqvist, Hans E. "Human SGBS Cells – a Unique Tool for Studies of Human Fat Cell Biology". *Obesity Facts* 1.4 (2008), pp. 184–189.

[248] Tomabechi, Yuka, Tsuruta, Takeshi, Saito, Shinichi, Wabitsch, Martin, and Sonoyama, Kei. "Extra-adrenal glucocorticoids contribute to the postprandial increase of circulating leptin in mice". *Journal of Cell Communication and Signaling* 12.2 (2018), pp. 433–439.

[249] Ernst, Jana et al. "Androstenedione changes steroidogenic activity of SGBS cells". *Endocrine Connections* 9.7 (2020), pp. 587–598.