THE UNIVERSITY
of EDINBURGH

Deep mutational scanning of mammalian loci using CRISPR-Cas9 and multiplex HDR

Martijn J.E. Kelder

THE UNIVERSITY of EDINBURGH

Doctor of Philosophy
The University of Edinburgh
2018

igmm
INSTITUTE OF GENETICS & MOLECULAR MEDICINE

MRC Medical Research Council

CANCER RESEARCH UK

Declaration of Authenticity

I hereby declare that this thesis is a result of my own work, unless otherwise stated, where the contributor has been fully acknowledged. This thesis has not been submitted in any form for any degree, qualification or similar at any other university or institution.

Martijn J.E. Kelder

September 2018

Acknowledgements

No one can finish a Ph.D. thesis by themselves and I too have had numerous people who made the last years both possible and enjoyable. I could write an entire thesis acknowledging everyone I am thankful to. Instead, I will highlight a few who have been essential in the process.

First and foremost, I would like to thank and acknowledge Andrew Wood for his supervision during the course of my Ph.D. It cannot be stressed enough that his endless enthusiasm for science and his unconditional support, believing in my work even at times when I did not, on both a professional and personal level have been the main pillars on which this thesis came to existence. Picking me up from the hospital after my bike accident is just one of many examples of his involvement with his lab members.

Other (former) members of the Wood lab, Gillian, Eirini, Lewis, Keerthi, Jessica, have been without a doubt crucial for both moral and technical support.

I would also like to acknowledge the input from my PhD committee members; Neil Carragher, Steven Pollard and Ian Adams. A special note to Ian for his involvement and advice throughout my Ph.D. and his generous time and patience in sharing his computational skills.

All things in life need to be balanced and finishing a Ph.D. is no different from that. It is hence in place to thank the people who have made me see the bright side of life during some of the dark days of my Ph.D. Gil and Thomas; every lunch or coffee break with you guys was a brilliant banter and valuable mirror at the same time. A special thank you to Chris for the numerous memories created during our escapes to the Highlands, mountain bothies and pre-work surf sessions. Louise, Sana, Marie-Jeanne, Matt, Oscar, Ross, James, Adam, Sam, Nezha, Mina, Sander, Nolwenn, Sam, Martino: unforgettable times in Edinburgh and up north, ranging from Edinburgh blizzards to Kearvaig bothy to forest discos.

Nicky, you were absolutely vital at keeping me sane during my write up, making me appreciate that there is more to life than work and that taking regular breaks only increase one's output. Your encouragement and support were unrivalled at times I needed it most.

Last but not least I acknowledge my parents Mariet and Eduard and my sister Paulien for their encouragement. They are one of the main reasons I became the person I am today and I am hence privileged to have these people unconditionally by my side every step of the way.

With 14 nationalities acknowledged above, I would finally like to emphasise the essential role of cultural diversity in academia.

Nothing has meaning
except for the meaning we give to it

- Max Tegmark

Abstract

Functional consequences of genetic variants are best studied in their endogenous chromosomal context. Gene editing by homology-directed repair can introduce such predetermined genetic changes into chromosomal DNA. In this thesis, I develop methods to generate tens to hundreds of genetic variants, expressed from a native chromosomal context, and simultaneously evaluate their phenotypic impact. This approach involves repair of Cas9-derived double strand breaks (DSBs) from oligonucleotide repair template libraries containing controlled levels of nucleotide heterogeneity. Cell populations are then purified based on a phenotypic assay and subjected to deep amplicon sequencing at the target site to link genotype with phenotype.

In the first chapter, I developed a bioinformatics pipeline for the processing of Illumina sequencing reads containing nucleotide variants, and validate this pipeline *in silico*. As a proof-of-principle, in the second chapter I then introduced nucleotide variants across 8 codons of a chromosomal GFP transgene in mouse embryonic stem cells. The functional impact of these variants was quantified, with the results benchmarked against an existing episomal dataset, and by *in silico* modelling of mutant protein structure. In the final chapter, I applied this pipeline to analyse a CRISPR deep mutational scanning dataset incorporating all possible amino acid substitutions within a region of β-catenin, a component of the Wnt signalling pathway, that is a mutational hotspot in many types of cancer. The functional impact of these clinically relevant variants was assessed using a fluorescent reporter of Wnt signalling. By combining the resulting functional scores with mutational signature data from genome sequencing of different tumour types, I finally dissect the relative contribution of mutational bias and natural selection to the different patterns of amino acid substitutions found in different tumour types.

Lay Abstract

The genome is an organism's complete set of DNA including all of its genes. The human genome consists of billions of nucleotides (A, C, G or T) and controls how the cells of an organism behave and function. Individuals have their unique set of genomic variations that make them unique, whilst they additionally acquire mutations throughout life due to errors in DNA replication and exposure to factors such as smoking and sunlight. These alterations underlie a wide variety of diseases including cancer. Understanding how genetic variations affect cell behaviour and diseases is crucial for improving medicine and biotechnology, although examining the effects of many genetic variations simultaneously remains difficult.

In this thesis, a DNA editing tool called CRISPR-Cas9 is used to make thousands to millions of cells in a dish each carry a different change in its DNA, such that the effects of these mutations on the behaviour of these cells can subsequently be measured. To achieve this, Cas9 cuts the DNA of these cells, after which the specific genetic variant is copied into the genome from a small strand of DNA that we introduce into the cell together with Cas9. After this, the fluorescence of these cells is assessed, which is indicative of their fitness as a result of the mutation. After isolation of groups of cells with different levels of fluorescence, we can then read ('sequence') the genetic code to see which mutation corresponds with which fluorescence level.

To analyse the sequencing data, programming scripts have been written and analysed to process the data and to subsequently make counts and graphs of nucleotide variants in the fluorescence groups. This methodology is first tested on cells that harbour a genetic element encoding for a green fluorescent protein (GFP) and subsequently applied to introduce and functionally assess single nucleotide variants introduced into GFP. The effects of many high-quality genetic variants on GFP fluorescence are assessed and validated, demonstrating that our approach works. In chapter 5, the developed approach is used to examine β-catenin, a protein often mutated in cancer. The effects of 380 clinically-relevant mutations in β-catenin were measured, after which this information is used to explain why different cancer types carry certain mutations. Through this, we better understand the mechanisms through which this protein works and how different genetic changes in this protein cause cancer. The technique developed in this thesis allows for systematic screening of the effects of thousands of mutations simultaneously and can be applied to a wide range of targets.

Table of Contents

List of figures

List of tables

List of Abbreviations

Abbreviation	Definition
ABE	adenosine base editor
alt-EJ	alternative end-joining
APC	adenomatous polyposis coli protein
ARM	armadillo repeat domain
β-TrCP	β-transducin repeat containing E3 ubiquitin protein ligase
BE	base editor
BER	base excision repair
BP	base pair
BWA	Burrows-Wheeler Aligner
c-NHEJ	canonical/classical NHEJ
Cas9n	nickase version of Cas9 endonuclease
CKI	casein kinase 1
COSMIC	Catalogue of Somatic Mutation in Cancer
CRISPR	clustered regularly interspaced short palindromic repeats
CRISPRi	CRISPR interference
dCas9	nuclease-deficient Cas9 variant
$\Delta\Delta G$	change in thermodynamic free energy of a protein
DDR	DNA damage response
DMS	deep mutational scanning
DNA	deoxyribonucleic acid
dNTPs	deoxynucleotide
DSB	double-strand break
dsDNA	double-stranded DNA
ESCs	embryonic stem cells
FA	Fanconi Anemia pathway
FACS	fluorescence-activated cell sorter
FIAU	fialuridine, or 1-(2-deoxy-2-fluoro-1-D-arabinofuranosyl)-5-iodouracil
GFP	green fluorescent protein
GSK-3β	glycogen synthase kinase 3β
GWAS	genome-wide association studies
HDR	homology-directed repair
HR	homologous recombination
LEF	lymphoid enhancer-binding factors
LIF	leukaemia inhibiting factor
MEFs	mouse embryonic fibroblasts
MMEJ	microhomology-mediated end joining

NGS	next-generation sequencing
NHEJ	non-homologous end-joining
NRS	nuclear retention signal
NT	nucleotide
PAM	protospacer adjacent motif
PCR	polymerase chain reaction
PE	paired end (sequencing)
RNA	ribonucleic acid
RNAi	RNA interference
SAM	Sequence Alignment/Map
SBS	sequencing by synthesis
SDSA	synthesis-dependent strand annealing
sgRNA	single guide RNA
SNP	single nucleotide polymorphism
SNV	single nucleotide variations
SpCas9	Cas9 from *Streptococcus pyogenes*
SSA	single strand annealing
ssDNA	single-stranded DNA
ssODN	short single-stranded oligodeoxynucleotides
ssODN	single-stranded oligodeoxynucleotides
SSTR	single-strand template repair
TALEN	transcription activator-like effector nucleases
TCF	T-cell factor
TCGA	The Cancer Genome Atlas
TILLING	Targeting induced local lesions in genomes
UV	ultra violet (radiation)
wtCas9	wildtype (nuclease) version of Cas9 endonuclease
ZFN	zinc-finger nucleases

1

Introduction

The human body is composed of trillions of cells (Bianconi *et al.*, 2013) that all contain nearly the same genetic information, which is stored in deoxyribonucleic acid (DNA) molecules that are collectively known as the genome. The genome contains the essential information for the production of all the proteins necessary for the functioning, growth and reproduction of an organism. Each of these proteins is tightly regulated by complex, tissue-specific processes that allow for the formation of different cell types while harbouring the same genome.

The last few decades have provided marvellous advantages in our understanding of our genome. The initial publication of the human genetic sequence by the Human Genome Project in 2001 revealed that there are approximately 19,000 coding genes and that most of the genome does not code for any functional elements and consists of repetitive stretches of code (Venter *et al.*, 2001; Ezkurdia *et al.*, 2014). In the subsequent years of research, scientists elucidated an astonishing amount about the function of genomic elements and showed the power of genomic maps by being able to map genes, regulatory elements, chromatin structures and conservation throughout evolution and between individuals (Lander, 2011). Although there is a high sequence similarity of approximately 99.9% between individuals (Altshuler *et al.*, 2012), this means that there is still variation across roughly 6 million single nucleotide sites. With the largest majority of these genetic variants not affecting functioning of the genome (Kimura, 1968), a major goal of 21st century genetics is to identify those that do.

With the current ease and cost-efficiency of sequencing experimental and clinical samples, vast amounts of genomic data are available on the incidence of single nucleotide variants (SNVs) and on their coincidence with disease and traits. A thorough understanding of the molecular consequences and mechanisms through which these variants work is however often missing and impair our ability to assign a functional significance to them. The reason for this is that systematic introduction and interrogation of vast numbers of genomic variants in their endogenous context has largely been unfeasible.

The discovery and development of CRISPR-based precise genome editing of eukaryotic cells has transformed life sciences. Geneticists can apply genome editing to introduce and study

the effects of genetic elements and sequence variations in their native genomic context to an unpreceded extent.

In this thesis, I will investigate the possibility of developing a systematic approach for the interrogation of the phenotypic consequences of genetic variants in their genomic context using genome editing.

1.1 The eukaryotic genome

1.1.1 Elements of the genome

Every organism is composed of cells, which are considered the basic units of life. Each cell stores its genetic information in deoxyribonucleic acid (DNA), which is a thread-like chain of nucleotides that in turn consist of a phosphate group, a deoxyribose sugar and a nitrogenous base (Watson and Crick, 1953) (see **Figure 1.1A**). Each strand of DNA is bound to a second, reverse strand of DNA, meaning that the two respective strands are going in opposite directions, from respectively 5' end to 3' end and vice versa. The hydrogen bonds between the two strands are formed between the nitrogenous bases of the respective molecules, which come in four types: adenosine (A), cytosine (C), guanine (G) and thymidine (T). Contrary to random occurrence, A will always pair with a T on the opposite strand (with two hydrogen bonds), as a C will similarly bind a G (with three hydrogen bonds). Thus, each strand of DNA is a template of the other.

All organisms can be described as either prokaryotes or eukaryotes, depending on their cellular structure. Whereas prokaryotes (comprising both archaea and bacteria) have a single, circular strand of DNA and contain all their cellular components in the cytoplasm (Doolittle, 1996), eukaryotes have membrane-separated, compartmentalised structures including the nucleus, which contain their genomic DNA. Eukaryotes, including plants and animals, typically organise their DNA in multiple linear molecules, which are enclosed in the nucleus.

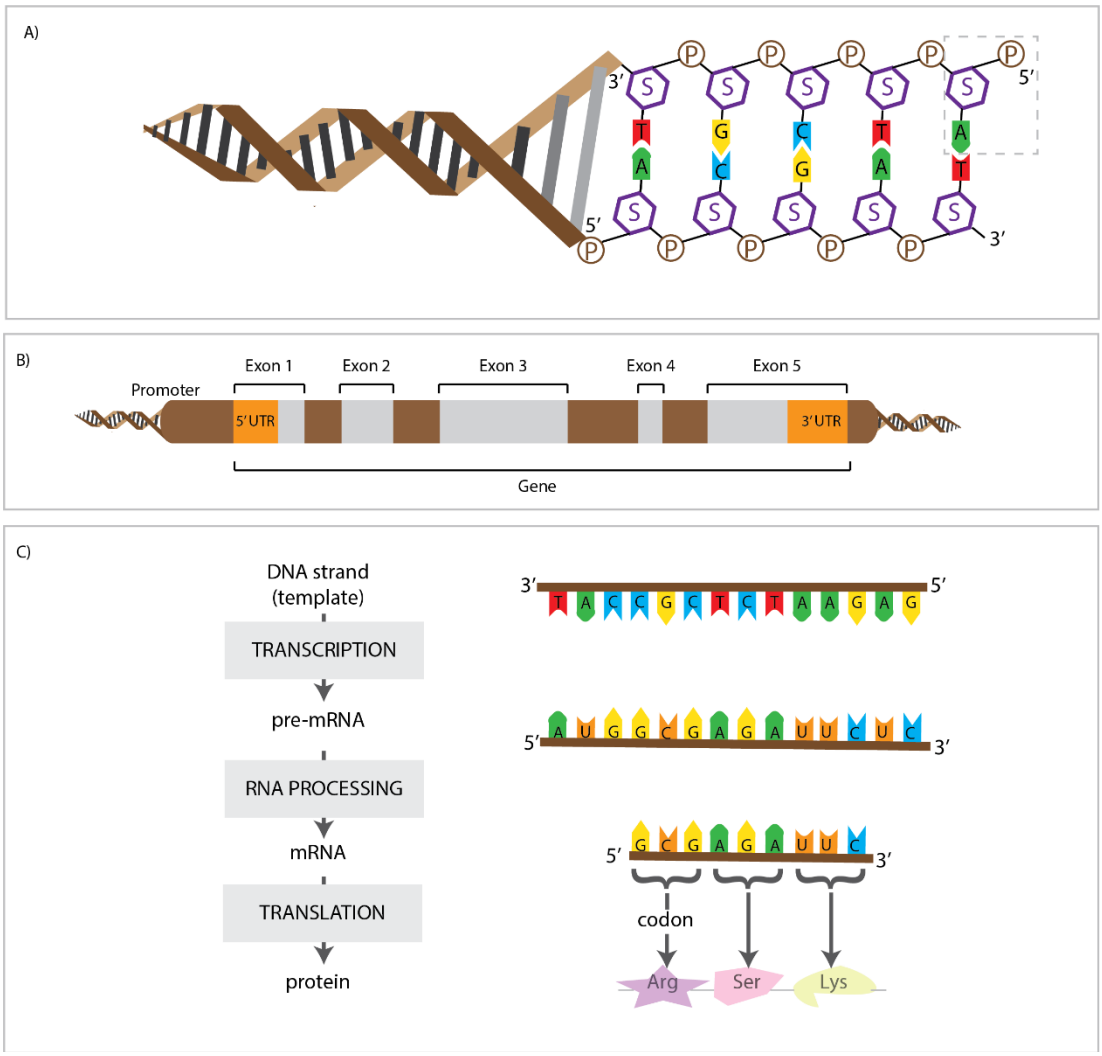


Figure 1.1 Structure of DNA and the central dogma of biology (A) At its primary structure, DNA consists of two nucleotide strands that are wound together in a helix. Each nucleotide (see dashed box) consists of a deoxyribose sugar group (indicated by S) bound at its 5th carbon to a phosphate group (termed 5' end) and to the phosphate group of a neighbouring nucleotide with its 3rd carbon (termed the 3' end). Each nucleotide also has either a nitrogenous base, adenosine (A), cytosine (C), guanine (G) or thymidine (T), at the central part of the helix, with which it binds to the other strand. A always pairs with T, whereas C always pairs with G. While each strand is oriented from 3' to 5', the strands are bound anti-parallel. Each turn of a DNA helix consists of 10.5 nucleotides. **(B)** A gene is a part of the genome that encodes for a functional transcript such as a protein. Expression of a gene is driven by a promoter 5' upstream from the gene and contains individual coding segments (exons) that are intervened by non-coding sequences (introns) and is embedded between untranslated regions (UTRs) at the 5'- and 3'-termini. The number and length of exons and introns varies within and between genes. **(C)** For protein synthesis, DNA is transcribed into pre-mRNA, which is processed into mRNA by removal of intronic regions, after which the mRNA is transported from the nucleus to the cytoplasm, where ribosomes translate it into a chain of amino acids (polypeptide chain) by corresponding each three nucleotides (codon) with an amino acid. This polypeptide chain folds into a protein.

1.1.2 Components of the cell and the central dogma

Each protein is encoded by a gene, which is a genomic region containing the information for the synthesis of a protein-coding or non-coding RNA, of which the transcription is regulated by a promoter (Pesole, 2008). A typical eukaryotic gene contains individual coding segments (exons) that are intervened by non-coding sequences (introns) and is embedded between untranslated regions (UTRs) at the 5'- and 3'-termini (see **Figure 1.1B**). Approximately 19,000 protein-coding genes are found in the human genome (Ezkurdia *et al.*, 2014), making up only ± 1.8 % of the total genomic sequence (Altshuler *et al.*, 2012). These genes can be transcribed into RNA, which can subsequently be translated into proteins (see **Figure 1.1C**). This is known as the central dogma of molecular biology (Crick, 1970).

RNA polymerase transcribes the full sequence of a gene from the DNA into pre-messenger RNA (pre-mRNA). In RNA, uracil (U) is used as a binding partner for A instead of the more stable thymidine T. After transcription, this pre-mRNA is processed and spliced to remove any redundant sequences such as introns, resulting in mRNA. This mRNA can now be transported out of the nucleus, after which it is captured and translated by ribosomal complexes into amino acid sequences that subsequently fold into three-dimensional protein structures. Each three-nucleotide sequence in an exon (codon) encodes for a specific amino acid. As there are (4 nucleotides * 3 three positions) 64 possible codons and 20 different amino acids, several codons encode the same amino acid and are known as synonymous codons (see **Table 1-1**).

In addition to protein-coding genes, other genes may encode for RNAs that themselves fulfil a role in the cell (Morris and Mattick, 2014; Marchese, Raimondi and Huarte, 2017). For example, transfer RNAs (tRNAs) are short strands of RNA that function as homing beacons, necessary for the binding between a specific triplet codon on mRNA and the corresponding amino acids in ribosomes (Rich and RajBhandary, 1976).

1.1.3 Codon usage bias

Whereas it follows from the central dogma that synonymous codons should not alter the eventual amino acid sequence and might hence be considered functionally irrelevant, they

Table 1-1 DNA codon to amino acid translation table. Each three-nucleotide sequence (codon) translates into any of the 20 amino acids or into a stop codon (indicated in red), which terminates translation. Several codons can encode for the same amino acid (i.e. synonymous codons).

First nucleotide	Second nucleotide: T	Second nucleotide: C	Second nucleotide: A	Second nucleotide: G	Third nucleotide
T	TTT Phe	TCT Ser	TAT Tyr	TGT Cys	T
	TTC Phe	TCC Ser	TAC Tyr	TGC Cys	C
	TTA Leu	TCA Ser	TAA Stop	TGA Stop	G
	TTG Leu	TCG Ser	TAG Stop	TGG Trp	A
C	CTT Leu	CCT Pro	CAT His	CGT Arg	T
	CTC Leu	CCC Pro	CAC His	CGC Arg	C
	CTA Leu	CCA Pro	CAA Gln	CGA Arg	G
	CTG Leu	CCG Pro	CAG Gln	CGG Arg	A
A	ATT Ile	ACT Thr	AAT Asn	AGT Ser	T
	ATC Ile	ACC Thr	AAC Asn	AGC Ser	C
	ATA Ile	ACA Thr	AAA Lys	AGA Arg	G
	ATG Met	ACG Thr	AAG Lys	AGG Arg	A
G	GTT Val	GCT Ala	GAT Asp	GGT Gly	T
	GTC Val	GCC Ala	GAC Asp	GGC Gly	C
	GTA Val	GCA Ala	GAA Glu	GGA Gly	G
	GTG Val	GCG Ala	GAG Glu	GGG Gly	A

are found at distinct frequencies in different organisms and even fluctuate widely in frequency between genes within the same genome, or more strikingly within the same gene, a phenomenon known as codon-usage bias (Ikemura, 1985; Sharp and Li, 1986; Plotkin and Kudla, 2011). Codon usage bias has a demonstrated effect on fitness and synonymous codons are hence under a nonnegligible selective pressure. Nucleotide changes due to codon bias can affect many processes including transcription (Zhou *et al.*, 2016) and folding and stability (Presnyak *et al.*, 2015) of an mRNA, whereby GC-rich mRNA sequences typically adopt strong secondary structures (de Smit and Van Duin, 1990; Kudla *et al.*, 2009). On the translational level, initiation rates may be affected by different codons (in *E. coli*) (Kudla *et al.*, 2009), whilst a resulting delay in translation at certain sites ensures correct folding of the protein (Sørensen, Kurland and Pedersen, 1989). Codon optimisation may thus hugely bias correct

functioning of mechanisms at both the transcriptional and translational level, which together affect global translation efficiency and cellular fitness.

The origin of differences in codon usage between genes in mammalian systems is still a point of debate. One underlying explanation is the observation that genes of different functional categories differ in their codon usage, providing evidence that the tRNA abundance in the cell types that these genes are expressed in is the largest determinant of codon usage between genes (Gingold *et al.*, 2014). Concentrations of different isoaccepting tRNAs – i.e. those binding the same amino acid – can vary in a cell (reviewed in Quax *et al.*, 2015). Highly-expressed genes hence frequently display a large skew towards codons with higher tRNA levels, thereby enhancing their elongation rate (Plotkin and Kudla, 2011). However, there still remains debate on whether codon usage adapts to tRNA abundance and as tRNA concentrations are hard to determine, this hypothesis remains based on many assumptions (Laurin-Lemay, Philippe and Rodrigue, 2018).

The most obvious pattern for the skew between genes in the same genome is the observation that in isochores (stretches >300kb with high uniformity in GC levels), both synonymous and non-synonymous codons are under constraints to reflect the GC composition (Bernardi, 2000). A recent study further supported this by demonstrating that codon-usage biases are explained by a correlation in the GC-content between the third codon positions of a gene and adjacent untranslated regions (Pouyet *et al.*, 2017), whereby the authors reasoned that this GC-content is largely driven by the variations in long-term recombination rates instead of expression, with regions with higher GC-content being more conserved. Furthermore, specific classes of genes are thought to fall within specific isochoric groups, e.g. housekeeping genes tend to be more GC-rich whereas genes important for development and differentiation tend to be AT-rich (Peden, 1999). It is hence postulated that GC-content is the major determinant of codon usage bias between genes (Rudolph *et al.*, 2016).

Codon usage bias affects many processes in fundamental and applied biology. For transgene or recombinant protein design, inter-species codon usage bias is a crucial consideration for correct expression (Gustafsson, Govindarajan and Minshull, 2004; Plotkin and Kudla, 2011). Synonymous mutations are known to underlie cancer and account for up to 8 % of the single

nucleotide changes found in oncogenes (Supek *et al.*, 2014). Thus, synonymous mutations emphasise the complexity of the genome and the challenges associated with predicting the functional impact of sequence variants.

1.2 Genomic variations

Genomic variations occur naturally in the genome of all organisms and can lead to alterations in cell function and disease. They are the basis for natural selection, with successive cycles of mutations and selective amplification of the fittest variants causing populations to adapt to their environment over time. Advantageous traits that emerge and provide a survival advantage are thus passed on, whereas deleterious mutations and their hosts gradually disappear.

Genomic variations are present in many forms, including substitutions, insertions, deletions, inversions and duplications and can vary from single nucleotide variations (SNVs) to large chromosomal rearrangements affecting several megabases (Feuk, Carson and Scherer, 2006). Regions of the genome that have no functional role are typically more permissive to the accumulation of mutation (Wolfe, Sharp and Li, 1989), since there is no selective pressure against such variants. In contrast, variants in regions of biological significance are more likely to have a negative impact on the fitness of an organism and would therefore be depleted in such a region. In rare cases, a variant may positively influence viability and would thus be selected for. Therefore, all the genomic variants found in the genome are a result of selective pressure or a lack thereof.

Variants occurring in the human genome, including their rate of prevalence, are described in the 1000-Genomes Project, which analysed over 2,500 haplotypes from 26 populations and comprises the total collection of allelic variants in the genome of an individual (Altshuler *et al.*, 2012). Over 88 million genetic variants occurring in at least 1% of the population have been described, ranging from single nucleotide polymorphisms (SNPs), indels and other small variants to large structural variants spanning over 100,000 bases. Every human individual caries approximately 300 rare, protein-encoding variants (Tennessen *et al.*, 2012), many of which are not functionally assessed.

1.3.1 Single nucleotide polymorphisms

Polymorphisms are genetic variants naturally occurring between individuals and are thought to comprise transient variations in a species that are either to be fixed or removed from a population over time. SNPs are variants found at a single nucleotide that occur in at least 1% of the population and comprise the most frequent type of sequence variation in the human genome, with over 84.7 million SNPs having been identified (Auton *et al.*, 2015). It is estimated that 93% of all protein coding gene loci are within 5kb of the nearest SNPs and that there are on average 2 exonic SNPs per gene (The International SNP Map Working Group, 2001). Thus, a large numbers of SNPs are within close proximity to a gene and are expected to alter the gene's function, by which it is not surprising that a large number of SNPs is associated with disease susceptibility. Besides Parkinson's disease (Nalls *et al.*, 2014) and sickle-cell anaemia (Ingram, 1956), a notable example is a missense mutation in *BRCA1*, which is carried by 10% of the female population and is thought to underlie breast cancer (Millot *et al.*, 2012). In addition to SNPs, two or three consecutive polymorphisms frequently appear in tumours (Rosenfeld, Malhotra and Lencz, 2010). Thus, SNPs play a large role in the variations between humans.

1.3.2 Single nucleotide variants

In contrast to SNPs, which occur in at least 1% of the population, SNVs are variations that can occur at any frequency in the population. SNVs can accumulate in somatic cells throughout the lifetime of an individual and are a main driver of tumorigenesis (Goya *et al.*, 2010). All cells in an organism are exposed to a continuous mutational burden proportional to their proliferation rate and age due to intrinsic infidelities such as the DNA replication and repair machinery (Stratton, Campbell and Futreal, 2009). For instance, over-activity of members of the APOBEC family of cytosine deaminases can cause C>U conversions that are typically repaired by base excision repair, which most often results in C>T conversions (Chen *et al.*, 1987; Powell *et al.*, 1987; Seeberg, Eide and Bjørås, 1995).

In addition to intrinsic factors, exposures to certain carcinogenic factors can result in accumulation of distinct mutations in specific tissues (Pfeifer, 2010). For example, ultraviolet radiation in sunlight is the most abundant factor in our environment but does not penetrate any further than our skin. The elevated abundance of C>T substitutions in melanomas are

hence thought to be the result of excessive exposure to ultraviolet light (Ravanat, Douki and Cadet, 2001; Pleasance *et al.*, 2010). In contrast, substitutions due to tobacco carcinogens are typically associated with C>A and CC>AA in lung cancer (Pfeifer *et al.*, 2002).

Due to the large array of sources of DNA damage, large numbers of SNVs are accumulated throughout the lifetime of an organism. In addition to the millions of SNPs that have been mapped over the last decades (The International SNP Map Working Group, 2001; Frazer *et al.*, 2007; Altshuler *et al.*, 2012; Auton *et al.*, 2015), this rapidly complicates the association of these variants with traits or disease. The 1000-Genomes Project greatly aided our understanding of the genetic variants naturally occurring in the human genome. However, as a minimum of 90% of the human genome is thought to be functionally inactive (Keightley, 2012; Scally and Durbin, 2012; Kellis *et al.*, 2014), the bulk of these variants will have no functional impact on the fitness of the organism (Kimura, 1968). Other studies have described the patterns of SNVs that arise due to certain carcinogenic processes in mutational signatures (Stratton, Campbell and Futreal, 2009; Nik-Zainal *et al.*, 2012; Alexandrov and Stratton, 2014), which further enhance our understanding of the sources and consequences of SNVs (mutational signatures are further discussed in Chapter 5). However, these are typically large association studies and do not systematically look at the effect of individual mutations, I will next showcase the methods typically used to study the effects of genetic variants.

1.3 Genotype-phenotype relationships

1.3.1 Forward genetic screens

Historically, associations of genetic variants with a phenotype have been examined by forward genetic screens (reviewed in Schneeberger, 2014). Hereby, DNA-damaging agents are used for random mutagenesis of an organism with a well-defined genetic background, after which mutants that express a characterised phenotype of interest are selected. Subsequently, genetic mapping is required to elucidate the locus harbouring the causative genetic variant, after which the exact mutation needs to be isolated and characterised. Identifying the loci of knock-outs typically occurred by screening differences in amplicon size by polymerase chain reaction (PCR) (Wienholds *et al.*, 2003), indicating a insertion or deletion, although this does not allow for the detection of missense mutations and requires

labour-intensive backcrossing to remove background mutations. Whilst for a long time it was the only approach to identify genes and mutations that underlie phenotypes, these forward genetic screens are costly and laborious. In addition, larger genes are more likely to be targeted by mutagenesis due to chance and these screens may thus be skewed.

1.3.1.1 Transposon-mediated mutagenesis

Transposon-mediated mutagenesis, i.e. the use of transposable elements to interfere with gene functions, often integrated through viral transduction, proved to be a more efficient forward genetic screening method than chemical mutagenesis in organisms such as *Caenorhabditis elegans* and *Drosophila melanogaster* due to higher mutation rates and lower lethality, although it had limited success in eukaryotes due to the lack of usable elements (Bellen *et al.*, 1989; Spradling *et al.*, 1995; Plasterk, 1996). Reconstruction of the *Sleeping Beauty* transposon however proved a successful method for insertional mutagenesis in higher eukaryotes (Ivics *et al.*, 1997), especially because it only requires TA as its target sequence. The use of *Sleeping Beauty*-mediated mutagenesis in systematic screening however remained limited due to low transposition efficiencies (Luo *et al.*, 1998), although some efforts addressed this limitation and demonstrated the use of this system in genome-wide cancer gene screening (Dupuy *et al.*, 2005). Other transposons such as the PiggyBac system have been developed and are now widely applied in genome-wide screening for cancer genes with TTAA specificity (Rad *et al.*, 2010). Hence, for a long time, transposons provided an effective alternative to radiation and chemical mutagens for the introduction of mutations underlying forward genetic screens.

1.3.1.2 Genome-wide association screens

With the rise of next-generation sequencing technologies, genome-wide association studies (GWAS) (and similarly familial studies) rapidly became a much-applied approach to link genetic variants with traits using samples from large human populations (reviewed in Visscher *et al.*, 2017). GWAS is based on linkage analysis and narrows the window on where a variant may be located, aiming to identify alleles that are non-randomly and differentially distributed between populations with and without the phenotype of interest. It offers a good system to associate genomic variations with traits on a large scale, but does not explicitly identify the causative variant. For common diseases, these studies however provide limited success, as many different loci may contribute modestly and associations may not directly be

informative as the association between the underlying molecular mechanism and phenotype is not always apparent. Experimental interrogations are thus still often required in addition to linkage analysis when validating larger number or co-occurring genomic variations.

1.3.2 Reverse genetic screens

In contrast to forward genetics, hypothesis-driven, reverse genetic screens start with the inactivation of a preselected gene, after which the function of the sequence is determined by the corresponding phenotype. Such 'genotype-to-phenotype' approaches thus start with prior knowledge of specific genetic elements to test the causal role of mutations herein. Next-generation sequencing technologies have greatly enhanced the speed with which random mutations can be mapped in classical forward genetic screens (Schneeberger, 2014), whereas reverse genetic screens improve the targeting scope by disrupting a known genetic element and thus have the advantage of not requiring costly genome-wide screens for genomic perturbations.

1.3.2.1 RNAi

The manipulation of endogenous RNA interference (RNAi) was first discovered in *Caenorhabditis elegans* (Fire *et al.*, 1998) and proved an effective way to disrupt gene function. With RNAi, double-stranded RNA sequences equal to mRNA transcripts are introduced into a cell, where it is recognised as exogenous genetic material and activates the RNAi pathway, which subsequently targets homologous endogenous transcripts for degradation. The resulting gene knockdown is now widely used to screen for functions of transcribed genetic elements (Ketting, 2011) and libraries targeting thousands of genes are available and allow for genome-wide loss-of-function screening. Despite the success of RNAi, substantial off-target activity and imperfect on-target activity obscure interpretation of phenotypes resulting from RNAi screens (Jackson and Linsley, 2010), thus having similar restrictions as classical genetic screens. In addition, RNAi cannot replace classic genetic screens, as the phenotypic effects are transient and not heritable, although the latter can be advantageous when studying essential genes.

1.3.2.2 TILLING

As the mapping of mutations underlying a phenotype of interest was one of the main bottlenecks in reverse genetic screening methods, Targeting Induced Local Lesions in

Genomes (TILLING) was quickly adopted as a general screening method after its publication (Till *et al.*, 2003). TILLING is a reverse-genetic approach that allows for the creation of an allelic series of induced point mutations in a gene of interest. It was initially developed in *Arabidopsis Thaliana* (Till *et al.*, 2003) and was later proven to work in other organisms including *C. elegans* (Gilchrist *et al.*, 2006) and *Danio rerio* (Moens *et al.*, 2008). TILLING is based on the introduction of point mutations in DNA with either ethylmethane sulphonate (EMS) or ethyl nitrosourea. A target gene or genetic region is PCR-amplified with fluorescent primers, after which denaturing and random reannealing of the amplicons occurs. If a mutant amplicon anneals with a wildtype or different amplicon, this reannealing will result in a mismatch. Digestion with an endonuclease (commonly Cel1 (Colbert *et al.*, 2001)) cleaves at mismatch-specific nucleotide sites, after which gel electrophoresis of the digested amplicons reveals the approximate position of the mismatch. As point mutations are less disrupting than large rearrangements, TILLING allowed for a more targeted screening of genomic variants, including the detection of gain-of-function mutations. Thus, TILLING allowed for a more effective and robust screening of mutations in contrast to classical genetic screens, although it still required high arrays of amplicon screening and the pre-selection of genomic regions (Gilchrist and Haughn, 2010).

1.3.3 Interrogating the effects of single nucleotide variants

In addition to characterising or enhancing the function of a gene, targeted mutagenesis of a gene can also be used to study the structure of a protein and the functioning of individual fragments of a genetic element. While genome-wide screening methods (e.g. GWAS) can assess the function of the most common genomic variants, measuring the effects of (multiple) amino acid substitutions on protein structure and function requires setup to diversify the target sequence and screen for a desired phenotype.

Alanine scanning, for example, systematically replaces all residues in a protein with alanine to assess their function (Cunningham and Wells, 1989). Alanine is used as a substitute as it does not introduce any electrostatic or steric effects, although it provides limited characterisation of sequence-function relationships and does not elucidate the effects of other amino acids.

As genetic variants are typically found in (the vicinity of) regulatory elements, the potential genomic locations of such regulatory elements are often determined by DNase I footprinting and more recently DNase I hypersensitive sites sequencing (DNase-seq) (Crawford *et al.*, 2006; Boyle *et al.*, 2008). As open chromatin regions are more accessible, DNAse I cleaves DNA in these regions more frequently, thereby leaving regions in a closed chromatin state intact. By sequencing and mapping the remaining DNA, regulatory elements appear from a depletion in sequence coverage. Such procedures are often used to narrow the window for finding causal variants.

A detailed mechanistic and predictive comprehension of the biophysical relationship between sequence variants and protein function is underlying a wide number of biological and biotechnological principles. In addition, non-synonymous mutations in protein-coding sequences are thought to underlie ± 57 % of the disease-linked genomic variants (Stenson *et al.*, 2017). To give meaning to genomic variants and to understand the consequences of each of these structural variants on the fitness and functioning of a cell or organism is crucial and comprises a large portion of the work of geneticists.

Predicting the function of genetic elements and how genomic variants can affect a phenotype, e.g. protein function, can prove complex. For example, a variant in an exon far away from a binding site or active site can affect the functioning of a protein due to a change in thermodynamic stability or enzymatic activity of the protein (Spiller *et al.*, 1999; Shimotohno *et al.*, 2001; Freeman *et al.*, 2011). Some highly conservative mutations (i.e. amino acid substitutions without a large change in biochemical properties) may be deleterious, neutral or hyperactivating (Fowler and Fields, 2014). Alternatively, a destabilising mutation can rescue or compensate for a destabilising mutation (Bloom *et al.*, 2005). Further contributing to the complexity of genotype-phenotype relationships, the effect of some mutations might require the presence of another, a phenomenon termed epistasis (Aita *et al.*, 2002; Hayashi *et al.*, 2006). If mutations have negative epistasis, their effects are less beneficial than expected from their effects when alone, whilst positive epistasis indicates that the effect of the mutations is synergistic, i.e. larger than the sum (reviewed in Khan *et al.*, 2011). *De novo* predictions on protein stability are often not able to

confidently predict the effects of such mutations on protein stability and thus require mutational data to discriminate between the predicted structures (Fowler and Fields, 2014).

Several computational approaches have been developed to determine the impact of genetic variants. SIFT and PolyPhen (Polymorphism Phenotyping) are such tool that annotates coding nonsynonymous SNPs and predicts the effects of the subsequent amino acid substitutions (Adzhubei *et al.*, 2010; Sim *et al.*, 2012). Both base these predictions on evolutionary sequence homology and the physical properties of amino acids. Similarly, the CADD (Combined Annotation Dependent Depletion) tool scores SNVs based on their deleteriousness, but rather than giving an absolute score it ranks variants based on their predicting effects without a defined cut-off and is therefore less suitable for the analysis of small numbers of variants (Rentzsch *et al.*, 2018). Whereas *in silico* predictions are a suitable method to obtain an initial prediction, they do not compete with *in vitro* and *in vivo* validations of these mutations.

1.3.4 Deep mutational scanning

Through the combination of next-generation sequencing (NGS) and single-cell analysis technologies, mutagenesis screenings can be performed to assess hundreds of thousands of unique variants in a protein on a large scale simultaneously. This process, known as deep mutational scanning (DMS) (Araya and Fowler, 2011), has been applied in several cases to study protein structures over the last years. Cells that express allelic variants of a protein-coding gene (typically introduced from a plasmid or virus) are coupled to a phenotypic screen. However, in addition to isolating variants that harbour the phenotype of benefit as with directed evolution (see section 1.3.4), DMS can give meaning to genomic variants in order to interrogate the protein structure. The phenotypic screen in DMS enriches cells with the desired variant activity whilst cells that harbour a negative variant are depleted. By determining the number of variants present in both the starting pool and the screened pool by deep sequencing, an enrichment score can be calculated for each variant (see **Figure 1.2**). While alanine scanning can reveal the specific catalytic residues in the primary structure of a protein, DMS additionally reveals information on secondary and tertiary structure by elucidating advantageous mutations in a parallel screen. A wide range of selection schemes has been used in the past, including fluorescence, binding affinities, growth and drug

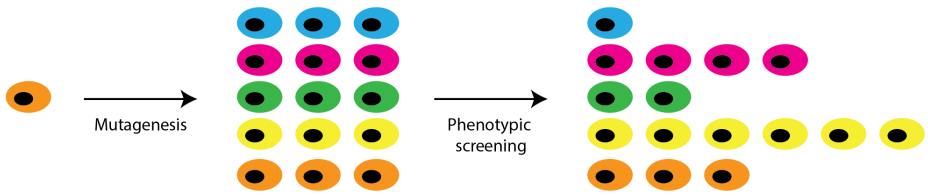
selection (Ge *et al.*, 2010; Sarkisyan *et al.*, 2016; Rocklin *et al.*, 2017), selecting for protein, RNA and regulatory elements (reviewed in Araya and Fowler, 2011). Library synthesis that is required for DMS is further discussed in section 1.6.3.

1.3.5 Directed evolution

The present-day sequence of a genome is the product of millions of years of successive genetic change and selection for a certain phenotype, e.g. fitness. These Darwinian principles have been applied to yield desired phenotypes over thousands of years to breed animals and plants with desired phenotypes. With the onset of molecular techniques, evolution can be applied in a directed fashion on the level of single cells or even single molecules rather than an entire organism. Hereby, successive rounds of mutagenesis and phenotypic selection are applied to yield a desired phenotype, e.g. protein function, which is the underlying principle of directed evolution (see **Figure 1.3**) (Romero and Arnold, 2009).

Directed evolution is an application of DMS and are performed by introducing different modifications in the genetic element of interest. The pool of cells bearing these different mutations can subsequently be exposed to a phenotypic screen on the protein level, such as catalysing a reaction or interactions with other proteins, to select for more beneficial phenotypes. Repetitive cycles allow for rapid evolution and the selection of a desired phenotype, thereby allowing for the optimisation of existing properties and evolution of proteins towards new goals.

As even for a small protein of 100 amino acids there are more possible amino acid sequences (20^{100}) than there are atoms in the universe (10^{80}) (Smith, 1970), there is a tremendous spectrum of proteins that can be engineered through directed evolution that are not found in nature. While most polypeptides do not encode a stable protein (and even more unlikely a protein with the desired phenotype) (Govindarajan and Goldstein, 1997; Taverna and Goldstein, 2002), studies have shown that most random amino acid substitutions are tolerated (Guo, Choe and Loeb, 2004; Bloom *et al.*, 2005). This is an important given for evolution as single nucleotide mutations, the most frequent mutation type occurring naturally, in genes important for fitness would otherwise always lead to a lower fitness and thus stall evolution (Smith, 1970). As most single amino acids substitutions are either



Variant	Mutation	Counts (input)	Counts (selected)	Functional score
Blue	T52R	3	1	0.33
Pink	S44I	3	4	1.33
Green	D60E	3	2	0.67
Yellow	S44E	3	6	2
Orange	Wildtype	3	3	1

Figure 1.2 Deep mutational scanning generates large-scale mutational data Deep mutational scanning relies on both high-throughput mutagenesis and high-throughput sequence analysis and allows for systematic structural analysis of protein variants. Starting with an initial pool, sequence variants are introduced at equal rates through mutagenesis, creating a heterozygous library of sequence variants. These sequences are then subjected to a phenotypic screen (e.g. survival) by which selection is imposed. By assessing the frequency of each variants in the pool before and after screening, a functional score can be calculated (bottom). Mutations that are enhancing fitness have a functional score larger than 1, whereas mutations that have a negative effect on fitness have a value between 0 and 1.

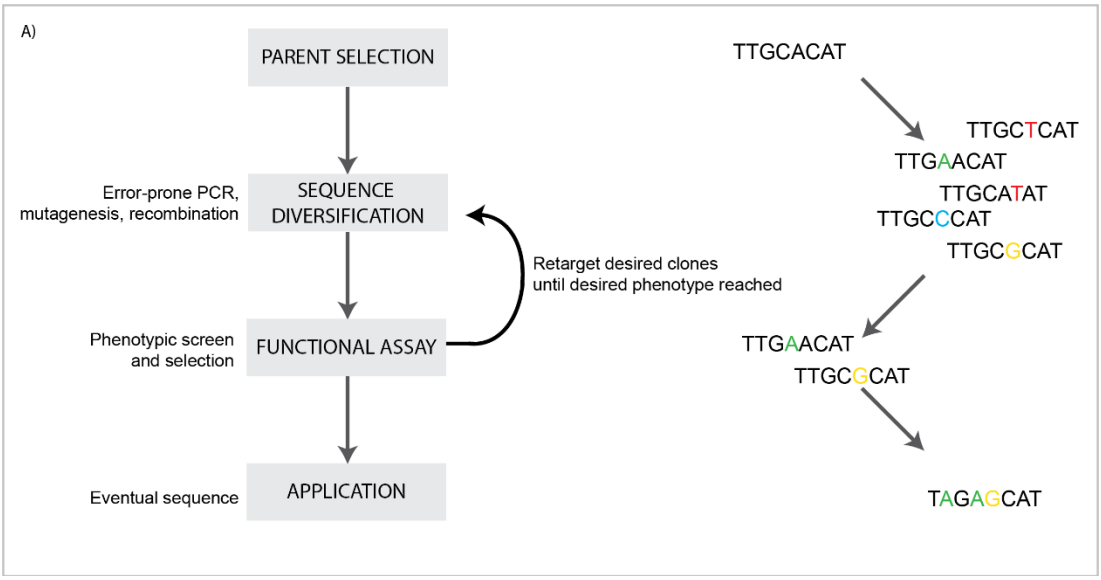


Figure 1.3 Directed evolution to yield desired phenotypes With directed evolution, the aim is to yield an enhanced or novel phenotype (e.g. enzyme activity or fluorescence) by sequential rounds of sequence diversification and selection for a desired phenotype. First of all, a suitable parent or target sequence (gene) and suitable phenotypic screen need to be selected. The target sequence is diversified, such that different clones each harbour a different sequence variant, after which the clones with the desired phenotype are selected.

These can be retargeted to further diversify the sequence and enhance the phenotypic selection until the desired phenotype has been met.

neutral or deleterious and only 0.01 - 1% is beneficial (reviewed in Romero and Arnold, 2009), directed evolution screens are most optimal by introducing only a single or two changes per cycle, by which the desired phenotype is often met after only five to ten cycles (Romero and Arnold, 2009).

Directed evolution has become a common tool for acquiring proteins or genetic elements with an optimised or novel function. Besides the eventual optimised product, it can also improve our understanding of genetic and protein structures.

1.3.6 Saturation mutagenesis

For the generation of sequence libraries used in directed evolution or DMS, *in situ* mutagenesis through chemical agents was initially the gold standard. However, limitations included the targeting specificity and on-target efficacy. As an alternative, saturation mutagenesis provides a sophisticated approach through which a variant library of all possible amino acids at a particular site in a genetic sequence is constructed. Many approaches have been developed to achieve this, including the treatment of PCR fragments with chemicals or error-prone PCR amplification to introduce nucleotide mismatches (Myers, Tilly and Maniatis, 1986; Cunningham and Wells, 1989). These DNA sequences are subsequently cloned into a plasmid backbone and introduced into mammalian cells by episomal or viral gene transfer. Saturation mutagenesis is often used to construct libraries which are subsequently used in directed evolution experiments in single cell organisms including budding yeast (*Saccharomyces cerevisiae*) and bacteria (*Escherichia coli)* (Derbyshire, Salvo and Grindley, 1986; Zheng, Baumann and Reymond, 2004; Reetz et al., 2010), though its application in higher organisms has been limited due to the difficulty of introducing and comparing genetic variants at large scale.

1.3.7 DNA library synthesis

In addition to saturation mutagenesis, several other strategies can be used to synthesise diverse repair template libraries in order to perform directed evolution. Gene synthesis can assemble dsDNA sequences *de novo* and also allows for the introduction of controlled

diversity at the nucleotide or codon level, whilst reducing costs and improving throughput (Wan *et al.*, 2017). Gene synthesis allows to minimise bias and can as such create libraries with equal proportions of codon variants, or avoid the introduction of stop codons or multiple non-consensus codons.

Nucleotide sequences can be synthesised in bulk and can be integrated into the genome through several techniques, which are further discussed in sections 1.4 and 1.5. Whilst single-stranded oligodeoxynucleotides (ssODNs) typically yield increased genomic integration rates compared to dsDNA donors (Chen *et al.*, 2011; F Ann Ran *et al.*, 2013; Renaud *et al.*, 2016), these strategies are less commonly used for library diversification and require alternative DNA repair mechanisms. For oligonucleotide synthesis, deoxyribonucleotide triphosphates (dNTPs) are sequentially coupled to the growing oligonucleotide chain in the order required by the sequence of the product, typically up to 200 bp (Stemmer *et al.*, 1995). In order to yield diversity, partially degenerate oligos are used in some studies, whereby during particular cycles of the synthesis, a number of possible dNTPs are mixed and incorporated at equal rates during synthesis, such that any of the nucleotides gets incorporated at random at the specified sites, whilst at other positions a fixed nucleotide gets incorporated. Instead of completely random codons, most often the complete set of standard amino acids is encoded using NNK or NNS codons, where K represents G or T and S represents C or G, as it lowers the probability of a premature stop codon and simultaneously induces a more uniform distribution of the 20 standard amino acids, compared to the distribution induced by the NNN codon (Patrick and Firth, 2005).

A similar process is oligonucleotide doping, whereby mixed base composition at specific sites are yielded by skewing the rate towards one dNTP, such that the other dNTPs are incorporated at much lower rates than with degenerate oligonucleotides. The disadvantages of degenerate oligonucleotides and oligonucleotide doping are that due to the random nature, (1) the co-occurrence of multiple mutations is likely and (2) there will be a skew towards codons that require only one or two nucleotide changes, thereby resulting in an unequal representation of codons in the library. The latter is advantageous as it reflects patterns of human mutations, while it simultaneously reduces the proportion of protein space, i.e. number of possible polypeptide sequences, that is available.

DMS has proven its applicability in a wide range of systems and is a suitable way to scan protein structures and the effect on both mRNA transcription and protein translation in several ways as discussed in section 1.1.3. Although plasmids and viral vectors provide a suitable way to drive directed evolution or DMS, transient expressions are not always able to capture the subtleties and complexity of cell regulation (Gibson, Seiler and Veitia, 2013). Genetic variances are hence best studied in their native sequence and epigenetic context, with the presence of endogenous transcription factors and normal expression levels the cell is found in. Hence, I will next discuss how genetic variants can be introduced into their genomic context and the historical context of mutagenesis.

1.4 Genome editing

1.4.1 Mutagenesis through radiation and chemical agents

Long before the discovery of the DNA double-strand helix, scientists pioneered ways to introduce mutations into the genome. When the term 'mutation' was coined as a process of 'discontinuous heritability' by De Vries (De Vries, 1914), he envisioned the induction of directed mutations, which would provide man with '*unlimited power over nature*'. After that, the field of genetics underwent rapid development through the discovery of heritable principles such as chromosomal crossing and genetic mapping (Sturtevant and Morgan, 1923), but their attempts to induce mutations with ether or ethanol remained unsuccessful. Indeed, the chemical nature of mutations remained a major and controversial question in the field until Hermann J. Müller discovered that X-rays had a mutagenic effect in *Drosophila* (Muller, 1927). Muller realised as a first that, in addition to inheritance and variability, genetic changes can spontaneously arise and passed on to subsequent generations, a fundamental insight for evolution. Whereas the previous decades of *Drosophila* work across the globe discovered approximately 200 mutations, Muller's first experiments with X-ray-induced mutations allowed him to find half of that number in less than two months (Carlson, 2007). This work greatly changed the concept of genetics, realising that our heritable material could be altered by man, therewith placing evolution in his control. Shortly after, Müller's collaborator Altenburg elucidated that ultraviolet (UV) radiation similarly induced alterations to genes (Altenburg, 1930).

Later, Delbrück, Timoféeff-Ressovsky and Zimmer concluded from experiments with X-rays that the gene might have a defined structure which could be decoded (Timoféeff-Ressovsky, Zimmer and Delbrück, 1935). This discovery subsequently lead to the establishment of the first phage group, considered the birth of molecular biology (Friedberg, 2002), which discovered that both X-rays and UV radiation could induce viral resistance in bacteria by increasing the mutation rate (Luria and Delbrück, 1943). In the same years, in addition to radiation, many chemical mutagens were discovered and thus provided initial means to alter the DNA prior to its discovery (reviewed in Auerbach, 1949).

1.4.2 Recombinant DNA

Recombinant DNA technologies in the 1970s provided scientists with the '*technical ability to join together, covalently, DNA molecules from diverse sources'* and provided huge advantages in the manipulation of DNA molecules (Singer and Soll, 1973). Recombinant DNA allowed for the interrogation of genetic function out of their native context and through the manipulation of homologous recombination (HR, further discusses in section 1.6) (Capecchi, 1989), exogenous DNA sequences with homology to the donor site could be integrated into the genome. Despite this technological breakthrough, the extremely low success rate of integration (1 in 10^6 - 10^9 cells) restricted the scale onto which these techniques can be applied. The use of the non-pathogenic adeno-associated virus type 2 (AAV) to transduct mammalian cells with recombinant DNA increased the rate of successful integration of DNA over 1% (Hermonat, Muzyczka, PNAS 1984), making it one of the most powerful methods for DNA manipulations for decades.

Similar approaches have proven more successful in other approaches, species and cell types. For example, manipulations of DNA in Saccharomyces cerevisiae were relatively effective compared early on in the field of genetics due to the use of shuttle vectors which were not effective in mammalian systems (Sikorski and Hieter, Genetics, 1989). Similarly, the chicken DT 40 B cell line has been a popular due to the ease of genetic manipulations exceeding any other mammalian cell line (Winding, Berchtold JIM 2001). The genetic editing of mouse of human cells would however remain laborious for the next decades.

1.4.3 Genome editing

In the past two decades, developments in the field of genome editing have allowed for the introduction of site-specific modifications in a genetic sequence in its endogenous context (Carroll, 2008). Genome editing enables specific and precise mutagenesis through the recruitment of DNA-binding factors that create a double-strand break (DSB) at the target site. The outcome of the subsequent, endogenous DNA repair mechanisms will be determinant for the eventual result of the genome editing strategy used, which is further discussed in section 1.6 (Bibikova *et al.*, 2003).

Technologies that can conveniently insert, delete and alter the nucleic sequence in a precise manner in either cells or whole organisms allow for the dissection and reconstruction of cellular networks in a wide range of applications. In biotechnology, the combination of genetic building blocks with genome editing tools allows for the development of disease-resistant crops or the production of biofuels. Within medicine, genome-manipulating tools allow for the repair of genetic alterations that underlie diseases or for the discovery of novel drug targets by selectively knocking out a single or multiple genetic element.

The tethering of the nuclease domain of the *Fok*I restriction enzyme to the DNA-binding domain of eukaryotic transcription factors known as zinc-finger proteins (ZFP), known as zinc-finger nucleases (ZFNs) (Bibikova et al., 2003; Porteus and Carroll, 2005; Urnov et al., 2010; Wood et al., 2011), gave birth to the first tools to yield site-specific DSBs (see **Figure 1.4A**). Soon after, fusion of the same *Fok*I nuclease domain to transcription activator-like effector (TALE) domains gave rise to the modular transcription activator-like effector nucleases (TALENs; **Figure 1.4B**) (Boch *et al.*, 2009; Morbitzer *et al.*, 2010; Miller *et al.*, 2011; Wood *et al.*, 2011). Due to its nickase activity, the *Fok*I domain requires dimerization in order to induce DSBs in both ZFN and TALEN approaches, thus requiring two separate binding events, which greatly enhances the targeting specificity.

The ZFP domain of a ZFN monomer contains a tandem array of up to six Cys-His2 fingers that each recognises a 3bp DNA region (Wolfe, Nekludova and Pabo, 2000). In TALENs, a single TALE repeat binds to a single base pair of DNA (Boch *et al.*, 2009). Due to design efficiencies, with the relationship between peptide sequence in the DNA binding domain and the

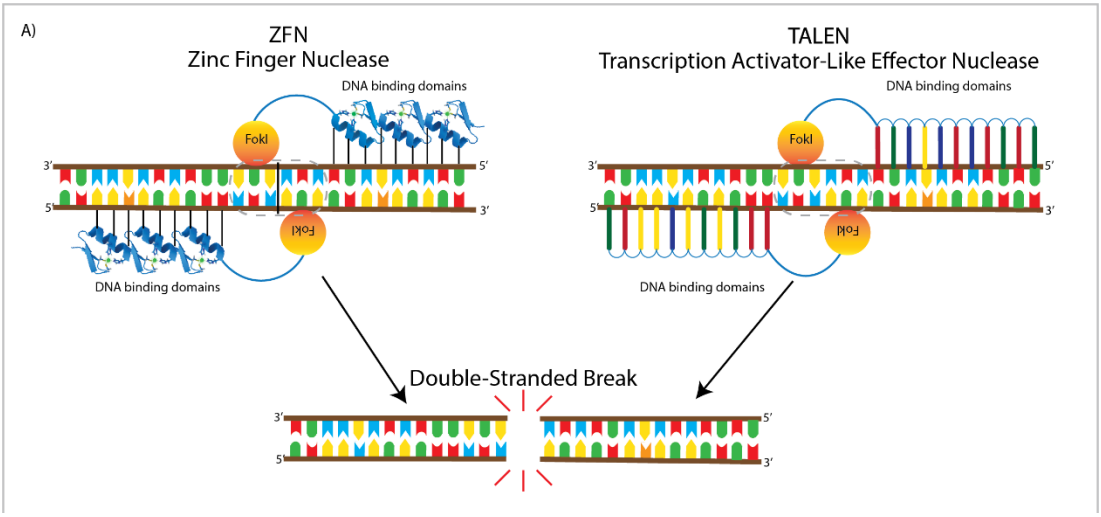


Figure 1.4 ZFNs and TALENs allow for precise introduction of genomic disruption Both ZFNs and TALENs utilise a non-specific FokI cleavage domain, which requires dimerisation in order to introduce a double-strand break in the spacer region (grey dashed box). ZFNs consist of DNA binding subdomains (i.e. fingers, 12-31 repeats), which each recognises a specific nucleotide triplet. By using two pairs of different fingers, ZFNs specifically introduce a break at the target site. Similarly, TALENs consist of subdomain that bind to specific single nucleotides, thereby having an increased programmability over ZFNs. TALENs can entail up to 18 repeats, thus recognising 36 nucleotides specifically, whereas ZFNs can be built from up to six fingers, recognising 24 nucleotides.

recognised nucleotide sequence being much easier to predict *in silico*, TALENs are highly programmable and considered to be advantageous over ZFNs or the less frequently used homing endonucleases (Smith *et al.*, 2006) for introducing targeted DSBs, but as a single TALE repeat (each repeat 33-34 amino acids in size) is needed for each target nucleotide, TALENs with single site specificity are larger and somewhat harder to deliver than ZFNs.

Genome editing with ZFNs and TALENs revolutionised the field of genetics, as precise editing of the genomic sequence had become scalable (compared to previous approaches) and efficient compared to recombinant DNA. However, as the construction of TALENs is still costly and requires labour-intensive protein engineering for each retargeting, it still limited the application of genome editing at a large scale.

1.5 Genome editing using CRISPR-Cas9

The adaptation of the bacterial CRISPR-Cas9 system from *Streptococcus pyogenes* (SpCas9) to eukaryotic cells provided large improvements in both efficiency and ease of design of precise genome editing (F Ann Ran *et al.*, 2013; Jinek *et al.*, 2013). In this section we will

examine the background of the CRISPR-Cas system and its manipulation for precise genome editing.

1.5.1 CRISPR-Cas systems mediate adaptive immunity in bacteria

Many bacterial genomes encode clustered regularly interspaced short palindromic repeats (CRISPRs), which are transcribed and processed into short RNAs that guide CRISPR-associated (Cas) proteins to cleave foreign nucleic acids of invading viruses or plasmids (or close relatives) that have previously been encountered (Ishino *et al.*, 1987; Jansen *et al.*, 2002; Barrangou *et al.*, 2007; Garneau *et al.*, 2010), thereby functioning as an adaptive immune system against these invaders (see **Figure 1.5A**).

Upon phage infection, CRISPR arrays in the bacterial host can acquire new repeat-spacer sequences that match the infecting virus, thereby guiding Cas proteins to cleave the foreign DNA upon a subsequent infection by the same or closely related viral strain. Only cells with these newly acquired sequences will survive infection with this agent, meaning that the spacer content of CRISPR arrays reflects the phages and plasmids it has been exposed to previously (Barrangou *et al.*, 2007; Doudna and Charpentier, 2014). This genetic memory allows the bacteria to rapidly develop an immune response to the agents it reencounters.

CRISPR-Cas systems are classified in three types (I, II and III), which use characteristic mechanisms to recognise and cleave the foreign DNA (Makarova *et al.*, 2011). Cas proteins require a CRISPR RNA (crRNA) that specifically recognises a foreign phage or plasmid DNA sequence and therewith guides Cas cleavage (Brouns *et al.*, 2008). Cas type I and II systems require an additional protospacer-adjacent motif (PAM) adjacent to the crRNA binding site on the invading DNA (Mojica, García-Martínez and Soria, 2005). Whereas type I and type III systems utilise a complex of multiple proteins for the cleavage of target strands, type II systems only require a single Cas protein (Gasiunas *et al.*, 2012; Jinek *et al.*, 2013), making these systems convenient for genome engineering.

The type II Cas9 system in *Streptococcus pyogenes* (SpCas9) is the best studied system and consists of two nuclease domains: HNH and RuVC-like (see **Figure 1.5B**) (Tang *et al.*, 2002; Bolotin *et al.*, 2005; Mojica, García-Martínez and Soria, 2005; Pourcel, Salvignol and

Vergnaud, 2005). The HNH domain of Cas9 is used for the cleavage of the DNA strand that is complementary to the sgRNA sequence, whilst the RuVC domain cleaves the adjacent strand at the same site to produce a blunt ended DSB.

As mentioned, for the recognition of the target strand, the crRNA requires a PAM sequence directly adjacent, typically NGG, further guiding the Cas9 nuclease to introduce a DSB between three and four nucleotides upstream of the PAM. In order for Cas9 to bind to the target site, it requires a dual RNA complex of the crRNA and another small trans-activating crRNA (tracrRNA) (Jinek *et al.*, 2012).

1.5.2 CRISPR-Cas9 system for highly efficient, targeted editing of genetic sequences

Through codon optimisations, the CRISPR-Cas9 system was optimised for the use in eukaryotic cells and proved its ability to cleave specified target strands in human cells (Cong *et al.*, 2013; F Ann Ran *et al.*, 2013; Jinek *et al.*, 2013; Mali *et al.*, 2013). This discovery promoted genome editing from a technical possibility to a more practical reality that could be routinely used by allowing highly efficient, programmable nucleases to be designed and assembled in a quicker and more efficient manner. The crRNA and tracrRNA were engineered into a single guide (sgRNA) that still contains both the 20-nt recognition sequence at its 5'-end and a more complex structure allowing for the binding of Cas9 on the 3'end (see **Figure 1.5B**). In contrary to laborious protein engineering necessitated by TALENs and ZFNs, the targeting of CRISPR-Cas9 can be adjusted to any genomic sequence adjacent to an NGG PAM by altering the 20-nt target sequence, making it a convenient two-component system for genome editing applications.

1.5.3 Alternatives to using SpCas9 for genome editing

Since the emergence of CRISPR-Cas9 as a targetable genome editing tool in a wide variety of species, genome editing became available for the wide audience and has been modified to be repurposed to allow for a wide range of applications. Besides SpCas9, which limits editing to a region containing the NGG PAM motif, a wide variety of optimised Cas9 proteins have broadened the scope of PAM sequences that can be targeted (summarised in (Komor, Badran and Liu, 2017)). In addition to Cas9, other bacterial CRISPR endonucleases have been

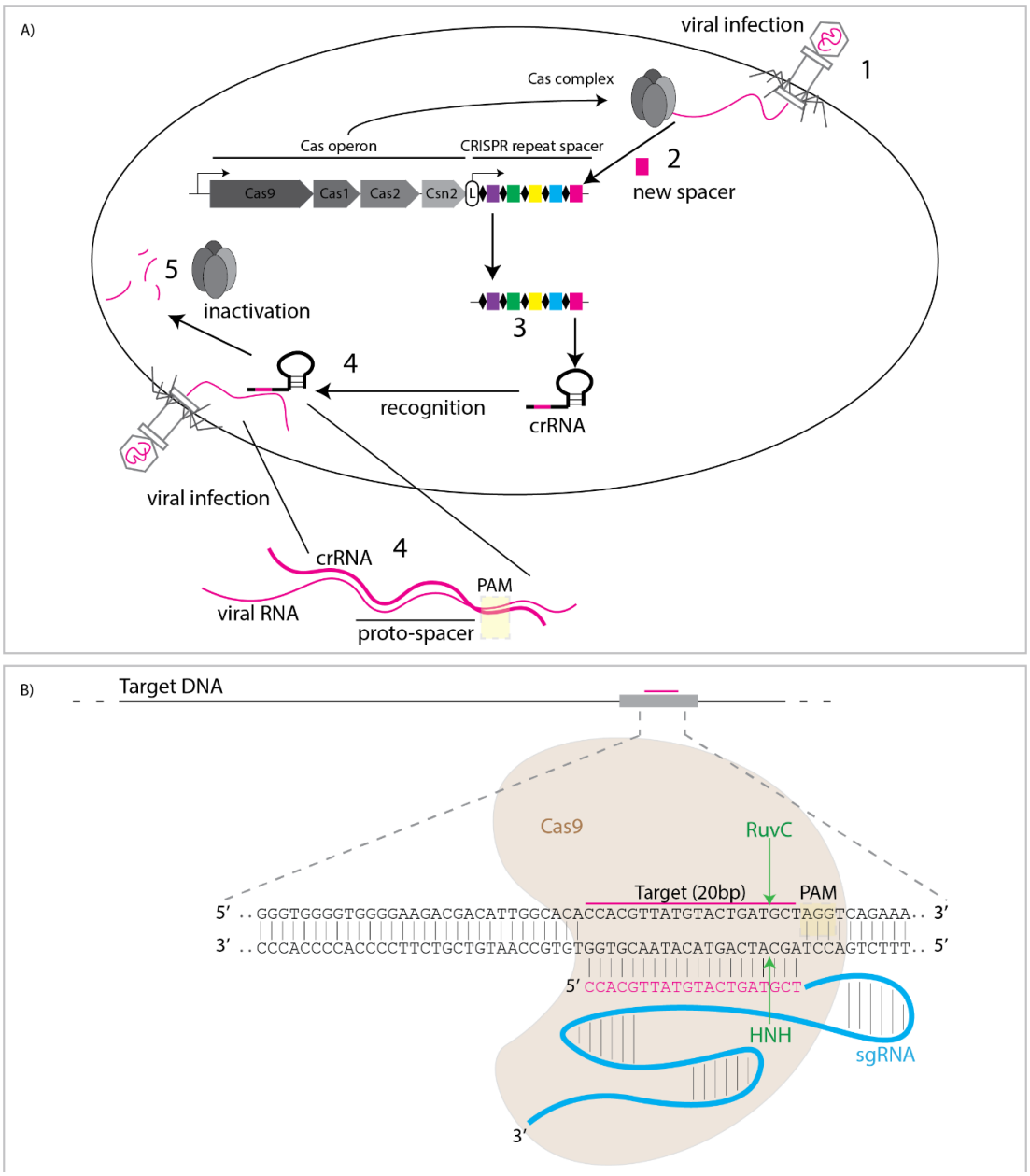


Figure 1.5 CRISPR is an adaptive immune system in bacteria and can be utilised for precise genome editing (A) In bacteria, the CRISPR locus contains clusters of both repeats (black diamonds) and spacers (coloured squares), which are preceded by a leader sequence (L) and a CRISPR-associated (Cas) operon. Upon viral infection (1), the Cas complex cleaves the viral genome and incorporates these new spacers into the CRISPR locus (2). After transcription of the CRISPR repeat spacer, the crRNA is cleaved into small crRNAs (3). Upon a novel infection, the crRNA recognises the viral genome (4) and binds with a 20-nt spacer directly adjacent to a protospacer adjacent motif (PAM). Through this binding, a tracrRNA (not shown) and the Cas complex can bind, after which Cas9 inactivates the foreign DNA through cleavage (5). Figure adapted from Barrangou *et al.*, 2015. **(B)** The CRISPR-Cas9 system from S. pyogenes has been adapted to allow precise genome editing in eukaryotic cells, requiring a single guide RNA (sgRNA) and Cas9 protein. This sgRNA contains a 20nt protospacer (pink) that binds to a complementary sequence adjacent to an NGG PAM, whilst the rest of the sgRNA allows for homing of Cas9 (brown). Cas9 has two cleavage domains, HNH and RuVC-like, which induce

cleavage on respectively the PAM-distal and PAM-adjacent strand, 3 nt upstream of the NGG PAM. Figure adapted from Ran *et al.*, 2013.

identified to have RNA-guided DNA cleavage activity in eukaryotic cells, amongst which Cpf1 notably requires a TTTN PAM at the 5' rather than the 3' end of the protospacer and induces staggered rather than blunt DSBs (Zetsche *et al.*, 2015). These variants to SpCas9 greatly broaden the scope and allow for more specific targeting to regions not accessible when restricting to an NGG motif, although SpCas9 remains the most frequently used nuclease.

A series of variant Cas9 proteins (xCas9) was similarly developed through directed evolution to recognise a broader range of NG PAMs (instead of the conventional NGG sequence), thereby expanding the targeting space of targeted nucleases (Hu *et al.*, 2018). These proteins can function in different contexts including DNA base editing showing lower off-target activity, but is outcompeted in recognising NGH PAMs *in vitro* by a recently published SpCas9 variant (Nishimasu *et al.*, 2018). Such advances in evolving and engineering Cas9 enzymes improve the applicability of genome editing in context difficult to target with conventional spCas9.

One limitation in the application of CRISPR-Cas9 for therapeutic purposes is the cleavage at genomic regions other than the targeted site, known as off-target effects, to which several solutions have been found. Firstly, truncation of the sgRNA can limit off-target DSBs, which is hypothesised to be due to the resulting decrease in tolerance for mismatches(Tsai *et al.*, 2015). Secondly, the delivery of pre-constructed Cas9:sgRNA ribonucleotide protein complexes (RNP) rather than plasmid expression results in a shorter, more transient effect, thus preventing Cas9 to linger around for too long and exert off-target activity (Lin *et al.*, 2014). By reducing the interaction between Cas9 and the substrate DNA, it was also possible to ensure Cas9 only cleaved the intended target (Slaymaker *et al.*, 2015). Additionally, whereas Cas9 by default creates a blunt double-strand break, interference with either of the nuclease domains, SpCas9 (HNH ($Cas9^{N863A}$) or RuVC ($Cas9^{D10A}$) (see **Figure 1.5B**)), results in the development of two nickase (Cas9n) variants that create a single-stranded DNA break on either the target or complementary strand (F. Ann Ran *et al.*, 2013; Mali *et al.*, 2013), thus requiring the independent binding of two different nickases in order to create a double stranded break (with overhang) (further discussed in section 1.6.2). Finally, FANCD2 is shown

to be recruited to Cas9-induced DSB sites and can therefore be allowed to specifically identify off-target sites (Richardson *et al.*, 2018). These approaches reduce potential off-target effects and thus further enhance the applicability of CRISPR-Cas9 as an editing tool in mammalian cells.

1.6 DNA damage repair in genome editing

As genome editing tools including CRISPR-Cas9 merely create lesions in the target DNA, the mutational outcome wholly relies on the subsequent response in the host cell. DNA-based genomes have developed in an environment set for aerobic metabolism and photosynthesis (Blankenship and Hartman, 1998; Falkowski *et al.*, 2005). As such, from the early stages of evolution, organisms already had to cope with DNA-damaging factors such as UV light and reactive oxidative species. Still in dividing mammalian cells today, an estimated average of ten DNA double-strand breaks (DSBs) occur per cell daily (Chang *et al.*, 2017). As cells cannot function with sustained DNA damage, evolution would not have been possible without the development of systems to dampen the most harmful effects of DNA damage. It is therefore not surprising that a wide array of DNA damage response (DDR) pathways can be found in eukaryotes, whereby each DDR pathway targets specific types of damage and errors. In this section, the main DDR pathways thought to act during CRISPR-Cas9 genome editing will be reviewed. However, much of our knowledge of these pathways comes from studies of DNA damage induced by agents other than genome editing nucleases and their relevance to CRISPR mutagenesis is thus still a very active area of research to this day (C. D. Richardson *et al.*, 2016; Brinkman *et al.*, 2018).

1.6.1.1 Double-strand break repair

Like most other genome editing tools, Cas9 typically creates a DSB at the target site, after which the subsequent DNA-repair machinery of the cell determines the outcome of editing. These DSBs can be highly damaging to a cell and can result in translocations, inversions or fragmentation of the DNA, especially when used with multiple sgRNAs (Canver *et al.*, 2014; Boroviak *et al.*, 2017; Kosicki, Tomberg and Bradley, 2018). As such, cells aim to minimise eventual scarring of the DNA by deploying either of the two major repair pathways: non-homologous end-joining (NHEJ) or homology-directed repair (HDR), in which the former does not use a homologous template and the latter does. Each of these are subdivided into multiple pathways, which will be discussed in the following sections.

1.6.1.2 Non-homologous end joining

In the absence of a homologous repair template, localised DSBs are repaired through canonical/classical NHEJ (c-NHEJ), which is the primary pathway for DSB repair. After detection of the break, the Ku70/Ku80 heterodimer and DNA-PK catalytic subunit (DNA-PKcs) stabilise the two ends of the break by forming the DNA-PK holoenzyme (see **Figure 1.6A**) (Meek, Dang and Lees-Miller, 2008). The DNA-PK holoenzyme recruits factors necessary to trim (nucleases) or fill in (polymerases) the termini after which these get directly ligated with DNA ligase IV and Xrcc4 (Li *et al.*, 1995; DiBiase *et al.*, 2000; Wang *et al.*, 2001). Hence, c-NHEJ often leads to insertions or deletions (indels) at the target site, or in rare cases translocations. However, perfectly repaired DNA is indistinguishable from DNA that has never been broken and thus cannot be measured *post hoc*, and successive cycles and cleavage and perfect repair could underlie a grossly overrated rate of errors. Published data has suggested that most Cas9-induced breaks are actually repaired via error-free repair (Bhargava *et al.*, 2018), whilst the accuracy of rejoining is still a major point of debate and has been disputed by others (Brinkman *et al.*, 2018). The vast majority of c-NHEJ repair of CRISPR-Cas9-induced breaks that can be measured through mutations typically results in indels not extending 20bp in length (Koike-Yusa *et al.*, 2014; Tan *et al.*, 2015; van Overbeek *et al.*, 2016), although large indels up to 25 Mb have been observed (Birling *et al.*, 2017; Kosicki, Tomberg and Bradley, 2018). This application of genome engineering is convenient in knock-out screens of genes, as indels in an exonic sequence often result in frameshifts or premature stop codons (Rouet, Smih and Jasin, 1994).

In addition to c-NHEJ, alternative end-joining (alt-EJ, including microhomology-mediated end joining (MMEJ)) repair refers to repair events that are independent of c-NHEJ factors such as Ku proteins and is instead mediated by PARP1 and DNA polymerase θ (Pol θ) following resection of the 5' ends of the break (see **Figure 1.6B**) (Howard, Yanez and Stark, 2015). Alt-EJ is considered highly mutagenic and its frequency increases as a result of the loss of c-NHEJ, a feature it shares with HR and SSA (discussed in the next section). One hallmark of alt-EJ is microhomology, although this is not absolutely essential for alt-EJ (Bennardo *et al.*, 2008) and can also be used by c-NHEJ (Pannunzio *et al.*, 2014). In general, alt-AJ is considered important for completing repair of aborted HR initiation events, e.g. in the absence of a sister chromatid, and thereby guard against chromosome loss (Howard, Yanez and Stark, 2015).

However, as mentioned before, the relative contribution of each DDR pathway remains largely uncharacterised.

1.6.1.3 Homology-directed repair

HDR comprises at least three sub-pathways: single strand annealing (SSA), synthesis-dependent strand annealing (SDSA) and homologous recombination (HR) (see **Figure 1.6B-D**). In contrast to NHEJ, HDR typically involves many more protein factors and extensive resection of the DNA ends to uncover sequence homology (Lieber, 2010); whereas NHEJ requires <4 bp homology, SSA requires >20 bp and the most conservative pathways (i.e. HR and SDSA) require >100 bp in length (reviewed in Chang *et al.*, 2017). HDR is the preferred pathway over NHEJ to minimise eventual scarring of the DNA. However, as with much of the work in the DNA repair field, it is important to note here that much of this work has been performed in *S. cerevisiae* and the extent to which these findings can be extrapolated to other species (and their specific cell types) remains to be determined. Furthermore, whereas HDR is an umbrella term covering SSA, SDSA and HR, in relation to genome editing, the abbreviation HDR is typically used for the introduction of an exogenous sequence.

During all HDR pathways, 5'-end resection results in 3' ssDNA overhangs, which are protected from further resection and secondary structure by Replication Protein A (RPA) (see **Figure 1.6**). RPA mediates the formation of a protein complex at the 3'-end including Rad51, Rad52 and BRCA2, after which Rad51 scans the rest of the genome for homology (Mehta and Haber, 2014). With help of Rad54, Rad51 catalyses homology-search, strand-pairing and strand-exchange.

SDSA requires a single DNA strand to find its complementary strand, which is used as a template for DNA polymerase for incorporation. A holiday junction is then formed from the 3'-end, invading the homologous template (see **Figure 1.6C**), after which DNA polymerase resynthesises the strand.

HR is highly regulated and also requires homology on the sister chromatid or exogeneous, double-stranded DNA donor as a template in genome editing experiments (see **Figure 1.6D**)(Bothmer *et al.*, 2017; Kan *et al.*, 2017). In contrast to SDSA, it forms a double holiday junction, thus requiring two strand invasions. The result is perfect repair of the DNA and can

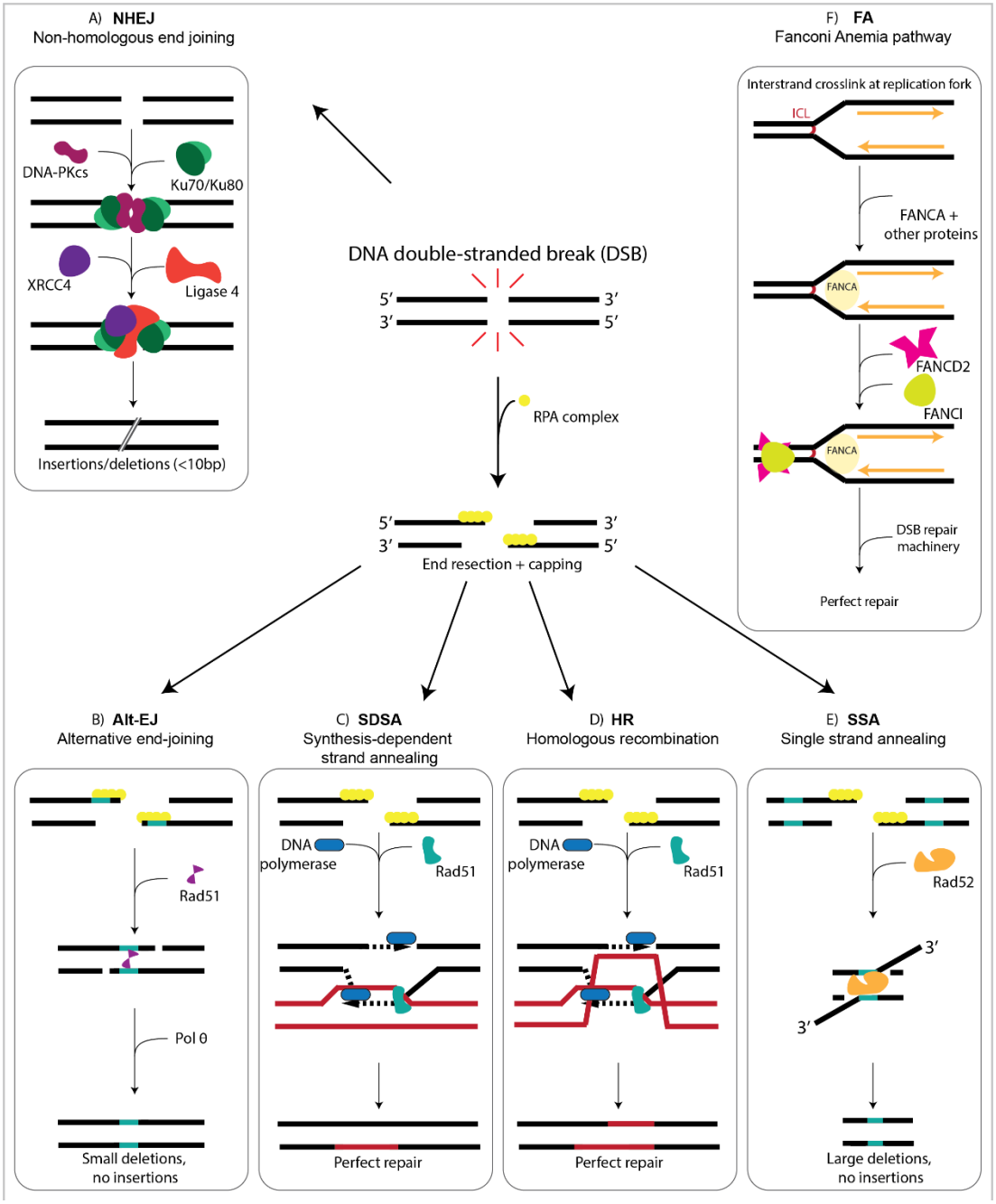


Figure 1.6 DNA double-strand breaks can be repaired through different repair mechanisms
When a double-strand break (DSB) occurs in a cell, different mechanisms result in different repair outcomes. **(A)** Ku70 and Ku80 proteins are typically present at high abundance and form heterodimers that bind at the DSB, where they form a DNA-PK holoenzyme that stabilises the two ends of the break together with DNA-PK catalytic subunit (DNA-PKcs). Once stabilised, these ends are either trimmed or filled in by respectively nucleases or DNA polymerases (not shown), after which XRCC4 and Ligase 4 together directly ligate the ends. NHEJ typically results in short (<10bp) insertions or deletions. In case Ku proteins do not bind to the DSB ends, both of the 5' ends are resected, creating free sticky ends, which are stabilised by capping with RPA proteins. The restricted ends then allow for SDSA, HR or SSA to repair the ends. **(B)** Alt-EJ is thought to comprise a pathway that, in absence of a sister

chromatid, reanneals the resected strands, leading to frequent or the loss of genetic information. One form of alt-EJ, MMEJ, can involve microhomology to guide the reannealing of the broken strands. **(C)** SDSA involves a set of proteins including Rad51, which bind to one of the free 3' ends and allows to search for homology, either on the sister chromatid (in red) or an introduced DNA template. Through the formation of this single Holliday Junction by Rad51 and co-factors, the homologous template sequence is copied in, after which the strand is ligated. In the meanwhile, the other 5' end is resynthesised. **(D)** Similar to SDSA, HR invades the sister chromatid, but contrastingly forms a double Holliday Junction., thereby copying in the sister chromatid on both strands. In addition to synthesis, a cross-over between the two sister chromatids can occur, through which the sister chromatid is also altered (not shown). **(E)** With SSA, a homologous region on the same strand (in cyan) is used by Rad52 to re-ligate the two strands together, thereby creating 3' overhangs, which are resected by nucleases, after which a ligase reattaches the strands. SSA results in the loss of information initially present between the two homologous regions. **(F)** The Fanconi Anemia (FA) pathway is involved in the repair of interstrand crosslinks (ICL) and stalled replication forks. Upon binding of FA-associated proteins including FANCA to the damaged site, these factors recruit and activate (ubiquitinate) FANCD2 and FANCI, which in turn can recruit other DDR factors.

hence be utilised to yield precise indels or substitutions by the introduction of repair templates with homology to the target sites.

Contrastingly to SDSA and HR, which both require Holliday Junctions and a sister chromatid or exogenous template with homology, SSA promotes annealing between interspersed stretches of homologies placed on the same chromosome (see **Figure 1.6E**), resulting in the loss of the intermediate sequence, and thus results in the loss of genetic information (Ceccaldi, Rondinelli and D'Andrea, 2016). In mammals, SSA is suppressed by other repair pathways and hence serves as a back-up pathway, and as such is considered to be suitable for repair of regions with tandem repeats (Dueva and Iliakis, 2013). Contrastingly, it has to be noted that SSA in *S. cerevisiae* is highly efficient (Ivanov *et al.*, 1996), illustrating the difficulty in generalising research findings from one system to another.

The Fanconi Anemia (FA) (see **Figure 1.6F**) pathway is a fourth DDR mechanism involving homology, which recently gained increased interest due to its involvement in single-strand template repair (SSTR) from ssODNs (Richardson *et al.*, 2018). SSTR is a poorly defined mechanism thought to be similar to SDSA, although using a single-stranded rather than a double-stranded template and being Rad51 independent (Gallagher and Haber, 2018). Normally, FA is associated with interstrand crosslink repair and stalled replication forks (reviewed in Palovcak *et al.*, 2017), but it is thought that Cas9-stimulated repair with an

ssODN mimics a stalled replication fork (Richardson *et al.*, 2018). Recruitment of FA-associated core proteins promotes the localisation of the FANCD2/FANCI heterodimer to Cas9-induced DSBs, where it is proposed to act as a 'traffic light' as to whether a DSB should be repaired through SSTR (Richardson *et al.*, 2018). As these are very novel findings and hypotheses, further investigations should elucidate the involvement of the FA pathway in the repair of DSBs from ssODNs.

1.6.1.4 Ratio of NHEJ to HDR

As sister chromatids are not produced until S phase, in a normal situation (i.e. when no exogenous homologous DNA template has been introduced) HDR can only take place in S or G2 phase in mitotic cells (Saleh-Gohari and Helleday, 2004). Whilst DNA damage in non-dividing, terminally differentiated cells gives rise to DNA damage repair, proliferating cells first stall the cell cycle at different checkpoints in order to repair DNA damage prior to cell division (Zhou and Elledge, 2000). As Ku proteins are typically present at high abundance in cells and prevent extensive restriction of the DNA-ends, repair through NHEJ is typically favoured (Mimitou and Symington, 2010). This skew towards NHEJ is likely to be a time constraint on damage control, as NHEJ is a rapid mechanism (several minutes), relatively error-free (Wang *et al.*, 2001) and thus limits the time that a DSB exists and thus the amount of additional damage occurring in the genome.

Cells in G1 favour NHEJ over HDR by more than 50-fold, whereas in other cell cycle phases this ratio is estimated to be closer to 4:1 (Beucher *et al.*, 2009; Chang *et al.*, 2017) (Chang *et al.*, 2017). Therefore, the rate of HDR differs between cell types and is typically out-numbered by NHEJ, typically resulting in HDR efficiencies < 5% after cleavage with Cas9 (Komor, Badran and Liu, 2017). This is currently a major limitation in the field of precise genome editing. Several chemical approaches are shown to have positive results on increasing the HDR to NHEJ ratio, for example, by using Scr7 to inhibit a DNA ligase (Lig4) necessary in NHEJ (Vartak and Raghavan, 2015) or by enhancing Rad51 involved in HDR using a dsDNA template (Pinder, Salsman and Dellaire, 2015), although these processes themselves are cell type and cell state dependent and are hence not broadly applicable. Further optimisation efforts are discussed in section 1.6.3.

1.6.2 Single-stranded break repair

By default Cas9 creates a blunt double-strand break, but interference with either of the nuclease domains, SpCas9 (HNH ($Cas9^{N863A}$) or RuVC ($Cas9^{D10A}$) (see **Figure 1.5B**)), results in the development of a nickase (Cas9n) variant that creates a single-stranded DNA break on either the target or complementary strand (F. Ann Ran *et al.*, 2013; Mali *et al.*, 2013). Damage to a single base can result naturally from several sources of DNA damage including ROS and chemicals inherent to cell metabolism, making this by far the most common form of DNA damage in a cell (Swenberg *et al.*, 2011). Whereas DSBs pose a complex challenge as both of the DNA strands are broken, the information required for correct repair of a nick is present on the intact strand. Several repair pathways are present to approach specific types of single-stranded breaks, although not mentioned in detail in this thesis.

The repair of a single-stranded nick is at a high-fidelity relative to DSB repair, with Cas9n creating genomic disruptions at lower frequencies than wildtype Cas9 (Hsu, Lander and Zhang, 2014). The use of two sgRNAs in opposite orientation at close proximity (paired nickases) however creates DSBs with sticky ends due to 5' resection (F. Ann Ran *et al.*, 2013; Mali *et al.*, 2013), which are thought to promote annealing with homologous strands when PAM sequences are outwards-oriented in case of a $Cas9^{D10A}$-induced paired nickases resulted in a 5' overhang (Bothmer *et al.*, 2017). Inwards-oriented, 3' overhangs induced by $Cas9^{N863A}$ theoretically create the same result, but the kinetics of Cas9-binding as such are thought to result in recognition of the breaks as two separate nicks rather than a single DSB (Bothmer *et al.*, 2017).

1.6.3 Effects of different donor templates on HDR

To utilise HDR following Cas9-induced breaks to yield specific modifications or large integrations at the target sites, a homologous donor template is required. While historically double-stranded DNA (dsDNA) with long homology-arms (500 - 1000nt) was often used as the donor template in CRISPR-Cas9 editing (Byrne *et al.*, 2014), it is considered inefficient in many cell types and requires Rad51-dependent HR, which is only active in S/G2/M phases (Chen *et al.*, 2011).

More recent studies demonstrated that single-stranded oligodeoxynucleotides (ssODNs) promote higher insertion rates compared to dsDNA whilst requiring shorter homology arms,

whilst also reducing toxicity and being effective at lower molarities (Miura *et al.*, 2015; Yoshimi *et al.*, 2016). In some cases, their use can increase HDR efficiencies up to 500-fold compared to dsDNA donors (Ferenczi *et al.*, 2017). ssODNs are typically used to introduce or modify sequences up to 200 bp, using 30-60 nt homology arms (Jacobi *et al.*, 2017). Long, single-stranded DNA (ssDNA) donors have more recently also shown their utility in introducing long stretches of DNA, without the need for long homology arms (Yoshimi *et al.*, 2016). In fact, perhaps surprisingly, longer homology arms in ssODNs were shown to reduce HDR efficiencies (Paix, Schmidt and Seydoux, 2016) and are optimal at 30-40 bp (Liang *et al.*, 2017).

When using paired nickases with outwards facing PAMs, efficiencies are enhanced up to 60% by binding of ssODNs to the strand opposite to the PAM (i.e. the strand complementary to the sgRNA) improve HDR efficiencies (Christopher D. Richardson *et al.*, 2016). The same study showed that asymmetric donor DNA with a longer PAM-proximal than PAM-distal homology arm (optimally 91-nt against 36-nt) even further enhances HDR efficiencies, indicating that the first-created 5' overhang created by the D10A paired nickases binds to the ssODNs, which is consistent with the hypothesis that ssODNs are copied in through SDSA rather than through direct integration (Liang *et al.*, 2017; Paix *et al.*, 2017), although a very recent study identified the FA pathway to be the pathway crucial for integration of ssODN (Richardson *et al.*, 2018). Their findings indicate that single-stranded DNA templates are incorporated in a Rad51-independent, FA-dependent manner, which indicates that Rad51-dependent SDSA cannot be the main repair pathway for ssODN-mediated repair.

In addition to the use of ssODN and the means discussed in section 1.6.1.4, other ways to improve HDR efficiencies have been sought after. For example, as degradation is thought to deplete the ssODN pool before DNA cleavage by Cas9 has occurred, elongating of the half-life of ssODNs by phosphorothioate-modification has been reported to increase HDR efficiencies, especially for long inserts (Renaud *et al.*, 2016). Another study demonstrated that Cas9-induced breaks can incorporate ssODN sequences without the need for HDR, termed homology-independent targeted integration (Suzuki *et al.*, 2016), although the rates for this are still considerably low (± 5%). Hence, precise editing efficiencies are slowly improving, continually improving the exact tailoring of precise genome editing to one's need.

In the meantime, as our knowledge of CRISPR–Cas9 gene editing and DNA repair is increasing, it is becoming apparent that the picture is not as clear and straightforward as we thought. With time, it will become more evident which of the combinations of nuclease and donor template are most suitable for each application, although at this stage single stranded donors appear to be most advantageous for many applications. Similarly, increasing our knowledge of the molecular mechanisms underlying precise genome editing may allow the outcome of cellular DNA repair pathways to be manipulated.

1.7 Genotype-phenotype screening with CRISPR-Cas9

As precise genome editing through CRISPR-Cas9 has become feasible in a wide range or targets and has shown to be easy to repurpose due to its modularity, it has been applied in many ways to study the effects of genomic variants on phenotypes. Hereafter I will discuss several means by which Cas9 is exploited for screening purposes.

1.7.1 Genome-wide screens mediated by CRISPR-Cas9

When a Cas9-induced double-strand break results in a frameshift or premature stop codon, this often results in a destabilised protein or nonsense-mediated decay of the mRNA transcript. As both of these processes result in similar gene-knockout, CRISPR-based genome-wide knockout screens (CRISPRi) rapidly developed into the gold standard for the gain-of-function and loss-of-function screening of gene functions as it proved more effective than traditional RNAi, as in the latter non-sense mediated decay (NMD) of the target RNA is not always entirely efficient (Shalem *et al.*, 2014; Wang *et al.*, 2014; Zhou *et al.*, 2014).

Moreover, a nuclease-deficient Cas9 variant (dCas9) for targeted recruiting of the Cas9-bound proteins to a target site without inducing a cut has been developed (Qi *et al.*, 2013), which can also be used in genome-wide screens without introducing permanent destabilising mutations (i.e. transient) when coupled to a transcriptional repressor or activator (Qi *et al.*, 2013; Konermann *et al.*, 2015). Hence, Cas9 can be utilised in a broad spectrum of applications to perform genomic interrogations, although these screens – as discussed in section 1.3 – only allow for loss of function or overexpression phenotypes, and not the systematic assessment of amino acid substitutions.

1.7.2 Saturation mutagenesis using CRISPR-Cas9

The advantage of using saturation mutagenesis in combination with Cas9-based genome editing is capturing the native context of the genetic element in contrast to episomal introduction or random integration. Site-directed mutagenesis of mouse and human enhancers using CRISPR-Cas9, whereby all individual nucleotides in the genomic region are disrupted through NHEJ, allows for the screening of functional mutations in non-coding genetic elements. By disrupting sequences within an enhancer in its endogenous context, essential elements can quickly be distinguished from other nucleotides (Canver *et al.*, 2014). Such site-directed mutagenesis using Cas9-induced indels is an effective approach to study the function of genetic elements, with targeting efficiencies up to 88% in *Drosophila* (Bassett *et al.*, 2013), although up to ± 40 % is typically reached in mammalian cells (Canver *et al.*, 2014; Wu *et al.*, 2017). Whilst effective in non-protein-coding functional elements, structure assessments of protein-coding genes are less suitable as for reasons mentioned in section 1.7.1, for which edits resulting in substitutions and hence not leading to frameshift mutations are more suitable.

1.7.3 Base editors

For the introduction of single nucleotide substitutions with Cas9, HDR was long thought to be the best approach since NHEJ frequently leads to indels. However, as the process of HDR can be inefficient in certain (non-dividing) cell types (Saleh-Gohari and Helleday, 2004), recent developments in the field of genome diversification have circumvented this process and resulted in Cas9n or dCas9-mediated base editors (BEs) (Hess *et al.*, 2016; Komor *et al.*, 2016). The most-used third generation of BEs recruits a single-strand-specific cytidine deaminase such as APOBEC that induce a C>U conversion in a 5-nt window at the PAM-distal end of the protospacer without introducing DSBs (Komor *et al.*, 2016). The addition of an uracil glycosylase inhibitor impedes base excision repair (BER), thereby favouring the permanent C>T conversion. Finally, by introducing a nick on the unedited strand, the DNA repair process resects the G-containing DNA strand and repairs it using the edited strand as a template. By directed evolution of RNA adenosine deaminase, a seventh generation of BEs termed adenosine base editors (ABEs) broadened the scope of base editors by allowing A>G

conversions in target DNA (Gaudelli *et al.*, 2017), thus allowing for all base transitions when combined with third generation BEs.

Comparable approaches to BE include (1) CRISPR-X, which involves dCas9 linked to MS2 RNA hairpins that recruit an MS2-binding protein fused to a hyperactive variant of activation-induced cytidine deaminase (AID) to the target site (Hess *et al.*, 2016), (2) AID fused to a ZFP or TALE (Yang *et al.*, 2016) and (3) an AID ortholog (PmCDA1) directly fused to dCas9 or nCas9 (Ma *et al.*, 2016; Nishida *et al.*, 2016). Base editing is hence becoming a frequently-used method for both the introduction of precise genome edits and targeted genome diversification.

As base editing rates are typically much higher than HDR rates, it has been used in recent DMS approaches (discussed in section 1.3.4). In addition, as no DSBs are introduced, there is typically no loss or gain of nucleotides. Therefore, BE is very effective for applications aiming to yield specific nucleotide changes. However, as the targeting window typically covers only a few bases adjacent to an NGG PAM motif (except CRISPR-X), its applicability is best suitable for the introduction of precise edits. Later studies addressed some of these issues by altering PAM compatibilities (Kim *et al.*, 2017) and enhancing on-target specificity (Rees *et al.*, 2017), although (intentionally) retaining a small window and only a single base conversion per experiment and as such has a limited applicability in DMS. As APOBEC requires ssDNA as a substrate, APOBEC-mediated diversification will probably remain limited to the protospacer region. In addition, xCas9 and spCas9-based base editing have been developed to broaden the range of targets and contexts in which these principles can be applied, as are discussed in section 1.5.3.

1.7.4 Nucleotide diversification through Cas9-guided polymerases

A very recent study describes an alternative to BE by guiding an error-prone, nick-translating DNA polymerase to a target site by Cas9n, thereby extending the targeting window up to 350 nt (Halperin *et al.*, 2018). This system, termed EvolvR, induces a nick at a specific site, after which the polymerase displaces the strand downstream of the nick. The window length, mutation rate and substitution bias are controlled by the fidelity and processivity of the polymerase variant used, which increase the mutagenesis window compared to BE by 70-

fold, its activity has so far only been demonstrated in *E. coli* and is thus not yet available for mammalian cells.

1.7.5 Deep mutational scanning using CRISPR-Cas9 and multiplex HDR

DMS (discussed in section 1.3.4) combined with CRISPR-Cas9 and multiplex HDR allows for broadening of the targeting region by using repair template pools that are saturated with (nigh to) all possible nucleotide variants without the loss or gain of nucleotides. Several studies have been published, of which I will list several here to give an overview of accomplishments and limitations.

Findlay and colleagues were one of the first to publish results on DMS with CRISPR-Cas9, demonstrating the assessment of the effects all possible SNVs across 6 nucleotides in an exon of *BRCA1* on transcriptional abundance, through the use of CRISPR-Cas9 and multiplex HDR using a pool of repair templates harbouring all possible variants (Findlay *et al.*, 2014a). They subsequently applied this approach to assess the effects of mutations in a conserved region of *DBR1* on proliferation, thereby proving its efficacy in assessing the phenotypic consequences of SNVs. The HDR libraries used for this study were plasmids with long (2 x ± 750 bp) homology arms to the target sites, generated using partially degenerate oligonucleotides, as explained in section 1.3.7. Despite their HDR rates comprising only ± 1-3 % in human embryoid kidney (HEK293T) cells, their repair templates introduced silent mutations that created PCR primer binding sites, such that they selectively amplified reads from loci that had undergone HDR, thus compensating for low HDR frequencies.

Similarly, CRISPR-Cas9-based mutagenesis was used to perform screenings for SNVs in *BRC-ABL1* underlying drug resistance (Ma *et al.*, 2017). Here, instead of degenerate oligonucleotides, a cassette ligation strategy (i.e. the ligation of several fragments with overhanging sequences)in a bacterial cloning plasmid was used to create library diversity including PCR, ligations and bacterial transformations, although this approach limited its targeting window to 30 nt. In the murine cancerous B-cell (Ba/F3) line they achieved 8% HDR, which for dsDNA donors is on the higher end. A recent study utilised site-directed mutagenesis with help of CRISPR-Cas9 in engineering antibodies in murine naïve B-cells (Mason *et al.*, 2018) using ssODN libraries 120nt in length, harbouring twelve NNK and NNS

degenerate codons (explained in section 1.3.7). Elegantly, they were able to achieve HDR efficiencies up to 36% by knockout of 53BP1, thereby disfavouring NHEJ.

One of the most elaborate studies describing Cas9-mediated DMS thus far interrogated the functional impact of 9,833 variants in the DNA-binding domain of *TP53*, the most frequently mutated gene in cancer (Kotler *et al.*, 2018). As libraries, synthetic ssODNs 200 nt in length were rationally designed to contain all possible single and dinucleotide (e.g. TT > CC) changes, premature stop codons, amino acid substitutions and deletions and commonly occurring SNVs across 1,160 nucleotide sites, and cloned into plasmid backbones. This study elucidated selective pressure on specific mutations in both evolution and tumorigenesis and derive a fitness score per mutation that corelates with their incidence in tumours. Interestingly, the authors identified SNPs appearing to be neutral in isolation, but which exhibited positive synergistic effects when cooccurring with other missense mutations, emphasising the merit of in-depth scanning of mutations.

Similarly, all possible TP53 variants were screened by Giacomelli *et al.* through parallel introduction of all possible amino acids and a stop codon at each codon position (Giacomelli *et al.*, 2018). Through massive parallel sequencing, it was determined that >99.8% of all mutant alleles were generated and assessed using 150-base oligonucleotides with 30 bases of complementary sequences flanking a 90-base variable region. Lentiviral transfection in isogenic cells with respectively $p53^{WT}$ and $p53^{NULL}$ allowed for the analysis of LOF alleles and depletion of alleles with wildtype-like activity. Their comprehensive study was related to transcriptional reporter assays, evolutionary conservation and mutational signatures found in tumours, making this study a comprehensive landmark in the field of DMS.

Finally, two recent paper describes the use of saturation genome editing to assay nearly 4,000 SNVs (96.5% of all possible SNVs) across 13 exons of the BRCA1 tumour suppressor gene (Findlay *et al.*, 2018; Starita *et al.*, 2018) to identify over 400 non-functional missense variants and 300 SNVs that disrupt expression. BRCA1 is essential in the haploid cell line HAP1, and their relative enrichment over time was hence used as proxy for their effect on cell survival. The authors used plasmid with homology arms 600-100bp in length and a fixed

mutation destroying the PAM, thereby preventing recleavage of edited alleles. With knockout of LIG4, a protein involved in NHEJ, increased HDR rates by 3.6-fold up to 75%.

These studies show that CRISPR-Cas9 with multiplex HDR can be used to generate endogenous libraries of sufficient size for both DMS and directed evolution, allowing for an unpreceded throughput and precision when combined with NGS. In contrast to base editors, where one type of base conversion can be achieved per experiment in a small window, saturation mutagenesis with multiplex HDR can achieve all possible single nucleotide or codon changes across a much larger window. Several techniques such as selective PCR and inhibitory drugs are deployed to improve HDR frequencies, reaching up to 36% (Mason *et al.*, 2018). As advances in the field of CRISPR-Cas9 progress, repair outcomes can be skewed to favour HDR over NHEJ, thus increasing the yield of possible variants that can be screened.

1.8 Thesis Objectives

There is a wide abundance of targets and approaches for DMS, which have proven powerful in providing a high-throughput screen for the phenotypic consequence of mutations frequently observed in nature. At the time of initiation of our study, none of the above screening methods were developed and published. In this thesis, I will therefore develop an experimental and data analysis pipeline to perform DMS using CRISPR-Cas9 and multiplex HDR in endogenous genes, which will provide a benchmark for oncogenic mutations found in β-catenin. The overall aims of this thesis are as follows:

- To develop experimental and bioinformatics pipelines to perform deep mutational scanning on single nucleotide and amino acid changes in chromosomally encoded genes with CRISPR-Cas9.
- To validate this approach for the assessment of single nucleotide variants using genomically-integrated green fluorescent protein (GFP) in mouse embryonic stem cells.
- To utilise our approach to assess the effects of amino acids substitutions in an endogenous gene to provide a benchmark for SNVs observed in tumours by using β-catenin as a proof-of-principle.

2

Materials & Methods

2.1 GFP

2.1.1 Embryonic stem cell culture

The RCNβH-B(t)-GFP cell line used in Chapter 4, also known as the RCN(t)-GFP cell line was previously characterised as an E14-derived mouse embryonic stem cell line with a single copy random integration of a *pCAG-Gfp* transgene with uniformly high GFP expression and fluorescence (Chambers *et al.*, 2007) (see **Figure 4.2A**). Mouse embryonic stem (ES) cells were retrieved from liquid nitrogen by rapid thawing in a 37°C water bath, after which they were resuspended in 9ml pre-warmed E14 media (GMEM (Life Technologies), 15% fetal calf serum, 1% L-glutamine, 1% penicillin/streptomycin (P/S), 1% sodium pyruvate, 1% non-essential amino acids, 0.1% β-mercaptoethanol and 0.2% LIF-conditioned media) and pelleted by centrifugation (at 300 x *g* for 5 minutes) to remove DMSO. These cells were cultured at 37°C with 5% CO_2 in pre-gelatinised culture flasks. Cells were passaged by 2x PBS wash and trypsinisation for 3 minutes at 37°C. 10 volumes of pre-warmed media were added to the cell-containing trypsin, after which the suspension was spun down at 300 x *g* for 5 minutes. Cells were then resuspended in E14 media and resuspended at approximately 1.0 x 10^5 cells/cm^2. For cryopreservation, cells were frozen at 4.0 x 10^6 cells / ml by adding 500µl of CryoStor® cell cryopreservation media CS5 (Sigma-Aldrich) to 500µl cell suspension at 8.0 x 10^6 cells/ml and stored at -70°C in a polystyrene box for 24 hours prior to long term storage in liquid nitrogen at -150°C.

2.1.2 Plasmid construction and validation

For targeting of cells with CRISPR-Cas9 (see section 4.2.6), sgRNA sequences (see **Table 2-1B**) were designed using the online version of the Azimuth sgRNA design software (Doench *et al.*, 2016) and cloned into plasmids utilising overhangs with sequence similarity to BbsI restriction site according to standard protocol (F Ann Ran *et al.*, 2013). For targeting experiments with Cas9 nickases, sgRNAs were cloned into pX335-U6-Chimeric_BB-CBh-hSpCas9n(D10A). For targeting with wild-type Cas9, sgRNAs were cloned into pSpCas9(BB)-2A-Puro (PX459) V2.0.

Plasmids were transformed into subcloning efficiency DH5α-competent *Escherichia coli* (*E. coli*) (Invitrogen). Cells were thawed on ice and 50µl of cell-suspension was mixed with 2µl of ligated plasmids and incubated for 3 minutes on ice. Cells were heat-shocked for 30 seconds at 42°C, left for 3 minutes on ice and resuspended in 250µl of Super Optimal broth with

Catabolite repression (SOC) media. The cell suspension was then incubated for 1 hour at 37°C while shaking at 400rpm before being spread on a LB-plate with 50µg/l ampicillin and left overnight at 37°C for selection of successful transformation.

Individual *E. coli* clones were picked from plates, transferred to 5ml LB medium with 50µg/l Ampicillin and left at 37°C while shaking at 225rpm overnight. Plasmids were extracted using the Plasmid Spin Miniprep Kit (QIAGEN) following manufacturer's instructions. Successful cloning of sgRNAs was confirmed by Sanger sequencing of the plasmid sequences using primers 'CRISPR_plasmids_U6_FP' and 'CRISPR_plasmids_RP' described in **Table 2-1A**.

2.1.3 sgRNA validation using CEL1 surveyor assay

For assessment of cleavage of target DNA by Cas9 using specific sgRNAs (see section 4.2.4), mutation rates were assessed by the Surveyor nuclease assay (F Ann Ran *et al.*, 2013) with several modifications. ES cells were transfected as described in section 2.1.6, after which genomic DNA extracted as described in section 2.1.8. After amplification of the target locus by PCR using 'GFP_CEL1' primers (see **Table 2-1A**) and reannealing of amplicon sequences, cleavage of mismatch DNA was performed by 7µl celery extract purified by Gillian Taylor. Bands were subsequently separated by standard gel electrophoresis on 2.5% (w/v) TBE agarose gels and stained with SYBR Gold (1 : 10,000, Thermo Fisher) for 45 minutes while agitated before visualisation.

2.1.4 ssODN repair template synthesis

ssODN repair templates used in Chapter 4 (described in

Figure *4.3*) were designed to be 100 nt in length with incorporation ratios of 94% for the genomic homology consensus nucleotide and 2% for each of the remaining nucleotides, or 97% and 1% for the respective classes, as further described in section 4.2.2 and synthesised by IDT. This resulted in four ssODN classes as described in **Table 2-1C**.

2.1.5 UltraRT construction

To synthesise long (820 bp) double-stranded repair templates with variable regions, 100nt ssODNs were used as a template. Two pairs of Ultramer primers 200 nt in length (see **Table 2-1A**) were designed with homology to the GFP locus (depicted in figure 4.16) and used in a two-step PCR amplification of the ssODN pool. First, a 10 µl PCR reaction with 0.5 µM ssODN

template DNA, 1 µM GFP_Ultra_F1 , 1 µM GFP_Ultra_R1 and 2X Phusion® High-Fidelity DNA Polymerase Master Mix (New England Biolabs) was prepared and placed in a thermal cycler (Tetrad) optimised to: initial denaturing 95°C 2 minutes, then 10 cycles of: second denaturing 95°C 30 seconds, annealing 55°C 30 seconds, extension 72°C 1 minute, before final extension 72°C 5 minutes. Correct band size (460 bp) of PCR products was subsequently separated by standard gel electrophoresis on 1.5% (w/v) TBE agarose gels, after which the product was cleaned with the QIAquick PCR Purification Kit (QIAGEN). The final product was used in a second PCR reaction at 0.5µM, with 1µM GFP_Ultra_F2, 1 µM GFP_Ultra_F2 and 2X Phusion® High-Fidelity DNA Polymerase Master Mix (New England Biolabs) at the same thermocycler protocol as the first reaction. After the second reaction, bands of 820bp were confirmed on agarose gels and extracted using the QIAquick Gel Extraction Kit (QIAGEN) before being used as repair templates for homology-directed repair.

2.1.6 Cell transfection

For transfection, cells were grown to 70% confluency to ensure logarithmic growth, after which 2.0 x 10^6 cells were seeded in a pre-coated well of a 6-well plate with 2ml of media. For transfection with paired Cas9 nickases, 2µg plasmid DNA (0.5 µg pSUPER.retro.puro and 0.75 µg px335 plasmid with sgRNA8 and 0.75µg px335 plasmid with sgRNA17) was added to 100µl Optimem (Invitrogen), after which 6µL FuGENE® HD (Promega) was added and mixed by pipetting. The FuGENE reagents pack the target DNA in vesicles that fuse with the cytoplasmic membrane, thereby delivering the vectors directly into the cell. For transfection with Cas9 nucleases (i.e. wtCas9), the 2 µg plasmid DNA consisted of 0.5 µg pSUPER.retro.puro and 1.5 µg px459 plasmid with the respective sgRNA (sgRNA8, sgRNA17 or sgRNA100, see **Table 2-1B**) After 10 minutes of incubation at room temperature the mixture was added dropwise to the well and cells were incubated with the mixture for 24 hours at 37°C. The media was then replaced by media containing 1.6µl/ml puromycin (Gibco) which was incubated for 24 hours, selectively killing cells that had not been transfected, after which this media was replaced by fresh media. Cells were cultured for 8 days post-transfection before being separated by fluorescence activated cell sorting (see section 2.1.7).

2.1.7 Fluorescence-activated cell sorting

Prior to fluorescence-activated cell sorting (FACS), cells were trypsinised as described in the previous section and pipetted up and down to ensure a single cell suspension. Cells were

subsequently stored on ice prior to sorting. Cells were subsequently analysed and sorted on the FACSAria™ (Becton Dickinson) cell sorter into respectively three populations based on GFP fluorescence. The GFP negative population was defined as the profile matching that of GFP-negative E14 embryonal stem cells, whereas GFP positive cells were those corresponding with untransfected RCN(t)-GFP cells. All cells falling in between these two populations were considered as intermediate.

2.1.8 Genomic DNA extraction

Genomic DNA was extracted from pelleted cells (centrifuged at 300 x *g* for 5 minutes) using the DNeasy Blood & Tissue Kit (QIAGEN) according to SOP. Quantification of DNA was performed using the Qubit Fluorometer (ThermoFisher) combined with the Qubit dsDNA BR (High Sensitivity) Assay Kit (ThermoFisher).

2.1.9 Digestions

The GFP target site was amplified in a 25 µl reaction with 1 µM 'GFP_157bp_F1', 1 µM 'GFP_157bp_R1', 50 µg genomic template DNA and 2X Phusion® High-Fidelity DNA Polymerase Master Mix (New England Biolabs) yielding a 157 bp product, which was subsequently placed in a thermal cycler (Tetrad) optimised to: initial denaturing 95°C 2 minutes, then 30 cycles of: second denaturing 95°C 30 seconds, annealing 58°C 30 seconds, extension 72°C 30 seconds, before final extension 72°C 3 minutes. Correct band size (460 bp) of PCR products was subsequently separated by standard gel electrophoresis on 1.5% (w/v) TBE agarose gels, after which the product was cleaned with the QIAquick PCR Purification Kit (QIAGEN). Subsequently, 1 µg of PCR amplicon was digested in a 50 µl reaction volume with 1 µl restriction enzyme (PsiI (New England Biolabs) or BsaXI (New England Biolabs), respectively) and 2X CutSmart® Buffer (New England Biolabs) for 2 hours at 37°C. Bands were subsequently separated by standard gel electrophoresis on 2.5% (w/v) TBE agarose gels and stained with SYBR Gold (1 : 10,000, ThermoFisher) for 45 minutes while agitated before visualisation.

2.1.10 Sequencing library preparation for genomic samples

For preparation of genomic samples for next-generation sequencing on the Illumina MiSeq platform, genomic loci were amplified and tagged in a two-step PCR as described in the Illumina '*16S Metagenomic Sequencing Library Preparation*' protocol. The first-round PCR

from 50 µg genomic DNA in a 25 µl reaction volume was performed using 2X Phusion® High-Fidelity DNA Polymerase Master Mix (New England Biolabs) with 3% DMSO and forward and reverse primers with Illumina adapters (see **Table 2-1A**, GFP_157bp_Illumina) at 0.5 µM in a thermocycler (Tetrad) with cycling parameters to standard protocol at 20 cycles. For direct sequencing of ssODN template pools, 'GFP_Oligo_Illumina_FP' and 'GFP_Oligo_Illumina_RP' were used. Clean-up using AmPure XP beads (Beckmann Coulter) was performed to standard protocol.

PCR amplicons were indexed by a second PCR round using Phusion® High-Fidelity DNA Polymerase Master Mix (New England Biolabs) with 3% DMSO and custom forward (i7) and reverse (i5) Nextera indexing primers (see **Table 2-1D**) at 0.25 µM with cycling parameters to standard protocol at 8 cycles. Indexed PCR amplicons were then quantified using the Qubit Fluorometer (ThermoFisher) and Qubit dsDNA BR (Broad Range) Assay Kit (ThermoFisher), and analysed by the Agilent 2100 Bioanalyzer system using high-sensitivity chips by the Wellcome Trust Clinical Research Facility. After quality control, samples were pooled equimolarly at 5 µM and sequenced by Edinburgh Genomics on the Illumina MiSeq platform at 150bp paired-end. Remaining template DNA was stored at -80°C

2.1.11 Determining *in silico* structure of GFP

The query GFP sequence was used to search for identical or related GFP structures using a BLAST (Altschul *et al.*, 1997) search against the Protein Data Bank (PDB) (Berman *et al.*, 2006) at the RCSB-PDB server (see section 4.2.14). Several GFP structure hits were returned from which the GFP crystal structure with the highest resolution (PDB ID: 2WUR; resolution 0.9 Å) was selected for analysis. Residue-specific solvent accessibility calculations were performed using ASAView (Ahmad *et al.*, 2004). Structure visualisation and analysis was conducted using PyMOL (http://www.pymol.org/; Schrödinger LLC, USA).

FoldX version 3b6 (Schymkowitz *et al.*, 2005) was used to predict energetic impact of mutations on protein stability, computed by subtracting the energy of the wild-type from that of the mutant (ΔΔG) in each case. In brief, the 'RepairPDB' option was used on PDB ID 2WUR (`FoldX --command=RepairPDB --pdb=2wur.pdb`) prior to mutating the residue using the 'Mutate residue' option under YASARA version 14.12.2 (Krieger, Koraimann and Vriend, 2002)) with the following parameters: Number of runs: 3; pH: 7; Temperature:

298 K; Ionic strength: 0.05 M; VdW design: 2. The resulting ΔΔG reported is the average of three runs. Typically, FoldX predicted ΔΔG values larger than 1.6 kcal/mol are considered highly significant and correspond to severely destabilising mutations (99% confidence interval), as they correspond to twice the standard deviation of the error in FoldX, while energy changes of > 0.8 kcal/mol are still considered significant (one standard deviation; 95% confidence interval) (Guerois, Nielsen and Serrano, 2002; Schymkowitz *et al.*, 2005; Rakoczy *et al.*, 2011; Kiel and Serrano, 2014).

2.2 β-catenin

All the experimental work performed on β-catenin, laying out the fundament to the bioinformatic work performed in Chapter 5, was entirely done by Anagha Krishna and Derya Özdemir from the Peter Hohenstein lab.

2.2.1 TCF/LEF cell line

mESCs with a TCF/LEF:: H2B-GFP reporter activity as described in (Ferrer-Vaquer *et al.*, 2010) was used as the parental cell line for saturation editing of β-catenin. Cells were cultured on gelatin-coated plates in 2i media as described in section 5.2.4.

To generate a β-catenin heterozygous knockout, the TCF/LEF cell line was transfected with a PuroΔtk selection cassette (pLCA.66/2272 plasmid; (Chen and Bradley, 2000)) and wtCas9 (pSpCas9(BB)-2A-Puro (PX459) V2.0) with integrated sgRNAs targeting intron 1 to intron 6 of the β-catenin-encoding gene *CTNNB1* (see **Table 2-1B**, CTNNB1_intron1_g1 and CTNNB1_intron6_g2), which were introduced through HDR in order to replace this fragment with a PuroΔtk selection cassette using puromycin (Gibco) positive selection for 24 hours (schematic can be found in Figure 5.2). Clonal cell lines were then validated for heterozygous knock-in by PCR and Sanger sequencing.

2.2.2 Repair plasmid construction

Double-stranded DNA oligonucleotides 200 bp in size containing all possible codon substitutions across 20 amino acid sites (L31-G50 of *CTNNB1*), BbSI restriction sites and PAM-synonymous mutations were synthesised by Twist Bioscience (see **Table 2-1**). To integrate the oligonucleotide library into a vector, a double-stranded backbone vector was generated by restricting the exonic region of CTNNB1 with type IIS restriction enzyme BbsI and having

5' and 3' β-catenin homology-arms using Gibson assembly. Double-stranded DNA oligonucleotides library were subsequently cloned into the previously generated β-catenin backbone vector by Golden Gate cloning (i.e. simultaneously and directionally assemble multiple DNA fragments into a single piece of DNA and validated by Sanger Sequencing.

2.2.3 Transfection

On day 1, 2.0 x 10^8 heterozygous knockout cells were cultured in 2i and transfected with the pooled targeting vectors, wtCas9 (pSpCas9(BB)-2A-Puro (PX459) V2.0) and sgRNAs targeting the PuroΔtk selection cassette (see **Table 2-1B**, CTNNB1_PKO_5'_g2 and CTNNB1_PKO_3'_g1). On day 3, media was changed to ES media containing LIF but without the 2i inhibitors to remove interference with Wnt/β-catenin signalling. FIAU negative selection was performed on this same day for 48 hours. On day 5, cells were FACS-sorted (see section 2.1.7) into 6 equally log-spaced bins based on GFP fluorescence (p2-p7). 200,000 cells were collected before sorting to make up the unbiased pool.

2.2.4 DNA extraction and PCR amplification

DNA from the unsorted cells and each FACS-isolated bin (p2-p7) was isolated using the using the DNeasy Blood & Tissue Kit (QIAGEN) and quantified using the Qubit Fluorometer (ThermoFisher, see **Table 5-1** for quantities) combined with the Qubit dsDNA BR (High Sensitivity) Assay Kit (ThermoFisher), after which it was used for further PCR amplifications. An initial PCR was performed with a forward primer outside the homology arms (Betacatenin_F1), to prevent amplification of random integration, and a handle-specific reverse primer (Betacatenin_R1), to selectively amplify clones that have undergone HDR from the template vector (see **Table 2-1A** for primer sequences). Next, amplicons from the first round of PCR were digested with DpnI (ThermoFisher) to get rid of any residual vector. The DpnI-digested PCR product was analysed on an 0.8% agarose gel after which correct band sizes were gel-eluted to further avoid amplification due to residual vector. This was followed by a second round of PCR (using the gel eluted first round amplicon as template) with an indexed forward primer (Betacatenin_F2; just outside the region of interest) and a primer (Betacatenin_R2). In parallel, another PCR was performed (referred to as HRM in Chapter 3) using the pooled DNA with the forward primer outside the homology arms (a described above) and a reverse primer outside the handle (Betacatenin_R3), after which the same

second round PCR step was performed. An overview of primer orientation can be found in figure 5.2a.

Second round PCR products were cleaned using AmPure XP magnetic beads (Beckmann Coulter), quantified using the Qubit Fluorometer (ThermoFisher) and Qubit dsDNA High Sensitivity Assay Kit (ThermoFisher). Samples were diluted to a concentration of 5nM and pooled, after which the integrity and concentration of the amplicons were checked using Bioanalyzer 2100. The pooled sample was sequenced by Edinburgh Genomics using 150bp paired-end MiSeq.

2.2.5 *CTNNB1* mutational likelihood analysis

Mutational likelihoods (discussed in section 5.2.16) were calculated by Ailith Ewing as follows. Tumour and normal (i.e. patient-derived non-tumour tissue) sequences were aligned to the reference exome (GRCh38), after which any mutations that occur in the tumour but not the normal were considered SNVs. Mutations were screened for by four SNV callers: MuSE, Mutect2, VarScan 2 and SomaticSniper (Larson *et al.*, 2011; Koboldt *et al.*, 2012; do Valle *et al.*, 2016). Only mutations that had been called by at least 2 callers were considered for further analysis. Samples were then filtered to only include those with a mutation in the analysed β-catenin region. Then, for the purposes of calculating relative frequencies and probabilities, non-synonymous mutations (i.e. missense and nonsense) were excluded in an effort to make these mutations mostly selectively neutral. Finally, all mutations in their trinucleotide context were counted across each of the tumour type to calculate their frequencies, which was subsequently used to calculate the probability of a codon change.

Table 2-1 Oligonucleotide sequences

(a) PCR primer sequences

Name	Target locus	Sequence	Purpose
Betacatenin_F1	Mm_CTNNB1	ATTGCCTTCGATGCGTCCGA	first round PCR all samples
Betacatenin_F2	Mm_CTNNB1	ATGGCCATGGAGCCGGA	second round PCR all samples
Betacatenin_R1	Mm_CTNNB1	GTCTTGATGATACCTCACAAG	first round PCR all samples
Betacatenin_R2	Mm_CTNNB1	TGTCAACATCTTCTTCTTCGGGA	second round PCR all samples, except HRM
Betacatenin_R3	Mm_CTNNB2	TTCATAAAGGACTTGGGAGGTGT	second round PCR HRM sample
CRISPR_plasmid_RP	PX335/PX459 plasmid	ACGACAGGTTTCCCGACTGG	Sanger sequencing for sgRNA integration
CRISPR_plasmids_U6_FP	PX335/PX459 plasmid	GACGTAATACGACTCACTATAGGGC	Sanger sequencing for sgRNA integration
GFP_157bp_FP	Mm_GFP	ACATGGAAGCAGCACGACTT	Digestions
GFP_157bp_Illumina_FP	Mm_GFP	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACATGGAAGCAGCACGACTT	MiSeq library preparation of genomic samples
GFP_157bp_Illumina_RP	Mm_GFP	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCGTCCTCCTTGAAGTCGATGC	MiSeq library preparation of genomic samples
GFP_157bp_RP	Mm_GFP	GTCCTCCTTGAAGTCGATGC	Digestions
GFP_CEL1_FP	Mm_GFP	CACATGGAAGCAGCACGACTT	CEL1 assay
GFP_CEL1_RP	Mm_GFP	CGTCCTCCTTGAAGTCGATGC	CEL1 assay
GFP_Oligo_Illumina_FP	ssODN/Mm_GFP	TCCAGGAGCGCACCATCTTC	MiSeq library preparation of ssODN
GFP_Oligo_Illumina_RP	ssODN/Mm_GFP	TTCAGCTCGATGCGGTTCAC	MiSeq library preparation of ssODN
GFP_Ultra_pair1_FP	Mm_GFP	GCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTACGGCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAAGCAGC	Creating dsONT from GFP ssONT

GFP_Ultra_pair1_RP	Mm_GFP	GTGTTCTGCTGGTAGTG GTCGGCGAGCTGCACGC TGCCGTCCTCGATGTTGT GGCGGATCTTGAAGTTC ACCTTGATGCCGTTCTTC TGCTTGTCGGCCATGATA TAGACGTTGTGGCTGTT GTAGTTGTACTCCAGCTT GTGC	Creating dsONT from GFP ssONT
GFP_Ultra_pair2_FP	Mm_GFP	AGGGGAGTGAGCTGGA TCCGGATAACTTCGTATA GCATACATTATACGAAG TTATCCTAGGGGATCCAC CGGTCGCCACCATGGTG AGCAAGGGCGAGGAGC TGTTCACCGGGGTGGTG CCCATCCTGGTCGAGCT GGACGGCG	Creating dsONT from GFP ssONT
GFP_Ultra_pair2_RP	Mm_GFP	AGCTTATCGAGCGGCCG CTTTACTTGTACAGCTCG TCCATGCCGAGAGTGAT CCCGGCGGCGGTCACGA ACTCCAGCAGGACCATG TGATCGCGCTTCTCGTTG GGGTCTTTGCTCAGGGC GGACTGGGTGCTCAGGT AGTGGT	Creating dsONT from GFP ssONT

(b) sgRNA sequences

Name	**Target locus**	**Sequence**
gRNA8	Mm_GFP	GCCGTCGTCCTTGAAGAAGA
gRNA17	Mm_GFP	CCGCGCCGAGGTGAAGTTCG
gRNA100	Mm_GFP	CAACTACAAGACCCGCGCCG
CTNNB1_intron1_g1	Mm_CTNNB1	GATGGAGTTGGACATGGCCA
CTNNB1_intron6_g2	Mm_CTNNB1	CAGGTGGGATGCAGGCACTG
CTNNB1_PKO_5’_g2	Mm_CTNNB1	TAGAGCCCACCGCATCCCCA
CTNNB1_PKO_3’_g1	Mm_CTNNB1	CCCCTACCCGGTAGTGAGGC

(c) ssODN donor templates. N1-N4 represent consensus nucleotides at 94% and the non-consensus nucleotides at 2%, whereby consensus in N1=A, N2=C, N3=G, N4=T.

Name	Target locus	Sequence
rGFP_VarSil_12nt	Mm_GFP	TCCAGGAGCGCACCATCTTCTTCAAGGACGACGGC(N1:94020202)(N1)(N2:02940202)(N4:02020294)(N1)(N2)(N1)(N1)(N3:02029402)(N1)(N2)(N2)CGCGCG(N3)(N1)(N3)(N3)(N4)(N3)(N1)(N1)(N3)(N4)(N4)(N2)GAGGGCGACACCCTGGTGAACCGCATCGAGCTGAA
rGFP_VarStop_12nt	Mm_GFP	TCCAGGAGCGCACCATCTTCTTCAAGGACGACGGC(N1:94020202)(N1)(N2:02940202)(N4:02020294)(N1)(N2)(N1)(N1)(N3:02029402)(N1)(N2)(N2)TTATAA(N3)(N1)(N3)(N3)(N4)(N3)(N1)(N1)(N3)(N4)(N4)(N2)GAGGGCGACACCCTGGTGAACCGCATCGAGCTGAA
rGFP_VarSil_100nt	Mm_GFP	(N4:01010197)(N2:01970101)(N2)(N1:97010101)(N3:01019701)(N3)(N1)(N3)(N2)(N3)(N2)(N1)(N2)(N2)(N1)(N4)(N2)(N4)(N4)(N2)(N4)(N4)(N2)(N1)(N1)(N3)(N3)(N1)(N2)(N3)(N1)(N2)(N3)(N3)(N2)(N1)(N1)(N2)(N4)(N1)(N2)(N1)(N1)(N3)(N1)(N2)(N2)CGCGCG(N3)(N1)(N3)(N3)(N4)(N3)(N1)(N1)(N3)(N4)(N4)(N2)(N3)(N1)(N3)(N3)(N3)(N2)(N3)(N1)(N2)(N1)(N2)(N2)(N2)(N4)(N3)(N3)(N4)(N3)(N1)(N1)(N2)(N2)(N3)(N2)(N1)(N4)(N2)(N3)(N1)(N3)(N2)(N4)(N3)(N1)(N1)
rGFP_VarStop_100nt	Mm_GFP	(N4:01010197)(N2:01970101)(N2)(N1:97010101)(N3:01019701)(N3)(N1)(N3)(N2)(N3)(N2)(N1)(N2)(N2)(N1)(N4)(N2)(N4)(N4)(N2)(N4)(N4)(N2)(N1)(N1)(N3)(N3)(N1)(N2)(N3)(N1)(N2)(N3)(N3)(N2)(N1)(N1)(N2)(N4)(N1)(N2)(N1)(N1)(N3)(N1)(N2)(N2)TTATAA(N3)(N1)(N3)(N3)(N4)(N3)(N1)(N1)(N3)(N4)(N4)(N2)(N3)(N1)(N3)(N3)(N3)(N2)(N3)(N1)(N2)(N1)(N2)(N2)(N2)(N4)(N3)(N3)(N4)(N3)(N1)(N1)(N2)(N2)(N3)(N2)(N1)(N4)(N2)(N3)(N1)(N3)(N2)(N4)(N3)(N1)(N1)
SRSF1-T7(linker)	Mm_SRSF1	TTACTCCCCAAGGAGAAGCAGAGGATCACCACGCTATTCTCCCCGTCATAGCAGATCTCGCTCTCGTACAGGATCCCCCGGCGCCGGCGCCATGGCATCGATGACAGGTGGCCAACAGATGGGTTAAGATGATTGGTGACACTTTTTGTAGAACCCATGTTGTATACAGTTTTCCTTTACTCAGTACAATCTTTTCA
SRSF1-NRS-T7 (BamHI)	Mm_SRSF1	AAGCAGAGGATCACCACGCTATTCTCCCCGTCATAGCAGATCTCGCTCTCGTACAGGATCCCCTCCGCCCGTGTCGAAGCGAGAGTCCAAGTCTAGGTCGCGGTCCAAGAGCCCACCCAAGTCTCCAGAAGAAGAGGGAGCAGTTTCTTCCATGGCATCGATGACAGGTGGCCAACAGATGGGTTAAGATGATTGGTGACACTTTTGTAGAACCCATGTTGTATACAGTTTTCCTTTACTC
SRSF1-NRS-T7 (EcoRI)	Mm_SRSF2	ACCACGCTATTCTCCCCGTCATAGCAGATCTCGCTCTCGTACAGAATCCCCTCCGCCCGTGTCGAAGCGAGAGTCCAAGTCTAGGTCGCGGTCCAAGAGCCCACCCAAGTCTCCAGAAGAAGAGGGAGCAGTTTCTTCCATGGCATCGATGACAGGTGGCCAACAGATGGGTTAAGATGATTGGTGACACTTTTTGTAGA
CTNNB1_Ultramer_200bp	Mm_CTNNB1	GCCATGGAGCCGGACAGAAAAGAAGACCCGCTGCTGTCAGCCACTGGCAGCAGCAGTCTTAC[TTGGATTCTGGAATCCATTCTGGTGCCACCACCACAGCTCCTTCCCTGAGTGGTAAAGGC]AATCCCGAA

		GAAGAAGATGTTGACACCTCCCAAGTCCTTTATGAATGGGCCGT CTTCAGCAAGGCTTTTCCCAGTCCT

(d) Nextera indexing primers and barcode sequences. In 'Index 1' and 'Index 2' reads, the adapter sequences are constant, whereas the barcode region (indicated with [i7] and [i5], respectively), are specific for each sample, thereby allowing for multiplexing of samples on the Illumina MiSeq platform.

Index 1 Read	CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG
Index 2 Read	AATGATACGGCGACCACCGAGATCTACAC[i5]TCGTCGGCAGCGTC
Forward indexing primer (i7)	
Name	**Sequence**
N701	TAAGGCGA
N702	CGTACTAG
N703	AGGCAGAA
N704	TCCTGAGC
N705	GGACTCCT
N706	TAGGCATG
N707	CTCTCTAC
N710	CGAGGCTG
N711	AAGAGGCA
N712	GTAGAGGA
N714	GCTCATGA
N715	ATCTCAGG
Reverse indexing primer (i5)	
Name	**Sequence**
S502	CTCTCTAT
S503	TATCCTCT
S205	GTAAGGAG
S206	ACTGCATA
S507	AAGGAGTA
S508	CTAAGCCT
S510	CGTCTAAT
S511	TCTCTCCG

3

Development and optimisation of a bioinformatic pipeline for the assessment of single nucleotide variances in CRISPR-Cas9 derived data

3.1 Introduction

3.1.1 Illumina MiSeq platform

Next-generation sequencing (NGS) technologies such as Illumina and IonTorrent have provided powerful tools to determine the sequence diversity of short DNA fragments. Due to cost effectiveness, fast-turnaround time and lowest error rate (Laehnemann, Borkhardt and McHardy, 2015), the Illumina MiSeq platform is considered one of the gold standards for the analysis of amplicon sequences. Illumina platforms rely on a chip-based bridge amplification procedure followed by sequencing-by-synthesis (SBS) using reversible terminator nucleotide dyes (Bentley *et al.*, 2008). Sequencing templates are fixed on a flow cell, after which solid-phase bridge amplification produces up to 1000 copies in a very close proximity (cluster generation) (see **Figure 3.1A**). Each cycle, the SBS technology adds a single fluorescently-labelled, reversible terminator-bound dNTP to the 3'-terminus of the sequence (see **Figure 3.1B**). The fluorophore is sequentially illuminated by a green laser for G and T and a red laser for C and A, after which imaging through different filters determines which nucleotide is incorporated. Removal of the fluorophore and 3'-terminator allow for the next cycle to commence. Successive cycles in parallel clusters can read millions of DNA fragments each up to 300-bp in length. As this technology will be used in the projects discussed in the subsequent chapters, the technical aspects and pitfalls of Illumina sequencing will here be discussed in more detail.

3.1.2 Illumina sequencing errors

Whereas Illumina sequencing allows for the analysis of large sets of data, sequencing errors resulting in incorrect output reads can occur. Overlapping emission spectra of the fluorophores and limitations of the detection filters used to distinguish signals cause G and T as well as C and A intensities to strongly correlate. As a result, the dominant sources of error on the Illumina platform are substitution-type miscalls (Schirmer *et al.*, 2015). In addition, lagging of molecules can occur by the incomplete removal of the 3'-terminator, or the skipping of nucleotides through the incorporation of dNTPs without an effective 3'-terminator. The number of affected sequences increases with each cycle, resulting in an exponential decline in confidence of base calling towards the end of a read (Cox, Peterson and Biggs, 2010). Several errors inherent to the Illumina platform cause specific patterns of errors in the output reads that are difficult to distinguish from true biological variation when

coverage is low. Conveniently, each base (*i*) in Illumina-derived reads is assigned an estimated error probability (p_i) represented as an ASCII-encoded Phred quality score calculated as $Q_i = -10 \log_{10} p_i$. This score thus gives an estimated probability of an error at that position, with a per base quality score (PBQS) of 10 indicating a probability of 1 in 10 in incorrect base calling (90% accuracy), PBQS of 20 indicating a probability of 1 in 100 (99%) etc., with a PBQS of 20 typically being used as a cut-off for low quality bases (Margulies *et al.*, 2005; Altshuler *et al.*, 2012).

Quality scores provide a key measurement for read errors as they are rarely associated with PBQS above 21 (Kozich *et al.*, 2013). Library preparation and sequencing primers are however thought to be the largest source of sequencing errors and as read quality scores do not reflect these errors, the confidence of bases is therefore thought to often be overestimated (Laehnemann, Borkhardt and McHardy, 2015; Schirmer *et al.*, 2015).

Illumina's paired-end (PE) sequencing mode can generate reads from both ends of the target DNA. If the DNA fragment is shorter than twice the read length, i.e. if there is overlap between the forward and reverse reads, these read pairs can be merged into a single larger fragment. In addition, bases that are overlapped by both forward and reverse read can furthermore be deployed to correct sequencing errors and yield higher quality sequence output. Therefore, designing amplicons to have overlapping PE reads for the bases key in an analysis greatly improves the confidence of calling these bases.

On the MiSeq platform, sequencing error rates increase substantially towards the distal end of a read and are significantly higher in the reverse read when doing PE sequencing; the overall substitution rate is known to be higher in reverse reads (1.07%) compared to forward reads (0.64%), whereas the insertion and deletion rates are approximately 4.2×10^{-2} % and 2.2×10^{-2} % (i.e. 100-fold less frequent than substitutions) for both reads (Schirmer *et al.*, 2015). The false discovery rate (FDR) of a single nucleotide substitution in a mate pair hence approximates 6.8×10^{-3}% per position. indicating the power of mate pair sequencing.

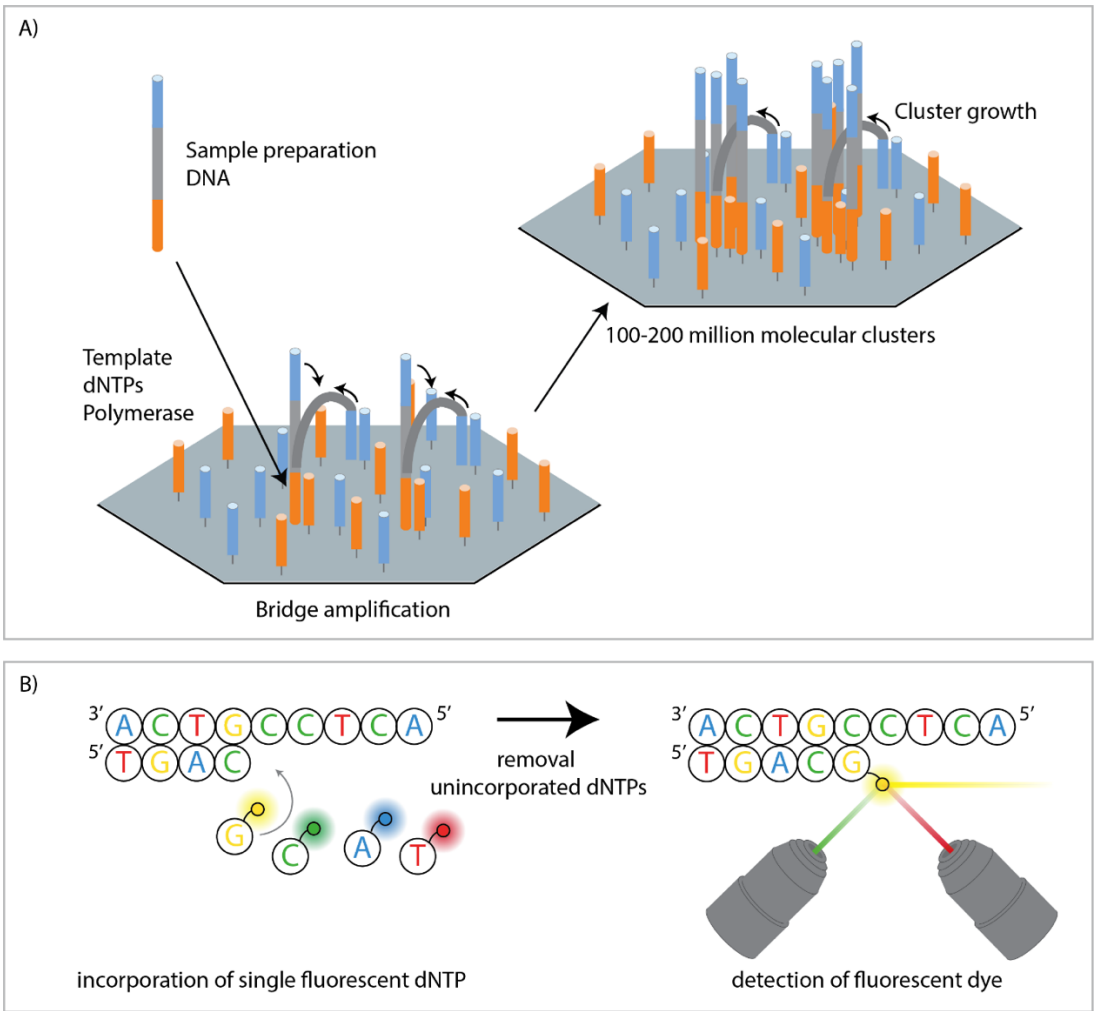


Figure 3.1 Schematic overview of sequencing-by-synthesis on the Illumina platform (A) Sequencing on the Illumina platform relies on immobilised probes on the flow-cell surface to which the adapter sequence (orange) of the single molecule template bind. Polymerases extend from the 3'-end of immobilised probes, after which the original template is removed by denaturing. The distal end of newly synthesised sequences (blue) binds to adjacent probes, thus performing bridge amplification. Clusters are formed through multiple cycles of annealing, extension and denaturation. After bridge amplification, the reverse strands are washed off, leaving only the forward strands, which can then be used the sequencing. **(B)** Each cycle, the complement to the single-stranded template molecule that is annealed to the cluster is extended by a single nucleotide with a fluorophore attached to its 3'-end. Next, the unincorporated dNTPs are washed off, after which sequential excitation with a green and red laser allows for the identification of the incorporated nucleotide (A, C, G or T). The fluorophore is then uncoupled from the 3'-end nucleotide, after which a new cycle can incorporate a new dNTP. After the desired read length is achieved, the same process is performed for the reverse of the reads.

Whilst slightly different, substitution and indel rates reported by another study are within the same range at respectively 0.25% and 9.1×10^{-2} % (Laehnemann, Borkhardt and McHardy, 2015). Errors are furthermore thought to not occur randomly but to rather be the result of

sequence motifs such as long homopolymer stretches, GC-rich regions and around inverted repeats, either due to altered enzyme preference or through backfolding during synthesis (Nakamura *et al.*, 2011; Quail *et al.*, 2012; Ross *et al.*, 2013). Because of high sequence similarity, amplicons are more susceptible to these motifs underlying the miscalling of substitutions, while in the calling of single nucleotide variants (SNVs) such systematic biases can result in high false positive rates. Hence, appreciating the various sources of sequencing errors are important when analysing sequencing data.

3.1.3 Motivation for study

The sheer volume and complexity of sequencing data generated by the MiSeq platform imposes a requirement for bioinformatic tools to facilitate downstream analyses. Sequences of poor quality can greatly impact the calling of sequencing variants and need to be identified and removed prior to analyses to minimise the FDR of nucleotide variants. Many hardware- and software-based correction methods have been developed to address errors in Illumina-derived sequencing data including quality trimming, error-correction and read overlapping. However, despite the availability of correction methods, most experimental biologists often lack a supportive bioinformatics infrastructure and often have to employ such software packages and pipelines, which results in computational inefficiency and inconsistencies between studies.

With the development of CRISPR-Cas9-based genome editing tools (F Ann Ran *et al.*, 2013; Jinek *et al.*, 2013), nucleotide diversification methods have become more readily available, while the demand for a convenient analysis pipeline for such amplicon data sets increases too. As discussed in the introduction, CRISPR-Cas9-induced breaks introduced at the target site will be repaired by either (error prone) end-joining or HDR (i.e. incorporating genetic information from an exogenous repair template) and are often used to yield specific outcomes. Several bioinformatics pipelines have been developed for the analysis of amplicon sequencing data from CRISPR (Güell, Yang and Church, 2014; Boel *et al.*, 2016; Pinello *et al.*, 2016; Park *et al.*, 2017; Wang *et al.*, 2017), but lack sufficient depth for the analysis and visual representation of the results. CRISPR-DAV is more sophisticated at visualisation (Wang *et al.*, 2017), but was not available at the initiation of this study and is not easily customised to analyse the DMS-based datasets which form the basis of the following chapters.

For the nucleotide diversification experiments on GFP (chapter 4), repair from an ssODN donor template through HDR will introduce a six-nucleotide invariant region and random nucleotide substitutions across 24 bases. A 157-bp region that will cover these sites will be amplified and will hence allow nearly full overlap of forward and reverse reads. The dataset is however expected to contain a variety of wildtype (including perfectly repaired alleles) sequences, sequences with inserts or deletions and reads containing the HDR core with additional mutations. As only HDR-derived reads are of interest, these reads need to be filtered based on the core sequence, whilst SNVs introduced by CRISPR-Cas9 and multiplex HDR need to be distinguished from sequencing errors.

In this chapter, I will therefore develop and optimise a bioinformatics pipeline to facilitate the processing of MiSeq-derived sequence libraries to detect and analyse pre-defined editing outcomes. These scripts will be tested on artificially-generated reads, representing a range of expected DNA repair outcomes, and will be provided in a step-by-step protocol for the utilisation of this bioinformatics package, which has been made publicly available. The package will subsequently be benchmarked on experimentally-derived data set from a genome editing project of *SRSF1* and demonstrate that it can be used to filter and analyse CRISPR-Cas9-induced mutations. This pipeline will lay the groundwork for analysis of empirical data in the subsequent chapters.

3.2 Results

3.2.1 Generation of artificial datasets

Optimisation of the data analysis pipeline required a data set containing sequences with pre-determined levels of variations and read quality. Artificial sequences were therefore generated based on the 157-nt reference sequence of the GFP amplicon used in chapter 4, sequenced with 150-bp PE on the MiSeq platform (see **Figure 3.2A**). Besides sequences with full consensus to the wildtype reference, additional sequences were generated containing nucleotide substitutions, deletions, insertions or ambiguous bases ranging from a single up to 15 mutations per sequence. As HDR-derived reads were expected to contain an invariable 6-nt sequence different from the wildtype sequence, sets of reads containing this 6-nt HDR core (respectively 'CGCGCG' for VarSil and 'TTATAA' for VarStop, further discussed in section 4.2.2) in the centre of the sequence were also generated, with additional nucleotide

substitutions mimicking the expected HDR-derived reads. From all these sequences, read pairs with variable PBQS for individual bases were then constructed and written to paired FASTQ files. The dataset furthermore contained a fixed number of reads with mismatched bases between the forward and reverse read to mimic sequencing errors.

The script 'GFP_dummygeneration.py' is publicly available on the GitHub repository 'deepmutationalscanning' (https://github.com/mkelder/deepmutationalscanning) and by adjusting the parameters at the start of the script, the pipeline can be used to generate sequences based on other FASTA sequences, sequences of different lengths or different numbers of mutations or output reads.

Thousands of unique reads were generated per mutation for either the VarSil or VarStop class, resulting in a total of 128,240 unique mutant reads in the final pool (see **Table 3-1A**). The sequence identifiers for each read contain information on the mutations it harbours with respect to the reference sequence, allowing for convenient analysis of reads dropping out at each stage of the pipeline (see **Table 3-1B**).

3.2.2 Trimming low quality bases using Trim Galore!

During preparation of Illumina sequencing libraries, short adapter sequences are typically attached to the sequence reads to allow for (1) binding to the Illumina flow cell using a universal adapter and (2) barcoding using sample-specific index adapters to allow for multiplexing of several libraries on a single flow cell (see **Figure 3.2A**). Whereas these adapter sequences are typically removed by the sequencing facility prior to data manipulation, this process may be incomplete and may obscure subsequent data analysis. Therefore, adapter trimming is often performed before further data-analysis.

Some tools such as FastQC (Andrews, 2010) allow for quality control with per-base and per-sequence quality profiling, while other methods for quality filtering such as Trim Galore! (Krueger, 2015) and Trimmomatic (Bolger, Lohse and Usadel, 2014) trim posterior bases with low quality scores and subsequently use *ad hoc* norms to filter reads with a minimum number of high-quality bases (Bokulich *et al.*, 2012). Average quality scores are calculated over a sliding window across the whole read during quality trimming, whereby the start of the read

is trimmed until the average quality score is above a given threshold and the end of the read is trimmed if the average quality score falls below the threshold.

Trim Galore! includes FastQC as well as adapter trimming from the 3'-end of reads using Cutadapt. By default, it uses the first 13-bp of Illumina standard adapters ('AGATCGGAAGAGC') for trimming, whereas other adapter sequences can manually be added. Trim Galore! is also able to remove low-quality bases, which (as mentioned in section 3.1.2) typically occur towards the end of a read.

Trim Galore! was used in its default settings for PE sequences, whereby bases are trimmed with a Phred score under 20 (p-value of FDR < 0.01) and auto-detects the Nextera adapter sequences used in this data set. By default, Trim Galore! is extremely stringent in adapter removal, removing sequences that have as few as 1-bp overlap with the adapter. As the used Nextera adapter sequences ('CTGTCTCTTATA') and the target amplicon sequence have a 3-nt homologous overlap, the stringency `-s` of overlap between adapter and read sequence was increased to 4 (see **Table 3-1**). As the pipeline will filter out short sequences further downstream, no size selection other than the default minimum fragment size of 20-bp was introduced. Final parameters: `trim_galore --stringency 4 --paired "$FILE1" "$FILE2"`

As is shown in Table 3.1c, different characteristics determine whether reads pass through the pipeline. Trim Galore! passes all reads regardless of the number of edits (within the assessed range of 1-15 edits). Once reads are aligned, reads with more than 6 missense bases (reflected by the "substitution", "silent" or "stop" categories), reads with more than 3 insertions or deletions (reflected by the "insertions" or "deletions" categories), reads with ambiguous bases (i.e. reads with "N" bases, reflected by the "ambiguous" category) or reads with non-consistent reads between the read pairs (considered read errors, reflected by "errors") are excluded from downstream analysis.

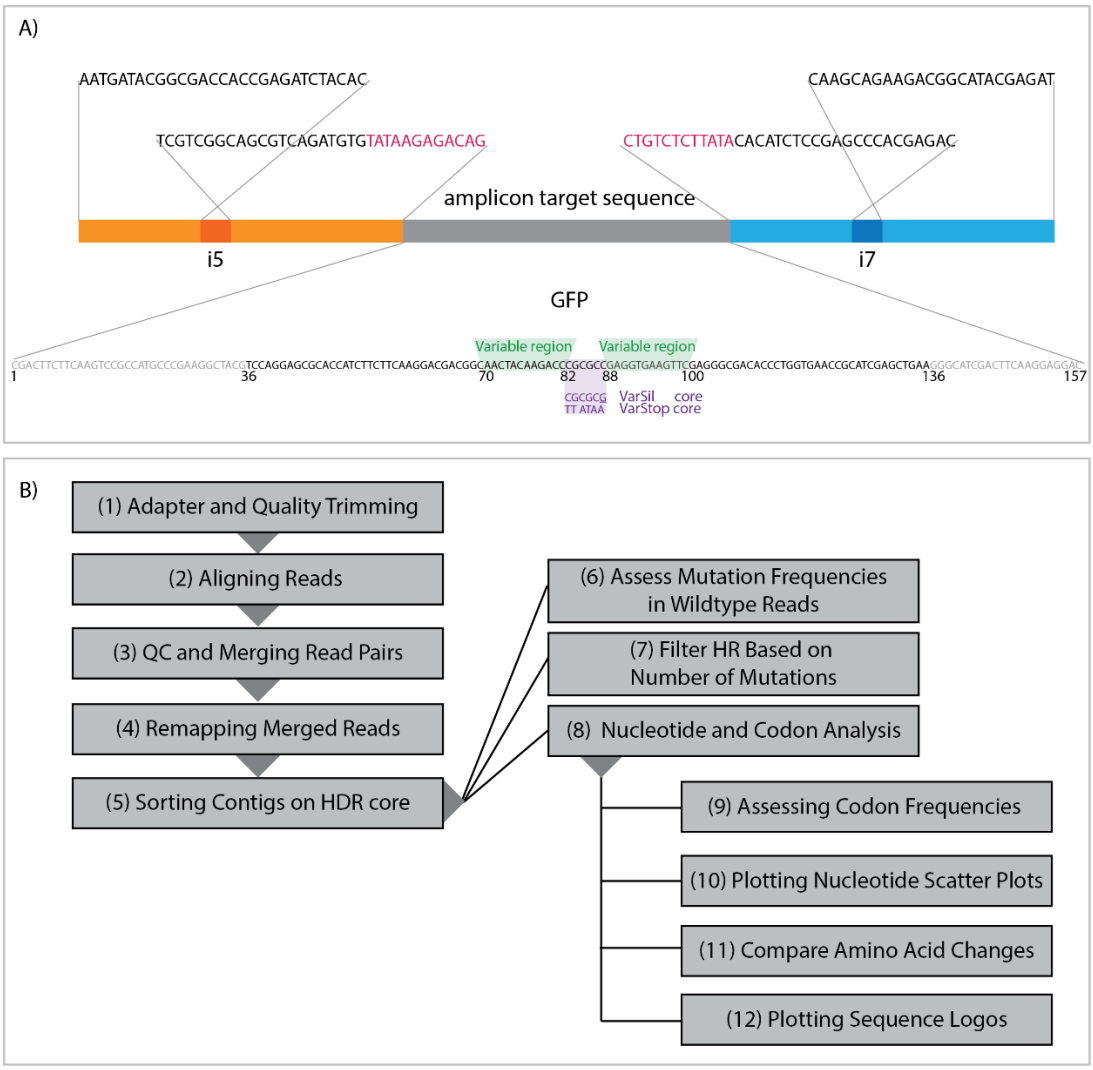


Figure 3.2 Overview of data analysis pipeline for the detection of HR-introduced mutations **(A)** Schematic representation of the GFP amplicon sequences tagged with Illumina adapters. In grey the amplicon-specific PCR sequence is depicted, whilst orange and blue represent the Illumina-specific sequences added to the distal ends of each amplicon. The sequences in light orange and light blue are constant Illumina-specific sequences, whilst the i5 and i7 index adapters (indicated in dark) are 8-nt sequences that are variable between libraries, thereby allowing for the multiplexing on a single sequencing lane. Sequences highlighted in pink are the Nextera-sequences recognised and trimmed by Trim Galore! Below, the 157-nt amplicon sequence for GFP is depicted (discussed in more detail in **Figure 4.3**), indicating the 100-nt region of interest in black font and the variable and core sequences highlighted in green and purple, respectively. The 6-nt VarSil or VarStop variants of the HDR core (purple text) differ respectively 1 and 6 nucleotides from the wildtype core and thereby allow for distinguishing the HDR-derived reads. **(B)** Overview of the steps involved in processing Illumina-derived sequencing reads. A full manual on the stepwise scrips is available in section 3.3.

Table 3-1 Overview of GFP dummy sequences and mapping parameters (A) Breakdown of read classes in the GFP dummy dataset A total of 152,360 reads were artificially generated to represent all possible editing outcomes in the deep mutational scanning project of GFP. Thousands of reads were generated for each type of mutation or read error. **(B)** Mapping efficiencies as a result of different BowTie2 parameters By default, 95,982 (74.8%) out of 152,360 of the dummy reads mapped. After optimisation, this number was increased to 104,493 (68.6%), mapping all reads with mismatches to the reference with 150M. For each parameter, the best setting (i.e. with the highest number of mapped reads are highlighted in green. Optimal parameters were determined to be `bowtie2 --threads 8 --phred33 --local -M 3 --rg-id 3 -x GFP_WT.fasta -1 dummy_1.fastq -2 dummy_2.fastq -S out.sam` **(C)** Number of reads passing each stage, broken down by mutation type. Wildtype reads and reads with substitutions (random or in the 6-nt core ('stop' and 'silent')) generally pass through all filtering stages, whereas reads with insertions, deletions, ambiguous bases or read errors are lost in the alignment or read-merging stages. Abbreviations: Sub = substitutions; Ins = insertions; Del = deletions, Amb = ambiguous, WT = wildtype (i.e. no mutations)

(A)

Total sequences	Substitutions	Insertions	Deletions	Stop	Silent	Ambiguous	Wildtype sequences	Read Errors
128,240	24,120	24,120	24,120	12,000	10,000	2,000	36,000	20,000

(B)

Default	Matching bonus (--ma)		Mapping penalty (--mp)		Ambiguous penalty (--np)		read/ref gap penalty (--rdg/rfg)		Seed extension attempts (-D)		Seed extension attempts (-R)		(-i x,1,0.1)		(-i S,x,0.1)		(-i S,1,x)		number of mismatches allowed (-N)	
95982	1	97015	2	100776	1	95982	5	99616	10	99616	1	99865	C	0	1	104043	0	0	0	104325
	2	99865	3	99807	2	95982	6	99779	11	99616	2	99865	L	95265	2	103751	0.1	104325	1	104493
	3	100849	4	98728	3	95982	7	99779	12	99616	3	99865	S	104043	3	103681	0.25	104167		
	4	101445	5	97833	4	95982	8	99779	13	99616	4	99865	G	104241	4	103499	0.5	104043		
	5	101450	6	96644	5	95982	9	99807	14	99616	5	99865			5	103381	1	103247		
	6	101491					10	99811	15	99616							2	101943		
	7	101546					11	99811									3	98555		
	8	101439					14	99811									4	98627		
	9	101438					17	99811									5	89095		
	10	101430					20	99811												
	20	96655					25	99811												
							30	99811												

(C)

				Read count								Percentage Total							
	Count	% Parent	% Total	Sub	Ins	Del	Stop	Silent	Amb	WT	Read Errors	Sub	Ins	Del	Stop	Silent	Amb	WT	Read Errors
Adapter and Quality Trimming	128,240	100.00	100.00	24,120	24,120	24,120	12,000	10,000	2,000	36,000	20,000	100	100	100	100	100	100	100	100
Aligning Reads	104,493	81.48	81.48	19,438	12,492	9,786	7,516	9,211	1,089	36,000	8,965	80.6	51.8	40.6	62.6	92.1	54.5	100	44.8
QC and Merging Read Pairs	72,165	69.06	56.27	19,438	-	-	7,516	9,211	-	36,000	-	100	-	-	100	100	-	100	-
Remapping Merged Reads	72,165	100.00	56.27	19,438	-	-	7,516	9,211	-	36,000	-	100	-	-	100	100	-	100	-

Amplicons were designed to only slightly exceed the length of a single 150-bp read, such that the overlap of both reads greatly reduces the base calling error rate. Genuine nucleotide variants in the dummy read set are expected to occur at rates higher than error rates. With quality trimming and quality control during the merging of reads already being implemented in the pipeline, several levels of noise reduction were already implemented. As error correction modules such as BayesHammer largely focus on indel errors and additionally often call low-frequency substitutions as false positives, these additional corrections were deemed to only interfere with the introduced nucleotide variances and were therefore not applied.

3.2.3 Optimising read alignments using BowTie2

After trimming of the raw reads, alignment to a reference genome is required for downstream variant calling and comparison between sequence individual reads and samples. Several alignment tools including Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2009) and BowTie2 (Langmead and Salzberg, 2012) are available for the mapping of reads in fastq format, the output format of most sequencers. Choosing an appropriate aligner can have drastic results on the accuracy of mapping. Spliced Transcripts Alignment to a Reference (STAR) (Dobin *et al.* Bioinformatics 2013) and other aligners are available but typically allow for alignment of reads other than amplicons, e.g. transcriptome data. As Bowtie2 allows end-to-end alignment that involves all read characters and is often used in the mapping of amplicon sequences, BowTie2 was used for the pipeline.

The BowTie2 parameters were optimised to correctly align the most possible reads in the pipeline. Bowtie2 outputs alignments in Sequence Alignment/Map (SAM) files, which, besides the sequence name, nucleotide and PBQS sequences, contain several fields describing information on the alignment. The FLAG is a bitwise number indicating whether a read and its mate mapped and in which order. Each reads also contains a CIGAR string that is used to indicate which bases align (either a match/mismatch) with the reference, are deleted from the reference, and are insertions that are not in the reference. As a readout for mapping efficiencies, the proportion of reads that were mapped with full consensus to the reference sequence were assessed (see **Table 3-1**), which is indicated in the CIGAR string by '`150M`'.

By default, Bowtie2 performs end-to-end read alignment, aiming to align every read character. Running Bowtie2 in --local mode allows for some characters to be soft clipped (i.e. ommitted) from the ends in order to achieve the greatest possible alignment score. It does hence not require the entire read to align from one end to the other. It was found that reads with single or multiple substitutions throughout the sequence were frequently soft clipped, thereby misaligning to the reference. It was hence decided that --local mode would not be optimal for the pipeline.

Bowtie2 runs multiple rounds of alignment and scores each alignment based on the similarity of the read sequence to the reference, from which the alignment with the highest alignment score is chosen. The score is determined by subtracting penalty scores for each difference (mismatch, gap, insertion), for which different parameters can be set to optimise alignments. All parameters were tested, which elucidated that a mapping penalty (`--mp`) of 4 and read and reference gap penalties (`--rdg` and `--rfg`) of 9 were yielding the highest number of reads mapped with full consensus (see **Table 3-1B**). In `--local` mode, it was found that a matching bonus (`--ma`) of 4 was optimal, whilst this is 0 by default in end-to-end mode with a low mapping score. Changing the ambiguous penalty (`--np`) or effort options (`-D` and `-R`) did not change how well sequences mapped. Configuring the minimum score using the function options had a big impact on the mapping efficiencies. Overall, constant and linear functions were giving very low mapping efficiencies, whilst natural log increase using `--i G,1,0.25` was shown to give to highest yield of mapping sequences.

Whilst the majority of the indels were correctly indicated by the CIGAR string with the optimised conditions, a small fraction (0.8%) of reads with indels at the periphery of the sequence were mapped with full consensus (`150M`). This may be due to the repetitive nature of the sequence, but since all of these mutations were outwith the central region covered by the repair template, within which mutations of interest reside, this would not pose an issue in downstream analysis. Final parameters that were used: '`bowtie2 --threads 8 --phred33 --no-mixed --mp 2 -i G,1,0.25 --rdg 9,3 --rfg 9,3 -N 1 --rg-id 3 -x [REFERENCE.FASTA]-1 [INPUT.MATE1.fastq] -2 [INPUT.MATE2.fastq] -S [OUTPUT.SAM]`'

The optimised BowTie2 parameters align Illumina-derived read pairs and allow up to 6 missense substitutions or 5 insertions or deletions to be mapped by BowTie2. For this reason, the majority of reads containing the 6-nt HDR core did not pass this stage. Therefore, reads were additionally mapped to the reference sequences containing the HDR core (respectively TTATAA and CGCGCG, discussed in sections 3.2.1 and 4.2.2) and finally combined the reads mapping to any of the three reference sequences into a single SAM file.

3.2.4 Concatenating reads into single amplicons

Reads processed by Bowtie2 are mapped in pairs, with each member of the pair expressed in separate lines of the output SAM file. PE sequencing greatly reduces the number of sequencing errors as each nucleotide covered by both mates is independently called twice as mentioned in the introduction. To further process amplicon sequences and correctly call single nucleotide mutations in downstream analysis, overlapping of forward and reverse reads into a single consensus contig was desired.

Several read merger scripts exist including FLASH (Magoč and Salzberg, 2011), PANDAseq (Masella *et al.*, 2012) and PEAR (Zhang *et al.*, 2014). Mergers such as FLASH and PEAR look at the maximum number of overlapping bases without taking the PBQS into account and hence do not perform any error correction. PANDAseq utilises PBQS but typically has issues with the merging of shorter fragments (i.e. amplicons) as it assumes all reads can be merged. None of these was therefore considered to be suitable for the merging of short fragments with high nucleotide variety, so a custom script was written to overlap sequences from both forward and reverse reads, utilising the base quality scores and CIGAR strings present in the SAM files generated by BowTie2.

At the start of the merger script, the read length (for the assays in this study 150-bp) and total amplicon sequence length covered by both the forward and reverse read (in case of GFP 157-bp) are determined. The script sets up empty read and Phred score sequence of the amplicon length and fills these in with ambiguous characters ('-' and '!', respectively). As we sought to minimize the FDR of SNVs, only read pairs in which both reads map to the reference (indicated by a FLAG tag of 99 and 147 for forward and reverse read, respectively) were

passed. Nucleotide sequences and Phred scores are read and copied into these new sequences, skipping positions in case the CIGAR string indicating a deletion while indicating an 'I' when there is an insertion, thereby retaining the amplicon length. When the forward and reverse read are scanned, the script compares both sequences in the overlapping region and writes high-quality consensus bases (i.e. Phred scores exceeding 20, corresponding with a p-value > 0.01) whilst assigning the highest of the two Phred scores. In case two conflicting nucleotides are called at high quality at the same position, the base with a Phred score above 20 or the one with the highest quality with a minimum difference of 5 in the Phred score, corresponding to half a log, is used. In case the two conflicting nucleotides are called at high confidence, it was considered an ambiguous nucleotide. As the deep mutational scanning projects require high confidence nucleotides in order to call SNVs, reads containing a deletion, insertion or ambiguous nucleotide in the region of interest (100-nt in case of GFP) are discarded at this stage. While stringent, these settings pass reads of the highest quality and minimise the FDR. These settings can however easily be adjusted at the start of the script.

Contigs written to the final output file are hence of very high quality and of full amplicon length, containing only mismatches. It was found that none of the reads containing deletions and insertions, indicated in the read names of the dummy sequences, have passed to this stage, thereby leaving only reads with wildtype sequences and substitutions. Contigs that contain conflicting or low-quality nucleotides are sorted into a separate output file.

For convenience, the output files containing the merged reads are remapped to the reference sequence to yield an output in SAM format. As at this stage, sequences do not contain insertions or deletions with respect to the reference, BowTie2 is here used at its default settings using `bowtie2 --local -x [REFERENCE.FASTA] –U [MERGED INPUT] –S [OUTPUT.SAM]`

3.2.5 Selecting and filtering reads with desired read outcomes

The remapped, high-quality contigs are at this point ready for analysis of nucleotide variances. As the sequences of interests are only those derived from HDR, those reads that contain the 6-nt HDR core that is unique for these sequences were filtered out. As the two experimental setups for GFP contained different cores for VarSil ('CGCGCG') and VarStop

('TTATAA') experiments, the script was set up in a dynamic fashion to bin reads based on having the expected HDR, wildtype ('CGCGCC') or other core sequence at the specified positions into separate files (discussed in section 3.2.1 and 4.2.2).

All dummy reads that were initially generated with HDR cores were correctly sorted into the HDR bin, confirming that the developed pipeline works efficiently at processing and isolating relevant reads. Whilst no additional sequences were found in this bin when selecting the STOP core sequence, a fraction of the reads containing substitutions in the wildtype read were found in the HDR bin. All these reads contained a C>G substitution converting a wildtype core sequence to a VarSil HDR core sequence. Whilst false positives through this specific mutation could theoretically also occur in experimentally-derived reads, it was reasoned that the developed bioinformatics pipeline cannot account for such a false positive. However, assuming that the proportion of HDR-reads will be sufficiently high in the experimental setup, such low frequency events were not expected to adversely affect downstream analysis. Indeed, missense mutations arising from a 1-nt substitution are very rare following NHEJ (van Overbeek *et al.*, 2016). Assessment of the proportion of VarSil cores in negative control samples should be performed to assess the FDR of this SNV in experimental samples.

3.2.6 Assessing frequencies of non-consensus nucleotides per repair template and in pool

After the isolation of contigs containing the 6-nt HDR core, these sequences can be used to count nucleotide frequencies at each position. A Python script was written to allow for a wide range of sequences to be processed and allow for the configuration of the reference and coding sequence, reading frame and region of interest at the start of the script. The script loops through all reads and counts nucleotide frequencies at each position, whilst also translating the nucleotide sequence to amino acids and highlighting any amino acid substitutions. It outputs both a tabular file of nucleotide frequencies at each position and a file listing each read containing a nucleotide substitution together with its amino acid substitution.

In addition, whereas nucleotide substitutions in an exonic region will often affect protein stability through a change in amino acid sequence, the script additionally translates SNVs to amino acid substitutions. It outputs a list with the SNVs and amino acid substitutions per read, allowing for the analysis of amino acid substitutions occurring in a sample.

Analysis of the nucleotide frequencies in the dummy sequences demonstrates that non-consensus nucleotides are equally occurring as is expected by the random nature of their origin (see **Figure 3.3A**). Processing a file containing dummy files with mostly nucleotide substitutions in two banks of 12-nt flanking the HDR core (i.e. the predetermined variable regions in the experimental setup of GFP) shows that these are also correctly processed and represented in the tables generated by pipeline (see **Figure 3.3B**). Some non-consensus nucleotides in the variable region were represented at a lower rate around 2%. To assess whether these mutations were misaligned, a grep search was performed to assess their frequencies. This assessment showed that these substitutions were occurring at a lower frequency in the generated dataset and that the read alignment is therefore not biased against these mutations. These findings indicate that the pipeline processes contigs with no bias towards specific nucleotide substitutions.

3.3 Step-by-step user guide for the use of the bioinformatics pipeline

A step-by-step user guide to the pipeline developed in this chapter (for schematic see **Figure 3.2B**) is provided in this section. In addition to the bioinformatic analyses described in the previous sections, additional (optional) scripts were provided to count and visualise nucleotide variants. Whilst these scripts have been optimised for the analysis of amplicon sequences derived from the GFP experiment, they can be adapted for the analysis of other amplicon sequences with predefined outcomes following the protocols laid out in this section.

3.3.1 Introduction to pipeline

The bioinformatics pipeline and supporting scripts described in this user guide are publicly available in the Github repository previously discussed (https://github.com/mkelder/deepmutationalscanning) (Dabbish *et al.*, 2012). The 'Pipeline.sh' Shell script contains all the modules necessary for full analysis of amplicon sequences and is optimised for the analysis of 157-nt amplicons of GFP. Each of the modules

functions independently and can as such be turned on or off at the start of the script (line 20-26), whilst the essential modules 1-5 are essential for subsequent analyses:

Essential modules:

Module 1: Adapter and quality trimming

Module 2: Aligning reads to reference

Module 3: Merging paired reads into a single contig

Module 4: Remapping reads into SAM files

Module 5: Sorting contigs based on core sequence

Optional modules:

Module 6: Determining mutation frequencies in wildtype reads

Module 7: Filter HR reads with single nucleotide change

Module 8: Assessment of nucleotide frequencies and mutations (recommended)

Module 9: Assessment of amino acid codon frequencies

Module 10: Plotting non-consensus nucleotide frequencies

Module 11: Compare amino acid changes between samples

Module 12: Plotting sequence logos

3.3.2 Pipeline configuration

Prior to the analyses, several reference files need to be in place in order for the scripts to function correctly:

- Ensure all provided scripts are placed into the /analysis directory, whilst 'Pipeline.sh' is positioned in the working directory.
- Place the FASTQ files compressed in the GNU zip (i.e. the standard output format of the MiSeq platform, indicated by the .gz file extension) in the working directory.
- Place the reference FASTA file, containing the sequence name in the first line in format ">SEQUENCE_NAME" and the reference sequence in the second line, in the working directory.
- Create a filenames_generic.txt file containing the desired sample names in the working directory.

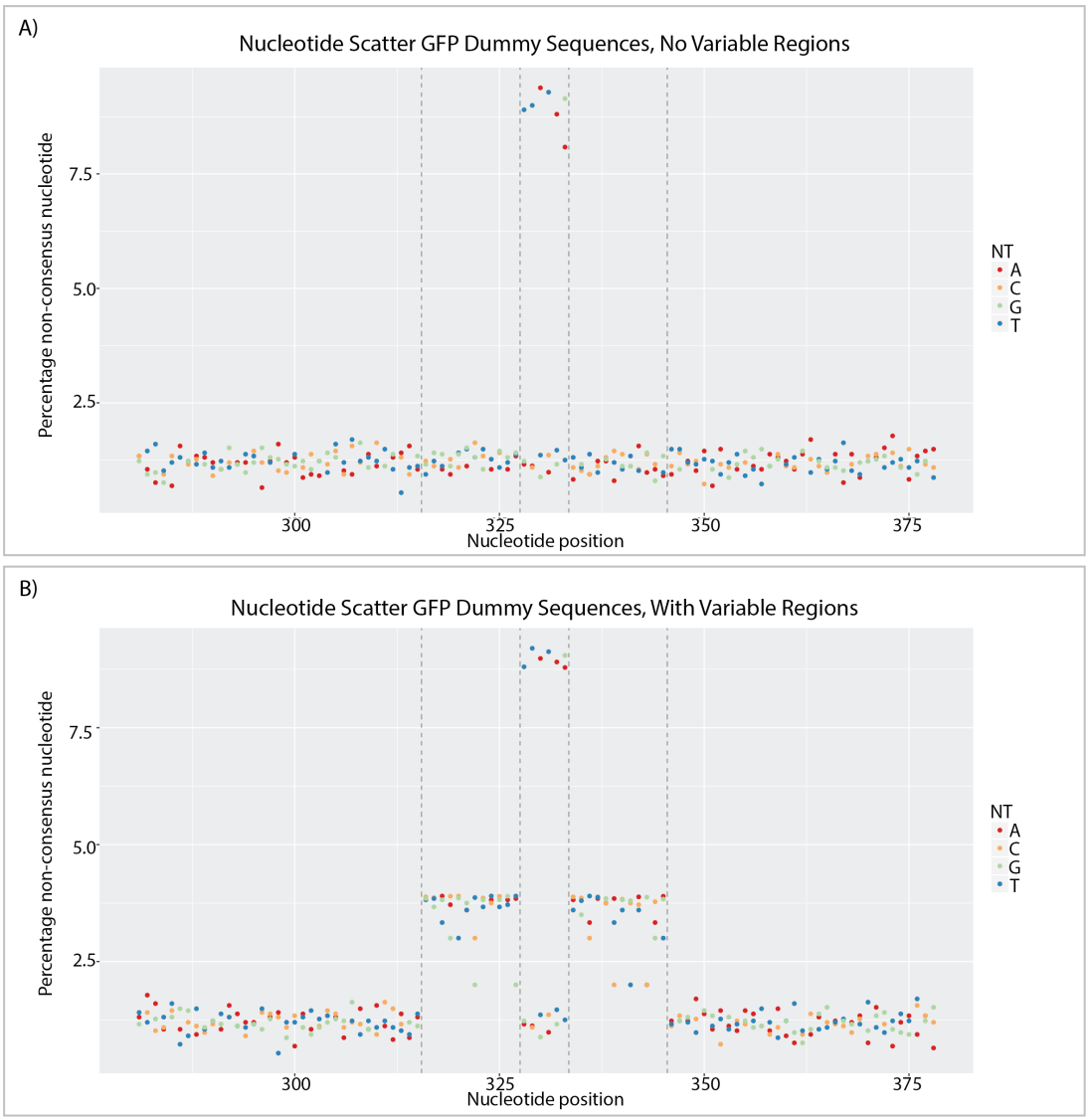


Figure 3.3 Dot plots of non-consensus nucleotides across nucleotide positions in dummy reads (A) Dot plot of non-consensus nucleotides at each position across the amplicon sequence. As nucleotide substitutions are generated randomly, there is no enrichment for specific nucleotide variants. In the centre, nucleotide variants representing the VarSil or VarStop core are enriched. **(B)** Dot plot of non-consensus nucleotides in dummy reads with non-consensus nucleotides enriched in the two banks of twelve nucleotides on either side. This shows that the developed pipeline is capable of detecting and visualising specific nucleotide variants. Non-consensus nucleotides occurring at reduced frequency in the variable region were underrepresented in the initial dataset (see previous page). Nucleotide position is with respect to the GFP ORF.

Grid engine options can be indicated at the start of the script and are by default at 4 Gb in 8 cores, with an expected run time less than 10 hours. Indicate the working directory in 'DIR' (line 13), reference file name in 'FASTA' (line 14) and toggle SILENT/STOP in 'HRCORE' (line 15) of the file. Finally, indicate which of the scripts should be turned on/off in the analysis (line 20-26).

For GFP, these configurations are sufficient to execute the full script from the command line by executing '`sh Pipeline.sh`' or by submitting the script to the cluster through '`qsub Pipeline.sh`'. Alternatively, modules of the script can be copied to the command line for manual execution of the scripts. Throughout the pipeline, intermediate files will be placed in the respective folders, whereas output files are stored in the analysis folder. For the analysis of amplicons other than GFP, individual scripts require manual configuration. Each of the scripts is therefore explained below with instructions for configuration.

3.3.3 Module 1: Adapter and quality trimming

This module will run Trim Galore! (a wrapper for Cutadapt and FastQC) on the list of files in 'filenames.txt' in the working directory, which will automatically be generated based on the unzipped '.fastq' files in the directory. In PE mode, the first two files in 'filenames.txt' are paired, then the next two etc.. The Trim Galore! command does not require configuration for the analysis of other amplicons but can be adjusted for its stringency or adapter sequence as explained in the Trim Galore! manual (Krueger, 2015). After trimming, summary reports will be written to .txt files in the /analysis directory.

3.3.4 Module 2: Aligning reads to reference sequence

This module will run the BowTie2 aligner on the list of files in 'filenames_trimmed.txt' in the target directory using the FASTA file set at the start of the script, which will automatically be generated based on the '.fq' files in the directory. Read pairs will be aligned to the FASTA reference sequence input at the start of the 'Pipeline.sh' script and outputs aligned reads in a SAM file. BowTie2 parameters were extensively optimised in section 0 and are suitable for a wide range of amplicon sequences but can be adjusted in the BowTie2 command line as can be read in the scripts manual (Langmead and Salzberg, 2012). Alignment summaries will be printed on the command line.

3.3.5 Module 3: Merging reads into single contig

This module merges forward and reverse reads into a single contig of a predefined amplicon length. When the forward read and reverse read are scanned, the script compares both sequences in the overlapping region and writes high-quality consensus nucleotides (i.e. Phred scores higher than 20, corresponding with a p-value for FDR < 0.01) whilst assigning the highest of the two Phred scores. For further details, see section 3.2.4.

Adjust parameters in 'M3.Merge_Reads_To_Contig.py' to configure this merger script for amplicons other than GFP. Set read length by parameter ‘ReadLen’, amplicon length by ‘AmpLen’ and toggle stringency ‘Stringent’ for the inclusion of indels and ambiguous nucleotide.

3.3.6 Module 4: Remapping reads into SAM files

Module 4 remaps the merged contigs to the reference sequence provided in the FASTA file. This module utilises the default parameters for BowTie2 and does not require any adjustments.

3.3.7 Module 5: Sorting contigs based on core sequence

This module sorts sequences based on their core sequence into three classes: NOHR (i.e. wildtype), HR and AMB (ambiguous or other). The HR core sequence will be based on the HDR core provided at the start of the ‘Pipeline.sh’ script. With this script, the proportions of predefined outcomes are hence measured and will be printed on the command line.

Reference sequence will be based on the predefined FASTA file, whilst core sequences (‘HRcore’ and ‘NOHRcore’) and positions (‘coreStart’ and ‘coreEnd’, with respect to the start of the amplicon sequence) can be adjusted at the start of the script ‘M5.Sort_SAM.py‘.

3.3.8 Module 6: Check mutation frequencies in WT reads

This script is optional and can be toggled on/off at the start of ‘Pipeline.sh’. It assesses the proportion of wildtype reads with full consensus to the FASTA reference sequence, thereby providing a gage for the number of reads unaffected by CRISPR-Cas9 and the quantification

of read errors. Similar to module 5, HDR core positions ('coreStart' and 'coreEnd') can be adjusted at the start of the 'M6.Filter_WT_mutations.py' script.

3.3.9 Module 7: Filter HR reads with single nucleotide change

This script is optional and can be toggled on/off at the start of 'Pipeline.sh'. Reads identified as having an HDR core in module 5 are analysed and further divided based on having either (1) no mutations, (2) a single mutation in the variable region and no additional mutations in the peripheral regions, (3) mutations in the peripheral regions but not in the variable regions, (4) multiple mutations in the variable region or (5) a single mutation in the variable region with additional mutations in the peripheral regions. Absolute read counts for each class are printed in the console, whilst reads from class 2 are written to '.HRunique.HR' files and reads from class 4 to '.moremuts.HR' files. Similar to module 5, HDR core positions ('coreStart' and 'coreEnd') can be adjusted at the start of the 'M6.Filter_WT_mutations.py'.

3.3.10 Module 8: Counting nucleotide frequencies and mutations

This script can be toggled on/off at the start of 'Pipeline.sh' and is optional, although highly recommended, as it scans through the HR reads and counts the number of non-consensus nucleotides at each position. It outputs a tabular file ('.NTcount') containing a summary of the absolute counts and proportions of each nucleotide at each position. This includes an overview of the reading frame and reference numbers of nucleotides and amino acids to the larger genomic open reading frame (ORF).

In addition, nucleotide substitutions are translated into amino acid substitutions. A list of reads with their respective nucleotide and amino acid substitutions (both with respect to the wildtype and HDR core sequence) are listed in a separate file ('.mutations').

This script requires several parameters to be configured. 'ReadingFrame' is the nucleotide frame (1, 2 or 3) use to translate the nucleotide sequence into amino acids, based on the amplicon reference sequence. 'outputstart' is the nucleotide position in the amplicon reference sequence from which the output should start reporting (inclusive), set at 0 to inactivate and thus output from the start of the amplicon reference sequence. 'outputstop' is the nucleotide in the amplicon reference sequence at which the output should stop

reporting (inclusive), set at 0 to inactivate and thus output to the end of the amplicon reference sequence. 'seqstart' is the start nucleotide of the amplicon reference sequence relative to a larger reference sequence (e.g. the GFP ORF) for reporting nucleotide and amino acid mutations.

3.3.11 Module 9: Assessing amino acid codon frequencies

This script is optional and can be toggled on/off at the start of 'Pipeline.sh'. It scans through all HDR reads and, based on the provided reading frame, makes counts for all possible codons at each position and writes it into a tabular summary file ('.codoncount'). Parameters are the same as those used in Module 8.

3.3.12 Module 10: Plotting non-consensus nucleotide frequencies

This script is optional and can be toggled on/off at the start of 'Pipeline.sh'. This module plots non-consensus nucleotide frequencies for all HR reads. It creates non-consensus nucleotide dot plots (as seen in **Figure 3.3**) and stacked bar plots based on the nucleotide frequencies in the '.NTcount' file generated in module 8. These graphic representations are output as PDF-files in the analysis directory.

Parameters are optimised for GFP amplicons but can be adjusted for other amplicon sequences. This however requires the adjustment of values in the section starting at line 41 for the boundary of nucleotides to report on. In addition, parameters in code lines 53 onwards can be adjusted for the tuning of nucleotide positions for the limits of the variable region indicated by the dashed lines in the plot.

3.3.13 Module 11: Compare amino acid changes between samples

This script is optional and can be toggled on/off at the start of 'Pipeline.sh'. It analyses the '.mutation' files generated in module 8 and outputs a tabular summary of the frequencies of amino acid substitutions, whilst additionally looking up the corresponding $\Delta\Delta$G value for the effect of these amino acid substitutions on GFP (discussed in more detail in section 4.2.14). These summaries are created for both the residues in the variable positions specifically ('.mutfreq.var.CSV') and for all amino acid positions ('.mutfreq.all.CSV') in the analysis directory.

3.3.14 Module 12: Plotting sequence logos

This script is optional and can be toggled on/off at the start of 'Pipeline.sh'. It translates all nucleotide sequences in the HR file into an amino acid multiple sequence alignment ('.AAs') in the AA directory and subsequently creates sequence logos in PDF-format the analysis directory. Parameters are the same as those used in Module 8.

3.4 Use Case: Detection of HDR-derived reads in SRSF1 targeting

As an initial assessment of how the established bioinformatics scripts perform on real data, albeit not from a deep mutational scanning experiment, the pipeline was utilised to assess HDR efficiencies in the repair of point mutations in *SRSF1* by CRISPR-Cas9 and HDR. The experimental design was conducted by Fiona Haward and Andrew Wood, and cell culture was entirely performed by Fiona Haward. SRSF1 is part of a serine/arginine-rich protein family of pre-mRNA splicing factors that is capable of nucleocytoplasmic signalling. The aim of this work is to study the physiological relevance of the cytoplasmic functions of SRSF1, by knocking in a nuclear retention signal (NRS) to create a SRSF1-NRS fusion protein that prevents the nucleocytoplasmic shuttling of SRSF1. In addition, a small T7 epitope tag was added, allowing for several biochemical analyses on the SRSF1 protein.

After several attempts to isolate homozygous clones through CRISPR-Cas9 (see **Figure 3.4A**), only a single viable clone with homozygous knock-in (clone 11) was obtained in mESCs. This clone however contained two single base deletions on one allele, resulting in frameshifts that inactivate the NRS (see **Figure 3.4B**). These results suggest that there is a strong selection against the obstruction of SRSF1 shuttling in mESCs. To test this, a CRISPR-Cas9-based assay was designed to assess changes in the frequencies of HDR-derived alleles during the retargeting of SRSF1-NRS-T7 clone 11. This assay was based on the premise that, if homozygous SRSF1-NRS-T7 alleles had a negative impact upon cell fitness, the frequency of HDR-derived alleles should decrease in the population over time.

Repair of the monoallelic single base pair deletions in the C-terminus and NRS of this clonal cell line was attempted by CRISPR-Cas9-mediated HDR. Fortunately, the 5' deletion created a new Cas9 PAM site, this was used to design a sgRNA to selectively target allele 1, thereby

leaving the other allelic copy with the correct insert unaffected. Cells were transfected with wtCas9, sgRNA and a modified ssODN with a silent nucleotide substitution destroying a BamHI restriction site, which was utilised to distinguish both allelic copies after successful repair. Transfected cells were edited for 24 hours and subsequently enriched by FACS for GFP fluorescence. Cells were harvested for gDNA extraction before targeting (T_0) and at multiple time points post-transfection (T_1-T_7, timepoint being 2 days apart) to assess whether there was a temporal reduction in presence of the desired knock-in see **Figure 3.4C**). The target site was amplified by PCR and subsequently deep-sequenced on the Illumina MiSeq platform.

In order to account for any proximal sequence modifications (especially in the terminal exon of SRSF1, i.e. the 5' portion of the insert) that could produce a compensatory mutation, the maximum PE read length of 300-bp was used to amplify across the entire C-terminal region, resulting in an amplicon of either 434-bp (unedited allele 1) or 436-bp (allele 2 or repaired allele 1) product. The bioinformatics pipeline developed in this chapter will be utilised to assess the frequency of successful HDR outcomes. This will furthermore assess the wider applicability of the pipeline.

3.4.1 Experimentally-generated reads are efficiently mapped to the reference

As the maximum difference between the expected repair outcomes (i.e. allele 1 unrepaired, allele 1 repaired and allele 2) is two single base pair deletions and a single base substitution, reads were mapped to a single reference sequence of repaired allele 1. For the proportions of allele 1 (either unedited or repaired), these were either represented as part of the entire sequencing pool or as proportion of all allele 1 reads by normalising to the number of allele 2 reads, as allele 2 was unaffected by CRISPR-Cas9 editing. Across all samples, on average 73.88% ± 0.07 of the reads mapped (see **Table 3-2**), with the highest proportion (85.0%) found in the T_0 (i.e. unedited) sample as expected. An interrogation of unmapped reads showed that these predominantly comprise short (47 – 62 bp, similar to CCTGTAGATGTTAGGCGCCAGGGGATATTCCCCACGCCTGCAATTCAACACTTTAGCCCCTT) reads that have little (10%) sequence similarity with the reference or the Nextera adapters, whilst a BLAST search for the sequences did indicate a 86.7% similarity to the *Srsf1* gene, with 8 A>C substitutions with respect to the reference (Altschul *et al.*, 1997). As this number of missense mutations is higher than the threshold (6 missense bases), this is in line with the stringency of the BowTie2 script in the pipeline. This indicates that the script optimised for the mapping

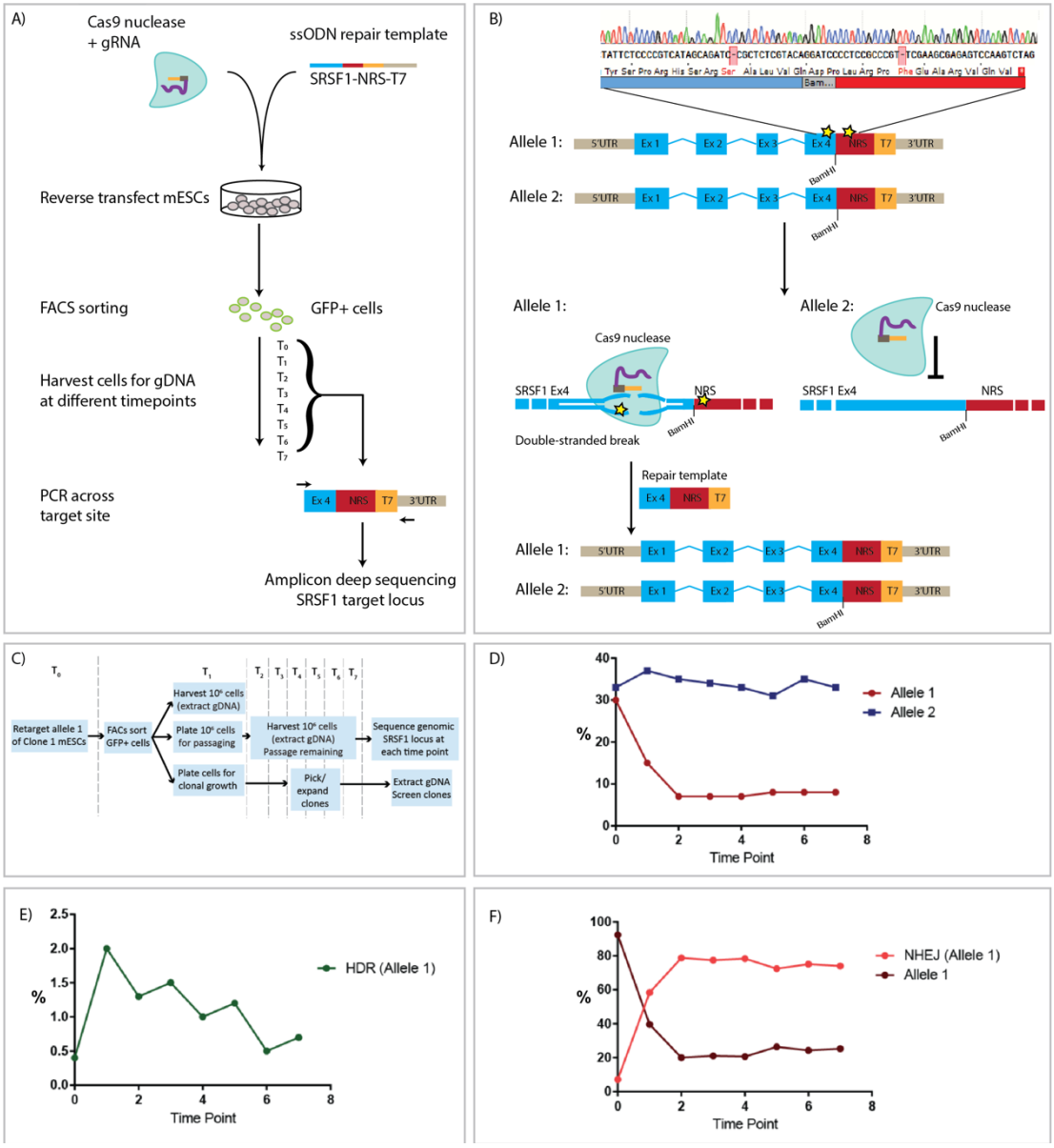


Figure 3.4 Schematic of SRSF1 editing strategy and sequencing outcomes (A) For the tagging of SRSF1 with an NRS and T7 sequence, mESCs were transfected with a Cas9-GFP plasmid, sgRNA and an ssODN containing the tag sequence and homology arms to the SRSF1 site. GFP-positive cells were enriched by flow cytometry and harvested at different time points after transfection and sequenced across the target site. **(B)** A single clone (Clone 11) was found to harbour a homozygous knock-in of the SRSF1-NRS-T7 construct but with two single base pair deletions on allele 1. This allele was specifically targeted with Cas9 and a ssODN containing the construct but with a silent mutation destroying a BamHI site, thereby allow to distinguish repaired allele 1 from allele 2. **(C)** Schematic overview of time points at which cells were harvested, which represent 2-day intervals. This schematic was adopted from the Ph.D. thesis of Fiona Haward. **(D)** Frequency of unedited allele 1 and 2 over time post-transfection. Whereas the frequency of allele 2 remains constant over time, the frequency of unedited

allele 1 sharply drops in the first two days after targeting, after which the frequency remains constant. This indicates that allele 1 but not allele 2 is targeted by CRISPR-Cas9. **(E)** Frequency of perfectly repaired HDR copies of allele 1. After editing, a sharp increase in the frequency of repaired copies of allele 1 is observed, indicating repair by HDR. Although some fluctuations are observed, this frequency gradually decreases in the time after editing, indicating a negative selection against perfectly repaired SRSF1-NRS-T7. **(F)** Frequency of copies of allele 1 containing full consensus and indels. Whereas there is a sharp drop-off in frequency of copies with full consensus, an inverted correlation is observed in the frequency of reads that have undergone NHEJ. This indicates that the majority of the copies of allele 1 undergo NHEJ.

of GFP reads can also be used for mapping different amplicons of a different length.

3.4.2 Perfect repair of SRSF1 is selected against in mESCs

For the merging of forward and reverse reads, parameters were optimised as described in the user guide to construct a 436-bp amplicon from 300-bp reads whilst allowing possible insertions/deletions to be included. Reads with bases that were not consistently called between both reads of a pair (i.e. read errors) were discarded. On average, 69.4 ± 3.04 % pairs merged with full consensus between forward and reverse reads.

Module 5 of the pipeline was configured to isolate reads based on the expected editing outcomes, i.e. allele 1 unrepaired, allele 1 repaired, allele 1 with indels, unaffected allele 2 and anything else. Editing efficiencies were determined by assessing the frequency of unrepaired allele 1 and allele 2, which was close to 1:1 at T_0 (see **Figure 3.4D**). The proportion of allele 2 remained broadly constant over the time, confirming that that allelic copy was not targeted by Cas9. In contrast, the proportion of unrepaired allele 1 dropped drastically from 40% to 15% in the first two days post-targeting, indicating a targeting efficiency approximating 62.5% (i.e. 100% - 15% / 40%). These results show that Cas9 specifically targeted a single allelic copy of *SRSF1*, which can be detected by the developed pipeline.

As mentioned in the introduction, cleaved DNA can be repaired by either NHEJ, leading to small insertions or deletions at the target site, or by HDR from the introduced template donor, which should result in the desired outcome with repair of both of the point mutations and the destruction of the BamHI restriction site through a silent G>A mutation. Assessment of the frequency of HDR shows that HDR reads increased to 2% in the first day after transfection (T_1) (see **Figure 3.4D** and **Table 3-3**), which corresponds with 6.7% when normalised to the number of wildtype allele 1 reads, demonstrating that the mutations have

Table 3-2 Mapping efficiencies for SRSF1 samples

Sample	Total reads	Aligned reads	Aligned %
T_0 (unedited)	928,541	789,301	85.00%
T_1	1,089,775	679,475	62.35%
T_2	1,134,650	769,863	67.85%
T_3	818,138	600,841	73.44%
T_4	415,908	302,323	72.69%
T_5	640,925	480,758	75.01%
T_6	919,511	640,531	69.66%
T_7	295,590	251,311	85.02%

been successfully repaired by HDR. Notably, the rate of HDR-reads decreases gradually over the first six days after transfection to frequencies similar those seen prior to transfection (i.e. T_0). This demonstrates a strong negative selection against a homozygous SRSF1-NRS-T7 cells.

Reads that matched neither the unrepaired or the 31-nt fully repaired length representing perfectly repaired allele 1 were considered as NHEJ reads (see **Figure 3.4E** and **Table 3-3**). These reads inversely increased with the decrease in unrepaired copies of allele 1 and plateaued after T_2, demonstrating that the majority of reads targeted by CRISPR-Cas9 are repaired by NHEJ. However, as a considerable number of reads fell into this class in the unedited sample, a closer examination revealed that these reads harboured additional nucleotide substitutions in the target region whilst the two restored nucleotide insertions, indicating that the initial pool was not entirely monoclonal or that errors originating from library preparation were at considerable levels. Regardless, the increase in NHEJ reads in the transfected pool indicates that these scripts can be used to detect rates of editing in these samples.

The analyses of the SRSF1 dataset shows that the developed pipeline can be utilised to analyse experimentally generated datasets with amplicon and read lengths different from those that the pipeline was initially designed for. By merely adjusting the parameters of the pipeline, the proportions of predefined CRISPR-Cas9 editing outcomes can be measured in different projects.

Table 3-3 Read proportions of SRSF1 targeted editing of allele 1 Samples were harvested before editing (T_0) and at 7 different time points after editing (T_1-T_7). Read proportions were determined by filtering reads for the occurrence of two base pair deletions (allele 1, unedited), repaired deletions but with a G>A substitution (allele 1, repaired by HDR), no deletions and no G>A substitution (allele 2) or other, which was classed as NHEJ. Proportions of reads were represented as either part of the read 1 pool (i.e. normalised to the unaffected allele 2) or as part of the total sequencing pool. After targeted, the proportion of both perfectly repaired allele 1 (HDR) and NHEJ increased largely, indicating editing. The proportion of perfectly repaired allele 1 decreases over time after transfection, indicating selection against this allele.

Sample	Allele 1 (unedited)	% Normalised to allele 2	% Total	Allele 1 (repaired G>A and deletions)	% Normalised to allele 2	% Total	NHEJ (i.e. not WT or correct HDR)	% Normalised to allele 2	% Total	Allele 2 (no dels and G)	% Total	All reads
T_0 (unedited)	142,351	85.4%	30%	702	0.3%	0.14%	65,842	6.03%	13.32%	166,688	33%	375,583
T_1	63,836	39.6%	15%	3,143	2.0%	0.72%	206,832	47.56%	58.41%	161,053	37%	434,864
T_2	34,489	20.0%	7%	2,202	1.3%	0.45%	283,350	57.51%	78.75%	172,670	35%	492,711
T_3	27,383	21.0%	7%	1,936	1.5%	0.50%	225,019	58.52%	77.48%	130,200	34%	384,538
T_4	12,971	20.6%	7%	628	1.0%	0.32%	116,962	60.45%	78.39%	62,926	33%	193,487
T_5	24,971	26.4%	8%	1,125	1.2%	0.37%	186,965	60.77%	72.42%	94,624	31%	307,685
T_6	34,701	24.3%	8%	755	0.5%	0.18%	231,772	56.54%	75.16%	142,712	35%	409,940
T_7	20,985	25.3%	8%	558	0.7%	0.22%	146,923	58.46%	74.00%	82,845	33%	251,311

3.5 Discussion

In this chapter, I developed and optimised a bioinformatics pipeline to analyse amplicon sequencing libraries with pre-defined outcomes generated on the Illumina MiSeq platform. These scripts were optimised on artificially generated sequences resembling all expected read outcomes and read errors based on the GFP experiment conducted in the following chapter, with the aim to distinguish SNVs introduced by CRISPR-Cas9 and multiplex HDR from sequencing errors. The analysis of data derived from an experiment aiming to repair monoallelic single nucleotide deletions in an SRSF1-NRS-T7 construct shows that this pipeline can conveniently be adapted to quantify pre-determined read outcomes in MiSeq-derived reads. Together, this data analysis pipeline provides a standardised and solid foundation for the analyses of multiple datasets in the subsequent chapters.

3.5.1 Read errors need to be prevented through adequate experimental design

Before being able to evaluate nucleotide variants in the dataset, several steps were required to ensure the exclusion of low-quality reads and to minimise the FDR by addressing the sources of error inherent to sample preparation and sequencing on the Illumina platform. One publication suggests that PCR, rather than sequencing, is the largest source of substitutions, insertions and deletions (Schirmer *et al.*, 2015). The Illumina MiSeq platform itself results in several kinds of errors, with substitutions occurring more often than insertions or deletions. Using high-fidelity polymerases and choosing the right primer sequences to avoid homopolymer stretches, GC-rich regions and inverted repeats are therefore essential steps prior to analysis by sequencing. Targeted amplicon sequencing as used in this project makes these sequences most sensitive for sequencing biases, whilst it simultaneously limits the design of primers for the exclusion of such nucleotide profiles. PCR-derived biases can hence not be accounted for through the bioinformatics pipeline and may result in false positives. It is therefore essential to include negative controls to identify possible biases inherent to analysis of the amplicon sequence and determine the minimal threshold to which to normalise SNV frequencies. Several steps that address these issues are included, with a focus on reducing the FDR of SNVs, which will be further addressed in the following paragraphs.

3.5.2 Pipeline is designed at analysing reads with less than 6 mismatches

Alignment is an essential step in the correct calling of substitutions. BowTie2 parameters were optimised to map the maximum number of GFP reads to the reference sequence (section 0), with a focus on the mapping of mismatched bases and reads containing the HDR core. The final parameters allowed for the alignment of up to six nucleotide variants, which posed a problem in HDR-derived reads from the VarStop experiment; as these inherently contain a 6-nt core different from the wildtype sequence and additional mutations therefore often did not map to the reference. Therefore, sequences from respective experiments can additionally be mapped to the reference sequences containing the VarSil or VarStop HR core to account for these discrepancies. As the pipeline was aimed to analyse single nucleotide and single codon variants, the exclusion of reads with more than 6 mismatches to the reference sequence would not pose any issues in downstream analyses.

The analysis of the SRSF1 data demonstrated that the BowTie2 parameters optimised for GFP also allow for the mapping of longer fragment including insertions and deletions (section 3.4). On average 26.12% of the SRSF1 reads did not successfully map. A further analysis revealed that these reads were shorter sequences with little overlap to the reference, which were presumably PCR-derived contaminations. As the developed pipeline mapped other sequences from the dataset correctly to the reference, this demonstrates that the used parameters align experimental data with an equal efficiency to the artificial GFP data.

3.5.3 Read pairs can be merged into a single contig at high quality

PE sequencing greatly reduces the FDR by analysing each nucleotide twice. The read merger developed in this chapter allows for the creation of a single contig whilst taking the PBQS into account. It only merges and calls a base when called at high confidence by at least one of the two reads. For the calling of SNVs in amplicon sequences used in the following chapters, reads with single nucleotide variants called at high confidence were required. Therefore, the most stringent parameters in the merger script only pass reads with every base called at high confidence by both reads in a pair, excluding reads with deletions or insertions. These settings hence greatly reduce the FDR of an SNV and only pass contigs with a very low likelihood of reads with nucleotides wrongly called due to sequencing errors. PCR-derived errors can however still be called at high confidence by both reads and can still pass through these filters. The merger file outputs a list of high-quality sequences with either full

consensus or with (a maximum of six) mismatches with respect to either the wildtype or HDR reference. As these reads are aligned and of equal length, they allow for convenient comparison and alignment.

Using the artificial GFP dataset, it was shown that the stringent filters of the merger script did not pass reads with insertions, deletions or low-quality scores. In the SRSF1 data, unedited copies of allele 1 contained indels with respect to the reference sequence and would hence be excluded when using the highest stringency settings (section 3.4). By lowering these stringencies, these reads were shown to be processed and resulted in an average of 69.4% of the read pairs to be merged into contigs. This shows that the settings of the scripts developed in this chapter can be adjusted to select for different sequencing outcomes.

3.5.4 Filtering of sequences for quantification of read outcomes

Subsequent quantification and binning of predefined sequences (e.g. those containing the HDR tag) in the contig pool allow for the convenient quantification of editing outcomes and in case of SRSF1 showed to be sufficient to finish the sequencing analysis (section 3.4). As reads are sorted into separate folders, reads of interest are isolated for further analysis.

It was demonstrated that in the quantification of HDR-derived mutations differing from the wildtype by a single C>G substitution, read errors can lead to the incorrect identification of HDR-derived reads, which could subsequently lead to the incorrect calling of a nucleotide variant as HDR-derived. However, as substitutions due to sequencing errors occur at a rate of 0.0107 and 0.0064 in respectively the forward and reverse read and thus at a rate of 6.8×10^{-5} in a read pair, the chance of a read containing the C>G mutation or additional substitutions remains low and will be expected to be outnumbered by true HDR reads that will occur at a substantially higher rate. These false positives can be addressed by only considering SNVs occurring above this 6.8×10^{-5} noise level threshold in the HDR read population.

The first five modules of the pipeline are sufficient for the processing and quantification of pre-determined sequence variants such as allelic variants and repair outcomes exemplified in the SRSF1 case study. For further analysis and visualisation of sequence variants additional

scripts are written and optimised for GFP. As stated in the user manual, most of these scripts are interactive and can easily be adapted to other amplicon sequences, which will be demonstrated in the following chapters.

Together, the results from this chapter demonstrate that the developed scripts allow for the analysis of predefined sequencing outcomes from amplicon sequencing. With the majority of the scripts allowing for convenient adaptation of the scripts to other sequencing formats, amplicons from different experiments can be analysed by the use of this pipeline. It provides a clear guide and is convenient to use whilst requiring little computational power. In the following chapters this pipeline will be utilised for the analysis of sequences derived from deep mutational scanning projects on GFP and β-catenin.

4

Deep mutational scanning of nucleotide substitutions in GFP

4.1 Introduction

In the previous chapter I developed a data analysis pipeline that can filter and assess the effects of nucleotide and amino acid variants introduced by CRISPR-Cas9 and multiplex HDR. The second aim of my thesis is to show that this pipeline can be applied to interrogate the effect of single nucleotide variants introduced into a chromosomally-encoded locus. In this chapter, GFP will be used as a proof-of-principle and show that a library of all possible single nucleotide substitutions can be created in a predefined window of a chromosomally encoded gene and assess their functional impact.

Green fluorescent protein (GFP) is the first naturally fluorescent protein that has been identified after it was isolated from the jellyfish *Aequorea Victoria* (Shimomura, Johnson and Saiga, 1962; Prasher *et al.*, 1992). Wildtype GFP is a 27kDa protein encoded by 240 codons and has an intrinsic fluorescence with excitation and emission maxima of respectively 395nm and 509nmm. This fluorescence is emitted from an α-helix chromophore that is surrounded by a β-barrel structure consisting of 11 β-strands (see **Figure 4.1**) (Tsien, 1998). Because the inward-facing side chains of amino acids in the β-barrel strongly interact with the chromophore, correct folding of these domains is essential for both spectroscopic and structural stability (Stepanenko *et al.*, 2013).

Since its discovery, GFP has been optimised through successive rounds of directed evolution both for use in eukaryotic systems and to enhance its fluorescence, resulting in a variant known as enhanced GFP (eGFP) (Cormack, Valdivia and Falkow, 1996; Yang, Cheng and Kain, 1996). This variant of GFP fluoresces 35 times brighter than the wildtype form and with its codons optimised for its translation in mammalian cells (Cormack, Valdivia and Falkow, 1996), eGFP rapidly turned into the mostly used variant for use in mammalian cells.

As GFP allows for a convenient distinction between folded (functional) and misfolded (dysfunctional) variants by its fluorescence and the structure of the protein is resolved up to a high resolution, this protein was considered to be an ideal first target to assess the deep mutational scanning assay. In addition, the crystal structure of GFP is resolved up to a resolution of 1.9 Å (Yang, Cheng and Kain, 1996) allowing for the *in silico* modelling of the effects of amino acid substitutions on protein stability, thereby providing means to validate

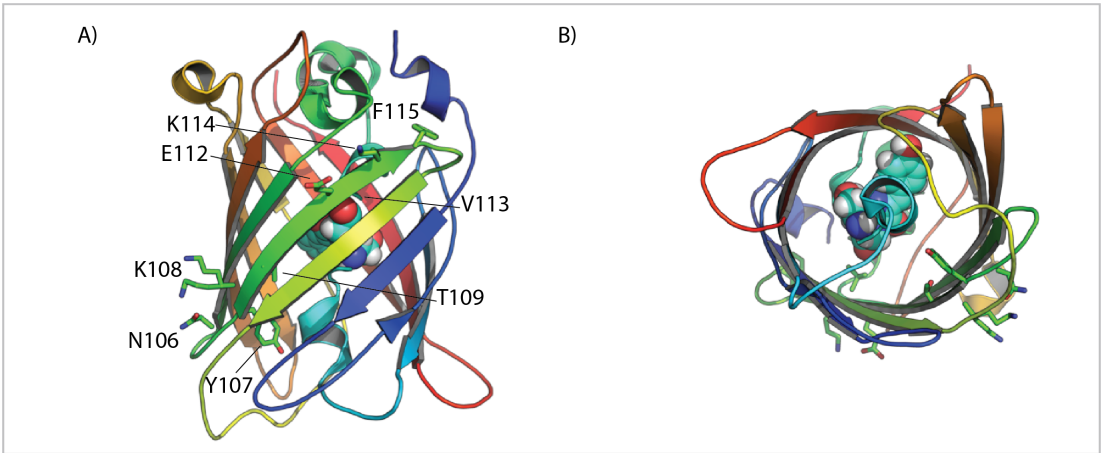


Figure 4.1 Structure of the GFP protein (A) Three-dimensional representation of the GFP protein. The protein consists of 11 β-sheets that together form a β-barrel in which the α-helix chromophore resides. Interactions with inward-facing amino acids in the β-sheets are essential for fluorescence. Side chains are shown for amino acid residues ^{106}NYKT--EVKF^{115}, which are the residues targeted in this chapter. Amino acid T109 and V113 are inward facing and potentially interact with functioning of the fluorophore. **(B)** Top view of the GFP protein showing the fluorophore present in the β-barrel and the side chain of V113 in close proximity to the fluorophore.

our experimental findings. The aim of this chapter is to interrogate the relationship between nucleotide variants and protein stability, so mutations were introduced into sequences coding for part of the β-barrel, which is the key structural scaffold, rather than the fluorophore itself (see **Figure 4.1**). Nucleotide variants will therefore be introduced into the codons for amino acids ^{106}NYKT--EVKF^{115}, which form the central part of one of the β-sheets.

In this chapter, an experimental work-flow will be setup and tested to introduce single nucleotide variants (SNVs) into a chromosomally encoded gene and apply the bioinformatic analysis pipeline developed and tested in the previous chapter to assign a functional score to mutations introduced by HR. This functional score will be benchmarked to both *in silico* predictions of protein stability as well as previously published experimental data (Sarkisyan *et al.*, 2016).

4.2 Results

4.2.1 Experimental workflow for the introduction and functional assessment of single nucleotide variances in GFP

The workflow used for the experiments in this chapter are schematically represented in ***Figure 4.2***. The RCNβH-B(t)-GFP (RCN(t)-GFP) cell line was previously characterised as a single copy random integration of a *pCAG-Gfp* transgene with uniformly high GFP expression (Chambers *et al.*, 2007) (see ***Figure 4.2*****A**). RCN(t)-GFP are transfected for 24 hours with two plasmids encoding Cas9 nickase (pX335-U6-Chimeric_BB-CBh-hSpCas9n(D10A)) (Cong *et al.*, 2013) and different sgRNAs targeting the GFP locus, a plasmid coding for a puromycin resistance cassette (pSUPER.retro.puro) and an ssODN pool. A majority of ssODNs in this pool harbour one or more nucleotide changes within a central region, such that Cas9-mediated breaks in different cells can yield HDR edits with distinct sequences. In the VarStop experiment (left panel), which serves as a control, all cells that undergo HDR will incorporate a stop codon, resulting in a loss of GFP fluorescence irrespective of the genotype within the adjacent variable regions.

In parallel, cells transfected with the VarSil ssODN pool will, in contrary to the VarStop experiment, incorporate a silent mutation instead of a stop codon. This should allow the effects of any single nucleotide changes within the variable regions to be unmasked, which is expected to result in a range of GFP expression. Cells will be subjected to puromycin selection for 24 hours, which will selectively kill cells that have not been transfected. 7 days post-transfection, the GFP populations will be sorted for genomic DNA isolation, after which PCR amplification across the target site and subsequent library preparation and indexing will allow for deep sequencing on the MiSeq platform. Each of the steps of the process will be further discussed in the following sections.

4.2.2 Design of repair template oligonucleotides to assess incorporation efficiencies and effects of nucleotide variants

As reports have shown that HDR efficiencies are increased with the use of ssODNs over double-stranded donors (Chen *et al.*, 2011; F Ann Ran *et al.*, 2013; Renaud *et al.*, 2016) this system was chosen as an initial system for the delivery of our templates. For homology-directed repair, two different ssODN repair template classes 100-nt in length were designed

with homology to the GFP locus, while having two variable regions of each 12-nt centred around a 6-nt invariable core sequence (see
Figure *4.3*A).

In the 24 variable nucleotide positions, non-consensus nucleotides (with consensus being the nucleotide found in the genomic sequence of *Gfp*) were introduced at controlled frequencies by 'doping' during the synthesis of the ssODNs (see
Figure *4.3*B). During synthesis of an exact ssODN sequence, a desired 2'-deoxynucleoside was added at each cycle of synthesis, resulting in the exact nucleotide being incorporated at that location in each ssODN molecule. However, different 2'-deoxynucleosides could be mixed at specific, pre-defined ratios in certain cycles, such that these positions contain different bases in different molecules synthesised in parallel. In contrast to degenerate oligonucleotides used in other studies (Findlay *et al.*, 2014b; Ma *et al.*, 2017; Mason *et al.*, 2018), whereby a completely random nucleotide is incorporated at certain positions, we will utilise 'doped' oligonucleotides, whereby the sequence is biased towards the genomic consensus nucleotide with only a small level of randomization. Mixing of the four 2'-deoxynucleosides at equal ratios would for example result in each nucleotide being incorporated in 25% of the synthesised molecules. However, by manipulating the ratio of consensus versus non-consensus nucleotides at each of the 24 variable positions in our ssODN, the level of diversity per position and per repair template could be controlled.

For this study, the genomic consensus nucleotides for the variable positions were added at either 97% or 94% with respectively 1% or 2% for each of the remaining non-consensus bases (in different ssODN pools that are henceforth referred to as 97:1:1:1 and 94:2:2:2), such that the highest possible proportion of ssODN molecules with a single non-consensus base in any of the 24 variable positions (8 codons) could be achieved (see
Figure *4.3*C). As the 'doping' process was random, a substantial proportion of molecules was expected to contain either no or multiple mutations as is shown in the graph. However, this way of synthesis allowed for the introduction of any of the 72 nucleotide variants into the genomic locus of GFP through HDR.

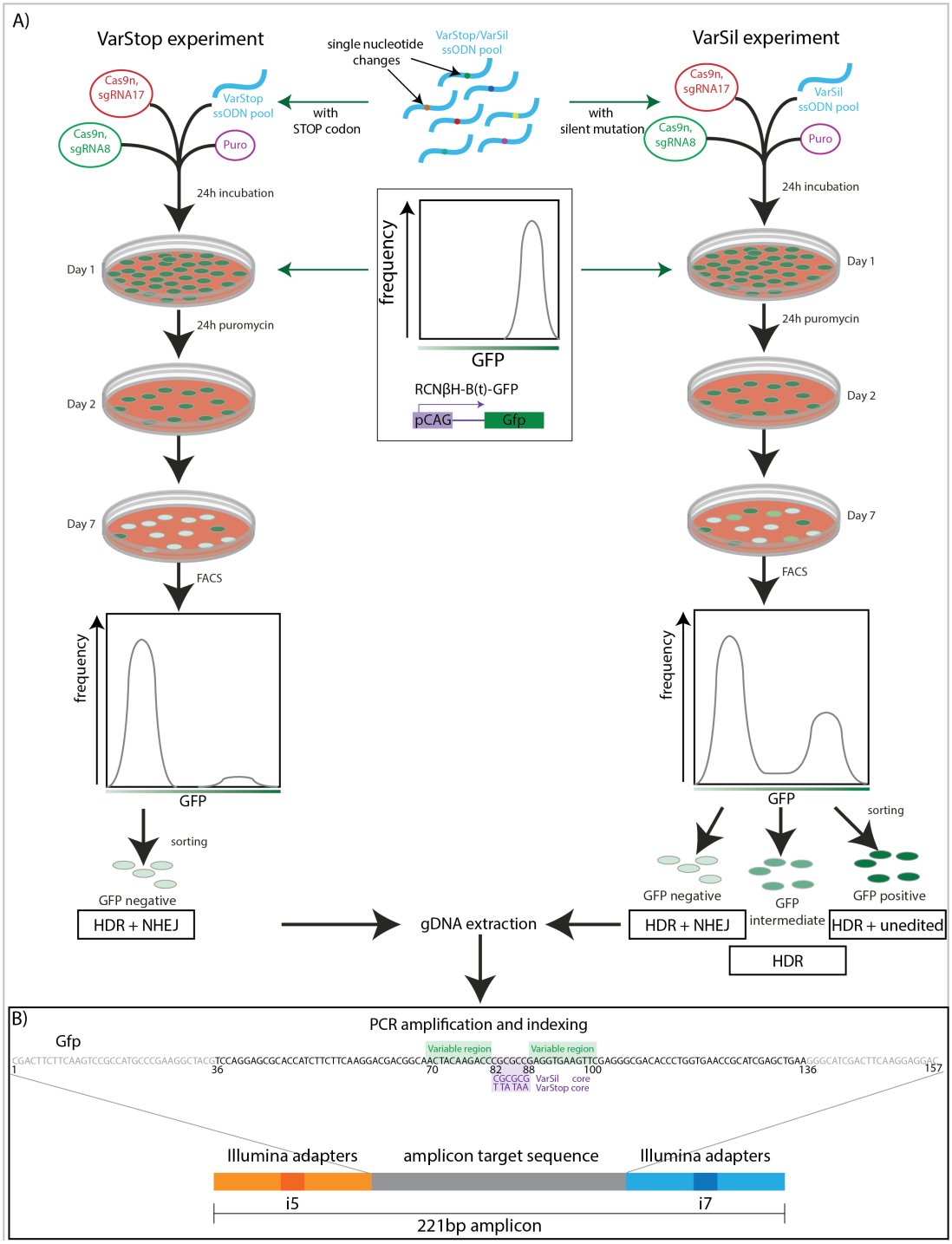


Figure 4.2 Experimental workflow for the introduction and functional assessment of single nucleotide variants in Gfp **(A)** For the introduction of amino acid substitutions, RCN(t)-GFP ES cells with a uniform high GFP fluorescence are transfected with plasmids encoding Cas9n and two sgRNAs (sgRNA8 and sgRNA17, see figure 4.6), a plasmid encoding a puromycin resistance gene and an 100nt ssODN repair template pool harbouring different single nucleotide substitutions per template (see figure 4.3). Two experiments (left and right hand of the figure) differ in the repair template used, which either contains an in-frame stop codon (VarStop) or a silent mutation (VarSil) at its centre. 24 hours after transfection, cells are exposed to puromycin for 24 hours, which kills cells that have not been transfected.

Afterwards, cells are cultured until 7-days post-transfection, at which cells are analysed by flow cytometry. In the VarStop experiment, all cells that have undergone incorrect repair through NHEJ and have thus incorporated indels will lose GFP fluorescence, whereas the same will occur for cells that have incorporated the VarStop ssODN through NHEJ. Meanwhile, in the VarSil experiment, cells that have undergone NHEJ will lose GFP expression, whilst cells that have undergone HDR have incorporated a silent mutation (not affecting GFP expression) and an additional mutation that might (partially) destabilise GFP expression, thus leading to a wider range of GFP expression. The increase in GFP intermediate in VarSil over VarStop is used as a measure for HDR efficiencies. Cells are sorted into either a single negative bin (VarStop) or into three bins (VarSil) after which gDNA is extracted. **(B)** PCR amplification of the target site results in a 157-bp amplicon that spans the ssODN homology region, including the variable regions and HDR-core (further explained in figure 4.3). Amplicons are extended with universal Illumina adapters that include i5 and i7 index sequences that allow for the multiplexing of different libraries on the same sequencing lane.

The 24 variable positions in the repair template covered 8 codons of the Gfp gene and could thus theoretically result in a total of (8 * 19) 152 amino acid substitutions. However, as the doping process makes the chance of two consecutive nucleotide substitutions very low, this approach would most likely be screening the effect of single nucleotide variants and thus limit the number of amino acid substitutions that can be interrogated. Whereas many of the introduced mutations were expected to result in the loss of a functional protein and hence lead to a complete loss in GFP fluorescence (termed null or amorphic mutations), some mutations might not affect GFP fluorescence compared to the wildtype fluorescence (silent mutations) (Muller, 1932). In addition, some nucleotide variants could result in a partial loss of gene function and would hence be expressing an intermediate level of fluorescence. By analysing the genomic variants present in the intermediate domain of GFP fluorescence, we hence expected to observe these so-termed hypomorphic mutations (Muller, 1932).

Two different repair template classes were designed, with each containing its own invariable 6 nucleotide core ('HDR core'), which differs from the genome and serves as a genetic marker to identify HDR-derived reads in the sequencing pool. Two variable regions, each comprising 12 nucleotides, are positioned either side of the HDR core (see **Figure 4.2B**). In the first repair template class, 'VarStop', this HDR core consist of an in-frame stop codon that result in premature translational termination (TAA), causing complete loss of the GFP protein – and hence fluorescence – in every cell that perfectly repaired Cas9-induced lesions through HDR. After editing, the pool of GFP negative cells would therefore contain both NHEJ and all HDR-derived alleles, regardless of the genotype within the variable positions. As the possible

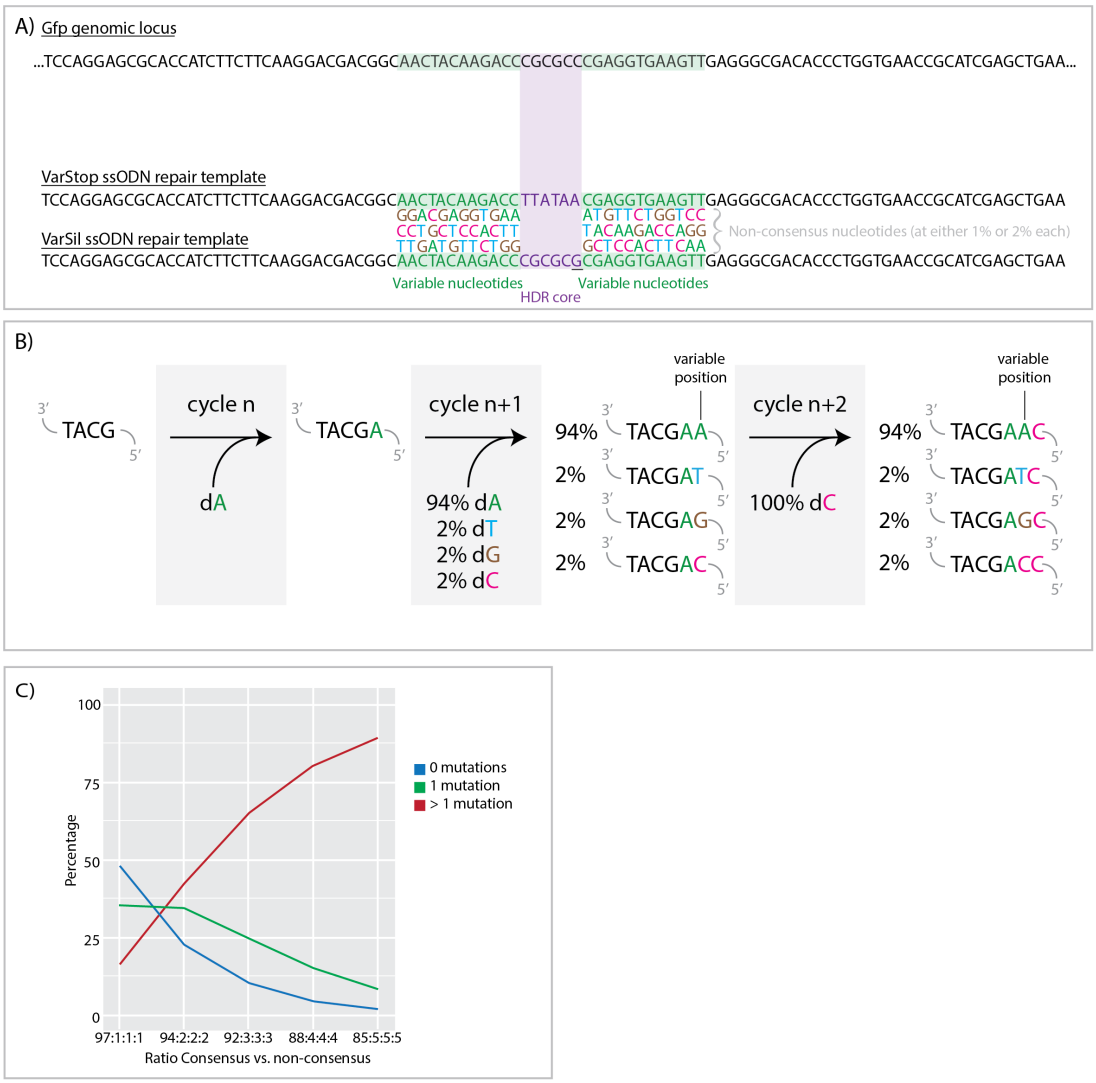


Figure 4.3 ssODN design and the doping by synthesis principle **(A)** Two classes of repair template ssODN molecules are designed, VarStop and VarSil, both with 100-nt in length and homology to the *Gfp* genomic locus with exception to the 6-nt HDR-core. This HDR-core sequence contains either an in-frame missense and stop codon (TTATAA, in VarStop) or a single, silent C>G substitution (CGCGCG, in VarSil). In addition, in the 12-nt flanking either side of the HDR-core (i.e. 24 nt in total), consensus nucleotides (with respect to the GFP locus) are randomly substituted with non-consensus nucleotides at either 1% or 2% for each non-consensus nucleotide during synthesis. **(B)** During ssODN synthesis, each cycle a specific 2'-deoxynucleoside (dN) is added and incorporated from 3' to 5' (exemplified by dA) in all molecules of the ssODN pool. However, the different 2'-deoxynucleosides can be mixed at pre-defined ratios in certain cycles, such that different nucleotides are incorporated at these ratios at the next residue. Mixing of the 2'-deoxynucleosides at 94% for the genomic consensus nucleotide and 2% for each of the other nucleotides, as exemplified in the figure, 6% of the ssODN molecules harbour a non-consensus nucleotide at this position. For this study, 24 positions on the ssODN were 'doped' in this way. **(C)** Theoretical plot showing the percentage of oligonucleotides containing 0, 1 or >1 nucleotide substitution with 24 variable ("doped") nucleotide positions. This shows that both 97:1:1:1 and 94:2:2:2 ratios of

consensus versus non-consensus bases give the highest probability of a single nucleotide substitution occurring in a template.

effects of the single nucleotide variants are masked, the VarStop sample thus allowed for assessment of HDR frequencies (by determining the number of reads with the TTATAA core) irrespective of phenotypic selection.

The second class of repair templates (VarSil) contain a 6-nt HDR core which differs from wildtype GFP by a single silent C>G substitution (CGCGCG, 'VarSil'), thereby still allowing for the identification of HDR-derived reads, whilst unmasking the phenotypic effects of nucleotide variants introduced at the variable positions. Whereas it would not be possible to use FACS to separate out cells harbouring a null mutation from cells with frameshift mutations and similarly silent mutations from cells without a mutation, hypomorphic mutations were expected to lead to a distinct level of intermediate GFP fluorescence not observed in the VarStop sample. By comparing the GFP profiles of cells transfected with VarSil to those transfected with VarStop, the HDR efficiency could hence be approximated by the increase in cells with intermediate GFP fluorescence.

4.2.3 Assessing nucleotide diversity on repair template oligonucleotides

As mentioned in the previous section, both 97:1:1:1 and 94:2:2:2 repair template ssODNs were ordered with either a VarStop or VarSil HDR core. The nucleotide diversity in these four repair template pools was empirically assessed by deep-sequencing of repair templates, which were PCR-amplified with adapters allowing for Illumina indexing (see **Figure 4.4A,B**). ssODN templates (containing either the VarStop or VarSil core sequence) were converted into double-stranded DNA and PCR-amplified with adapter-linked primers in a single step, such that the entire ssODN was amplified and primer sequences did not cover the variable positions (for a detailed description see Materials & Methods). The amplicon was subsequently indexed and sequenced on the Illumina platform as discussed in the previous chapter.

Focussing solely on reads harbouring a single nucleotide substitution, each possible non-consensus nucleotide was shown to be represented in each of the ssODN pools (see **Figure 4.4C**). The average frequency of individual non-consensus nucleotides in 97:1:1:1 and

94:2:2:2 were respectively 1.04 ± 0.055% and 2.24 ± 0.11 % as opposed to 0.22% outwith the variable positions (see **Figure 4.4A** and **Figure 4.5**), thus approximating overall substitution rates of approximately 3% and 6% per position at the variable sites (see **Figure 4.5**). Positional biases and consistent differences favouring the incorporation of thymidine over other bases were hereby observed. Correspondence with the manufacturer of these ssODN libraries (IDT) confirmed a slight skew towards the incorporation of guanine and thymidine during the synthesis process. As all single nucleotide substitutions were represented in the ssODN, this pool however demonstrated to be useful for the nucleotide diversification of our genomic locus of interest.

Assessment of the number of non-consensus nucleotides per template additionally confirmed that both repair template classes harbour the same percentage (35%) of reads with a single nucleotide variant (see **Figure 4.4B**), indicating that both repair templates were equally suitable for introducing single nucleotide variants across 24 nucleotide residues. The 94:2:2:2 repair template pools were used in subsequent experiments.

4.2.4 Design of sgRNAs for the editing of GFP using CRISPR-Cas9

As discussed in the introduction, individual nicks in the genome are frequently repaired with high fidelity to ensure genome integrity, whereas simultaneous nicking via appropriately offset guide RNAs is believed to yield lesions which are processed as double-stranded breaks whilst extend the number of specifically recognized bases for target cleavage by the created overhang (F. Ann Ran *et al.*, 2013).

Whilst these paired nickase-variants of Cas9 (Cas9^{D10A} as discussed in section 1.5.3, hereafter referred to as Cas9n) were thought to increase HDR-efficiencies when using ssODNs to insert small fragments of up to 9 nt (F. Ann Ran *et al.*, 2013; Christopher D. Richardson *et al.*, 2016), we wanted to assess the effects of Cas9n and wildtype Cas9 (wtCas9) on introducing nucleotide diversity into a genomic locus. A side by side comparison of Cas9n and wtCas9 was therefore performed.

In order to test editing efficiencies of wtCas9, multiple sgRNAs were used that induced cleavage at different distances with respect to the ssODN center to determine the optimal

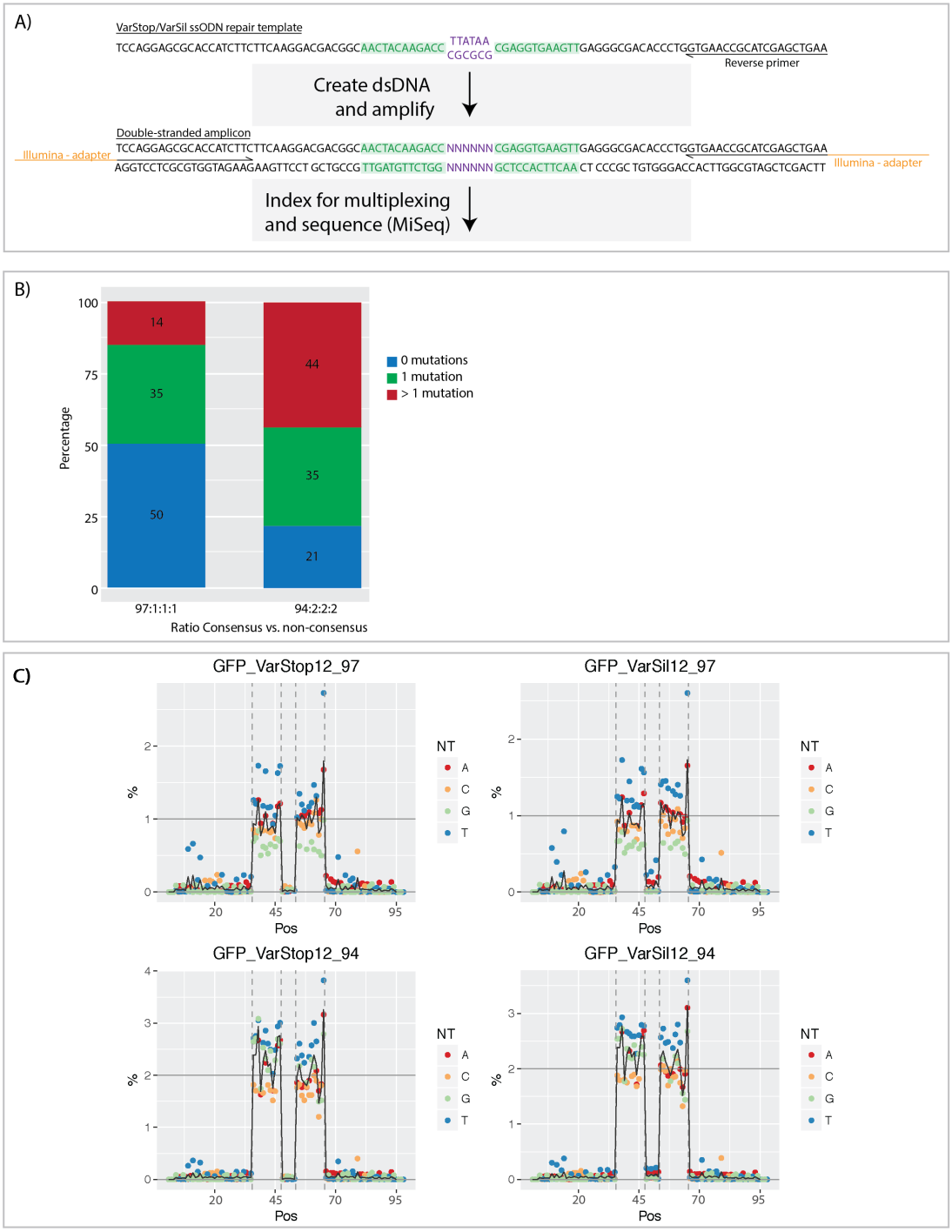


Figure 4.4 Assessment of nucleotide diversity of ssODN pools through deep-sequencing (A) For the sequencing of ssODNs (either with a VarStop (TTATAA) or VarSil (CGCGCG) HDR-core), the single-stranded molecules are PCR-amplified with a reverse primer that transforms it into double-stranded DNA, after which amplification with primers extended with Illumina adapters for sequencing is being performed. These amplicons can then be indexed (as mentioned in figure 4.2) and sequenced. **(B)** Analysis of ssODN shows that in both 97:1:1:1 and 94:2:2:2 ssODNs 35% harbours a single mutation, whilst half of the oligo's in the 97:1:1:1 pool harbours no mutation in the variable region. **(C)** Dot plot of rate of non-consensus nucleotides at each position of the ssODN pool shows that in both VarSil and VarStop with

97% consensus, these rates average around 1% at the variable positions, whereas the 94% consensus pool averages around 2% as expected. T and G are consistently higher present across the variable positions as a result of synthesis bias.

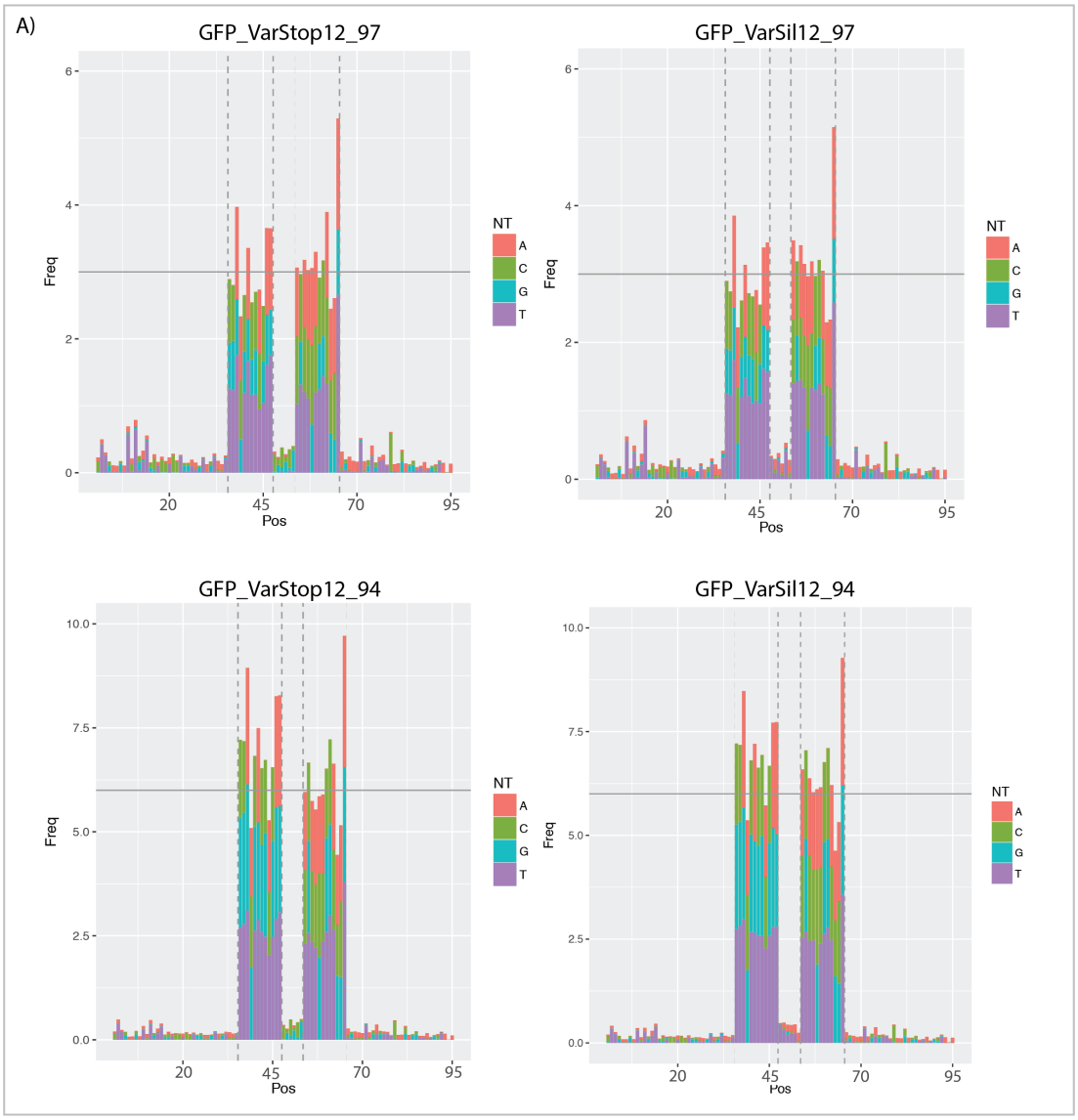


Figure 4.5 Stacked bar plots of nucleotide diversity across ssODN template pools Stacked bar plots showing distribution of non-consensus nucleotides and average substitution rate per position. Horizontal line indicates expected substitution rate in the variable positions (3% and 6% for bottom and top panes, respectively) and shows positional biases that are consistent between the different repair template pools. Non-consensus nucleotides in the non-variable regions are consistent between different experiments, suggesting PCR or sequencing biases.

system for introducing nucleotide diversity into *Gfp* (see **Figure 4.6A**). Cleavage of Cas9 with the respective sgRNAs was assessed by the Cel-1 surveyor assay (see **Figure 4.6B,C**), which

allowed for the detection of nucleotide variants (i.e. NHEJ) at a specific locus (F Ann Ran *et al.*, 2013)

This demonstrated that sgRNA17, sgRNA100 and sgRNA8 promote cleavage of the target DNA by wtCas9, and that sgRNA17 and sgRNA8 together promote the cleavage of DNA by paired nickases. Having validated these reagents for basic editing procedures, their suitability for introducing nucleotide variants through HDR was tested next.

4.2.5 Determining the optimal time between mutagenesis and selection by flow cytometry

Analysis of the RCN(t)-GFP cell line by flow cytometry confirmed a consistent level of high GFP fluorescence in the untransfected cells (see **Figure 4.7A**). The experimental workflow was performed as described in section 4.2.1. Untransfected cells that were subjected to puromycin selection were killed, indicating that the drug selection scheme worked. Transfection of RCN(t)-GFP cells with plasmids (pX335-U6-Chimeric_BB-CBh-hSpCas9n(D10A) (Cong *et al.*, 2013)) and puromycin selection plasmid (pSuper.Puro) slightly affected GFP fluorescence in the absence of sgRNAs (see **Figure 4.7B**), which is most likely due to cellular stress associated with transfection and/or puromycin selection. However, a considerably larger loss in GFP fluorescence was observed in the presence of sgRNA8 and sgRNA17 (82.3% versus 1.3% GFP negative, see **Figure 4.7B**), demonstrating that the change in fluorescence was the result of cleavage in *Gfp*, and that a large majority of surviving cells had received editing reagents.

Cells were then transfected with a puromycin selection plasmid, sgRNA8 and sgRNA17 (both at equimolar ratios from a Cas9n plasmid) and respectively the VarSil12 or VarStop12 ssODN. Fluorescence profiles were assessed at 7- and 14-days post-transfection (see **Figure 4.7C**). For gating, we defined the population corresponding to the GFP negative control as 'GFP negative', whilst the 'GFP positive' population was corresponding with the untransfected RCN(t)-GFP cells. The area between these two populations was called 'GFP intermediate'. In cell populations transfected with the VarStop ssODN, separation into GFP positive and negative populations was clearly noticeable at 7 days post-transfection, with almost no GFP intermediate population (± 0.5 %). However, transfections using the VarSil ssODN donor yielded a ± 7-fold increase in the GFP intermediate population, indicating that the effects of

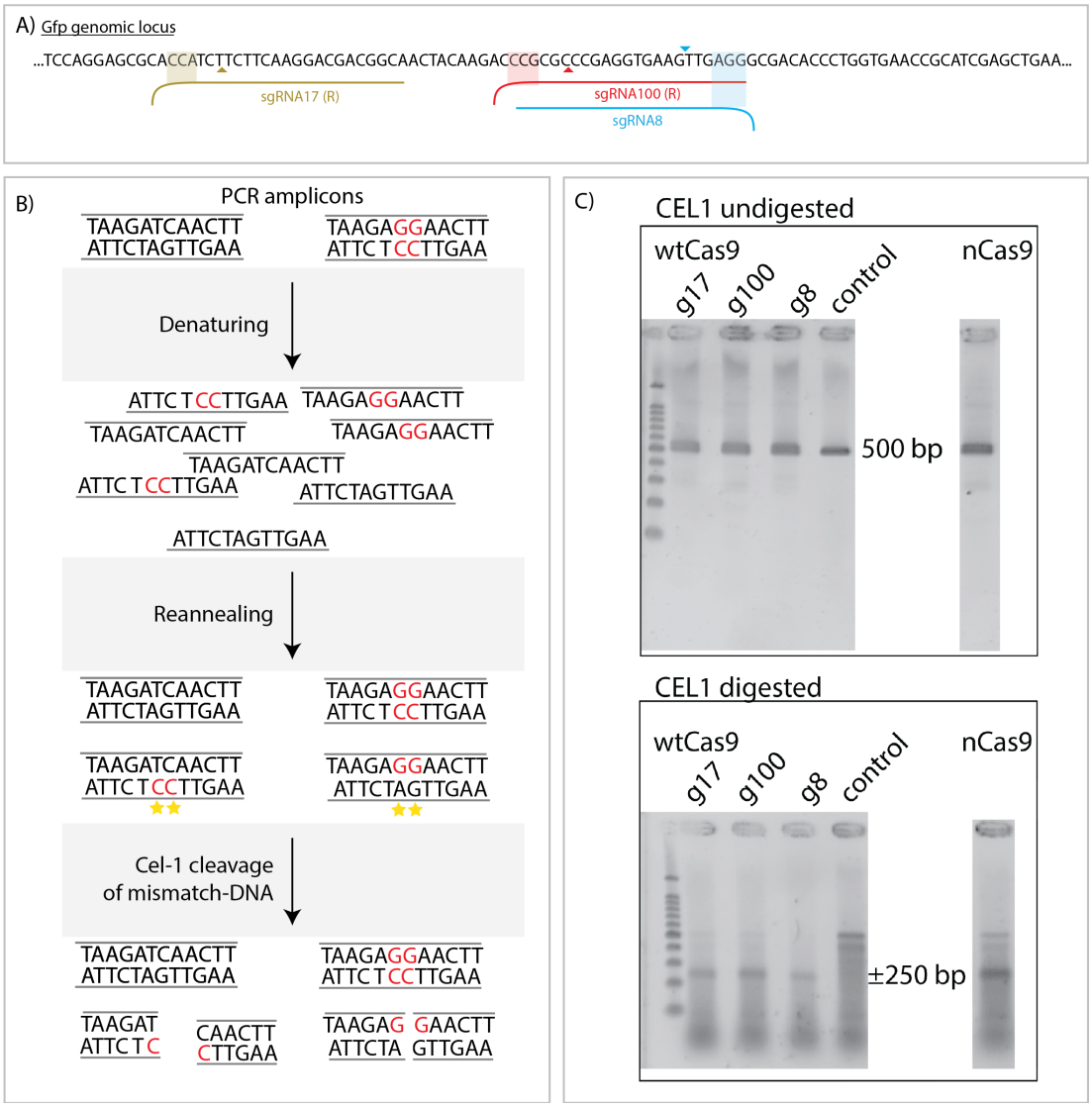


Figure 4.6 Design and Cel1 assay assessing cleavage of sgRNA (A) sgRNA17 and sgRNA100 are anti-sense oriented (indicated by R for reverse), thereby inducing a nick on the antisense strand (as the nickase encoded by the PX335 plasmid has the RuvC domain of Cas9 inactivated), whereas sgRNA8 induces a nick on the sense strand. All three sgRNAs were used for experiments with wtCas9, whilst sgRNA17 and sgRNA8 were additionally used for double Cas9n. **(B)** Schematic overview of Cel-1 surveyor assay. After PCR amplification of gDNA from a targeted cell pool, the amplicon fragments are of approximately equal length regardless of sequence variants (due to NHEJ or HDR). These fragments get denatured and slowly reannealed, which occurs in a random fashion. The Cel-1 enzyme now recognizes DNA that contains mismatches and cleaves this DNA at the target site, hence creating smaller DNA fragments. This DNA pool can now be analysed by gel electrophoresis, in which the presence of smaller DNA fragments indicates the presence of sequence variation and, in the case of Cas9-edited cells, editing. **(C)** Cel-1 surveyor assay showing that wtCas9 cleaves the target site in combination with any of the sgRNAs, and Cas9n cleaves the target DNA in combination with sgRNA17 and sgRNA8. Control is genomic DNA from untargeted cells.

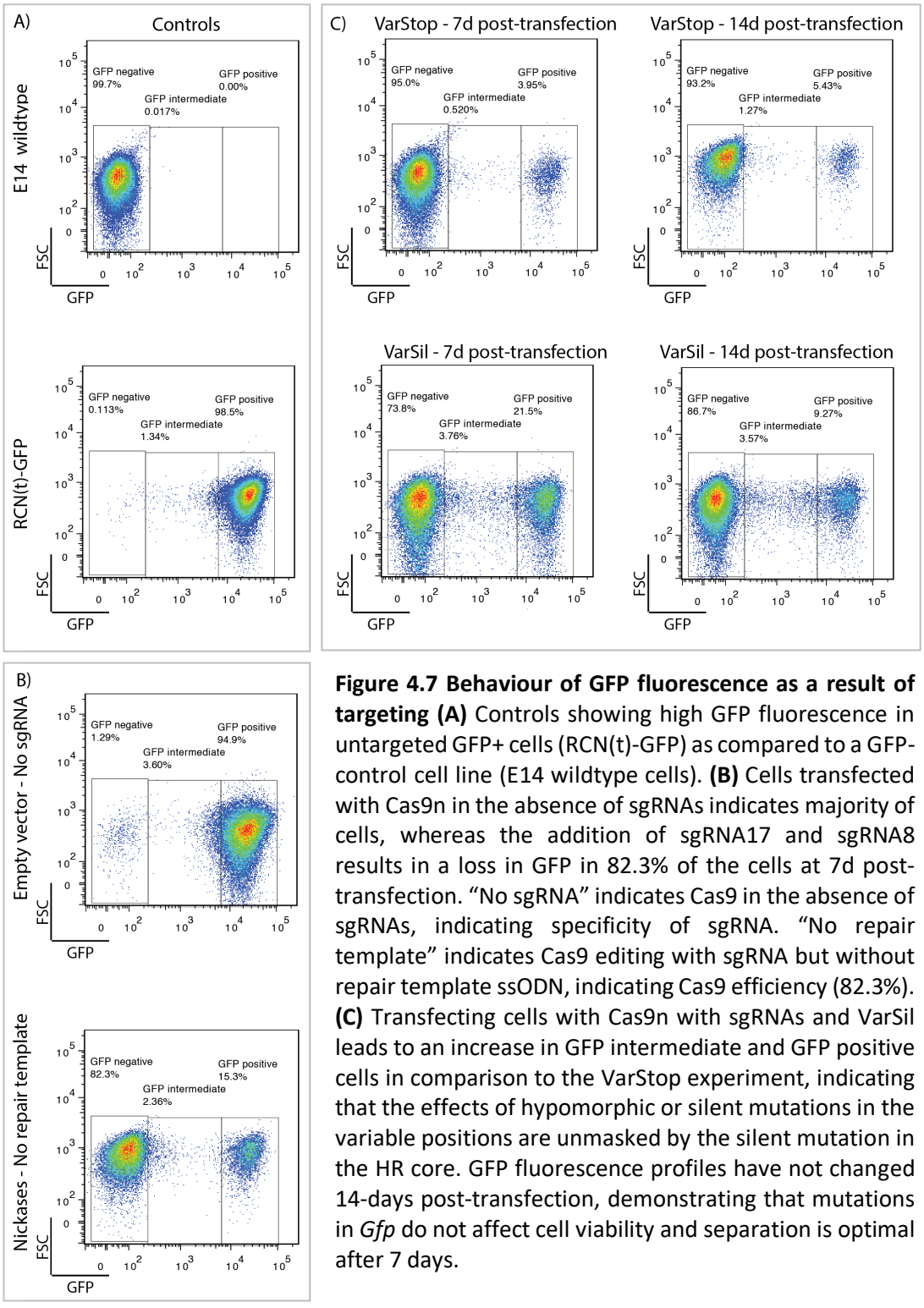


Figure 4.7 Behaviour of GFP fluorescence as a result of targeting (A) Controls showing high GFP fluorescence in untargeted GFP+ cells (RCN(t)-GFP) as compared to a GFP- control cell line (E14 wildtype cells). **(B)** Cells transfected with Cas9n in the absence of sgRNAs indicates majority of cells, whereas the addition of sgRNA17 and sgRNA8 results in a loss in GFP in 82.3% of the cells at 7d post-transfection. "No sgRNA" indicates Cas9 in the absence of sgRNAs, indicating specificity of sgRNA. "No repair template" indicates Cas9 editing with sgRNA but without repair template ssODN, indicating Cas9 efficiency (82.3%). **(C)** Transfecting cells with Cas9n with sgRNAs and VarSil leads to an increase in GFP intermediate and GFP positive cells in comparison to the VarStop experiment, indicating that the effects of hypomorphic or silent mutations in the variable positions are unmasked by the silent mutation in the HR core. GFP fluorescence profiles have not changed 14-days post-transfection, demonstrating that mutations in *Gfp* do not affect cell viability and separation is optimal after 7 days.

hypomorphic mutations in the variable positions are unmasked by the removal of the stop codon inherent to the VarStop ssODN. The fluorescence profile of VarStop-transfected cells

did not change between 7 and 14 days, whereas a slight decrease in GFP positive cells was observed in the VarSil-transfected cells over the second week post-transfection. This was possibly due to second rounds of cleavage. As the intermediate population did however not change, 7 days was considered to be sufficient to observe the effects of mutations on GFP fluorescence and was therefore used as the time point for the isolation of cells for gDNA extraction.

4.2.6 Targeting cells with CRISPR-Cas9 and multiplex library to introduce single nucleotide variants

Transfections of RCN(t)-GFP cells with VarSil and VarStop ssODNs were performed simultaneously in biological duplicates and analysed by flow cytometry (see **Figure 4.8**). Whilst a complete loss of GFP is considered to be the result of either HDR-induced mutations or indels caused by NHEJ, cells with intermediate GFP expression are considered to be exclusively originating from nucleotide variants introduced from the VarSil template through HDR. As mentioned in section 4.2.1, the increase in intermediate population in the VarSil over VarStop condition was used as a proxy to assess the efficiency of HDR in a given experiment. In the nickase experiment, up to 2.43% of the VarSil cells harboured intermediate GFP fluorescence, a 7.3-fold increase over the VarStop experiment suggesting that some cells had undergone incorporation of the HDR repair template. Respectively 82.7% and 95.6% of the cells had lost GFP expression in the VarStop experiment which is thought to reflect overall editing efficiencies (i.e. all HDR plus NHEJ outcomes). As transfections were performed in parallel and editing and HDR efficiencies were thus assumed to be similar between VarSil and VarStop experiments, the HDR efficiency was approximated to be at least the difference in proportion in GFP negative populations, i.e. 3.6% and 5.4%.

Whereas paired Cas9n creates DSBs with sticky (i.e. overhanging) ends, wtCas9 creates blunt DSB ends (as discussed in the main introduction). In parallel experiments, Cas9n was replaced with wtCas9 (pSpCas9(BB)-2A-Puro (PX459) V2.0 (F Ann Ran *et al.*, 2013)) and three different sgRNAs (as depicted in **Figure 4.6A**) to determine whether these different DSB end-structures had an effect on introducing nucleotide diversity through HDR. Whilst there were considerable fluctuations between replicates, wtCas9 with sgRNA100 (cutting centrally with respect to the ssODN template) showed similar increases in intermediate cells, indicating similar rates of HDR compared to the double nickases (nickases vs. sgRNA100, Bonferroni-

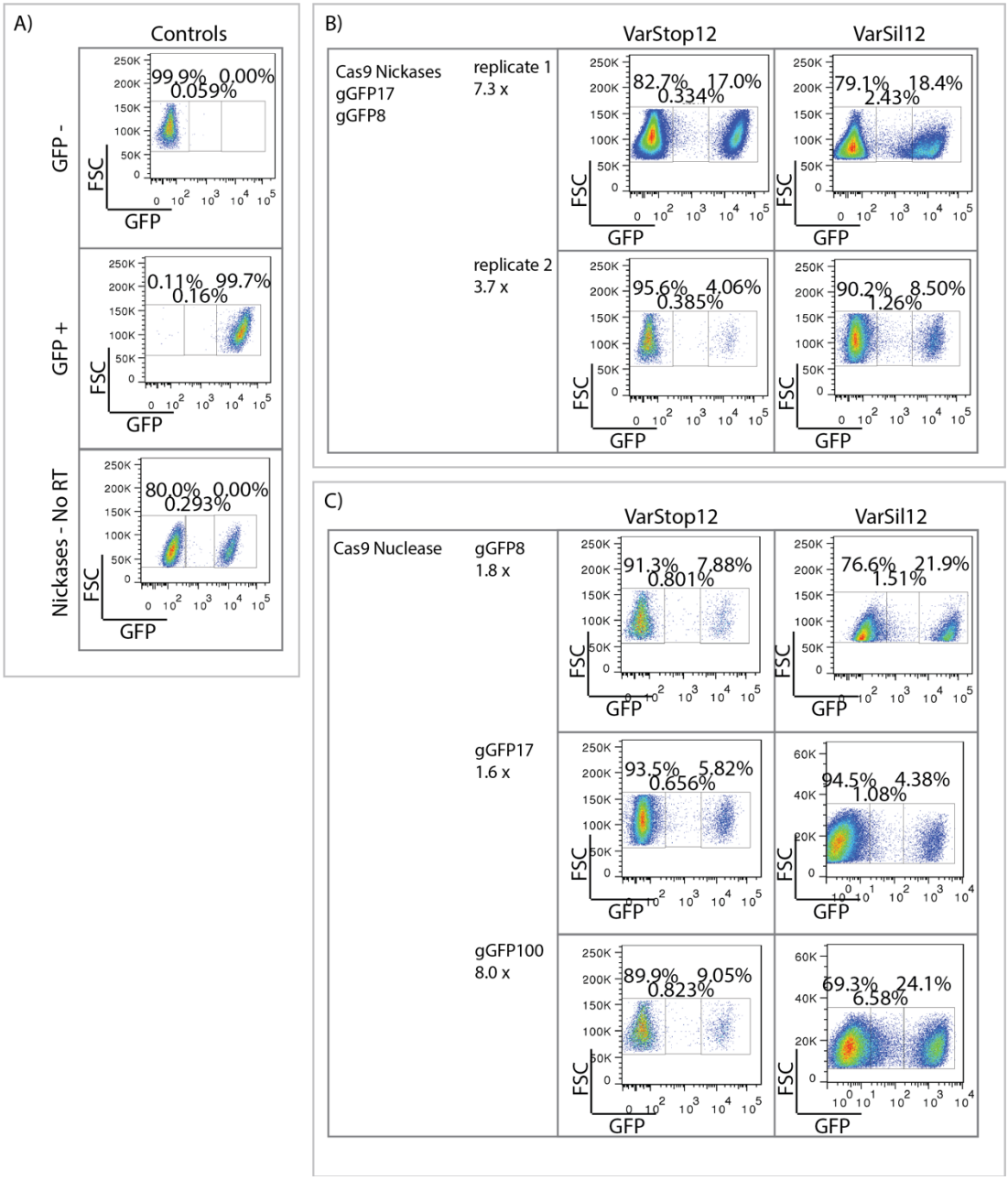


Figure 4.8 Silent mutations unmask the effects of single nucleotide variability (A) Transfection controls as explained in 4.5.a. Percentages on right, centre and left are for the proportion of cells fallin in respectively the negative, intermediate and positive bin. **(B)** Fluorescence profiles of Cas9n experiments. Whilst there is variation between experiments, VarSil fluorescence profiles consistently show an increase in cells with intermediate GFP fluorescence ranging from 3.7X to 7.3X compared to VarStop. **(C)** Fluorescence profiles of wtCas9 experiments show similar patterns to the experiments using Cas9n. Intermediate populations are consistently higher in the VarSil experiment compared to the VarStop experiment.

adjusted t-test p-value = 0.65, see **Figure 4.9**). wtCas9 with either of the distally-binding sgRNAs showed a lower frequency of intermediate cells compared to the use of double

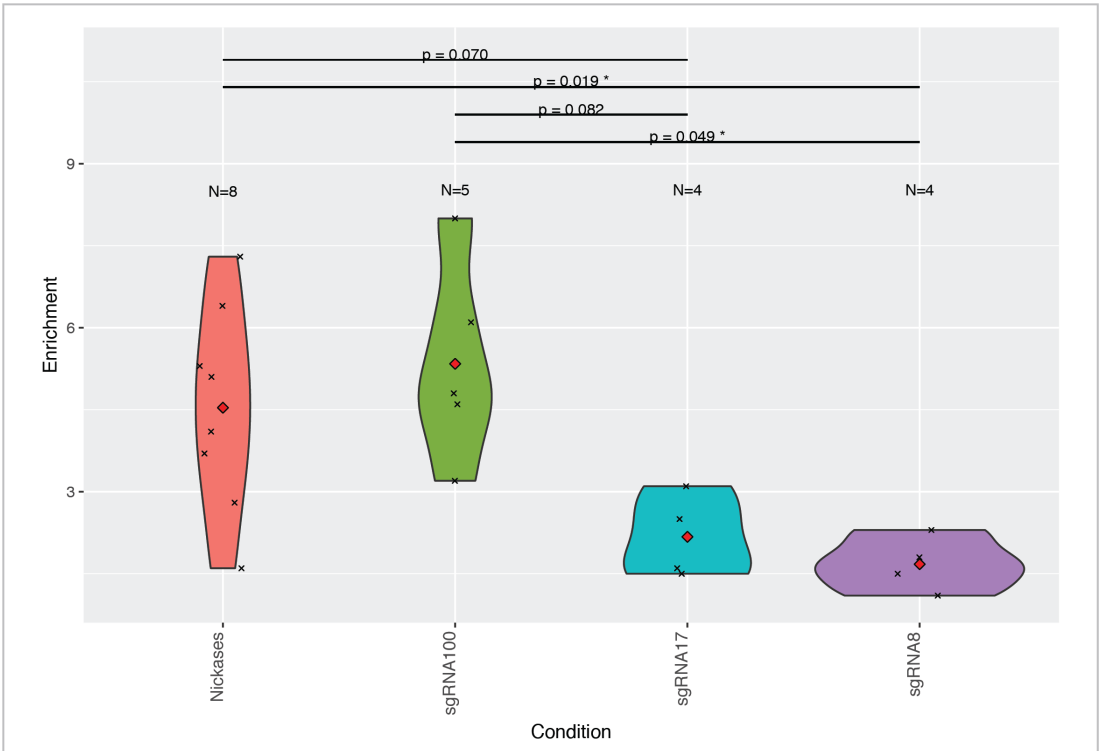


Figure 4.9 Fold increase in intermediate populations VarSil over VarStop in using Cas9 nuclease and nickases Fold increase in intermediate population in VarSil over VarStop experiments of simultaneous transfections using either Cas9n (with sgRNA8 and sgRNA17) or wtCas9 with the sgRNA denoted under the graph, which is used as a measure for HDR efficiencies. The orientation of sgRNA is depicted in Figure 4.6a. The use of wtCas9 with sgRNA8 shows less of an increase in intermediate population compared to either the nickases or wtCas9 combined with sgRNA100.

nickases (nickases vs. sgRNA8 and sgRNA17 p-values of 0.019 and 0.070, sgRNA100 vs. sgRNA8 and sgRNA17 p-values of 0.049 and 0.082), indicating lower efficiencies of integration of the repair template. Hence, HDR efficiencies in the use of wtCas9 are largely dependent on the sgRNA used.

Cells were sorted into three bins based on GFP expression, with a minimum of 100,000 cells per bin. While the voltages between the experiments fluctuated slightly, gates for sorting were kept as constant as possible between experiments to ensure consistency. After isolation of genomic DNA for each of the samples, the target locus was PCR-amplified using primers binding outwith the region covered by the ssODN template (see **Figure 4.2B**), thereby preventing amplification of randomly-integrated oligonucleotide templates. HDR-core sequences were designed to introduce restriction sites that allow detection of HDR-derived sequences in samples by digestion with either PsiI (VarStop) or BsaXI (VarSil) (see **Figure**

4.10A,B). Digestion of PCR amplicons from different bins indicate that HDR cores are present in the respective samples, with the exception of the VarStop bins for the wtCas9 (GFP negative population of wtCas9 with g100 (i.e. centrally cleaving Cas9) and other) experiments. As this is consistent with the FACS profile of the wtCas9 not indicating high HDR efficiencies in these samples, these samples were excluded from further analyses. The remaining samples, including biological duplicates for the VarStop and VarSil Cas9n samples, were prepared in technical triplicates (i.e. separately PCR-amplified) for deep sequencing.

For each sample, 100 nanograms of template DNA was used. As a single diploid murine genome contains approximately 2.86 picograms of DNA and a single copy of *Gfp* is found per genome, approximately 35,000 copies of *Gfp* would hence be sampled in each reaction. Assuming the maximum number of single nucleotide variants being 72 (3 non-consensus bases at 24 sites) and at an HDR efficiency of at least 5% (see first paragraph of this section), this would suggest an average representation of 24 clones per variant. However, as only 35% of the ssODN molecules in the pool contained a single nucleotide substitution in the variable positions, 100 nanograms of template DNA would on average give a minimal of 8 clones per single nucleotide variant per sample. This would be mostly applicable for VarStop negative, where a roughly equal representation of each variant is expected.

4.2.7 Data analysis indicates contaminations across samples

Sequencing data was analysed using the bioinformatics pipeline described in Chapter 3. Contrary to our expectations, analysis of an initial sequencing run revealed that all samples consistently harboured large proportions of reads (> 52%) with full consensus (whilst allowing mismatches) to the wildtype reference sequence in all bins (see **Figure 4.10C**). As wildtype reads were not expected to be present at high rates in any of the GFP negative bins, each step in the processing of the samples was interrogated to elucidate where these anomalies originated.

PCR reactions with primers used for library preparation indicated a product in the absence of template DNA (see **Figure 4.10B**), revealing contamination of one or more reagents used in the preparation of the sequencing libraries. We hypothesised that the broad usage of GFP

plasmids in molecular biology labs could be the potential source of contaminants. Therefore, PCR amplification was cautiously prepared in a chemistry lab using new batches of reagents.

4.2.8 Nucleotide variances can repeatedly be detected in replicates

Sequencing of libraries prepared under these sterile conditions demonstrated that the frequency of wild-type reads in libraries generated from GFP negative cells were largely reduced, although still present at (see **Table 4-1**). This shows that contaminants present in the lab were the origin of these high rates of wildtype reads and that preparation of GFP amplicons requires care to avoid contamination. Each of the experiments was performed twice independently (i.e. two biological replicates per sample), whereas gDNA samples were PCR amplified in three independent technical replicates.

Using the bioinformatics pipeline optimised and explained in the previous chapter, only reads with full consensus between both reads of a read pair (110,228 ± 42,399 contigs per sample) were considered for analysis. Typically, fewer reads from samples originating from the negative bin passed our pipeline, which is in line with the expected increased proportion of reads with indels in this population.

Frequencies of non-consensus nucleotides at the 24 variable positions (i.e. 72 in total) correlated well between technical replicates with an average R^2 of 0.94 (see **Figure 4.11**), indicating high technical reproducibility of our results. Sequencing data from technical replicates were therefore pooled prior to the assessment of correlation between biological replicates. Good correlations were observed (average R^2 of 0.94), these correlations were substantially lower between the biological replicates of the different VarSil bins (average R^2 of 0.45), indicating that different bins harbour different non-consensus nucleotides. Reassuringly, the correlation of VarStop with any of the respective VarSil bins was higher (R^2 of 0.59 ± 0.038), which is in line with our expectation as the VarStop sample should contain all the nucleotide variants also present in the respective VarSil bin. The biological replicates were hence also pooled for downstream analysis, thus resulting in 6 samples per bin for the Cas9 experiments.

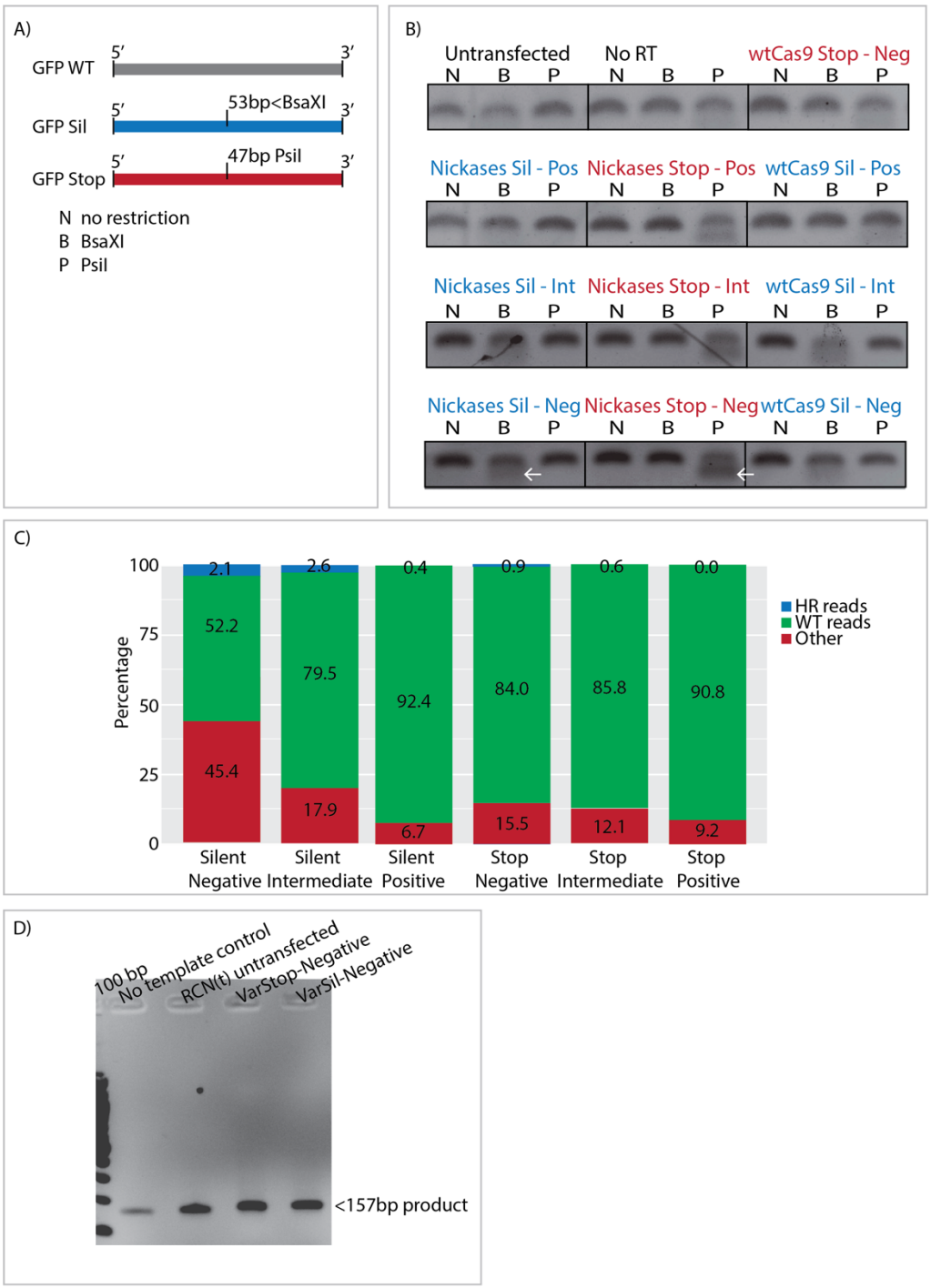


Figure 4.10 Assessing HDR efficiencies using restriction enzymes and contamination of reads (A) Schematic representation of restriction sites. BsaXI specifically recognises and cleaves the HDR core sequence introduced by the VarSil ssODN repair template. PsiI similarly recognises and cleaves sequences introduced by the HDR core sequence on VarStop repair templates. **(B)** Digestion of PCR amplicons from gDNA derived from Cas9-exposed cells. VarSil-derived samples specifically show digestion products when exposed to BsaXI but not PsiI, indicating the presence of HDR sequences. Reciprocally VarStop samples are digested by PsiI samples providing an estimate for the level of HDR reads in these experiments. **(C)** Breakdown of sequence reads based on both the HDR core sequence and consensus with the

reference sequence indicates very high levels of reads with full wildtype sequences, indicating contamination of the sequenced samples. **(D)** Analysis of PCR amplicons by gel electrophoresis shows the presence of a PCR product despite the absence of template DNA, indicating contamination of other reagents with GFP sequences.

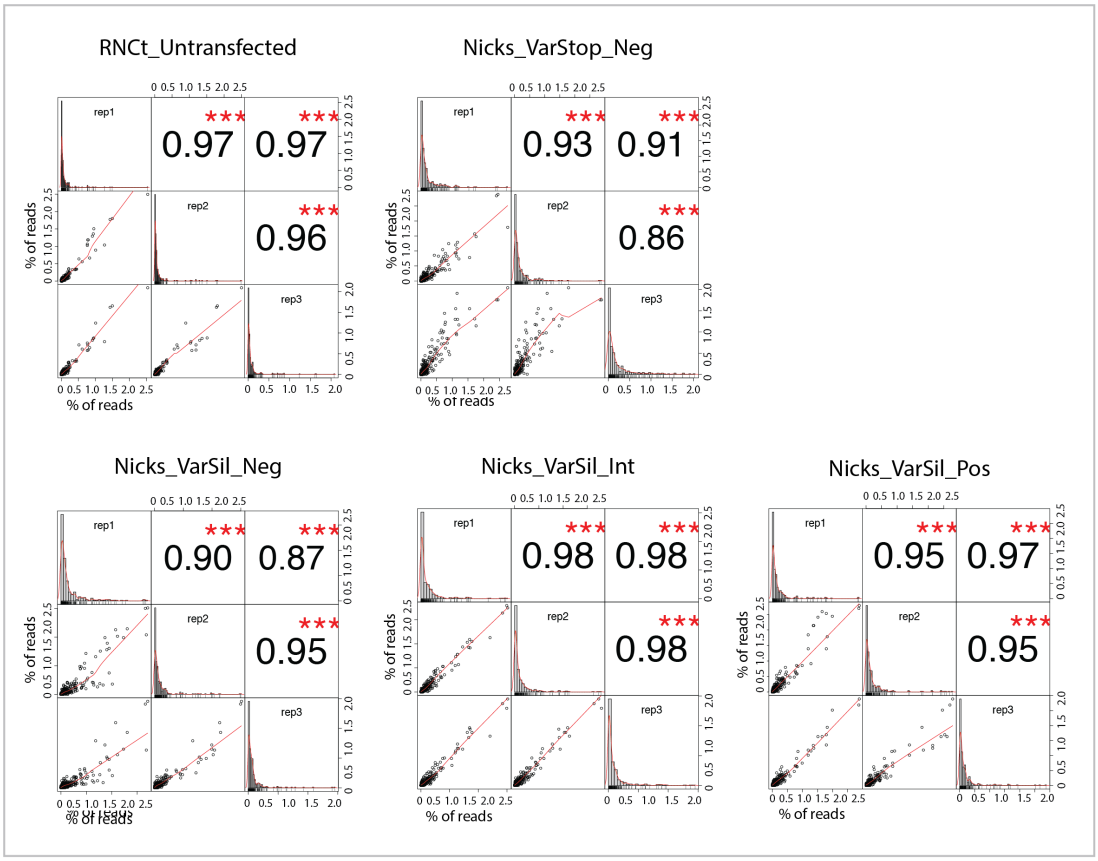


Figure 4.11 Correlations technical replicates Correlation of frequencies of non-consensus nucleotides show high correlations between technical replicates. Reproducibility was highest in the untransfected and VarSil positive samples. Correlation between cbiological replicates can be found in table 4.2.

Table 4-1 Overview of samples and sequence read breakdown All samples were sequenced in two biological replicates (br) and three technical PCR replicates (tr), with the exception of RCNt_untransfected and the wtCas9 experiments respectively. Full consensus read pairs are reads with concordance between forward and reverse reads at each overlapping position. Hr frequencies are consistently higher in the VarStop negative samples. Abbreviations: RCN(t) = name of parent cell line; Neg = GFP negative bin, Int = GFP intermediate bin, Pos = GFP positive bin, Nicks = nickase variant of Cas9, wtCas9 = nuclease variant of Cas9;

Sample	Repair template ssODN	sgRNA	Plasmid backbone	GFP bin	Figure reference	Full consensus read pairs	Reads with HR core	%	Reads with WT core	%	Other	%
RCN(t)_Untransfected_tr1	-	-	-	unsorted	-	97,116	134	0%	78,174	80%	18,808	19%
RCN(t)_Untransfected_tr2	-	-	-	unsorted	-	78,634	471	1%	68,427	87%	9,736	12%
RCN(t)_Untransfected_tr3	-	-	-	unsorted	-	103,296	430	0%	86,885	84%	15,981	15%
Nicks_VarSil12_Neg_br1_tr1	VarSil12	sgRNA8, sgRNA17	PX335-nickase Cas9	negative	fig. 4.9; 4.10; 4.11; 4.13; 4.14; 4.16	102,561	3,235	3%	23,606	23%	75,720	74%
Nicks_VarSil12_Neg_br1_tr2	VarSil12	sgRNA8, sgRNA17	PX335-nickase Cas9	negative	fig. 4.9; 4.10; 4.11; 4.13; 4.14; 4.17	133,931	4,932	4%	31,364	23%	97,635	73%
Nicks_VarSil12_Neg_br1_tr3	VarSil12	sgRNA8, sgRNA17	PX335-nickase Cas9	negative	fig. 4.9; 4.10; 4.11; 4.13; 4.14; 4.18	115,395	3,269	3%	27,477	24%	84,649	73%
Nicks_VarSil12_Neg_br2_tr1	VarSil12	sgRNA8, sgRNA17	PX335-nickase Cas9	negative	fig. 4.9; 4.10; 4.11; 4.13; 4.14; 4.19	102,561	3,235	3%	23,606	23%	75,720	74%
Nicks_VarSil12_Neg_br2_tr2	VarSil12	sgRNA8, sgRNA17	PX335-nickase Cas9	negative	fig. 4.9; 4.10; 4.11; 4.13; 4.14; 4.20	133,931	4,932	4%	31,364	23%	97,635	73%
Nicks_VarSil12_Neg_br2_tr3	VarSil12	sgRNA8, sgRNA17	PX335-nickase Cas9	negative	fig. 4.9; 4.10; 4.11; 4.13; 4.14; 4.21	115,395	3,269	3%	27,477	24%	84,649	73%
Nicks_VarSil12_Int_br1_tr1	VarSil12	sgRNA8, sgRNA17	PX335-nickase Cas9	intermediate	fig. 4.9; 4.10; 4.11; 4.13; 4.14; 4.22	171,960	12,384	7%	105,473	61%	54,103	31%

Nicks_VarSil12_Int_br1_tr2	VarSil12	sgRNA8, sgRNA17	PX335-nickase Cas9	intermediate	fig. 4.9; 4.10; 4.11; 4.13; 4.14; 4.23	181,367	14,505	8%	112,272	62%	54,590	30%
Nicks_VarSil12_Int_br1_tr3	VarSil12	sgRNA8, sgRNA17	PX335-nickase Cas9	intermediate	fig. 4.9; 4.10; 4.11; 4.13; 4.14; 4.24	146,499	9,623	7%	93,503	64%	43,373	30%
Nicks_VarSil12_Int_br2_tr1	VarSil12	sgRNA8, sgRNA17	PX335-nickase Cas9	intermediate	fig. 4.9; 4.10; 4.11; 4.13; 4.14; 4.25	171,960	12,384	7%	105,473	61%	54,103	31%
Nicks_VarSil12_Int_br2_tr2	VarSil12	sgRNA8, sgRNA17	PX335-nickase Cas9	intermediate	fig. 4.9; 4.10; 4.11; 4.13; 4.14; 4.26	181,367	14,505	8%	112,272	62%	54,590	30%
Nicks_VarSil12_Int_br2_tr3	VarSil12	sgRNA8, sgRNA17	PX335-nickase Cas9	intermediate	fig. 4.9; 4.10; 4.11; 4.13; 4.14; 4.27	146,499	9,623	7%	93,503	64%	43,373	30%
Nicks_VarSil12_Pos_br1_tr1	VarSil12	sgRNA8, sgRNA17	PX335-nickase Cas9	positive	fig. 4.9; 4.10; 4.11; 4.13; 4.14; 4.28	195,895	8,164	4%	164,091	84%	23,640	12%
Nicks_VarSil12_Pos_br1_tr2	VarSil12	sgRNA8, sgRNA17	PX335-nickase Cas9	positive	fig. 4.9; 4.10; 4.11; 4.13; 4.14; 4.29	162,247	6,601	4%	141,901	87%	13,745	8%
Nicks_VarSil12_Pos_br1_tr3	VarSil12	sgRNA8, sgRNA17	PX335-nickase Cas9	positive	fig. 4.9; 4.10; 4.11; 4.13; 4.14; 4.30	139,699	5,475	4%	121,895	87%	12,329	9%
Nicks_VarSil12_Pos_br2_tr1	VarSil12	sgRNA8, sgRNA17	PX335-nickase Cas9	positive	fig. 4.9; 4.10; 4.11; 4.13; 4.14; 4.31	195,895	8,164	4%	164,091	84%	23,640	12%
Nicks_VarSil12_Pos_br2_tr2	VarSil12	sgRNA8, sgRNA17	PX335-nickase Cas9	positive	fig. 4.9; 4.10; 4.11; 4.13; 4.14; 4.32	162,247	6,601	4%	141,901	87%	13,745	8%
Nicks_VarSil12_Pos_br2_tr3	VarSil12	sgRNA8, sgRNA17	PX335-nickase Cas9	positive	fig. 4.9; 4.10; 4.11; 4.13; 4.14; 4.33	139,699	5,475	4%	121,895	87%	12,329	9%
Nicks_VarStop12_Neg_br1_tr1	VarStop12	sgRNA8, sgRNA17	PX335-nickase Cas9	negative	fig. 4.9; 4.10; 4.11; 4.12; 4.14; 4.16	41,052	5,332	13%	9,988	24%	25,732	63%
Nicks_VarStop12_Neg_br1_tr2	VarStop12	sgRNA8, sgRNA17	PX335-nickase Cas9	negative	fig. 4.9; 4.10; 4.11; 4.12; 4.14; 4.17	106,127	13,642	13%	27,520	26%	64,965	61%

Nicks_VarStop12_Neg_br1_tr3	VarStop12	sgRNA8, sgRNA17	PX335-nickase Cas9	negative	fig. 4.9; 4.10; 4.11; 4.12; 4.14; 4.18	135,708	17,218	13%	42,825	32%	75,665	56%
Nicks_VarStop12_Neg_br2_tr1	VarStop12	sgRNA8, sgRNA17	PX335-nickase Cas9	negative	fig. 4.9; 4.10; 4.11; 4.12; 4.14; 4.19	56,663	6,936	12%	47,308	83%	2,419	4%
Nicks_VarStop12_Neg_br2_tr2	VarStop12	sgRNA8, sgRNA17	PX335-nickase Cas9	negative	fig. 4.9; 4.10; 4.11; 4.12; 4.14; 4.20	128,408	13,024	10%	109,533	85%	5,851	5%
Nicks_VarStop12_Neg_br2_tr3	VarStop12	sgRNA8, sgRNA17	PX335-nickase Cas9	negative	fig. 4.9; 4.10; 4.11; 4.12; 4.14; 4.21	79,653	10,772	14%	65,118	82%	3,763	5%
wtCas9_g17_VarSil12_Neg_br1	VarSil12	sgRNA17	PX459V2.0 – WT Cas9	negative	fig 4.15	46,653	1662	4%	12542	27%	32449	70%
wtCas9_g17_VarSil12_Neg_br2	VarSil12	sgRNA17	PX459V2.0 – WT Cas9	negative	fig 4.15	70,494	3948	6%	21781	31%	44765	64%
wtCas9_g17_VarSil12_Int_br1	VarSil12	sgRNA17	PX459V2.0 – WT Cas9	intermediate	fig 4.15	106,310	522	0%	58812	55%	46976	44%
wtCas9_g17_VarSil12_Int_br2	VarSil12	sgRNA17	PX459V2.0 – WT Cas9	intermediate	fig 4.15	72,133	204	0%	42759	59%	29170	40%
wtCas9_g100_VarSil12_Neg_br1	VarSil12	sgRNA100	PX459V2.0 – WT Cas9	negative	fig 4.15	89,232	684	1%	28913	32%	59635	67%
wtCas9_g100_VarSil12_Neg_br2	VarSil12	sgRNA100	PX459V2.0 – WT Cas9	negative	fig 4.15	88,203	4080	5%	26828	30%	57295	65%
wtCas9_g100_VarSil12_Int_br1	VarSil12	sgRNA100	PX459V2.0 – WT Cas9	intermediate	fig 4.15	61,854	264	0%	36066	58%	25524	41%
wtCas9_g100_VarSil12_Int_br2	VarSil12	sgRNA100	PX459V2.0 – WT Cas9	intermediate	fig 4.15	55,672	240	0%	34488	62%	20944	38%
wtCas9_g100_VarSil12_Pos_br1	VarSil12	sgRNA100	PX459V2.0 – WT Cas9	positive	fig 4.15	102,440	1284	1%	66916	65%	34240	33%
wtCas9_g100_VarSil12_Pos_br2	VarSil12	sgRNA100	PX459V2.0 – WT Cas9	positive	fig 4.15	127,153	1350	1%	77330	61%	48473	38%
wtCas9_g8_VarSil12_Neg_br1	VarSil12	sgRNA8	PX459V2.0 – WT Cas9	negative	fig 4.15	46,141	3339	7%	16923	37%	25879	56%
wtCas9_g8_VarSil12_Neg_br2	VarSil12	sgRNA8	PX459V2.0 – WT Cas9	negative	fig 4.15	74,999	5241	7%	38651	52%	31107	41%
wtCas9_g8_VarSil12_Int_br1	VarSil12	sgRNA8	PX459V2.0 – WT Cas9	intermediate	fig 4.15	42,544	504	1%	23310	55%	18730	44%
wtCas9_g8_VarSil12_Int_br2	VarSil12	sgRNA8	PX459V2.0 – WT Cas9	intermediate	fig 4.15	84,438	690	1%	50280	60%	33468	40%

wtCas9_g8_VarSil12_Pos_br1	VarSil12	sgRNA8	PX459V2.0 – WT Cas9	positive	fig 4.15	71,571	2868	4%	60153	84%	8550	12%
wtCas9_g8_VarSil12_Pos_br2	VarSil12	sgRNA8	PX459V2.0 – WT Cas9	positive	fig 4.15	74,204	2016	3%	54171	73%	18017	24%

Table 4-2 Coefficients of determination of biological replicates After pooling of the technical replicates, correlation coefficients for all biological replicates for all samples were calculated and transformed into coefficients of determination (i.e. R^2). Biological replicates of the four samples (VarStop, VarSil negative, VarSil intermediate, VarSil positive) highly correlate, indicating sufficient sampling and sequencing depth across the samples. The correlation of VarStop with any of the respective VarSil bins was higher than between the VarSil bins, which is in line with our expectation as the VarStop sample should contain all the nucleotide variants also present in the respective VarSil bin.

	GFP_Nov_VarStop12_Neg_br1	GFP_Nov_VarStop12_Neg_br2	GFP_Nov_VarSil12_Neg_br1	GFP_Nov_VarSil12_Neg_br2	GFP_Nov_VarSil12_Int_br1	GFP_Nov_VarSil12_Int_br2	GFP_Nov_VarSil12_Pos_br1	GFP_Nov_VarSil12_Pos_br2
Nicks_VarStop12_Neg_br1		0.94	0.58	0.56	0.65	0.62	0.52	0.53
Nicks_VarStop12_Neg_br2	0.94		0.62	0.59	0.63	0.60	0.57	0.55
Nicks_VarSil12_Neg_br1	0.58	0.62		0.89	0.52	0.53	0.45	0.44
Nicks_VarSil12_Neg_br2	0.56	0.59	0.89		0.43	0.47	0.43	0.44
Nicks_VarSil12_Int_br1	0.65	0.63	0.52	0.43		0.97	0.43	0.43
Nicks_VarSil12_Int_br2	0.62	0.60	0.53	0.47	0.97		0.40	0.41
Nicks_VarSil12_Pos_br1	0.52	0.57	0.45	0.43	0.43	0.40		0.97
Nicks_VarSil12_Pos_br2	0.53	0.55	0.44	0.44	0.43	0.41	0.97	

4.2.9 Nucleotide variants on variable regions are present at lower but equal proportions in genome compared to ssODN

As discussed in section 4.2.2, the VarStop sample allowed for calculating both the HDR efficiencies and nucleotide substitution frequencies irrespective of phenotypic selection. Across the samples, an average HDR frequency of 12.3 ± 0.010% was seen in the VarStop Negative sample (see **Table 4-1**), as determined by the proportion of reads containing the 6-nt HDR core sequence that was specifically introduced from the ssODN donor.

In the HDR-derived reads, the average rates of substitution (i.e. presence of non-consensus bases) across the 24 variable positions and 70 constant positions were respectively 4.4% and 0.065% per nucleotide position (considered the background noise level) as compared to 6.2%

and 0.22% in the ssODN donor pool. The rate of substitutions in the constant positions was likely due to either PCR or sequencing error as has been previously described (Austinat *et al.*, 2008) and discussed in the previous chapter. The decrease in non-consensus nucleotides in the variable positions in the experimental samples compared to the ssODN could indicate short incorporation tracts that include the HDR core but not the variable positions.

Reassuringly however, all 72 possible single nucleotide substitutions were detected above background noise level (0.065, see previous paragraph) in the VarStop Negative sample, demonstrating that our pipeline can be used to introduce all possible single nucleotide substitutions into a specified genomic region (see **Figure 4.12A**). In addition, comparisons between the edited VarStop Negative sample and the VarStop ssODN repair template demonstrated that the proportions of non-consensus nucleotides did not differ at the variable positions (see **Figure 4.12B,D** and **Table 4-3**; p-values for z-scores for the variable positions were > 0.05 after false discovery rate correction). While the previous paragraph showed that the overall frequency of non-consensus nucleotides is lower, these results demonstrate that the ratio of non-consensus nucleotide variants per position did not differ between VarStop Negative and ssODN.

4.2.10 Unmasking the effects of genomic variants using VarSil allows for phenotypic selection of destabilising mutations

Whereas the VarStop experiment allowed for the assessment of HDR integration efficiencies, the VarSil experiment was intended to assess the functional consequences of the single nucleotide variants introduced by HDR. HDR frequencies in the VarSil bins reached 2.3% - 6.7% (see **Table 4-1**), indicating that HDR efficiencies were slightly lower in comparison to the VarStop experiment. The majority of reads in the GFP negative population of VarSil were reads that had a length different to that of the reference sequence, which presumably harboured indels, and non-consensus nucleotides (see 'Other', **Table 4-1**), whilst the majority of the reads in the positive population had full consensus with the wild type sequence. This demonstrates that the phenotypic selection scheme used in this chapter correctly identifies the effects of mutations on GFP fluorescence.

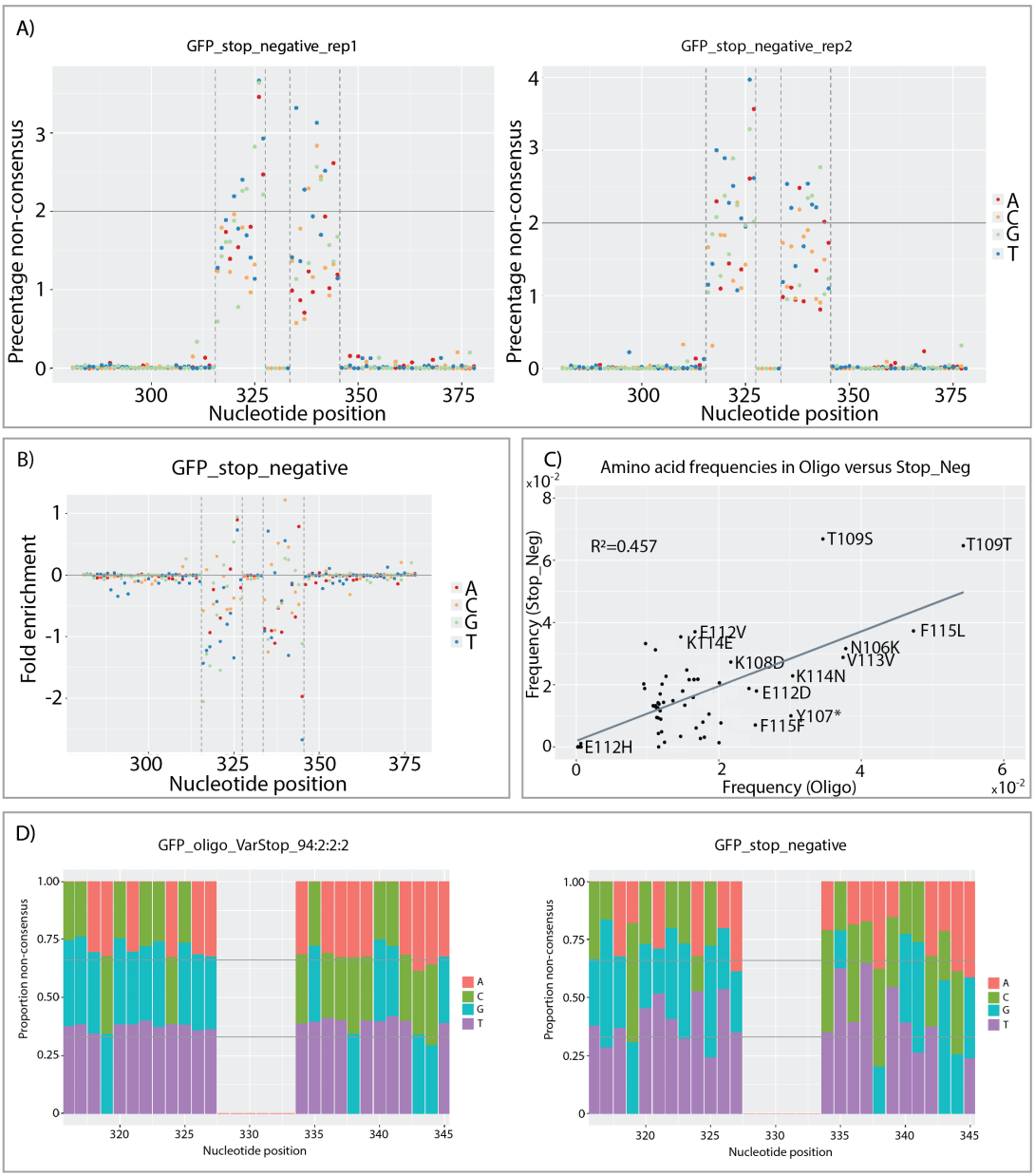


Figure 4.12 Distribution of mutations in HR reads from VarStop Negative (A) Dot plot of the rate of non-consensus nucleotides at each position of the targeted region in HR reads. This shows an enrichment of non-consensus nucleotides in the variable regions but not in the constant flanking region. These findings are consistent across both biological replicates. **(B)** Fold enrichment of non-consensus nucleotides in VarStop sample over the VarStop oligo was calculated by dividing the frequency of each non-consensus nucleotide in the VarStop ssODN repair template from the VarStop genomic sample. Dots below the horizontal grey line are non-consensus nucleotides underrepresented in the genomic sample, whereas dots above the line are found more frequently in the genomic sample compared to the ssODN. These results show that the majority of the nucleotide frequencies are within 1% difference. **(C)** Correlation plot of amino acid substitutions rates at the variable positions in VarStop ssODN donor and genomic sample. Nucleotide variants in the respective samples were translated into amino acid substitutions, after which their frequency in the total HDR pool was calculated. These findings show that the amino acids in the VarStop sample weakly correlate

with those in the ssODN pool. **(D)** Proportions of non-consensus bases per variable position show variances between the VarStop oligo and genomic sample. All non-consensus nucleotides are however represented in both the oligo and genomic sample.

Similar to the VarStop experiment, the VarSil samples showed an increase in non-consensus nucleotides in the variable regions (see **Figure 4.13A**). Assessment of the number of nucleotide variances per HDR template revealed that in VarStop Negative, 28% of the reads containing the HR core did not harbour any additional mutations, which is higher than the 21% that was observed in the ssODN template pool (see **Table 4-4**). In the VarSil samples, the number of HDR reads without additional mutations ranged from 3% in the Negative bin to 33% in the Positive bin. The VarSil Negative bin harboured the highest percentage of reads with multiple mutations (see **Table 4-4**), which is to be expected as multiple mutations are assumed to have a destabilising effect on the stability of the GFP protein. Assessment of the classes of individual nucleotide substitutions reveals that nonsense mutations are predominantly found in the VarSil Negative class while silent mutations most frequently occur in the VarSil positive population, although also still present in the Negative and Intermediate populations (see **Figure 4.13B**). Together, these results demonstrate that replacing the VarStop with VarSil HR core unmasks the effects of adjacent mutations on GFP fluorescence.

4.2.11 Cas9 nuclease is less effective than Cas9 nickase in introducing nucleotide diversity

We next wanted to assess the effect of wtCas9 compared to the Cas9 paired nickases, for which both the HDR frequency and nucleotide diversity were assessed (see **Table 4-1** and **Figure 4.14**). sgRNAs were designed to be in different orientations and proximities to the HDR core sequence. Based on read analysis, HDR frequencies were consistently lower using wtCas9 regardless of the position of the sgRNA with respect to the ssODN (see **Table 4-1**). This is in contrast to the increase in intermediate population in VarSil over VarStop found in WT Cas9 with sgRNA100 (see **Figure 4.9**), which was also used as a measure for HDR efficiencies (see section 4.2.6).

Nucleotide diversity, as measured by the proportions of non-consensus nucleotides in the variable region (**Figure 4.14**), was lower compared to the double nickases when using WT Cas9 with sgRNA8 and sgRNA17 but not WT Cas9 with sgRNA100. Combined with the fact

Table 4-3 P-values for z-scores for non-consensus nucleotides in HR reads In order the ascertain the differences in proportions of each non-consensus nucleotide at each position between the VarStop negative sample and VarStop ssODN, Z-scores were calculated for each non-consensus nucleotide. Afterwards, significance of Z-scores was assessed corrected for false-discovery. Whilst nucleotides in the constant regions (nucleotide positions 208-315 and 345-378) were significantly different (highlighted in red) between the genomic VarStop ssODN and genomic VarStop sample, the non-consensus nucleotide frequencies in the variable region (316-327 and 334-345) did not differ. As HR reads were selected based the 6-nt core sequence, these positions could not be calculated. NA indicates consensus nucleotides.

Nucleotide	281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297	298	299	300	301	302	303	304	305
A	3E-01	4E-01	5E-01	NA	2E-04	2E-04	NA	2E-10	5E-22	6E-01	4E-01	NA	3E-02	2E-15	NA	1E-02	4E-01	2E-15	1E-01	1E-06	1E-03	6E-01	5E-01	NA	NA
C	4E-01	NA	NA	1E-01	2E-01	3E-01	6E-02	3E-04	NA	8E-05	NA	1E+00	NA	NA	7E-02	5E-02	NA	2E-14	2E-16	NA	2E-08	1E+00	NA	2E-01	1E-03
G	NA	4E-01	NA	6E-04	NA	NA	5E-01	NA	3E-01	NA	6E-01	2E-01	1E-04	NA	8E-03	5E-01	9E-03	2E-01	8E-20	NA	3E-05	3E-01	NA	2E-01	1E-03
T	NA	7E-01	2E-01	6E-03	2E-05	2E-03	6E-02	4E-14	3E-19	4E-04	3E-01	1E-01	3E-04	2E-15	7E-02	NA	7E-03	NA	NA	1E-06	NA	NA	5E-01	5E-01	3E-01

Nucleotide	306	307	308	309	310	311	312	313	314	315	316	317	318	319	320	321	322	323	324	325	326	327	328	329	330
A	6E-13	3E-12	NA	3E-03	4E-04	NA	3E-02	2E-02	9E-04	2E-03	NA	NA	8E-01	4E-01	NA	9E-01	NA	NA	1E+00	NA	2E-01	5E-01	NA	NA	NA
C	7E-02	2E-02	NA	NA	3E-07	2E-01	NA	5E-02	3E-05	NA	5E-01	4E-01	NA	4E-01	8E-01	NA	5E-01	9E-01	3E-01	9E-01	NA	NA	NA	NA	NA
G	NA	NA	9E-10	4E-02	NA	2E-03	5E-02	NA	NA	7E-01	6E-01	2E-01	7E-01	9E-01	4E-01	2E-01	6E-01	7E-01	NA	4E-01	5E-01	6E-01	NA	NA	NA
T	4E-15	2E-15	9E-10	3E-01	6E-01	6E-03	3E-01	3E-01	7E-01	3E-03	1E+00	4E-01	8E-01	NA	6E-01	3E-01	1E+00	7E-01	4E-01	3E-01	8E-01	9E-01	NA	NA	NA

Nucleotide	331	332	333	334	335	336	337	338	339	340	341	342	343	344	345	346	347	348	349	350	351	352	353	354	355
A	NA	NA	NA	5E-01	NA	4E-01	3E-01	8E-01	2E-01	NA	NA	1E+00	4E-01	9E-01	4E-01	2E-14	NA	1E-08	5E-12	1E-02	9E-01	9E-16	NA	6E-01	NA
C	NA	NA	NA	4E-01	6E-01	4E-01	5E-01	6E-01	9E-01	8E-01	9E-01	9E-01	8E-01	1E+00	NA	6E-01	1E-01	4E-02	4E-02	9E-02	NA	7E-04	NA	NA	7E-01
G	NA	NA	NA	NA	1E-01	NA	NA	3E-01	NA	8E-01	1E-01	NA	3E-01	8E-01	5E-01	NA	9E-03	NA	NA	NA	8E-01	NA	1E-09	1E-01	2E-01
T	NA	NA	NA	8E-01	9E-02	9E-01	2E-01	NA	4E-01	1E+00	1E-01	9E-01	NA	NA	1E-01	4E-13	2E-02	2E-11	1E-10	7E-04	1E+00	1E-09	1E-09	6E-01	3E-01

Nucleotide	356	357	358	359	360	361	362	363	364	365	366	367	368	369	370	371	372	373	374	375	376	377	378
A	9E-05	6E-05	3E-07	4E-01	3E-09	1E-02	9E-03	4E-05	NA	NA	6E-01	2E-01	1E+00	7E-06	NA	4E-04	5E-07	2E-19	NA	2E-02	2E-02	NA	3E-21
C	NA	NA	NA	2E-01	5E-05	2E-01	6E-02	3E-03	NA	4E-03	NA	NA	6E-01	NA	NA	1E-03	NA	3E-01	5E-14	9E-01	NA	4E-02	1E-03
G	1E-02	1E-02	NA	1E-01	NA	NA	7E-01	NA	7E-01	5E-01	4E-01	NA	NA	NA	5E-06	8E-01	NA	NA	1E-14	NA	8E-01	4E-13	NA
T	8E-08	9E-06	3E-07	NA	8E-04	2E-02	NA	1E-02	7E-01	2E-02	9E-01	2E-01	1E+00	7E-06	5E-06	NA	5E-07	3E-21	6E-01	7E-01	7E-01	NA	NA

Table 4-4 Number of non-consensus nucleotides in variable positions in HR reads The number of non-consensus nucleotides per HR read varies per bin. In VarStop Negative, nearly half of the reads have a single mutation. Reads with no mutation predominantly reside in the VarSil Positive bin, whereas reads with multiple mutations predominantly fall in the VarSil Negative bin. There is a skew in incorporation of reads with one mutation over reads with multiple mutations when comparing rates in the VarStop Negative sample with the VarStop oligo.

Sample	Total no. HR reads	Reads with no mutation		Reads with single mutation		Reads with multiple mutations	
Nicks_VarSil12_Neg	22873	893	4%	14856	65%	7124	31%
Nicks_VarSil12_Int	73024	3602	5%	53167	73%	731	22%
Nicks_VarSil12_Pos	40480	13591	34%	14491	38%	12398	28%
Nicks_VarStop12_Neg	66924	19451	29%	31623	47%	15850	24%
Nicks_VarStop12_oligo94:2:2:2	146874	30843	21%	51406	35%	64625	44%

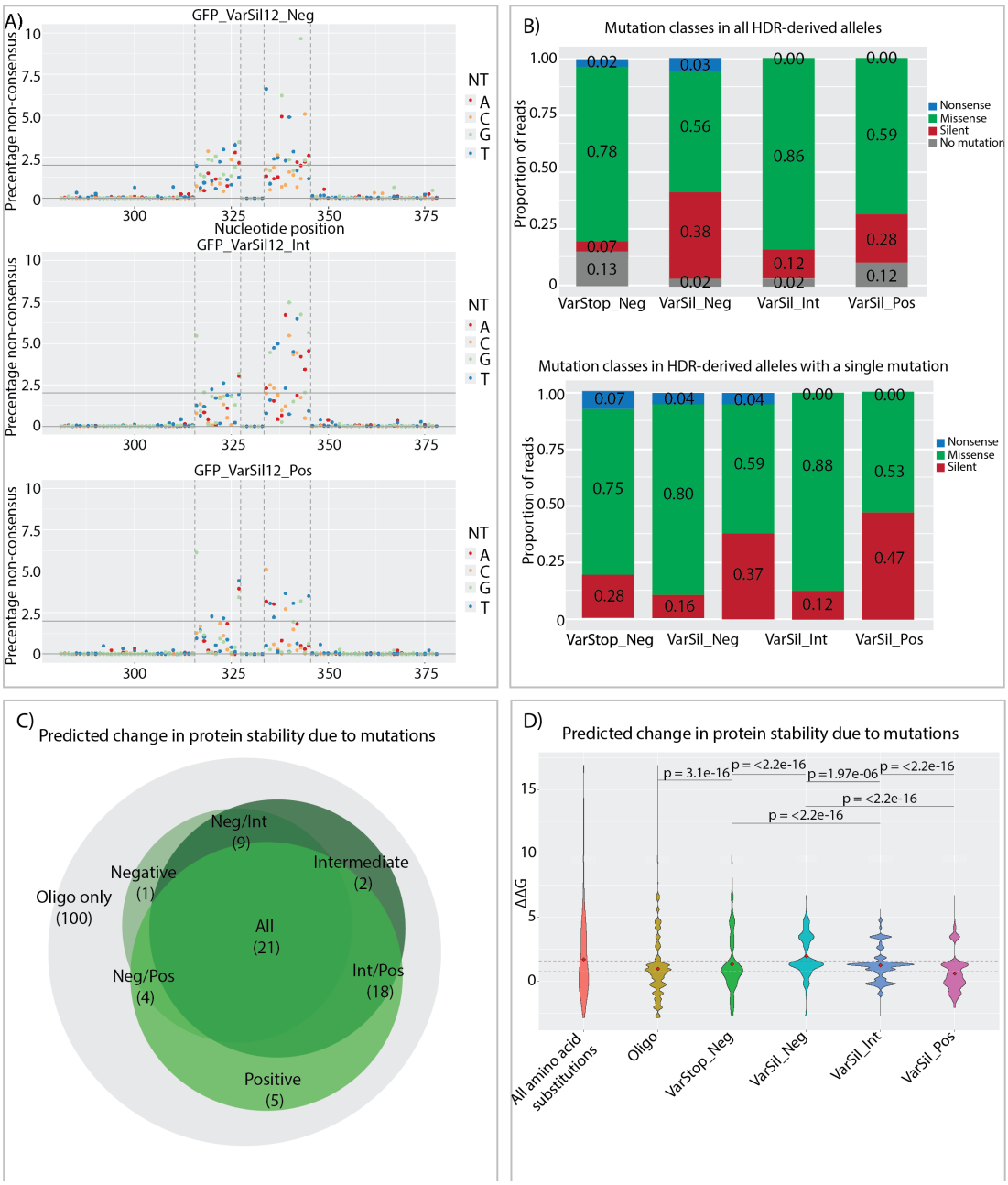


Figure 4.13 Distribution of mutations in HR reads from the VarSil samples (A) Dot plots indicate an enrichment of non-consensus nucleotides in the variable positions over the constant positions and a slight increase in the right variable region over the left variable region (see table 4.3) **(B)** Bar plots of amino acid substitution classes. Non-sense mutations are predominantly observed in the negative bin of VarSil, whereas missense mutations form the predominant class in the intermediate bin and the positive bin shows the highest number of reads without a mutation. These differences become more apparent when observing reads harbouring only a single mutation. **(C)** Non-consensus amino acids at the variable positions that are observed in at least a single read across the VarSil samples. All 160 amino acids were observed in the repair template oligo, but only 60 were observed in the VarSil genomic samples. The majority of the amino acid substitutions were observed in multiple bins, although at different rates (see table 4.6). **(D)** Violin plots of ΔΔG values for the amino

acids found in the different classes. Most left, distribution of all $\Delta\Delta G$ values theoretically possible for amino acid substitutions at the variable positions. Oligo represents the amino acid substitutions found on the ssODN repair templates, whereas VarStop Negative shows a similar distribution, indicating there is no selection for destabilising mutations. Amino acid substitutions found in the VarSil samples show different distributions, suggesting a selective distribution of amino acid substitutions based on their effect on protein stability. P-values were calculated using Wilcox signed-rank sum test.

that the HDR efficiencies were consistently higher in the nickase experiments, these results suggest that double nickases are favoured over nucleases for introducing single nucleotide variants from ssODNs. Analyses hereafter are hence performed solely on samples from the nickase dataset.

4.2.12 Amino acid substitutions are limited by the required number of nucleotide substitutions

After having confirmed that single nucleotide variants could be introduced into our region of interest, the next step was to assess whether the functional consequences of mutations could be assessed across the VarSil bins by predicting their effect on protein stability. Therefore, nucleotide variants were translated into amino acid substitutions. Theoretically, 19 possible amino acid substitutions could be found at any of the 8 codons that were targeted by the variable positions, hence totalling 152 possible amino acid substitutions. Out of these, 67 substitutions (44.1%) were found in the VarStop negative bin and 94 (61.8%) across the VarSil bins (see **Table 4-5**). Although the HDR efficiencies were higher in the VarStop Negative sample than for any of the VarSil samples, the number of reads was considerably higher for the latter, thereby explaining the substantial higher number of amino acid substitutions found in the VarSil bins.

As some amino acids are encoded by more codons than others and amino acid substitutions at each of the target residues can require up to three separate nucleotide substitutions, the 'distance' from one codon to another in terms of nucleotide changes was assessed for each site (see **Table 4-6**). All 51 amino acid substitutions that can be achieved by a single nucleotide change and half (43/87) of all the amino acids requiring two nucleotide changes were found in the VarSil samples, whilst none of the amino acids that required 3 nucleotide changes were observed. These findings show that the range of amino acid substitutions that can be assessed by the pipeline is dictated by the required number of nucleotide

Figure 4.14 Dot plots of non-consensus nucleotides in wtCas9 samples (A) Dot plots of the wtCas9 samples with sgRNA17 shows an enrichment of nucleotide substitutions in the right variable region over the left variable region as was observed in the Cas9n experiment, albeit at a higher frequency. In the intermediate population, a much lower nucleotide diversity was observed. The positive sample was not sequenced. (**B)** sgRNA100 shows similar distributions to sgRNA17. (**C)** sgRNA8 results in low sequence diversity, likely due to lower HR frequencies.

substitutions at each position.

Table 4-5 Amino acid substitutions found in variable regions in HR reads List of all amino acid substitutions and consensus amino acids found in repair template ssODN, VarStop Negative and VarSil Negative, Intermediate and Positive samples, respectively. For each sample, numbers indicate absolute read counts (left column) and proportion (right column) for each sample. ΔΔG values were extracted from table 4.9. While all 160 possible amino acids were detected in the ssODN, 67 unique amino acid substitutions were found in the VarStop negative bin and 94 substitutions across the VarSil bins. Table was sorted based on read count in priority VarSil_Neg, VarSil_Int, VarSil_Pos and amino acid.

Refer-ence	Posi-tion	Alter-native	Mutation class	ddG	VarStop ssODN Read count	VarStop ssODN Proportion	VarStop_Neg Read count	VarStop_Neg Proportion	VarSil_Neg Read count	VarSil_Neg Proportion	VarSil_Int Read count	VarSil_Int Proportion	VarSil_Pos Read count	VarSil_Pos Proportion
N	106	*	nonsense	-0.93	79	8.5E-04	-	-	-	-	-	-	-	-
N	106	A	missense	-0.93	35	3.7E-04	-	-	-	-	-	-	-	-
N	106	C	missense	-1.06	40	4.3E-04	-	-	-	-	-	-	-	-
N	106	E	missense	-0.83	97	1.0E-03	-	-	-	-	-	-	-	-
N	106	F	missense	-1.53	40	4.3E-04	-	-	-	-	-	-	-	-
N	106	G	missense	0.09	44	4.7E-04	-	-	-	-	-	-	-	-
N	106	L	missense	-2.83	25	2.7E-04	-	-	-	-	-	-	-	-
N	106	M	missense	-2.86	37	4.0E-04	-	-	-	-	-	-	-	-
N	106	P	missense	0.95	10	1.1E-04	-	-	-	-	-	-	-	-
N	106	Q	missense	-1.43	62	6.6E-04	-	-	-	-	-	-	-	-
N	106	R	missense	-1.53	106	1.1E-03	-	-	-	-	-	-	-	-
N	106	V	missense	-1.43	40	4.3E-04	-	-	-	-	-	-	-	-
Y	107	S	missense	6.46	979	1.0E-02	114	1.5E-02	-	-	-	-	-	-
Y	107	A	missense	5.28	24	2.6E-04	-	-	-	-	-	-	-	-
Y	107	E	missense	-0.84	54	5.8E-04	-	-	-	-	-	-	-	-
Y	107	I	missense	2.71	27	2.9E-04	-	-	-	-	-	-	-	-
Y	107	K	missense	4.60	32	3.4E-04	-	-	-	-	-	-	-	-
Y	107	L	missense	1.86	93	1.0E-03	-	-	-	-	-	-	-	-

Y	107	P	missense	4.59	20	2.1E-04	-	-	-	-	-	-	-	-
Y	107	Q	missense	4.25	53	5.7E-04	-	-	-	-	-	-	-	-
Y	107	R	missense	3.62	31	3.3E-04	-	-	-	-	-	-	-	-
Y	107	T	missense	5.37	19	2.0E-04	-	-	-	-	-	-	-	-
Y	107	V	missense	3.54	27	2.9E-04	-	-	-	-	-	-	-	-
Y	107	W	missense	7.66	37	4.0E-04	-	-	-	-	-	-	-	-
K	108	A	missense	1.18	26	2.8E-04	-	-	-	-	-	-	-	-
K	108	D	missense	1.11	39	4.2E-04	-	-	-	-	-	-	-	-
K	108	F	missense	-0.09	2	2.1E-05	-	-	-	-	-	-	-	-
K	108	G	missense	1.83	43	4.6E-04	-	-	-	-	-	-	-	-
K	108	H	missense	0.44	38	4.1E-04	-	-	-	-	-	-	-	-
K	108	I	missense	-0.47	68	7.3E-04	-	-	-	-	-	-	-	-
K	108	L	missense	-0.28	65	7.0E-04	-	-	-	-	-	-	-	-
K	108	P	missense	3.62	11	1.2E-04	-	-	-	-	-	-	-	-
K	108	S	missense	1.23	84	9.0E-04	-	-	-	-	-	-	-	-
K	108	V	missense	0.06	31	3.3E-04	-	-	-	-	-	-	-	-
K	108	W	missense	-0.16	33	3.5E-04	-	-	-	-	-	-	-	-
K	108	Y	missense	0.12	58	6.2E-04	-	-	-	-	-	-	-	-
T	109	*	nonsense	0.93	2	2.1E-05	-	-	-	-	-	-	-	-
T	109	C	missense	-0.24	30	3.2E-04	-	-	-	-	-	-	-	-
T	109	D	missense	3.25	37	4.0E-04	-	-	-	-	-	-	-	-
T	109	E	missense	-0.04	1	1.1E-05	-	-	-	-	-	-	-	-
T	109	F	missense	10.19	46	4.9E-04	-	-	-	-	-	-	-	-
T	109	G	missense	1.96	42	4.5E-04	-	-	-	-	-	-	-	-
T	109	H	missense	7.71	38	4.1E-04	-	-	-	-	-	-	-	-

T	109	K	missense	2.00	58	6.2E-04	-	-		-	-		-	-		-	-
T	109	L	missense	-1.49	35	3.7E-04	-	-		-	-		-	-		-	-
T	109	M	missense	-2.42	41	4.4E-04	-	-		-	-		-	-		-	-
T	109	Q	missense	-0.38	2	2.1E-05	-	-		-	-		-	-		-	-
T	109	R	missense	3.76	111	1.2E-03	-	-		-	-		-	-		-	-
T	109	V	missense	-2.03	44	4.7E-04	-	-		-	-		-	-		-	-
T	109	Y	missense	16.90	34	3.6E-04	-	-		-	-		-	-		-	-
E	112	M	missense	-1.18	31	3.3E-04		11	1.5E-03	-	-		-	-		-	-
E	112	*	nonsense	0.61	1335	1.4E-02	-	-		-	-		-	-		-	-
E	112	F	missense	-1.67	1	1.1E-05	-	-		-	-		-	-		-	-
E	112	I	missense	-1.11	1	1.1E-05	-	-		-	-		-	-		-	-
E	112	K	missense	-0.91	1076	1.2E-02	-	-		-	-		-	-		-	-
E	112	L	missense	-1.27	64	6.9E-04	-	-		-	-		-	-		-	-
E	112	P	missense	0.49	18	1.9E-04	-	-		-	-		-	-		-	-
V	113	*	nonsense	2.96	17	1.8E-04	-	-		-	-		-	-		-	-
V	113	D	missense	4.40	25	2.7E-04	-	-		-	-		-	-		-	-
V	113	F	missense	6.60	57	6.1E-04	-	-		-	-		-	-		-	-
V	113	H	missense	3.32	1	1.1E-05	-	-		-	-		-	-		-	-
V	113	I	missense	-0.63	62	6.6E-04	-	-		-	-		-	-		-	-
V	113	K	missense	3.41	15	1.6E-04	-	-		-	-		-	-		-	-
V	113	N	missense	3.36	1	1.1E-05	-	-		-	-		-	-		-	-
V	113	P	missense	4.70	11	1.2E-04	-	-		-	-		-	-		-	-
V	113	Q	missense	3.92	20	2.1E-04	-	-		-	-		-	-		-	-
V	113	R	missense	5.77	42	4.5E-04	-	-		-	-		-	-		-	-
V	113	S	missense	4.09	25	2.7E-04	-	-		-	-		-	-		-	-

V	113	T	missense	2.29	15	1.6E-04	-	-	-	-	-	-	-	-
V	113	W	missense	11.86	24	2.6E-04	-	-	-	-	-	-	-	-
K	114	A	missense	4.77	30	3.2E-04	-	-	-	-	-	-	-	-
K	114	C	missense	0.38	2	2.1E-05	-	-	-	-	-	-	-	-
K	114	D	missense	2.25	68	7.3E-04	-	-	-	-	-	-	-	-
K	114	F	missense	-1.55	1	1.1E-05	-	-	-	-	-	-	-	-
K	114	G	missense	0.88	26	2.8E-04	-	-	-	-	-	-	-	-
K	114	H	missense	0.14	53	5.7E-04	-	-	-	-	-	-	-	-
K	114	I	missense	2.59	99	1.1E-03	-	-	-	-	-	-	-	-
K	114	L	missense	-0.35	88	9.4E-04	-	-	-	-	-	-	-	-
K	114	P	missense	2.89	20	2.1E-04	-	-	-	-	-	-	-	-
K	114	S	missense	0.42	88	9.4E-04	-	-	-	-	-	-	-	-
K	114	V	missense	2.75	52	5.6E-04	-	-	-	-	-	-	-	-
K	114	W	missense	0.81	45	4.8E-04	-	-	-	-	-	-	-	-
K	114	Y	missense	-1.39	59	6.3E-04	-	-	-	-	-	-	-	-
F	115	*	nonsense	3.54	84	9.0E-04	-	-	-	-	-	-	-	-
F	115	A	missense	3.54	12	1.3E-04	-	-	-	-	-	-	-	-
F	115	D	missense	5.70	21	2.2E-04	-	-	-	-	-	-	-	-
F	115	G	missense	4.46	7	7.5E-05	-	-	-	-	-	-	-	-
F	115	H	missense	3.02	10	1.1E-04	-	-	-	-	-	-	-	-
F	115	K	missense	2.95	1	1.1E-05	-	-	-	-	-	-	-	-
F	115	M	missense	1.02	30	3.2E-04	-	-	-	-	-	-	-	-
F	115	N	missense	3.80	17	1.8E-04	-	-	-	-	-	-	-	-
F	115	P	missense	2.23	12	1.3E-04	-	-	-	-	-	-	-	-
F	115	Q	missense	3.55	1	1.1E-05	-	-	-	-	-	-	-	-

F	115	R	missense	2.75	16	1.7E-04	-	-		-	-		-	-		-	-	
F	115	T	missense	3.84	21	2.2E-04	-	-		-	-		-	-		-	-	
F	115	W	missense	0.60	21	2.2E-04	-	-		-	-		-	-		-	-	
E	112	R	missense	-0.60	59	6.3E-04	-	-		-	-		-	-			4	7.7E-04
E	112	Q	missense	-0.10	1001	1.1E-02	-	-		-	-		-	-			7	1.3E-03
E	112	T	missense	0.19	19	2.0E-04	-	-		-	-		-	-			18	3.5E-03
E	112	S	missense	0.97	18	1.9E-04	-	-		-	-		1	5.8E-05		-	-	
E	112	Y	missense	-1.57	55	5.9E-04	8		1.1E-03	-	-		1	5.8E-05		1		1.9E-04
E	112	N	missense	0.36	40	4.3E-04	-	-		-	-		1	5.8E-05		12		2.3E-03
N	106	H	missense	0.33	994	1.1E-02	105		1.4E-02	-	-		1	5.8E-05		25		4.8E-03
K	108	K	silent	0.00	934	1.0E-02	98		1.3E-02	-	-		2	1.2E-04		15		2.9E-03
K	114	R	missense	-1.42	1303	1.4E-02	172		2.3E-02	-	-		2	1.2E-04		91		1.7E-02
K	108	N	missense	0.92	1878	2.0E-02	202		2.7E-02	-	-		3	1.7E-04		8		1.5E-03
E	112	H	missense	-0.72	31	3.3E-04	-	-		-	-		4	2.3E-04		12		2.3E-03
Y	107	Y	silent	-0.01	1573	1.7E-02	133		1.8E-02	-	-		4	2.3E-04		19		3.6E-03
N	106	T	missense	-1.39	992	1.1E-02	93		1.2E-02	-	-		4	2.3E-04		23		4.4E-03
K	108	T	missense	0.49	962	1.0E-02	231		3.1E-02	-	-		6	3.5E-04		31		5.9E-03
K	108	E	missense	0.89	1091	1.2E-02	168		2.2E-02	-	-		8	4.6E-04		31		5.9E-03
K	108	R	missense	0.26	1420	1.5E-02	118		1.6E-02	-	-		13	7.5E-04		8		1.5E-03
N	106	K	missense	-2.07	3276	3.5E-02	234		3.1E-02	-	-		13	7.5E-04		43		8.2E-03
Y	107	G	missense	6.83	31	3.3E-04	-	-		-	-		15	8.6E-04		-	-	
E	112	V	missense	-0.46	1443	1.5E-02	274		3.6E-02	-	-		20	1.2E-03		177		3.4E-02
Y	107	F	missense	0.62	1431	1.5E-02	160		2.1E-02	-	-		61	3.5E-03		68		1.3E-02
T	109	A	missense	0.93	1328	1.4E-02	98		1.3E-02	-	-		129	7.4E-03		-	-	
T	109	T	silent	0.00	4705	5.0E-02	479		6.4E-02	-	-		202	1.2E-02		504		9.7E-02

Y	107	C	missense	4.77	1373	1.5E-02	160	2.1E-02	-	-		269	1.5E-02	11	2.1E-03
F	115	Y	missense	-0.42	1019	1.1E-02	86	1.1E-02	-	-		380	2.2E-02	17	3.3E-03
T	109	N	missense	1.63	1479	1.6E-02	161	2.1E-02	1	1.6E-04	-	-		-	-	
E	112	W	missense	-1.47	20	2.1E-04	-	-	1	1.6E-04	-	-		-	-	
N	106	S	missense	-0.31	1557	1.7E-02	23	3.1E-03	1	1.6E-04	5	2.9E-04	18	3.5E-03		
K	108	M	missense	-0.23	1294	1.4E-02	133	1.8E-02	1	1.6E-04	10	5.8E-04	48	9.2E-03		
F	115	F	silent	-0.03	2177	2.3E-02	52	6.9E-03	1	1.6E-04	25	1.4E-03	224	4.3E-02		
N	106	N	silent	-0.02	1758	1.9E-02	57	7.6E-03	2	3.3E-04	1	5.8E-05	27	5.2E-03		
K	114	T	missense	0.24	1175	1.3E-02	110	1.5E-02	2	3.3E-04	216	1.2E-02	66	1.3E-02		
T	109	S	missense	0.82	2998	3.2E-02	495	6.6E-02	3	4.9E-04	85	4.9E-03	30	5.8E-03		
Y	107	N	missense	5.61	978	1.0E-02	70	9.3E-03	4	6.5E-04	-	-		-	-	
K	108	Q	missense	0.57	995	1.1E-02	69	9.2E-03	4	6.5E-04	3	1.7E-04	46	8.8E-03		
K	108	*	nonsense	1.18	1457	1.6E-02	45	6.0E-03	5	8.1E-04	7	4.0E-04	-	-		
K	114	Q	missense	0.13	842	9.0E-03	246	3.3E-02	5	8.1E-04	692	4.0E-02	40	7.7E-03		
N	106	I	missense	-1.73	1613	1.7E-02	78	1.0E-02	6	9.8E-04	1	5.8E-05	18	3.5E-03		
F	115	C	missense	3.47	820	8.8E-03	150	2.0E-02	6	9.8E-04	346	2.0E-02	2	3.8E-04		
E	112	E	silent	-0.01	1037	1.1E-02	36	4.8E-03	9	1.5E-03	4	2.3E-04	67	1.3E-02		
V	113	A	missense	2.96	1042	1.1E-02	149	2.0E-02	10	1.6E-03	129	7.4E-03	1	1.9E-04		
E	112	A	missense	0.61	1027	1.1E-02	66	8.8E-03	12	2.0E-03	198	1.1E-02	115	2.2E-02		
K	114	E	missense	1.33	1270	1.4E-02	262	3.5E-02	13	2.1E-03	1888	1.1E-01	14	2.7E-03		
Y	107	H	missense	4.48	965	1.0E-02	98	1.3E-02	15	2.4E-03	4	2.3E-04	-	-		
N	106	Y	missense	-0.93	1539	1.6E-02	59	7.8E-03	15	2.4E-03	6	3.5E-04	26	5.0E-03		
K	114	M	missense	-1.03	1735	1.9E-02	10	1.3E-03	16	2.6E-03	307	1.8E-02	156	3.0E-02		
K	114	N	missense	0.92	2630	2.8E-02	169	2.2E-02	17	2.8E-03	1963	1.1E-01	18	3.5E-03		
F	115	I	missense	2.62	975	1.0E-02	93	1.2E-02	21	3.4E-03	531	3.1E-02	15	2.9E-03		

Y	107	*	nonsense	5.28	2609	2.8E-02	74	9.8E-03	24	3.9E-03	4	2.3E-04	1	1.9E-04
V	113	M	missense	0.54	998	1.1E-02	32	4.2E-03	24	3.9E-03	17	9.8E-04	5	9.6E-04
Y	107	D	missense	6.72	942	1.0E-02	97	1.3E-02	26	4.2E-03	-	-	2	3.8E-04
K	114	K	silent	0.00	1231	1.3E-02	126	1.7E-02	29	4.7E-03	30	1.7E-03	134	2.6E-02
V	113	E	missense	4.20	1006	1.1E-02	102	1.4E-02	46	7.5E-03	135	7.8E-03	76	1.5E-02
E	112	D	missense	1.18	2190	2.3E-02	133	1.8E-02	51	8.3E-03	679	3.9E-02	125	2.4E-02
F	115	L	missense	1.26	4099	4.4E-02	276	3.7E-02	53	8.6E-03	1345	7.7E-02	34	6.5E-03
E	112	G	missense	1.82	1268	1.4E-02	25	3.3E-03	69	1.1E-02	914	5.3E-02	-	-
V	113	G	missense	4.62	1073	1.1E-02	11	1.5E-03	72	1.2E-02	11	6.3E-04	1	1.9E-04
T	109	I	missense	-2.74	1738	1.9E-02	153	2.0E-02	79	1.3E-02	7	4.0E-04	-	-
F	115	S	missense	4.16	1061	1.1E-02	106	1.4E-02	163	2.7E-02	6	3.5E-04	-	-
K	114	*	nonsense	4.77	1343	1.4E-02	183	2.4E-02	212	3.5E-02	-	-	3	5.8E-04
T	109	P	missense	2.75	1020	1.1E-02	126	1.7E-02	246	4.0E-02	29	1.7E-03	-	-
V	113	L	missense	-0.21	2099	2.2E-02	139	1.8E-02	296	4.8E-02	2130	1.2E-01	41	7.9E-03
F	115	V	missense	3.45	831	8.9E-03	139	1.8E-02	733	1.2E-01	1214	7.0E-02	408	7.8E-02
N	106	D	missense	1.32	1509	1.6E-02	20	2.7E-03	1592	2.6E-01	1518	8.7E-02	895	1.7E-01
V	113	V	silent	0.00	3243	3.5E-02	213	2.8E-02	2251	3.7E-01	1777	1.0E-01	1434	2.7E-01

Table 4-6 Nucleotide per codon overlap table For each alternative triplet at each target codon, the number of overlapping nucleotides with the reference codon were calculated and depicted in this graph. These values can be used to calculate the likelihood of an amino acid substitution to arise from nucleotide substitutions at each triplet. Corresponding codons (i.e. 3 nucleotides overlap) are highlighted in dark grey.

			A	A	A	A	C	C	D	D	E	E	F	F	G	G	G	G	H	H	I	I	I	K	K	L	L	L	L	L	L	Met	N	N
Pos	Ref AA	Ref codon	GCA	GCC	GCG	GCT	TGC	TGT	GAC	GAT	GAA	GAG	TTC	TTT	GGA	GGC	GGG	GGT	CAC	CAT	ATA	ATC	ATT	AAA	AAG	TTA	TTG	CTA	CTC	CTG	CTT	ATG	AAC	AAT
106	N	AAC	0	1	0	0	1	0	2	1	1	1	1	0	0	1	0	0	2	1	1	2	1	2	2	0	0	0	1	0	0	1	3	2
107	Y	TAC	0	1	0	0	2	1	2	1	1	1	2	1	0	1	0	0	2	1	0	1	0	1	1	1	1	0	1	0	0	0	2	1
108	K	AAG	0	0	1	0	0	0	1	1	1	2	0	0	0	0	1	0	1	1	1	1	1	2	3	0	1	0	0	1	0	2	2	2
109	T	ACC	1	2	1	1	1	0	1	0	0	0	1	0	0	1	0	0	1	0	1	2	1	1	1	0	0	0	1	0	0	1	2	1
112	E	GAG	1	1	2	1	0	0	2	2	2	3	0	0	1	1	2	1	1	1	0	0	0	1	2	0	1	0	0	1	0	1	1	1
113	V	GTG	1	1	2	1	0	0	1	1	1	2	1	1	1	1	2	1	0	0	1	1	1	0	1	1	2	1	1	2	1	2	0	0
114	K	AAG	0	0	1	0	0	0	1	1	1	2	0	0	0	0	1	0	1	1	1	1	1	2	3	0	1	0	0	1	0	2	2	2
115	F	TTC	0	1	0	0	2	1	1	0	0	0	3	2	0	1	0	0	1	0	1	2	1	0	0	2	2	1	2	1	1	1	1	0

			P	P	P	P	Q	Q	R	R	R	R	R	R	S	S	S	S	S	S	T	T	T	T	V	V	V	V	W	Y	Y	Stop	Stop	x
Pos	Ref AA	Ref codon	CCA	CCC	CCG	CCT	CAA	CAG	CGA	CGC	CGG	CGT	AGA	AGG	TCA	TCC	TCG	TCT	AGC	AGT	ACA	ACC	ACG	ACT	GTA	GTC	GTG	GTT	TGG	TAC	TAT	TAA	TAG	TGA
106	N	AAC	0	1	0	0	1	1	0	1	0	0	1	1	0	1	0	0	2	1	1	2	1	1	0	1	0	0	0	2	1	1	1	0
107	Y	TAC	0	1	0	0	1	1	0	1	0	0	0	0	1	2	1	1	1	0	0	1	0	0	0	1	0	0	1	3	2	2	2	1
108	K	AAG	0	0	1	0	1	2	0	0	1	0	1	2	0	0	1	0	1	1	1	1	2	1	0	0	1	0	1	1	1	1	2	0
109	T	ACC	1	2	1	1	0	0	0	1	0	0	1	1	1	2	1	1	2	1	2	3	2	2	0	1	0	0	0	1	0	0	0	0
112	E	GAG	0	0	1	0	1	2	0	0	1	0	0	1	0	0	1	0	0	0	0	0	1	0	1	1	2	1	1	1	1	1	2	0
113	V	GTG	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	0	0	0	1	0	2	2	3	2	1	0	0	0	1	0
114	K	AAG	0	0	1	0	1	2	0	0	1	0	1	2	0	0	1	0	1	1	1	1	2	1	0	0	1	0	1	1	1	1	2	0
115	F	TTC	0	1	0	0	0	0	0	1	0	0	0	0	1	2	1	1	1	0	0	1	0	0	1	2	1	1	1	2	1	1	1	1

4.2.13 Synonymous mutations can affect GFP fluorescence differently

The Venn diagram in **Figure 4.13C** shows that the majority of amino acid substitutions occur in multiple bins of VarSil, albeit the frequencies are different (see **Table 4-5**). It was therefore tested whether synonymous codons could have different effects on GFP fluorescence, which would be obscured by translating nucleotide variants into amino acids (see **Table 4-7**). All possible nucleotide variances apart from T343G and A325T were present in the VarStop Negative bin and in at least one of the VarSil bins. The ratios between nucleotide variances resulting in the same amino acid change were compared between bins (see **Table 4-7**).

The ratios between other codons encoding the same amino acid did not differ between different bins, demonstrating that the occurrences of amino acids across the different bins are not the result of different mutations on the nucleotide level and are therefore likely not due to biological differences.

4.2.14 Validation of mutations using *in silico* modelling

Because the β-barrel region constitutes the structural scaffold of GFP, I hypothesised that loss of fluorescence arising from missense mutations in this region should occur through loss of structural stability. In order to test this hypothesis, I performed *in silico* modelling to predict the effect of each variant on protein stability. For this, amino acid substitutions were modelled onto the resolved crystal structure as was described in the Material & Methods chapter. For each possible single amino acid substitution, the $\Delta\Delta G$ was calculated, providing a measure for the change in free energy relative to the wildtype sequence (see **Table 4-8**). Of the 152 possible amino acid substitutions, 67 have a neutral or positive effect on the thermodynamic stability of the protein (arbitrarily defined as $\Delta\Delta G < 0.8$ kcal/mol), whilst 65 have a severely destabilising effect ($\Delta\Delta G > 1.6$ kcal/mol). The remaining 20 amino acid substitutions were intermediate in their effect on protein stability, leading to a partially destabilised protein.

Table 4-7 Nucleotide substitution read counts in HR-derived reads with a single mutation. Nucleotide substitutions that encode the same amino acid are highlighted in blue (two synonymous codons) or yellow (three synonymous codons). All possible nucleotide substitutions were found in both the VarStop Negative and either of the VarSil samples. T109T mutations are found at significantly different proportions between the three bins, indicating that A and G residues affect fluorescence. Other amino acid substitutions occurring in all three samples do not show differences in read frequencies for the different nucleotide substitutions, suggesting that the simultaneous occurrences of amino acids across different bins are likely due to overlapping bins or sequencing errors. Substitutions were ordered alphabetically.

NT mutation	# synonymous codons	Substitution	Mutation class	StopNeg	SilNeg	SilInt	SilPos
A316G	single	N106D	missense	93	0	4	46
A316C	single	N106H	missense	106	1	0	4
A316T	single	N106Y	missense	77	32	6	35
A317T	single	N106I	missense	73	2	1	34
A317G	single	N106S	missense	155	0	13	57
A317C	single	N106T	missense	36	31	6	36
C318A	two	N106K	missense	181	79	8	0
C318G	two	N106K	missense	161	1	1	0
C318T	single	N106N	silent	113	2	129	1
T319G	single	Y107D	missense	180	6	346	2
T319C	single	Y107H	missense	10	16	308	268
T319A	single	Y107N	missense	41	2	787	8
A320G	single	Y107C	missense	52	116	25	305
A320T	single	Y107F	missense	107	72	531	15
A320C	single	Y107S	missense	110	90	216	150
C321A	two	Y107*	nonsense	125	66	30	212
C321G	two	Y107*	nonsense	194	29	2	127
C321T	single	Y107Y	silent	156	21	1179	10
A322T	single	K108*	nonsense	288	38	1889	15
A322G	single	K108E	missense	105	88	30	300
A322C	single	K108Q	missense	183	212	98	4
A323T	single	K108M	missense	59	7295	2070	1587
A323G	single	K108R	missense	64	2	2	84
A323C	single	K108T	missense	246	6	692	72
G324A	single	K108K	silent	124	4	108	2
G324C	two	K108N	missense	133	0	7	31
G324T	two	K108N	missense	58	2	0	1
A325G	single	T109A	missense	89	1	4	0
A325C	single	T109P	missense	16	24	4	0
A325T	two	T109S	missense	114	0	0	0

C326T	single	T109I	missense	94	15	2	2
C326A	single	T109N	missense	160	48	61	90
C326G	two	T109S	missense	160	0	269	11
C327A	three	T109T	silent	38	202	240	60
C327G	three	T109T	silent	72	103	198	195
C327T	three	T109T	silent	274	92	20	288
A335C	single	E112A	missense	77	237	686	40
A335G	single	E112G	missense	62	92	1520	16
A335T	single	E112V	missense	33	24	17	5
G336C	two	E112D	missense	36	132	4	116
G336T	two	E112D	missense	95	0	440	194
G336A	single	E112E	silent	25	69	918	0
G337C	two	E112L	missense	86	2006	380	17
G337T	two	E112L	missense	106	163	7	0
G337A	single	E112M	missense	108	114	559	24
T338C	single	V113A	missense	12	63	721	25
T338A	single	V113E	missense	156	22	66	5
T338G	single	V113G	missense	139	1	1304	491
G339A	three	V113V	silent	20	6065	1793	1029
G339C	three	V113V	silent	119	32	1	65
G339T	three	V113V	silent	100	51	1	18
A340T	single	V114*	nonsense	406	1	19	21
A340G	single	V114E	missense	138	246	29	0
A340C	single	V114Q	missense	104	64	66	48
A341T	single	K114M	missense	233	35	64	388
A341G	single	K114R	missense	188	34	95	273
A341C	single	K114T	missense	127	5	43	255
G342A	single	K114K	silent	11	11	18	2
G342C	two	K114N	missense	120	26	138	108
G342T	two	K114N	missense	148	62	129	1
T343A	single	F115I	missense	69	8	3	103
T343C	three	F115L	missense	118	0	14	35
T343G	single	F115V	missense	87	0	0	0
T344G	single	F115C	missense	188	43	8	42
T344C	single	F115S	missense	122	1	3	8
T344A	single	F115Y	missense	96	1	2	20
C345T	single	F115F	missense	230	42	6	45
C345A	three	F115L	missense	45	5	7	0
C345G	three	F115L	missense	141	1	10	53

4.2.15 Multiple amino acid substitutions per read obscure the effect on GFP fluorescence

As mentioned in section 4.2.10, the VarSil negative bin harboured the highest proportion of reads with multiple mutations. To assess whether reads with multiple mutations could be used to assess the effects of individual amino acid substitutions on GFP stability, the distribution of change in thermodynamic free energy ($\Delta\Delta G$) for amino acid substitutions per bin was plotted using the independent values that were predicted in the previous section (see **Figure 4.15**), as *in silico* predictions did not allow for multiple mutations. Larger numbers of amino acid substitutions occurring on the same read resulted in smaller differences in $\Delta\Delta G$ values between the different VarSil bins. This signifies that reads with a single amino acid substitution are most informative on interrogating the effects of these mutations on GFP stability.

Nonsense mutations were excluded from this analysis as their free energy changes could not be assessed *in silico*. Mutations found in the VarStop Negative class show a wide range of $\Delta\Delta G$ values comparable to the distribution found in the ssODN repair template, indicating that both the genomic integration and depth of our screening are sufficient to detect all introduced amino acid substitutions.

The VarSil samples in the violin plot reveal that the GFP negative bin harbours mutations with high $\Delta\Delta G$, indicating a negative effect on GFP stability (see **Figure 4.13D**). Mutations found in the GFP intermediate bin show a wide distribution of predicted effects on stability, with a mean $\Delta\Delta G$ value that is nonetheless intermediate between GFP negative and positive. We reasoned that the arbitrary boundaries of the gates for sorting (defined in section 4.2.5) could allow for the collection of cells harbouring null, hypomorphic and neutral mutations in the intermediate bin explaining the wide range of $\Delta\Delta G$ values found in this bin. Whilst overlapping, the distribution of free energy predictions per bin follow the expected pattern of distribution and therewith validate our ability to classify mutations by functional effect.

4.2.16 Mutational effect score provides a single value per amino acid substitution

Despite the different distributions of $\Delta\Delta G$ values between bins, the majority of the amino acid substitutions were observed in more than a single bin. Therefore, this section will assign

Table 4-8 Predicted effect of amino acid substitutions on thermal stability of GFP protein FoldX was used to calculate the change in thermal protein stability (ΔΔG) due to each of the 152 possible amino acid substitutions at the target residues in GFP. A higher ΔΔG indicates a more severely reduced effect on protein stability. Three classes of amino acid substitutions were classified with thresholds at 0.8 and 1.6 kcal/mol, resulting in 67 mutations having no effect on protein stability, 20 mutations causing a reduced protein stability and 65 mutations causing a severely reduced protein stability.

Position	Calculated ΔΔG	Substituted AA: Alanine (A)	Cysteine (C)	Aspartic acid (D)	Glutamic acid (E)	Phenylalanine (F)	Glycine (G)	Histidine (H)	Isoleucine (I)	Lysine (K)	Leucine (L)	Methionine (M)	Asparagine (N)	Proline (P)	Glutamine (Q)	Arginine (R)	Serine (S)	Threonine (T)	Valine (V)	Tryptophan (W)	Tyrosine (Y)
106	N	-0.93	-1.06	1.32	-0.83	-1.53	0.09	0.33	-1.73	-2.07	-2.83	-2.86		0.95	-1.43	-1.53	-0.31	-1.39	-1.43	1.46	-0.93
107	Y	5.28	4.77	6.72	-0.84	0.62	6.83	4.48	2.71	4.60	1.86	1.48	5.61	4.59	4.25	3.62	6.46	5.37	3.54	7.66	
108	K	1.18	0.93	1.11	0.89	-0.09	1.83	0.44	-0.47		-0.28	-0.23	0.92	3.62	0.57	0.26	1.23	0.49	0.06	-0.16	0.12
109	T	0.93	-0.24	3.25	-0.04	10.19	1.96	7.71	-2.74	2.00	-1.49	-2.42	1.63	2.75	-0.38	3.76	0.82		-2.03	13.73	16.90
112	E	0.61	0.24	1.18		-1.67	1.82	-0.72	-1.11	-0.91	-1.27	-1.18	0.36	0.49	-0.10	-0.60	0.97	0.19	-0.46	-1.47	-1.57
113	V	2.96	2.12	4.40	4.20	6.60	4.62	3.32	-0.63	3.41	-0.21	0.54	3.36	4.70	3.92	5.77	4.09	2.29		11.86	5.41
114	K	4.77	0.38	2.25	1.33	-1.55	0.88	0.14	2.59		-0.35	-1.03	0.92	2.89	0.13	-1.42	0.42	0.24	2.75	0.81	-1.39
115	F	3.54	3.47	5.70	4.36		4.46	3.02	2.62	2.95	1.26	1.02	3.80	2.23	3.55	2.75	4.16	3.84	3.45	0.60	-0.42

		Count
No effect on structural stability	ΔΔG < 0.8 kcal/mol (cyan)	67
Reduced structural stability	ΔΔG > 0.8 & < 1.6 kcal/mol (magenta)	20
Severely reduced structural stability	ΔΔG > 1.6 kcal/mol (red)	65
	Total	152

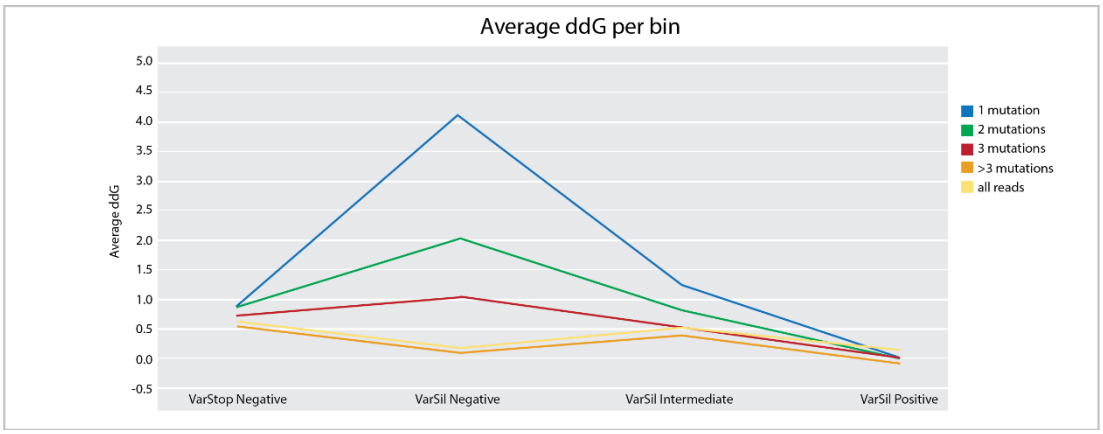


Figure 4.15 Average ΔΔG dependent on the number of mutations per read HDR reads from each bin were sorted based on the number of amino acid substitutions in the variable region of the read, after which the average ΔΔG per bin was calculated. Reads with a single substitution show the largest difference in ΔΔG between VarSil negative, intermediate and positive, indicating that the presence of multiple amino acid substitutions in a single read obscures the effects of hypomorphic and null mutations.

several scores on both the mutational effect and confidence to each amino acid substitution, such that the substitutions with a highly-confident effect remain.

Three amino acid substitutions and one silent mutation (N106D, ΔΔG = 1.32; F115V, ΔΔG = 3.45; V113L, ΔΔG = -0.21; V113V, ΔΔG = -0.01) were amongst the most abundantly found mutations, ubiquitously present in all 3 bins (see **Table 4-5**). These mutations did not occur at a higher frequency in VarStop, raising the possibility that this could be a VarSil-specific issue down to the synthesis of these ssODN. However, as these mutations were also not enriched in the ssODN pool, these artefacts were most likely introduced during library synthesis and were thus excluded from further analysis.

A single functional score was assigned to each amino acid substitution based on the relative enrichment of these mutations across each bin relative to the other bins. A previously published equation (Kosuri *et al.*, 2013) was found to calculate a single score per mutation allowing for this, while also accounting for the differences in abundance between the different codons in the overall pool.

To estimate mutational effect scores, we first calculated, for each codon *i* and each bin *j*, the normalised fractional contribution of each bin per codon a_{ij} using the formula

$$a_{ij} = {f_j \cdot c_{ij}}\Big/{\sum_j f_j \cdot c_{ij}} \quad (4.1)$$

, where f_i is the fraction of cells harboured in each bin based on the VarSil FACS plot and c_{ij} is the number of amino acid *i* in each bin *j*. For a given substitution, $\sum_j a_{ij} = 1$ (i.e. each codon has one a_{ij} per bin, and these six values together add up to 1).

The final mutational effect score, E_i, was calculated as:

$$\mathrm{E_i} = \exp\left[\sum_j a_{ij} \log(m_j)\right] \quad (4.2)$$

, where m_j is the median fluorescence measurement of bin j (45.9, 1,338 and 23,800 for negative, intermediate and positive, respectively).

After calculating the effect scores for the amino acid substitutions, we next sought to eliminate mutations that were expressed at low levels or that were not differentially expressed in different bins. To minimise the noise arising from this effect, a threshold was set by subtracting the read error (0.0064%, as published in Schirmer *et al.*, 2015, and discussed in the previous chapter) from the proportions each amino acid *i* made out of the whole HR pool in each bin, which left a total of 26 amino acid substitutions and an additional four synonymous mutations that were detected above the level of background noise and could be functionally classified (see **Table 4-9**).

To remove amino acid substitutions with a low confidence mutational effect score, we sought to ensure mutations were not evenly distributed across all the bins and hence assessed the enrichment of amino acids across the respective bins. Enrichment of amino acids in any of the bins was represented by an F value following from an F distribution, with corresponding p value and demonstrated that 15 amino acid substitutions were significantly enriched in one of the bins (see **Table 4-9**). This demonstrated that 15 amino acid substitutions were

significantly enriched in one of the bins. The synonymous mutations are all highly significant with a low mutational effect score, serving as internal controls for our assay.

For the significantly enriched amino acid substitutions, the calculated mutational effect scores correlated with the predicted $\Delta\Delta$G values (R^2=0.60, see Table 4.9). As $\Delta\Delta$G values only gages the effect of an amino acid substitution on protein stability, the scores were additionally benchmarked to another experimental dataset, the most comprehensive study that previously quantified the effect of individual amino acid substitutions on GFP fluorescence (Sarkisyan *et al.*, 2016). This 'Kondrashov score' for single mutations correlated similarly with the $\Delta\Delta$G values (R^2=0.58, see Table 4.9) but relatively poorly with our mutational effect score (R^2=0.43, see Table 4.9). The lower correlation between the data from this study and the Kondrashov score is likely due to the different nature of the proteins used, as we use the optimised eGFP variant in comparison to the native form from *Aequorea victoria* (avGFP). However, as this was the most comprehensive study systematically studying the effects of amino acid substitutions on GFP, this demonstrates that our deep mutational effect score is equally sophisticated at predicting mutational effects of amino acid substitutions as Sarkisyan *et al.*, whilst acknowledging that their study assessed a much wider set of both single and double mutations.

4.2.17 Use of long double-stranded DNA donor templates result in high HDR frequencies but impairs phenotypic selection

As ssODNs and double-stranded donors are both suggested to be suitable as a repair template for the introduction of specific variants through DSB repair (Radecke *et al.*, 2006, 2010; Chen *et al.*, 2011; Soldner *et al.*, 2011), we sought to assess whether double-stranded repair template (dsRT) donors would either improve HDR efficiencies or reduce the positional bias during the introduction of nucleotide variants through HDR as was demonstrated in section 4.2.11. Therefore, ssODNs donor libraries with variable positions were PCR-amplified into 820-bp products using four ssODNs, each 200-bp in length with homology to the *Gfp* gene, as primers (see **Figure 4.16**). As starting templates, VarStop and VarSil repair templates with either 24 variable positions at 94:2:2:2 consensus versus non-consensus ratios (VarStop12 and VarSil12), or 94 variable positions (i.e. all positions except for the HDR core) with 97:1:1:1 ratios (VarStop100 and VarSil100) were used. As there is a 20-bp overlap

between each primer and the template, there is expected to be some loss of diversity at the distal ends of the repair templates with full diversity.

In section 4.2.6 it was demonstrated that HDR repair from the VarSil12 ssODN template leads to a spectrum of GFP intensities, including intermediate GFP fluorescence, that was not observed in the experiment using VarStop12. In the samples using any of the dsRTs, analysis by flow cytometry at 7-days post-transfection showed that the majority of the targeted cells had completely lost GFP fluorescence, with a complete absence of a GFP intermediate population (see **Figure 4.17**). This indicates that whilst editing efficiencies are high, phenotypic selection for specific variants is completely lost when compared to the use of ssODNs in section 4.2.6. With the lack of an intermediate and positive population in the fluorescence profile, these results show that under the conditions tested, long double stranded repair templates are not suitable for the integration and functional assessment of single nucleotide variants in *Gfp*. Therefore, these experiments were not followed through.

Table 4-9 Mutational effect scores for amino acid substitutions occurring above background level Mutational effect scores were calculated for the 30 mutations that occurred above background noise frequency (0.0064). 15 of these are significantly enriched in at least one of the three populations as depicted by the F score and p value. Highly stabilising mutations (N106K, K114R, E112V), partially destabilising (K108M, K114T) and highly destabilising (F115L, K114N, F115I) mutations typically correspond with low, intermediate and high ΔΔG values, but do not correlate well with previously published work (Kondrashov score, whereby a lower score indicates a more destabilising mutation).

Substitution	Mutational Effect	ΔΔG	Kondrashov	F value	p value	
E112A	**1383.88**	0.61	3.63	8.22	0.019	*
E112D	**261.65**	1.18	-	1.60	0.278	
E112G	**83.71**	1.82	3.54	4.21	0.072	
E112V	**22823.83**	-0.46	3.75	24.31	0.001	**
F115C	**260.10**	3.47	3.62	2.45	0.167	
F115F	**18496.97**	-0.03	-	37.84	0.000	***
F115I	**168.51**	2.62	3.57	13.62	0.006	**
F115L	**163.84**	1.26	3.67	14.60	0.005	**
F115S	**45.99**	4.16	3.14	6.07	0.036	*
F115Y	**2788.10**	-0.42	3.66	3.66	0.091	
K108K	**22655.34**	0.00		682.93	0.00	***
K108M	**9328.18**	-0.23	3.65	7.48	0.023	*
K108Q	**1834.74**	0.57	3.73	9.96	0.012	*
K114*	**46.41**	-	-	9.17	0.015	*
K114E	**572.08**	1.33	3.49	3.11	0.118	
K114M	**1415.86**	-1.03	3.63	1.04	0.408	
K114N	**497.48**	0.92	3.69	14.53	0.005	**
K114Q	**1193.52**	0.13	3.63	6.46	0.032	*
K114R	**23604.63**	-1.42	3.72	30.54	0.001	**
K114T	**4602.96**	0.24	3.71	0.78	0.502	
N106K	**21333.91**	-2.07	3.70	106.39	0.000	***
T109A	**1338.00**	0.93	3.61	26.05	0.001	**
T109I	**46.12**	-2.74	3.42	2.52	0.161	
T109P	**46.20**	2.75	-	2.03	0.212	
V113A	**87.57**	2.96	3.56	2.47	0.165	
V113E	**146.42**	4.20	1.34	0.45	0.660	
V113G	**46.79**	4.62	-	1.14	0.381	
Y107C	**2657.28**	4.77	3.40	2.82	0.137	
Y107F	**17596.27**	0.62	3.79	12.96	0.007	**
Y107Y	**22034.81**	-0.01		1067.48	0.00	***

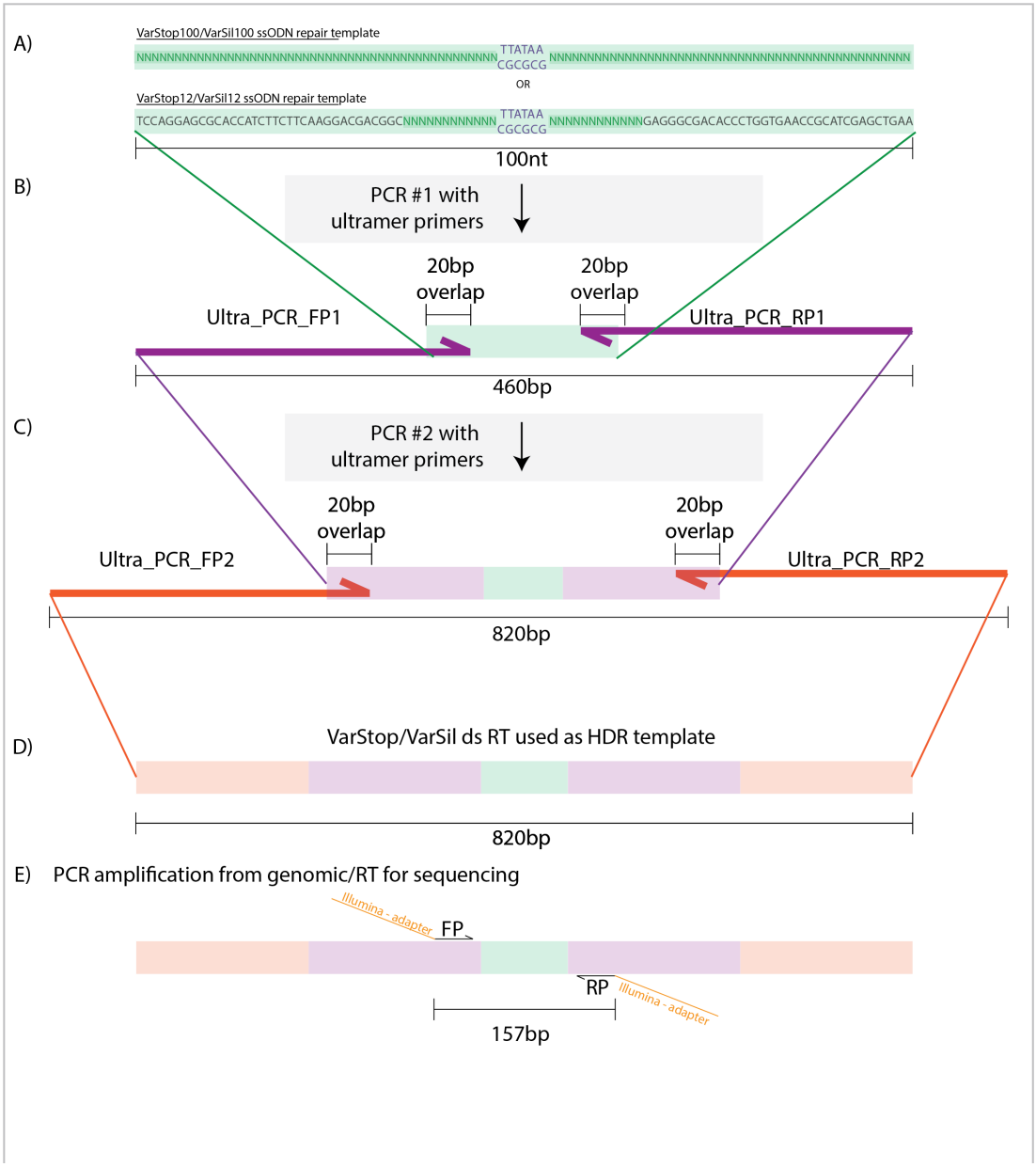


Figure 4.16 PCR schematic of synthesis and library preparation of double-stranded RT pool **(A)** Both VarStop and VarSil ssODN molecules are used as templates for the construction for the double-stranded (ds) RT pool. In addition to the repair templates with variable nucleotides at 2 x 12-nt (VarSil12), other pools or ssODNs have variable positions across the entire 100-nt ssODN except for the 6-nt HDR-core. All ssODN pools are prepared in 4 different reactions, hence leading to 4 different pools of ds RTs. **(B)** In an initial reaction, ssODNs are PCR-amplified with 200-nt Ultramer primers, which each have a 20-nt overlap with the ssODN, hence resulting in an 460-bp ds product. **(C)** In an second reaction, 460-bp amplicons are again amplified with 200-nt Ultramer primers, this time resulting in a **(D)** final 820-bp ds product with complete homology (with the exception for the 6-nt HDR-core and substitutions in the variable positions) to the GFP locus. This product was gel-purified and used as an HDR-template. **(E)** For sequencing library preparation of both the ds RT pool and gDNA from transfected cells, the same primers were used as described in figure 4.2, resulting in a 157-bp amplicon with Illumina-adapters.

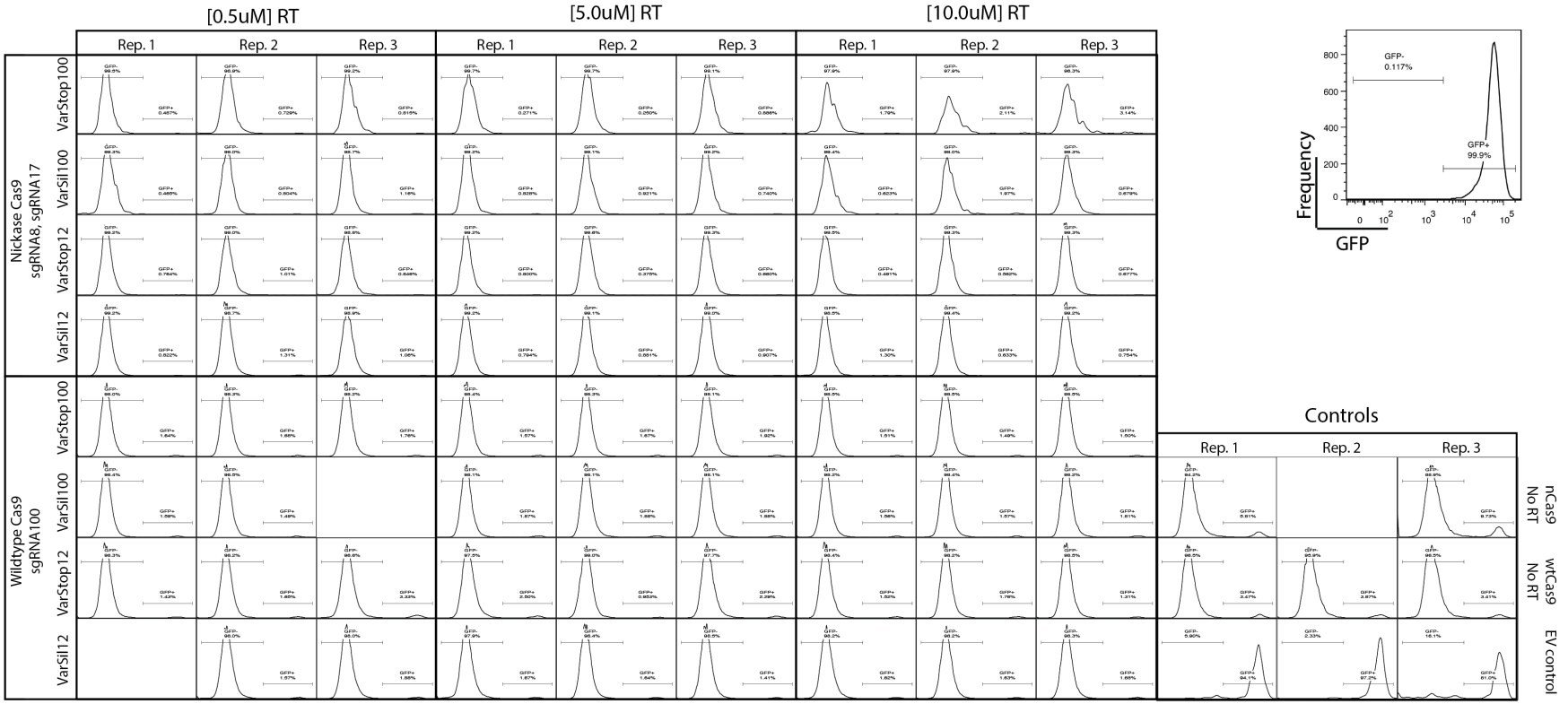


Figure 4.17 Effect of double-stranded repair templates on GFP fluorescence 7-days post-transfection with double-stranded long DNA repair templates, cells were analysed for GFP fluorescence by flow cytometry. Regardless of the use of wtCas9 or Cas9n or the concentration of repair template DNA, the majority of the cells lost GFP expression.

4.3 Discussion

In this chapter, I have shown that a genomic library containing all possible single nucleotide variants in a predefined genomic region can be created using CRISPR-Cas9 and multiplex HDR. Multiple conditions for introducing these genomic variants were tested and it was demonstrated that the use of paired Cas9n and ssODN templates as a donor is the most suitable system for this application. The effect of these nucleotide changes on GFP fluorescence can be determined by translating them into amino acid substitutions and subsequently calculating a single weighted score from the proportion of HDR reads encoding that substitution in each bin. Ultimately, these scores are benchmarked by calculating the predicted effect on protein stability using the resolved protein structure of GFP *in silico*. In the presence of an appropriate functional classification system, this pipeline can be used for the interrogation of single nucleotide and amino acid variants in endogenous genes.

4.3.1 Deep-mutational scanning on the nucleotide level can be used to interrogate gene function

As an example, the direct fusion of a fluorescent gene to an endogenous open reading frame which would be diversified could be used to infer something about transcription or stability of the encoded protein. Whilst supporting data is not shown in this thesis, we made a fusion of an mCherry fluorophore to endogenous *Nanog* and yielded consistent high (± 30 %) HDR rates. However, due to both the complexity of the transcriptional regulation of *Nanog* and issues with the screening assay, it was decided to not continue with this project. Yet, this provides one example for the application of the pipeline developed in this chapter.

4.3.2 Effects of nucleotide variants can be explained by their translation into amino acid substitutions

Whilst on the outset the aim was to test the effect of nucleotide variants, the findings in this chapter demonstrate that the effects of these variants can all be explained by their effect on the codon sequence and do not underlie codon-specific differences (with the exception of T109T, see section 4.2.13). All 51 amino acid substitutions that can be achieved with a single nucleotide change and 43/87 amino acid that can be achieved by two nucleotide substitutions were detected above noise levels in the VarSil samples (section 0). This shows that the range of amino acid substitutions that can be assessed is largely limited by the

number of nucleotide changes required for that amino acid change. To study the effect of all possible amino acid substitutions using CRISPR-Cas9 and multiplex HDR, it would therefore be more suitable to design repair templates with codons for all amino acid substitutions, although this was not feasible using the oligonucleotide doping strategy deployed in this chapter.

4.3.3 Depth of analysis of HDR-derived variants is limited by a variety of factors

Another limiting factor in studying the effects of amino acid substitutions using this pipeline is the depth of detecting single changes in HR reads, which is due to several reasons. Firstly, whilst NGS allows for thousands to millions of reads per sample (depending on the number of samples multiplexed on a single lane), the overall rates of HDR in the experiments in this chapter do not exceed 14% (section 4.2.8), meaning that the large majority of the reads are rejected from downstream analysis. Therefore, low HDR frequencies, and the amplification of wildtype and indel containing alleles are the main limiting factor on the extent to which single nucleotide variations can be detected.

HDR efficiencies were increased by limiting the level of contamination with wildtype GFP sequences, which arose from plasmid sequences present in the laboratory environment (section 4.2.7). It is however expected that these levels of contamination are characteristic of GFP, as it is a widely used selective marker and abundantly present on plasmid sequences. We therefore expect that this will not be a major issue in the study of other genes.

Secondly, the design of sgRNAs and repair templates allowed for the re-cleavage of alleles that have already undergone HDR as the donor template does not destroy either of the PAM sequences. The elevated HDR rates in VarStop (14 %) compared to VarSil (4 – 8 %) could also be explained by the respective HDR core sequence destroying the sgRNA binding site. To increase HDR efficiencies without causing a bias in nucleotide frequencies, future HDR donor sequences should therefore ideally destroy the PAM sequence with a constant mutation in the HDR core or the constant region outwith the variable regions.

A third factor limiting the sequencing depth is the low number of sequencing reads harbouring only a single mutation. As was observed in both the ssODN donor templates and

genomic samples, nearly a third of the HDR reads did not harbour any mutations whilst nearly half (47%) contained a single nucleotide variant (section 4.2.10). Whilst in some cases multiple nucleotide substitutions occurred in a single codon and thus in a single amino acid change, the majority of these reads also contained multiple amino acid changes. As I demonstrated that the $\Delta\Delta G$ values correlating with the amino acid substitutions are best interrogated without the presence of additional amino acid substitutions (section 4.2.15), this means that the majority of the HDR sequences could not be used in downstream analysis. Non-random parallel synthesis of donor templates using technology available through companies such as Twist and Agilent, could perhaps further increase the number of reads with a single nucleotide or amino acid variant and thus the depth of analysis.

4.3.4 Mutational effect score provides a weighted measure for the phenotypic effect

Sequence reads encoding 94 out of 152 possible amino acid substitutions were discovered in the pool of amplicons with a VarSil HR core (section 0). Many amino acid substitutions were detected in more than a single bin and a large overlap was detected in the distribution of $\Delta\Delta G$ values, which are likely due to inherent variability in fluorescence levels from mutant proteins, thus leading to a decreased separability between the bins. This overlap was resolved by using a weighted score encapturing the frequencies of amino acids and median fluorescence in each bin (section 4.2.16). The mutational score was eventually calculated for the 30 amino acid substitutions that occurred above noise level and were enriched in one of the bins, which demonstrated that these scores highly correlate with the predicted effect on protein stability.

Whilst the use of three bins allows for the detection of mutational effects on a large scale, it does limit the sensitivity of our screen in determining the effect of amino acid substitutions. For example, both neutral mutations (e.g. Y107Y and K108K) and stability-enhancing mutations, i.e. with a predicted negative $\Delta\Delta G$ (e.g. N106K and K114R), have similar mutational effect scores of approximately 22 as they are most abundantly present in the GFP positive bin (section 4.2.16). Similarly, amino acid substitutions with different grades of predicted destabilising effects (e.g. F115L ($\Delta\Delta G$ = 1.26) and F115I ($\Delta\Delta G$=2.62)) have a similar mutational effect score of 91. The increased resolution of multiple gates would allow for the

detection of these minor differences and would hence allow for a better assessment of mutational effects.

The change in thermodynamic stability ($\Delta\Delta G$) is of course an *in silico* modelling and not benchmarked and thus might itself not be as sensitive and accurate for all values. In application of this assay in studying endogenous genes, benchmarking such a score to monoclonal cell lines harbouring specific mutations would hence be an appropriate control.

4.3.5 Mutational effect scores indicate a positional bias

Most of the mutational effect scores can be explained by the effect of the amino acid substitution on protein stability as is shown by the correspondence of mutational effect scores with the predicted $\Delta\Delta G$ values (section 4.2.16). Y107C and T109I are the largest outliers and were predicted to not have an effect on protein stability. As inward-facing residues in the β-sheet are known to interact with the fluorophore (Stepanenko *et al.*, 2013), mutations on residues 107, 109 and 113 would be expected to have a negative effect on fluorescence whilst they would not directly destabilise the GFP protein. Mutational effect scores were proportional to the predicted $\Delta\Delta G$ of the 6 amino acid substitutions at these residues found in our screen, whilst the effect on Y107C and T109I could be explained by the difference in chemical properties destroying the interactions of these residues with the fluorophore.

4.3.6 Efficiency of HDR is dependent on sgRNA position

The difference in HDR-efficiencies using paired Cas9 nickases (Cas9n) and wildtype Cas9 showed that the increase in HDR efficiency with the use of Cas9n over wtCas9 largely depends on the sgRNA that is used (section 4.2.11). By assessing the fluorescence profiles and difference in intermediate populations between the VarSil and VarStop experiments, it was shown that wtCas9 used with sgRNA100 had similar levels of HDR to the use of paired nickases, whilst sgRNA8 and sgRNA17 showed slightly lower increases in intermediate populations. Lower nucleotide diversity in the variable region was observed between the experiments with different sgRNAs, which can also be attributed to these variations in HDR frequencies, as samples with lower diversity were those with lower levels of HDR. Previous studies have shown that following Cas9 cleavage, binding of ssODNs to the strand opposite

to the PAM (i.e. the strand complementary to the sgRNA) largely improve HDR efficiencies (Christopher D. Richardson *et al.*, 2016), which in our design is the case for sgRNA17 and sgRNA100 but not for sgRNA8. The same study showed that asymmetric donor DNA with a longer PAM-proximal than PAM-distal homology arm (optimally 91-nt against 36-nt) even further enhances HDR efficiencies, which is closer resembled with sgRNA100 (52-nt and 48-nt) than with sgRNA17 (18-nt and 82-nt). This provides an explanation as to why the highest HDR efficiencies are achieved with sgRNA100 and demonstrates that careful design of Cas9 editing experiments can largely enhance targeting efficiencies. These results hence demonstrate that complementary strandedness of sgRNA and ssODN donor provides the best targeting efficiencies with wtCas9 but that this does not exceed the efficiencies yielded with the double nickases.

4.3.7 Use of long double-stranded DNA donor templates impairs phenotypic selection

The addition of a dsRT consistently lead to a nearly complete loss of both GFP fluorescence compared to the no-template controls (section 4.2.17). This suggests that CRISPR editing efficiencies are highly increased due to the presence of a double-stranded molecule, but that these variances are not incorporated in frame of the initial GFP locus. One possibility is that the dsRTs were incorporated in tandem, thereby impeding correct expression or folding of GFP. Therefore, future experiments with the use of double-stranded donors should ensure a suitable selection scheme to avoid amplification from random integrands, ideally through a PCR selection outwith the region covered by the homology arms of the donor whilst also selectively amplifying loci with HDR-specific mutations.

An alternative explanation for the loss in GFP is that the addition of linear DNA (either single- or double-stranded) skews the repair of CRISPR-Cas9-induced breaks to NHEJ, as is suggested in previously published work (C. D. Richardson *et al.*, 2016). As short ssODNs are suggested to be efficiently incorporated through the FA pathway (Richardson *et al.*, 2018), double-stranded donors would thus be more efficient when introduced from a plasmid.

4.3.8 Concluding remarks

The experimental pipeline optimised in this chapter combined with the bioinformatics pipeline developed in the previous chapter provide a suitable method for the systematic

interrogation of large numbers of nucleotide variants in a genomic copy of GFP. Fluorescence is a convenient measure to detect different gradients of activity as a result of genetic variants. To apply this pipeline to an endogenous gene, a selectable marker would have to be coupled to that gene to assess the phenotypic consequences of the introduced genetic variant. Design of a suitable marker depends on the hypothesis that needs testing; whereas stability would require a direct fusion of a fluorophore to the target gene, signalling activity would require the coupling of a fluorophore to a downstream promoter and fitness could be assessed by the relative survival rate of individual clones. In the following chapter I will test the applicability of this system by performing deep mutational scanning on the signalling activity of *CTNNB1* on the codon level.

5

Deep mutational scanning of amino acid substitutions in β-catenin

5.1. Introduction

The canonical Wnt/β-catenin signalling pathway is one of the primary regulatory mechanisms balancing self-renewal, differentiation and apoptosis in several adult stem cell niches, thereby maintaining tissue homeostasis in the adult organism (Reya and Clevers, 2005).

Wnt proteins are highly conserved throughout evolution and can be found in many phyla across the animal kingdom, ranging from sea anemones (Kusserow *et al.*, 2005) to insects (Nusslein-Volhard and Wieschaus, 1980) and mammals including human and mouse. The *Wnt1* proto-oncogene encodes a 40kDa secreted protein containing many conserved residues (Takada R, Satomi Y, Kurata T, Ueno N, Norioka S, Kondoh H, Takao T, 2006). Whilst signalling of the Wnt-homolog Wg in *Drosophila* is best known for the tissue-wide concentration gradient it forms (Zecca, Basler and Struhl, 1996), Wnt in other organisms predominantly signals between cells that are in each other's vicinity (Sato *et al.*, 2011).

The signal transduced by the canonical Wnt cascade is largely regulated through the destruction complex that regulates proteolytic turnover of cytoplasmic β-catenin (see **Figure 5.1A**). β-catenin (encoded by the *CTNNB1* gene) is a dual function protein, being involved in both cell-cell adhesion, linking intercellular cadherin proteins in simple epithelia, and as a signal-transducer in the Wnt signalling pathway (Peifer *et al.*, 1991; Noordermeer *et al.*, 1994; Peifer, Berg and Reynolds, 1994). It is expressed across almost all tissues in vertebrates, with elevated expression found in highly proliferating cells including a wide variety of stem cells.

When Wnt receptors are not engaged, Axin, WTX and adenomatous polyposis coli (APC) form a scaffolding complex which can dock cytoplasmic β-catenin (Reya and Clevers, 2005; Major *et al.*, 2007). This complex now allows for casein kinase 1 (CKI) and glycogen synthase kinase 3β (GSK-3β) to phosphorylate β-catenin on a set of highly conserved serine and threonine residues in the amino terminus of the protein (Cohen, 1986; Tuazon, 1991; Fish *et al.*, 1995; Aberle *et al.*, 1997; Orford *et al.*, 1997; Winston *et al.*, 1999). CKI phosphorylates serines on several components of the Wnt signalling pathway including β-catenin at Ser45 (Davidson *et al.*, 2005).

Subsequently, GSK-3β binds this phosphorylated residue with a positively charged pocket neighbouring its active site in order to sequentially phosphorylate Thr41, Ser37 and Ser33. The active site of GSK-3β binds the terminal phosphate of ATP and transfers it to each of the target locations on the substrate (Patel and Woodgett, 2017). A fully phosphorylated 'degron' motif will then be recognised by β-transducin repeat containing E3 ubiquitin protein ligase (β-TrCP), an E3 ubiquitin ligase, which subsequently targets β-catenin for proteosomal degradation in the *26S* proteasome (Aberle *et al.*, 1997). In the absence of Wnt signalling, this process is constitutively active resulting in a high turnover rate of β-catenin protein, thereby preventing accumulation and nuclear translocation of cytoplasmic β-catenin.

Wnt proteins bind to a heterodimeric receptor complex consisting of Frizzled and LRP5/6 proteins that contain a large extracellular cysteine-rich domain to which Wnt can bind (Bhanot *et al.*, 1996; Clevers and Nusse, 2012) (see **Figure 5.1A**). This binding allows the receptor complex to interact with cytoplasmic proteins including Dishevelled (Chen, 2003). This protein in turn passes on the signal through a downstream intracellular signalling cascade, inhibiting the kinase activity of CKI and GSK-3β and hence preventing the proteolytic turnover of β-catenin (Hatsell *et al.*, 2003). Cytoplasmic levels of β-catenin now accumulate, eventually resulting in the protein being transported to the nucleus with the help of Rac1 (Wu *et al.*, 2008).

Once in the nucleus, β-catenin binds to DNA-binding proteins of the T-cell factor (TCF) and lymphoid enhancer-binding factors (LEF) families (Behrens *et al.*, 1996; Molenaar *et al.*, 1996). In the absence of Wnt signalling, TCF/LEF proteins are bound by Groucho transcriptional repressors (Cavallo *et al.*, 1998; Roose *et al.*, 1998), but interaction with β-catenin converts TCF/LEF into transcriptional activators, targeting specific downstream genes and affecting subsequent cellular decisions. Over a hundred genes are activated by β-catenin, some of which are oncogenes such as c-MYC (*MYC*), Cyclin D1 (*CCND1*), Survivin (*BIRC5*), Axin2 (*AXIN2*), matrix metalloproteinases (MMPs) and other factors involved in proliferation and stem cell maintenance (Arend *et al.*, 2013). While different stem cell types require distinctive extrinsic signals to maintain their 'stemness', Wnt signalling stands out as it is required in the maintenance of many stem cell types (Reya *et al.*, 2001).

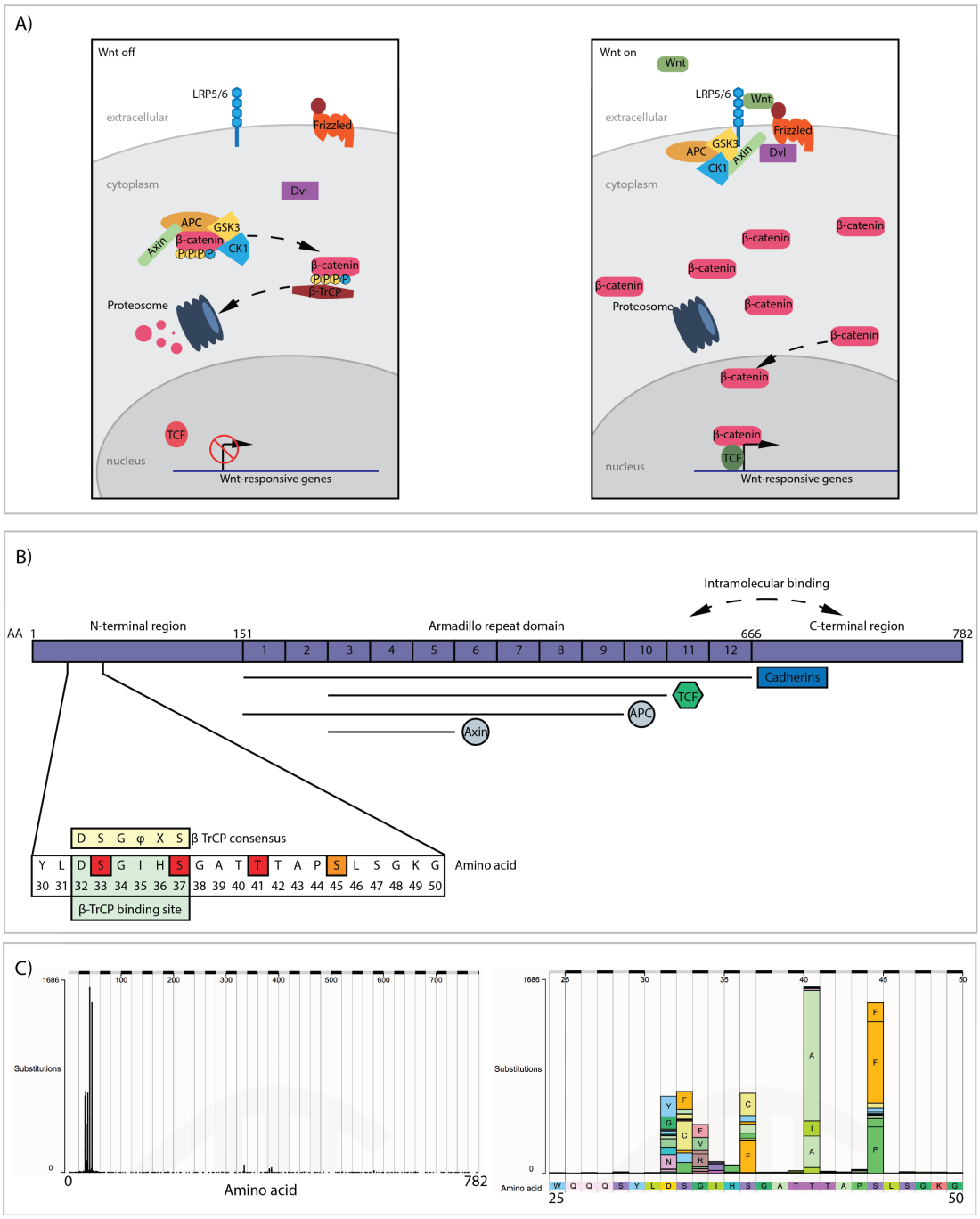


Figure 5.1 The canonical Wnt/β-catenin signalling cascade, simplified (A) In the absence of Wnt, cytoplasmic β-catenin protein is bound by APC and Axin, which allows CKI and GSK-3β to phosphorylate β-catenin. The phosphorylated protein is now recognised by β-TrCP and ubiquitinated subsequently degraded by the proteasome. When Wnt is present, it binds transmembrane proteins LRP5/6 and Frizzled, which now form an intracellular complex which inhibits the activity of kinases CKI and GSK-3β, thereby inhibiting the turnover of β-catenin protein, allowing for it to accumulate in the cytoplasm and translocate to the nucleus. Once in the nucleus, it binds TCF/LEF factors that activate transcription of Wnt-responsive genes. **(B)** The β-catenin protein consists of an N-terminal region, a central Armadillo repeat domain and a C-terminal region. The N-terminal region has four phosphorylation sites that

can be phosphorylated by CKI (orange) or GSK-3β (red). β-TrCP subsequently recognises a 6-nucleotide binding motif including pS33 and pS37 (highlighted in green, ϕ indicating a hydrophobic and X any amino acid) and result in the ubiquitination of β-catenin necessary for proteolytic degradation. The C-terminal region acts as a trans-activator that may interact with the armadillo domain to regulate interactions with other proteins. **(C)** Distribution of amino acid substitutions affecting β-catenin found in cancer. Left: gene-wide overview of substitution frequencies affecting β-catenin shows a clustering of substitutions in the N-terminal region of the protein. Right: enlargement of the N-terminal mutational cluster shows mutations happen most frequently in residues T41, S45 and residues D32-S37. These residues additionally harbour different frequencies of amino acid substitutions. Figure adapted from the Catalogue of Somatic Mutations in Cancer (COSMIC).

5.1.1. β-catenin structure and function

The β-catenin protein consists of 782 amino acids, of which the bulk is structured in twelve characteristic armadillo repeats that together form the rigid armadillo repeat domain (ARM) (Huber, Nelson and Weis, 1997; Loh *et al.*, 2006) (see **Figure 5.1B**). The highly structured ARM region forms a wedge-shaped curvature, of which the inner surface serves as a ligand-binding site for the various interaction partners of β-catenin. These proteins bind to adjacent but not identical residues in the ARM region, allowing partners from the same complex to bind β-catenin simultaneously. This domain has affinity with several binding partners and binding sites for these proteins thus overlap, causing competition and requiring regulation at several levels to maintain delicate equilibria between the different roles of β-catenin.

The N-terminal and far C-terminal domains that flank the ARM domain do not adopt any structure by themselves but do play crucial roles in the regulation of β-catenin function. The C-terminus functions as a transcriptional activator that allows interaction with TCF/LEF (Van de Wetering *et al.*, 1997). The N-terminal domain contains a region that is highly conserved in evolution with four residues that can be phosphorylated. Interaction with APC and Axin in the degradation complex initiate phosphorylation of Ser45 by CKI (Aberle *et al.*, 1997; Orford *et al.*, 1997; Winston *et al.*, 1999). The resulting pSer45 priming site now allows GSK-3β to phosphorylate three residues adjacent to this, each one separated by three other residues in its primary structure (Hagen and Vidal-Puig, 2002). This phosphorylation cascade results in a six amino acid sequence in the N-terminal region that forms a ^{32}DpSGϕXpS^{37} binding motif (ϕ indicating a hydrophobic and X any amino acid) for β-TrCP (see **Figure 5.1B**) necessary for proteolytic degradation (Wu *et al.*, 2003). In addition to degradation, nuclear phosphorylated β-catenin is known to be stable and present at low concentrations, where it binds to

centromeres (Huang, Senga and Hamaguchi, 2007). Although unstructured, the β-TrCP binding motif was determined to require a tight bend of residues 30 to 40 in order for β-TrCP to bind (Megy *et al.*, 2006). The phosphate groups of pSer37 and especially pSer33 make the largest number of contacts with β-TrCP, whilst Asp32 is an invariant residue making the most rigid contacts (Wu *et al.*, 2003). Ile35, of which the hydrophobic nature is conserved in the binding motif, makes Van der Waals contacts with β-TrCP. His36 is the only variable position of the motif as its side chain points outward and therefore has no interactions with β-TrCP. The backbone amide and carbonyl groups of His36 however make a pair of hydrogen bonds with β-TrCP in the absence of Wnt signalling. The essential role of this region in turnover of β-catenin makes it hence unsurprising that residues in this part have been highly conserved throughout evolution.

With β-catenin having these different roles, it raises questions with regards to its regulation: How does β-catenin choose among its different binding partners and can it switch from one complex to another once it is bound? Whereas the role of β-catenin in Wnt signalling has been most extensively described in the literature, at low cytoplasmic concentrations β-catenin interacts with adherens junction proteins such as E-cadherin (Nelson and Nusse, 2004), indicating that affinity for these partners is highest. In adherens junctions, β-catenin connects cadherin adherens receptors in the cell membrane with α-catenins that are bound to the cytoskeleton. Whilst the half-life of cytoplasmic β-catenin is in the order of minutes, the protein in these complexes is very stable, presumably as it is protected from regulation through β-TrCP (Peifer, M., McCrea, P.D., Green, K.J., Wieschaus, E., and Gumbiner, 1992).
When protein levels increase, adherens junction partners are saturated and β-catenin becomes available in the cytoplasm for binding by the destruction complex components APC/axin (Nelson and Nusse, 2004) and destroyed in the absence of Wnt signalling (Harris and Peifer, 2005). As described earlier, β-catenin only becomes available for transcriptional regulation when turnover by the destruction complex is inhibited through Wnt signalling. Together these different affinities provide a principal model for regulating the different functions of β-catenin.

Whilst the described regulatory mechanisms would in principle be sufficient to explain the non-overlapping roles of β-catenin, one study found that β-catenin by default has a similar

affinity to its transcriptional partner TCF/LEF as to cadherins (Gottardi and Gumbiner, 2004). The authors found that the C-terminus of the protein can fold back to interact with the armadillo repeats to yield a closed conformation, thereby selectively preventing cadherins to interact with β-catenin and favour interactions with TCF/LEF (Gottardi and Gumbiner, 2004). This is due to the difference in interaction, requiring all twelve armadillo repeats for cadherin binding as opposed to TCF/LEF which only requires the central eight repeats (Gottardi and Gumbiner, 2004). It is suggested that this conformational change of β-catenin might also be regulated by the Wnt pathway, although this has not been confirmed. In addition, a similar mechanism could be in place to block APC and axin interactions, but this has again not been confirmed.

Another mechanism skewing the function of β-catenin towards interaction with its transcription factor partners is through the actions of BCL9 (Kramps, 2002). Through phosphorylation of Y142 in the ARM domain of β-catenin, BCL9 weakens the interactions of the protein with cadherins, thereby dissociating it from adherens junctions and shifting the balance to transcription complexes (Brembeck *et al.*, 2004).

Hence, whilst differences in binding affinity play a role in regulating the binding of partner proteins to β-catenin and its subsequent function, there are several additional factors that regulate the protein's function post-transcriptionally. Wnt signalling can hence manipulate β-catenin functioning in several different ways to modulate its downstream role in transcriptional activation.

5.1.2. Mutations in Wnt/β-catenin signalling pathway

Underscoring the relevance of the Wnt/β-catenin pathway, this pathway is frequently hijacked in cancers. This most notably occurs in cell types that are normally Wnt-dependent (Polakis, 2000; Giles, Van Es and Clevers, 2003). Besides initiating tumours, mutations in the Wnt/β-catenin pathway are also found to positively influence tumour progression through processes such as invasion, tubular branching, tumour growth and epithelial to mesenchymal transitions in colorectal tumours (Hao *et al.*, 1997; Kirchner and Brabletz, 2000). Upregulation of β-catenin signalling is therefore correlated with aggressive and metastatic colorectal tumours. Whilst all oncogenic mutations eventually lead to stabilisation of β-catenin levels,

many components involved in the signalling cascade have mutated in tumours as is summarised in Clevers & Nusse, 2012.

The majority of colorectal tumours, for example, have inactivating mutations on both copies of the *APC* gene, the main negative regulator of the Wnt/β-catenin pathway (Polakis, 2000; Kramps, 2002). Loss of function mutations are frequently found in APC, thereby perturbing differentiation and triggering tumour formation (Fodde, Smits and Clevers, 2001; Kielman *et al.*, 2002; Gaspar and Fodde, 2004). Loss of APC by itself leads to cytoplasmic accumulation of β-catenin but is not sufficient to promote its nuclear translocation (Phelps *et al.*, 2009), whereby only 1-2% of the tumours harbouring APC additionally upregulate signalling through mutations in *CTNNB1* (the gene encoding for β-catenin) itself (Polakis, 2000; Kramps, 2002), thus suggesting there is no strong selection for mutations in both components of the pathway. Additional mutations resulting in oncogenic K-ras are thought to enhance Rac1-dependend nuclear relocalisation of β-catenin and these two targets are hence frequently mutated together (Phelps *et al.*, 2009).

5.1.3. Oncogenic mutations found in *CTNNB1* cluster in the N-terminus

While in some developmental disorders loss-of-function mutations are found in *CTNNB1* (Kuechler *et al.*, 2014), most mutations found in tumours stabilise β-catenin protein levels. The non-uniform distribution of predominantly missense mutations in *CTNNB1* reveals a mutation hotspot frequently seen in oncogenes such as *TP53.* This clustering is consistent with oncogenic mutations in *CTNNB1* acting through a gain-of-function mechanism (Walker *et al.*, 1999; Glazko *et al.*, 2006).

Mutations in *CTNNB1* are found in a wide range of tumours, including soft tissue (42.5% of all tumours), pituitary (37.0%), liver (21%) and endometrial tumours (18.2%) (Forbes *et al.*, 2018). Several common themes are apparent across cancer types. For instance, 88.8% of all mutations found in *CTNNB1* are missense, of which 92.4% occur within residues 32-45 in exon 3 (see **Figure 5.1C**) (Forbes *et al.*, 2018). In-frame deletions of the entire exon (Rebouissou *et al.*, 2016), or residues 32-37 (Cairo *et al.*, 2008) have also been described in some tumours, although at much lower frequency. As explained, this N-terminal region covers the phosphorylation residues necessary for protein turnover and mutations in this

region hence desensitise the protein to proteolytic degradation (see **Figure 5.1B**) (Polakis, 2000; Fodde, Smits and Clevers, 2001; Giles, Van Es and Clevers, 2003; Luo *et al.*, 2007).

Measurements of signalling activity in colorectal cell lines show that some mutations in *APC* and *CTNNB1* only cause a subtle increase in β-catenin and that particular sites cause specific levels of signalling (Rosin-Arbesfeld *et al.*, 2005). It has also been shown that certain nucleotide substitutions encoding for the same amino acid in *CTNNB1* can result in different levels of β-catenin signalling in hepatocellular carcinoma (Austinat *et al.*, 2008). Asp32, Ser33, Gly34 and Ser37 are often found to be mutated in cancers, consistent with the most conserved residues in the ^{32}DpSGϕXpS^{37} binding motif for β-TrCP as mentioned in section 5.1.2. In contrast, Ile35 is rarely mutated in cancer (Polakis, 2000; Rebouissou *et al.*, 2016). Despite the previously mentioned phosphorylation cascade, mutations in the Ser45 residue of β-catenin have been shown to not prevent phosphorylation at the other residues (Wang, Vogelstein and Kinzler, 2003), suggesting that, at least in colorectal cancer, this site is not always essential for phosphorylation. It is suggested that mutations affecting the β-TrCP binding motif have a higher transcriptional activity compared to mutations in Thr41, which are in turn more disruptive than mutations in Ser45 (Rebouissou *et al.*, 2016). As S45 is the most frequently altered in tumours after T41, these results suggest that partially increased signalling is preferred over a complete loss of β-catenin turnover (Forbes *et al.*, 2018).

5.1.4. Just right signalling model suggests selection for an optimal level of β-catenin signalling

Whilst the previous section discussed that oncogenesis can involve selection for mutations enhancing β-catenin expression, overexpression of *CTNNB1* with missense mutations frequently found in hepatocellular carcinoma suggests that these variants result in different levels of downstream signalling and that optimal levels of β-catenin signalling are not consistent between tumour types (Austinat *et al.*, 2008). This in turn suggests that, whilst upregulation of β-catenin signalling promotes tumour progression, β-catenin levels need to be optimal in each cell type. Albuquerque *et al.* hypothesised this and stated that cells with mutations blocking β-catenin degradation over a certain threshold (either through β-catenin directly or through another partner such as APC) may trigger apoptosis and are hence not clonally expanded inside a tumour. This 'just right' model is also consistent with mutations

in *CTNNB1* rarely being found biallelically (Björklund *et al.*, 2008; Le Guellec *et al.*, 2012), although this could also partially be explained by individual missense mutations being rare (compared to the likelihood of loss-of-function mutations) and are hence less likely to happen on both alleles in a cell. By now, this 'just right' model has been well described for colorectal cancers (Lamlum *et al.*, 1999; Albuquerque *et al.*, 2002, 2011), whereby it is shown that classes in mutations in APC differ along the colorectal tract. In addition, hypomorphic *Apc* mutations in mice remarkably result in liver tumours but not intestinal tumours (Gaspar and Fodde, 2004), giving additional support to the idea that tumorigenesis happens in a tissue-specific way depending on Wnt/β-catenin signalling levels.

5.1.5. Mutational signatures arise from different mutational processes

Tumours arise from somatic cells undergoing stepwise mutagenesis. This creates genetic diversity amongst cells on which natural subclonal selection can subsequently act, which is a characteristic of Darwinian evolution (Curtis, 1965; Nowell, 1976). All cells in an organism are exposed to a continuous mutational burden proportional to their proliferation rate and age due to intrinsic infidelities in the DNA replication and repair machinery (Stratton, Campbell and Futreal, 2009). For instance, over-activity of members of the APOBEC family of cytosine deaminases can cause C>U conversions that typically get repaired by base excision repair, which when mutagenic most often results in C>T conversions (Chen *et al.*, 1987; Powell *et al.*, 1987; Seeberg, Eide and Bjørås, 1995).

In addition to intrinsic factors, exposures to exogenous carcinogenic factors can result in accumulation of distinct mutations in specific tissues (Pfeifer, 2010). For example, ultraviolet radiation in sunlight is a common source of DNA damage but does not typically penetrate further than our skin. The elevated abundance of C>T substitutions in melanomas are hence thought to be the result of excessive exposure to ultraviolet light (Ravanat, Douki and Cadet, 2001; Pleasance *et al.*, 2010). In contrast, substitutions due to tobacco carcinogens are typically associated with C>A and CC>AA in lung cancer (Pfeifer *et al.*, 2002). Hence, each carcinogenic process operating on a tissue generates a characteristic imprint, or mutational signature, on the cancer genome. With multiple processes acting on a given tissue and generating characteristic substitutions, cancer genomes were previously deconvoluted to

elucidate the mutational contribution of up to 21 different processes (Alexandrov *et al.*, 2013; Alexandrov and Stratton, 2014).

To identify a mutational signature, single nucleotide substitutions in cancer genomes cancer type are enumerated in their trinucleotide context, i.e. accompanied by their 5' and 3' neighbouring bases. For example, T[C>A]G would denote a C to A substitution in its trinucleotide context of TCG. As each base can be mutated into the 3 other possible bases, 4 possible neighbouring bases can be present on either side of the target base and mutations on the Watson and Crick strands cannot be distinguished (i.e. C[G>T]A on the Watson strand has the same result as T[C>A]G on the Crick strand), there are (3 x 4 x 4 x 2) 96 possible substitution mutations. This thus allows one to distinguish cases where the same mutation occurs in different contexts. This is crucial as it has been shown that mutation frequencies are heavily influenced by nucleotide context (Ellegren, Smith and Webster, 2003).

With the knowledge of the underlying mutational processes (e.g. infidelities in DNA maintenance and replication, the carcinogens to which specific tissues are exposed and additional information on the age of collection of tumours), the contribution of each mutational process to a specific tissue, the mutational signature, can be computed through mathematical analyses (Alexandrov *et al.*, 2013). These mutational signatures together make up the eventual palette of genetic changes found in a tumour, which is referred to as the mutational portrait (Helleday, Eshtad and Nik-Zainal, 2014). Hence, mutational signatures provide insights on (1) the mechanistic bases of endogenous and exogenous carcinogenic burdens and (2) the likelihood of a nucleotide substitution to arise in a particular cell lineage.

Mutational signatures explain how certain tissues may acquire specific mutations more frequently than others, affecting the raw material available for natural selection to work upon. In *CTNNB1*, specific tissues show strong biases for residues between tumour types. For example, whilst the serine phosphorylation residues Ser45, Ser37 and Ser33 are all encoded by TCT, colorectal cancers mainly present mutations affecting Ser45 and Thr41, while tumours in the endometrium and ovaries predominantly harbour mutations at Ser37 and Ser33 (Polakis, 2000; Albuquerque *et al.*, 2011). A study investigating *CTNNB1* mutational patterns in desmoid tumours (a rare subtype soft tissue tumour) by Lazar *et al.* found that

T41A (59%), S45F (33%), and S45P (8%) were the most frequently occurring in desmoids, with a significantly poorer five-year survival rate in S45F-mutated desmoids (23%, $p < 0.0001$) versus either T41A (57%) or tumours without a mutation (65%) (Lazar *et al.*, 2008). These results thus show that the observed frequency of mutations is different between tissues that the tumours originate from, but the extent to which either the background mutation rate or selective pressure contributes to the eventual mutational pattern remains a major open question in the field and is crucial in order to understand the biology of tumours.

5.1.6. Motivation for study

Oncogenic mutations in CTNNB1 cluster in the N-terminal region of the protein, typically impeding phosphorylation by CKI or GSK-3β or by destroying the β-TrCP binding motif. Many questions in the field remain, as to why different tumours preferentially harbour mutations at certain residues and to what the functional consequences of specific point mutations are. Most studies thus have not systematically assessed the effects of specific amino acid substitutions in *CTNNB1* on downstream signalling or cell fitness.

Since the mechanism of Wnt/β-catenin signalling is not known to differ between tumour types or between tissues, other cellular mechanisms such as the level of signalling required to trigger apoptosis might underlie the different responses to specific levels of β-catenin signalling. As yet, there no evidence to suggest that the same *CTNNB1* variant can elicit different phenotypic consequences in different cell types, although this has not been rigorously tested. While previous efforts have assessed the effects of mutations in CTNNB1 on downstream TCF/LEF signalling have been tested, only a small number of mutations has been assessed (Austinat *et al.*, 2008). As the hotspot region is confined to a few residues of *CTNNB1* (Figure 5.1C) and tumour-associated mutations are primarily missense, this region was therefore considered as a suitable subject for our deep mutational scanning pipeline. In this chapter, I have adapted the pipeline described in chapters 1 & 2 to assess the effects of single amino acid substitutions in this region on β-catenin signalling.

The experimental design for this study was done by Derya Özdemir and Peter Hohenstein with help from Andrew Wood, and all the experimental work was performed by Anagha Krishna and Derya Özdemir. I performed all of the data analyses throughout the project with

one exception: the background mutation rates in tumours with *CTNNB1* mutation were calculated by Dr Ailith Ewing with help from Colin Semple.

5.2. Results

5.2.1. Using double-stranded repair templates containing codon replacements

In contrast to our study on GFP, we here wanted to assess the consequences of all possible amino acid substitutions within the *CTNNB1* hotspot region. Therefore, the proposed pipeline was reassessed for its applicability to perform deep mutational scanning on the amino acid level. Because the oligo doping approach to template synthesis is biased against amino acid substitutions requiring more than one nucleotide substitution, it was instead opted to synthesise double stranded templates using the ssDNA-synthesis platform provided by Twist Bioscience that allows for the specific incorporation of alternative codons instead of alternative nucleotides, as discussed in the main introduction to this thesis. This should theoretically result in predefined codon substitutions without biasing against those requiring multiple simultaneous nucleotide changes. As described more extensively in section 2.2.2, HDR template plasmids contained ±2-kb homology arms to *CTNNB1*, with residues L31 to G50 of the gene encoding β-catenin replaced with codons encoding for all 19 possible amino acid substitutions (see **Figure 5.2**). As sgRNA binding sites are within the puΔtk selection cassette, the wildtype allele (Allele 1 in figure 5.2a) will remain uncleaved, limiting subsequent edits to the puΔtk allele. In addition, the sgRNA binding sites are removed upon repair by HDR, such that Cas9 is prevented from cleaving either the episomal HDR templates, or genomic DNA from cells that have already undergone HDR. This served to maximise the frequency of HDR-derived alleles in the cells undergoing selection, which was a large bottleneck in the screen on GFP described in the previous chapter. In addition, repair templates contained silent mutations situated just outside the variable amino acid positions, which create novel binding sites for PCR primers, making it possible to selectively amplify from HDR-derived template strands while ignoring unedited alleles or template strands containing indels. Together, these modifications would maximise the percentage of HDR reads available for downstream analyses.

5.2.2. Monoallelic replacement of *CTNNB1* with a *puΔtk* cassette allows for targeting of a single allele and for the selection of targeted cells.

The presence of biallelic copies of *CTNNB1* could obscure both targeting and the downstream analysis of allelic variants. Whilst an approach utilising a single copy proved useful in our analysis of GFP, oncogenic mutations in tumours typically occur in the presence of a wildtype copy in *trans*. To mimic the activity of mutations found in cancer, it was reasoned that the allelic series would have to be introduced onto one of the allelic copies whilst leaving the other copy intact.

For this reason, Anagha Krishna monoallelically replaced exon 1-6 of *CTNNB1* with a cassette containing a fusion protein between puromycin N-acetyltransferase and a truncated form of herpes simplex virus thymidine kinase (*puΔtk*) driven by its own promoter (Chen and Bradley, 2000). This positive/negative selection cassette facilitates positive selection of cells harbouring this cassette in-frame using puromycin, whilst making these cells sensitive to 1-(-2-deoxy-2-fluoro-1-β-D-arabino-furanosyl)-5-iodouracil (FIAU). A schematic overview of our pipeline is displayed in **Figure 5.2B**.

Subsequent targeting of these cells with sgRNAs that direct Cas9 cleavage at either end of the *puΔtk* cassette lead to its excision and replacement with sequence from the HDR donor template library and reconstitution of the CTNNB1 open reading frame (see **Figure 5.2B**). These cells, but not cells that are unedited or have undergone NHEJ, lose sensitivity to FIAU, thereby allowing for the selection of cells that have undergone desired editing outcomes. In addition, silent mutations introduced onto the targeted allele by this final HDR step allow for the selective amplification of this allele by PCR. With this, an approach was set up to introduce putative gain of function mutations specifically into one allele of *CTNNB1* while leaving a wildtype copy intact, and to selectively amplify HDR-derived reads in downstream sequencing.

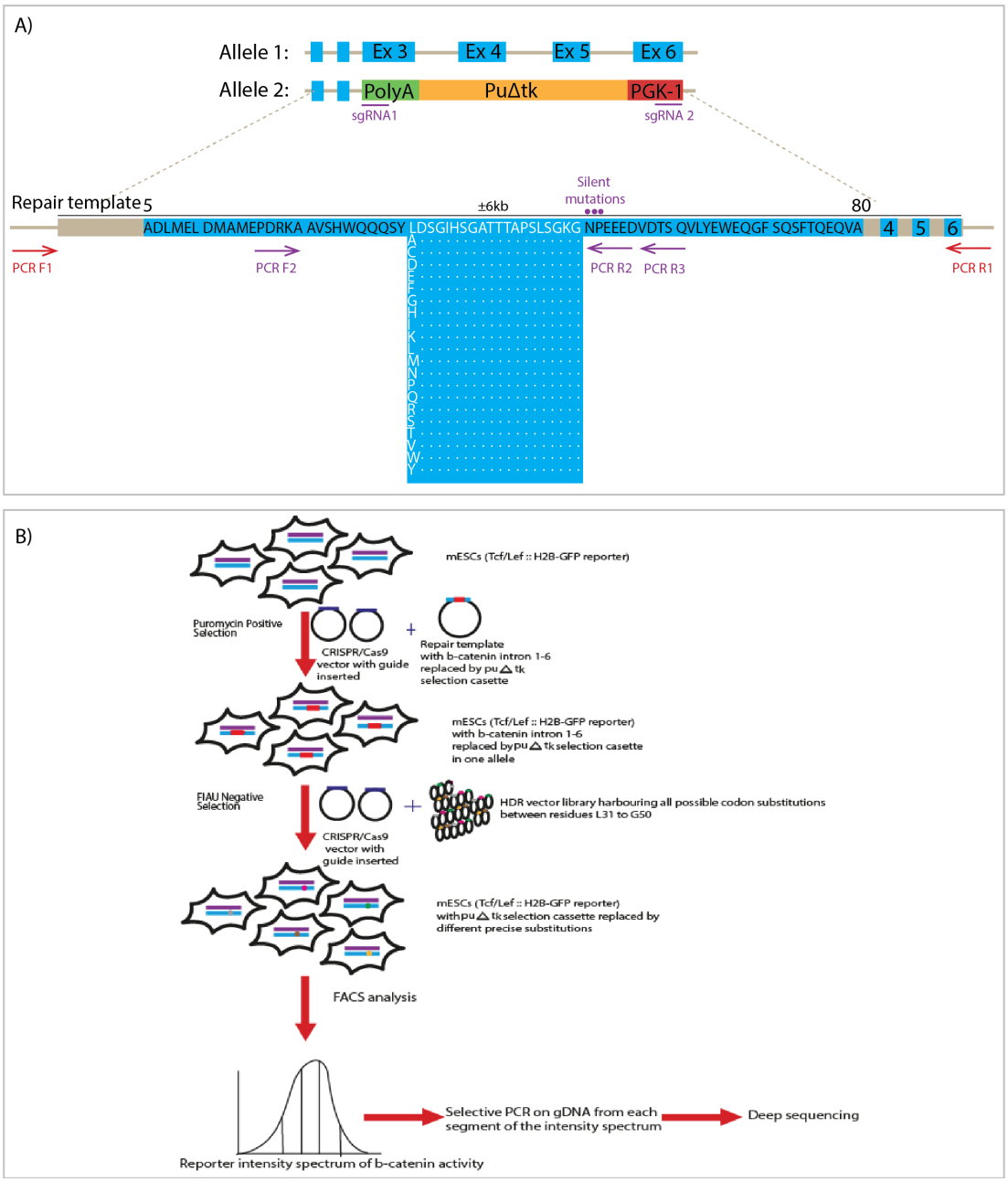


Figure 5.2 Schematic representation of experimental design to perform deep mutational scanning of CTNNB1 (A) Schematic overview of cell line harbouring a single wildtype copy of *CTNNB1*, whilst the other allelic copy has exon 3-6 replaced by an insert containing a puromycin N-acetyltransferase and a truncated form of herpes simplex virus thymidine kinase (*puΔtk*) cassette, a promoter (PGK-1) and bGH polyadenylation signal (polyA), ±2.4kb in size. sgRNAs guide wtCas9 to induce double-stranded DNA breaks at the termini of the cassette that are hence specific for allele 2. HDR repair template plasmids contain endogenous exon 3-6 with a single codon substitution in any of the 20 targeted triplets (amino acid residues 31-50 in exon 3), together making up a total of 380 possible variants, and flanking homology arms of ±2kb. Amplification of the samples by PCR is performed by a primary amplification from outwith the homology arms (F1 & R1) to exclude off-target integrations, whereas a second round of PCR amplification selectively amplifies from HDR-derived alleles harbouring silent mutations at the 5'-end of the targeted region (F2 & R2). In

parallel, the pool ('HRM') was PCR-amplified with an alternative reverse primer (R3) binding 20bp downstream from R2 that was not specific for HDR-derived reads and hence also amplified from non-HDR reads, thereby allowing for the assessment of the frequency of HDR reads in the pool. **(B)** Schematic of *CTNNB1* targeting strategy. TCF/LEF::H2B-GFP ES cells are targeted with CRISPR-Cas9 and repair template targeted with *puΔtk* selection cassette and screened for heterozygous replacements to achieve the cell line described in (A). Targeting of these cells with CRISPR-Cas9 and an HDR vector library followed by selection for successful integration by FIAU results in a pool of cells harbouring biallelic codon variants of *CTNNB1*. After fluoresce-based sorting of cells, genomic DNA was isolated and selectively amplified as described in (A). These amplicons were finally subjected to deep sequencing.

5.2.3. The TCF/LEF::H2B-GFP reporter system allows for the efficient measuring of fluctuations in Wnt/β-catenin signalling activity

In the previous chapter I described how a direct readout of the effects of mutations on the targeted protein was used by either measuring its fluorescence directly (GFP). While I have shown that this is a suitable way to assess loss-of-function or hypomorphic mutations, in the case of β-catenin we wished to assess the effect of gain of function mutations on signalling pathway activation. For this reason, our phenotypic assay utilised a well-established transcriptional reporter of Wnt/β-catenin signalling (Ferrer-Vaquer *et al.*, 2010).

This reporter comprises a transgene containing six copies of a TCF/LEF-responsive element and an hsp68 minimal promoter driving expression of a cDNA encoding human histone H2B fused to GFP (Ferrer-Vaquer *et al.*, 2010). This TCF/LEF::H2B-GFP system has been widely used as a sensitive reporter of the Wnt/β-catenin signalling pathway at single cell resolution both *in vitro* and *in vivo* (Faunes *et al.*, 2013; Kuwajima *et al.*, 2013). It was hypothesised that cells harbouring mutations stabilising β-catenin would lead to an increase in GFP fluorescence and could subsequently be separated out by FACS, similar to the experiments done in the previous chapter. Therefore, mouse embryonic stem cells harbouring this reporter were used as the model system for this study, which were derived specifically for this project by the lab of Anna-Katerina Hadjantonakis at the Memorial Sloan-Kettering Cancer Center.

5.2.4. Culturing embryonic stem cells under 2i conditions allows for an unbiased integration of amino acid substitutions into *CTNNB1*

A key aim of culturing embryonic stem (ES) cells *in vitro* is to maintain pluripotency and avoid differentiation. With the establishment of the first murine ES cell culture in 1981, these cells

were cultured on irradiated mouse embryonic fibroblasts (MEFs) in media containing fetal bovine serum (FBS) (Martin, 1981), until it was discovered that leukaemia inhibiting factor (LIF) could replace MEFs (Smith *et al.*, 1988; Williams *et al.*, 1988).

Three decades later, media containing both serum and LIF, either on feeder cells or gelatinized surfaces, are still routinely used for the culturing of ES cells in many labs. Since sera such as FBS are animal derived, a shortcoming of these culture conditions is that the media is not completely defined, may show batch variability and contain differentiation factors.

In 2008 an alternative, defined culturing method was developed containing inhibitors of two key signalling pathways; the mitogen-activated protein kinase/extracellular-signal-regulated kinase (MEK) inhibitor PD0325901 and GSK-3β inhibitor CHIR-99021 (Ying *et al.*, 2008). These two inhibitors ('2i') enable the maintenance of self-renewal and pluripotency marker expression in mouse ES cells, without the use of animal-derived factors. 2i is increasingly used in the field and is thought to better represent 'ground state pluripotency' than culturing under serum and LIF (Ying *et al.*, 2008).

As any oncogenic mutation affecting β-catenin would immediately give these cells a proliferative advantage in the presence of Wnt signalling, culturing cells using serum and LIF would adversely bias our screen towards these faster replicating cells. Therefore, cells were cultured under 2i conditions to avoid the confounding effect on cell fitness. As stated in the Materials & Methods chapter, two days before selecting cells based on their activity by FACS, the GSK-3β inhibitor CHIR-99021 was removed from the culture media, thereby unmasking the effects of amino acid substitutions in β-catenin on signalling without giving these cells enough time to selectively proliferate.

5.2.5. Targeting cells with CRISPR-Cas9 and HDR library causes an increase in TCF/LEF signalling

Analysis by FACS revealed that the targeted cell population showed a wider range in increased GFP fluorescence compared to the unedited population (compare **Figure 5.3A** with **Figure 5.3C,D**), suggesting an increase in signalling in a subpopulation of edited cells.

As it was aimed to detect subtilties in effects of mutations on β-catenin signalling, cells were sorted into six equally log-spaced bins (referred to as bin P2-P7) based on their fluorescence (with **Figure 5.3C,D** and **Table 5-1**) and genomic DNA was isolated from each bin. In addition, unsorted cells (referred to as pool) were also processed for genomic DNA isolation. This experiment was performed in duplicate to account for experimental noise. In the sections hereafter, I will refer to the directly sequenced dsDNA repair template pool as "plasmid", the entire unsorted population of targeted cells as "pool" and each of the six fluorescence-restricted populations as "bins".

After selectively amplifying HDR-derived alleles from the genomic DNA (from samples P2-P7 and the pool, amount of gDNA denoted in **Table 5-1**) by two sequential PCR steps (see **Figure 5.2A**), samples were indexed for multiplexing. In parallel, the KO allele was amplified and indexed with a different reverse primer (P3) that did not selectively amplify HDR reads (referred to as HRM, see **Figure 5.2A**) in order to ascertain the proportion of HDR reads as part of the whole pool. For some samples the amount of gDNA suggested a higher number of biallelic copies (e.g. 133 ng represents ± 2.2 x 10^4 biallelic copies in P7.1) than the number of cells sorted (e.g. 4,075 cells in P7.1) and the gDNA concentrations are thus likely to be overestimated. Therefore, the experimental design hereby does not allow for the accurate estimation of the number of HDR alleles that were amplified. Samples were sequenced on the Illumina MiSeq platform using 150bp paired-end sequencing.

5.2.6. Data analysis shows high HDR efficiencies and a high sequencing depth per sample

After trimming reads for both adapter sequences and bases with a low-quality score (as described in Chapter 3), pairs of sufficient read length were aligned to the reference sequence. For the biological replicates of HRM (i.e. the sample that had not undergone a selective PCR-step for HDR-derived alleles and hence includes all HDR and non-HDR reads), perfect FIAU selection resulting in 100% cells having undergone HDR would yield 50% wildtype reads. Respectively 497,470 and 430,202 reads (64.8% and 67.9%) aligned to the wildtype reference sequence (of 162 nt), whereas 205,698 and 162,663 reads (26.8% and 25.6%) mapped to the reference sequence containing the silent mutations introduced from

Table 5-1 Number of cells sorted per bin Cells were sorted into a total of six bins which were equally log-scaled, with the exception of p2 which spanned a relatively larger range. Amounts of genomic DNA (gDNA) used per PCR reaction, whereby the lower values (i.e. p7) were at very low concentrations and correspond with higher numbers of cells and are therefore likely to be overrepresented. Values were provided by Anagha Krishna.

Bin	Fluorescence range	Replicate 1		Replicate 2	
		Cells sorted	gDNA used in PCR (ng)	Cells sorted	gDNA used in PCR (ng)
p2	5.0×10^2 - 5.0×10^3	100,000	2,156	100,000	2,254
p3	5.0×10^3 - 1.0×10^4	100,000	2,646	100,000	3,469
p4	1.0×10^4 - 5.0×10^4	100,000	2,744	100,000	3,577
p5	5.0×10^4 - 1.0×10^5	30,000	475	50,000	10,098
p6	1.0×10^5 - 5.0×10^5	234,000	5,929	211,000	13,573
p7	5.0×10^5 - 1.0×10^6	4,075	133	4,234	87
Pool	All cells	200,000	2,500	200,000	5,941

the HDR repair template, which is 142 nt in length, while not mapping to the non-HDR reference sequence. This was expected, as the HRM samples (i.e. the sample that had not undergone a selective PCR-step for HDR-derived alleles and hence includes all HDR and non-HDR reads) were amplified with F2 and R3 rather than F2 and R2 primers (see **Figure 5.2A**) and therefore contained both HDR and non-HDR reads. A small proportion (64,301 and 41,064 (8,4% and 6,6%)) of the reads did not map to either of the reference sequences, and a closer examination demonstrated that these were reads with more than 6 mismatches. Sequences mapping to the wildtype reference did not contain the silent mutations specific for HDR-derived alleles whereas the reads mapping to the HDR reference did (see **Figure 5.2A**), indicating that the overall frequency of HDR-derived reads approximates 26.2% in both biological replicates. Considering that only HDR-derived alleles and non-targeted allele from both edited ('HDR cells') and unedited cells are amplified, an equal proportion of the reads (26.2%) should result from the unedited allele from 'HDR cells' and (100 - 26.2 - 26.2 =) 47.6% of the reads come from the unedited allele of non-HDR cells. Thus, about a third (26.2 * (100/(47.6+26.2)) =) 35.5% of the cells after FIAU selection harboured an HDR allele. While the overall HDR efficiencies (i.e. before FIAU selection) cannot be assessed, these results suggest that FIAU selection is efficient in enriching cells that have undergone HDR.

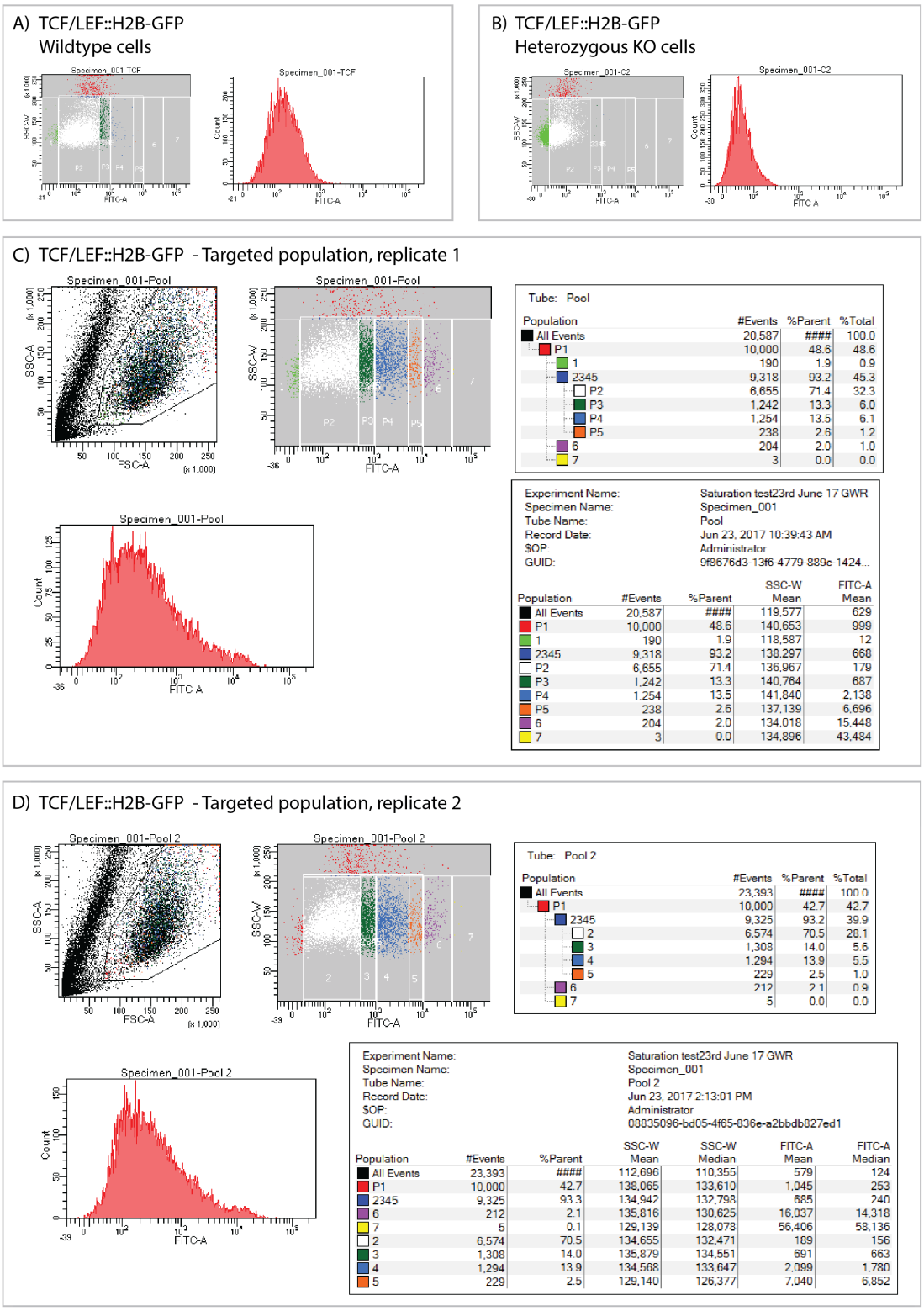


Figure 5.3 FACS analysis of TCF/LEF::H2B-GFP post-targeting (A) Fluorescence profile of wildtype TCF/LEF::H2B-GFP cells harbouring biallelic wildtype copies of *CTNNB1* shows that fluorescence does not exceed ± 5 x 10^4 fluorescence units. **(B)** Fluorescence profile of TCF/LEF::H2B-GFP cells harbouring heterozygous replacement of *CTNNB1* with *puΔtk* cassette show reduced fluorescence with respect to the wildtype cells depicted in (A). **(C,D)**

Fluorescence profiles of cells (C: replicate 1; D: replicate 2) targeted harbouring allelic variants shows increased range of GFP fluorescence. Fluorescence increased beyond the baseline levels shown in (A). Cells were sorted into bins 2-7, which represent equally log-spaced bins as described in table 5.1.

For the other samples (i.e. P2-P7 and Pool, all amplified with F2 and R2 primers, see **Figure 5.2A**), consistent mapping efficiencies of >98% to the HDR reference sequence were observed, with an average of 679,845 ± 76,509 mapped reads per sample. As this mapping only occurs for reads containing silent mutations introduced by HDR, this indicates that the selective PCR efficiently discriminates HDR-derived reads from wildtype. Read pairs without insertions or deletions and with no more than 6 conflicting nucleotides with respect to the reference sequence were considered for further analysis. On average 463,583 ± 72,680 reads (68.2%) passed this threshold. Because 381 possible variants were assessed (19 codon substitutions at 20 positions and the wildtype allele), this suggests an average 10X coverage for variants occurring at a frequency as low as 2.16×10^{-6}. However, as for p7 only ± 4,000 cells were sorted, this reflects ± 100X coverage, whereas in P2 not all cells will have undergone HDR (as P2 reflects wildtype expression levels) and with 100,000 there will thus be < 10-fold coverage. The overall coverage was however deemed sufficient to correctly assess variances in proportions between the bins.

5.2.7. Amino acid substitutions are efficiently integrated through homology-directed repair

Whilst multiple codons can encode the same amino acid, only a single codon was used per amino acid change in our repair templates. Amino acid replacement frequencies are displayed in **Figure 5.4**. The average percentage of non-consensus codons in the non-variable region (residues 19 – 30) across all samples is 0.02%, which is consistent with the error rate of the Illumina MiSeq platform (0.0064% * 3 nucleotides = 0.0192%) (Schirmer *et al.*, 2015). The observed frequency of non-consensus codons per site is 5.92 ± 1.34% in the variable region (residues L31 – G48) of the pool (i.e. the total population of targeted unsorted cells) and closely resembles the frequency found on the plasmid (5.59 ± 1.14%), whilst the frequencies of non-consensus codons correlate between plasmid and the sequenced pool (R^2 = 0.737, see **Figure 5.5**). Whilst there is some degree of variability, this demonstrates that genomic integration of the plasmid sequences through HDR is largely unbiased and all codons present in the plasmid pool are represented in the pool.

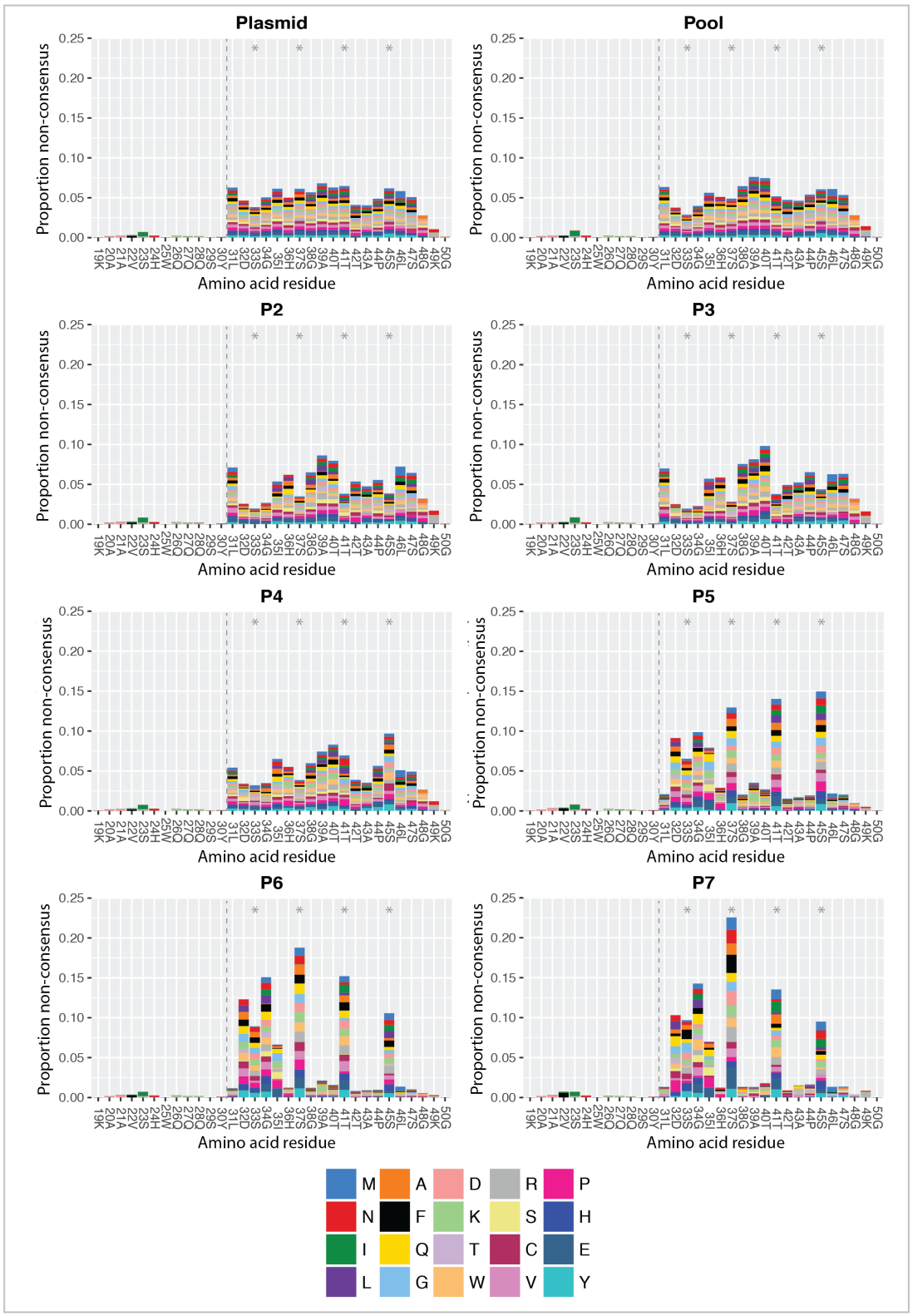


Figure 5.4 Stacked bar plot of amino acid substitution rates per codon Colours represent specific amino acid substitutions as explained in the legend at the bottom. Dashed line separates saturated residues (right) from constant residues (left), whereas asterisks indicate phosphorylation sites. In the repair template plasmid, non-consensus amino acid

substitutions show elevated levels in targeted residues L31-48G as compared to non-targeted residues 19K-30Y. This heterogeneity was not observed in residues 49K and 50G. Amino acid substitution profiles are consistent between the plasmid and unselected pool, suggesting an unbiased genomic integration of the nucleotide diversity found in the plasmid library. Amino acid substitutions at residues 32D, 33S, 34G, 37S, 41T, 45S are depleted in bins P2, P3, P4 but show an enrichment for amino acid substitutions in bins P5, P6, P7.

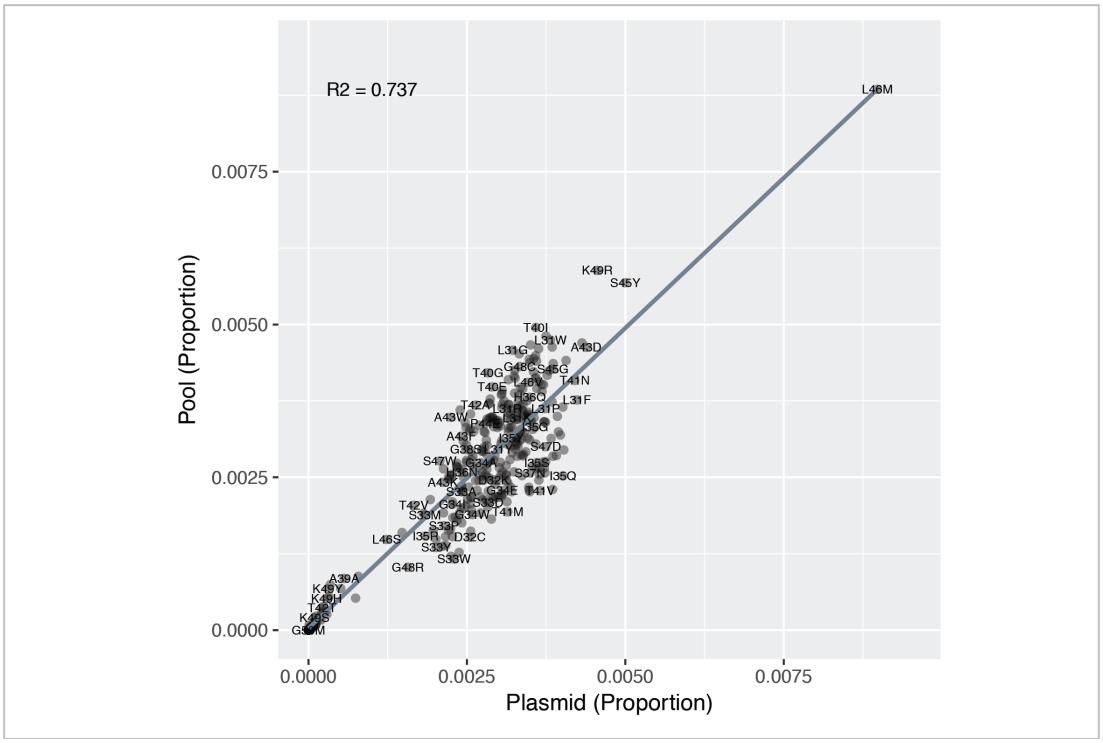


Figure 5.5 Correlation of codon frequencies in repair template plasmid and cell pool Codon frequencies as proportion of the total number of HDR reads were analysed in both the repair template plasmids and in the unsorted cell pool to assess the level of integration. Codons occurring at low proportions were typically those at residues 49 and 50, as is also visible in figure 5.4. L46M was observed as the most frequent codon in both the plasmid and pool. Analysis showed a strong correlation (R^2=0.737), indicating efficient genomic integration of the plasmid pool through HDR.

Codon frequency analysis of the template plasmid confirms that, whilst there are differences between the residues, all intended codon substitutions are represented across targeted residues L31 – G48 (see **Figure 5.4**), while an underrepresentation of codon substitutions was observed at residues K49 and G50. Frequencies of these codons are consistently lower in the repair template plasmid similarly to the targeted pool of cells (see **Figure 5.5**), thus excluding the possibility of a skew in editing. This underrepresentation was found to be consistent throughout all sequenced samples (see **Figure 5.4** and **Figure 5.5**). This raised the possibility of a bias in our alignment pipeline, a skew in library preparation or an inherent decrease in

diversity at these distal residues, resulting in the loss of these codons in the eventual pool. To distinguish between these possibilities, an unbiased search using the UNIX command grep in the unprocessed data files for sequences containing codon substitutions at these positions, confirmed that these sequences are scarcely present in all sequenced samples. Finally, frequencies are also consistently low in the HRM sample, which uses a sequencing primer different from the other samples. From this it was deduced that these sequences are truly underrepresented in the sequencing pool and that the low diversity stems from an issue during construction of the original plasmid donor library used for this screen. For this reason, only amino acid substitutions at residues L31 – G48 were considered for further analyses.

5.2.8. Technical replicates show reproducibility of the data

Frequencies of non-consensus codons show a very high correlation between technical replicates of the plasmids (R^2 = 0.92; see **Figure 5.6**), indicating that the sequencing depth of the samples is sufficient. The biological replicates of the pool and P2-P7 consistently show a high correlation (see **Figure 5.6**) (R^2 = 0.83 ± 0.11). The highest correlation found (P6, R^2 = 0.99) can be explained by the fact that many amino acid substitutions were not represented in this bin and their values hence approximate 0, thereby increasing the correlation between the replicates. Following this reasoning, P7 would be expected to have a similar correlation between replicates, but counterintuitively here the lowest correlation between replicates was observed (R^2 = 0.653). As the sequencing depth for these samples was not different from the other bins, this outlier was considered to likely be due to the low number of cells that was sorted for this bin (see **Table 5-1**).

To further ascertain whether a high enough coverage was obtained in each bin, the extent to which the combined bins resemble the pool was assessed. The frequency of each codon in the pool c_{ir} was therefore artificially reconstructed by multiplying the frequency of a given codon *i* in each bin *j* (c_{ij}) by the fraction of cells sorted into that bin f_j, such that:

$$c_{ir} = \sum_j c_{ij} \cdot f_j \tag{5.1}$$

The codon frequencies in the reconstructed pool highly correlate with the frequencies from the sequenced pool (R^2 = 0.835, see **Figure 5.7**), suggesting that both sampling and sequence

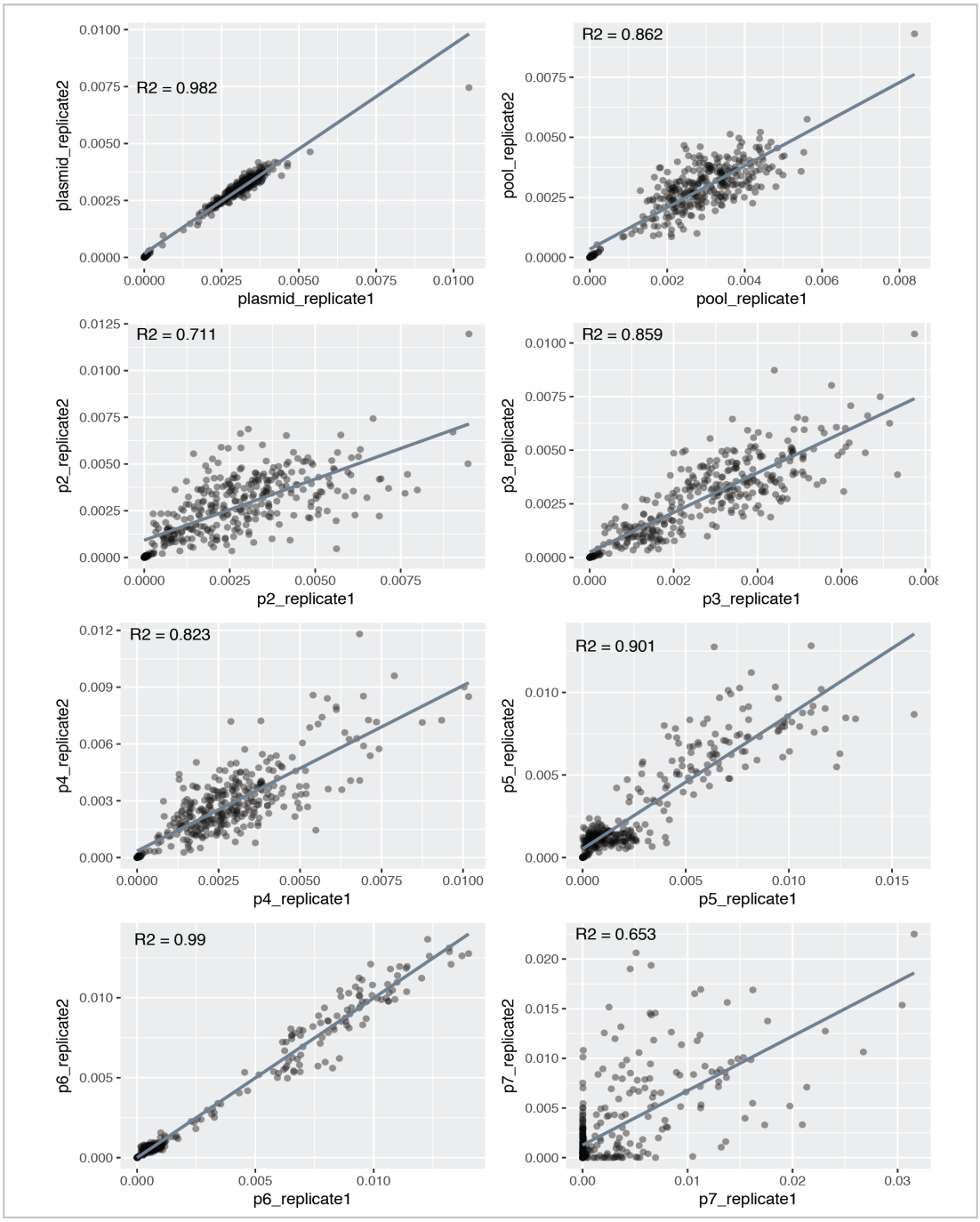


Figure 5.6 Correlation of codon frequencies in sequenced replicates Technical (plasmid sample) and biological (pool and P2-P7) replicates show high correlation, indicating a low variation between biological replicates. P7 shows the highest variation which is likely due to the low number of cells sorted as displayed in table 5.1.

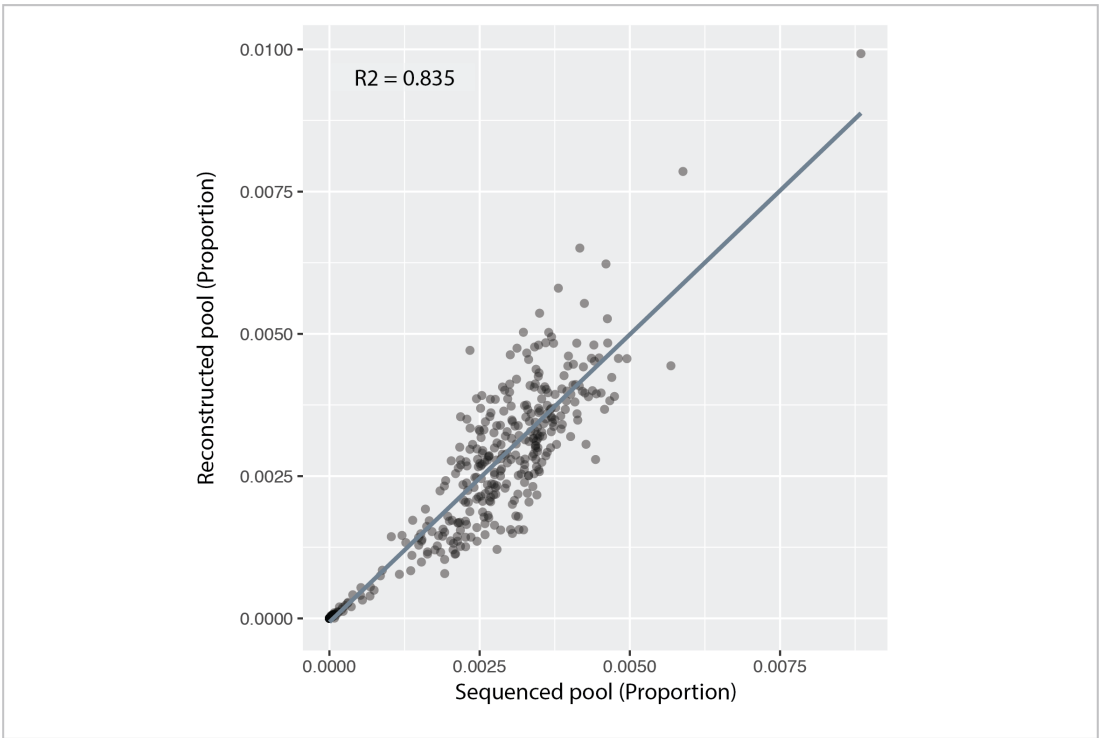


Figure 5.7 Correlation of codon frequencies in sequenced pool and reconstructed pool Codon frequencies from individual bins P2-P7 were combined proportionally to the number of cells sorted in that bin in order to assess the extent to which they together represent the unsorted pool. Codon frequencies show a high correlation between the sequenced pool and the reconstructed pool, indicating that there is no loss in diversity through separation of cells into individual bins.

coverage within each bin is sufficient to accurately represent the true allelic frequencies. With correlation for all replicates being high, the biological replicates were merged for further analyses.

5.2.9. Reads predominantly harbour a single alternative codon

As mentioned, the aim is to assess the effects of all single amino acid substitutions across the targeted residues. To ensure only assessing alleles with a unique amino acid change were assessed, the number of non-consensus codons per read in each bin were analysed (see **Table 5-2**). The highest frequencies of wildtype reads (i.e. with no codon substitutions) were found in bins P2 and P3, which correspond with the baseline fluorescence levels of TCF/LEF::H2B-GFP in unedited cells. Reads with no mutations were rare across all populations (i.e. < 10%) and as expected, were even lower in bins with β-catenin signalling levels exceeding wildtype (i.e. < 3% in bins 6 and 7). With a sequencing error rate of 0.02%, (54 * 0.02 % ≈) 1.08% double mutations would be expected due to sequencing error. Reads with

Table 5-2 Number of non-consensus codons per read. In each sample, the number of non-consensus codons per read were counted. Whilst in all samples the vast majority of the reads contained a single codon change, approximately 5% of the reads in the plasmid and 7% in the pool contained no mutation and less than 3% contained more than a single mutation. Across the bins, reads with no mutation are more frequently found in the bins of low GFP fluorescence. The proportion of reads with more than a single mutation did not show large fluctuations between the bins.

	Number of non-consensus codons (read count)						
Sample	**0**	**1**	**2**	**3**	**4**	**> 4**	**Total**
Plasmid_1	22,935 5.2%	412,429 92.6%	9,671 2.2%	116 0.0%	2 0.0%		445,153
Plasmid_2	22,224 5.9%	344,630 92.1%	7,225 1.9%	91 0.0%	 0.0%		374,170
pool_1	36,941 7.2%	459,119 89.9%	14,370 2.8%	226 0.0%	2 0.0%		510,658
pool_2	32,280 7.1%	406,863 89.9%	13,132 2.9%	214 0.0%	4 0.0%		452,493
p2_1	42,540 9.4%	395,869 87.9%	11,938 2.6%	172 0.0%	2 0.0%		450,521
p2_2	42,804 8.2%	459,333 88.3%	17,434 3.4%	547 0.1%	11 0.0%	1 0.0%	520,130
p3_1	37,304 8.0%	414,470 89.0%	13,759 3.0%	185 0.0%	3 0.0%	1 0.0%	465,722
p3_2	27,965 6.4%	397,381 90.5%	13,579 3.1%	244 0.1%	4 0.0%		439,173
p4_1	31,880 6.5%	444,303 90.5%	14,339 2.9%	219 0.0%	1 0.0%		490,742
p4_2	30,644 6.5%	428,893 90.5%	14,307 3.0%	261 0.1%	1 0.0%		474,106
p5_1	16,050 3.6%	413,139 93.3%	13,296 3.0%	213 0.0%	3 0.0%		442,701
p5_2	16,891 3.3%	474,162 93.7%	14,777 2.9%	230 0.0%	3 0.0%		506,063
p6_1	12,328 2.5%	461,180 94.2%	15,618 3.2%	261 0.1%	3 0.0%		489,390
p6_2	12,692 2.8%	423,905 93.8%	14,886 3.3%	236 0.1%	5 0.0%		451,724
p7_1	5,638 1.3%	425,364 95.0%	16,734 3.7%	243 0.1%	3 0.0%		447,982
p7_2	8,852 1.9%	434,828 95.2%	12,713 2.8%	199 0.0%	 0.0%		456,592

more than a single mutation were not enriched in any of the populations, which could be explained by combinations of codons either disrupting β-catenin phosphorylation (expected to result in higher GFP) or disrupting stability of the protein all together (which would result in low GFP). As the aim was to assess the effects of single amino acid changes, only reads containing only a single alternative codon in the targeted residues were passed.

5.2.10. Amino acid substitutions are found at different proportions in the bins

With the knowledge that our assay had efficiently integrated the desired genomic variants at positions 31-48 in an unbiased manner, it was assessed whether the frequencies of particular substitutions differed among the cell populations sorted into 'bins' based on transcriptional activation of TCF/LEF::H2B-GFP (see **Figure 5.3C,D**). Plotting the codon substitution frequency per residue (**Figure 5.4**), it is clear that bins P2, P3 and P4 show a spread of substitutions across all residues while slightly depleted at residues D32, S33, G34, S37, T41 and S45 (with respect to the pool), whereas the bins with higher transcriptional activity (P5, P6, P7) show an enrichment over these latter mentioned residues. This boundary in mutational patterns between bins P4 and P5 is consistent with the GFP fluorescence of the unedited samples, which does not extend beyond 1.0 x 10^3 fluorescence units (see **Figure 5.3C,D**), corresponding to the upper limit of P4. This suggests that cells with fluorescence higher than P4 (i.e. P5, P6, P7) harbour mutations which enhance reporter expression. This threshold identifies positions that are permissive for mutations (those mutated in P2, P3, P4) and positions at which disturbances tend to lead to an increase in transcriptional activity, i.e. the CKI and GSK3β phosphorylation sites and residues in the ^{32}DpSGϕXpS^{37} binding motif for β-TrCP. As previous studies have shown that these residues are involved in the proteolytic turnover of β-catenin (see section 5.1.3), this confirms that our pipeline is able to detect amino acid substitutions that enhance the effect of β-catenin on TCF/LEF expression.

5.2.11. Hydrophobic residues do not affect β-catenin activity at I35

Amino acids have distinct physicochemical properties that, in combination, determine the structure and function of a protein. For example, hydrophobic residues are typically buried inside a protein whereas hydrophilic amino acids are typically found at the surface of a protein. Similarly, amino acids are crucial in the formation of binding pockets for interactions with other proteins or nucleic acids. As some amino acids are more similar than others (see

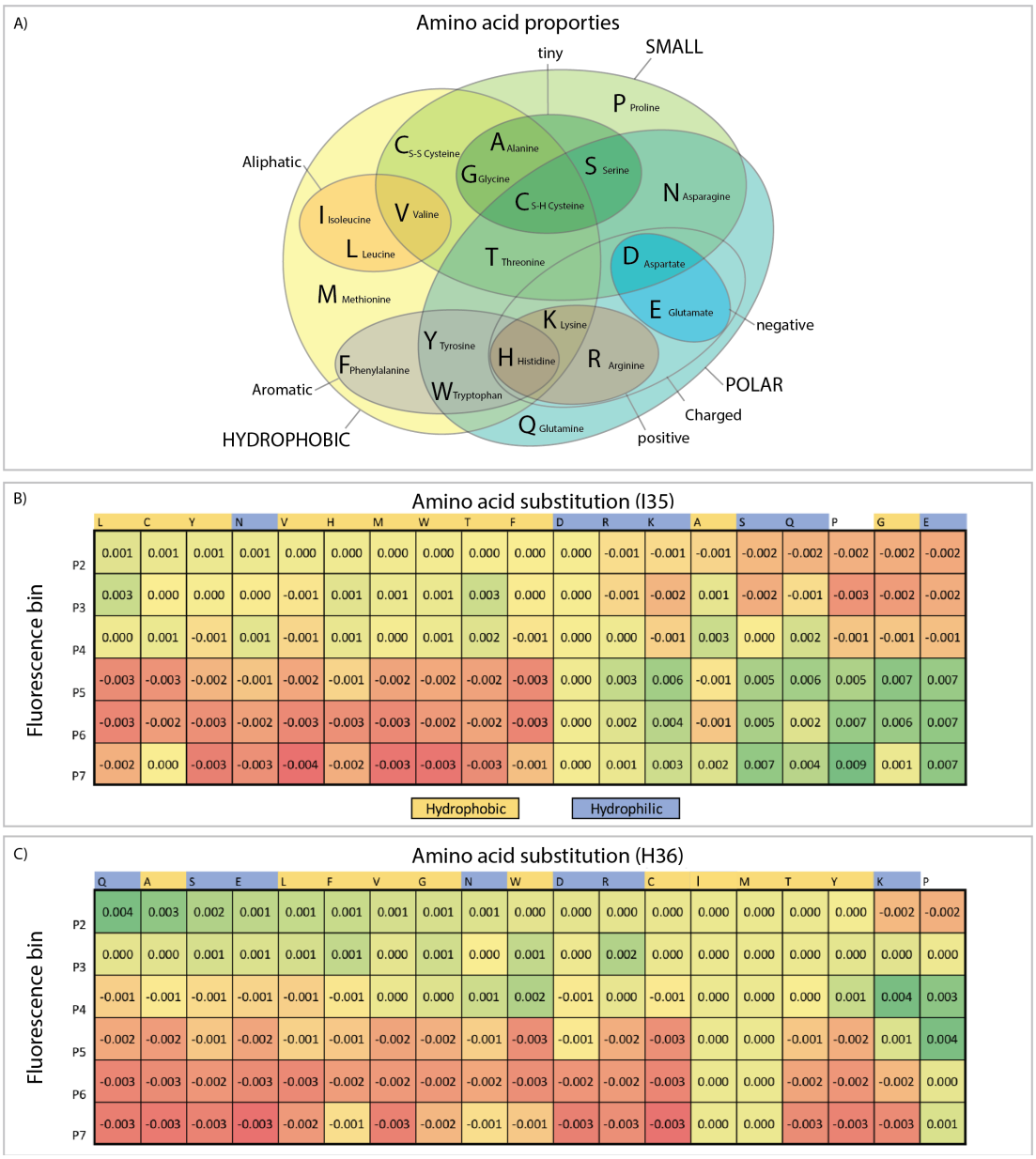


Figure 5.8 Codon enrichment scores of amino acid substitutions at different sites (A) Overview of the 20 different amino acids and their properties. Each amino acid has specific characteristics defining its role in protein structure. The thiol side chain of cysteine is susceptible to oxidation resulting in the disulphide derivative cysteine ("S-S Cysteine"), which serves an important role in many proteins. Therefore, both forms of cysteine are depicted. These characteristics aid in understanding why some amino acid substitutions have a larger effect on protein stability and function than others. Figure adapted from (Livingstone and Barton, 1993). **(B)** Enrichment of non-consensus amino acids at residue 35 across bins. Amino acids with hydrophobic properties (highlighted in yellow) and asparagine (N) are mostly enriched in populations P2-P4 indicating that they do not affect β-catenin activity. Amino acid substitutions with hydrophilic properties proline (P) are underrepresented in populations P2-P4 but enriched in P5-P7, indicating that these substitutions may interfere with the turnover of β-catenin. Alanine (A) and glycine (G) are hydrophobic but show a positive effect on β-catenin signalling, which could be explained by their tiny side chain (see

figure A). **(C)** Enrichment of non-consensus amino acid substitutions at residue 36 across the bins. Most amino acid substitutions show an enrichment in the lower fluorescence bins, indicating a neutral effect o β-catenin signalling, which supports earlier studies describing that this residue does not interact with β-TrCP upon binding. Lysine (K) and proline (P) show some enrichment in higher populations indicating that these amino acids (partially) increase β-catenin signalling.

Figure 5.8), certain substitutions are more likely than others to conserve protein stability and function. β-TrCP binds a motif with a hydrophobic residue at its core, in the case of β-catenin at residue I35 (Wu *et al.*, 2003). In order to test the hypothesis that hydrophobic substitutions are better tolerated than hydrophilic substitutions, the relative frequency of each substitution in each bin compared to the frequency in the pool was calculated.

Analysis of all substitutions at position I35 across all bins revealed that hydrophobic amino acid substitutions are consistently depleted at position I35 in bins P5-P7, whereas proline and polar amino acids are enriched in these bins with higher β-catenin activity (see **Figure 5.8B**). Alanine (A) and glycine (G) are hydrophobic but show a positive effect on β-catenin signalling at residue 35, which could be explained by their small side chain not being able to make Van der Waals interactions. Amino acid replacements at the variable position H36 (see **Figure 5.8C**), with the exception of proline, do not affect β-catenin signalling. The behaviour of substitutions at these sites in our dataset therefore confirms and extends previously published work on β-TrCP (Wu *et al.*, 2003) and suggests that our dataset provides sufficient power to assess the effect of individual amino acid substitutions.

5.2.12. Serine and threonine are interchangeable at the phosphorylation sites of β-catenin

As CKI and GSK-3β can phosphorylate proteins at both serine and threonine residues (in the case of GSK-3β only adjacent to a phosphorylated serine or threonine four residues towards the C-terminus) (Cohen, 1986; Tuazon, 1991; Fish *et al.*, 1995), it was hypothesised that these two amino acids should be interchangeable without affecting the activity of the TCF-LEF:H2B-GFP reporter (see **Figure 5.9A**). The frequency of serine/threonine substitutions at the phosphorylation sites are generally depleted in the bins with higher activity, indicating that these mutations do not enhance β-catenin signalling. S33T and S37T show highest activity in P3 and P4 rather than in P2, which could indicate a decreased binding affinity for β-TrCP. In

addition, these mutations show an elevated occurrence in P7 whilst depleted in P6. We do not have an immediate explanation for this unexpected occurrence and hypothesise that the hike in P7 could be due to fluctuations as a result of the low number of cells that was sampled from this bin (table 5.1). Despite this caveat, the data nonetheless suggests that serine/threonine substitutions at the phosphorylation sites do not disrupt CTNNB1-mediated signalling, presumably because the ability of CKI or GSK-3β to add a phosphate group to these sites is preserved.

5.2.13. The behaviour of phosphomimetics is context dependent

Some amino acids show chemical resemblance to a phosphorylated serine or threonine and replacement of a phosphorylation site with such a phosphomimetic is thought to mimic a constitutively phosphorylated residue. Aspartic acid (D) and glutamic acid (E) are the best known phosphomimetics for both serine and threonine residues (Hiscott *et al.*, 1999; Konson *et al.*, 2010) and would, in the case of β-catenin, theoretically lead to a continuous turnover of the protein and subsequent low steady-state GFP levels. Our dataset provides an opportunity to determine how effectively these phosphomimetic substitutions operate and to test their relative efficacy at different positions in a single phosphodegron motif.

Our data show that aspartic acid (D) leads to an elevated (or in the case of S45D a partially elevated) transcriptional activity at all four phosphorylation sites (see **Figure 5.9B,C**). This is inconsistent with phosphomimetic function. Glutamic acid at site S33 is most abundant in the bins of lower activity (P2-P4) suggesting that it mimics a phosphorylated serine at this site. As this trend is not seen at the other three phosphorylation sites, it shows that the behaviour of phosphomimetic substitutions is context dependent.

5.2.14. Mutational effect scores: Single metrics for effects of substitutions on β-catenin activity

Whilst the relative abundances of codon sequences across each of the six bins allow for an initial assessment of effect size for a small number of mutations on β-catenin activity, six separate values per substitution hampered any systematic comparisons of phenotypic effects between different codon sequences. It was therefore examined whether a single functional score could be assigned to each codon substitution that reflected its enrichment

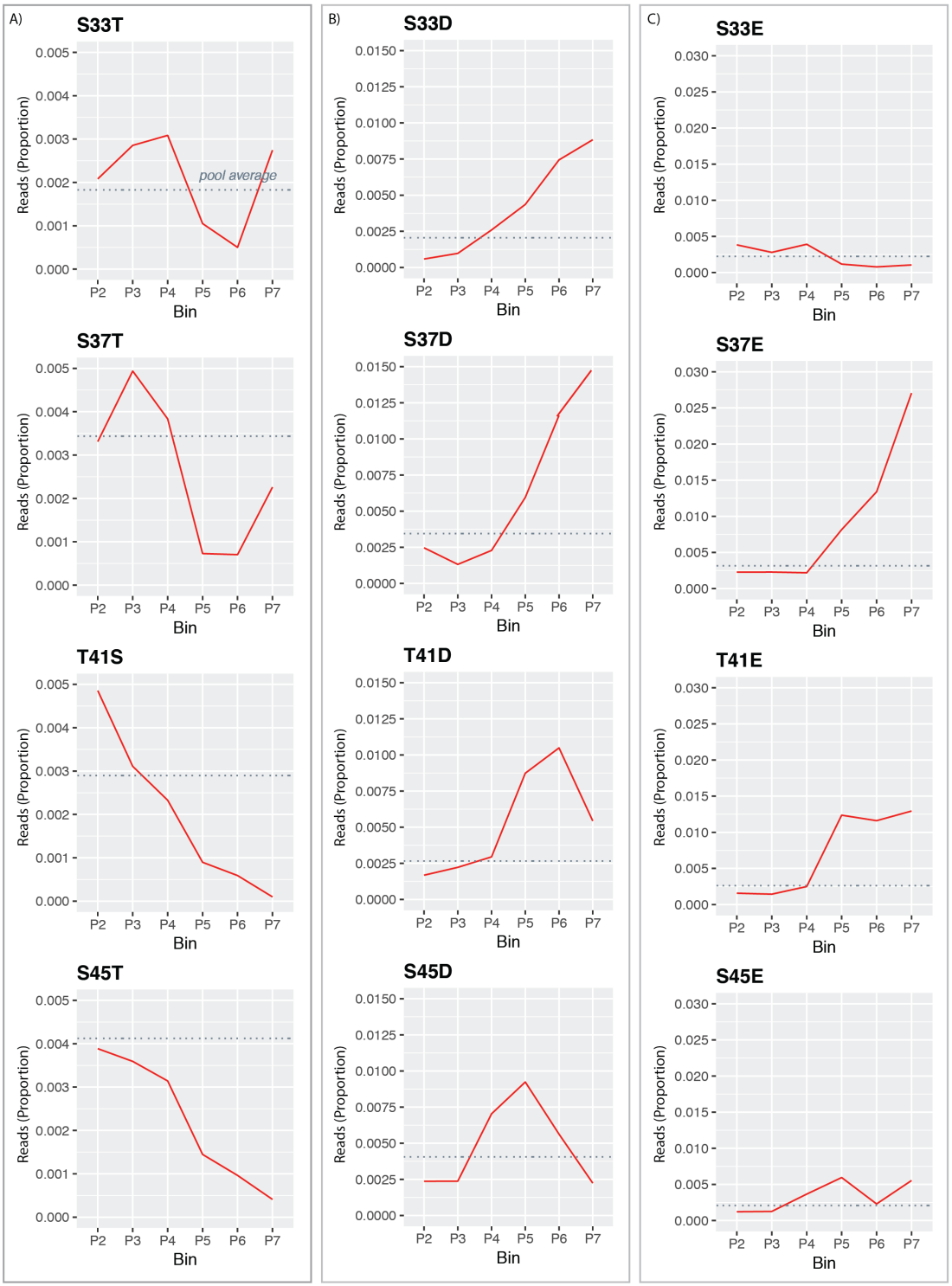


Figure 5.9 Behaviour of serine, threonine and phosphomimetics across phosphorylation sites (A) Reciprocal replacement of serine and threonine at phosphorylation sites shows a steady decline towards bins with higher GFP fluorescence, indicating that these changes do not affect the phosphorylation of β-catenin. P7 shows an interruption of this trend for S33T and S37T. **(B)** Behaviour of phosphomimetic aspartic acid (D) at phosphorylation sites. The enrichment of these frequencies in the higher populations indicates that aspartic acid does not mimic phosphorylation at these sites. **(C)** Behaviour of phosphomimetic glutamic acid (E)

at phosphorylation sites. Glutamic acid enhances β-catenin activity at sites S37, T41 and S45 but not at S33E, suggesting that the behaviour of this amino acid as a phosphomimetic is context dependent.

or depletion across all six sorted bins relative to the unsorted pool. To achieve this, a previously published equation that calculates mutational effect scores from codon frequencies in each FACS-purified bin based on GFP fluorescence levels was adapted (Kosuri *et al.*, 2013).

To estimate mutational effect scores, we first calculated, for each codon *i* and each bin *j*, the normalised fractional contribution of each bin per codon a_{ij} using the formula

$$a_{ij} = {f_j \cdot c_{ij}} \Big/ {\sum_j f_j \cdot c_{ij}} \tag{5.2}$$

where f_i is the fraction of cells harboured in each bin and c_{ij} is the number of amino acids *i* in each bin *j*. For a given substitution, $\sum_i a_{ij} = 1$ (i.e. each codon has one a_{ij} per bin, and these six values together add up to 1).

The final mutational effect score, E_i, was calculated as:

$$\mathrm{E_i} = \exp\left[\sum_j a_{ij} \log(m_j)\right] \tag{5.3}$$

A heatmap based on mutational effect scores provides a straightforward way to visualise which residues are most sensitive to perturbations (see **Figure 5.10A**). Reassuringly, this highlights the phosphorylation sites (S33, S37, T41 and S45) and residues comprising the β-TrCP binding motif (D32, G34, I35 and, to a lesser extent, H36). The mutational effect score hence provides a reliable metric with which to compare the relative effect of each amino acid substitution on β-catenin activity.

A histogram was plotted to show the distribution of all calculated effect scores, which revealed a bimodal spread across the sampled codon substitutions (see **Figure 5.10B**). In order to classify the different amino acid substitutions based on their effect score, the scores were divided into two categories based on where the slope of the curve approximated 1 (see **Figure 5.10B**, by eye), binning the majority of the mutations to not have an effect on β-catenin activity (212; 62%), whereas approximately a third enhanced signalling activity (130; 38%) (see **Figure 5.10C**).

5.2.15. Different mutations in CTNNB1 are found in specific cancer types

Previous studies have suggested that cancers arising from distinct tissues have mutations at specific residues in the CTNNB1 N-terminus (Polakis, 2000; Albuquerque *et al.*, 2011), while even different subtypes of liver cancer were previously shown to harbour distinct amino acid substitutions (Austinat *et al.*, 2008; Rebouissou *et al.*, 2016). To ascertain the extent of differences between tumour types, all tumours with mutations in codons 31-48 of CTNNB1 were identified across 9 different cancer types from the Catalogue of Somatic Mutation in Cancer (COSMIC) (Forbes *et al.*, 2018) (see **Figure 5.11A**). COSMIC was the largest database for missense mutations in cancer at the time of writing.

Plotting the fraction of tumours with a mutation at each position, broken down by tissue-of-origin, supports the hypothesis that different tumour types preferentially harbour CTNNB1 mutations at different N-terminal residues. For instance, whereas cancers from adrenal gland (in adrenal gland vs. other, Pearson's Chi-squared $p = 1.41 \times 10^{-12}$), kidney ($p = 2.2 \times 10^{-14}$) and skin ($p = 0.0201$) predominantly harbour mutations in S45, mutations in soft tissue cancers are predominantly found in T41 ($p = 2.2 \times 10^{-16}$).

Endometrial and liver cancer contrastingly show a much wider distribution of mutated residues. When focussing on the most frequently mutated residue across all cancers, S45, variance between cancer types is less apparent with the majority of cancer types predominantly harbouring a phenylalanine (see **Figure 5.11B**). However, it is notable that although many cancers harbour a tyrosine at residue 45, this is completely absent in soft tissues despite the relatively high number of samples (N=135, $p = 4.46 \times 10^{-4}$). These analyses

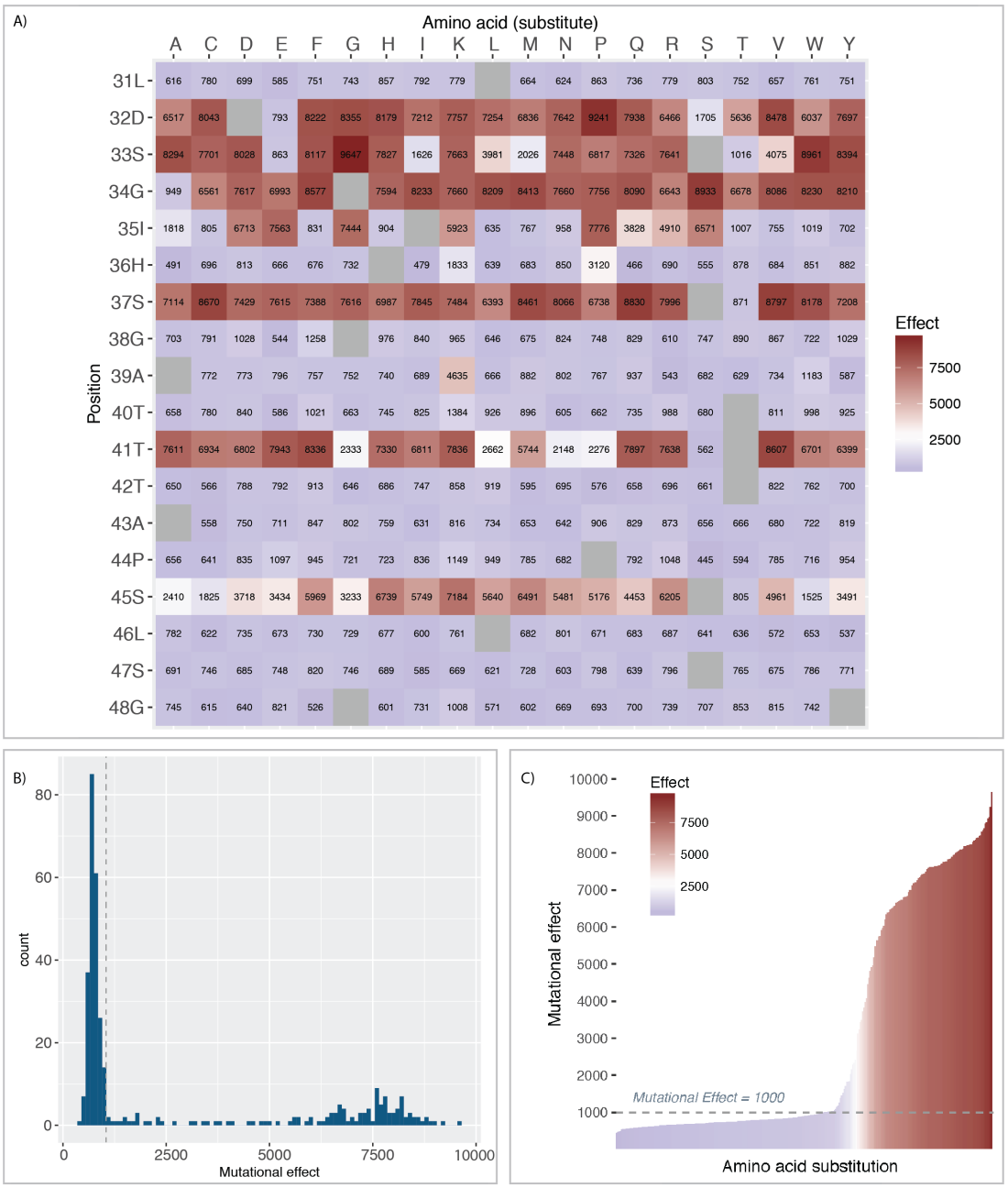


Figure 5.10 Mutational effect score per amino acid substitution (A) Heat map of mutational effect score per amino acid substitution. Amino acid substitutions at residues 32D, 33S, G34, S37, T41 and S45 show the highest effect on enhancing β-catenin activity. Scale runs from blue indicating the lowest (445), to white indicating median (2500) to red indicating the highest (9647) score. **(B)** Histogram of frequencies of mutational effect scores in bins with a size of 100. Most amino acid substitutions have a mutational effect score < 1000, indicating no positive effect on β-catenin signalling. **(C)** Bar plot of mutational effects scores sorted from high to low. Mutations are categorised based on their scores into no effect (N=212) and positive effect (N=130) using cut-off at 1000 where the slope of the curve approximates 1 (by eye).

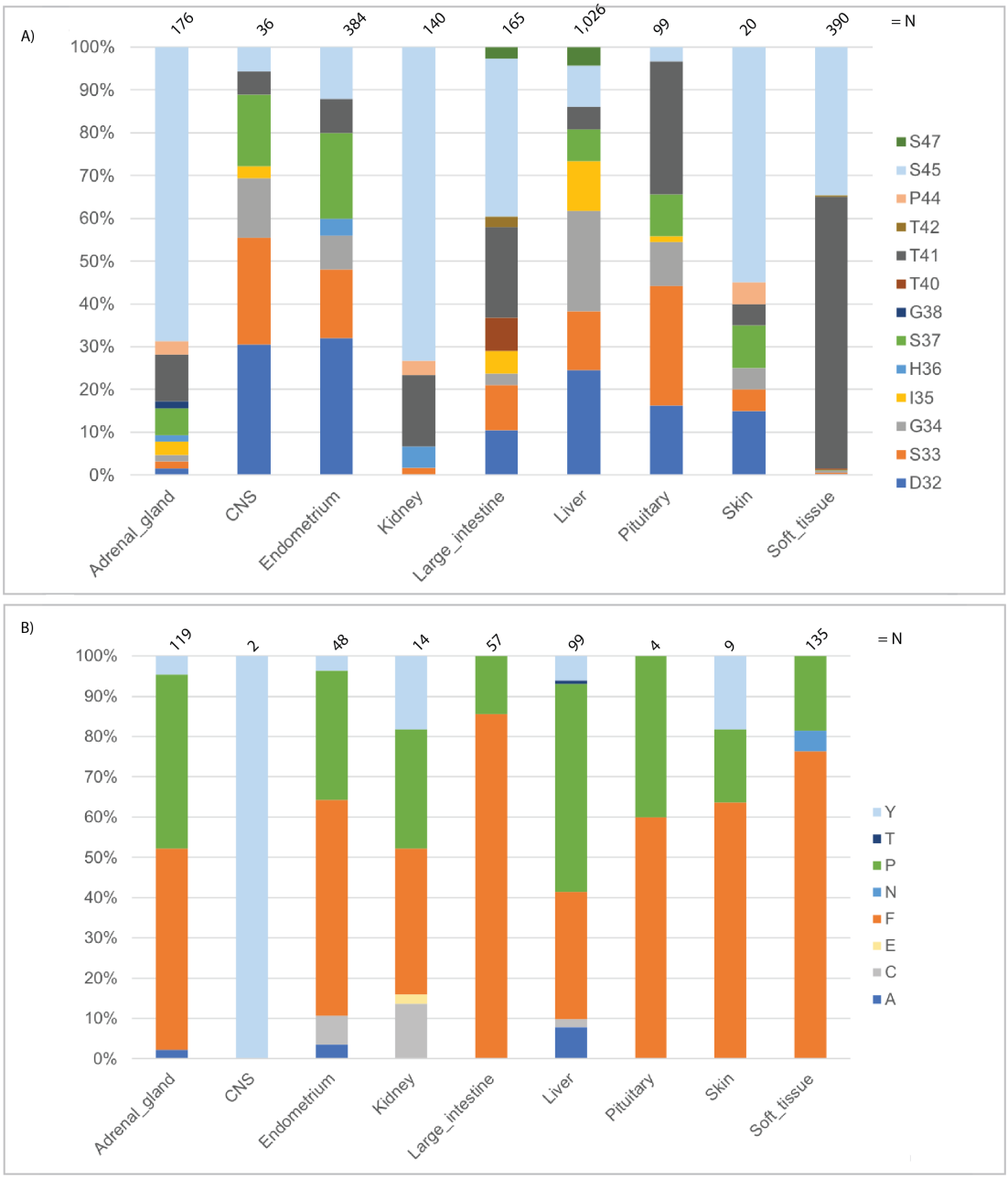


Figure 5.11 Frequency of residues and mutations in CTNNB1 found across cancers (A) Distributions of mutated residues across 9 cancer types show that different cancer types harbour mutations in specific residues. Cancers from adrenal gland, kidney and skin tissues predominantly harbour mutations in S45 (light blue), while mutations in soft tissue cancers are predominantly found in S37 (green). Endometrial and liver cancers show a much wider distribution of mutations, with S45 mutations completely absent in the former. Sample sizes are denoted at the top of each distribution, whereas the legend on the right indicates the colour indexes. **(B)** Distributions of amino acids found at residue 45 across the same cancer types as in A show that variance between cancer types is less apparent and most cancer types most often harbour a phenylalanine (F, orange) followed by proline (P, green). However, notable is that whereas many cancers harbour a tyrosine (Y, light blue) regardless of tissue

type, this amino acid is completely absent in soft tissues despite the relatively high number of samples (N=135).

confirm that different tumour types preferentially harbour different amino acid substitutions within the CTNNB1 N-terminal region.

The observed difference in mutated residues between cancer types could theoretically result from either tissue-specific dissimilarities in selection, i.e. selection for effects on Wnt signalling that are optimal for tumorigenesis in that tissue, or tissue-specific differences in mutational signatures that cause particular amino acid substitutions to arise more frequently in certain tissue contexts. The extent to which these two influences contribute to the overall pattern of mutation is not clear. Given that the mutational effect scores provide a metric of Wnt signalling strength, it was assessed whether predominant mutations in each tumour type were characterised by different mutational effects.

To assess whether different tumours preferentially harboured mutations of particular effect size, I next plotted the distribution of mutational effect scores extended to all 31 tissues in which at least one mutation in residues 31 – 48 was found (see **Figure 5.12**). Whereas the distribution of all possible amino acid substitutions (left column) shows that most mutations do not affect β-catenin activity (average mutational effect score of 2730; see **Figure 5.12**, most left column), tumour samples show a clear enrichment for high mutational effect scores. This suggests that cancers preferably contain mutations that enhance Wnt/β-catenin signalling.

Next, inter-tumour differences in β-catenin activity were ascertained. Whereas tumours in the central nervous system (CNS) and endometrium predominantly carry mutations with very high effect score (averages of respectively 7407 and 7436), other tissues such as large intestine (5359), liver (6679) and soft tissue (6921) harbour mutations with high effects on β-catenin activity, whereas stomach (1212) and kidney (4857) tumours show no high activity mutations, although these could be explained by their small sample sizes and that these tumours are not driven by these mutations. Statistical analyses furthermore show that some tumours with larger sample sizes such as pituitary gland and soft tissue are significantly different (average mutational effects of 7387 and 6921, unpaired t-test with Bonferroni-corrected p-value = 3.26×10^{-5}) (see

Table 5-3), suggesting that different tumour types harbour mutations with different levels of activity.

5.2.16. Coupling mutational effect scores and mutational probability of amino acid substitutions

The mutational effect scores provide a quantitative measure for the effect an amino acid substitution in *CTNNB1* has on TCF-LEF signalling in ESCs. While acknowledging that the same mutation could conceivably influence the reporter in a distinct way in a different cell type, these experiments are done under the assumption that it can be used as a proxy to infer the functional effect (i.e. what is being "selected" for) in tumours. In the context of tumorigenesis, the availability of amino acid changes for selection depends upon the probability of such amino acid changes arising through nucleotide changes. With this in mind, we wanted to determine the extent to which the different patterns of amino acid substitutions observed in different tumour types, could be attributed to different mutational processes influencing the likelihood these mutations to arise. Therefore, the probability of each amino acid substitution to arise in tumours that specifically harboured a mutation in *CTNNB1* was calculated.

To assign a value to the probability for a given amino acid substitution to occur at each residue, I first assessed the nucleotide changes necessary to achieve an amino acid at a given position. Whilst the amino acid sequence for the targeted region is fully conserved between human and mouse, synonymous substitutions mean that their codon sequences differ at 6 out of 54 nucleotides. As the clinical data is available for human, the human coding sequence was used as the reference sequence.

As discussed in the introduction (section 5.1.5) it is well established that mutational signatures differ between tissues (Alexandrov and Stratton, 2014), meaning that different nucleotide changes are likely to occur at different rates between tumour types. In order to account for this mutational bias, publicly available data from primary tumours harbouring mutations in *CTNNB1* was used to derive these values using the trinucleotide context approach pioneered by the Stratton laboratory (Alexandrov *et al.*, 2013).

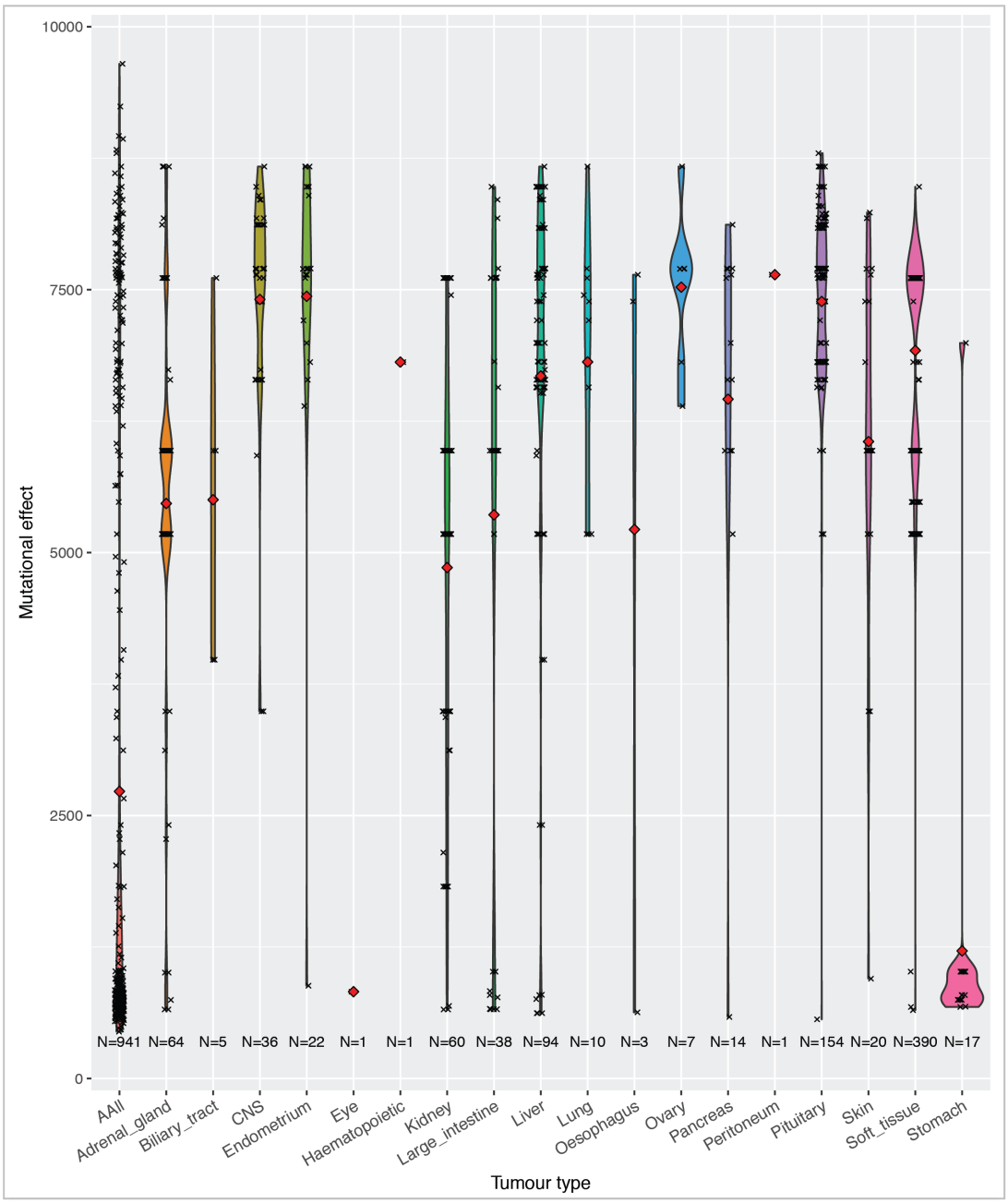


Figure 5.12 Distribution of mutational effects as a result of amino acid substitutions across different cancers Plots were depicted for all cancers types from which primary tumour data was available in COSMIC. A distribution of mutational effect scores of all possible amino acid substitutions ('All', left-most plot) shows that most mutations do not have a positive effect on β-catenin signalling. Distributions of mutational effect scores across tumour types show that tumours most frequently harbour mutations with a high mutational effect score, but that different tumours harbour different missense mutations resulting in specific levels of signalling. Mutations found in central nervous system (CNS) and endometrium are associated with high mutational effect scores, whereas soft tissue cancers harbour mutations with intermediate mutational effect.

Table 5-3 Differences between average mutational effect scores in cancer types Significance levels are shown for those samples between which the mutational effect scores were significant and sorted by the p-value (unpaired t-test), which was Bonferroni-corrected. Levels of significance are indicated by * (0.05), ** (0.01) and *** (0.001). Mean and sample size (N) are indicated for sample 1 (1) and sample 2 (2).

Sample 1	Sample 2	P-value (Bonferroni)		Mean (1)	N (1)	Mean (2)	N (2)
Stomach	Endometrium	1.88E-12	***	1213	159	7437	384
Stomach	CNS	4.38E-12	***	1213	159	7407	252
Pituitary	Kidney	5.74E-12	***	7387	99	4857	140
Liver	Stomach	3.41E-11	***	6679	1026	1213	159
Stomach	Pituitary	5.49E-10	***	1213	159	7387	99
CNS	Kidney	5.89E-10	***	7407	252	4857	140
Stomach	Ovary	6.39E-10	***	1213	159	7524	138
Stomach	Soft tissue	2.94E-09	***	1213	159	6921	1408
Pituitary	Adrenal gland	3.76E-09	***	7387	99	5467	178
Stomach	Adrenal gland	6.15E-09	***	1213	159	5467	178
Soft tissue	Kidney	7.26E-09	***	6921	1408	4857	140
Stomach	Skin	2.01E-08	***	1213	159	6055	133
Stomach	Lung	6.76E-08	***	1213	159	6813	47
Stomach	Kidney	1.94E-07	***	1213	159	4857	140
Large intestine	Stomach	2.89E-07	***	5359	165	1213	159
CNS	Adrenal gland	8.25E-07	***	7407	252	5467	178
Stomach	Pancreas	1.53E-06	***	1213	159	6458	133
Liver	Kidney	2.06E-05	***	6679	1026	4857	140
Pituitary	Soft tissue	3.26E-05	***	7387	99	6921	1408
Endometrium	Kidney	3.58E-05	***	7437	384	4857	140
Ovary	Kidney	1.19E-04	**	7524	138	4857	140
Adrenal gland	Endometrium	3.24E-03	**	5467	178	7437	384
Adrenal gland	Ovary	3.65E-03	**	5467	178	7524	130
Large intestine	Pituitary	6.94E-03	**	5359	165	7387	99
Large intestine	CNS	1.19E-02	*	5359	165	7407	252
Liver	Adrenal gland	1.97E-02	*	6679	1026	5467	178
CNS	Soft tissue	0.019744	*	7407	252	6921	1408
Large intestine	Ovary	2.41E-02	*	5359	165	7524	138

A given amino acid may be encoded by up to six different triplets, which each differ at up to three nucleotides from the reference codon, maximising the number of combinations of nucleotide changes per amino acid substitution to 36 (i.e. (3 * 2 * 1 =) 6 paths for 6 codons,

see **Figure 5.13A-C**). In some cases, multiple nucleotide changes are required to yield a single codon change paths or multiple codon changes can lead to the same amino acid change. Double and triple nucleotide polymorphisms can theoretically arise from independent sequential lesions or from a single event resulting in the change of multiple nucleotides and may hence be more complex than the sum of individual mutations (Rosenfeld, Malhotra and Lencz, 2010). However, as we had no insight into the extent to which nucleotide substitutions co-occur, it is assumed that all codon changes are the result of independent successive single nucleotide changes. This resulted in an overview (see **Table s5.1**, available via https://uoe-my.sharepoint.com/:x:/g/personal/s1478317_ed_ac_uk/EWFH4G4WPDlDnhqppDlPh0IBQEBGtFKnXVwm4Rebs70W6Q?e=Sg570S) of all the possible codon replacement 'paths' and their respective single nucleotide substitution 'steps' in their trinucleotide context for each analysed codon in *CTNNB1*.

Ailith Ewing extracted whole exome-based somatic mutation data from The Cancer Genome Atlas (TCGA) Research Network (https://cancergenome.nih.gov/, 01-02-2018) from the two largest databases that harboured mutations in *CTNNB1*, i.e. liver hepatocellular carcinoma (TCGA-LIHC, N = 74 tumour exomes) and uterine corpus endometrioid carcinoma (TCGA-UCEC, N = 103 tumour exomes). From this data, Ailith calculated the observed frequency for all 64 possible nucleotide changes (relative to the human reference genome sequence) in their trinucleotide context as described in detail in Material & Methods.

As mentioned, I denoted each single nucleotide change (step) in its trinucleotide context, i.e. in the context of the nucleotides present at its 3' and 5' end (see **Figure 5.13A-C**). Knowing the possible paths to each codon change, the calculated background probabilities for each step within a path were then multiplied together to ascertain the probability of a certain codon substitution. After calculating probability values (*P*) for each individual path, the probabilities of all individual paths were combined to calculate the overall probability of an amino acid substitution to occur, which would require at least one of the *n* paths to that amino acid to occur (see **Figure 5.13D**). The probability of *at least* one of the paths to occur and the probability of none of the paths to occur should sum up to 1,

$$P_{at\ least\ one\ path\ occurring} + P_{none\ of\ the\ paths\ occurring} = 1 \qquad (5.4)$$

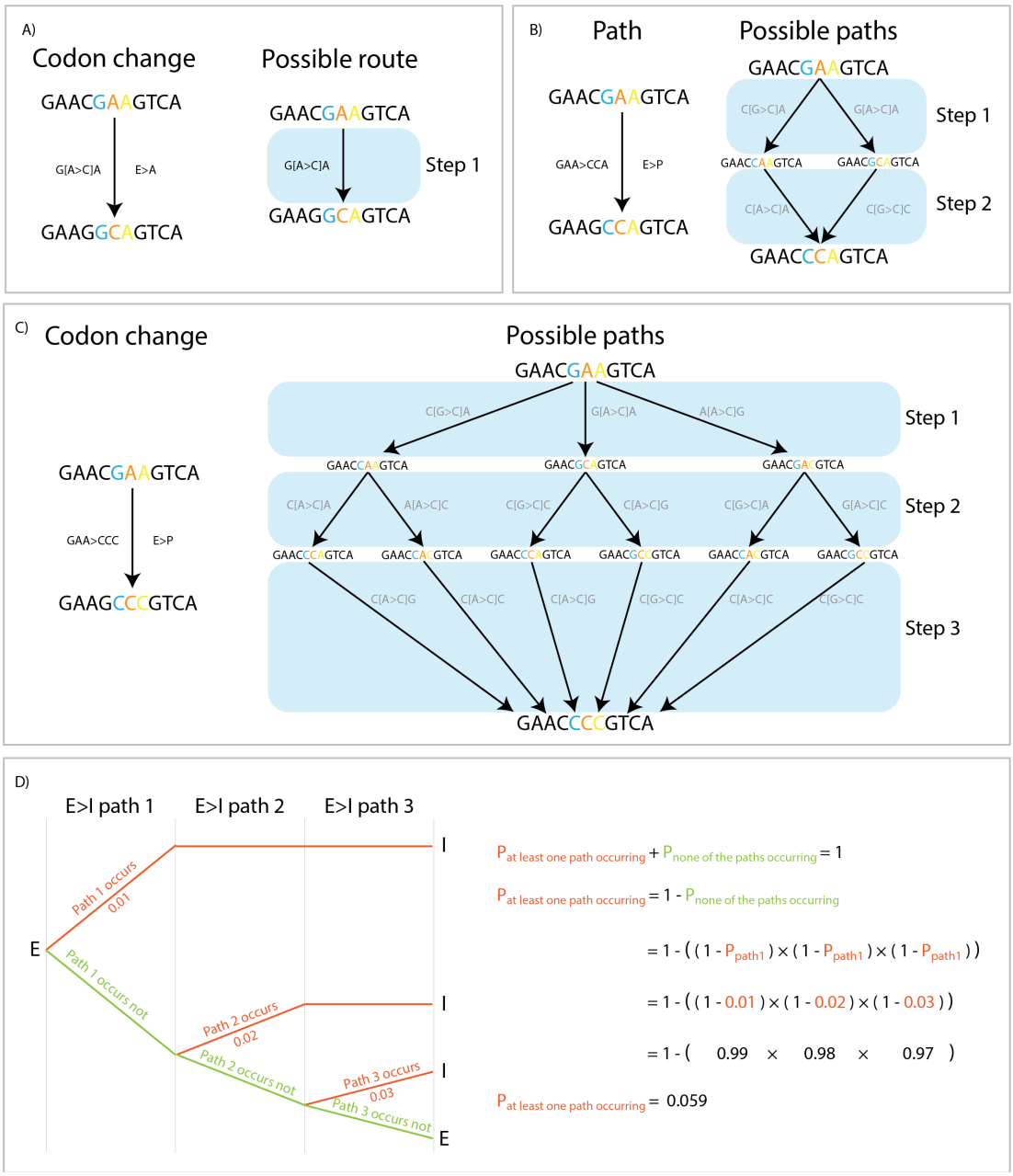


Figure 5.13 Calculating mutational signatures and amino acid substitution rates in tumours harbouring a CTNNB1 mutation All possible paths of single nucleotide substitutions to achieve each possible alternative codon were determined as exemplified in this panel. In the figures, G[A>C]A indicates an A substituted by a C in a GAA trinucleotide context. **(A)** In some cases, an alternative codon (in this case GCA) differs only by a single nucleotide from the reference (GAA) and can hence by achieved by only a single path of a single nucleotide change (step; in this case G[A>C]A). **(B)** Other codons may differ from the reference by two nucleotide changes (CCA), which can be achieved by two paths, which each consists of two steps. **(C)** Finally, a codon may differ at all three nucleotides from the reference requiring three individual amino acid substitutions and can be achieved by six possible routes. **(D)** The probability of a path to occur is calculated by multiplying the probabilities of each mutation in that path. To combine all possible paths to arrive at a single likelihood score per codon change, we calculated the probabilities of any of the paths to occur (sufficient for that codon

substitution to arise) as depicted in figure 5D. Here, three possible paths with exemplary probabilities of 0.01, 0.02 and 0.03 would together lead to an eventual probability of 0.059. As several amino acids are encoded by different codons, probabilities of the individual codons were combined by a similar rationale as in this figure.

, which means that the probability of any one of the paths to occur equals to:

$$P_{at\ least\ one\ path\ occurring} = 1 - P_{none\ of\ the\ paths\ occurring} \quad (5.5)$$

The probability of none of the paths occurring equals to the product of the probabilities of *n* paths to not occur, i.e.:

$$P_{at\ least\ one\ of\ the\ paths\ occurring} = 1 - ((1 - P_{path1}) \times \dots (1 - P_{path\ n})) \quad (5.6)$$

Using this equation, the probability of each codon change was determined (as exemplified in **Figure 5.13D**) occurring in respectively liver and endometrial tumours using clinical data. As multiple codon changes could lead to the same amino acid change, the probabilities of different codon changes were combined in exactly the same manner as the paths (see **Figure 5.13D**) to yield the probability of an amino acid change.

These calculations were performed for each possible amino acid substitution at each of the 18 target residues, resulting in 342 probability scores for respectively endometrial and hepatocellular carcinoma harbouring a mutation in *CTNNB1*. The probability scores for both cancer types poorly correlate (R^2 = 0.243), demonstrating that there are substantial differences in mutational likelihood between tumour types.

5.2.17. Mutations found in cancers not consistently explained by probability and β-catenin activity

The combination of mutational effect scores and amino acid probabilities then allowed us to determine the extent to which either of these factors explain the amino acid changes seen in cancers. To visualise their particular weights, I extracted amino acid substitution frequencies from the same tumour databases from which the probabilities were calculated (TCGA) and combined these values in a single plot per tumour type (see **Figure 5.14**).

All amino acid changes observed in both liver and endometrial tumours are those that have a high likelihood of occurrence based on background substitution rates, and an intermediate to large effect on β-catenin signalling. No mutations are found that have a low mutational effect score (<0.23, see **Figure 5.10**), despite this group comprising the majority of possible substitutions. Similarly, no mutations are observed with a very low probability (i.e. $< 3 \times 10^{-4}$). These results indicate that neither the background mutation rate or selectional drive alone accounts for the observed pattern of mutations in *CTNNB1*, which is also supported by the similar distributions in mutational scores and likelihood in **Figure 5.15**.

Several residues are however not consistent with this trend. Substitutions including H36P (N=5) and S45P (N=11), which have a high likelihood of occurrence in both tumour types, are observed in liver hepatocellular carcinoma, yet are absent in the analysed endometrioid carcinomas (Fisher's exact test, for H36P and S45P respectively $p = 0.0118$ and $p = < 0.0001$), raising the possibility that these mutations could favour clonal outgrowth in the liver but not in the endometrium. Several incidences of S45P in endometrioid carcinomas in the larger COSMIC database were observed (see **Figure 5.11** and **Figure 5.12**). As the mutational effect score distributions were furthermore broadly similar between both the endometrial cohorts (7749 ± 304 (N = 103) vs. 7385 ± 228 (N = 487)) and liver cohorts (6790 ± 779 (N = 79) vs. 6682 ± 354 (N = 1105)), sample data from clinal samples from COSMIC was included, under the assumption that the likelihood scores calculated with the smaller sample sizes can be extrapolated to other tumours from the same type (see **Figure 5.16**).

The larger datasets show that virtually all mutations that have a large mutational effect and high likelihood are found in both endometrial and liver cancer. Several mutations at residue G34, an essential residue in the β-TrCP binding motif, (e.g. G34I, G34S, G34K, G34D) are contrastingly absent in both databases whilst falling in the same range as many other mutations that are occurring in endometrial cancer. H36P, which is relatively frequent in hepatocellular carcinoma (43/1105) despite a modest effect score (3120), was completely absent from endometrial cancer (0/487, Fisher's exact $p = < 0.00001$). This is consistent with the broader range of effect scores for mutations found in hepatocellular carcinoma compared to endometrium. We speculate that H36P may not be sufficiently potent to drive endometrial carcinogenesis. However, the possibility that the effect size of H36P measured

in mESCs (Figure 5.10) does not accurately reflect the effect size in both endometrial cells and hepatocytes.

Three mutations with intermediate effect scores, H36P, S45P and T41A, were highly enriched in liver over endometrial tumour samples (chi-squared $p = 1.87 \times 10^{-9}$, 1.44×10^{-14} and 1.64×10^{-7}), whilst S37F and S37C, both with mutational effect scores above 8500, were more highly enriched in endometrial cancer ($p = 1.18 \times 10^{-11}$ and 1.13×10^{-7}, i.e. higher in endometrium). Supporting these findings, four additional mutations with mutational effect scores below 7000 were statistically enriched in liver (I35S, S45F, G34V, S45A) and three others with mutational effect scores > 0.68 were enriched in endometrium over liver (T41I, G34E, S33F) This demonstrates that different cancer types harbour different mutations that are largely driven by mutational effect score and not by mutational likelihood.

Notably, the larger sample sizes reveal many mutations with low mutational effect but high likelihood of occurrence (i.e. bottom right of the plot) occurring at a low frequency (see **Figure 5.16**). The mutational pattern indeed proved to be the result of both of these factors (multiple linear regression, p-value: $< 3.5 \times 10^{-21}$ and 7.9×10^{-16} for liver and endometrium, respectively). However, neither likelihood or mutational effect alone can account for the mutational pattern in liver (Kruskal-Wallis $p = 0.5914$ and $p = 0.253$, respectively) and endometrial cancer (Kruskal-Wallis $p = 0.5440$ and $p = 0.253$, respectively) cancer and are both not different between liver and endometrial tumours (**Figure 5.15**). This hence provides comprehensive evidence for the theory that patterns of mutations found in tumours arise from natural selection acting on the available genetic variation. In addition, the data more tentatively suggest that selective constraints can vary between tissues.

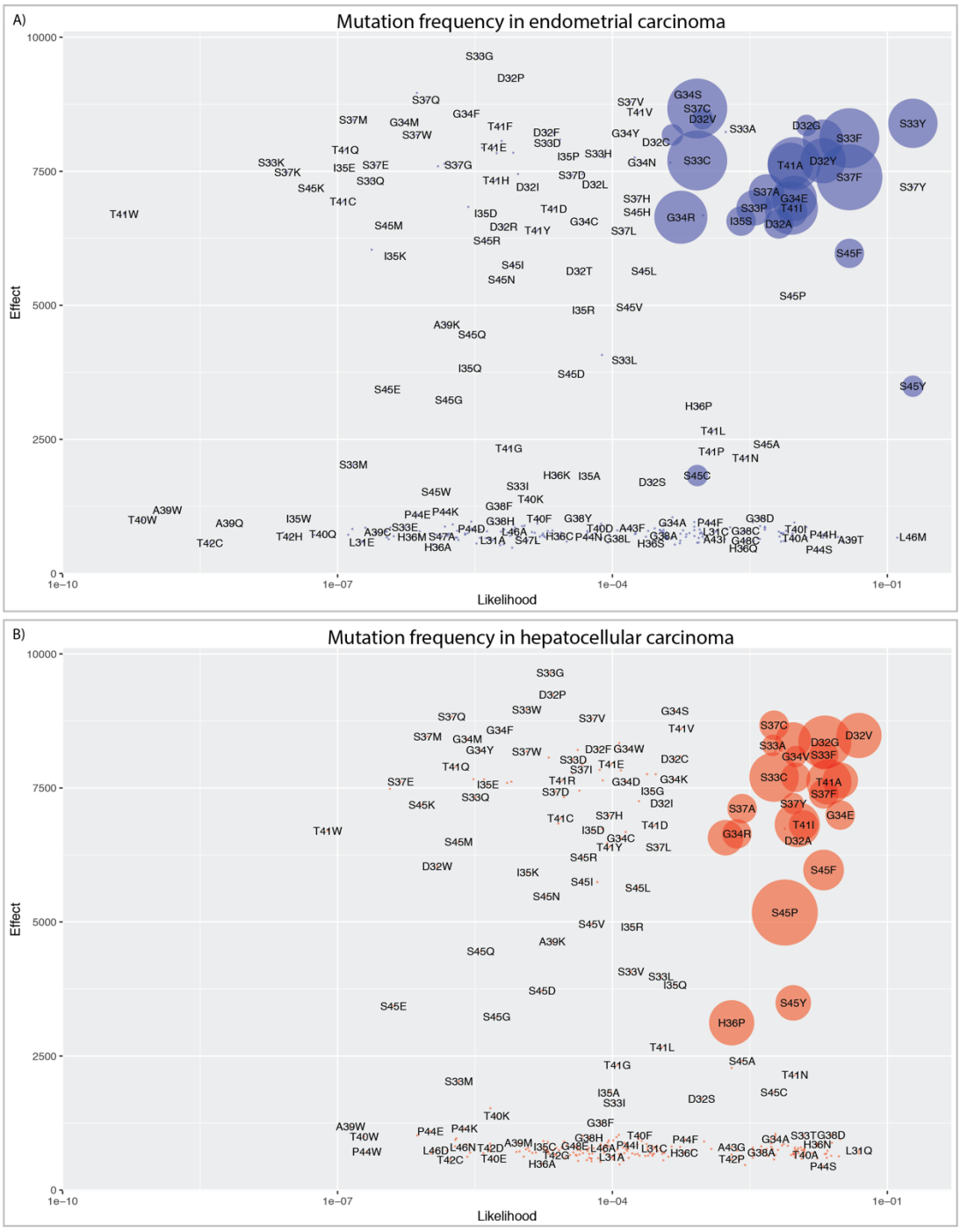


Figure 5.14 Frequency of amino acid substitutions as a result of their probability and mutational effect score in endometrial and hepatocellular carcinoma Amino acid substitutions were plotted as a result of their probability (calculated from whole exome sequencing data) and their mutational effect scores (determined from our assay) for endometrial carcinoma **(A)** and hepatocellular carcinoma **(B)**. The frequency of amino acid substitutions in tumours is indicated by the size of the circle, as a proportion of the total frequency of missense mutations found in that database (endometrial carcinoma N=103, hepatocellular carcinoma N=74), with the largest circle in each plot represents 11 tumours (liver hepatocarcinoma S45P, endometrial carcinoma S37F). All mutations occurring in both

tumour types have high probabilities and medium to high mutational effect scores. As was shown in figure 5.10, endometrial tumours show a higher average mutational score whereas liver tumours show a spread of tumours. Some substitutions with a high probability and high mutational effect score, e.g. S36Y, G34I and S45P, are not found in endometrial cancer but are frequently observed in hepatocellular tumours. Vice versa, S37P and D32A are solely observed in endometrial cancer based on these two databases.

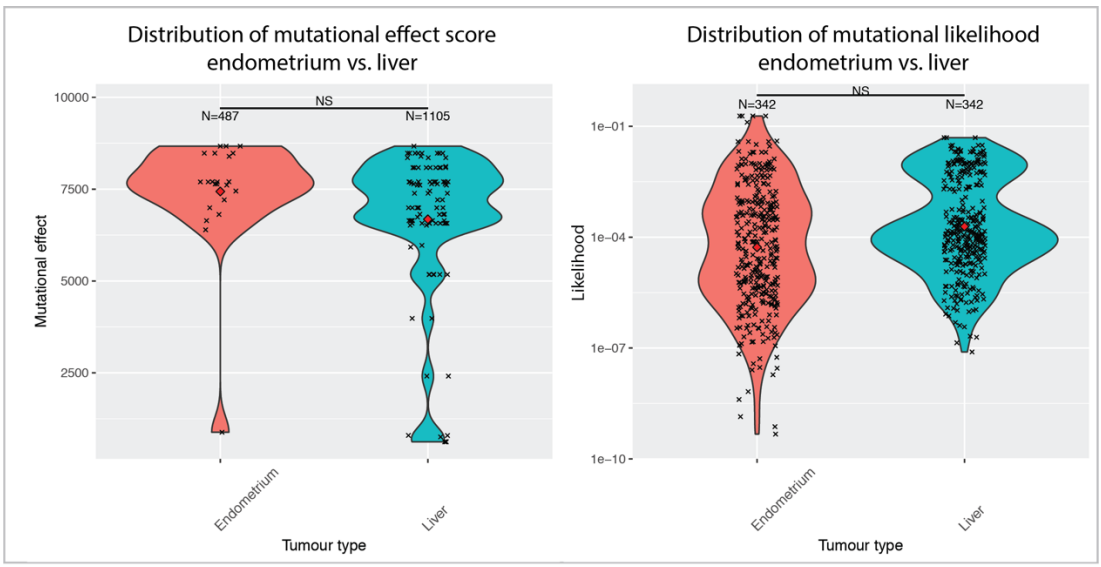


Figure 5.15 Distribution of mutational effect scores and mutational likelihood across endometrium and liver cancer Using the same samples as used in 5.14, distribution plots show similar distributions across both cancers types (left) and mutational likelihood (right) (NS = not significant). Differences in mutational effect score and mutational likelihood between endometrial and liver tumours are not significant.

Figure 5.16 Frequency of amino acid substitutions as a result of their probability and mutational effect score in endometrial and hepatocellular carcinoma, including COSMIC Similar to plot 5.14, amino acid substitutions were plotted as a result of their probability (calculated from a subset of samples and their mutational effect scores for endometrial carcinoma **(A)** and hepatocellular carcinoma **(B)**. The frequency of amino acid substitutions in tumours is indicated by the size of the circle, as a proportion of the total frequency of missense mutations found in that database (endometrial carcinoma N=487, hepatocellular carcinoma N=1105), with the largest circle in each plot represents 149 tumours in liver (T41A) and 60 in endometrium (S37F). The additional samples in this plot as compared to 5.14 did

not change the distribution, whilst revealing some mutations with a low mutational effect but high likelihood, suggesting that the mutational likelihood has a larger influence on the eventual mutational pattern than the mutational effect.

5.3. Discussion

In this chapter, I analysed experimental data from work performed in the Hohenstein lab and assessed whether amino acid substitutions can be introduced into *CTNNB1* and their effects on β-catenin signalling could be functionally assessed. Initially, I have shown that all possible amino acid substitutions can be efficiently integrated over a range of 18 different codons. Secondly, I have shown the high reproducibility between biological replicates and the combined bins and pool demonstrated the robustness of the editing procedures conducted by the Hohenstein lab (section 5.2.8). I demonstrate that a single mutational effect score can be derived per amino acid substitution and assess the phenotypic consequences of all possible amino acid substitution at a given site (section 5.2.14). The probability of an amino acid substitution occurring in different tumour types was calculated and combined this with the mutational effect score to show that both of these factors are required to explain mutational patterns observed in different tumours (section 5.2.16). This study demonstrates that deep mutational scanning can be applied to study the effects of amino acid changes in disease genes in their endogenous genomic context.

5.3.1. Genomic variants are efficiently integrated from plasmid into genome

While codon variants at 2 out of 20 residues (K49 and G50) were shown to be underrepresented in both the plasmid and genomic samples (section 5.2.7), our data confirmed that all variants at the other sites found in the plasmid are present at approximately equal proportions in the genomic pool, indicating that variants were incorporated from HDR repair templates in an unbiased manner. The reconstruction of the pool from the individual samples and the high correlation with the unsorted pool additionally shows that heterogeneity is not lost by binning the samples (section 5.2.8).

Equal representation of codons in the plasmid and unselected cell pool furthermore implies that there is no competitive selection for genomic variants during the culture period. By culturing cells in the presence of GSK-3β inhibitor until two days prior to analysis by flow cytometry, individual clones will not have sufficient time to establish substantial differences

in clonal expansion while it allows for β-catenin levels to stabilise and alter GFP expression accordingly.

5.3.2. Separating cells into multiple bins allows for better assessment of mutational consequences

In the behaviour of codons across the bins, P7 frequently exhibited outliers. In addition, this bin showed the lowest correlation between the biological replicates (R^2 = 0.653, section 5.2.8). Since the sequencing depth in each sample is consistently exceeding a 1000X for each codon, the sequencing should not introduce any bottleneck that could explain these discrepancies. With P7 having the lowest number of sorted cells (respectively 4,075 and 4,234, see **Table 5-1**), on average a maximum of 10 cells were sampled per amino acid substitution. This is however under the assumption that all cells are represented in the final sequencing data, not considering loss of cells or DNA during the processing steps. Therefore, this likely explains the variability between the replicates and would hence suggest that these proportions are not truly representative of the activity in that bin. However, as the fraction of cells binning in P7 is small (0.1% of the total cells is sorted into this bin), its contribution to the eventual mutational effect scores is insubstantial and fluctuations in this bin are largely obscured by the 99.9% of cells in other bins. Together with the correlation still being high, including of this bin was therefore justified.

5.3.3. Perturbations of phospho-residues are context dependent

We have shown that serine and threonine are interchangeable at phosphorylation sites (section 5.2.12), which is supported by our understanding that CKI and GSK-3β can phosphorylate both amino acids. The data moreover demonstrates that phosphomimetics do not behave consistently across all four phosphorylation sites (section 5.2.13). Whereas S45D is the main phosphomimetic used in literature, being minimally effective in phosphomimicking (Nejak-Bowen *et al.*, 2010) and one papers used S33/S37/T41E triple substitutions to mimic phosphorylation (Huang, Senga and Hamaguchi, 2007), the majority of these mutations have not been used in isolation. Whereas substitutions with aspartic or glutamic acid impede turnover of β-catenin in all other cases, S33E mutants retain a low TCF/LEF activity level (see **Figure 5.9C**), suggesting that phosphomimetics are context dependent.

GSK-3β phosphorylates serines or threonines four residues in the N-terminal direction of another phosphorylated serine or threonine (Cohen, 1986; Tuazon, 1991; Fish *et al.*, 1995). As all phosphomimetics at sites S37, S41 and S45 enhance activity of β-catenin, this suggests that GSK-3β is not able to dock on an aspartic or glutamic acid as it would on a phosphorylated serine or threonine. This has not been described before and further supports our understanding of the behaviour of this central kinase.

β-TrCP establishes identical hydrogen bonds with pS33 and pS37 upon binding β-catenin (Wu *et al.*, 2003). The data presented in this chapter suggests that these bonds can still be formed with glutamic acid (S33E), as it retains fairly low levels of TCF/LEF signalling, but not with aspartic acid (S33D) (section 5.2.13, figure 5.9). S37E does not follow this trend, but this may be due to its inference with the phosphorylation of S33 through the GSK3-3β cascade, whereas the reciprocal is not likely as explained in the previous paragraph. Under this assumption, a double mutant of S33E/S37E should however be turned over, as these residues can both be recognised as phosphorylated residues by β-TrCP. Additional experiments such as protein complex immunoprecipitation (Co-IP) for β-TrCP and mutant β-catenin proteins would be required to confirm these findings.

As aspartic acid and glutamic acid substitutions at any of the phospho-sites require a minimum of respectively two and three nucleotide changes, this, together with the fact that all of them requiring at least two nucleotide substitutions (**Table s5.1**), explains why these mutations are not found in tumour databases (COSMIC or TCGA) despite their positive effect on β-catenin activity.

5.3.4. Mutational effect score combines multi-dimensional data into a single value

I have shown that a single mutational effect score can be assigned to each amino acid substitution, which incorporates the proportions across all six bins and the pool and the relative sizes and fluorescence values of each bin (section 5.2.14). As both a higher number of bins and a higher number of HDR reads were obtained, the scores are much more sensitive than with the initial mutational effect scores calculated for GFP.

The mutational effect score represents the effect of each amino acid substitution on the transcriptional activity of β-catenin. Previous studies have shown that missense substitutions in *CTNNB1* have a range of effects on β-catenin activity *in vitro*, but only assessed a small number of mutations. This is the first systematic and quantitative interrogation of the phenotypic consequences of each possible substitution in this N-terminal region of *CTNNB1*.

The sensitivity of the scoring system cannot be benchmarked with the data currently available. Can anything for example be inferred about the differences between S33H (scoring 7827) and S33P (scoring 6817)? Future experimental work using clonal cell lines should therefore independently evaluate the activity of individual mutations in order to validate these findings and establish the sensitivity of the proposed system. Determining β-catenin protein concentrations in these clonal cell lines should furthermore confirm the hypothesis that mutations directly affect the turnover of this protein.

5.3.5. Mutational effect scores should be benchmarked by *in vitro* studies

The mutational effect scores provide a framework that can improve our understanding of the phenotypic consequences of mutations observed in *CTNNB1* in cancers. However, this assumes that mouse ESCs harbouring TCF/LEF::H2B-GFP are a suitable model system to infer phenotypes in other cell type and tissues. Firstly, while there could be intrinsic differences between the activity of β-catenin in mouse and human, the targeted amino acid residues are fully conserved between human and mouse and Wnt/β-catenin signalling itself is thought to be highly conserved between the two species (Sato *et al.*, 2004). It was therefore assumed that these findings could be extrapolated to the activity in human.

Secondly, whilst ESCs make a good system for the proposed deep mutational scanning approach due to their high HDR efficiencies and the ability to conditionally inhibit GSK3β signalling to mask the effects, our limited understanding of the differences in β-catenin signalling between tissues could obscure predictions on the behaviour of amino acid substitutions in specific cancers. Future work should therefore evaluate the activity of amino acid changes in different cell types, for example by differentiating earlier mentioned clonal cell lines harbouring amino acid substitutions into different lineages. This would be of particular interest for mutations found frequently in some cancer types but rarely in others,

such as H36P and S45P (see **Figure 5.14** and **Figure 5.16**). This should elucidate the extent to which the mutational effect scores can be extrapolated to infer anything about β-catenin signalling in other cell types.

The distribution of activity scores across all possible amino acid substitutions indicates that most mutations do not activate β-catenin signalling. Many of these mutations could disturb protein stability altogether, thereby completely stalling β-catenin protein accumulation and signalling. It is worthy to note that, taking this in consideration, all of the amino acid substitutions at the phosphorylation sites or β-TrCP binding motif are found in the higher GFP-expressing populations and these mutations thus cannot destabilise the protein. This is possibly because these residues have exposed side chains necessary for binding to one of the cofactors, reducing the effect of a substitution at these residues on protein stability.

5.3.6. Mutational effect scores in comparison to previous studies

When comparing the findings from this study to previous studies on quantifying the effects of specific mutations in *CTNNB1*, several parallels can be drawn. A study investigating *CTNNB1* mutational patterns in soft tissue tumours by Lazar *et al.* found that T41A (59%), S45F (33%), and S45P (8%) were the most frequently occurring (Lazar *et al.*, 2008). These mutations all have intermediate mutational effects (7611, 5969 and 5176, respectively) and corresponds with the histogram for soft tissue in **Figure 5.11** and mutational effect scores in **Figure 5.12**. This study found that S45F-mutated desmoids have a significantly poorer survival than T41A-mutated tumours (65%), which is inconsistent with the former having a higher effect score than the latter (5969 vs. 7611).

A study quantifying the effects of several mutations in a hepatocellular cancer cell line (HuH7) on TCF/LEF signalling shows some parallels, P44A (their luciferase score 10 versus a mutational effect score 656) S45P (15 vs. 5176) S45F (22 vs. 5969), although they found that H36P was the strongest mutation (24 vs. 3120) (Austinat *et al.*, 2008). Hence, the findings in the screen presented in this chapter are largely consistent with previous studies, but the discrepancies confirm that validation of these findings is essential as this could be a source of tissue specificity.

5.3.7. Different cancer types have distinct distributions of mutations

I have shown that different tumour types statistically harbour different mutations with specific levels on TCF/LEF activity (section 5.2.15). This is consistent with the 'just right' signalling model proposed by Albuquerque et al., which states that cells with too high β-catenin activity might be prone to trigger apoptosis and are hence not clonally expanded inside a tumour.

Alternative to a strong selection for mutations of intermediate effect, as is the case with the just right model, the difference in activity levels between tumour types (see

Table 5-3) could indicate that some tumours select for high levels of Wnt/β-catenin activity, whereas a broader range of activity levels may confer the same effect on proliferation in the context of other tumour types. Whilst various tumours harbour mutations with different levels of TCF/LEF signalling (see

Table 5-3), activity levels of these mutations should be addressed in the relevant tissues to confirm the extent to which the findings from ESCs can give meaningful information on other tissues.

In addition, findings from Austinat *et al.* suggest that some mutations (H36P) can enhance TCF/LEF signalling whilst leaving glutamine synthetase, another important downstream target of β-catenin targeting unchanged (Austinat *et al.*, 2008). When relating the mutational effect score from our study to cell fitness one should this be aware that β-catenin signalling is more complex than just the TCF/LEF activity.

5.3.8. Mutational patterns in *CTNNB1* are explained by both probability and activity

Using endometrial and liver cancer as case studies (section 5.2.16), it was furthermore deduced that neither the likelihood of mutations nor activity score alone can fully explain the observed difference in mutational patterns between tumours. By quantifying both factors, it was shown that both the likelihood and selective pressure are key for the eventual occurrence of missense mutations in tumours. This can be expected, as mutational processes lay out the mutational diversity that selection can work on. This is best seen in liver cancer (with a large number of samples) where many of the mutations with a high likelihood are observed, but mutations with a higher effect are most more frequent.

Liver and endometrial cancer were shown to harbour different mutations that are largely driven by mutational effect score and not by mutational likelihood (section 5.2.17), i.e. mutations that are enriched in liver cancers compared to endometrial are those with intermediate effect scores (e.g. H36P, S45P) whereas endometrial cancer is enriched in mutations with a high effect score (e.g. S37F, S37C). No difference was observed in the overall distribution in mutational effect score or likelihood between liver and endometrial cancer (**Figure 5.12**, **Figure 5.16** and

Table 5-3), which I expect to be the result of the limited number of mutations with both a high enough selective advantage and high likelihood, such that any advantageous mutation might be selected for in either of these tissues regardless of specific effect. As more clinical data becomes available, the accuracy of these findings should improve.

Calculating the probability of mutations for other tumours is currently limited by the availability of whole genome or exome sequencing (WGS/WES) data for samples containing *CTNNB1* mutations. Additional samples could be included by adding tumours without mutations in *CTNNB1*, although this could lead to confounding effects if CTNNB1 mutant tumours have different background mutation frequencies compared to other histologically similar tumours.

Finally, whilst the screen demonstrated in this chapter interrogated the effect of genetic variants on TCF/LEF signalling, the same strategy of generating a genetic library of cells could be applied to perform a survival screen as depicted in **Figure 5.2**; comparing the prolific advantage with the mutational effect score calculated in this study would reveal the extent to which signalling through TCF/LEF is proportional to survival.

6

Discussion & concluding remarks

6.1 Discussion

The main aim of this thesis was to utilise both CRISPR-Cas9 and multiplex HDR to establish a deep-mutational scanning (DMS) pipeline for mammalian genomic loci, and to validate this approach by assessing the effects of single nucleotide variants in GFP. We furthermore aimed to utilise the developed approach to assess the phenotypic consequences of amino acid substitutions in β-catenin.

A wide range of targeted genome diversification tools have been developed to date and allow for the interrogation of the phenotypic effects of genomic variants. Base editing tools provide a powerful way of introducing precise edits into a targeted window (Hess *et al.*, 2016; Komor *et al.*, 2016; Rees *et al.*, 2017), but as the targeting window is narrow and base editing often only allows one or two nucleotide positions to be modified per sgRNA, its application for DMS remains limited. The most fruitful attempts at DMS using CRISPR-Cas9 have been in the use of multiplex HDR (Findlay *et al.*, 2014a; Ma *et al.*, 2017; Kotler *et al.*, 2018; Mason *et al.*, 2018), whereby a repair template library is introduced to a pool of cells, such that each cell integrates a unique variant into the genomic locus of interest. Each of the approaches that has been developed to achieve this has its own advantages and limitations, as discussed in section 1.7.5. The two strategies for DMS presented in this thesis provide parallel approaches that unlock the potential of introducing hundreds of genomic variants simultaneously, either on the nucleotide or amino acid level.

This pipeline was used to characterise the effect of nucleotide changes that encode for 28 amino acid substitutions in GFP, which included a range of neutral, hypomorphic and null mutations (see **Table 4-9**), and furthermore characterised all 19 possible amino acid substitutions across 18 residues in a mutational hotspot region in β-catenin (see **Figure 5.10**). The latter scores were used to explain the mutational patterns observed in cancer by separating its effects from that of mutational signatures in certain tissues (**Figure 5.12** and **Figure 5.14**), thereby providing an innovative approach to acquiring novel insights into mutational drives in cancer.

6.1.1 Bioinformatics pipeline prove effective at filtering and analysing predefined outcomes

In chapter 3, a bioinformatics pipeline was developed and optimised using dummy reads representing repair outcomes from the GFP experiment. It was subsequently used to filter and analyse experimental datasets from different projects including SRSF1, GFP and β-catenin.

Analysis of the SRSF1 data (section 3.4) and GFP data (chapter 4) demonstrated that the pipeline is able to efficiently merge read pairs into a single contig and filter contigs based on a predefined editing outcome (e.g. HDR handles). Through this strategy, reads are typically sorted into multiple files (e.g. wildtype-like reads, HDR-like reads and other reads). Whilst this approach is well tailored to the needs in these projects, it does not allow for the detection of unknown outcomes, e.g. the quantification of specific repair outcomes following NHEJ, especially because the merging of read pairs only passes contigs that have the same length as the reference (i.e. no indels) and up to 6 nt substitutions. Other pipelines are however dedicated to the analysis of such read outcomes (Wang *et al.*, 2017). As many DMS methods are based on substitutions rather than indels, the pipeline optimised in chapter 3 would be applicable for a wide variety of amplicon-based screens, providing a pipeline with low memory usage that can be used without (substantial) prior knowledge of bioinformatics.

6.1.2 Selection strategies allow for increased assessment of HDR-derived variants

One of the limitations in the nucleotide diversification study on GFP was the low HDR efficiency from the ssODN pool. As described in **Table 4-1**, the maximum HDR efficiency obtained was 14%, but these proportions did typically not exceed 4 – 8%. As it was shown that the most effective analysis processes reads with a single variant (section 4.2.8), the proportion of usable reads is even lower than this. Consequently, the bulk of the sequence reads were discarded, thereby greatly decreasing the depth of analysis.

DSB repair from ssODNs is typically thought to yield higher insertion rates compared to dsDNA donors, especially for shorter stretches of repair (Miura *et al.*, 2015; Yoshimi *et al.*, 2016; Ferenczi *et al.*, 2017). During the time that this study was conducted, many solutions have been issued to increase the proportion of HDR-derived alleles in the eventual pool.

First of all, various small molecules have been shown to skew the ratio of NHEJ to HR towards the latter (Pinder, Salsman and Dellaire, 2015; Vartak and Raghavan, 2015), although these are species and cell type dependent and are often optimised on the integration of dsDNA (e.g. by enhancing Rad51). Much remains to be elucidated about the exact mechanisms through which ssODNs are integrated into the genome. Rather than classical homologous recombination (HR), the Fanconi Anemia (FA) pathway and Rad51-independent single-strand template repair (SSTR) have recently been suggested to be responsible for the introduction of nucleotide substitutions from ssODNs after Cas9 cleavage (Richardson *et al.*, 2018). Therefore, chemical enhancers of HDR using ssODN developed in the future could focus on factors involved in these pathways. As our understanding of the mechanisms involved in repair of Cas9-induced breaks develops, so will the ability to enhance integration (either through SSDA or direct integration, discussed in section 1.6 and Kan *et al.*, 2017) of donor templates.

Another way to increase the proportion of HDR-derived alleles in the pool is by preventing the recleavage of alleles that have already undergone HDR by destroying the protospacer and/or PAM. This strategy was performed in the targeting of *CTNNB1*, whereby the entire (puΔtk) selection cassette, therewith in addition the targeted protospacers and PAM, was replaced by the HDR template (see **Figure 5.2**). As the protospacer and PAM were absent from the dsDNA template, this furthermore prevented cleavage of the donor plasmid. In addition, the *CTNNB1* targeting project employed an FIAU selection step which removed a large fraction of cells that had not undergone HDR. Although it was not possible to ascertain HDR efficiency prior to this step, the obtained HDR efficiency of 52.4% in the HRM sample indicates that the FIAU step is likely to substantially increase the fraction of useful alleles in the PCR template (section 5.2.6). However, a drawback of this approach is that it requires a laborious genome engineering step to introduce the selection cassette.

To further enhance the eventual proportion of sequencing reads representing HDR-derived alleles, repair from the donor plasmid introduced silent mutations in *CTNNB1*. This creates a PCR priming site to enable selective amplification of HDR-derived template strands (Findlay *et al.*, 2014a), thereby also excluding the non-targeted allelic copy (see **Figure 5.2**). This

increased eventual mapping efficiencies to the HDR template to >98%. In comparison to the maximum yield of 15% with the GFP project described in chapter 4, these results demonstrate that HDR efficiencies can be largely increased by preventing recleavage of HDR-derived alleles, by selection against non-targeted cells and through selective PCR of HDR-derived reads.

6.1.3 Efficacy of the use of a double-stranded template was project-dependent

The use of a dsDNA donor template proved efficient for the functional assessment of β-catenin variants. However, the integration of genomic variants into GFP from a long (±820-bp) dsDNA donor led to a complete loss of GFP expression, making the phenotypic separation of genetic variants impossible (section 4.2.17). This discrepancy between both screens could be due to several reasons. Previous work has shown that the addition of random (i.e. non-homologous), linear DNA to editing reactions can increase the rate of error-prone repair (C. D. Richardson *et al.*, 2016), providing a possible explanation for the complete loss of GFP fluorescence whereby, in contrast to the targeting of β-catenin, linear templates were used. This likely provides a likely explanation for the loss in GFP fluorescence.

6.1.4 The advantages of nucleotide versus amino acid substitutions

Whereas the DMS of GFP was conducted by introducing single nucleotide substitutions from doped ssODNs, β-catenin was saturated at the codon level using a rationally designed repair template pool. Codon substitutions were not performed on systematic scale (i.e. all amino acid substitutions across a long range of residues) until recently (Kotler *et al.*, 2018), as the synthesis of such libraries was technically unfeasible, making the scale at which the β-catenin study is performed thus relatively novel. Substituting codons rather than nucleotides has an advantage considering the equal representation of all possible amino acid variants in the eventual pool, without biasing against those requiring multiple nucleotide changes. In addition, a protein sequence has functional information that is not directly visible in the nucleotide sequence, which is of predominant interest when comparing sequence conservation between species, considering that codon bias might obscure sequence conservation on the nucleotide level (discussed in section 1.1.3).

Contrastingly, while the introduction of variants on the nucleotide level through oligo doping predominantly results in amino acid substitutions that can be obtained by a single nucleotide change, the advantage is that this more closely reflects the variants that arise naturally in biological systems. In addition, when viewing these variants through the scope of codon usage bias, silent mutations might affect cell fitness in various ways, elaborately discussed in section 1.1.3, which are not assessed by the approach using a single codon per amino acid utilised in chapter 5. This was illustrated in section 4.2.13, where the synonymous codons encoding T109T in GFP represented different mutational effect scores. This provisional finding however needs corroborating evidence through experiments with clonal cell lines. This furthermore raises a possible limitation on the 'one codon, one amino acid' strategy used in the β-catenin, which uses the most frequently used codons in mouse and thereby possibly obscuring some codon usage biases when comparing the data to mutational data in tumours, where a different codon may underlie the amino acid change.

Whether DMS on the nucleotide or codon level is the best way forward thus depends on the hypothesis being tested. Current library synthesis methods however allow for the rational design of sequencing libraries encompassing very large numbers of sequence variants, as is demonstrated in a recent study (Kotler *et al.*, 2018), which can therefore entail all possible single variants on both the nucleotide and amino acid level and thus encompass the best of both worlds. In my view, this scale of synthesis will be the gold standard for DMS and will be seen more often in the near future.

6.1.5 Mutational effect score encompasses multi-dimensional data in a single value

The formula derived from Kosuri *et al.*, (2013) was used to assign a single functional value per amino acid substitution entailing both its proportionality and the fluorescence value per bin. Because of the low number of bins, the scores in GFP were the same for variants at high and low frequency, which needed to be weighted by an F score to indicate the power of these mutational effect scores. The mutational effect score was more robust in the β-catenin study (see **Figure 5.10**) compared to the GFP study (see **Table 4-9**) because of (1) the depth of analysis (i.e. number of reads) and (2) the larger number of bins, which allowed scoring at a higher resolution. Nonetheless, the validity of our GFP study is underscored by the agreement between mutational effect scores and $\Delta\Delta G$ values (R2 = 0.47, see **Table 4-9**).

In the β-catenin study, the largest difference in frequencies of introduced codons was detected between the lower (P2-P4) and higher (P5-P7) segment of TCF/LEF-GFP fluorescence (section 0), which is consistent with the upper limit of GFP expression in the untargeted cells (see **Figure 5.3A**). This difference is focussed around residues involved in the turnover of β-catenin, i.e. the phosphorylation sites and β-TrCP binding site more frequently have mutations in the higher GFP bins, supporting the idea that each of these sites is essential for the degradation of the protein. By exploring the individual bins in more detail, it was however shown that the behaviour of codon substitutions exhibits differences across the bins (see **Figure 5.4** and **Figure 5.9**), confirming that the higher number of bins increases the resolution of the screen and is necessary to elucidate subtleties between amino acid substitution.

The mutational effect scores calculated provide single values for the relative effect of different variants and correspond with the mutational patterns observed in cancers. Future efforts could interrogate whether the data from these experiments could be analysed by applying different models, for example through regression analysis, principle component analysis or a different form of multivariate statistics that reduce the number of dimensions in the data and emphasise strong patterns in the dataset. These approaches would be worth investigating and should receive a high priority, but have not been performed for this thesis due to time constraints. A suitable benchmarking strategy would be necessary to judge which model performs best, which is further discussed in the next section. However, as the mutational effect scores calculated in this thesis show a distribution as expected and correlate with mutations that are occurring frequently in tumours, these scores currently provide sufficient detail for our purposes.

6.1.6 Validations are essential in determining sensitivity and robustness of mutational effect score

The DMS screen developed in this thesis quantified the phenotypic consequences of genetic variants in both GFP and β-catenin. The mutational effect scores determined in GFP were compared against *in silico* predicted effects on protein stability (see **Table 4-9**). This approach to calculate the ΔΔG was however less suitable in β-catenin, as (1) the region is inherently

unstructured (Xing *et al.*, 2008) and (2) *gain*-of-function mutations were measured in contrast to loss-of-function, which was moreover done with (3) a downstream reporter rather than direct functional measurements of the targeted protein. Whilst mutational effect scores are consistent with clinical prevalence of substitutions in the mutational region of *CTNNB1* (see **Figure 5.12**), it is not clear that the highest possible levels of *CTNNB1* signalling are always optimal for tumour development (see section 5.1.4). Fluorescence profiles of clonal cell lines of at least a subset of the mutations interrogated in β-catenin should ideally be generated to validate the sensitivity of the mutational effect screen. However, studies of clonal cell lines carrying the same GFP substitutions have been performed by other members of the Wood laboratory. In some cases, substantial differences in fluorescence were observed, which complicates this approach.

Different cell types reside under specific environmental conditions (e.g. exposure to certain growth factors) and have specific gene expression patterns. Whilst the phenotypic quantifications in this thesis provide a powerful tool to assess the effects of these mutations under these specific conditions, the extent to which anything about their effects in other tissues can be inferred from these effect scores is not known. Therefore, the DMS screen of β-catenin should be repeated in a different cell type. Indeed, this caveat is broadly applicable to studies applying DMS to understand the genetic basis of phenotypic traits in multicellular organisms: the findings in one system are not necessarily transferrable to other systems or cell lines.

6.1.7 Other applications of DMS using CRISPR-Cas9 and multiplex HDR

In this thesis, the developed DMS pipeline was utilised to study the effects of genetic variants in coding sequences at the codon level. In addition, analysis touched upon the effects of synonymous mutations (section 4.2.14), albeit the depth of the GFP screen was not sufficient to further elaborate on this. Until recently a bottleneck in laboratory-based studies of evolution was the generation of targeted genetic variation in the genome. CRISPR-Cas9 allows for an increase in variation, and thereby the rate of evolution, in a highly localised fashion whilst leaving the rest of the genome relatively unscathed. When studying the effects of genetic variants on protein stability, DMS combined with HDR as proposed here is especially powerful for the interrogation of essential genes, as substitutions allow for the

introduction and assessment of hypomorphic mutations, something less feasible when basing the mutagenesis on indels.

Whilst not utilised to this purpose here, DMS using CRISPR-Cas9 and template libraries can interrogate non-coding elements such as upstream open reading frames (uORFs), UTRs or other regulating regions including enhancers (Canver *et al.*, 2014). With the ease and affordability of next-generation sequencing, vast amounts of genomic data are being produced and made publicly available through initiatives including Ensembl (Zerbino *et al.*, 2017), the 1000 Genomes Project (Auton *et al.*, 2015) and dbSNP (Frazer *et al.*, 2007). A vast majority of the variants identified in these projects have not been phenotypically characterised, which could be exploited by DMS. Rationally, a future target of DMS should hence be a genomic region with a high density of uncharacterised variants, as was aimed to do on *CTNNB1* mutations from COSMIC in this thesis (Forbes *et al.*, 2018). However, even interrogations of non-coding elements require a suitable phenotypic screen, either through transcriptional or fluorescent measurement of an element in *cis* or through a fitness screen, and this hence limits the number of suitable targets, as was experienced with analysis of *Nanog* (not shown in this thesis).

Alternatively, as published before (Hess *et al.*, 2016; Ma *et al.*, 2017), DMS can also be exploited in cancer drug resistance screens, whereby the heterogeneous pool of cells is subjected to a chemotherapeutic agent, by which clones harbouring a variant underlying resistance can be identified after survival. Similarly, it can be used to engineer valuable therapeutics such as antibodies (Mason *et al.*, 2018). DMS using CRISPR-Cas9 is hence a potent tool in analysing the phenotypic consequences in a wide variety of genomic elements and in driving directed evolution. Indeed, now we have been able to generate vast amounts of genomic data, the next step in genetics is to understand and manipulate alterations in the genetic code.

6.2 Closing remarks

With the relative ease and cost efficiency of both precise genome editing and sequence analysis, deep mutational scanning has become a widely used approach to systematically study the effects of genetic elements and variations therein. A wide range of approaches to this have been developed, each with its limitations and advantages.

The pipeline presented in this thesis provides a means to introduce both nucleotide and codon substitutions into a genomic locus using CRISPR-Cas9 and multiplex HDR and to assess their phenotypic consequences in parallel. The experimental and bioinformatics pipelines can be used to study coding elements. Combined with the analysis of mutational signatures found in cancer databases, this pipeline was used to reveal novel insights into β-catenin variants that increase our understanding of mutations found in tumours. However, the availability of more tumour genome sequences and DMS datasets from clinically relevant cell lines will be required to separate the relative contribution of mutational bias and natural selection on the eventual mutational pattern observed in different tumours.

7

Bibliography

Aberle, H. *et al.* (1997) 'Beta-Catenin Is a Target for the Ubiquitin-Proteasome Pathway.', *The EMBO journal*, 16(13), pp. 3797–804. doi: 10.1093/emboj/16.13.3797.

Adzhubei, I. A. *et al.* (2010) 'A method and server for predicting damaging missense mutations', *Nature methods*. Nature Publishing Group, 7(4), p. 248.

Ahmad, S. *et al.* (2004) 'ASAView: database and tool for solvent accessibility representation in proteins', *BMC bioinformatics*. BioMed Central, 5(1), p. 51.

Aita, T. *et al.* (2002) 'Surveying a local fitness landscape of a protein with epistatic sites for the study of directed evolution', *Biopolymers: Original Research on Biomolecules*. Wiley Online Library, 64(2), pp. 95–105. doi: 10.1002/bip.10126.

Albuquerque, C. *et al.* (2002) 'The "just-right" signaling model: APC somatic mutations are selected based on a specific level of activation of the beta-catenin signaling cascade.', *Human molecular genetics*, 11(13), pp. 1549–1560. doi: 10.1093/hmg/11.13.1549.

Albuquerque, C. *et al.* (2011) 'Colorectal cancers choosing sides', *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*. Elsevier B.V., 1816(2), pp. 219–231. doi: 10.1016/j.bbcan.2011.07.005.

Alexandrov, L. B. *et al.* (2013) 'Signatures of mutational processes in human cancer', *Nature*. Nature Publishing Group, 500(7463), p. 415.

Alexandrov, L. B. and Stratton, M. R. (2014) 'Mutational signatures: The patterns of somatic mutations hidden in cancer genomes', *Current Opinion in Genetics and Development*. Elsevier Ltd, 24(1), pp. 52–60. doi: 10.1016/j.gde.2013.11.014.

Altenburg, E. (1930) 'The effect of ultraviolet radiation on mutation', *Anat. Rec*, 47, p. 383.

Altschul, S. F. *et al.* (1997) 'Gapped BLAST and PSI-BLAST: a new generation of protein database search programs', *Nucleic acids research*. Oxford University Press, 25(17), pp. 3389–3402.

Altshuler, D. M. *et al.* (2012) 'An integrated map of genetic variation from 1,092 human genomes', *Nature*, 491(7422), pp. 56–65. doi: 10.1038/nature11632.

Andrews, S. (2010) 'FastQC: a quality control tool for high throughput sequence data'.

Araya, C. L. and Fowler, D. M. (2011) 'Deep mutational scanning: assessing protein function on a massive scale', *Trends in Biotechnology*, 29(9), pp. 435–442. doi: 10.1016/j.tibtech.2011.04.003.Deep.

Arend, R. C. *et al.* (2013) 'The Wnt/β-catenin pathway in ovarian cancer: A review', *Gynecologic Oncology*. Elsevier Inc., 131(3), pp. 772–779. doi: 10.1016/j.ygyno.2013.09.034.

Auerbach, C. (1949) 'Chemical Mutagenesis', *Biological Reviews*, 24(3), pp. 355–391. doi: 10.1146/annurev.bi.51.070182.003255.

Austinat, M. *et al.* (2008) 'Correlation between β-catenin mutations and expression of Wnt-signaling target genes in hepatocellular carcinoma', *Molecular Cancer*, 7, pp. 1–9. doi: 10.1186/1476-4598-7-21.

Auton, A. *et al.* (2015) 'A global reference for human genetic variation', *Nature*, 526(7571), pp. 68–74. doi: 10.1038/nature15393.

Barrangou, R. *et al.* (2007) 'CRISPR provides acquired resistance against viruses in prokaryotes', *Science*. American Association for the Advancement of Science, 315(5819), pp. 1709–1712.

Bassett, A. R. *et al.* (2013) 'Highly Efficient Targeted Mutagenesis of Drosophila with the CRISPR/Cas9 System', *Cell Reports*. The Authors, 4(1), pp. 220–228. doi: 10.1016/j.celrep.2013.06.020.

Behrens, J. *et al.* (1996) 'Functional interaction of β-catenin with the transcription factor LEF-1', *Nature*, pp. 638–642. doi: 10.1038/382638a0.

Bellen, H. J. *et al.* (1989) 'P-element-mediated enhancer detection: a versatile method to study development in Drosophila.', *Genes & development*. Cold Spring Harbor Lab, 3(9), pp. 1288–1300.

Bennardo, N. *et al.* (2008) 'Alternative-NHEJ is a mechanistically distinct pathway of mammalian chromosome break repair', *PLoS genetics*. Public Library of Science, 4(6), p. e1000110.

Bentley, D. R. *et al.* (2008) 'Accurate whole human genome sequencing using reversible terminator chemistry', *nature*. Nature Publishing Group, 456(7218), p. 53.

Berman, H. M. *et al.* (2006) 'The protein data bank, 1999–', in *International Tables for Crystallography Volume F: Crystallography of biological macromolecules*. Springer, pp. 675–684.

Bernardi, G. (2000) 'Isochores and the evolutionary genomics of vertebrates', *Gene*, 241(1), pp. 3–17. doi: 10.1016/S0378-1119(99)00485-0.

Beucher, A. *et al.* (2009) 'ATM and Artemis promote homologous recombination of radiation-induced DNA double-strand breaks in G2', *The EMBO journal*. EMBO Press, 28(21), pp. 3413–3427.

Bhanot, P. *et al.* (1996) 'A new memeber of the frizzled family from Drosophila funcions as a Wingless receptor', *Nature*, 382, pp. 225–31.

Bhargava, R. *et al.* (2018) 'C-NHEJ without indels is robust and requires synergistic function of distinct XLF domains', *Nature Communications*. Springer US, 9(1). doi: 10.1038/s41467-018-04867-5.

Bianconi, E. *et al.* (2013) 'An estimation of the number of cells in the human body', *Annals of human biology*. Taylor & Francis, 40(6), pp. 463–471.

Bibikova, M. *et al.* (2003) 'Enhancing gene targeting with designed zinc finger nucleases.', *Science (New York, N.Y.)*, 300(5620), p. 764.

Birling, M.-C. *et al.* (2017) 'Efficient and rapid generation of large genomic variants in rats and mice using CRISMERE', *Scientific Reports*. Nature Publishing Group, 7, p. 43331.

Björklund, P. *et al.* (2008) 'Stabilizing mutation of CTNNB1/beta-catenin and protein accumulation analyzed in a large series of parathyroid tumors of Swedish patients', *Molecular cancer*. BioMed Central, 7(1), p. 53.

Blankenship, R. E. and Hartman, H. (1998) 'The origin and evolution of oxygenic photosynthesis', *Trends in biochemical sciences*. Elsevier, 23(3), pp. 94–97.

Bloom, J. D. *et al.* (2005) 'Thermodynamic prediction of protein neutrality', *Proceedings of the National Academy of Sciences*. National Acad Sciences, 102(3), pp. 606–611.

Boch, J. *et al.* (2009) 'Breaking the code of DNA binding specificity of TAL-type III effectors.', *Science (New York, N.Y.)*, 326(5959), pp. 1509–1512.

Boel, A. *et al.* (2016) 'BATCH-GE: Batch analysis of Next-Generation Sequencing data for genome editing assessment', *Scientific reports*. Nature Publishing Group, 6, p. 30330.

Bokulich, N. A. *et al.* (2012) 'Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing', *Nature Methods*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 10(1), p. 57. doi: 10.1038/nmeth.2276.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: A flexible trimmer for Illumina sequence data', *Bioinformatics*, 30(15), pp. 2114–2120. doi: 10.1093/bioinformatics/btu170.

Bolotin, A. *et al.* (2005) 'Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin', *Microbiology*. Microbiology Society, 151(8), pp. 2551–2561.

Boroviak, K. *et al.* (2017) 'Revealing hidden complexities of genomic rearrangements

generated with Cas9', *Scientific reports*. Nature Publishing Group, 7(1), p. 12867.
Bothmer, A. *et al.* (2017) 'Characterization of the interplay between DNA repair and CRISPR/Cas9-induced DNA lesions at an endogenous locus', *Nature Communications*. Nature Publishing Group, 8(May 2016), pp. 1–12. doi: 10.1038/ncomms13905.
Boyle, A. P. *et al.* (2008) 'High-resolution mapping and characterization of open chromatin across the genome', *Cell*. Elsevier, 132(2), pp. 311–322.
Brembeck, F. H. *et al.* (2004) 'Essential role of BCL9-2 in the switch between β -catenin ' s adhesive and transcriptional functions service Essential role of BCL9-2 in the switch between ␣-catenin ' s adhesive and transcriptional functions', *Genes and Development*, 18, pp. 2225–2230. doi: 10.1101/gad.317604.
Brinkman, E. K. *et al.* (2018) 'Kinetics and Fidelity of the Repair of Cas9-Induced Double-Strand DNA Breaks', *Molecular Cell*. Elsevier Inc., 70(5), p. 801–813.e6. doi: 10.1016/j.molcel.2018.04.016.
Brouns, S. J. J. *et al.* (2008) 'Small CRISPR RNAs guide antiviral defense in prokaryotes', *Science*. American Association for the Advancement of Science, 321(5891), pp. 960–964.
Byrne, S. M. *et al.* (2014) 'Multi-kilobase homozygous targeted gene replacement in human induced pluripotent stem cells', *Nucleic acids research*. Oxford University Press, 43(3), pp. e21–e21.
Cairo, S. *et al.* (2008) 'Hepatic Stem-like Phenotype and Interplay of Wnt/β-Catenin and Myc Signaling in Aggressive Childhood Liver Cancer', *Cancer Cell*, 14(6), pp. 471–484. doi: 10.1016/j.ccr.2008.11.002.
Canver, M. C. *et al.* (2014) 'Characterization of genomic deletion efficiency mediated by CRISPR/Cas9 in mammalian cells', *Journal of Biological Chemistry*. ASBMB, p. jbc-M114.
Capecchi, M. R. (1989) 'Altering the genome by homologous recombination', *Science*. American Association for the Advancement of Science, 244(4910), pp. 1288–1292.
Carlson, E. A. (2007) 'Genes, radiation, and society; the life and work of HJ Muller'.
Carroll, D. (2008) 'Progress and prospects: zinc-finger nucleases as gene therapy agents', *Gene therapy*. Nature Publishing Group, 15(22), p. 1463.
Cavallo, R. A. *et al.* (1998) 'Drosophila Tcf and Groucho interact to repress wingless signalling activity', *Nature*, 395(6702), pp. 604–608. doi: 10.1038/26982.
Ceccaldi, R., Rondinelli, B. and D'Andrea, A. D. (2016) 'Repair Pathway Choices and Consequences at the Double-Strand Break', *Trends in Cell Biology*. Elsevier Ltd, 26(1), pp. 52–64. doi: 10.1016/j.tcb.2015.07.009.
Chambers, I. *et al.* (2007) 'Nanog safeguards pluripotency and mediates germline development.', *Nature*, 450(7173), pp. 1230–1234.
Chang, H. H. Y. *et al.* (2017) 'Non-homologous DNA end joining and alternative pathways to double-strand break repair', *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, 18(8), pp. 495–506. doi: 10.1038/nrm.2017.48.
Chen, F. *et al.* (2011) 'High-frequency genome editing using ssDNA oligonucleotides with zinc-finger nucleases', *Nature methods*. Nature Publishing Group, 8(9), p. 753.
Chen, S.-H. *et al.* (1987) 'Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon', *Science*. American Association for the Advancement of Science, 238(4825), pp. 363–366.
Chen, W. (2003) 'Dishevelled 2 Recruits -Arrestin 2 to Mediate Wnt5A-Stimulated Endocytosis of Frizzled 4', *Science*, 301(5638), pp. 1391–1394. doi: 10.1126/science.1082808.
Chen, Y. and Bradley, A. (2000) 'A new positive/negative selectable marker, puΔtk, for use in embryonic stem cells', *genesis*. Wiley Online Library, 28(1), pp. 31–35.

Clevers, H. and Nusse, R. (2012) 'Wnt/β-catenin signaling and disease', *Cell*, 149(6), pp. 1192–1205. doi: 10.1016/j.cell.2012.05.012.

Cohen, P. (1986) '11 Muscle Glycogen Synthase', in *The enzymes*. Elsevier, pp. 461–497.

Colbert, T. *et al.* (2001) 'High-throughput screening for induced point mutations', *Plant physiology*. Am Soc Plant Biol, 126(2), pp. 480–484.

Cong, L. *et al.* (2013) 'Multiplex Genome Engineering Using CRISPR/Cas System', *Science*, 819(February), pp. 819–823. doi: 10.1126/science.1231143.

Cormack, B. P., Valdivia, R. H. and Falkow, S. (1996) 'FACS-optimized mutants of the green fluorescent protein (GFP)', *Gene*, 173(1), pp. 33–38. doi: 10.1016/0378-1119(95)00685-0.

Cox, M. P., Peterson, D. A. and Biggs, P. J. (2010) 'SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data', *BMC Bioinformatics*, 11. doi: 10.1186/1471-2105-11-485.

Crawford, G. E. *et al.* (2006) 'Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS)', *Genome research*. Cold Spring Harbor Lab, 16(1), pp. 123–131.

Crick, F. (1970) 'Central dogma of molecular biology.', *Nature*, 227(5258), pp. 561–3. doi: 10.1038/227561a0.

Cunningham, B. C. and Wells, J. A. (1989) 'High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis', *Science*, 244(4908), pp. 1081–1085. doi: 10.1126/science.2471267.

Curtis, H. J. (1965) 'Formal discussion of: somatic mutations and carcinogenesis', *Cancer research*. AACR, 25(8), pp. 1305–1309.

Dabbish, L. *et al.* (2012) 'Social coding in GitHub: transparency and collaboration in an open software repository', in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, pp. 1277–1286.

Davidson, G. *et al.* (2005) 'Casein kinase 1?? couples Wnt receptor activation to cytoplasmic signal transduction', *Nature*, 438(7069), pp. 867–872. doi: 10.1038/nature04170.

Derbyshire, K. M., Salvo, J. J. and Grindley, N. D. (1986) 'A simple and efficient procedure for saturation mutagenesis using mixed oligodeoxynucleotides.', *Gene*, 46(2–3), pp. 145–152.

DiBiase, S. J. *et al.* (2000) 'DNA-dependent protein kinase stimulates an independently active, nonhomologous, end-joining apparatus', *Cancer research*. AACR, 60(5), pp. 1245–1253.

Doench, J. G. *et al.* (2016) 'Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9', *Nature biotechnology*. Nature Publishing Group, 34(2), p. 184.

Doolittle, W. F. (1996) 'Some aspects of the biology of cells and their possible evolutionary significance', in *SYMPOSIA-SOCIETY FOR GENERAL MICROBIOLOGY*, pp. 1–22.

Doudna, J. A. and Charpentier, E. (2014) 'The new frontier of genome engineering with CRISPR-Cas9', *Science*, 346(6213). doi: 10.1126/science.1258096.

Dueva, R. and Iliakis, G. (2013) 'Alternative pathways of non-homologous end joining (NHEJ) in genomic instability and cancer', *Translational Cancer Research*, 2(3), pp. 163–177. doi: 10.3978/j.issn.2218-676X.2013.05.02.

Dupuy, A. J. *et al.* (2005) 'Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system', *Nature*. Nature Publishing Group, 436(7048), p. 221.

Ellegren, H., Smith, N. G. C. and Webster, M. T. (2003) 'Mutation rate variation in the mammalian genome', *Current opinion in genetics & development*. Elsevier, 13(6), pp. 562–568.

Ezkurdia, I. *et al.* (2014) 'Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes', *Human molecular genetics*. Oxford University Press,

23(22), pp. 5866–5878.
Falkowski, P. G. *et al.* (2005) 'The rise of oxygen over the past 205 million years and the evolution of large placental mammals', *Science*. American Association for the Advancement of Science, 309(5744), pp. 2202–2204.
Faunes, F. *et al.* (2013) 'A membrane-associated β-catenin/Oct4 complex correlates with ground-state pluripotency in mouse embryonic stem cells', *Development*. Oxford University Press for The Company of Biologists Limited, 140(6), pp. 1171–1183.
Ferenczi, A. *et al.* (2017) 'Efficient targeted DNA editing and replacement in Chlamydomonas reinhardtii using Cpf1 ribonucleoproteins and single-stranded DNA', *Proceedings of the National Academy of Sciences*. National Acad Sciences, 114(51), pp. 13567–13572.
Ferrer-Vaquer, A. *et al.* (2010) 'A sensitive and bright single-cell resolution live imaging reporter of Wnt/ss-catenin signaling in the mouse', *BMC developmental biology*. BioMed Central, 10(1), p. 121.
Feuk, L., Carson, A. R. and Scherer, S. W. (2006) 'Structural variation in the human genome', *Nature Reviews Genetics*. Nature Publishing Group, 7(2), p. 85.
Findlay, G. M. *et al.* (2014a) 'Saturation editing of genomic regions by multiplex homology-directed repair.', *Nature*. Nature Publishing Group, 513(7516), pp. 1–2. doi: 10.1038/nature13695.
Findlay, G. M. *et al.* (2014b) 'Saturation editing of genomic regions by multiplex homology-directed repair', *Nature*. Nature Publishing Group, 513(7516), pp. 1–2. doi: 10.1038/nature13695.
Findlay, G. M. *et al.* (2018) 'Accurate classification of BRCA1 variants with saturation genome editing', *Nature*. Nature Publishing Group, 562(7726), p. 217.
Fire, A. *et al.* (1998) 'Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans', *nature*. Nature Publishing Group, 391(6669), p. 806.
Fish, K. J. *et al.* (1995) 'Isolation and characterization of human casein kinase Iε (CKI), a novel member of the CKI gene family', *Journal of Biological Chemistry*. ASBMB, 270(25), pp. 14875–14883.
Fodde, R., Smits, R. and Clevers, H. (2001) 'APC, Signal transduction and genetic instability in colorectal cancer', *Nature Reviews Cancer*, 1(1), pp. 55–67. doi: 10.1038/35094067.
Forbes, S. A. *et al.* (2018) 'COSMIC : somatic cancer genetics at high-resolution', 45(May), pp. 777–783. doi: 10.1093/nar/gkw1121.
Fowler, D. M. and Fields, S. (2014) 'Deep mutational scanning: A new style of protein science', *Nature Methods*, 11(8), pp. 801–807. doi: 10.1038/nmeth.3027.
Frazer, K. A. *et al.* (2007) 'A second generation human haplotype map of over 3.1 million SNPs', *Nature*, 449(7164), pp. 851–861. doi: 10.1038/nature06258.
Freeman, A. M. *et al.* (2011) 'Action at a distance: amino acid substitutions that affect binding of the phosphorylated CheY response regulator and catalysis of dephosphorylation can be far from the CheZ phosphatase active site', *Journal of bacteriology*. Am Soc Microbiol, p. JB-00070.
Friedberg, E. C. (2002) 'The intersection between the birth of molecular biology and the discovery of DNA repair', *DNA Repair*, 1(10), pp. 855–867. doi: 10.1016/S1568-7864(02)00112-X.
Gallagher, D. N. and Haber, J. E. (2018) 'Repair of a Site-Specific DNA Cleavage: Old-School Lessons for Cas9-Mediated Gene Editing', *ACS Chemical Biology*, 13(2), pp. 397–405. doi: 10.1021/acschembio.7b00760.
Garneau, J. E. *et al.* (2010) 'The CRISPR/cas bacterial immune system cleaves bacteriophage and plasmid DNA', *Nature*, 468(7320), pp. 67–71. doi: 10.1038/nature09523.

Gasiunas, G. *et al.* (2012) 'Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria', *Proceedings of the National Academy of Sciences*. National Acad Sciences, 109(39), pp. E2579–E2586.

Gaspar, C. and Fodde, R. (2004) 'APC dosage effects in tumorigenesis and stem cell differentiation', *International Journal of Developmental Biology*, 48(5–6), pp. 377–386. doi: 10.1387/ijdb.041807cg.

Gaudelli, N. M. *et al.* (2017) 'Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage', *Nature*. Macmillan Publishers Limited, part of Springer Nature. All rights reserved., 551, p. 464. Available at: http://dx.doi.org/10.1038/nature24644.

Ge, X. *et al.* (2010) 'Rapid construction and characterization of synthetic antibody libraries without DNA amplification', *Biotechnology and bioengineering*. Wiley Online Library, 106(3), pp. 347–357.

Giacomelli, A. O. *et al.* (2018) 'Mutational processes shape the landscape of TP53 mutations in human cancer', *Nature Genetics*. Springer US, 50(October). doi: 10.1038/s41588-018-0204-y.

Gibson, T. J., Seiler, M. and Veitia, R. A. (2013) 'The transience of transient overexpression', *Nature Methods*. Nature Publishing Group, 10(8), pp. 715–721. doi: 10.1038/nmeth.2534.

Gilchrist, E. and Haughn, G. (2010) 'Reverse genetics techniques: engineering loss and gain of gene function in plants', *Briefings in functional genomics*. Oxford University Press, 9(2), pp. 103–110.

Gilchrist, E. J. *et al.* (2006) 'TILLING is an effective reverse genetics technique for Caenorhabditis elegans', *BMC genomics*. BioMed Central, 7(1), p. 262.

Giles, R. H., Van Es, J. H. and Clevers, H. (2003) 'Caught up in a Wnt storm: Wnt signaling in cancer', *Biochimica et Biophysica Acta - Reviews on Cancer*, 1653(1), pp. 1–24. doi: 10.1016/S0304-419X(03)00005-2.

Gingold, H. *et al.* (2014) 'A dual program for translation regulation in cellular proliferation and differentiation', *Cell*. Elsevier, 158(6), pp. 1281–1292.

Glazko, G. V. *et al.* (2006) 'Mutational hotspots in the TP53 gene and, possibly, other tumor suppressors evolve by positive selection', *Biology Direct*, 1, pp. 1–9. doi: 10.1186/1745-6150-1-4.

Gottardi, C. J. and Gumbiner, B. M. (2004) 'Distinct molecular forms of beta-catenin are targeted to adhesive or transcriptional complexes.', *The Journal of Cell Biology*, 167(2), pp. 339–349. doi: 10.1083/jcb.200402153.

Govindarajan, S. and Goldstein, R. A. (1997) 'Evolution of model proteins on a foldability landscape', *Proteins: Structure, Function, and Bioinformatics*. Wiley Online Library, 29(4), pp. 461–466.

Goya, R. *et al.* (2010) 'SNVMix: Predicting single nucleotide variants from next-generation sequencing of tumors', *Bioinformatics*, 26(6), pp. 730–736. doi: 10.1093/bioinformatics/btq040.

Güell, M., Yang, L. and Church, G. M. (2014) 'Genome editing assessment using CRISPR Genome Analyzer (CRISPR-GA)', *Bioinformatics*. Oxford University Press, 30(20), pp. 2968–2970.

Le Guellec, S. *et al.* (2012) 'CTNNB1 mutation analysis is a useful tool for the diagnosis of desmoid tumors: a study of 260 desmoid tumors and 191 potential morphologic mimics', *Modern Pathology*. Nature Publishing Group, 25(12), p. 1551.

Guerois, R., Nielsen, J. E. and Serrano, L. (2002) 'Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations', *Journal of molecular biology*. Elsevier, 320(2), pp. 369–387.

Guo, H. H., Choe, J. and Loeb, L. A. (2004) 'Protein tolerance to random amino acid change', *Proceedings of the National Academy of Sciences*. National Acad Sciences, 101(25), pp. 9205–9210.

Gustafsson, C., Govindarajan, S. and Minshull, J. (2004) 'Codon bias and heterologous protein expression', *Trends in biotechnology*. Elsevier, 22(7), pp. 346–353.

Hagen, T. and Vidal-Puig, A. (2002) 'Characterisation of the phosphorylation of β-catenin at the GSK-3 priming site Ser45', *Biochemical and Biophysical Research Communications*, 294(2), pp. 324–328. doi: 10.1016/S0006-291X(02)00485-0.

Halperin, S. O. *et al.* (2018) 'CRISPR-guided DNA polymerases enable diversification of all nucleotides in a tunable window', *Nature*. doi: 10.1038/s41586-018-0384-8.

Hao, X. *et al.* (1997) 'Reciprocity between membranous and nuclear expression of β-catenin in colorectal tumours', *Virchows Archiv*, 431(3), pp. 167–172. doi: 10.1007/s004280050084.

Harris, T. J. C. and Peifer, M. (2005) 'Decisions, decisions: β-catenin chooses between adhesion and transcription', *Trends in Cell Biology*, 15(5), pp. 234–237. doi: 10.1016/j.tcb.2005.03.002.

Hatsell, S. *et al.* (2003) 'β-Catenin and Tcfs in mammary development and cancer', *Journal of mammary gland biology and neoplasia*. Springer, 8(2), pp. 145–158.

Hayashi, Y. *et al.* (2006) 'Experimental rugged fitness landscape in protein sequence space', *PLoS ONE*, 1(1). doi: 10.1371/journal.pone.0000096.

Helleday, T., Eshtad, S. and Nik-Zainal, S. (2014) 'Mechanisms underlying mutational signatures in human cancers', *Nature Reviews Genetics*. Nature Publishing Group, 15(9), pp. 585–598. doi: 10.1038/nrg3729.

Hess, G. T. *et al.* (2016) 'Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells', *Nature Methods*, 13(12), pp. 1036–1042. doi: 10.1038/nmeth.4038.

Hiscott, J. *et al.* (1999) 'Triggering the interferon response: the role of IRF-3 transcription factor', *Journal of interferon & cytokine research*. Mary Ann Liebert, Inc., 19(1), pp. 1–13.

Howard, S. M., Yanez, D. A. and Stark, J. M. (2015) 'DNA Damage Response Factors from Diverse Pathways, Including DNA Crosslink Repair, Mediate Alternative End Joining', *PLoS Genetics*, 11(1), pp. 1–25. doi: 10.1371/journal.pgen.1004943.

Hsu, P. D., Lander, E. S. and Zhang, F. (2014) 'Development and applications of CRISPR-Cas9 for genome engineering', *Cell*. Elsevier, 157(6), pp. 1262–1278. doi: 10.1016/j.cell.2014.05.010.

Hu, J. H. *et al.* (2018) 'Evolved Cas9 variants with broad PAM compatibility and high DNA specificity', *Nature*. Nature Publishing Group, 556(7699), p. 57.

Huang, P., Senga, T. and Hamaguchi, M. (2007) 'A novel role of phospho-β-catenin in microtubule regrowth at centrosome', *Oncogene*, 26(30), pp. 4357–4371. doi: 10.1038/sj.onc.1210217.

Huber, A. H., Nelson, W. J. and Weis, W. I. (1997) 'Three-Dimensional Structure of the Armadillo Repeat Region of ␣-Catenin', 90, pp. 871–882.

Ikemura, T. (1985) 'Codon usage and tRNA content in unicellular and multicellular organisms.', *Molecular biology and evolution*, 2(1), pp. 13–34.

Ingram, V. M. (1956) 'A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin', *Nature*. Nature Publishing Group, 178(4537), p. 792.

Ishino, Y. *et al.* (1987) 'Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product.', *Journal of bacteriology*. Am Soc Microbiol, 169(12), pp. 5429–5433.

Ivanov, E. L. *et al.* (1996) 'Genetic requirements for the single-strand annealing pathway of double-strand break repair in Saccharomyces cerevisiae', *Genetics*, 142(3), pp. 693–704.

Ivics, Z. *et al.* (1997) 'Molecular reconstruction of sleeping beauty, a Tc1-like transposon from fish, and its transposition in human cells', *Cell*, 91(4), pp. 501–510. doi: 10.1016/S0092-8674(00)80436-5.

Jackson, A. L. and Linsley, P. S. (2010) 'Recognizing and avoiding siRNA off-target effects for target identification and therapeutic application', *Nature reviews Drug discovery*. Nature Publishing Group, 9(1), p. 57.

Jacobi, A. M. *et al.* (2017) 'Simplified CRISPR tools for efficient genome editing and streamlined protocols for their delivery into mammalian cells and mouse zygotes', *Methods*. The Authors, 121–122, pp. 16–28. doi: 10.1016/j.ymeth.2017.03.021.

Jansen, R. *et al.* (2002) 'Identification of genes that are associated with DNA repeats in prokaryotes', *Molecular Microbiology*, 43(6), pp. 1565–1575. doi: 10.1046/j.1365-2958.2002.02839.x.

Jinek, M. *et al.* (2012) 'A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity', *Science*. American Association for the Advancement of Science, p. 1225829.

Jinek, M. *et al.* (2013) 'RNA-programmed genome editing in human cells', *elife*. eLife Sciences Publications Limited, 2, p. e00471.

Kan, Y. *et al.* (2017) 'Mechanisms of precise genome editing using oligonucleotide donors', *Genome Research*, 27(7), pp. 1099–1111. doi: 10.1101/gr.214775.116.

Keightley, P. D. (2012) 'Rates and fitness consequences of new mutations in humans', *Genetics*, 190(2), pp. 295–304. doi: 10.1534/genetics.111.134668.

Kellis, M. *et al.* (2014) 'Defining functional DNA elements in the human genome', *Proceedings of the National Academy of Sciences*. National Acad Sciences, 111(17), pp. 6131–6138.

Ketting, R. F. (2011) 'The many faces of RNAi', *Developmental cell*. Elsevier, 20(2), pp. 148–161.

Khan, A. I. *et al.* (2011) 'Negative Epistasis Between Beneficial Mutations in an Evolving Bacterial Population Negative Epistasis Between Beneficial Mutations in an Evolving Bacterial Population Evolutionary theory predicts that epistatic', *Cooper Source: Science, New Series*, 332(60343), pp. 1193–1196. Available at: http://www.jstor.org/stable/27977986%0Ahttp://www.jstor.org/stable/27977986?seq=1&cid=pdf-reference#references_tab_contents%0Ahttp://about.jstor.org/terms.

Kiel, C. and Serrano, L. (2014) 'Structure-energy-based predictions and network modelling of RASopathy and cancer missense mutations', *Molecular systems biology*. EMBO Press, 10(5), p. 727.

Kielman, M. F. *et al.* (2002) 'Apc modulates embryonic stem-cell differentiation by controlling the dosage of β-catenin signaling', *Nature Genetics*, 32(4), pp. 594–605. doi: 10.1038/ng1045.

Kim, Y. B. *et al.* (2017) 'Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions', *Nature Biotechnology*. Nature Publishing Group, 35(4), pp. 371–376. doi: 10.1038/nbt.3803.

Kimura, M. (1968) 'Evolutionary rate at the molecular level', *Nature*, pp. 624–626. doi: 10.1038/217624a0.

Kirchner, T. and Brabletz, T. (2000) 'Patterning and Nuclear β-Catenin Expression in the Colonic Adenoma-Carcinoma Sequence', *The American Journal of Pathology*. American Society for Investigative Pathology, 157(4), pp. 1113–1121. doi: 10.1016/S0002-9440(10)64626-3.

Koboldt, D. C. *et al.* (2012) 'VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing', *Genome research*. Cold Spring Harbor Lab.

Koike-Yusa, H. *et al.* (2014) 'Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library', *Nature biotechnology*. Nature Publishing Group, 32(3), p. 267.

Komor, A. C. *et al.* (2016) 'Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage', *Nature*. Nature Publishing Group, 533(7603), pp. 420–424. doi: 10.1038/nature17946.

Komor, A. C., Badran, A. H. and Liu, D. R. (2017) 'CRISPR-Based Technologies for the Manipulation of Eukaryotic Genomes', *Cell*. Elsevier Inc., 169(3), p. 559. doi: 10.1016/j.cell.2017.04.005.

Konermann, S. *et al.* (2015) 'Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex', *Nature*. Nature Publishing Group, 517(7536), p. 583.

Konson, A. *et al.* (2010) 'Pigment epithelium-derived factor and its phosphomimetic mutant induce JNK-dependent apoptosis and p38-mediated migration arrest', *Journal of Biological Chemistry*. ASBMB, p. jbc-M110.

Kosicki, M., Tomberg, K. and Bradley, A. (2018) 'Repair of CRISPR–Cas9-induced double-stranded breaks leads to large deletions and complex rearrangements', *Nature Publishing Group*. Nature Publishing Group, (June). doi: 10.1038/nbt.4192.

Kosuri, S. *et al.* (2013) 'Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*', *Proceedings of the National Academy of Sciences*, 110(34), pp. 14024–14029. doi: 10.1073/pnas.1301301110.

Kotler, E. *et al.* (2018) 'A Systematic p53 Mutation Library Links Differential Functional Impact to Cancer Mutation Pattern and Evolutionary Conservation', *Molecular Cell*. Elsevier Inc., 71(1), p. 178–190.e8. doi: 10.1016/j.molcel.2018.06.012.

Kozich, J. J. *et al.* (2013) 'Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform', *Applied and environmental microbiology*. Am Soc Microbiol, 79(17), pp. 5112–5120. doi: 10.1128/AEM.01043-13.

Kramps, T. et al. (2002) 'Wnt/wingless signaling requires BCL9/ legless-mediated recruitment of pygopus to the nuclear beta-catenin- TCF complex.', *Cell*, 109(23), pp. 47–60.

Krieger, E., Koraimann, G. and Vriend, G. (2002) 'Increasing the precision of comparative models with YASARA NOVA—a self-parameterizing force field', *Proteins: Structure, Function, and Bioinformatics*. Wiley Online Library, 47(3), pp. 393–402.

Krueger, F. (2015) 'Trim galore', *A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files*.

Kudla, G. *et al.* (2009) 'Coding-sequence determinants of gene expression in Escherichia coli', *science*. American Association for the Advancement of Science, 324(5924), pp. 255–258.

Kuechler, A. *et al.* (2014) 'De novo mutations in beta-catenin (CTNNB1) appear to be a frequent cause of intellectual disability: expanding the mutational and clinical spectrum', *Human Genetics*, 134(1), pp. 97–109. doi: 10.1007/s00439-014-1498-1.

Kusserow, A. *et al.* (2005) 'Unexpected complexity of the Wnt gene family in a sea anemone', *Nature*, 433(7022), pp. 156–160. doi: 10.1038/nature03158.

Kuwajima, T. *et al.* (2013) 'ClearT: a detergent-and solvent-free clearing method for neuronal and non-neuronal tissue', *Development*. Oxford University Press for The Company of Biologists Limited, 140(6), pp. 1364–1368.

Laehnemann, D., Borkhardt, A. and McHardy, A. C. (2015) 'Denoising DNA deep sequencing data—high-throughput sequencing errors and their correction', *Briefings in bioinformatics*. Oxford University Press, 17(1), pp. 154–179.

Lamlum, H. *et al.* (1999) 'The type of somatic mutation at APC in familial adenomatous

polyposis is determined by the site of the germline mutation: A new facet to Knudson's "two-hit" hypothesis', *Nature Medicine*, 5(9), pp. 1071–1075. doi: 10.1038/12511.
Lander, E. S. (2011) 'Initial impact of the sequencing of the human genome', *Nature*. Nature Publishing Group, 470(7333), pp. 187–197. doi: 10.1038/nature09792.
Langmead, B. and Salzberg, S. L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nature methods*. Nature Publishing Group, 9(4), p. 357.
Larson, D. E. *et al.* (2011) 'SomaticSniper: identification of somatic point mutations in whole genome sequencing data', *Bioinformatics*. Oxford University Press, 28(3), pp. 311–317.
Laurin-Lemay, S., Philippe, H. and Rodrigue, N. (2018) 'Multiple Factors Confounding Phylogenetic Detection of Selection on Codon Usage.', *Molecular biology and evolution*, 35(6), pp. 1463–1472. doi: 10.1093/molbev/msy047.
Lazar, A. J. F. *et al.* (2008) 'Specific mutations in the β-catenin gene (CTNNB1) correlate with local recurrence in sporadic desmoid tumors', *The American journal of pathology*. Elsevier, 173(5), pp. 1518–1527.
Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows–Wheeler transform', *Bioinformatics*. Oxford University Press, 25(14), pp. 1754–1760.
Li, Z. *et al.* (1995) 'The XRCC4 gene encodes a novel protein involved in DNA double-strand break repair and V (D) J recombination', *Cell*. Elsevier, 83(7), pp. 1079–1089.
Liang, X. *et al.* (2017) 'Enhanced CRISPR/Cas9-mediated precise genome editing by improved design and delivery of gRNA, Cas9 nuclease, and donor DNA', *Journal of biotechnology*. Elsevier, 241, pp. 136–146.
Lieber, M. R. (2010) 'The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway', *Annual review of biochemistry*. Annual Reviews, 79, pp. 181–211.
Lin, S. *et al.* (2014) 'Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery.', *eLife*, 3, pp. 1–13. doi: 10.7554/eLife.04766.
Livingstone, C. D. and Barton, G. J. (1993) 'Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation', *Bioinformatics*. Oxford University Press, 9(6), pp. 745–756.
Loh, Y.-H. *et al.* (2006) 'The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells.', *Nature genetics*, 38(4), pp. 431–440.
Luo, G. *et al.* (1998) 'Chromosomal transposition of a Tc1/mariner-like element in mouse embryonic stem cells', *Proceedings of the National Academy of Sciences*. National Acad Sciences, 95(18), pp. 10769–10773.
Luo, J. *et al.* (2007) 'Wnt signaling and human diseases: what are the therapeutic implications?', *Laboratory Investigation*, 87(2), pp. 97–103. doi: 10.1038/labinvest.3700509.
Luria, S. E. and Delbrück, M. (1943) 'Mutations of bacteria from virus sensitivity to virus resistance', *Genetics*. Genetics Society of America, 28(6), p. 491.
Ma, L. *et al.* (2017) 'CRISPR-Cas9–mediated saturated mutagenesis screen predicts clinical drug resistance with improved accuracy', *Proceedings of the National Academy of Sciences*, p. 201708268. doi: 10.1073/pnas.1708268114.
Ma, Y. *et al.* (2016) 'Targeted AID-mediated mutagenesis (TAM) enables efficient genomic diversification in mammalian cells', *Nature Methods*. Nature Publishing Group, 13(12), pp. 1029–1035. doi: 10.1038/nmeth.4027.
Magoč, T. and Salzberg, S. L. (2011) 'FLASH: fast length adjustment of short reads to improve genome assemblies', *Bioinformatics*. Oxford University Press, 27(21), pp. 2957–2963.
Major, M. B. *et al.* (2007) 'Wilms Tumor Suppressor WTX Negatively Regulates WNT/b-

Catenin Signaling', *Science*, 316, pp. 1043–1046. doi: 10.1007/s13398-014-0173-7.2.
Makarova, K. S. *et al.* (2011) 'Evolution and classification of the CRISPR–Cas systems', *Nature Reviews Microbiology*. Nature Publishing Group, 9(6), p. 467.
Mali, P. *et al.* (2013) 'RNA-Guided Human Genome Engineering via Cas9 Prashant', *Science*, 339(6121), pp. 823–826. doi: 10.1126/science.1232033.RNA-Guided.
Marchese, F. P., Raimondi, I. and Huarte, M. (2017) 'The multidimensional mechanisms of long noncoding RNA function', *Genome biology*. BioMed Central, 18(1), p. 206.
Margulies, M. *et al.* (2005) 'Genome sequencing in microfabricated high-density picolitre reactors', *Nature*. Nature Publishing Group, 437(7057), p. 376.
Martin, G. R. (1981) 'Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells.', *Proceedings of the National Academy of Sciences*, 78(12), pp. 7634–7638. doi: 10.1073/pnas.78.12.7634.
Masella, A. P. *et al.* (2012) 'PANDAseq: paired-end assembler for illumina sequences', *BMC bioinformatics*. BioMed Central, 13(1), p. 31.
Mason, D. M. *et al.* (2018) 'High-throughput antibody engineering in mammalian cells by CRISPR / Cas9-mediated homology-directed mutagenesis', (June), pp. 1–14. doi: 10.1101/285015.
Meek, K., Dang, V. and Lees-Miller, S. P. (2008) 'DNA-PK: the means to justify the ends?', *Advances in immunology*. Elsevier, 99, pp. 33–58.
Megy, S. *et al.* (2006) 'STD and TRNOESY NMR studies for the epitope mapping of the phosphorylation motif of the oncogenic protein β-catenin recognized by a selective monoclonal antibody', *FEBS Letters*, 580(22), pp. 5411–5422. doi: 10.1016/j.febslet.2006.08.084.
Mehta, A. and Haber, J. E. (2014) 'Sources of DNA double-strand breaks and models of recombinational DNA repair', *Cold Spring Harbor perspectives in biology*. Cold Spring Harbor Lab, p. a016428.
Miller, J. C. *et al.* (2011) 'A TALE nuclease architecture for efficient genome editing', *Nature Biotechnology*, 29(2), pp. 143–150. doi: 10.1038/nbt.1755.
Millot, G. A. *et al.* (2012) 'A guide for functional analysis of BRCA1 variants of uncertain significance', *Human mutation*. Wiley Online Library, 33(11), pp. 1526–1537.
Mimitou, E. P. and Symington, L. S. (2010) 'Ku prevents Exo1 and Sgs1-dependent resection of DNA ends in the absence of a functional MRX complex or Sae2', *The EMBO journal*. EMBO Press, 29(19), pp. 3358–3369.
Miura, H. *et al.* (2015) 'CRISPR/Cas9-based generation of knockdown mice by intronic insertion of artificial microRNA using longer single-stranded DNA', *Scientific Reports*. Nature Publishing Group, 5(August), pp. 1–11. doi: 10.1038/srep12799.
Moens, C. B. *et al.* (2008) 'Reverse genetics in zebrafish by TILLING', *Briefings in Functional Genomics and Proteomics*. Oxford University Press, 7(6), pp. 454–459. doi: 10.1093/bfgp/eln046.
Mojica, F. J. M., García-Martínez, J. and Soria, E. (2005) 'Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements', *Journal of molecular evolution*. Springer, 60(2), pp. 174–182.
Molenaar, M. *et al.* (1996) 'XTcf-3 transcription factor mediates β-catenin-induced axis formation in xenopus embryos', *Cell*, 86(3), pp. 391–399. doi: 10.1016/S0092-8674(00)80112-9.
Morbitzer, R. *et al.* (2010) 'Regulation of selected genome loci using de novo-engineered transcription activator-like effector (TALE)-type transcription factors.', *Proceedings of the National Academy of Sciences of the United States of America*, 107(50), pp. 21617–21622.

Morris, K. V and Mattick, J. S. (2014) 'The rise of regulatory RNA', *Nature Reviews Genetics*. Nature Publishing Group, 15(6), p. 423.
Muller, H. J. (1927) 'Artificial Transmutation of the Gene', *Science*, 66(1699), pp. 84–87.
Muller, H. J. (1932) 'Further studies on the nature and causes of gene mutations.', *Proc. Sixth Int. Cong. Genet., Ithaca, New York, USA*, 1, pp. 213–255.
Myers, R. M., Tilly, K. and Maniatis, T. (1986) 'Fine structure genetic analysis of a beta-globin promoter.', *Science (New York, N.Y.)*, 232(4750), pp. 613–618. doi: 10.1126/science.3457470.
Nakamura, K. *et al.* (2011) 'Sequence-specific error profile of Illumina sequencers', *Nucleic Acids Research*, 39(13). doi: 10.1093/nar/gkr344.
Nalls, M. A. *et al.* (2014) 'Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease', *Nature genetics*. Nature Publishing Group, 46(9), p. 989.
Nejak-Bowen, K. N. *et al.* (2010) 'Accelerated liver regeneration and hepatocarcinogenesis in mice overexpressing serine-45 mutant β-catenin', *Hepatology*. Wiley Online Library, 51(5), pp. 1603–1613.
Nelson, W. J. and Nusse, R. (2004) 'Convergence of Wnt, b-Catenin, and Cadherin Pathways', *Science*, 303(March), pp. 1483–1488.
Nik-Zainal, S. *et al.* (2012) 'Mutational processes molding the genomes of 21 breast cancers', *Cell*. Elsevier, 149(5), pp. 979–993.
Nishida, K. *et al.* (2016) 'Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems', *Science*. American Association for the Advancement of Science, 353(6305), p. aaf8729. doi: 10.1126/science.aaf8729.
Nishimasu, H. *et al.* (2018) 'Engineered CRISPR-Cas9 nuclease with expanded targeting space', *Science*. American Association for the Advancement of Science, 361(6408), pp. 1259–1262.
Noordermeer, J. *et al.* (1994) 'dishevelled and armadillo act in the wingless signalling pathway in Drosophila.', *Nature*, 367(6458), pp. 80–83. doi: 10.1038/367080a0.
Nowell, P. C. (1976) 'The clonal evolution of tumor cell populations', *Science*. American Association for the Advancement of Science, 194(4260), pp. 23–28.
Nusslein-Volhard, C. and Wieschaus, E. (1980) 'Mutations affecting number and polarity in Drosophila', *Nature*, 287(5785), pp. 795–801.
Orford, K. *et al.* (1997) 'regulated Ubiquitination and Degradation of ᴺʟ -Catenin *', *October*, pp. 24735–24738.
van Overbeek, M. *et al.* (2016) 'DNA repair profiling reveals nonrandom outcomes at Cas9-mediated breaks', *Molecular cell*. Elsevier, 63(4), pp. 633–646.
Paix, A. *et al.* (2017) 'Precision genome editing using synthesis-dependent repair of Cas9-induced DNA breaks', *Proceedings of the National Academy of Sciences*, (14), p. 201711979. doi: 10.1073/pnas.1711979114.
Paix, A., Schmidt, H. and Seydoux, G. (2016) 'Cas9-assisted recombineering in C. elegans: Genome editing using in vivo assembly of linear DNAs', *Nucleic Acids Research*, 44(15), p. e128. doi: 10.1093/nar/gkw502.
Palovcak, A. *et al.* (2017) 'Maintenance of genome stability by Fanconi anemia proteins', *Cell and Bioscience*. BioMed Central, 7(1), pp. 1–18. doi: 10.1186/s13578-016-0134-2.
Pannunzio, N. R. *et al.* (2014) 'Non-homologous end joining often uses microhomology: implications for alternative end joining', *DNA repair*. Elsevier, 17, pp. 74–80.
Park, J. *et al.* (2017) 'Cas-analyzer: an online tool for assessing genome editing results using NGS data', *Bioinformatics*. Oxford University Press, 33(2), pp. 286–288.

Patel, P. and Woodgett, J. R. (2017) 'Glycogen synthase kinase 3: a kinase for all pathways?', in *Current topics in developmental biology*. Elsevier, pp. 277–302.

Patrick, W. M. and Firth, A. E. (2005) 'Strategies and computational tools for improving randomized protein libraries', *Biomolecular Engineering*, 22(4), pp. 105–112. doi: 10.1016/j.bioeng.2005.06.001.

Peden, J. F. (1999) 'Analysis of codon usage', *Bio Systems*, 5(1), pp. 45–50. doi: 10.1016/j.biosystems.2011.06.005.

Peifer, M., McCrea, P.D., Green, K.J., Wieschaus, E., and Gumbiner, B. M. (1992) 'The vertebrate adhesive junction proteins beta-catenin and plakoglobin and the Drosophila segment polarity gene armadillo form a multigene family with similar properties', *Journal of Cell Biology*, 118, pp. 681–691.

Peifer, M. *et al.* (1991) 'The segment polarity gene armadillo interacts with the wingless signaling pathway in both embryonic and adult pattern formation.', *Development (Cambridge, England)*, 111(4), pp. 1029–43. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1879348.

Peifer, M., Berg, S. and Reynolds, A. B. (1994) 'A repeating amino acid motif shared by proteins with diverse cellular roles', *Cell*, pp. 789–791. doi: 10.1016/0092-8674(94)90353-0.

Pesole, G. (2008) 'What is a gene? An updated operational definition', *Gene*. Elsevier, 417(1), pp. 1–4.

Pfeifer, G. P. *et al.* (2002) 'Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers', *Oncogene*. Nature Publishing Group, 21(48), p. 7435.

Pfeifer, G. P. (2010) 'Environmental exposures and mutational patterns of cancer genomes', *Genome medicine*. BioMed Central, 2(8), p. 54.

Phelps, R. A. *et al.* (2009) 'A Two-Step Model for Colon Adenoma Initiation and Progression Caused by APC Loss', *Cell*. Elsevier Ltd, 137(4), pp. 623–634. doi: 10.1016/j.cell.2009.02.037.

Pinder, J., Salsman, J. and Dellaire, G. (2015) 'Nuclear domain "knock-in"screen for the evaluation and identification of small molecule enhancers of CRISPR-based genome editing', *Nucleic acids research*. Oxford University Press, 43(19), pp. 9379–9392.

Pinello, L. *et al.* (2016) 'Analyzing CRISPR genome-editing experiments with CRISPResso', *Nature biotechnology*. Nature Publishing Group, 34(7), p. 695.

Plasterk, R. H. A. (1996) 'The Tc1/mariner transposon family', in *Transposable elements*. Springer, pp. 125–143.

Pleasance, E. D. *et al.* (2010) 'A small-cell lung cancer genome with complex signatures of tobacco exposure', *Nature*. Nature Publishing Group, 463(7278), p. 184.

Plotkin, J. B. and Kudla, G. (2011) 'Synonymous but not the same: The causes and consequences of codon bias', *Nature Reviews Genetics*. Nature Publishing Group, 12(1), pp. 32–42. doi: 10.1038/nrg2899.

Polakis, P. (2000) 'Wnt signaling and cancer Wnt signaling and cancer', *Genes Dev.*, 14(650), pp. 1837–1851. doi: 10.1101/gad.14.15.1837.

Porteus, M. H. and Carroll, D. (2005) 'Gene targeting using zinc finger nucleases.', *Nature biotechnology*, 23(8), pp. 967–973.

Pourcel, C., Salvignol, G. and Vergnaud, G. (2005) 'CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies', *Microbiology*. Microbiology Society, 151(3), pp. 653–663.

Pouyet, F. *et al.* (2017) 'Recombination, meiotic expression and human codon usage', *eLife*, 6, pp. 1–19. doi: 10.7554/eLife.27344.

Powell, L. M. *et al.* (1987) 'A novel form of tissue-specific RNA processing produces

apolipoprotein-B48 in intestine', *Cell*. Elsevier, 50(6), pp. 831–840.
Prasher, D. C. *et al.* (1992) 'Primary structure of the Aequorea victoria green-fluorescent protein', *Gene*. Elsevier, 111(2), pp. 229–233.
Presnyak, V. *et al.* (2015) 'Codon optimality is a major determinant of mRNA stability', *Cell*. Elsevier, 160(6), pp. 1111–1124.
Qi, L. S. *et al.* (2013) 'Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression', *Cell*. Elsevier, 152(5), pp. 1173–1183.
Quail, M. A. *et al.* (2012) 'A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers', *BMC genomics*. BioMed Central, 13(1), p. 341.
Quax, T. E. F. *et al.* (2015) 'Codon bias as a means to fine-tune gene expression', *Molecular cell*. Elsevier, 59(2), pp. 149–161.
Rad, R. *et al.* (2010) 'PiggyBac transposon mutagenesis: a tool for cancer gene discovery in mice', *Science*. American Association for the Advancement of Science, 330(6007), pp. 1104–1107.
Radecke, F. *et al.* (2006) 'Targeted chromosomal gene modification in human cells by single-stranded oligodeoxynucleotides in the presence of a DNA double-strand break', *Molecular Therapy*. Elsevier, 14(6), pp. 798–808.
Radecke, S. *et al.* (2010) 'Zinc-finger nuclease-induced gene repair with oligodeoxynucleotides: wanted and unwanted target locus modifications', *Molecular Therapy*. Elsevier, 18(4), pp. 743–753.
Rakoczy, E. P. *et al.* (2011) 'Analysis of disease-linked rhodopsin mutations based on structure, function, and protein stability calculations', *Journal of molecular biology*. Elsevier, 405(2), pp. 584–606.
Ran, F. A. *et al.* (2013) 'Double nicking by RNA-guided CRISPR cas9 for enhanced genome editing specificity', *Cell*. Elsevier, 154(6), pp. 1380–1389. doi: 10.1016/j.cell.2013.08.021.
Ran, F. A. *et al.* (2013) 'Genome engineering using the CRISPR-Cas9 system.', *Nature protocols*, 8(11), pp. 2281–308. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3969860&tool=pmcentrez&rendertype=abstract.
Ravanat, J.-L., Douki, T. and Cadet, J. (2001) 'Direct and indirect effects of UV radiation on DNA and its components', *Journal of Photochemistry and Photobiology B: Biology*. Elsevier, 63(1–3), pp. 88–102.
Rebouissou, S. *et al.* (2016) 'Genotype-phenotype correlation of CTNNB1 mutations reveals different ß-catenin activity associated with liver tumor progression', *Hepatology*, 64(6), pp. 2047–2061. doi: 10.1002/hep.28638.
Rees, H. A. *et al.* (2017) 'Improving the DNA specificity and applicability of base editing through protein engineering and protein delivery', *Nature Communications*. Nature Publishing Group, 8, pp. 1–10. doi: 10.1038/ncomms15790.
Reetz, M. T. *et al.* (2010) 'Iterative Saturation Mutagenesis Accelerates Laboratory Evolution of Enzyme Stereoselectivity: Rigorous Comparison with Traditional Methods', *Journal of the American Chemical Society*. American Chemical Society, 132(26), pp. 9144–9152. doi: 10.1021/ja1030479.
Renaud, J.-B. *et al.* (2016) 'Improved genome editing efficiency and flexibility using modified oligonucleotides with TALEN and CRISPR-Cas9 nucleases', *Cell reports*. Elsevier, 14(9), pp. 2263–2272.
Rentzsch, P. *et al.* (2018) 'CADD: predicting the deleteriousness of variants throughout the human genome', *Nucleic acids research*.

Reya, T. *et al.* (2001) 'Stem cells, cancer, and cancer stem cells', 414(November), pp. 105–111.

Reya, T. and Clevers, H. (2005) 'Wnt signalling in stem cells and cancer', *Nature*, 434(7035), pp. 843–850. doi: 10.1038/nature03319.

Rich, A. and RajBhandary, U. L. (1976) 'Transfer RNA: molecular structure, sequence, and properties', *Annual review of biochemistry*. Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, 45(1), pp. 805–860.

Richardson, C. D. *et al.* (2016) 'Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA', *Nature Biotechnology*. Nature Publishing Group, 34(3), pp. 339–344. doi: 10.1038/nbt.3481.

Richardson, C. D. *et al.* (2016) 'Non-homologous DNA increases gene disruption efficiency by altering DNA repair outcomes', *Nature Communications*. Nature Publishing Group, 7, pp. 1–7. doi: 10.1038/ncomms12463.

Richardson, C. D. *et al.* (2018) 'CRISPR-Cas9 genome editing in human cells works via the Fanconi Anemia pathway', *Nature Genetics*. Springer US, 50(August). doi: 10.1101/136028.

Rocklin, G. J. *et al.* (2017) 'Global analysis of protein folding using massively parallel design, synthesis, and testing', *Science*. American Association for the Advancement of Science, 357(6347), pp. 168–175.

Romero, P. A. and Arnold, F. H. (2009) 'Exploring protein fitness landscapes by directed evolution', *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, 10(12), pp. 866–876. doi: 10.1038/nrm2805.

Roose, J. *et al.* (1998) 'The Xenopus Wnt effector XTcf-3 interacts with Groucho-related transcriptional repressors', *Nature*, 395(6702), pp. 608–612. doi: 10.1038/26989.

Rosenfeld, J. A., Malhotra, A. K. and Lencz, T. (2010) 'Novel multi-nucleotide polymorphisms in the human genome characterized by whole genome and exome sequencing', *Nucleic Acids Research*, 38(18), pp. 6102–6111. doi: 10.1093/nar/gkq408.

Rosin-Arbesfeld, R. *et al.* (2005) 'Nuclear export of the APC tumour suppressor controls beta-catenin function intranscription.', *The EMBO Journal*, 22(5), pp. 1101–1113.

Ross, M. G. *et al.* (2013) 'Characterizing and measuring bias in sequence data', *Genome biology*. BioMed Central, 14(5), p. R51.

Rouet, P., Smih, F. and Jasin, M. (1994) 'Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease.', *Molecular and cellular biology*. Am Soc Microbiol, 14(12), pp. 8096–8106.

Rudolph, K. L. M. *et al.* (2016) 'Codon-driven translational efficiency is stable across diverse mammalian cell states', *PLoS genetics*. Public Library of Science, 12(5), p. e1006024.

Saleh-Gohari, N. and Helleday, T. (2004) 'Conservative homologous recombination preferentially repairs DNA double-strand breaks in the S phase of the cell cycle in human cells', *Nucleic acids research*. Oxford University Press, 32(12), pp. 3683–3688.

Sarkisyan, K. S. *et al.* (2016) 'Local fitness landscape of the green fluorescent protein', *Nature*. Nature Publishing Group, 533(7603), pp. 397–401. doi: 10.1038/nature17995.

Sato, N. *et al.* (2004) 'Maintenance of pluripotency in human and mouse embryonic stem cells through activation of Wnt signaling by a pharmacological GSK-3-specific inhibitor', *Nature medicine*. Nature Publishing Group, 10(1), p. 55.

Sato, T. *et al.* (2011) 'Paneth cells constitute the niche for Lgr5 stem cells in intestinal crypts', *Nature*. Nature Publishing Group, 469(7330), pp. 415–418. doi: 10.1038/nature09637.

Scally, A. and Durbin, R. (2012) 'Revising the human mutation rate: implications for understanding human evolution', *Nature Reviews Genetics*. Nature Publishing Group, 13(10), p. 745.

Schirmer, M. *et al.* (2015) 'Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform', *Nucleic Acids Research*, 43(6). doi: 10.1093/nar/gku1341.

Schneeberger, K. (2014) 'Using next-generation sequencing to isolate mutant genes from forward genetic screens', *Nature Reviews Genetics*. Nature Publishing Group, 15(10), pp. 662–676. doi: 10.1038/nrg3745.

Schymkowitz, J. *et al.* (2005) 'The FoldX web server: An online force field', *Nucleic Acids Research*, 33(SUPPL. 2).

Seeberg, E., Eide, L. and Bjørås, M. (1995) 'The base excision repair pathway', *Trends in biochemical sciences*. Elsevier, 20(10), pp. 391–397. doi: 10.1016/S0968-0004(00)89086-6.

Shalem, O. *et al.* (2014) 'Genome-scale CRISPR-Cas9 knockout screening in human cells', *Science*. American Association for the Advancement of Science, 343(6166), pp. 84–87.

Sharp, P. M. and Li, W.-H. (1986) 'The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications', *Nucleic Acids Research*, 16(2), pp. 719–738.

Shimomura, O., Johnson, F. H. and Saiga, Y. (1962) 'Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan, Aequorea', *Journal of Cellular Physiology*. Wiley Online Library, 59(3), pp. 223–239.

Shimotohno, A. *et al.* (2001) 'Demonstration of the importance and usefulness of manipulating non-active-site residues in protein design', *The Journal of Biochemistry*. The Japanese Biochemistry Society, 129(6), pp. 943–948.

Sim, N.-L. *et al.* (2012) 'SIFT web server: predicting effects of amino acid substitutions on proteins', *Nucleic acids research*. Oxford University Press, 40(W1), pp. W452–W457.

Singer, M. and Soll, D. (1973) 'Guidelines for DNA Hybrid Molecules Author', *Science*, 181(4105), p. 1114.

Slaymaker, I. M. *et al.* (2015) 'Rationally engineered Cas9 nucleases with improved specificity', *Science*. American Association for the Advancement of Science, p. aad5227.

de Smit, M. H. and Van Duin, J. (1990) 'Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis.', *Proceedings of the National Academy of Sciences*. National Acad Sciences, 87(19), pp. 7668–7672.

Smith, A. G. *et al.* (1988) 'Inhibition of pluripotential embryonic stem cell differentiation by purified polypeptides', *Nature*, 336(6200), pp. 688–690. doi: 10.1038/336688a0.

Smith, J. *et al.* (2006) 'A combinatorial approach to create artificial homing endonucleases cleaving chosen sequences', *Nucleic acids research*. Oxford University Press, 34(22), pp. e149–e149.

Smith, J. M. (1970) 'Natural selection and the concept of a protein space', *Nature*. Nature Publishing Group, 225(5232), p. 563.

Soldner, F. *et al.* (2011) 'Generation of isogenic pluripotent stem cells differing exclusively at two early onset parkinson point mutations', *Cell*, 146(2), pp. 318–331. doi: 10.1016/j.cell.2011.06.019.

Sørensen, M. A., Kurland, C. G. and Pedersen, S. (1989) 'Codon usage determines translation rate in Escherichia coli', *Journal of molecular biology*. Elsevier, 207(2), pp. 365–377.

Spiller, B. *et al.* (1999) 'A structural view of evolutionary divergence', *Proceedings of the National Academy of Sciences*. National Acad Sciences, 96(22), pp. 12305–12310.

Spradling, A. C. *et al.* (1995) 'Gene disruptions using P transposable elements: an integral component of the Drosophila genome project', *Proceedings of the National Academy of Sciences*. National Acad Sciences, 92(24), pp. 10824–10830.

Starita, L. M. *et al.* (2018) 'ARTICLE A Multiplex Homology-Directed DNA Repair Assay Reveals

the Impact of More Than 1 , 000 BRCA1 Missense Substitution Variants on Protein Function', *The American Journal of Human Genetics*. ElsevierCompany., 103(4), pp. 498–508. doi: 10.1016/j.ajhg.2018.07.016.

Stemmer, W. P. C. *et al.* (1995) 'Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides', *Gene*. Elsevier, 164(1), pp. 49–53.

Stenson, P. D. *et al.* (2017) 'The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies', *Human genetics*. Springer, 136(6), pp. 665–677.

Stepanenko, O. V *et al.* (2013) 'Beta-barrel scaffold of fluorescent proteins: folding, stability and role in chromophore formation', in *International review of cell and molecular biology*. Elsevier, pp. 221–278.

Stratton, M. R., Campbell, P. J. and Futreal, P. A. (2009) 'The cancer genome', *Nature*. Nature Publishing Group, 458(7239), p. 719.

Sturtevant, A. H. and Morgan, T. H. (1923) 'Reverse mutation of the Bar gene correlated with crossing over', *Science*. American Association for the Advancement of Science, 57(1487), pp. 746–747.

Supek, F. *et al.* (2014) 'Synonymous mutations frequently act as driver mutations in human cancers', *Cell*, 156(6), pp. 1324–1335. doi: 10.1016/j.cell.2014.01.051.

Suzuki, K. *et al.* (2016) 'In vivo genome editing via CRISPR/Cas9 mediated homology-independent targeted integration', *Nature*, 540(7631), pp. 144–149. doi: 10.1038/nature20565.

Swenberg, J. A. *et al.* (2011) 'Endogenous versus exogenous DNA adducts: Their role in carcinogenesis, epidemiology, and risk assessment', *Toxicological Sciences*, 120(SUPPL.1), pp. 130–145. doi: 10.1093/toxsci/kfq371.

Takada R, Satomi Y, Kurata T, Ueno N, Norioka S, Kondoh H, Takao T, T. S. (2006) 'Monounsaturated fatty acid modification of Wnt protein: its role in Wnt secretion.', *Developmental Cell*, 11(6), pp. 791–801.

Tan, E. *et al.* (2015) 'Off-target assessment of CRISPR-C as9 guiding RNA s in human i PS and mouse ES cells', *genesis*. Wiley Online Library, 53(2), pp. 225–236.

Tang, T.-H. *et al.* (2002) 'Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus', *Proceedings of the National Academy of Sciences*. National Acad Sciences, 99(11), pp. 7536–7541.

Taverna, D. M. and Goldstein, R. A. (2002) 'Why are proteins so robust to site mutations? 1', *Journal of molecular biology*. Elsevier, 315(3), pp. 479–484.

Tennessen, J. A. *et al.* (2012) 'Evolution and functional impact of rare coding variation from deep sequencing of human exomes', *science*. American Association for the Advancement of Science, 337(6090), pp. 64–69.

The International SNP Map Working Group (2001) 'A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms', *Nature*. Macmillian Magazines Ltd., 409(6822), pp. 928–933. Available at: http://dx.doi.org/10.1038/35057149.

Till, B. J. *et al.* (2003) 'Large scale discovery of induced point mutations with high throughput TILLING', *Genome Research*, 13, pp. 524–530. doi: 10.1101/gr.977903.

Timoféeff-Ressovsky, N., Zimmer, K. and Delbrück, M. (1935) 'The nature of genetic mutations and structure of the gene', *Nachrichten aus der Biologie der Gesellschaft der Wissenschaften Göttingen*, 1, pp. 189–245.

Tsai, S. Q. *et al.* (2015) 'GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases', *Nature biotechnology*. Nature Publishing Group, 33(2), p. 187.

Tsien, R. Y. (1998) 'The green fluorescent protein'. Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA.

Tuazon, P. T. (1991) 'Casein kinase I and II-multipotential serine protein kinases: structure, function, and regulation.', *Adv. Second Messenger Phosphoprotein Res.*, 23, pp. 123–164.

Urnov, F. D. *et al.* (2010) 'Genome editing with engineered zinc finger nucleases', *Nature Reviews Genetics*. Nature Publishing Group, 11(9), pp. 636–646. doi: 10.1038/nrg2842.

do Valle, Í. F. *et al.* (2016) 'Optimized pipeline of MuTect and GATK tools to improve the detection of somatic single nucleotide polymorphisms in whole-exome sequencing data', *BMC bioinformatics*. BioMed Central, 17(12), p. 341.

Vartak, S. V and Raghavan, S. C. (2015) 'Inhibition of nonhomologous end joining to increase the specificity of CRISPR/Cas9 genome editing', *The FEBS journal*. Wiley Online Library, 282(22), pp. 4289–4294.

Venter, J. C. *et al.* (2001) 'The sequence of the human genome.', *Science (New York, N.Y.)*, 291(5507), pp. 1304–51. doi: 10.1126/science.1058040.

Visscher, P. M. *et al.* (2017) '10 Years of GWAS Discovery: Biology, Function, and Translation', *American Journal of Human Genetics*. ElsevierCompany., 101(1), pp. 5–22. doi: 10.1016/j.ajhg.2017.06.005.

De Vries, H. (1914) 'The principles of the theory of mutation', *Science*. JSTOR, 40(1020), pp. 77–84.

Walker, D. R. *et al.* (1999) 'Evolutionary conservation and somatic mutation hotspot maps of p53: Correlation with p53 protein structural and functional features', *Oncogene*, 18(1), pp. 211–218. doi: 10.1038/sj.onc.1202298.

Wan, W. *et al.* (2017) *Immobilized muts-mediated error removal of microchip-synthesized DNA*, *Methods in Molecular Biology*. doi: 10.1007/978-1-4939-6343-0_17.

Wang, H. *et al.* (2001) 'Genetic evidence for the involvement of DNA ligase IV in the DNA-PK-dependent pathway of non-homologous end joining in mammalian cells', *Nucleic acids research*. Oxford University Press, 29(8), pp. 1653–1660.

Wang, T. *et al.* (2014) 'Genetic screens in human cells using the CRISPR-Cas9 system', *Science*. American Association for the Advancement of Science, 343(6166), pp. 80–84.

Wang, X. *et al.* (2017) 'CRISPR-DAV: CRISPR NGS data analysis and visualization pipeline', *Bioinformatics*, 33(23), pp. 3811–3812. doi: 10.1093/bioinformatics/btx518.

Wang, Z., Vogelstein, B. and Kinzler, K. W. (2003) 'Phosphorylation of β-catenin at S33, S37, or T41 can occur in the absence of phosphorylation at T45 in colon cancer cells', *Cancer Research*, 63(17), pp. 5234–5235.

Watson, J. D. and Crick, F. H. C. (1953) 'Molecular structure of nucleic acids', *Nature*, 171(4356), pp. 737–738.

Van de Wetering, M. *et al.* (1997) 'Armadillo coactivates transcription driven by the product of the Drosophila segment polarity gene dTCF', *Cell*, 88(6), pp. 789–799. doi: 10.1016/S0092-8674(00)81925-X.

Wienholds, E. *et al.* (2003) 'Efficient target-selected mutagenesis in zebrafish', *Genome research*. Cold Spring Harbor Lab, 13(12), pp. 2700–2707.

Williams, R. L. *et al.* (1988) 'Myeloid leukaemia inhibitory factor maintains the developmental potential of embryonic stem cells.', *Nature*, 336(6200), pp. 684–687. doi: 10.1038/336684a0.

Winston, J. T. *et al.* (1999) 'The SCF ␤ -TRCP – ubiquitin ligase complex associates specifically with phosphorylated destruction motifs in I ␬ B ␣ and ␤ -catenin and stimulates I ␬ B ␣ ubiquitination in vitro Jefferey T . Winston , Peter Strack , Peggy Beer-Romero , Claire Y . Chu , S', *Genes & Development*, 283, pp. 9369–9369. doi: 10.1016/j.immuni.2005.02.009.

Wolfe, K. H., Sharp, P. M. and Li, W.-H. (1989) 'Mutation rates differ among regions of the mammalian genome', *Nature*. Nature Publishing Group, 337(6204), p. 283.
Wolfe, S. A., Nekludova, L. and Pabo, C. O. (2000) 'DNA recognition by Cys2His2 zinc finger proteins', *Annual review of biophysics and biomolecular structure*. Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, 29(1), pp. 183–212.
Wood, A. J. *et al.* (2011) 'Targeted genome editing across species using ZFNs and TALENs', *Science*, 333(6040), p. 307. doi: 10.1126/science.1207773.
Wu, G. *et al.* (2003) 'Structure of a β-TrCP1-Skp1-β-catenin complex: Destruction motif binding and lysine specificity of the SCFβ-TrCP1ubiquitin ligase', *Molecular Cell*, 11(6), pp. 1445–1456. doi: 10.1016/S1097-2765(03)00234-X.
Wu, Q. *et al.* (2017) 'In situ functional dissection of RNA cis-regulatory elements by multiplex CRISPR-Cas9 genome engineering', *Nature communications*. Nature Publishing Group, 8(1), p. 2109.
Wu, X. *et al.* (2008) 'Rac1 Activation Controls Nuclear Localization of β-catenin during Canonical Wnt Signaling', *Cell*, 133(2), pp. 340–353. doi: 10.1016/j.cell.2008.01.052.
Xing, Y. *et al.* (2008) 'Crystal Structure of a Full-Length β-Catenin', *Structure*, 16(3), pp. 478–487. doi: 10.1016/j.str.2007.12.021.
Yang, L. *et al.* (2016) 'Engineering and optimising deaminase fusions for genome editing', *Nature communications*. Nature Publishing Group, 7, p. 13330.
Yang, T. T., Cheng, L. and Kain, S. R. (1996) 'Optimized codon usage and chromophore mutations provide enhanced sensitivity with the green fluorescent protein', *Nucleic Acids Research*, 24(22), pp. 4592–4593. doi: 10.1093/nar/24.22.4592.
Ying, Q. L. *et al.* (2008) 'The ground state of embryonic stem cell self-renewal', *Nature*, 453(7194), pp. 519–523. doi: 10.1038/nature06968.
Yoshimi, K. *et al.* (2016) 'ssODN-mediated knock-in with CRISPR-Cas for large genomic regions in zygotes', *Nature communications*. Nature Publishing Group, 7, p. 10431.
Zecca, M., Basler, K. and Struhl, G. (1996) 'Direct and long-range action of a wingless morphogen gradient', *Cell*, 87(5), pp. 833–844. doi: 10.1016/S0092-8674(00)81991-1.
Zerbino, D. R. *et al.* (2017) 'Ensembl 2018', *Nucleic acids research*. Oxford University Press, 46(D1), pp. D754–D761.
Zetsche, B. *et al.* (2015) 'Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system', *Cell*. Elsevier, 163(3), pp. 759–771.
Zhang, J. *et al.* (2014) 'PEAR: A fast and accurate Illumina Paired-End reAd mergeR', *Bioinformatics*, 30(5), pp. 614–620. doi: 10.1093/bioinformatics/btt593.
Zheng, L., Baumann, U. and Reymond, J.-L. (2004) 'An efficient one-step site-directed and site-saturation mutagenesis protocol.', *Nucleic acids research*, 32(14), p. e115.
Zhou, B.-B. S. and Elledge, S. J. (2000) 'The DNA damage response: putting checkpoints in perspective', *Nature*. Nature Publishing Group, 408(6811), p. 433.
Zhou, Y. *et al.* (2014) 'High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells', *Nature*. Nature Publishing Group, 509(7501), p. 487.
Zhou, Z. *et al.* (2016) 'Codon usage is an important determinant of gene expression levels largely through its effects on transcription', *Proceedings of the National Academy of Sciences*. National Acad Sciences, 113(41), pp. E6117–E6125.