THE UNIVERSITY
of EDINBURGH

Protein Binding as a Selective Filter for New Mutations at Regulatory Sites in the Germline and in Cancers

Author:
Lana Talmane

Supervisor:
Prof. Martin Taylor

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy
in the
University of Edinburgh

2019

MRC Human Genetics Unit
The University of Edinburgh
igmm Institute of Genetics & Molecular Medicine

Declaration of Authorship

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree. Work done in collaboration with, or with assistance of, others, is indicated as such.

Signed:

Date:

Abstract

Genetic mutations provide the raw material for evolution, they are responsible for heritable disease and drive the development of cancer. It has been previously shown that the binding of chromatin and regulatory proteins to DNA can interfere with replication, surveillance and repair processes but the proposed mechanisms presume the loading of sequence-specific binding factors over nucleotide mismatches and other lesions. This seems paradoxical for binders that recognise their docking sites by motif with defined sequence. In this work I propose the biased mask model where the binding of some transcription factors can tolerate mismatch substitutions or other lesions strand specifically at some sites, acting as a selective filter of new mutations. I provide electrophoretic mobility shift assay support for the biased mask, and illustrate how it is shaping the mutation patterns of both cancers and the human germline. Being replication associated, the mutational burden of this biased mask predicts that the protein binding sites occupied during germline replication are hotspots for functionally important mutations, which will be exacerbated by increased paternal age. Exploring this, in collaboration with other group we have isolated and applied chromatin accessibility assay, ATAC-seq, to primary human and mouse spermatogonial cells, which account for up to 80% of human and 30% of mouse germline DNA replication. I have used this data to develop a custom ATAC-seq processing pipeline and map protein binding landscape of the germline, and also of a number of somatic tissues for which ATAC-seq data was available. By combining this map with human and mouse population variation data I confirm sequence specific binding sites in germline as hotspots of deleterious mutations, and provide evidence that this mutational effect is dependent on protein binding.

Lay Abstract

Mutations are changes in the identity of individual building blocks of DNA, which is a molecule that carries instructions for the correct structure and function of a cell. Mutations occurring in cells that lead to the formation of sperm in males and oocytes in females (germline cells) can cause diseases that will be passed on to the next generation. The occurrence of mutations in other cells of the body can lead to various diseases, including cancer. Preservation of special sites on the DNA where proteins can attach (protein binding sites) is important for the correct function of the cell. In this work we isolated highly dividing germline cells where we anticipate that the majority of heritable mutations are occurring. I have identified locations of protein binding sites in those cells, as well as in a number of other tissue types. Here I show that binding sites that are active in the germline exhibit a high frequency of mutations. I provide evidence that those mutations are deleterious and are likely to cause hereditary disease. I provide examples of such mutations in cancers and demonstrate that those mutations are likely caused by the physical interaction of protein and DNA. I propose a model that explains how proteins are able to cause harmful mutations at the sites on the DNA with which they interact, and provide experimental evidence for this model.

vi

Acknowledgements

First and foremost, I would like to acknowledge my supervisor, Martin Taylor, for giving me an opportunity to learn (many!) new things I had no idea about, and guiding me through my scientific journey over the past three and a half years. Your immense enthusiasm for this work and support at every step of the way can not be given enough credit. Thank you for inspiring confidence, having faith in me, and being kind.

I would like to thank Robert Young, for his day-to-day supervision, infinite support (both research-related and moral), answering in detail to (a lot!) of silly questions, and all the chats. And, of course, your meaningful and insightful inputs into this project. Other members of the Taylor group: Craig Anderson - for your helpful and insightful comment on my thesis; Thomas Williams - for your moral support and occasional sharing of PhD life frustrations; Juliet Luft - for putting up with all my fidgeting during the write-up process. My second supervisor - Wendy Bickmore, and rest of my thesis committee - Ian Adams and Martin Reijns, for all your helpful suggestions and comments. Everyone in Evogen group for all the help, and all the friends within IGMM.

Many thanks go to Abdenour Soufi and whole of Soufi group, who have adopted me for the length of wet-lab part of this project, and in particular Gareth Roberts, who baby-set me in the lab and taught me how to perform EMSAs. Yatendra Kumar, Lizzie Freyer, Marie MacLennan, Fiona Semple and Susan Campbell - for experimental work that went into this thesis. Surgeons Roland Donat and CJ Shukla, as well as all the patients that have kindly donated their tissue. MRC, for funding me during the course of past four years.

Janahan, for putting up with and getting me through my occasional mental melt-downs, and also for all the dinners waiting at home. My parents, Tatjana and Jurijs, who have provided me with every opportunity to succeed, and without whose support I would never be where I am now. You have always given me freedom me do

what I wanted and supported me in these decisions, even if that meant moving away to an island far away, and for that I am immensely grateful.

Contents

List of Figures

List of Tables

Abbreviations

(f:) Formula.

AF All Fragment.

ATAC-seq Assay for Transposase-Accessible Chromatin using sequencing.

ChIP-seq Chromatin Immunoprecipitation Sequencing.

DHSF DNase-seq Overlap.

DNA Deoxyribonucleic Acid.

DNase-seq DNase I hypersensitive sites sequencing.

EMSA Electrophoretic Mobility Shift Assay.

FACS Fluorescence-Activated Cell Sorting.

FLOP Fragment Length Occurrence Propensity.

FMM Forward Mismatch.

GERP Genomic Evolutionary Rate Profiling.

GST Glutathione S-transferase.

ICGC International Cancer Genome Consortium.

LAU Linear Arbitrary Units.

MMR Mismatch Repair.

MMRd Mismatch Repair deficient.

MUT Mutation.

NER Nucleotide Excision Repair.

PE Phycoerythrin.

PWM Position Weight Matrix.

RMM Reverse Mismatch.

RNA Ribonucleic Acid.

SF Short Fragment.

SSC Spermatogonial Stem Cell.

TF Transcription Factor.

TSS Transcriptional Start Site.

UV Ultraviolet.

WGS Whole Genome Sequencing.

YFP Yellow Fluorescent Protein.

CHAPTER 1

Introduction

The central dogma of genetics postulates that the flow of information in a cell is a step-wise conversion of instructions from the level of deoxyribonucleic acids (DNA) to ribonucleic acids (RNA) and then to proteins (Crick, 1958). The storage and inheritance of information in the cell is a prerogative of the DNA in the form of a quaternary code. Individual units that the DNA consists of are four types of bases – A (*adenine*), T (*thymine*), C (*cytosine*) and G (*guanine*). Those bases come together to form long stretches of instructions encoding the components and organisation of a living organism. This is our genome.

Genomes are normally faithfully replicated and inherited - the reason progeny tend to resemble parents. However, changes to the order or identity of DNA bases do occur and are termed mutations. While mutations provide a new material for selection to act upon and drive evolution, they are more often deleterious than advantageous and can result in disease (Kimura, 1968, 1991). Mutations come in different shapes and sizes, including base changes (substitutions); the insertion or deletion of a single or small number of bases (indels), to larger scale re-ordering with sequences 'flipped around' (inversions); or relocated to a different part of the genome altogether (translocations). Normal endogenous processes within a healthy cell can lead to the generation of new mutations but the rate of mutation can be dramatically increased when DNA is damaged by exogenous agents, such as tobacco smoke or ultraviolet radiation (Chatterjee and Walker, 2017). There are processes within the cells that act to counteract this and repair the damage.

The distribution of mutations across the genome is far from uniform (Makova and Hardison, 2015), and uncovering the mutational patterns and processes that shape them is important for understanding how diversity arises, how various types of genetic diseases occur (*e.g.* cancer) (Alexandrov, 2018) or in what ways the process of ageing is likely to affect the organism's fitness (Garinis et al., 2008).

Our genome is akin to a globe, with continents and oceans, countries, seas and deserts. Mutations can be meteors falling down from the sky and depending on the type of area they impact, the consequences will differ. Mutations occurring within the coding regions of genes, termed exons, can affect the functionality of the resulting protein, while the ones occurring outside coding sequences might have an effect on regulation of expression of a nearby gene, or possibly no biological consequence at all (falling into the ocean far away from anything). Some areas get hit much more frequently than

others. The frequency with which a certain position or an area gets mutated is termed the mutation rate. The human germline mutation rate has previously been estimated to be $\approx 1 - 1.5$ x10^{-8} mutations per site per generation: a child is expected to have between 70 and 80 new mutations that were not inherited by their parents (Michaelson et al., 2012; Kong et al., 2012; Francioli et al., 2015), however if we were to travel along the genome, we would find that there is a great regional variability (Makova and Hardison, 2015). To understand fluctuations in the mutation rate we must consider a manifold of states that can be ascribed to every region of the genome, spatial and temporal organisation of different cellular processes and the context of the underlying sequence, as well as the physical nature and properties of bases.

In the introduction to this work I will cover the general overview of DNA and genome structure, how genetic information is read and expressed, and how it is organised within a cell (Section 1.1). The vast part of the genome is thought to be non-coding (International Human Genome Sequencing Consortium, 2001) and a portion of it plays a role in regulation of gene expression (Dunham et al., 2012), acting as docking sites for the proteins that bind to specific sequences and manage the expression of genes. Hence I will describe why non-coding sequence preservation can be important and how the sequence-specific binding of certain proteins regulates gene expression and thereby a phenotypic outcome (Section 1.2). I will describe how cells ensure that high fidelity copy of the genome gets passed on to the daughter cells (Section 1.3), and discuss the variability in the mutation rate across the genome associated with aspects of replication, but also with other features that underlie the difference in the mutation rates (Section 1.4). I shall then go on to introduce the role of mutations in a particular group of cells termed 'germline cells', the only cell population in multicellular organisms that is able to propagate its genomic information through to the next generation. I will also describe the importance of the mutations that occur in cells of the body termed 'somatic cells' that never make it to the progeny, but are hugely important for fitness and survival of the organism itself. I will give an overview of how multiple repair mechanisms work to try and counteract those mutational processes, how their activity fluctuates across the genome and hence contributes to the variability of the genome-wide mutational landscape (Section 1.5).

Genetics poses many questions. How does all the variability we see around us come about? What are the role of specific or stochastic processes in shaping variability?

And how can we utilize knowledge of those processes to improve the quality of human life? The work described here touches on all aspects of this, looking at whether certain changes in the composition of our genome are likely to have any functional impact (in relation to the regulation of the gene expression by the proteins such as transcription factors), at how those changes come about (what are the mechanistic causes of those changes), and provides some information that would help to identify where those changes generally are more likely to happen.

1.1 Genome structure

1.1.1 Basic blocks of life

One hallmark of a living organism, together with prerequisites of consisting of one or multiple cells, responding to environmental stimuli and capacity to convert energy, is an ability to reproduce, creating another life form that is similar, and in some cases, identical to the original entity (Koshland Jr., 2002). This is achieved by utilization of a molecule termed DNA, which is commonly referred to as a 'blueprint' of the cell, because it contains information that is necessary to 'build' a cell/organism, maintain it, and respond to external stimuli. Information encoded within the DNA can be copied, which allows supply of identical, or nearly-identical, DNA molecules to each of individual entities of a divided cell.

DNA consists of two strands of polymers that are made up of four types of monomer units termed 'nucleotides' - adenines (A), thymines (T), guanines (G), and cytosines (C) (Figure 1.1) in a manner that has been proposed and described by Watson and Crick (1953) more than 60 years ago. Each of the nucleotides consists of a 5-carbon sugar, a nitrogenous base, and one phosphate group, and can be classified as ether *purine*, where nucleotide base contains two carbon-nitrogen rings, or *pyrimidine*, containing one. Strands that make up the DNA duplex have opposite directionality. Those two polymeric strands are connected to each other by hydrogen bonds in a specific manner, where a pyrimidine-containing nucleotide on one strand only pairs up with a complimentary purine on the other strand. Thymine and cytosine will normally pair up only with adenine and guanine, respectively. The total set of those nucleotides, or base pairs, present in a cell is termed a *genome* and the human genome contains approximately 3 billion of those base pairs, and most cells contain two copies of the genome.

1.1.2 Higher order structures

The complete genome is not distributed randomly across the cell, nor is it contained within a single molecule. In humans, it is unevenly divided between 23 intricately compacted units, termed *chromosomes*. The number of chromosomes varies between

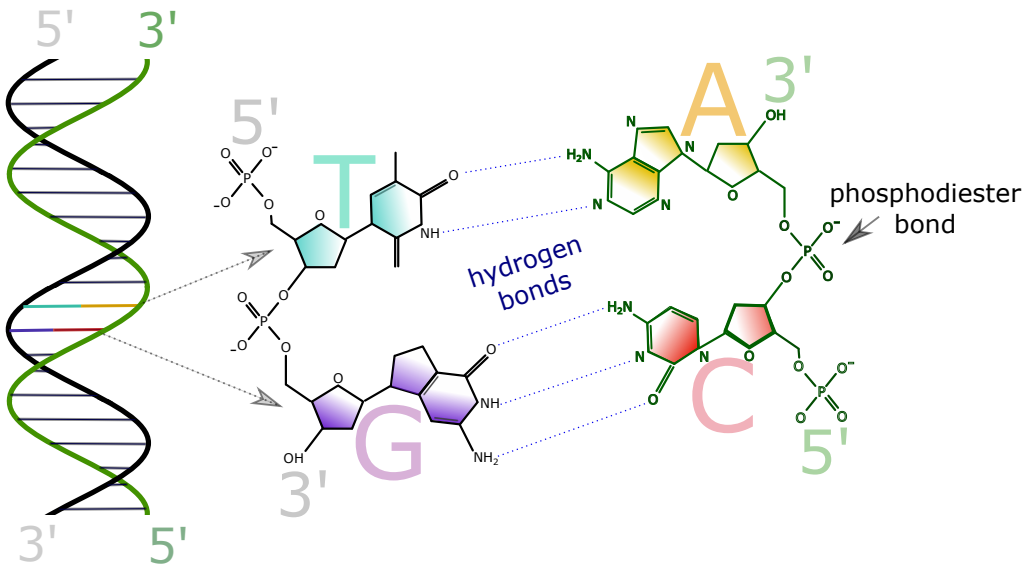


Figure 1.1: DNA consists of two strands of polymers that are made up of four types of monomer nucleotide units - adenines (yellow), thymines (blue), guanines (purple), and cytosines (red). Each of the nucleotides consists of a 5-carbon sugar, a nitrogenous base, and one phosphate group. Nucleotides are connected to each other in a linear strand by covalent phosphodiester bonds that form between the phosphate group of one nucleotide and the hydroxyl group of another. The end containing the hydroxyl group is termed 3', and the end containing phosphate group is termed 5'. Those two strands are connected to each by hydrogen bonds in a specific manner, where thymine normally only pairs with adenine, and cytosine pairs with guanine.

species, *e.g.* mice have 20. Each cell of the human organism contain two copies of each chromosome, in which case the genome is said to be *diploid*. Exceptions to this are egg and sperm cells, where only one copy of each chromosome is present, in which case it is said to be in a *haploid* state. While 22 of the chromosome pairs, which are called *autosomes*, are more or less similar to each other, the individual chromosomes of the 23^{rd} pair (*allosomes*) can differ, and are involved in determination of sex. There are two types of allosomes - X and Y. Individuals that possesses two copies of the X chromosome are defined genetically as a female, while individuals that have one of both the X and Y chromosomes are defined genetically as a male.

Chromosomes are compacted into ordered structures with a help of *nucleosomes*, that in turn consist of *histones* (reviewed in Cutter and Hayes (2015)) (Figure 1.2). Each nucleosome contains 8 histone proteins, which are positively charged. Due to the fact that DNA is negatively charged, it interacts with the histones well. Ap-

proximately 147 base pairs (bp) of DNA wrapped around a nucleosome, together with a special type of linker histone (H1/H5) that binds at $\approx 20-80$ bp region between nucleosomes, forms a *chromatosome*. Those structures then fold up to form 30 nm fibres, which in turn form higher order structures that are tightly coiled to form a chromatid (Li and Reinberg, 2011). Collectively, the DNA-protein complex is termed *chromatin*. In humans, sister chromatids are held together at the region called *centromere*. Ends of the chromosomes are called *telomeres*, which are DNA-protein complexes that act to prevent fusion between chromosomes, and their degradation (O'Sullivan and Karlseder, 2010).

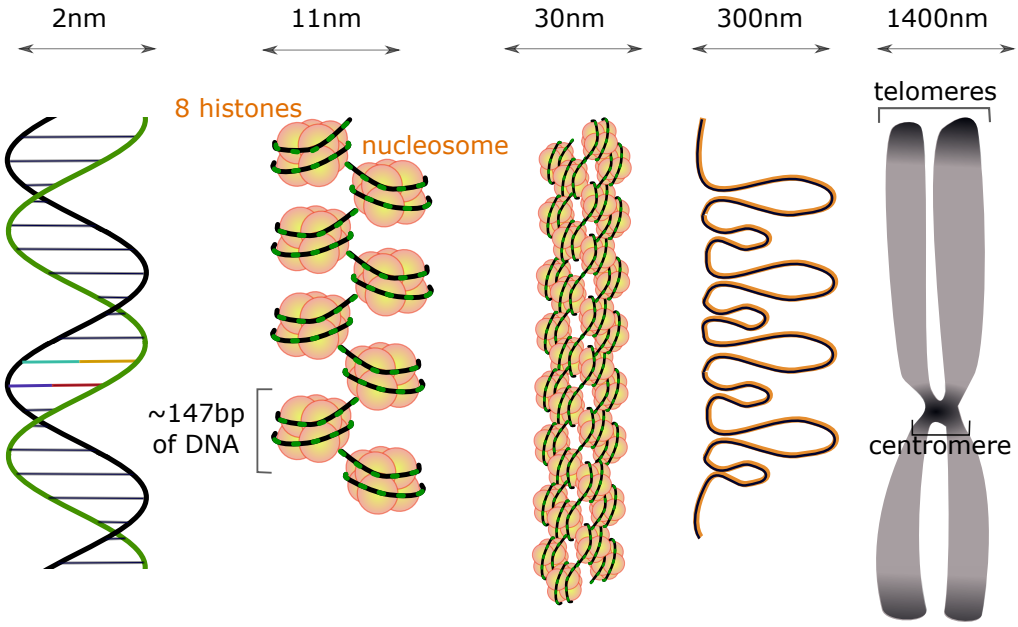


Figure 1.2: Higher order chromatin structure. Approximately 147 base pairs of DNA wrap around eight core histone proteins, which make-up a nucleosome unit, and these form 11 nm fibres. Nucleosomes are compacted to form 30 nm chromatin fibres which then form loops averaging $\approx$ 300 nm in length and are further compressed and coiled to form a chromosome.

1.1.3 The coding and non-coding genome

DNA encodes instructions for the synthesis of proteins out of amino acids. A segment of DNA that codes for a particular type of protein is called a *protein-coding gene*. For production of that protein one of the DNA strands containing the gene (template strand) is 'read' by an *RNA polymerase II* enzyme to produce a *messenger RNA* molecule (pre-mRNA) (Brenner et al., 1961), which is complementary to that DNA strand using

similar base pairing rules to double-stranded DNA, in a process called *transcription*. Genetic positions towards the 5' end of the template strand are referred to as being *upstream*, while the ones towards the 3' end - *downstream*. Transcription can only proceed from 3' to 5' on the template strand, with RNA growing in 5' $\longrightarrow$ 3' direction (nucleotides are always added to the 3' end). The chemical difference between DNA and RNA is that the latter contains a ribose sugar backbone, instead of deoxyribose, and it is single-stranded. In addition to that, thymine is replaced by its analogue uracil, which does not contain a methyl group.

Pre-mRNA contains within itself two types of sequences - *exons* and *introns* (Chow et al., 1977; Berk and Sharp, 1977). Exons are part of the transcript that contain information about the structure of a protein, while introns are intervening sequences in between exons that do not code for protein products. Introns are excised from pre-mRNA in a process termed *splicing*. In the process of splicing, not only introns, but also exons can be excised, which leads to a single gene being able to produce several alternative *transcripts*, and therefore multiple protein products. An mRNA molecule is then processed by a *ribosome* (Palade, 1955), which constructs a protein from the polymerisation of amino acids in an order dictated by the RNA nucleotide sequence, in a process termed *translation*. Each nucleotide triplet sequence falling into non-overlapping windows, termed *codons*, defines that a specific amino acid be added to the growing peptide (Nirenberg et al., 1966). Considering that there are 4 nucleotide types and 64 possible triplet sequences, they could potentially code for that many amino acids. However, there is a degree of redundancy, and multiple codons represent the same amino acid, thereby limiting the number of possible options to 20. This redundancy mainly comes from the third position within the codon, and is commonly referred to as *third position wobble*, as in most cases change in the nucleotide identity at that position does not alter the incorporated amino acid. Some of the codons act as a mechanism that instructs translation to cease. Conversely, a special type of codon, called a 'start codon', that codes for the amino acid methionine, signals the position from which a ribosome should initiate translation.

Only a small proportion of the human genome is thought to be coding (less than 2%) (International Human Genome Sequencing Consortium, 2001). The rest does not code for any currently known proteins and consists of various different elements (Dunham et al., 2012), and a large portion of the genome is transcribed and generates

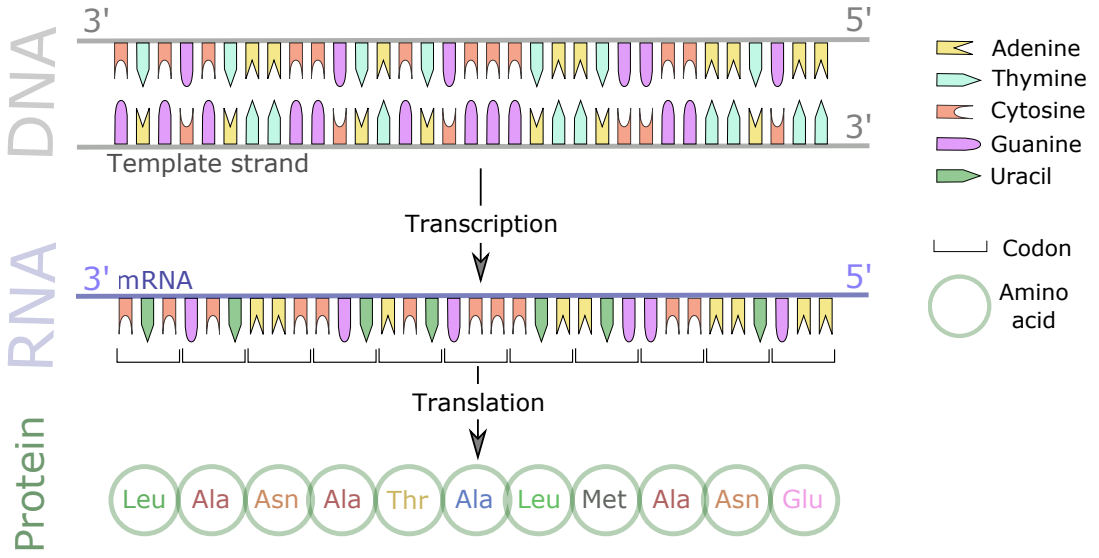


Figure 1.3: Transcription and translation. One of the two strands of DNA, which contains a gene going in the 3'⟶5' direction - a template strand - is copied by RNA polymerase to produce a messenger RNA (mRNA) molecule that is single-stranded and contains uracils in place of thymines. mRNA is then processed by a ribosome that assembles amino acids into a peptide to form a protein according to the information encoded on mRNA in chunks of three nucleotides termed codons.

different forms of RNA molecules (Forrest et al., 2014). This includes ribosomal RNAs (rRNAs), which play an important part in the ribosome structure, transfer RNAs (tRNAs) that serve as a link between mRNA and amino acids, as well as several others such as long non-coding RNAs (lncRNAs) and small nuclear RNAs (snRNAs). Those regulate many processes, including gene expression, with many functions still unknown (reviewed in Cech and Steitz (2014)). Transposable elements - sequences that have an ability to move to different locations in the genome; pseudogenes - genes that have lost their functionality; and satellite DNA - arrays of tandemly repeating sequences also contribute towards a vast part of human non-coding genome. One of the most important functions of the non-coding genome is regulation. With a limited number of protein-coding genes and their similarity between species, it has been proposed that the differences between organisms stems in a large part from the variation in regulation (King and Wilson, 1975; Wittkopp and Kalay, 2012), therefore regulatory sequence elements within the non-coding genome are recognized to carry an important role.

1.1.4 Chromosomal domains, compartments and their interactions

Not all of the chromatin is compacted in the same way and to the same degree. It can be broadly divided into *euchromatin* and *heterochromatin*. Euchromatin is less compacted permitting DNA accessibility for frequently transcribed genes and as a result, most of the genes tend to be located in euchromatic regions. Heterochromatin is a more tightly compacted and tends to be located towards the periphery of the nucleus. Constitutive heterochromatin tends to assemble at repetitive elements, while facultative heterochromatin can harbour developmentally regulated genes, compaction of which can vary depending on the stimuli (Wang et al., 2016).

Chromosomes are thought to reside within distinct locations throughout the nucleus, *chromosome territories* (CT) (Cremer et al., 2001). Individual chromosome segments have also been suggested to be organised into *topologically associated domains* (TADs), though to be conserved across different cells (Dixon et al., 2012). TADs are regions of the genome that tend to interact with each other more than with the other regions. It is not clear whether larger scale TADs are similar to smaller-scale structures present within them, such as sub-TADs and individual chromatin loops, leading to the idea that variable organization on smaller level structures might be associated with cell-specific gene expression (Dixon et al. (2016)). Binding sites of the main architectural protein of chromosomes - CTCF - have been found to be enriched at TAD boundaries, consistent with its role as an insulator, which together with cohesin protein complex brings close distant genomic regions to form loops, and is a key player in TAD organisation (Dixon et al., 2016). Loss of cohesin complex has been shown to eliminate formation of loops and TAD domains, but only modestly affects expression of active genes (Rao et al., 2017; Schwarzer et al., 2017).

Recent studies have revealed that distinct genomic megabase-sized compartments (which are roughly equivalent to closed and open chromatin) form without CTCF/cohesin involvement (Rao et al., 2017; Schwarzer et al., 2017; Seitan et al., 2013). Those compartments are characterised by particular chromatin marks and have been speculated to arise due to formation of a 'compartment globule' mediated by either bridging of proximal nucleosomes by proteins (polymer-polymer phase separation), or stabilization by proteins that exhibit multivalent interactions among each other (liquid-liquid phase-separation) (Erdel and Rippe, 2018). Association of compartments with

particular chromatin marks is a possible underlying reason for recruitment of various elements, such as DNA repair systems whose targeting is known to be linked to histone modifications (House et al., 2014), to certain sub-compartment regions.

Therefore, these two modes of 3D architecture (loops/TADs and compartments) are thought to comprise two distinct layers of chromatin organisation, which can be acting in opposing manner. For example, a TAD might be bringing together two loci that would normally reside in separate sub-compartments (Rao et al., 2017; Schwarzer et al., 2017).

1.2 Transcription factors and regulation of gene expression

1.2.1 Regulatory elements in the genome

There are several types of trans- and cis- regulatory elements, previously suggested to constitute up to 80% of the human genome (Dunham et al., 2012). Trans-regulators mainly consist of proteins and various non-protein coding RNAs (reviewed in Cech and Steitz (2014)), while cis-acting elements contain, among others, promoters, enhancers, insulators and silencers. Those elements can often be characterised by the presence of specific chromatin marks - modifications on the histones that are in an immediate proximity to those elements, what is often referred to as a *histone code* (Strahl and Allis, 2000). More than 500 different histone modification have been characterised (Zhao and Garcia, 2015). Different histone marks might be present on a particular type of elements depending on whether they are in active or inactive states, or compartments. Those elements, when active, are located within the accessible part of the genome because various proteins, termed *transcription factors* (TFs), require access to the DNA sequence to bind. Recent attempts at characterisation of the regulatory elements in the non-coding genome have resulted in large-scale generation and aggregation of multiple types of data in project such as ENCODE (Dunham et al., 2012) and Roadmap Epigenomics (Roadmap Epigenomics Consortium et al., 2015).

1.2.2 Transcription factors

TFs are proteins that regulate transcription so that it occurs in the right cells at the right time. While different TFs can be simplistically classified as either activators or repressors, depending on whether they promote or block transcription, some TFs have been found to be both, with their function being dependent on context (Schmitges et al., 2016). They interact with DNA in a sequence-specific manner through their DNA-binding domain (DBD). Modes of action among different TFs can vary - while some act by recruitment of co-factors, or even RNA polymerase itself, binding of some is actually aimed at occlusion of specific sequences to prevent other TFs or nucleosomes from binding.

While TFs are generally thought of as binding to the nucleosome depleted DNA, where binding motifs tend to be more exposed, there is a specific class of TFs, termed *pioneer factors* (Takahashi and Yamanaka, 2006), which are able to bind nucleosomal DNA and promote chromatin opening. Ability of those proteins to bind might be due to the position of their binding motif, that is displayed on the nucleosome surface (Soufi et al., 2015). This process permits subsequent binding of other TFs, and demonstrate their dynamic interplay for controlling gene expression.

Evolution of TF activity can be driven by alterations in protein-protein interactions and sequence specificity, and levels of TF expression (Schmitges et al., 2016). Proteins that contain a C2H2 zinc finger DBDs are the largest family of TFs in vertebrates (Hughes, 2011). In a recent review by Lambert et al. (2018), authors have undertaken a survey of human TFs, and have re-analysed and extended the previously compiled lists of 355 (Fulton et al., 2009) and 1,391 (Vaquerizas et al., 2009) human TFs to 1,639 known or likely transcription factors, most of which are C2H2 zinc finger proteins.

1.2.3 Promoters

Promoters are a class of regulatory element that are normally within $\approx$ 500bp upstream of TSS and are regions where a transcription initiation complex forms. Promoters of expressed genes generally show open conformation and are bound by multiple TFs that recruit RNA polymerase II and other components necessary for transcription to the *core promoter* (approximately 50bp upstream and downstream of the TSS) (Haberle and Stark, 2018). Sites of transcription initiation can be defined by various techniques, such as cap analysis of gene expression (CAGE), which works by capture of the 5' end cap of mRNAs (Shiraki et al., 2003).

Because active promoters harbour a dynamic range of TFs and transcription machinery, they are considered to be nucleosome-depleted. Though rather than being completely devoid of them, they might instead harbour more 'dynamic' nucleosomes, that contain specific histone variants (Jin et al., 2009; Young et al., 2017). For differential regulation of gene expression, some promoters exhibit open conformation, to allow gene transcription, while others remain closed to suppress it. This is frequently accompanied by a change in the histone modifications such as trimethylation of histone H3 Lys4 (H3K4me3) and acetylation of H3 Lys27 (H3K27ac), which are active marks,

or trimethylation of H3 Lys27 (H3K27me3), an inactive mark (reviewed in Elkon and Agami (2017)). Interestingly, promoters at 5' ends of genes encoding for key developmental TFs have been found to be kept in a 'poised' state in embryonic stem cells, marked with both active and repressive chromatin marks (Bernstein et al., 2006). A subset of *housekeeping* promoters can be consistently found open/active in multiple cell types, and those are likely to accompany genes that are essential for basic cell maintenance and function. Promoters that are only found to be active in a specific tissue are termed *tissue-specific* promoters and are more likely to control expression of genes with a specialized role. Promoter regions tend to be enriched in CpG dinucleotides relative to the rest of the genome, and overlap with CpG islands (Gardiner-Garden and Frommer, 1987; Deaton and Bird, 2011). This tends to be more pronounced at housekeeping promoters, presumably because of their hypomethylated status in the germline (Saxonov et al., 2006), as methylated CpGs frequently undergo C $\longrightarrow$T change (Youssoufian et al., 1986), which is predominantly a mammalian feature.

1.2.4 Enhancers

The role of promoters in the regulation of transcription is complemented by activity of more distal elements termed *enhancers*. Enhancers are regions that are not necessarily located in linear proximity to the TSS, or co-oriented with the direction of transcription (Bulger and Groudine, 2011). Instead they are thought to interact with promoters through the three-dimensional organisation of chromatin, regulated through TADs (Subsection 1.1.4). Enhancers may also be bound by TFs and co-factors, and interactions with promoters can have an effect on the transcription of promoter-proximal genes (Figure 1.4).

In recent years the notion of *super enhancers* has emerged, which defines large genomic regions several kilobases in length, which bind multiple TFs at increased density and are responsible for driving cell type-specific gene expression (reviewed in Pott and Lieb (2015)). Initially defined as master regulators of pluripotent embryonic stem cells (ESCs), and characterised by a large number of binding sites for pluripotency-associated TFs (such as KLF4, OCT4, SOX2, NANOG, and ESRRB), and Mediator complex, super-enhancers have also been found in other cell types (Whyte et al., 2013). Enhancer elements are commonly considered to exhibit acetylation of histones, lower ratio of tri- versus mono- methylation on the histone H3 lysine 4 (H3K4me3

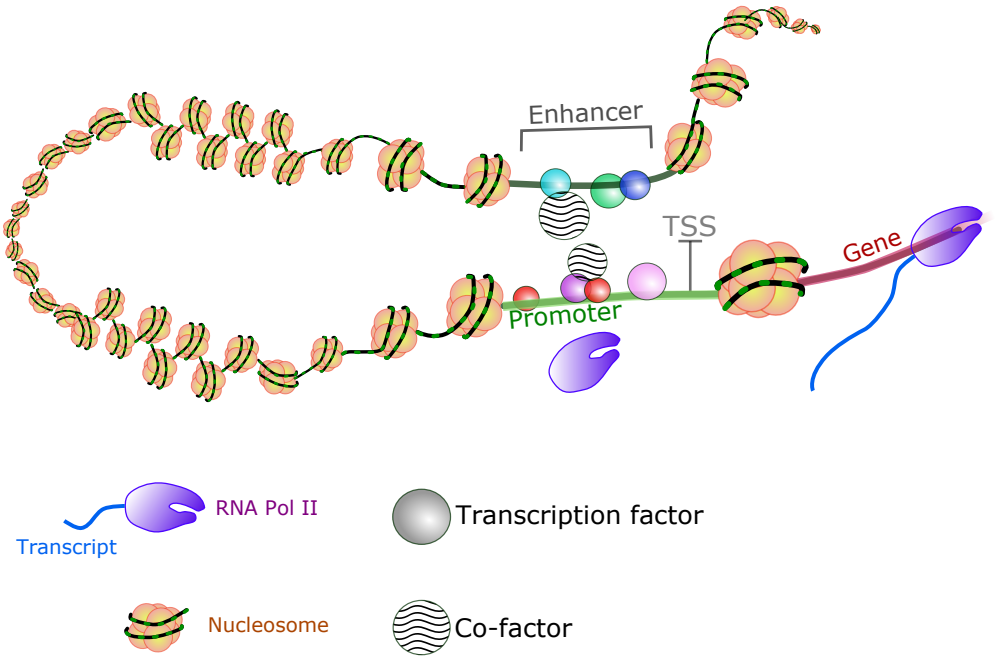


Figure 1.4: Regulation of gene expression. The core promoter constitutes a region approximately 50bp upstream and downstream of the transcription start site (TSS), where RNA polymerase and TFs bind. TFs regulate gene expression and can recruit other co-factors. Enhancers are regions that can be bound by TFs, which can be located far from the TSS on linear DNA, but are brought into proximity to a promoter through the three-dimensional arrangement of chromatin. Enhancer-promoter interaction can further regulate gene expression.

and H3K4me1, respectively), and presence of particular histone variants, such as H3.3 and H2A.Z (Calo and Wysocka, 2013). Transcription, often bidirectional and resulting in production of short and unstable transcripts, has been found to occur at enhancers (Kim et al., 2010). While the presence of bidirectional transcription has previously been proposed to mark enhancer activity, it has also been shown to be a general property of accessible regions, rather than a mark of functionality, and not specific to enhancer elements (Young et al., 2017).

1.2.5 Identification of protein-binding sites

There have been multiple techniques and methods developed for identification of transcription factor binding sites (TFBSs) and sequence motifs associated with those. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is one of the most widely used techniques for TF binding site identification (Johnson et al., 2007) and involves

cross-linking a protein to DNA using formaldehyde following precipitation with an antibody that is specific in recognising the target TF. It does have a caveat in the form of detection of indirect binding due to protein-protein interactions, and does not measure equilibrium binding due to the cross-linking step. ChIP-seq results constitute rather wide 'peaks' that define sites where TFs are bound, but lack single nucleotide-resolution of protein-DNA interactions. ChIP-exo is a variation of ChIP-seq which with addition of the exonuclease digestion step gives a much more precise measure of protein binding (Rhee and Pugh, 2011), but to date has been implemented in relatively small number of studies. Regions defined by ChIP-seq can then be scanned to look for recurrent sequences, or motifs, that TFs preferentially bind to.

More general techniques, aimed at identification of the open chromatin include DNase I hypersensitive sites sequencing (DNase-seq) (Song and Crawford, 2010), formaldehyde-assisted isolation of regulatory elements (FAIRE-seq) (Simon et al., 2012), and more recently developed assay for transposase-accessible chromatin using sequencing (ATAC-seq) (Buenrostro et al., 2013). Those techniques take advantage of the sensitivity of nucleosome-bound DNA to formaldehyde cross-linking (in case of FAIRE-seq), and enzyme accessibility of the open chromatin (DNase-seq and ATAC-seq), whereas DNA that is bound by nucleosomes or other proteins is protected from digestion. These techniques can also be used for the identification of the transcription-factor binding sites.

The ability of proteins to protect interacting DNA from digestion is a general idea employed in methods for identification of TFBSs, and their detection is commonly called 'footprinting'. Multiple methods and software have been developed for identification of binding sites from DNase-seq data (comparative analysis of such software has been done by Gusmao et al. (2016)), and less so for ATAC-seq, mainly due to former being available for longer. While theoretically DNase-seq methods for footprinting can be applied to ATAC-seq data as well, properties of two types of data can differ, primarily due to the different sequence biases associated with the enzymes used for digestion. Difficulty in evaluating the performance of any of the methods is the lack of a ultimate 'gold standard': ChIP-seq results are generally useful in this respect, but the variability in site occupation dynamics between cells, and the limited number of TFs for which ChIP-seq has been performed, can be prohibitive.

1.3 DNA Replication

As mentioned in the beginning, one of the hallmarks of live organism is its ability to procreate. In order to achieve this, genetic information in form of DNA has to be copied in the process termed *replication*, followed by a cell division. Replication takes place at the *S phase* (synthesis) of the interphase that precedes *mitosis*, the division into two daughter cells (Flemming, 1879). During S phase, all of the chromosomes are duplicated to form two identical copies. This is followed by prophase, when chromosomes condense, metaphase, when chromosomes align along the cell equator, anaphase, when sister chromatids separate and are pulled apart, and telophase, after which point one cell becomes two individual *daughter cells* upon undergoing cytokinesis.

1.3.1 Replicative asymmetry

The process of replication occurs in a semi-conservative manner (Meselson and Stahl, 1958), which means that each of the two DNA strands serve as a template, are paired with newly synthesised strands of DNA and separated into daughter cells. Replication initiates in an organised manner at sites within the genome, termed the *origins of replication*, which fire at specific times during S phase of the cell cycle (Jacob et al., 1963; Burgers and Kunkel, 2017). Initiation of this process leads to separation of two strands with the help of specific protein complexes (Moyer et al., 2006). As replication origins are not located at the ends of the chromosomes, this causes the formation of the *replication bubbles* with two *replication forks* proceeding in opposite directions (Figure 1.5). As new nucleotides can only be added to the 3' end of the growing strand, replication is also said to be *asymmetric*. While synthesis of one of the new strands can proceed continuously, essentially following the replication fork, the other strand has to be synthesised in a discontinuous manner, in stretches of short *Okazaki fragments*, a couple of hundred base pairs at a time (Okazaki et al., 1968). The former is termed as the *leading strand*, while the latter is known as the *lagging strand*.

Class of enzymes termed *DNA polymerases* are responsible for synthesis of new DNA (reviewed in Johansson and Dixon (2013)). There are several DNA polymerases that are primarily responsible for replication in eukaryotes - Polα, Polδ, and Polε. Possessing a primase activity - the ability to create a RNA primer, Polα is responsible

for initiation of both leading and lagging strands at the origins of replication, and also for every Okazaki fragment on the lagging strand thereafter (Perera et al., 2013). Initial stretches synthesised by Polα are then extended by Polε on the leading strand and Polδ on the lagging strand (Stith et al., 2008). Some evidence suggests that Polδ might also be involved in initial steps of leading strand replication (Daigaku et al., 2015; Garbacz et al., 2018). On lagging strand, Polα-synthesised DNA is normally replaced by Polδ which is synthesising preceding Okazaki fragment. It has been shown that a substantial proportion of Polα-synthesised DNA ($\approx$ 1.5%) is retained in the mature genome post-replication (Reijns et al., 2015).

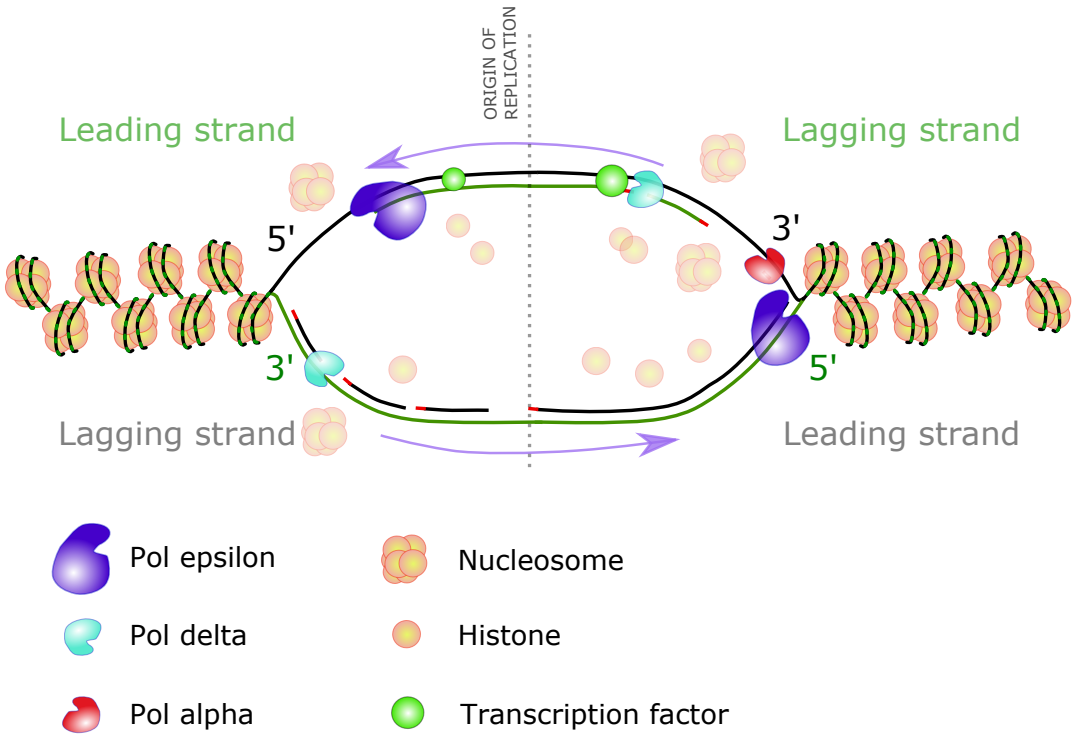


Figure 1.5: Replication initiates bidirectionally from multiple origins across the genome. Nucleotides can only be added to the 3' end of the growing strand, therefore one of the strands (leading) can be synthesised continuously, while the other one (lagging) has to be synthesised in stretches of discontinuous Okazaki fragments. Nucleosomes have to be disassembled in front of the replication fork and reassembled on the newly synthesised duplex. Transcription factors similarly have to be removed as the fork progresses.

1.3.2 Replication and chromatin

As mentioned in Section 1.1, DNA is compacted by being wrapped around nucleosomes, and in nucleosome-depleted regions it is bound by other proteins such as TFs (Section 1.2). All of those present obstacles to the progression of the replication fork if they remain bound to DNA during replication. Therefore they must dissociate and then re-associate on the newly synthesised duplexes. Due to the doubling of the DNA, twice as many histones are needed to preserve the same chromatin compaction as was present on the original duplex and a mixture of old and newly synthesised histones are laid down behind the replication fork (Lai and Pugh, 2017). Histone marks present on the old histones are copied over to the new ones, thereby preserving the histone code (Petryk et al., 2018). In yeast, Okazaki fragment termini are enriched at the nucleosome dyad positions and at some of the protein-binding sites, suggesting that Okazaki fragment processing is connected to nucleosome assembly and TF binding (Smith and Whitehouse, 2012). The presence of only one copy of the TF per two newly-synthesised genomes has lead to speculation that the distinct patterns of gene expression in daughter cells can be the result (Whitehouse and Smith, 2013). Reassembled nucleosomes have been shown to lack precise positioning immediately following replication, and ascertain a more stable location after undergoing maturation (Ramachandran and Henikoff, 2016). The same study reported deposition of histones at normally nucleosome-depleted regions, such as promoters and enhancers, shortly after replication, and necessity of the TFs to compete with histones to access their binding site. Interestingly, TFs at the tissue-specific enhancers were less efficient at competing, while broadly active ones had TFs that were quicker to associate with DNA.

1.4 Mutational heterogeneity

1.4.1 Mutation types

Mutations are changes in the identity, order, or number of nucleotides within the DNA sequence. Mutations can be classified into several different categories based upon the type of change, or the effect that the mutation produces. The simplest type of mutation is a *single nucleotide substitution*, where nucleotide identity is changed at a genomic position. While the initial change likely occurred on only one of the complementary strands, after a single round of replication this change can lead to altered nucleotide identities on both of the strands in one of the daughter cells, with potentially no change in the second daughter cell. Single nucleotide changes can result from a variety of DNA lesions. Loss of the nucleotide's nitrogenous base leads to formation of an *abasic site*, which are more likely to affect purines, rather than pyrimidines, with adenine preferentially incorporated opposite the lesion (Randall et al., 1987). Incorporation of the non-complementary base opposite the template during replication, or a spontaneous change in the nucleotide identity on one of the strands, leads to formation of a *mismatch*. Some bases can be modified by endogenous or exogenous factors, such as oxidative stress, UV radiation and tobacco smoke (Chatterjee and Walker, 2017).

Other types of mutations include insertions and deletions, collectively termed *indels*; duplications, inversions, translocations, and others. When occurring within genes, mutations can be classified on the type of effect that they have on the resulting protein product. *Missense* mutations lead to the formation of a codon that encodes for a different amino acid. *Nonsense* mutations result in the occurrence of a STOP codon, that signals for termination of translation and formation of a truncated protein product. *Silent* mutations have no effect on the product, due to some of the different codons encoding for the same amino acid. *Frameshift* mutations, as a consequence of indels, result in shifting of the *reading frame* in a way that subsequent codons are changed. Analysis in this work is largely focused on single nucleotide substitutions, and term *mutation* will be further referred to in this context.

1.4.2 Mutations, selection, and evolution

Mutations provide the raw material for *selection* and *evolution*. Differences in genome sequences between living organisms, species, and individuals of the same species have all originated as mutations at some point. Mutations that result in phenotypic traits that are advantageous to the organism are more likely to persist within the population and rise to a higher frequency, while deleterious mutations that hinder the ability of the carrier organism to reproduce are less likely to be propagated. Single nucleotide changes that have risen to a particular frequency within the population are termed *single nucleotide polymorphisms* (SNPs). Alternative nucleotides that can be found to be present at a single location in the genome are termed *alleles*. Allelles that were present in the common ancestor are *ancestral alleles*, while a nucleotide with an alternate identity is a *derived allele* (Figure 1.6).

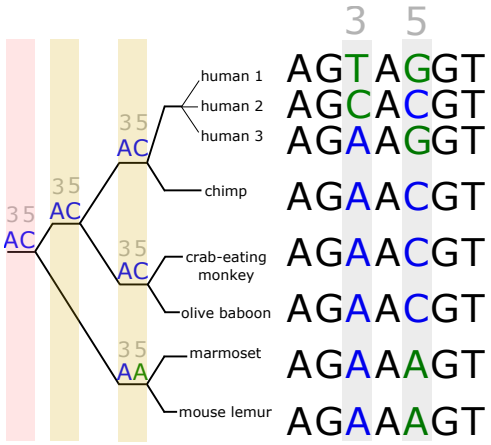


Figure 1.6: Alleles observed at polymorphic sites (3 and 5; grey shading) can be separated into either derived (green) or ancestral (purple) based on the reconstructed ancestral sequences (orange and pink shading).

The process of an allele, and therefore associated traits, becoming more or less frequent within a population in response to environmental perturbation is termed *selection*. Negative, or *purifying* selection, is a process by which a certain detrimental allele is removed from a population, while positive, or *diversifying* selection refers to an allele rising in frequency due to it conferring a *selective advantage*. Mutations that have no effect on the fitness of an organism or its ability to procreate are said to be evolving neutrally and can be subjected to *genetic drift*, a stochastic process of changes in allele frequency due to random sampling inherent in a finite population (Kimura, 1968, 1991).

Genomic elements such as exons are frequently found to be highly conserved

as a result of purifying selection - precise sequences within those are so important for the formation of the final functional product, that deviation would be deleterious, while introns are not subjected to purifying selection to the same extent, therefore exhibit more variability. Hence, conservation is sometimes used to define functionality - if a sequence is conserved it is thought to be important, and therefore functional (Lindblad-Toh et al., 2011). At the same time, sequences that exhibit large degrees of variation are also sometimes likely to be functional. After all, without diversifying selection there would be no evolution and no variation of phenotypes. Therefore, significant deviations from the pattern expected for neutral evolution are indicative of selection and thus organism level function.

1.4.3 Determinants of regional mutation rates

Estimation of the neutral rate of mutations is important, as it provides a reference point relative to which sites could be compared and defined as either conserved or evolving at increased rates. This is complicated by the fact that the mutation rates are not uniform, but vary at different scales across the genome, leading to a notion of *regional mutation rates* (Wolfe et al., 1989; Makova and Hardison, 2015). Regional mutation rates can be affected by the frequency of the lesion occurrences, and by the frequency of their repair. There are multiple features and genomic states that correlate with both of these (Figure 1.7).

One of the most striking examples is dependency of the mutation rate on sequence, and more specifically on the presence of CpG dinucleotides. CpGs have 10-18 times more C$\longrightarrow$T mutations than any other dinucleotides in the genome (Campbell and Eichler, 2013). Cytosines in the CpG context are frequently methylated, and methylated cytosines have a high propensity to undergo deamination due to its unstable nature, resulting in formation of thymine in their stead (Youssoufian et al., 1986). While some genomic regions are often enriched in CpGs (CpG islands), they exhibit lower levels of mutability than the rest of the genome. This is potentially due to reduced methylation in those regions, higher selectional pressures, or lower levels of deamination due to stronger binding between strands (Ségurel et al., 2014; Acuna-Hidalgo et al., 2016).

Clusters of mutations can occur as a result of the activity of "apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like" (APOBEC) enzymes. It has

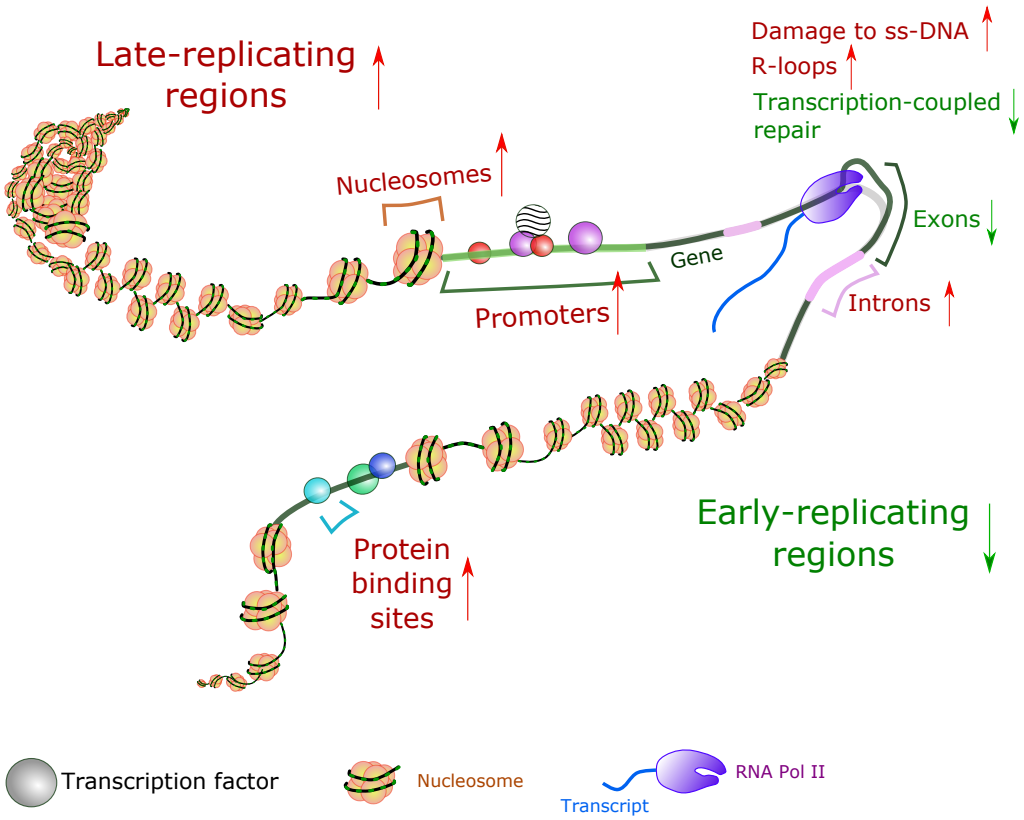


Figure 1.7: Some of the determinants of regional mutation rates. Factors that are thought to promote increased numbers of mutations ↑ are in red, while those that suppress mutations ↓ are in green.

been proposed to result from APOBEC enzymes attacking single-stranded DNA, as such can occur during transcription, upon formation of double-stranded breaks, and at dysfunctional replication forks (Chan and Gordenin, 2016). Mutations are also more likely to arise at the regions where double-stranded breaks occur at high frequency, such as recombination hotspots (regions 1-2kb in size), with nucleotide composition at those sites likely to be shaped through biased transmission of alleles by gene conversions (Tiemann-Boege et al., 2017).

Replication across the genome occurs in a temporarily non-uniform manner. There are certain replication origins that fire at early stages of the S phase, while others fire later. It has been previously shown that late-replicating regions tend to exhibit higher rates of single nucleotide substitutions due to the lower level of mismatch repair (Supek and Lehner, 2015), depletion of free nucleotides (Watt et al., 2015), and accumulation of single-stranded DNA (Stamatoyannopoulos et al., 2009).

The process of transcription can influence the propensity of the genomic region to mutate. While a lot of transcriptional activity occurs at generally conserved regions of the genome, such as protein-coding genes, the process itself leads to transcription-associated mutagenesis (TAM) (Jinks-Robertson and Bhagwat, 2014). Exposure of susceptible single-stranded DNA is associated with increased damage. A study by Lodato et al. (2015) found that numbers of single-nucleotide variants were increased within the transcribed gene regions of neurons. The process of transcription can also be mutagenic due to the occurrence of structures termed R-loops - as formation of hybrids derived from nascent RNA and a template strand of DNA leave the non-template strand in the single-stranded conformation. At the same time, transcription tends to be associated with transcription-coupled repair (TCR), thereby partly counterbalancing the mutagenic effects of TAM (Hanawalt and Spivak, 2008). In addition to the TAM and TCR that goes on in the genic regions, exons and introns have been shown to exhibit differences in mutation rate due to their differential targeting by mismatch repair machinery (Frigola et al., 2017).

Nucleosome occupancy has been shown to be correlated with an increase in genetic variation, with more substitutions observed at nucleosome dyads (Semple and Taylor, 2009; Tolstorukov et al., 2011; Reijns et al., 2015), and this relies upon intact DNA repair machinery (Yazdi et al., 2015). Interestingly, rates of C$\longrightarrow$T mutations have been found to be reduced in the core regions of nucleosomes (Prendergast and Semple, 2011; Sasaki et al., 2009). This is possibly due to the fact that packaged DNA undergoes less local conformational fluctuations within double-stranded DNA, which leads to transient single-stranded DNA exposure, commonly referred to as 'DNA breathing' (Fei and Ha, 2013). Lower levels of this would make cytosines less likely to undergo deamination (Chen et al., 2012b). This has also been proposed to be due to selection acting to maintain optimal GC composition at the nucleosomal sites and linker regions (Prendergast and Semple, 2011).

At larger scale, open chromatin regions (defined by DNase-seq) have been associated with reduced densities of mutations in cancer, attributed to higher accessibility of DNA repair machinery in those regions (Polak et al., 2015, 2014). On smaller scales, Perera et al. (2016) has shown that there is an increased number of mutations at the midpoint of accessible regions, and more specifically at promoters, associated with transcriptional activity and reduced repair by nucleotide excision repair in UV- and

tobacco smoke- induced cancers. Promoters have previously been observed to exhibit increased variation and high rates of evolutionary turnover, in particular housekeeping and testis-specific ones (Young et al., 2015; Taylor et al., 2006).

Mutation rates have been shown to vary even at relatively small scales such as single protein binding sites. Following analysis of sequence divergence between species around particular sequence-specific protein-binding sites, Reijns et al. (2015) demonstrated that there is an unusual increase in numbers of single nucleotide substitutions immediately surrounding highly conserved human TF binding motifs (Figure 1.8). This was proposed to be a reflection of the increased mutational pressure that is acting all across the region physically occupied by the protein. While the motif itself is being preserved through action of purifying selection, stretches of sequence just immediate to it are not, and are thus mutated at a higher rate then the flanking regions that are not occupied by the protein. This has lead to the proposition of the *lagging strand hypothesis*, where enrichment of mutations at protein binding sites is attributed to retained Polα-synthesised tracks post-replication, due to being trapped by the fast binding proteins that hinder track removal (Figure 1.9). However, this increase in numbers of between-species single nucleotide substitutions surrounding highly conserved protein binding motifs has not been directly shown to be a consequence of increased mutation rate, rather than evidence of, for example, diversifying selection favouring variation at the protein binding site edges. Thus, separation of individual contributions of mutation rate and selection would be necessary to demonstrate difference between these possibilities.

Perturbation of DNA replication has similarly been suggested as the reason for an increased numbers of mutations in colon cancers across CTCF binding sites (Katainen et al., 2015). Also, Sabarinathan et al. (2015) demonstrated an increased number of melanoma mutations at TF binding sites, but found that a decrease in nucleotide excision repair was responsible. However, cancer cells can be subject and driven by processes distinct from those taking place in the germline. For example, melanomas are known to arise from high levels of UV damage that is by enlarge repaired by nucleotide excision repair, while mutational burden of germline cells is not driven primarily by UV and role of nucleotide excision repair at protein binding sites in germline might not be as important. Thus, separate investigation of the mutational burden of the protein binding sites in germline is necessary. Also, models that rely on

the interference of proteins with normal cellular maintenance processes do not explain how DNA-interacting proteins that are known to be sequence-specific binders, such as TFs, would be able to bind to their mutated target sequences to do so, a discrepancy that needs to be reconciled.

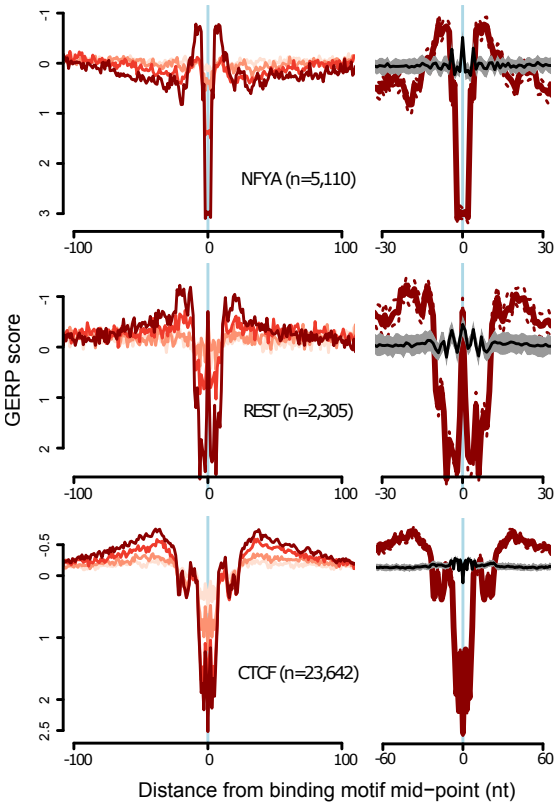


Figure 1.8: Increase in between-species divergence around NFYA, REST and CTCF TF motifs. There is an increase in sequence divergence immediately proximal to the motifs. The so-called 'shoulders' of increased substitutions are proposed to result from an increased mutation rate in the protein-occupied region. Darker red lines corresponds to the top quartiles of ChIP-exo signal strength at the motif region, taken as proxy for binding strength and occupancy. Positive correlation between the level of divergence and ChIP-exo signal is consistent with protein binding being causal of increased mutation rate. Grey lines represent the trinucleotide expectation. Figure adapted from Reijns et al. (2015).

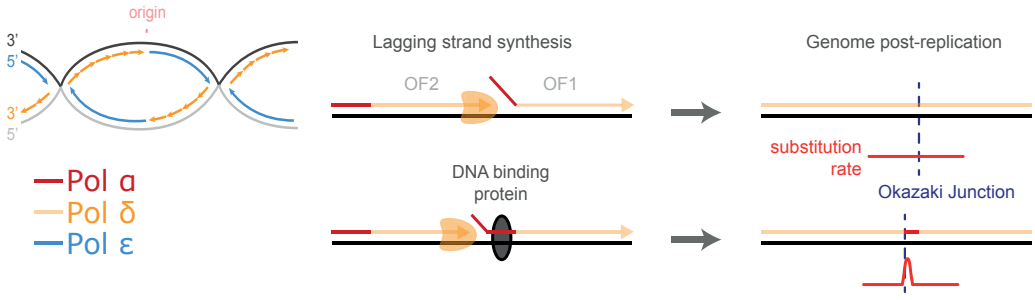


Figure 1.9: Lagging strand hypothesis. Retention of the error-prone Polα-synthesised Okazaki primers by TFs bound shortly post-replication is proposed to increased mutation rate at those protein-binding sites. Figure adapted from Reijns et al. (2015).

1.4.4 Germline mutations

The germline is the only population of cells in which new mutations have a chance of being passed on to the organism's progeny. Therefore any differences in genome sequence between individuals must have arisen at some point as mutations in a cell of the germ lineage. The germline mutation rate has previously been estimated at $\approx 1 - 1.5$ x10^{-8} mutations per site per generation: $\approx 70 - 80$ new mutations that occurred in parental germ cells are inherited by a child (Michaelson et al., 2012; Kong et al., 2012; Francioli et al., 2015). Importantly, germline mutations form the basis for heritable disease, therefore estimating the impact of these mutations is important. Models of mutation retention in germline cells are relatively unknown.

The occurrence of germline mutations can be inferred from polymorphisms present within the population, and between species. On average, single human genome is estimated to have $\approx 40,000 - 200,000$ variants with allele frequency of less than 0.5%, which constitutes 1-4% of all variants, with the rest being more common (Gibbs et al., 2015). The most direct way of identifying germline mutations is to look at the *de novo* variants that occur in offspring relative to genetic variants observed in the parents, which is typically done by comparing genomes of the family trios (Roach et al., 2011). Presence of a variant in the proband at a frequency close to 0.5 (as the genome is diploid), but not in either of the parents means that the mutation must have occurred in a germ lineage cell of one of the parents. This assumption does not always hold true when taking into account *post-zygotic mosaicism* (Campbell et al., 2015). In this instance, mutations that occur at an early stage of the post-zygotic cell divisions are likely to be present at high frequency in the tissue of one type, with about 7% of *de novo* mutations found in blood being mosaic (Acuna-Hidalgo et al., 2015).

There are differences in mutation types and rates between male and female germlines. This is primarily due to the inherent differences in germ cell formation and progression between the two genders. Female germline cells undergo fewer divisions and are formed before birth (Drost and Lee, 1995). This makes them less susceptible to mutations that are replication-associated in nature, but more likely to undergo formation of lesions due to cells having to maintain themselves while in meiotic arrest (Herbert et al., 2015). There is an association between the maternal age and chromoso-

mal abnormalities in the offspring that is thought to result from an inability to properly segregate the sister chromatids, leading to aneuploidies (Sherman et al., 1994).

The paternal germline, on the other hand, undergoes multiple rounds of cell division, and therefore multiple round of genome replication over the lifetime of a male (Drost and Lee, 1995). This has lead to the notion of a male mutational bias - a preposition that the majority of the mutations in offspring are paternal in origin. What more, the numbers of *de novo* mutations in offspring have been found to increase with the age of the father at conception (Kong et al., 2012; Francioli et al., 2015). Age-related increase in both male and female germline mutations is also likely to stem from lower levels of repair - general failure to repair non-replicative damage between cell divisions and the lower efficiency of repair mechanisms (Guo et al., 2015). Additionally, mutations in offspring of younger fathers tend to localize in late-replicating regions, while more *de novo* mutations in the genic regions has been observed in older fathers (Francioli et al., 2015). Some mutations can confer a selective advantage to the germ cell, such as clonal expansion of sperm progenitors, and at the same time be causing disease in the offspring. These 'selfish' mutations in genes such as *PTPN11*, *HRAS*, *FGRF2*, *FGFR3*, and others, have been implicated in various, often developmental, disorders (Acuna-Hidalgo et al., 2016).

1.4.5 Somatic and cancer mutations

Mutations that occur in the cells of the soma are not going to be passed on to the progeny and will cease to exist with the death of an organism. Mutations occur in each somatic cell during the lifetime of a person. While most of those have virtually no effect, others may confer a selective advantage leading to expansion of that cell population, resulting in *somatic mosaicism*. Somatic mosaicism can occur during early development, and is a known cause of developmental disorders, but also has been recognised to a phenotype of ageing (De, 2011). Somatic mutations that confer a selective advantage to the cell through increased proliferative ability and the loss of programs characteristic of normal cellular function can lead to cancer (Hanahan and Weinberg, 2011).

Mutations in cancer can be defined as either *drivers* or *passengers*. Driver mutations are though to be a causative for cancer initiation, or contributing to its progression. They could be mutations that result in hyper-activation of oncogenes,

leading to an increase in cell growth, proliferation and metabolism. In conjunction with inactivating mutations in tumour-suppressor genes, such as those that are responsible for initiation of apoptosis (programmed cell death) that renders its products inactive, cancer cells typically undergo uncontrollable proliferation. Identification of driver genes within tumours is not always straightforward due to the presence of large numbers of passenger mutations. Those mutations do not confer any selective advantage to the cancer cells, but are rather 'hitch-hikers', and are able to rise to the same frequency in the tumour as driver mutations because they happened to co-occur in the same cell.

Most known driver mutations are located in the protein-coding part of the genome, in genes such as *TP53*, *PIK3CA*, *KRAS*, *BRAF* and others, which is not surprising considering that most of studies have been conducted using whole-exome sequencing (Campbell, 2016). One of the well-known examples of the non-coding cancer driver is a mutation in the promoter of the *TERT* gene, which creates a binding site for a TF ETS and leads to gene's over-expression (Vinagre et al., 2013).

Most cancers carry 1,000-20,000 single nucleotide substitutions, however numbers can vary greatly depending on the type (Campbell, 2016). While some cancers occur through the activity of external mutagens, the variation in cell divisions of affected tissues has been shown to explain variation in cancer risk (Tomasetti and Vogelstein, 2015).

Somatic mutations can be the result of germline predisposition. An example of this would be a germline mutation in genes associated with maintaining genome integrity, such as those coding for the components of the mismatch repair machinery (*MLH1*, *MLH2*, *MLH6*, *PMS2*, and *PMS1*) (Peltomäki, 2001). Heterozygous germline mutations in those genes can lead to accumulation of the DNA replication errors and cause Lynch syndrome, which is characterised by occurrence of early onset colorectal, endometrial, ovarian and other cancers (Cohen and Leininger, 2014).

As different cancers are driven and dominated by distinct mutational processes, they exhibit different *mutational signatures* - consistent patterns of changes occurring on the local sequence background, and with variable distribution across the genome. More than 20 different mutational signatures have been initially classified by Alexandrov et al. (2013) and this number is growing constantly. Some are attributed to specific mutational processes, such as APOBEC mutagenesis or exposure to UV, while others reflect the 'clock-like' mutational process that operate in normal human

cells (Alexandrov et al., 2015), and others are of unknown aetiology.

1.5 Repair processes

As mentioned previously, heterogeneity of mutation rates across the genome results from both the differences in the rate of lesion formation, and in the efficiency and accuracy of lesion repair. Unlike other cellular components, DNA exists as necessary entity whose encoded information cannot be regenerated if lost, and therefore must be maintained and repaired upon formation of damage. There are multiple mechanisms in the cell that act to try and preserve the integrity of the genome including mismatch repair (MMR), nucleotide excision repair (NER), base excision repair (BER), as well as single and double stranded break repair (SSBR and DSBR, respectively) (reviewed in Ciccia and Elledge (2010)).

1.5.1 Replication-coupled repair

Despite its remarkable fidelity, the process of replication is inherently mutagenic. The necessity to faithfully replicate 3 billion base pairs in a relatively short period of time calls for strict quality-control. DNA polymerases are responsible for replicating the genome, and the various families of this enzyme vary in their fidelity. On average, for every $\approx 10^4 - 10^5$ bases that they incorporate during replication, one mismatch is produced (Kunkel, 2009). Polδ and Polε, which are the main replication polymerases in eukaryotes, make fewer mistakes than Polα, as they posses 3'$\longrightarrow$5' *exonuclease* activity, an ability to cleave and remove a wrongly incorporated nucleotide, triggered by the abnormal geometry of base in the active site and slowing of polymerisation (reviewed in Ganai and Johansson (2016)).

When normal replicative polymerases are unable to deal with a lesion, a special type of *translesion polymerase* might be recruited to the site of the damage (reviewed in Vaisman and Woodgate (2017)). Those polymerases are quite often good at bypassing damage to allow replication to proceed, but do not possess high levels of fidelity. Interestingly, the error-prone nature of some polymerases can be utilized by cells to drive adaptive capability as in bacteria (Janion, 2008), and in processes such as somatic hypermutation and diversification of the immunoglobulin variable regions in humans for ability to recognise and rapidly respond to infection (Faili and Gue, 2009).

Most mistakes that go unnoticed by replicative polymerases can be repaired

by MMR (reviewed in Kunkel and Erie (2015)). MMR is broadly involved in repair of short indels and single base-base mismatches. In MMR-deficient yeast, high rates of CG⟶TA mismatches are observed to be generated at a highest rate, probably through higher occurrence of G:T mismatches that are most efficiently corrected by MMR (Lujan et al., 2014). In addition to that, lower level of MMR activity are apparent on the lagging strand *versus* leading strand, which corresponds to the increased number of mismatches generated on the former (Lujan et al., 2014; Andrianova et al., 2017). MMR is thought to be temporarily coupled with DNA replication, suggesting that some types of replication errors, such as the ones generated during particular DNA repair and recombination processes, would not be readily available to MMR to fix (Kunkel and Erie, 2015). This temporal coupling of MMR to the replication fork, accessibility of DNA to the repair machinery, or opportunity for repair might be responsible for lower levels of MMR observed in late-replicating regions (Supek and Lehner, 2015). Interestingly, exons have been shown exhibit higher levels of MMR, surveillance and activity than introns, due to the recruitment of MMR components to H3K36me3 histone mark enrichment within exons (Frigola et al., 2017).

1.5.2 Nucleotide-excision repair

Nucleotide excision repair (NER) works by recognising and removing DNA-helix-distorting lesions. NER has the ability to remove a large number of different types of lesions, such as UV-induced lesions, bulky chemical adducts, inter-strand crosslinks, and damage induced by reactive oxygen species. NER repairs those by excising a ≈30bp fragment on a single strand around the lesions followed by re-synthesis of the stretch. Germline inactivating mutation in the components of NER lead to an extreme phenotypes that are highly sensitive to UV damage and elevated risk of various tumours, resulting in a condition termed *Xeroderma Pigmentosum* (Cleaver, 1968; Lehmann et al., 2011). NER can be classified into the transcription-coupled NER (TC-NER) and global genomic NER (GG-NER). GG-NER can recognize a wide variety of different lesions, owing to its ability to detect the distortion of the DNA helix, while TC-NER is recruited to the sites of stalled RNA polymerase II (reviewed in Marteijn et al. (2014)). Histone modifications, in particular acetylation and ubiquitylation, that are suggested to promote histone displacement, are thought be play a role in allowing access of NER to lesions (Wang et al., 2006; Lans et al., 2012). Recently, GG-NER complexes have been

found to be located at the nucleosome-free regions at gene promoters that can be found at the boundaries of the higher order nucleosome-nucleosome interacting domains in undamaged cells, and is initiated from there upon damage (Eijk et al., 2018).

1.6 Aims and research outline

1.6.1 Specific motivations and importance

Large proportion of studies, particularly those concerned with human disease aetiology, are still mainly confined to a frame of protein-coding part of the genome. In recent years, in particular due to advances in whole-genome sequencing capability and availability, the importance of the role played by the non-coding fraction is becoming more evident. The ability to sift out mutations and variants that are causal for the specific phenotype is in a large part hampered by the lack of means to pinpoint the genomic regions where functionally consequential mutations are more likely to occur, and this is confined by our estimations of selection and mutation pressures. There are differences in regional mutation rates, and this variability is attributed to a growing number of genomic features, such as protein-binding sites. However, the mutational cost of TF-DNA interactions and its interplay with genome maintenance is not fully understood.

The focus of the work described herein concerns the investigation of the mutational burden at protein-binding sites. Increased numbers of mutations at protein-binding sites has been supported by several recent studies, and is attributed to the TF-DNA interactions interfering with DNA replication and repair. More specifically, Reijns et al. (2015) has previously proposed that occlusion of the lesion-containing stretch of error-prone polymerase-synthesised DNA by bound TFs interferes with normal cellular maintenance processes, leading to retention of mutations in the mature genome. If true, his would then suggest that (1.) mutation rate are elevated *within* as well as adjacent to the binding sites, and that (2.) only sites that are bound by proteins within the cells where mutations are occurring are expected to exhibit elevated mutation rates.

The main focus of the work here are mutations that occur in the cells of the germline, and therefore have a chance of being passed on to the next generations and cause hereditary disease, and estimation of the impact that those mutations have. This is of particular importance in light of the evidence for replication-related nature of those mutations and potential association with the increasing numbers of *de novo* mutations

in offspring with advancing paternal age. Identification of the protein-binding landscape of the most highly-dividing germline cells can provide a map of the mutational hotspots where disease-causal variants are likely to be found. As no inference is made here about the function of proteins that interact with analysed binding sites, the range of binders investigated in this work includes, but is not limited to, transcription factors. Thus, identified binding sites might be occupied by sub-nucleosomal-sized proteins that do not regulate transcription, but could nevertheless be important for proper cell function, such as those playing a role in organisation of chromatin structure (for example, CTCF).

A mechanistic basis for the paradoxical occlusion of the mutagenic lesions by the sequence-specific binding proteins is unexplained and intriguing. Here I aim to address this and propose a model that could resolve this. The resulting observations of protein-DNA interaction in connection with mutagenesis will provide an insight into the mechanistic basis of the occurrence of mutations at regulatory sites.

1.6.2 Main research questions

The research project described in this thesis aims to answer the following questions, also illustrated in Figure 1.10:

1) Are protein binding sites subjected to increased mutational burden adjacent to and within the sites? If so, are the mutations occurring at binding sites in the germline deleterious and do they have the potential to lead to heritable disease?

2) Can sequence-specific DNA-binding proteins be responsible for inducing increased numbers of single-nucleotide changes within their binding motifs and if so, how can this be reconciled with the sequence specificity of their binding?

1.6.3 Thesis structure

Throughout this thesis I aim to address the first question through identification of the protein-binding sites in the human and mouse germlines and, for comparison, in somatic cells by generation and analysis of ATAC-seq data (Chapter 2), followed by estimation of the selection and mutation pressures acting up on those sites using between-species sequence divergence and alleles in human and mouse populations (Chapter 3). The second question is addressed through analysis of cancer mutation rates at sequence-specific protein binding sites of several TFs (Chapter 4), and experimental testing of

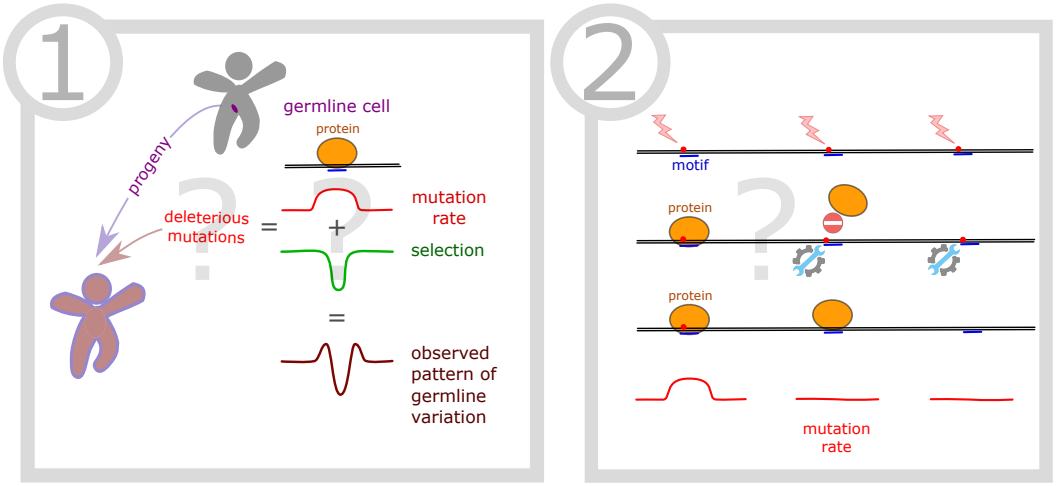


Figure 1.10: In the current work I aim to address the following questions: (1) Is the increased variation next to TF motifs shaped by the combined actions of increased mutation rate and purifying selection? If so, is the mutation rate elevated within, as well as adjacent to the protein-binding sites, making them mutational hotspots and leading to hereditary disease if they occur in the germline? (2) Is protein binding causal for retention of mutations at their binding sites and how can this be reconciled with sequence-specificity of TF binding?

the tolerance of the KLF4 protein to its binding motif-containing sequences harbouring lesions (Chapter 5).

CHAPTER 2

Identification and separation of germline and somatic protein binding sites

2.1 Introduction

2.1.1 Male germline is more mutagenic than female due to large number of cell divisions

For most cells of the living entity, the set of mutations that it acquires during its lifetime will disappear with the death of the organism and have no effect on any progeny. The exception to this are germline cells (sequence of cells which develop into eggs and sperm), which provide the genetic basis for the next generation. Mutations acquired during the life of germ lineage cell that will go on and form a zygote will be present in every cell of the subsequent generation. Therefore all variation is shaped by mutations that have occurred at some point along germ cell lineage, and termed *germline variation*.

It is known that germline variation, in the form of *de novo* mutations in offspring, is correlated with parental age (Ségurel et al., 2014). While some of the individual contributions of both male and female germline are known, the overall burden associated with each is not clear. Generally, the male germline is widely regarded as a major contributor to mutation load, estimated to add 1.5-2 mutations per year of the father's life prior to conception, compared to 0.25-0.37 mutations per year of the mother's age (Kong et al., 2012; Francioli et al., 2015; Michaelson et al., 2012; Jónsson et al., 2017; Goldmann et al., 2018), with most mutations found in the paternally inherited haplotype (Gao et al., 2018). Differences in estimated mutation rates between males and females are in accordance with differing numbers of zygote-to-zygote divisions that male and female germlines undergo (Figure 2.1). The view of replication process as primary contributor to human mutations stems from this observations of increased variation with rising numbers of germline cell divisions (Ségurel et al., 2014).

The male germline undergoes continuous replication cycles aimed at production of large number of sperm cells and self-maintenance of stem cell populations (Drost and Lee, 1995). In males, spermatogenesis initiates at the onset of puberty (≈13-14 years after birth), at which point each cell has been estimated to have already undergone 33-34 divisions (Drost and Lee, 1995; Ségurel et al., 2014). Thereafter, a set of stem-like spermatogonial cells (SSCs) divide every 16 days, both in order to maintain

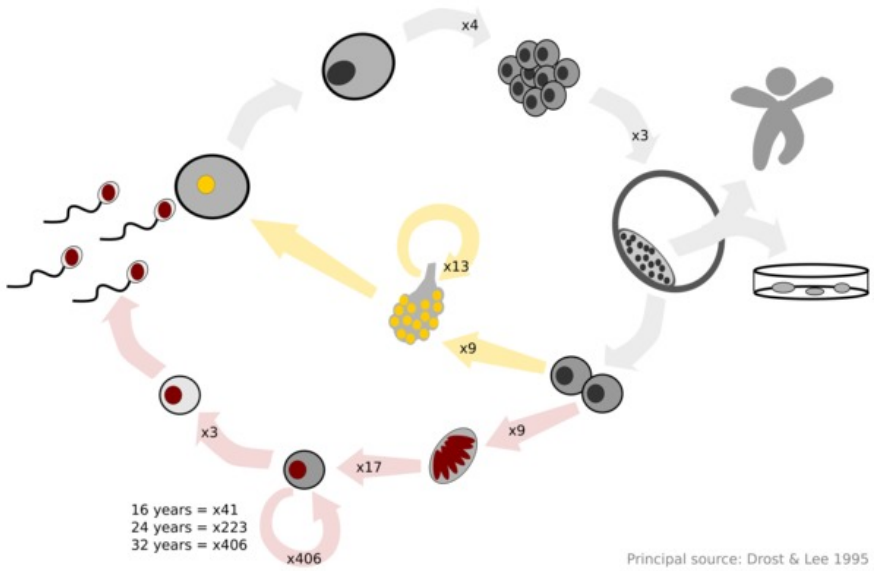


Figure 2.1: Overview of human female and male germlines. Numbers on arrows indicate the estimated numbers of divisions each cell type undergoes before transformation into the next stage cell. Female germline divisions are shown in yellow and male germ cell ones in pink. On the bottom are estimations of the numbers of divisions each spermatogonial cell would have had to have undergone at indicated age of a male. (Figure adapted from Prof Martin Taylor; principal source Drost and Lee (1995))

a their pool, and to produce cells committed to differentiation, which then undergo further 3 divisions before formation of sperm. Therefore, in humans, about 80% of all germline replications occur in the spermatogonial cells, while in mice this number is around 30% due to a shorter generation time (see Box 2.1 for calculations).

Box 2.1: Calculation of spermatogonial cell divisions

General formula for the calculation of the number of spermatogonia stem cell (SSC) divisions (Div) in male:

$$\text{Div}_{SSC} = \frac{(Age_{years}^{at\ conception} - Age_{years}^{at\ puberty} - (\frac{\text{Spermatogenesis}_{days}}{365})) \times 365}{1\ \text{SSC division}_{days}} \quad (f\text{:}2.1.1)$$

Calculation of SSC divisions as a proportion of total cell divisions in both parents:

$$\text{Div}_{total} = \text{Div}_{pre-birth}^{male} + \text{Div}_{pre-birth}^{female} + \text{Div}_{SSC}^{male} + \text{Div}_{post\ SSC}^{male} \quad (f\text{:}2.1.2)$$

Then the percentage of SSC divisions is:

$$SSC_{\%} = \frac{\text{Div}_{SSC}}{\text{Div}_{total}} \times 100 \quad (f\text{:}2.1.3)$$

Assuming the average age of father at conception is 30 (Jónsson et al., 2017), puberty occurs at 14 years, spermatogenesis takes 74 days, and SSC divide every 16 days thereafter (Drost and Lee, 1995), then:

$$\text{Div}_{SSC} = \frac{(30 - 14 - (74 \div 365)) \times 365}{16} = 360.375 \quad (f\text{:}2.1.4)$$

$$\text{Div}_{total} = 33 + 29 + 360.375 + 3 = 425.375 \quad (f\text{:}2.1.5)$$

$$SSC_{\%} = \frac{360.375}{425.375} \times 100 = 84.7\% \quad (f\text{:}2.1.6)$$

While in mice:

$$\text{Div}_{SSC} = \frac{274_{days} - 6_{days} - 43_{days}}{8.6} = 26 \quad (f\text{:}2.1.7)$$

$$\text{Div}_{total} = 21 + 18 + 26 + 9 = 74 \quad (f\text{:}2.1.8)$$

$$SSC_{\%} = \frac{26}{74} \times 100 = 35.1\% \quad (f\text{:}2.1.9)$$

Conversely, in females the eggs undergo all rounds of DNA replication prior to birth following relatively small number of divisions (approximately 31 in human) (Drost and Lee, 1995). The integrity of the bivalent chromosomes (homologous chromosomes that are physically held together) within an oocyte has to be preserved for many years in humans, and while there is no replication taking place, mutations can arise as a result of such features as structural chromosome fragility due to cohesin depletion (Herbert et al., 2015). Therefore, maternal age effects tend to be associated with risks of aneuploidies and found to be correlated with the numbers of clustered mutations at particular locations linked to processes involving the formation of double-stranded breaks (Goldmann et al., 2018; Jónsson et al., 2017). Recently it has also been argued that the maternal age contributes more substantially to the parental age effect than thought previously, as until the 4-cell stage the zygote relies on the repair and replication machinery of maternal oocytes, which may undergo degradation and lose fidelity in older mothers (Gao et al., 2018). This could also imply that lesions acquired in the paternal germline due to the larger number of divisions would be less likely repaired in older females.

2.1.2 Most germline mutations at protein-binding sites are expected to occur at spermatogonia-active sites

We hypothesise that the increase in germline variation proximal to TF binding motifs observed by Reijns et al. (2015) (Figure 1.8) represents increased mutation rate at the genomic regions that are physically occupied by bound proteins. If the increase in the mutation rate is indeed protein binding dependent, we would expect to see this elevation in germline variation at protein binding sites that are bound by proteins ("active" binding sites) in germline cells, where those mutations are occurring, in contrast to those that are not occupied by proteins in the germline, such as somatic-specific protein binding sites (Figure 2.2). This is testable, assuming we know where proteins in germline and soma bind. Increased variation at germline-active, but not at the somatic-specific binding sites would support our hypothesis. Increase at both categories of sites would indicate that the elevated variation is not dependent on the protein binding, but instead is just a general feature of the binding sites. No elevation at either of the categories would mean that the protein binding sites do not exhibit increased mutation

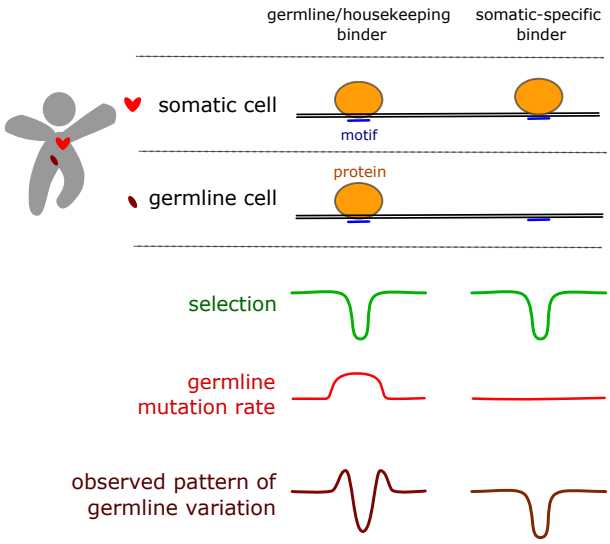


Figure 2.2: Expected differences in germline mutation rate, selection and between-species divergence in germline, housekeeping and somatic binding sites, if binding of proteins does induce increased mutation rate.

rate as such.

This mutational phenomenon has been proposed to be related to replication (Reijns et al., 2015). Even though the replication machinery has evolved to have an impressive level of fidelity (Kunkel and Erie, 2015), some mutations escape surveillance and repair. As mentioned in Subsection 1.5.1, most mismatches that would lead to the formation of mutations are normally repaired by mismatch repair machinery and exonuclease activity of the replicative polymerases. Obstruction of those processes would be expected to raise mutational load substantially. Proteins could conceivably present such an obstruction if they bound DNA shortly after synthesis of the daughter strand. This has lead to proposition of the *lagging strand hypothesis*, where binding proteins hinder the process of error-prone Polα-synthesised Okazaki primer replacement (Figure 1.9) (Reijns et al., 2015). One would then expect the majority of mutations to occur at binding sites active in germline cells that are most highly replicating, which are spermatogonial cells.

2.1.3 Spermatogonial cells sub-populations and challenges in their isolation

In humans, spermatogonia are classified into several categories of A-type stem-like (A_{pale} - active, euchromatic, A_{dark} - reserve, heterochromatic) and differentiating B-type cells (Di Persio et al., 2017). In mice there are several additional populations of A_{paired} and $A_{aligned}$ spermatogonia, which form clusters of interconnected cells (Wa-

heeb and Hofmann, 2011). While in humans during the pre-pubertal period A-type spermatogonia comprise 80-90% of germ cells, after onset of puberty they comprise a small percentage of the total germ cell population (Wu et al., 2009). In part due to this, there is currently limited amount of data available that would allow comprehensive identification of the protein-binding sites in spermatogonial cells. Hence, here we have set out to describe the protein-binding landscape for both mouse and human that represents the primary tissue in which the majority of germline DNA replication occurs.

There is wide interest in spermatogonial cell isolation due to the high value of those cell populations in fertility preservation of males undergoing chemotherapy and fertility treatment of patients with azoospermia (lack of sperm in the semen) (Forbes et al., 2018). Isolation of pure spermatogonial cell populations is challenging due to a number of reasons. First, as already mentioned, spermatogonial cells, and particularly stem-like sub-populations (SSCs), which are of the main interest here, represent a relatively small percentage of cells in testis post-puberty . The sub-populations of stem-like and differentiating cells also become difficult to morphologically differentiate after the onset of puberty (Wu et al., 2009). Due to the challenges in obtaining large quantities of tissue and cell isolation, human spermatogonial cell specific markers have not been robustly established. More is known about the spermatogonial cell populations in rodents, as tissue is more readily available (Kopylow and Spiess, 2017). While the general germ cell progression is somewhat similar between humans and rodents, they exhibit differences in marker expression (He et al., 2010). Additionally, some of the recent attempts at comprehensive compilation of published markers and cross-study intersection analysis found that most marker expression is not uniform between the morphologically identical spermatogonial cell types and even the few markers that are common to all types do not always show uniform expression within the subtypes (Kopylow and Spiess, 2017).

2.1.4 Questions addressed in the current Chapter

The main aim of the work described in this Chapter is to (1.) isolate populations of the highly replicative germ cells from specimens of mouse and human adult testicular tissues, and to (2.) define sets of binding sites that are active specifically in germline or somatic cells, or in both. With main focus in obtaining spermatogonial cell populations

free of somatic cell contamination, we ideally would want to obtain the sub-population of most highly dividing stem-like spermatogonial cells (SSCs). However, partially differentiated cells are also replicative cells of the germline lineage which are likely to exhibit commonality of binding landscape, so are likewise relevant for this analysis. As SSCs constitute a rare population within adult testes with no clear-cut marker expression, we extended our scope to a wider range of spermatogonial subcategories.

2.2 Methods

2.2.1 Spermatogonial cell marker selection

For isolation of the live cell populations we used fluorescent-activated cell sorting (FACS), as this method should not majorly perturb chromatin composition and gene expression (Richardson et al., 2015) and allows for a fine-scale resolution gating of the sorted populations based on the marker signal, but also size and shape of the cells (Valli et al., 2014). FACS requires fluorescently-labelled antibodies against at least one of cell type specific marker, which has to meet several criteria. First, the epitope (part of the marker molecule to which the antibody attaches itself) recognised by the antibody has to be a surface marker (located outside the cell membrane), as that is the only way to label cells without the need for permeabilization. Secondly, spermatogonial cells reside within the seminiferous tubules and are surrounded by somatic cells, such as Sertoli and Leidig cells, that play roles in supporting spermatogenesis. Therefore, any marker used must not only be expressed in the germ cells, but also not be expressed in any of the somatic cell populations present within the testes. Thirdly, the antibody against the epitope of a marker has to be commercially available, and ideally have been validated for cell isolation before. Wu et al. (2009) previously compared expression of multiple candidate markers within the cells of human pre-pubertal spermatogonial cells with their expression levels in somatic cells of the testes, and reported some of the surface markers to show enrichment in the former, mostly in accordance with a recent comprehensive review of the published spermatogonial cell markers by Kopylow and Spiess (2017). Among those there were markers that have been reported to be more biased towards the expression in differentiating cells (KIT) (Di Persio et al., 2017), or show contradictory results across multiple studies (some reviewed in Kopylow and Spiess (2017)), with likely expression in somatic cells within the testes (THY1, ITGA6, GPR125, GFRalpha1R) (He et al., 2010; Altman et al., 2014). Of the remaining markers, we have chosen to use FGFR3 (fibroblast growth factor receptor 3), which has been reported as a protein not expressed in gonocytes or somatic cells within testes (Ewen et al., 2013). Kossack et al. (2013) also confirmed FGFR3 to be the only specific cell surface biomarker in agreement with the earlier Von Kopylow et al. (2010) study.

Wu et al. (2009) has reported 85-fold increase in FGFR3 expression in human spermatogonia *versus* somatic cells within testes. Figure 2.3 shows an example of FGFR3 staining from the Human Protein Atlas (https://www.proteinatlas.org/). The commercially available antibody used here (FAB766P, clone 136334, R&D systems) recognises the IIIc isoform that is expressed in testis (Ewen et al., 2013). *FGFR3* encodes a cell surface protein that spans the cell membrane and is speculated to be involved in spermatogonial survival signalling (Ewen et al., 2013). At the time of our initial use of FGFR3 as a marker for the spermatogonial cells there had not been any other studies that successfully implemented it for FACS. Since then FGFR3 has been used in another study reporting successful isolation of spermatogonial cells via magnetic cell isolation (Von Kopylow et al., 2016), where isolated cells were shown to express FGFR3 mRNA. In addition, they described absence of somatic cell transcripts, such as WT1, ACTSA2 and INSL3, and co-expression of the pluripotency-associated protein UTF1, an established human spermatogonial intracellular marker (Valli et al., 2014).

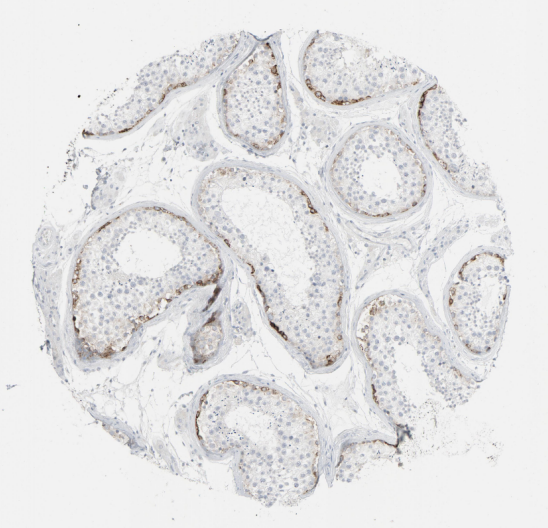


Figure 2.3: FGFR3 staining (brown) in seminiferous ducts of adult testis sections from Human Protein Atlas (Thul et al., 2017) (image available at https://v18.proteinatlas.org/ENSG00000068078-FGFR3/tissue/testis#imid_1393709)

2.2.2 Germline tissues

Specimens of human testicular tissues were collected from patients undergoing orchiectomies at the Western General Hospital, Edinburgh. Tissue was obtained after informed consent through the Lothian NRS BioResource, and study was approved by NHS Lothian (Lothian R&D Project Number 2015/0370TB). After extraction, testis was cut open by the surgeon and one or more fragments (typically $\approx 0.5cm^3$) of the inner tissue were excised and put into ice-cold phosphate-buffered saline (PBS) and kept on

ice. It would be most desirable to obtain phenotypically and genetically 'normal' tissue from healthy individuals; however obtaining such samples would present difficulties in finding suitable donors and obtaining ethical approval. The majority of patients from which the tissue was obtained were undergoing orchiectomies due to presence of a suspected testicular tumour, and in rarer cases due to undiagnosed testicular pain, or gender reassignment surgery. In cases of testicular tumours, tissues used here were taken away from the site of the tumour and appeared to be exhibit grossly normal structure. Tissue was then rapidly transported downstream analysis, with total processing completed within ≈ 5-7 hours of explant.

Human analyses were complemented by those in mice. In collaboration with the research group of Dr Ian Adams (IGMM, HGU MRC, Edinburgh), we were able to obtain pure mouse spermatogonial cell populations from mice hemizygous for a germ cell-specific Cre driver Ddx4-Cre (Tg(Ddx4-cre)1Dcas) (Gallardo et al., 2007), and heterozygous for a yellow fluorescent protein (YFP) reporter conditionally expressed from the ROSA26 locus (Gt(ROSA)26$Sor^{tm1(EYFP)Cos}$) (Srinivas et al., 2001). At 6 days *post partum* these mice express YFP specifically in spermatogonia.

2.2.3 Cell isolation and FACS

Tissue desegregation was performed by Dr Marie MacLennan or Dr Fiona Semple (trained in biohazard safety containment level 2 procedures for primary human tissues). First, tissue was subjected to mechanical desegregation using a scalpel. Cells were then pelleted by spinning in the centrifuge for 3 minutes at 3000 RPM. Enzymatic degradation was carried out by re-suspending tissue in 5 ml of 0.25% trypsin (in PBS), followed by 10 minute incubation at 37 °C. 5 ml of 20% fetal bovine serum (10% final concentration) was then added to stop digestion. After mixing by inversion, tissue was spun for 3 minutes at 3000 RPM. PBS was then discarded and tissue re-suspended in 5 ml of PBS with 10% FBS. Cells were then filtered through 35 μm strainer (Becton Dickinson) using a pastette. Cells were the spun for 3 minutes at 3000 RPM and supernatant discarded. Cells were then re-suspended in total of 1 ml of PBS+10%FBS. 20 μl of the cell suspension was used as a negative control for FACS analysis, rest was used for the antibody labelling. 10 μl of the PE-conjugated FGFR3 antibody was added to ≈1 ml of the cell suspension and incubated at room temperature for 20 minutes with rotation. Antibody was then washed away by spinning cells at 3000 RPM, discarding

supernatant and re-suspending in 1 ml of PBS+10%FBS. This was repeated twice. Finally, cells were re-suspended in 200ul PBS+10%FBS for FACS analysis.

FACS was performed by Dr Elizabeth Fryer on *Aria II* cell sorter (BD biosciences). Mouse cells were sorted based on YFP fluorescence. Most PE-positive (human) cells gated as 'live' were then sorted into three separate populations and labelled as 'small', 'large', and 'large with high side-scatter' based on their relative position along Forward Scatter (FSC) vs Side Scatter (SSC) axis.

2.2.4 ATAC-seq

To define the chromatin landscape of spermatogonial cells we utilized Assay for Transposase-Accessible Chromatin using Sequencing (ATAC-seq) (Buenrostro et al., 2013). ATAC-seq has benefits over other methods, such as the DNase-seq, because it requires fewer cells as an input. The cell numbers isolated from the limited amounts of the human testicular tissue that we were obtaining were expected to be relatively small (<50,000), as spermatogonial cell population represents a minor proportion of cells in adult testes. Another advantage of ATAC-seq is that it requires fewer steps and is less time-consuming than competing accessibility and footprinting methods. Previously published protocol for ATAC-seq was followed (Buenrostro et al., 2013). ATAC-seq was carried out by Dr Yatendra Kumar. In short, during the ATAC-seq procedure a hyperactive Tn5 transposase, preloaded with DNA adaptors, is added to the cell lysate. This modified Tn5 enzyme works by interacting with DNA at the sites of accessible chromatin, creating a double-stranded break separated by 9 base pairs before ligating Illumina sequencing adaptors (with one free end). The proximal activity of two transposases results in the formation of genomic DNA fragments flanked by Illumina sequencing adaptors. Highly compacted DNA or regions protected by bound nucleosomes/other proteins are protected from digestion. Since Tn5 binds to DNA as a dimer, it has been reported to require at least 38bp of accessible DNA due to steric hindrance (Adey et al., 2010). After PCR amplification, fragments were subjected to the paired-end sequencing (with a typical read length of 75bp and ≈100-230 million fragments per library). The details of the sequencing platforms (sequencing carried out by Edinburgh Genomics) used can be found in Tables 2.1 and 2.2. Samples sent in the same batch were multiplexed and each sample was run on two lanes to allow for correction of lane-specific bias.

2.2.5 Computational analysis of ATAC-seq data

The `.fastq.gz` files containing raw read sequences for our primary data were downloaded from the Edinburgh Genomics delivery server and `md5sums` were checked. The human somatic tissue ATAC-seq reads were obtained from the ENCODE project repository (`https://www.encodeproject.org`). Data from the Guo et al. (2017) were obtained from the NCBI GEO repository (accession GSE92280). Pre-aligned reads in `.bam` format were obtained for mouse lung, bone marrow and large intestine ATAC-seq from Cusanovich et al. (2018). Raw `.fastq` reads from other mouse somatic tissue ATAC-seq were obtained for cerebellum (Feng et al., 2017) (NCBI GEO accession GSE76984), B cells (Minnich et al., 2016) (NCBI GEO accession GSE71698), and mammary gland (Dravis et al., 2018) (NCBI GEO accession GSE116386). See Table 2.3 for descriptions of all the human spermatogonial datasets, Table 2.4 for other publicly available human ATAC-seq datasets, and Table 2.5 for mouse datasets that were processed.

Reads were then trimmed to remove any retained adaptor sequences at both 3' and 5' ends using the command `cutadapt -n3 --format=fastq --overlap=3 -g GAGATGTGTATAAGAGACAG -g CAGATGTGTATAAGAGACAG -a CTGTCTCTTATACACATCTG -a CTGTCTCTTATACACATCTC` (Martin, 2011). Trimmed reads were aligned to the reference genome (*GRCh38/hg38* for human; *NCBI37/mm9* for mouse) with `Bowtie2` (Langmead and Salzberg, 2012) in paired-end mode and default settings, with the exception of limiting the insert size to 4Kb. The resulting `.sam` files were then compressed to `.bam` format using the `samtools view` (Li et al., 2009) command with –`q 30` flag, which only kept the reads with map-quality score >30. Then `samtools flagstat` (Li et al., 2009) was used to get information about numbers of reads. Reads were sorted by name using `samtools sort` (Li et al., 2009). The resulting file was fed to the `bamToBed` (Quinlan and Hall, 2010) command in –`bedpe` mode to convert the `.bam` file to the paired-end `.bed` file. The starting coordinate of the 5' read and end coordinate of the 3' read were then extracted to get a `.bed` file with the coordinates of the insert fragment (subsequently termed *fragments*). Upon encountering fragments with exactly matching start and end coordinates, only one was retained, as those were likely to be PCR duplicates produced during amplification, rather than truly distinct fragments coming from separate chromosomes. Any fragments with reads that aligned to the mitochondrial genome were filtered out and excluded from further analysis. Fragments

overlapping with the regions previously blacklisted as mitochondrial homologs were discarded as well. Reads from the same sample, but sequenced on separate lanes were initially processed individually to ensure that they correlate well with each other, and then were merged together for further analysis.

2.2.6 Peakcalling

All filtered fragment files were converted to `.bampe` format for peakcalling using `bedpeToBam` (Quinlan and Hall, 2010). Peaks were called using MACS2 (Zhang et al., 2008) with the following arguments: `callpeak -f BAMPE -g hs --keep-dup all -B --SPMR --nomodel`. Separate sets of peaks were called either from all the fragments for each individual biological replicate (data coming from technical replicates has been combined after confirming high degree of correlation), further termed peaks **A**ll **F**ragment (**AF**) peaks; or from only short fragments of sub-nucleosomal length (<100bp), with some of the datasets combined together based on the same tissue type and, in the case of human spermatogonial cells, similar cell morphology (see Tables 2.3 and 2.4 for details on which replicates were combined), further termed peaks **S**hort **F**ragment (**SF**) peaks. For visualization, `.bedGraph` format files of the fragment pileups output by MACS2 (genome-wide normalized coverage of fragments) were converted to the binary `.bigWig` format and uploaded to the UCSC Genome Browser mirror installed on a local secure machine. Separate files with the non-normalized fragment coverages were created using the `bedtools genomecov -bg` (Quinlan and Hall, 2010) command for downstream analysis where required.

2.2.7 Peak classification

AF peaks were classified as either *'tissue-specific'* or *'common'* based on the number of tissues and replicates each peak has been called in. A schematic demonstrating this classification is shown in Figure 2.4. Peaks defined as 'present' in all of the datasets were classified as *'common'*. In order for the region to be classified as *'tissue specific'*, the AF peak had to be found 'present' in at least one of the datasets belonging to the tissue type and 'not present' in all of the other datasets belonging to different tissue types. For each of the individual datasets, the SF peaks that intersected with each of the AF-peak defined category of regions were used to get the exact coordinate and the peak scores. MACS2-derived peak scores represent fold enrichment for the peak summit

against a random Poisson distribution with local (5Kb and 10Kb) lambda (Zhang et al., 2008). These scores were then used to match each of the peaks in the *'tissue-specific'* category to a corresponding peak in the *'common'* category, thereby creating two files with peak coordinates with matched distributions of peak scores, that being done for each of the datasets. A separate ('ultimate') set of *'common'* peaks was constructed by defining the genomic coordinates covered by SF peaks from all the tissue datasets analysed (see Figure 2.4 bottom left panel).

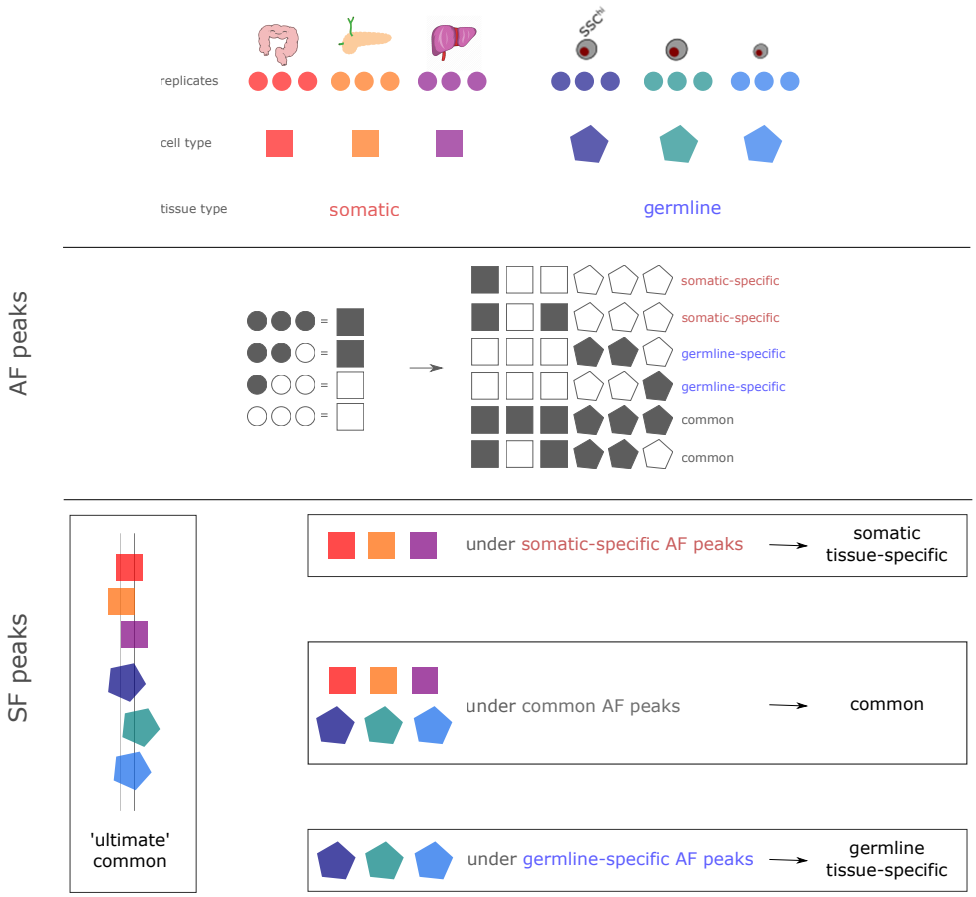


Figure 2.4: Peak classification. From the top: AF peaks were called from individual replicates (circles) for each of the cell types. Overlapping peaks were merged (using `bedtools merge`*) and considered as one. Any somatic tissue or germline cell type (squares and pentagons, respectively) with > 1 biological replicate (n) required the peak to be present in at least $n - 1$ replicates to be considered as 'present'. AF peaks that were 'present' only in somatic or only in germline tissue/cell types were classed as either 'somatic-specific' or 'germline-specific', respectively. SF peaks were called from the set of fragment where replicates for each tissue/cell type were combined together. SF peaks that overlapped with a particular category of the AF peaks were defined as SF peaks belonging to that category. An "ultimate" set of common peaks was defined as genomic coordinates covered by SF peaks from all the tissue/cell type datasets analysed.*

2.2.8 An alternative method for protein-binding site identification

The exact location of the Tn5 insertion site can be a useful measure when looking for single-nucleotide precision level span of DNA-protein interaction. As mentioned before, open chromatin regions that are easily accessible to Tn5 enzyme would exhibit a high frequency of adaptor insertion, while sites protected by TFs bound to DNA are likely to be depleted from the Tn5 enzyme activity. Those short insertion-depleted regions within broader regions of frequent insertion would form sites similar to what are commonly termed 'footprints' when looking at DNase-seq data. Footprints are not always strongly pronounced and, depending on the depth of the fragment coverage, are not easily identifiable. From visual inspection of patterns of the Tn5 insertion frequency, it came to our attention that insertion sites tend to cluster at one or both edges of the potential protein-binding sites (as defined by measures such as ChIP-seq and presence of the TF-specific motif; see Figure 2.5 for examples).

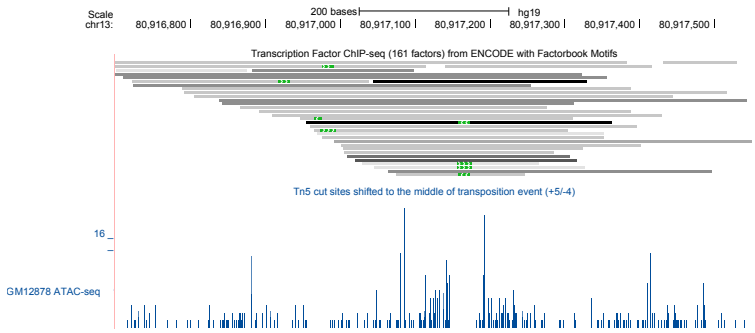


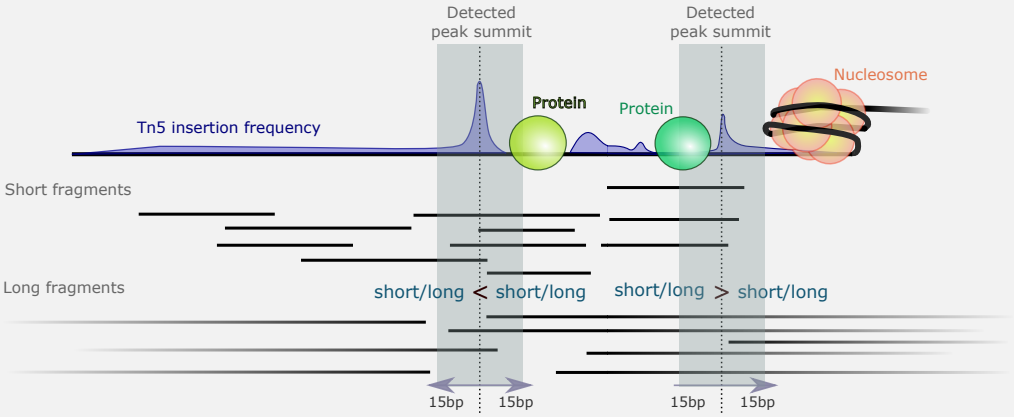
Figure 2.5: UCSC browser snapshot showing an example of Tn5 insertion frequency (blue bars) around a potential binding site defined by ChIP-seq signal (grey) and presence of binding motif (green).

I have therefore employed this observation to develop a **F**ragment **L**ength **O**ccurrence **P**ropensity method (FLOP) to identify edges of the bound proteins and define the location of protein relative to those edges. Initially peaks of increased frequency of Tn5 insertion are defined. Tn5 insertion sites are defined as coordinates of the 5' and 3' ends of the insert fragments. The Tn5 enzyme inserts two adaptors at the same transposition event, one on the Crick and one on the Watson strand, and those insertions sites are known to be separated by 9 base pairs (Adey et al., 2010; Buenrostro et al., 2013). It

is currently a common practice when analysing ATAC-seq data to shift the coordinates of the adapter insertion sites at the 5' end of the fragment and 3' end of the fragment by +4 and -5 bp, respectively (Buenrostro et al., 2013). The resulting coordinate would then represent the middle of transposition event. Those 9 base pairs between the insertion sites must be physically occupied by the Tn5 enzyme, and therefore represents an unoccupied region on DNA, 9 base pairs in length. Therefore, when calling peaks of the insertion sites, each mid-transposition event coordinate was shifted in the 5' direction by 4 bp and then extended by 9 bp towards 3' (`macs2 callpeak -f BED --keep-dup all --nomodel --shift 4 --extsize 9 --call-summits`). Assignment of edges of bound proteins is described in Box 2.2. Results validating this novel approach are presented in Subsection 2.3.7.

Box 2.2: Description of the FLOP method

Peak summits were assigned to represent edges of interacting proteins, and compared the ratios of short (<100bp, sub-nucleosomal length) and long (>100bp) fragments on either side of the peak summit to define the edge as being located either to the left or to the right of the DNA bound protein. Due to the increased insertion frequency around the interacting protein, a higher number of the fragments spanning it would be expected to be short. Lower number of short fragments would be present immediately outside the edge with the presence of anything as large as a nucleosome next to the protein. The total coverage of long and short fragments within the 15 bp windows immediately to the right and to the left of the peak summit were determined and the ratio of short vs long coverage was calculated for each window.



If the absolute ratio was higher in the right window, then the peak was classified as the left side of the protein and vice versa, as in formula:

SF: *Short fragment (<100bp) coverage* LW: *Left window (15 bp)*

LF: *Long fragment (>100bp) coverage* RW: *Right window (15 bp)*

$$FLOP_{rat} = \frac{SF_{LW} \times LF_{RW}}{LF_{LW} \times SF_{RW}} \qquad (f\text{:}2.2.10)$$

$FLOP_{rat} > 1$: right edge peak

$FLOP_{rat} < 1$: left edge peak

$FLOP_{rat} = 1$: ambiguous peak

2.3 Results

2.3.1 Data description and quality assessment

Details of each individual human specimen that has been collected and processed can be found in Table 2.1, along with the numbers of cells of every morphology type that were FAC-sorted from individual samples and details of the sequencing platforms that were used. Details of mouse sorted cells are in Table 2.2.

Accession	FACS population	Number of cells	Patient info	Date of collection	Platform	Read length
H1.1	Large, high SSC	24,000	Patient 2 (cancer)	23.02.2016	Illumina HiSeq2500	125
H1.2	Small	50,000			Illumina HiSeq2500	125
H2.1	Large, high SSC	25,000	Patient 3 (cancer)	02.03.2016	Illumina HiSeq2500	125
H2.2	Mixed large and small	60,000			Illumina HiSeq2500	125
H5.1	Large, high SSC	58,000	Patient 10 (non-cancer)	06.06.2016	Illumina HiSeq4000	75
H5.2	Large	36,000			Illumina HiSeq4000	75
H5.3	Small	65,000			Illumina HiSeq4000	75
H5.4	Large, high SSC	42,000			Illumina HiSeq4000	75
H5.5	Large	23,000			Illumina HiSeq4000	75
H7.2	Small	69,000	Patient 11 (cancer)	22.08.2016	Illumina HiSeq4000	75
H7.3	Large	69,000			Illumina HiSeq4000	75
H7.4	Large, high SSC	5,000			Illumina HiSeq4000	75
H8.3	Large and high SSC mixed	15,500	Patient 13 (non-cancer)	21.09.2016	Illumina HiSeq4000	75
H10.2	Large	24,000	Patient 18 (cancer)	19.10.2016	Illumina HiSeq4000	75

Table 2.1: Tissues collected and numbers of cells FAC-sorted into populations based on distinct morphological appearance: small (yellow), large (purple), and large cells with high side scatter (SSC) (blue). Colour correspond to the cell populations on Figure 2.6. Some populations contained cells of multiple morphological types (H2.2 and H8.3)

Accession	FACS population	Number of cells	Mouse number	Date of collection	Platform	Read length
m1	YFP-positive	55,000	1	28.07.2015	Illumina HiSeq2500	125
m5	YFP-positive	55,000	7	28.07.2015	Illumina HiSeq2500	125
m7	YFP-positive	75,000	7.1	07.09.2015	Illumina HiSeq2500	125

Table 2.2: Numbers of mouse-derived cells FAC-sorted based on YFP-fluorescence and sequencing platforms used for each sample.

Across human samples, on average, the 'small' cell populations had the largest number of sorted cells, followed by the 'large' category. Figure 2.6 shows an example of the FACS plots with the gates used for the sorting of separate cell populations.

23,000 - 69,000 cells were used for the subsequent ATAC-seq processing. Tables 2.3 and 2.4 show the details of the each of the processed ATAC-seq samples. Generally, our primary data showed little mitochondrial contamination (5-21%), which is known to be a problem with ATAC-seq analysis in some tissues (Montefiori et al, 2017). Mitochondrial sequences are normally discarded as unrelated to the scope of the experiment (Tsompana and Buck, 2014) and a high percentage of these would reduce the numbers of usable nuclear genomic reads. The proportion of PCR duplicates was estimated to be 10-24%. The number of resulting fragments for each of the individual samples varied between 100 and 150 million (just below 100 million for mouse data), which is a sufficient depth to be able to perform peakcalling for identification of potential binding sites. Figures 2.7 and 2.8 show the fragment length distribution for each human and mouse samples. All of the samples analysed show pattern of mono-, di- and tri- nucleosomal fragment length enrichment, indicating that the chromatin structure of cells was intact when the assay was performed.

In some samples (H1.1, H1.2, H2.1, H2.2, H5.3, H8.2, H7.2, H7.4, H8.3), visual inspection of the fragment pileups revealed lack of the localized enrichment of fragments that would be sufficient to form distinct peaks. Figure 2.9 shows a genome-wide-normalized fragment coverage around the promoter regions of some highly expressed housekeeping genes from Eisenberg and Levanon (2013). While some of the datasets show a clear enrichment of fragments over the promoter regions, others do not.

If the 'noisiness' reflects distressed chromatin state of cells, one would intuitively expect to see a difference in the fragment length distributions between those that exhibit it and those that do not. However, there did not seem to be any definitive difference. To quantitatively measure this 'noisiness', I have calculated the percentage of fragments that overlapped with the 'housekeeping' DNase-seq footprints defined by Reijns et al. (2015). Those measures for each of the samples can be seen in Tables 2.3 and 2.4. While not an absolute measure, the comparison of signal at these housekeeping-like binding sites as defined from many cell types to background signals provides a useful relative comparator of signal to noise between datasets. I postulate that such 'noisy' data could represent cells being subjected to high level of stress during the desegregation and, in particular, FACS procedures, possibly also indicating loss of nuclear integrity during preparation. Additionally, during the ATAC-seq processing of the samples H1 and H2, the amount of Tn5 enzyme was not titrated to

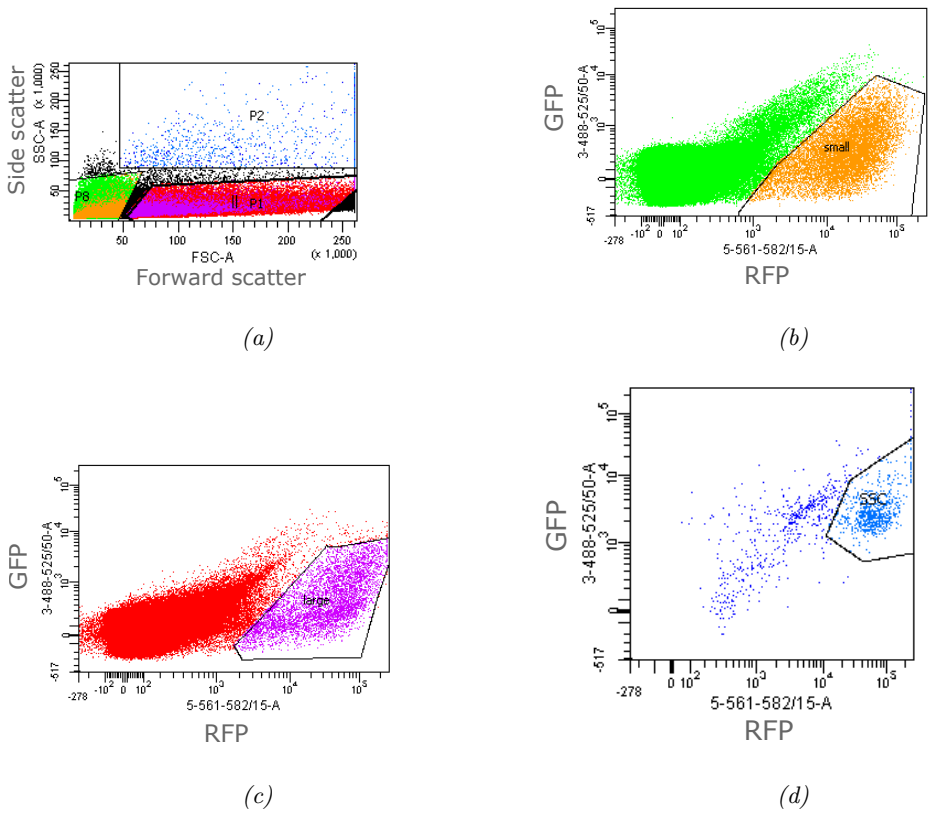


Figure 2.6: FACS plots showing the gates used for the sorting of different cell populations. (a) Gating of the cell populations based on the forward (FSC) and side (SSC) scatter (measures of cell size and granularity, respectively). Three cell populations were defined based on those measures - P2 (blue), P1 (red and purple), and P8 (yellow and green). Subpopulation of those have then been sorted based on the level of fluorescence (RFP - red fluorescence; GFP - green fluorescence), and "Small" (b), "Large" (c) and "SCC" (d) subpopulations forming bright clusters were isolated.

the numbers of cells. This could potentially have caused over-digestion of the DNA in those samples, even at the regions of DNA that were highly compacted or protected by bound proteins/nucleosomes. Percentages of the fragment-derived bases falling within the DNase-seq footprints was higher for almost all other human samples (Tables 2.3 and 2.4). For all the downstream analysis I have excluded the datasets that did not show formation of distinct peaks by the visual inspection on the fragment coverage and showed <3% overlap with the DNase-seq footprints. An example of the data can be seen on the Genome Browser snapshot in Figure 2.10.

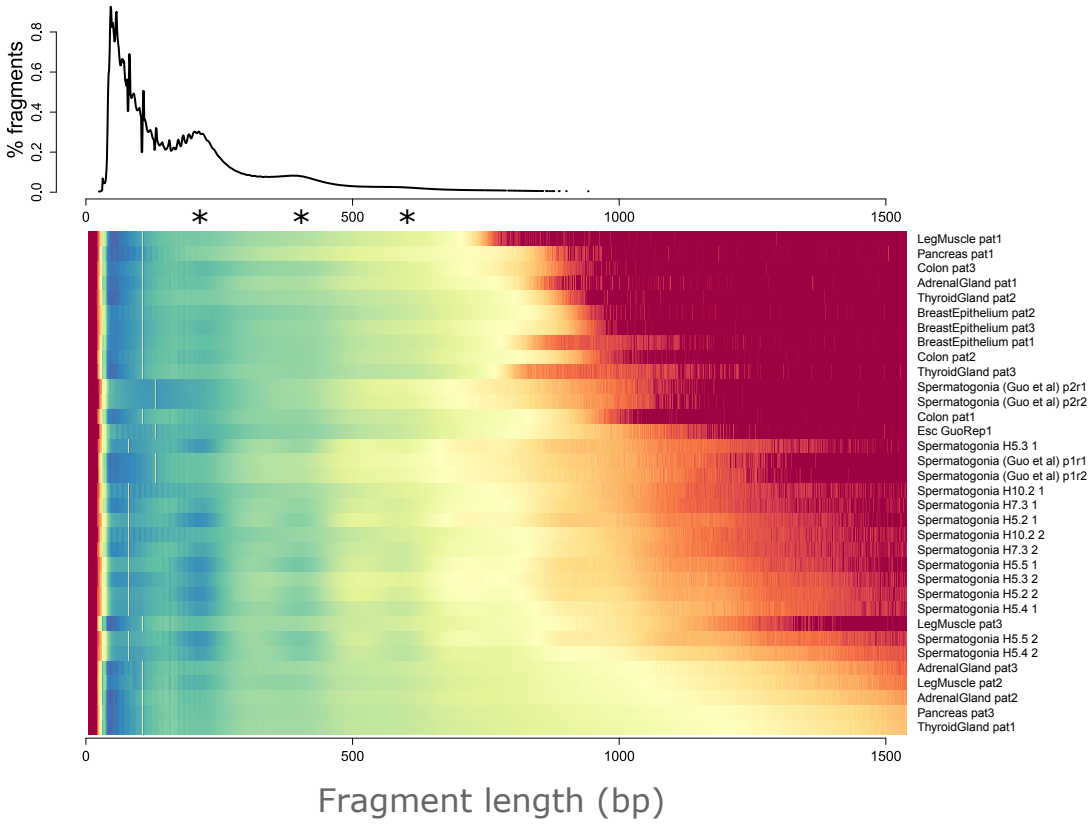


Figure 2.7: Distribution of the fragment length for each of the ***human*** *datasets. Histogram on the top represents the average percentage of fragments of each length among all the datasets. The dips at 48-49bp, 73-74bp and 123-124bp followed by peaks in the top histogram (also observable in the heatmap) result from the inability of the* `cutadapt` *software to trim the adaptor sequences at ends of reads that are <3bp when actual fragments are 1-3bp shorter than read length. For better color discrimination in the heatmap, numbers of fragments have been* log_2*-transformed. There is a substantial variability in the maximal lengths of the fragments, but most of the datasets show an observable enrichment in the mono-, di- and tri- nucleosomal length fragments (indicated by *).*

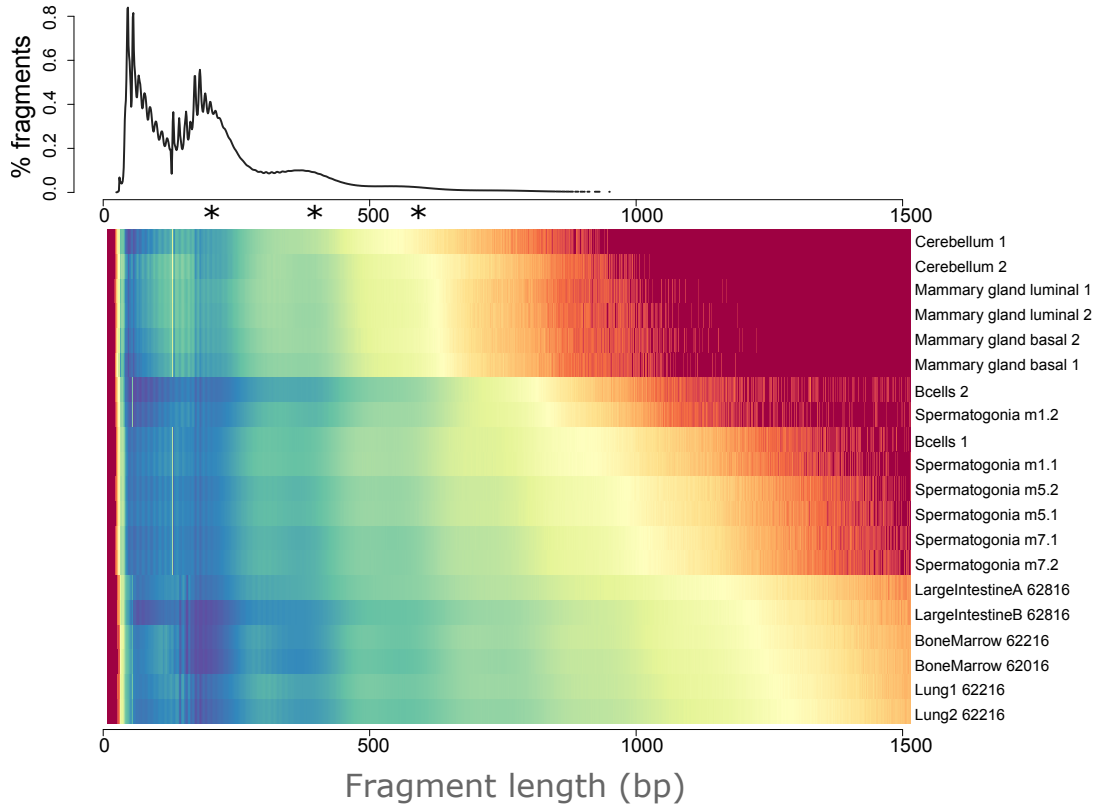


Figure 2.8: Distribution of the fragment length for each of the ***mouse*** *datasets. Histogram on the top represents the average percentage of fragments of each length among all the datasets. The dip at 123-124bp followed by peaks in the top histogram (also observable in the heatmap) result from the inability of the* `cutadapt` *software to trim the adaptor sequences at ends of reads that are <3bp when actual fragments are 1-3bp shorter than read length. For better color discrimination in the heatmap, numbers of fragments have been* log_2*-transformed. There is a substantial variability in the maximal lengths of the fragments, but most of the datasets show an observable enrichment in the mono-, di- and tri- nucleosomal length fragments (indicated by *).*

Sample	Reads	Quality < 30	Mapped	chrM	Duplicates	Fragments	DHSF overlap	Combined
Spermatogonia H1.1	230,098,226	18.35%	98.13%	6.93%	2.46%	85,088,415	1.14%	-
Spermatogonia H1.2	233,877,636	17.32%	97.99%	2.10%	2.47%	92,097,644	1.24%	-
Spermatogonia H10.2 1	132,637,829	16.41%	96.69%	5.40%	14.87%	44,190,459	4.62%	Sprmt H107
Spermatogonia H10.2 2	136,598,517	16.61%	96.52%	5.24%	16.12%	44,758,906	4.90%	Sprmt H107
Spermatogonia H2.1	249,846,463	17.54%	98.16%	8.00%	2.57%	92,139,778	0.74%	-
Spermatogonia H2.2	215,196,712	16.61%	97.91%	1.68%	2.04%	86,174,010	0.80%	-
Spermatogonia H5.1 1	137,721,470	16.78%	98.01%	13.80%	11.29%	43,620,699	3.54%	Sprmt H5.14
Spermatogonia H5.1 2	129,722,995	17.09%	97.90%	14.19%	13.04%	39,911,943	3.48%	Sprmt H5.14
Spermatogonia H5.2 1	190,262,479	15.32%	97.87%	7.24%	14.53%	63,574,959	3.81%	Sprmt H5.25
Spermatogonia H5.2 2	178,853,134	15.89%	97.60%	7.65%	15.82%	58,117,781	3.75%	Sprmt H5.25
Spermatogonia H5.3 1	136,285,039	16.42%	97.06%	9.72%	21.07%	40,215,216	2.17%	-
Spermatogonia H5.3 2	141,122,757	15.58%	97.74%	9.43%	21.09%	42,352,220	2.09%	-
Spermatogonia H5.4 1	167,165,475	17.64%	97.90%	20.32%	23.55%	41,701,746	5.28%	Sprmt H5.14
Spermatogonia H5.4 2	165,009,328	18.28%	97.43%	20.80%	23.85%	40,325,604	5.20%	Sprmt H5.14
Spermatogonia H5.5 1	119,428,755	14.97%	97.61%	6.70%	14.29%	40,397,045	3.26%	Sprmt H5.25
Spermatogonia H5.5 2	111,489,253	15.70%	97.13%	7.05%	14.98%	36,850,218	3.24%	Sprmt H5.25
Spermatogonia H7.2 1	111,302,579	18.54%	95.77%	7.44%	13.77%	35,619,573	1.74%	-
Spermatogonia H7.2 2	105,023,790	18.30%	96.12%	7.75%	12.43%	34,190,077	1.60%	-
Spermatogonia H7.3 1	191,743,346	18.52%	96.72%	11.05%	9.58%	62,126,764	6.58%	Sprmt H107
Spermatogonia H7.3 2	205,508,091	18.73%	96.51%	10.72%	10.66%	65,770,645	7.07%	Sprmt H107
Spermatogonia H7.4 1	129,286,318	18.03%	96.35%	6.72%	16.46%	40,757,994	1.52%	-
Spermatogonia H7.4 2	124,091,564	17.76%	96.68%	6.77%	15.29%	39,855,725	1.41%	-
Spermatogonia H8.3 1	145,550,935	16.33%	96.42%	3.07%	29.98%	40,782,928	1.09%	-
Spermatogonia H8.3 2	142,682,719	16.06%	96.74%	3.11%	29.00%	40,733,402	1.04%	-

Table 2.3: Summary statistics of primary spermatogonial cell ATAC-seq datasets that were analyzed. ***Sample****: the name of individual sequenced samples (same samples run on different lanes are not combined here);* ***Reads****: raw numbers of reads ;* ***Quality*** $<$ ***30*** *: percentage of reads with map-quality* $<$ *30 ;* ***Mapped*** *: percentage of reads mapped (without quality filter) ;* ***chrM*** *: percentage of reads aligned to the mitochondrial genome ;* ***Duplicates*** *: percentage of fragments that were discarded as PCR duplicates ;* ***Fragments*** *: total number of fragments (de-duplicated);* ***DHSF overlap*** *: percentage of fragment bases overlapping with housekeeping DNase-seq footprints ;* ***Combined*** *: samples with the same combined dataset name were merged for SF-peak identification*

Sample	Reads	Quality < 30	Mapped	chrM	Duplicates	Fragments	DHSF overlap	Combined
AdrenalGland pat1	64,084,767	25.68%	98.27%	45.25%	3.11%	12,567,420	6.47%	AdrenalGland
AdrenalGland pat2	143,535,298	15.69%	98.15%	15.09%	3.00%	49,587,069	10.39%	AdrenalGland
AdrenalGland pat3	75,905,207	21.35%	98.60%	42.89%	1.67%	16,700,709	9.39%	AdrenalGland
BreastEpithelium pat1	166,177,596	17.27%	96.13%	7.13%	12.95%	55,125,612	6.21%	BreastEpithelium
BreastEpithelium pat2	167,832,058	16.46%	96.55%	5.02%	12.78%	57,450,225	8.39%	BreastEpithelium
BreastEpithelium pat3	193,556,412	18.92%	96.79%	7.70%	12.31%	63,025,465	4.50%	BreastEpithelium
LegMuscle pat1	98,841,236	15.79%	96.92%	4.89%	4.12%	37,663,826	5.75%	LegMuscle
LegMuscle pat2	58,955,334	11.58%	98.21%	6.31%	2.44%	23,729,383	11.95%	LegMuscle
LegMuscle pat3	57,095,849	13.55%	97.55%	3.25%	2.79%	23,156,311	8.23%	LegMuscle
Pancreas pat1	49,509,733	21.23%	97.99%	23.86%	4.04%	14,175,734	1.98%	Pancreas
Pancreas pat2	87,463,109	19.53%	97.28%	15.02%	3.89%	28,545,561	3.83%	Pancreas
Pancreas pat3	126,973,192	16.23%	98.23%	20.92%	4.94%	39,761,314	18.71%	Pancreas
SigmoidColon pat1	173,532,487	17.65%	95.43%	7.19%	14.52%	56,268,440	4.97%	SigmoidColon
SigmoidColon pat2	179,741,814	16.74%	95.90%	12.11%	16.43%	54,438,179	20.55%	SigmoidColon
SigmoidColon pat3	174,419,995	18.41%	94.14%	6.78%	12.09%	57,919,070	6.20%	SigmoidColon
Spermatogonia (Guo et al) p1r1	47,692,375	17.98%	98.03%	11.04%	2.68%	16,865,889	7.45%	Sprmt Guo
Spermatogonia (Guo et al) p1r2	65,269,203	18.28%	97.84%	11.69%	4.08%	22,481,064	6.76%	Sprmt Guo
Spermatogonia (Guo et al) p2r1	71,378,610	19.19%	97.82%	13.43%	3.33%	24,011,100	3.00%	Sprmt Guo
Spermatogonia (Guo et al) p2r2	64,975,329	17.95%	97.49%	8.93%	1.76%	23,700,990	2.34%	Sprmt Guo
ThyroidGland pat1	99,075,248	13.16%	98.26%	9.51%	4.13%	37,119,172	14.10%	ThyroidGland
ThyroidGland pat2	107,085,475	17.60%	97.01%	13.16%	16.70%	31,727,379	7.99%	ThyroidGland
ThyroidGland pat3	91,374,702	20.16%	97.27%	24.89%	15.69%	22,959,997	10.93%	ThyroidGland

Table 2.4: Summary statistics of ATAC-seq datasets from other studies that were analyzed. ***Sample****: the name of individual sequenced samples (same samples run on different lanes are not combined here);* ***Reads****: raw numbers of reads ;* ***Quality*** $<$ ***30*** *: percentage of reads with map-quality* $<$ *30 ;* ***Mapped*** *: percentage of reads mapped (without quality filter) ;* ***chrM*** *: percentage of reads aligned to the mitochondrial genome ;* ***Duplicates*** *: percentage of fragments that were discarded as PCR duplicates ;* ***Fragments*** *: total number of fragments (de-duplicated);* ***DHSF overlap*** *: percentage of fragment bases overlapping with housekeeping DNase-seq footprints ;* ***Combined*** *: samples with the same combined dataset name were merged for SF-peak identification*

Sample	Reads	Quality < 30	Mapped	chrM	Duplicates	Fragments	Combined
B Cells 1	144,485,716	18.46%	97.63%	15.21%	11.24%	44,101,843	B Cells 1
B Cells 2	190,621,773	17.61%	97.92%	11.44%	10.83%	61,698,280	B Cells 2
Bone Marrow 62016	156,854,820	9.87%	100.00%	0.00%	0.77%	69,981,835	Bone Marrow
Bone Marrow 62216	142,049,329	9.59%	100.00%	0.00%	0.79%	63,532,973	Bone Marrow
Cerebellum 1	70,528,067	18.45%	98.44%	15.11%	7.10%	22,580,486	Cerebellum
Cerebellum 2	57,569,155	18.81%	98.54%	22.73%	20.62%	14,272,990	Cerebellum
Cerebellum 3	73,176,730	18.86%	98.53%	22.75%	17.17%	18,916,302	Cerebellum
Large Intestine A 62816	100,203,602	9.98%	100.00%	0.00%	0.48%	44,734,526	Large Intestine
Large Intestine B 62816	191,293,803	9.86%	100.00%	0.00%	0.63%	85,473,658	Large Intestine
Lung 1 62216	112,036,458	11.31%	100.00%	0.00%	1.71%	48,655,303	Lung
Lung 2 62216	134,206,386	9.93%	100.00%	0.00%	1.34%	59,471,676	Lung
Mammary gland basal 1	87,340,224	22.16%	96.74%	24.09%	11.00%	22,682,989	Mammary gland basal
Mammary gland basal 2	68,699,952	22.74%	96.86%	27.24%	8.64%	17,438,676	Mammary gland basal
Mammary gland luminal 1	64,685,078	25.83%	96.70%	30.44%	13.00%	14,344,765	Mammary gland luminal
Mammary gland luminal 2	62,573,029	25.60%	96.15%	27.15%	11.77%	14,741,166	Mammary gland luminal
Spermatogonia m1.1	85,748,528	19.38%	97.83%	3.61%	3.25%	31,994,578	Spermatogonia
Spermatogonia m1.2	91,986,843	19.42%	98.31%	3.58%	3.50%	34,317,952	Spermatogonia
Spermatogonia m5.1	100,668,906	19.93%	97.84%	3.53%	4.25%	36,942,918	Spermatogonia
Spermatogonia m5.2	108,782,025	19.97%	98.34%	3.51%	4.62%	39,872,695	Spermatogonia
Spermatogonia m7.1	124,039,281	20.21%	98.16%	2.96%	6.75%	44,555,521	Spermatogonia
Spermatogonia m7.2	116,574,885	20.23%	97.54%	2.98%	6.32%	41,926,746	Spermatogonia

Table 2.5: Summary statistics of mouse ATAC-seq datasets that were analyzed. ***Sample****: the name of individual sequenced samples (same samples run on different lanes are not combined here);* ***Reads****: raw numbers of reads ;* ***Quality*** $<$ ***30*** *: percentage of reads with map-quality* $<$ *30 ;* ***Mapped*** *: percentage of reads mapped (without quality filter; reads from lung, large intestine and bone marrow were obtained pre-mapped, and therefore have value of 100%) ;* ***chrM*** *: percentage of reads aligned to the mitochondrial genome ;* ***Duplicates*** *: percentage of fragments that were discarded as PCR duplicates ;* ***Fragments*** *: total number of fragments (de-duplicated);* ***Combined*** *: samples with the same combined dataset name were merged for SF-peak identification*

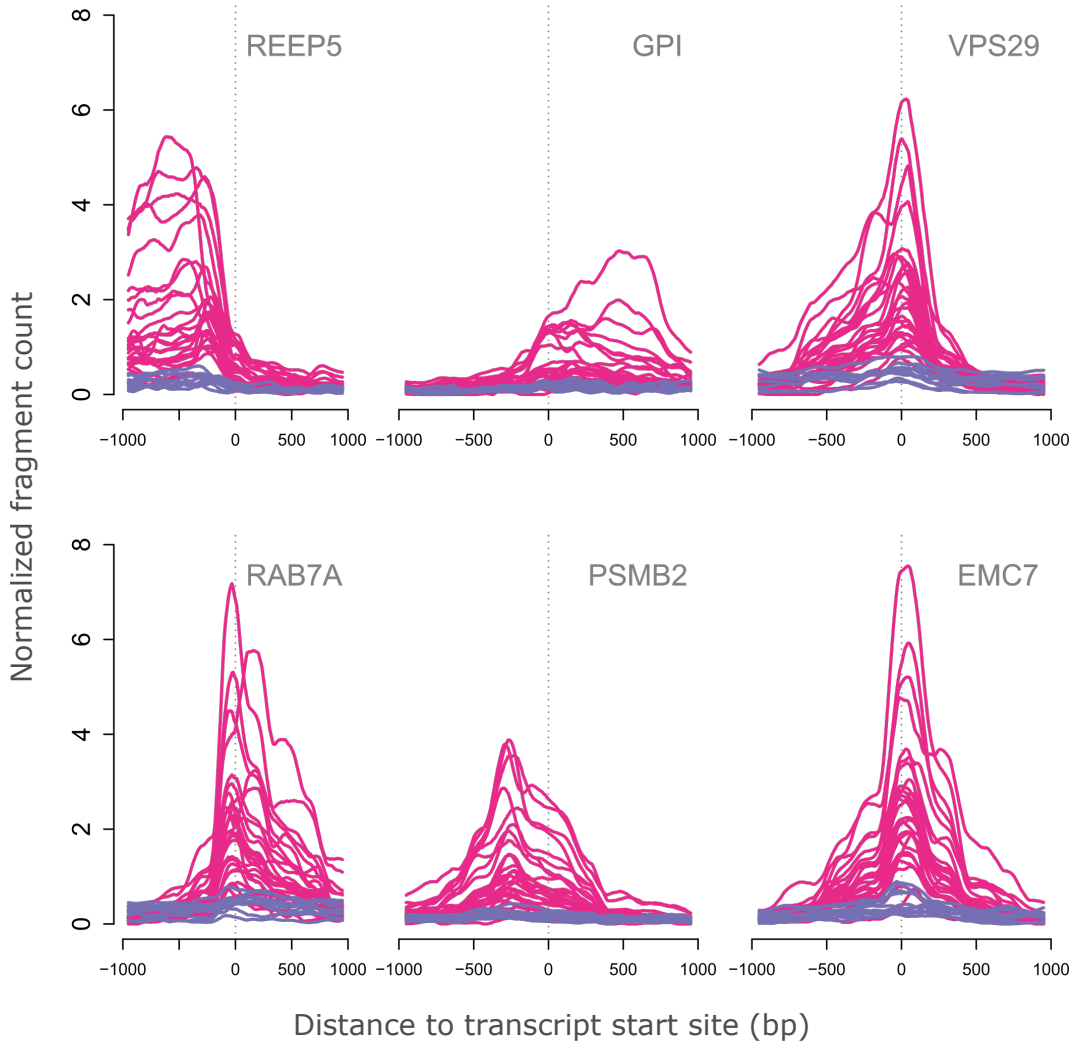


Figure 2.9: Normalized ATAC-seq fragment coverage of all of the human datasets across a number of highly expressed housekeeping gene transcription start sites (-1000:+1000bp from the TSS). Datasets that have been used for subsequent analysis (Spermatogonia H10.2, H5.1, H5.2, H5.4, H5.5, H7.3; data from all patient from adrenal gland, breast epithelium, leg muscle, pancreas, sigmoid colon, thyroid gland, spermatogonia (Guo et al)) are in pink, while the ones that have been excluded from analysis (Spermatogonia H1.1, H1.2, H2.1, H2.2, H5.3, H7.2, H7.4, H8.3) are in purple. Kept datasets show a more prominent open state of the promoters that are expected to be accessible in all of these cell types, indicating a better quality of data. Datasets that have been excluded failed to form detectable peaks and therefore would not be suited for analysis.

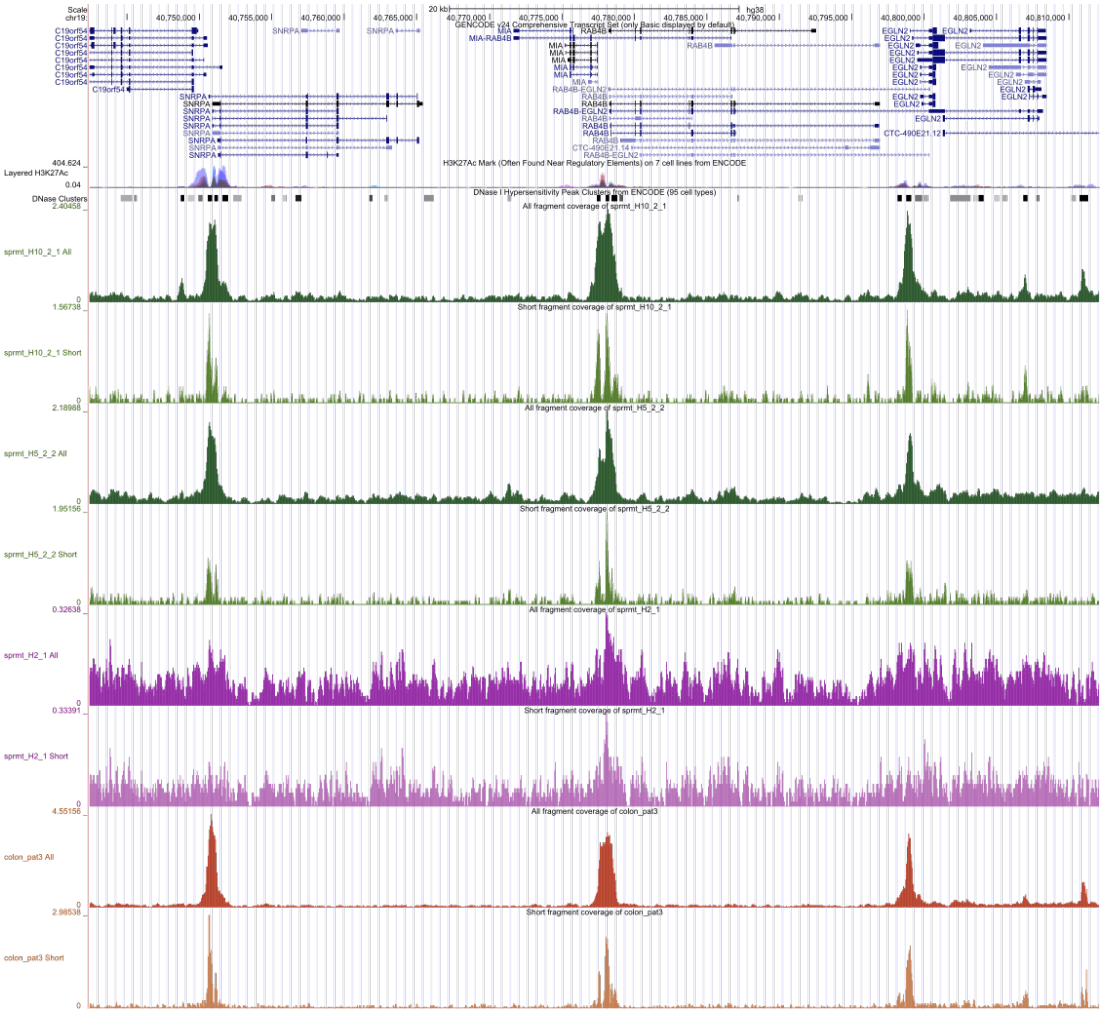
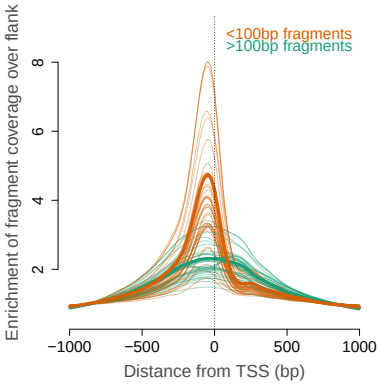


Figure 2.10: An example Genome Browser track snapshot showing the coverage of all fragment lengths (dark) and short fragment coverage (light) for two of high quality spermatogonia datasets (green), a spermatogonia dataset with 'noisy' fragment distribution (purple), and colon dataset (orange). Clear formation of peaks is observable at the promoter regions of several genes (blue, top), except for the purple dataset, corresponding to the patterns of H3K27Ac (histone mark associated with active transcription) and DNase-seq Clusters (black). H3K27Ac marks are enriched in the gaps between SF peaks, where nucleosomes would be expected to reside.

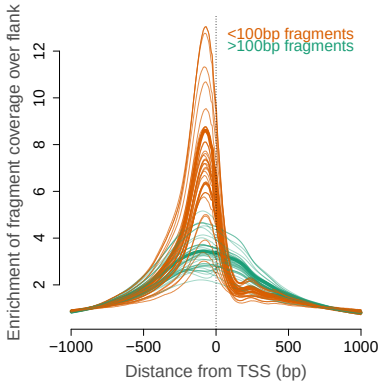
2.3.2 Enrichment of sub-nucleosomal length fragments mark regulatory regions, while Tn5 insertion sites can mark some individual protein-binding sites

Plotted coverage of fragments around RefSeq (O'Leary et al., 2016) and CAGE-defined transcription start sites (TSSs) from Young et al. (2015) can be seen in the Figures 2.11a and 2.11b, along with the fragment coverage over previously-defined DNase-seq footprints (Figure 2.11d). ATAC-seq fragment coverage shows expected patterns of enrichment around TSSs and DNase-seq footprints. These TSS locations are either an aggregate of all defined transcription start sites (as in RefSeq), or are expressed specifically in testes (as in CAGE). Both datasets, however, contain regions that represent housekeeping regulatory sites and patterns of accessibility at those is a good check for the quality of data and its ability to identify the housekeeping and, hopefully, cell-type specific regulatory regions.

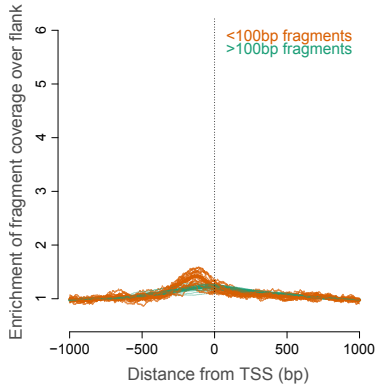
To assess the suitability of the ATAC-seq data for identification of the individual protein-binding sites, I looked at the distribution of the Tn5 insertion sites in the colon tissue around the motifs of TFs found under ChIP-seq peaks in the matching cell type (Figure 2.12). Colon was chosen because both ATAC-seq from the tissue and ChIP-seq data (Yan et al., 2013) from a tissue-derived cell line could be obtained for multiple different TFs. I also looked at the distribution of the insertion sites around the protein-binding sites defined by DNase-seq footprinting (Figure 2.13). While there is heterogeneity in the patterns of Tn5 insertion frequencies around motifs of different TFs, many sites show peaks on both sides of the motif. For some TFs, the insertion frequency does not form obvious footprints with strong depletion at and around the motif, but instead exhibit formation of ragged 'insertion signatures' within the motif, similar to what has previously been observed for DNase-seq data (Neph et al., 2012; He et al., 2014; Sung et al., 2014; Baek et al., 2017).



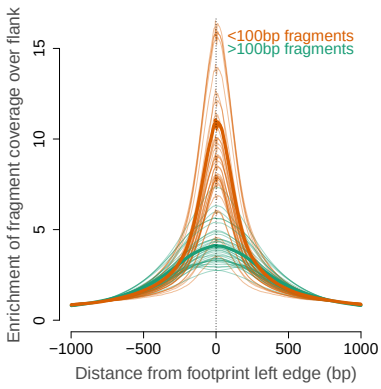
(a) TSSs defined by RefSeq



(b) Promoter regions defined by CAGE-tag enrichment in human testis tissue



(c) Muscle-restricted promoters (by CAGE) and spermagonial ATAC-seq fragments



(d) DNase-seq footprints

Figure 2.11: Coverages of short (<100bp) and long (>100bp) fragments over TSSs (a, b, and c) and DNase-seq footprints (d). Promoters have been oriented so that the transcription would initiate at x=0 and proceed towards the right. Each individual dataset is represented by a thin line, while the mean of all the datasets is represented by a bold line. There is greater enrichment of the sub-nucleosomal length fragments at the promoter regions with depletion at the site where the +1 nucleosome would be expected to be located (a,b). Enrichment of spermatogonia-derived ATAC-seq fragments is absent at the muscle-restricted promoters (c), as expected. Similarly, there is an enrichment of the sub-nucleosomal length fragments over the DNase-seq footprints (d). This shows that sub-nucleosomal length fragments are indeed enriched at regions where proteins are expected to bind and are depleted at closed non-accessible regions and at sites occupied by nucleosomes.

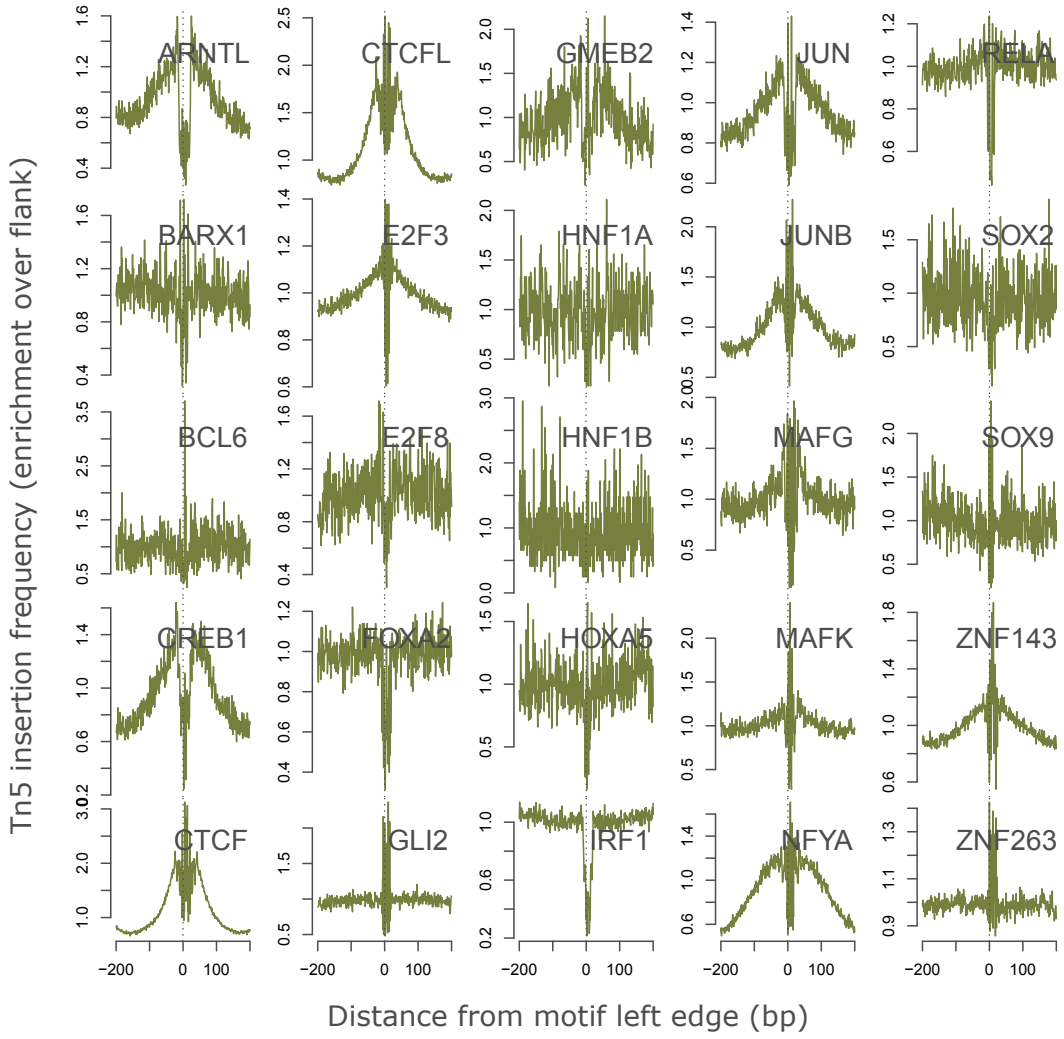


Figure 2.12: Enrichment (relative to the flank) in the frequency of Tn5 insertions from one of the colon ATAC-seq samples. Individual plots are centred on aggregate of motif mid-points, and a TF would be expected to bind in the middle. These motifs were found under ChIP-seq peaks of corresponding TFs in LoVo cell line (colorectal adenocarcinoma-derived cell line). Tn5 insertion pattern shows identifiable peaks on both sides of some of the TF binding sites (such as CREB1 and CTCF), while binding sites of others are more poorly captured (such as HNF1B and SOX2).

At least in relation to DNase-seq footprints, the observed specific profiles of the nuclease digestion over the binding sites of different TFs have previously been shown to reflect the physical interaction of protein with DNA (Neph et al., 2012), or argued to reflect nature of the DNA structure (He et al., 2014; Sung et al., 2014; Baek et al., 2017), rather than protein binding property. The depth of the footprint has also been reported to be dependent on the dynamic nature of protein binding, with TFs exhibit-

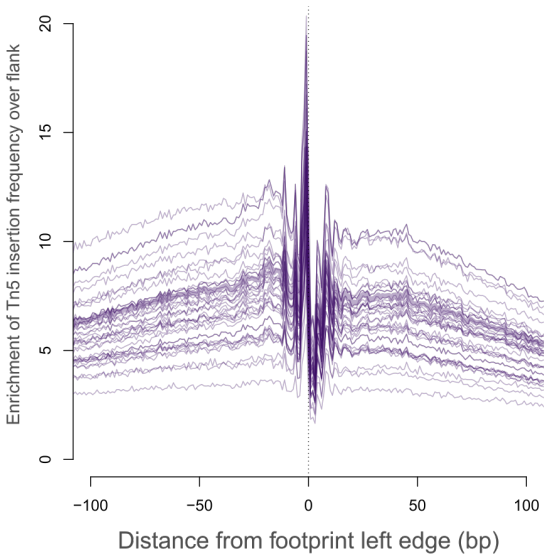


Figure 2.13: Enrichment of the Tn5 insertion sites from all the analyzed datasets (enrichment over the average calculated from -1000:-800 and +800:+1000 regions away from DNase-seq footprint left edge) centered on the left edge of footprints. Tn5 insertion sites show a distinct peak of enrichment coinciding with the edge of DNase-seq footprint. Average width of a footprint is ≈10bp.

ing short residency time not producing deep footprints (Sung et al., 2014). ATAC-seq data is likely to be similar to DNase-seq data in this respect, although exhibiting a distinct enzyme-specific bias and therefore generating different footprint shapes (Calviello et al., 2018). Therefore, not all of the TF binding sites would be identifiable by Tn5 insertion, especially for TFs with short residence time, and some erroneous assignments are possible, particularly with the TFs that exhibit strong insertion signatures within the motif.

2.3.3 Isolated cells show open spermatogonial cell promoters by ATAC-seq

While the human cells that we have isolated are likely to be spermatogonial cells, as they were all FGFR3-positive, those cells also represent distinct populations of size and shape and potentially amount to various stages of the spermatogonial lineage. To better understand the identity of these cellular sub-populations, I investigated the promoter regions of some of the genes that are thought to be active or poised in our cells of interest. Figure 2.14 shows ATAC-seq fragment coverage from spermatogonial cell datasets over the promoter regions of some of the pluripotency-related genes that have been shown to be expressed in spermatogonia, along with the germ cell-specific markers. While accessible chromatin defined by ATAC-seq peaks is often correlated with transcription initiation activity and gene expression (Wang et al., 2018), it is not

a perfect correlate.

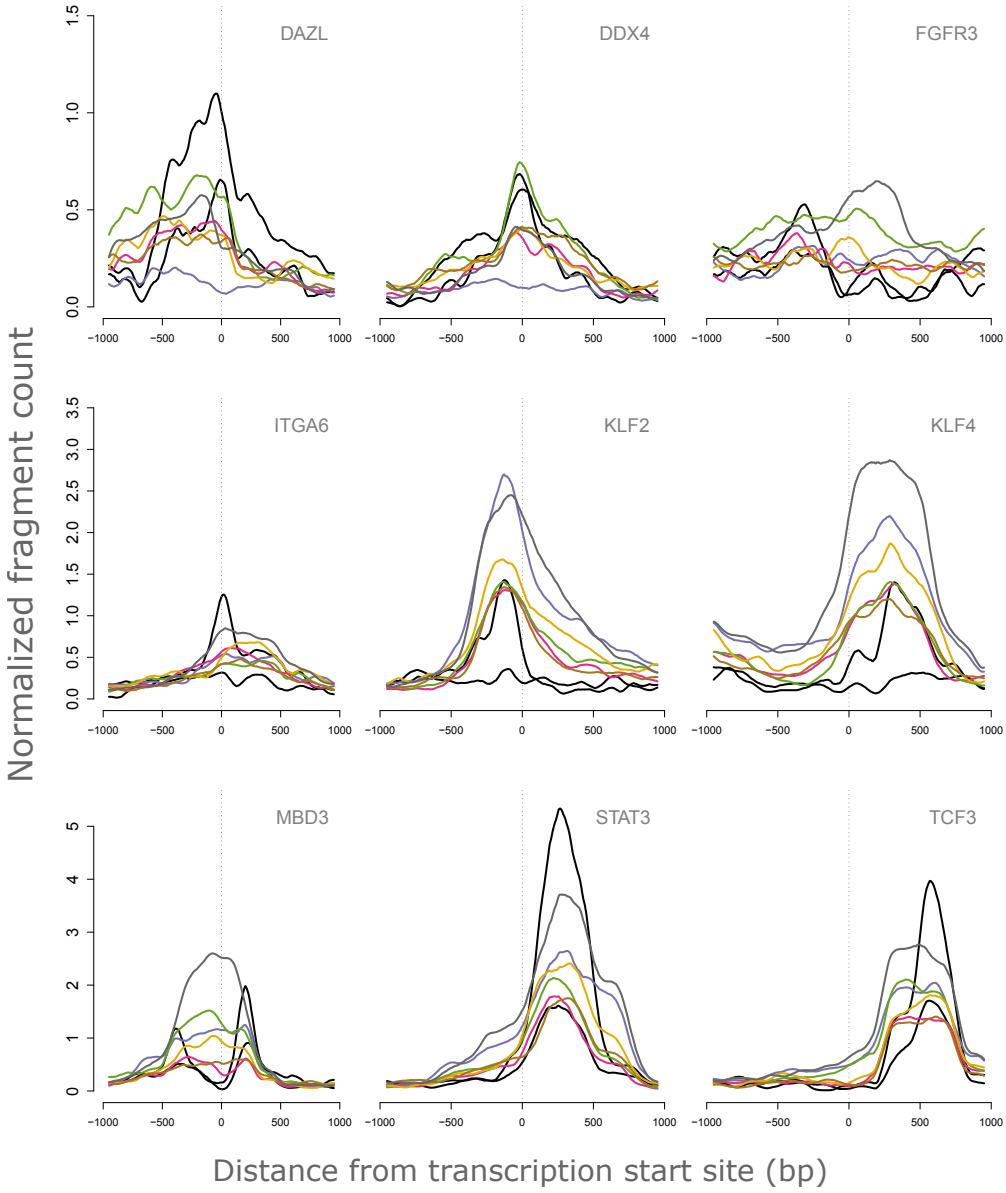


Figure 2.14: Fragment coverage over the germ-cell (DAZL, DDX4 and FGFR3) and pluripotency (ITGA6, KLF2, KLF4, MBD3, STAT3 and TCF3) gene promoters. Coloured lines represent our primary data, while black lines are hSCCs from two patients from Guo et al. (2017).

Surprisingly, the promoter region of the *FGFR3* gene does not show a peak, as would be expected, as that is the marker that was used to FAC-sort the cells in question. Promoters of *DAZL* and *DDX4* genes, which are germ-specific markers show some degree of openness (in all datasets except for H10.2, in purple), while most of the pluripotency markers that have been previously been observed to be expressed in both

spermatogonial stem cells and those committing to differentiation (Guo et al., 2017) show clear ATAC-seq peaks.

2.3.4 Analysis of peaks

The numbers of AF peaks called from each of the datasets can be seen in the Tables 2.6 (human), and 2.7 (mouse).

Source tissue	Total AF peaks	Common AF peaks	Specific AF peaks	Common	Specific
Spermatogonia H5.2	175,407	11,378	5,235	6.49%	2.98%
Spermatogonia H7.3	146,700	12,188	13,166	8.31%	8.97%
BreastEpithelium pat1	129,397	15,018	6,696	11.61%	5.17%
Spermatogonia H10.2	127,210	11,017	8,436	8.66%	6.63%
BreastEpithelium pat3	110,632	14,285	7,417	12.91%	6.70%
SigmoidColon pat2	105,205	15,356	11,927	14.60%	11.34%
Colon pat2	105,204	15,381	11,909	14.62%	11.32%
BreastEpithelium pat2	104,982	14,091	7,307	13.42%	6.96%
Spermatogonia H5.4	96,036	11,599	4,216	12.08%	4.39%
SigmoidColon pat1	89,523	14,523	10,172	16.22%	11.36%
Colon pat1	89,472	14,541	10,142	16.25%	11.34%
Spermatogonia H5.5	82,804	11,020	2,461	13.31%	2.97%
AdrenalGland pat2	81,050	15,066	13,042	18.59%	16.09%
ThyroidGland pat1	79,637	15,091	18,729	18.95%	23.52%
Spermatogonia H5.1	74,259	11,546	3,083	15.55%	4.15%
Colon pat3	73,973	14,271	8,298	19.29%	11.22%
Pancreas pat3	61,051	13,519	2,247	22.14%	3.68%
ThyroidGland pat3	53,414	14,270	17,254	26.72%	32.30%
ThyroidGland pat2	52,832	14,001	12,416	26.50%	23.50%
LegMuscle pat2	52,016	13,295	12,276	25.56%	23.60%
LegMuscle pat3	50,794	13,474	12,384	26.53%	24.38%
AdrenalGland pat3	47,818	13,162	9,985	27.53%	20.88%
Spermatogonia (Guo et al) p2	46,961	8,802	1,711	18.74%	3.64%
LegMuscle pat1	43,105	13,190	7,003	30.60%	16.25%
Spermatogonia (Guo et al) p1	34,073	11,497	3,754	33.74%	11.02%
AdrenalGland pat1	28,824	11,931	4,765	41.39%	16.53%
Pancreas pat2	21,093	11,828	2,185	56.08%	10.36%
Pancreas pat1	6,070	4,484	119	73.87%	1.96%

*Table 2.6: Numbers of **human** AF (all fragment) peaks identified in all of the analyzed datasets with counts and proportions of peaks that have been classified as 'common' or 'tissue-type specific' based on comparisons with replicates and other tissues.*

Datasets H5.2 and H5.5; H5.1 and H5.4; H10.2 and H7.3 were combined for the SF peak calling based on similar size and shape appearance of those cells in FACS. All of the mouse spermatogonial datasets (m1, m5 and m7) were combined for SF peak calling. Matching cell types tend to exhibit better correspondence between peak scores in linear regression (Figures 2.15 and 2.16). Neither of the three human pancreas

Source tissue	Total AF peaks	Common AF peaks	Specific AF peaks	Common	Specific
Lung 2 62216	107,443	9,979	36,640	9.29%	34.10%
Spermatogonia m1	98,941	9,572	5,481	9.67%	5.54%
Lung 1 62216	89,496	9,807	36,051	10.96%	40.28%
Cerebellum 3	82,223	9,807	30,862	11.93%	37.53%
Bone Marrow 62016	79,129	8,894	12,408	11.24%	15.68%
Large Intestine B 62816	71,241	9,628	17,447	13.51%	24.49%
Bone Marrow 62216	65,609	8,744	11,075	13.33%	16.88%
Spermatogonia m5	65,289	9,050	5,586	13.86%	8.56%
Cerebellum 1	61,018	9,459	26,821	15.50%	43.96%
Mammary gland basal 1	56,954	9,296	28,831	16.32%	50.62%
Cerebellum 2	54,875	9,301	26,033	16.95%	47.44%
Mammary gland basal 2	45,151	9,077	22,587	20.10%	50.03%
Bcells 1	40,477	8,639	21,093	21.34%	52.11%
Mammary gland luminal 1	30,914	7,964	10,424	25.76%	33.72%
Mammary gland luminal 2	28,216	8,229	10,904	29.16%	38.64%
Spermatogonia m7	24,042	8,018	1,487	33.35%	6.19%
Large Intestine A 62816	17,220	6,262	4,589	36.36%	26.65%

*Table 2.7: Numbers of **mouse** AF (all fragment) peaks identified in all of the analyzed datasets with counts and proportions of peaks that have been classified as 'common' or 'tissue-type specific' based on comparisons with replicates and other tissues.*

datasets show much similarity to each other, probably owing to the low numbers of peaks called in two out of three samples (Table 2.6). All of the H5 datasets correlate best with each other ($0.85 > R^2 > 0.75$), likely due to cells having come from the same patient (even though some of those cells were of different morphologies).

Source tissue	Total SF peak number	Common SF peaks	Specific SF peaks	Common	Specific
Adrenal gland	156,521	31,198	17,084	19.93%	10.91%
Thyroid gland	145,020	30,188	26,610	20.82%	18.35%
Colon	120,234	28,582	16,110	23.77%	13.40%
Sigmoid colon	118,911	28,390	16,069	23.87%	13.51%
Breast epithelium	115,527	26,108	11,238	22.60%	9.73%
Leg muscle	101,187	25,561	17,502	25.26%	17.30%
Pancreas	98,506	25,762	3,170	26.15%	3.22%
Spermatogonia H107	87,982	22,532	16,064	25.61%	18.26%
Spermatogonia H5.14	44,037	16,412	5,548	37.27%	12.60%
Spermatogonia (Guo et al)	42,141	16,668	5,140	39.55%	12.20%
Spermatogonia H5.25	37,902	15,421	5,940	40.69%	15.67%

*Table 2.8: Numbers of **human** "Short Fragment" (SF) peaks (short fragment peaks) identified in each of the tissue types with counts and proportions of peaks that fall into 'common' or 'tissue-type specific' category based on intersections with AF peaks (all fragment peaks) of the corresponding category.*

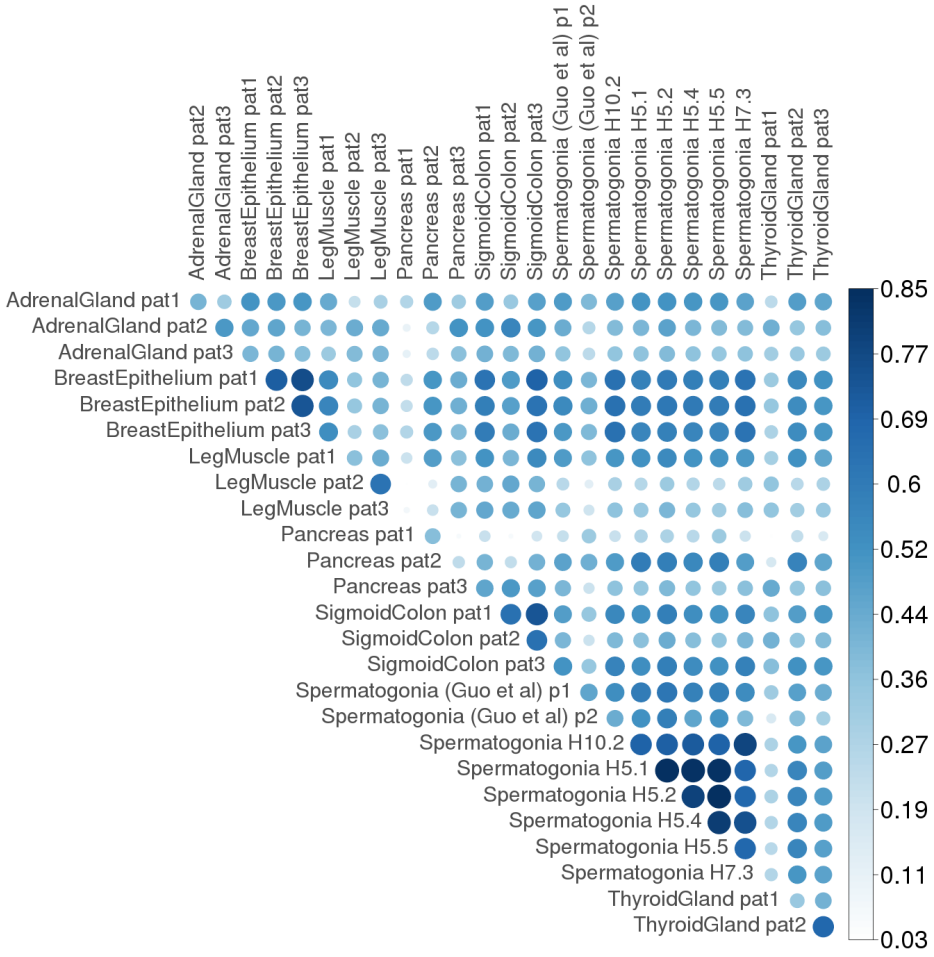


*Figure 2.15: Matrix representing R^2 scores from linear regression of the **human** AF peaks called from each of the individual datasets (only technical replicates combined, not biological replicates). Only scores for peaks that were detected in both compared datasets were used.*

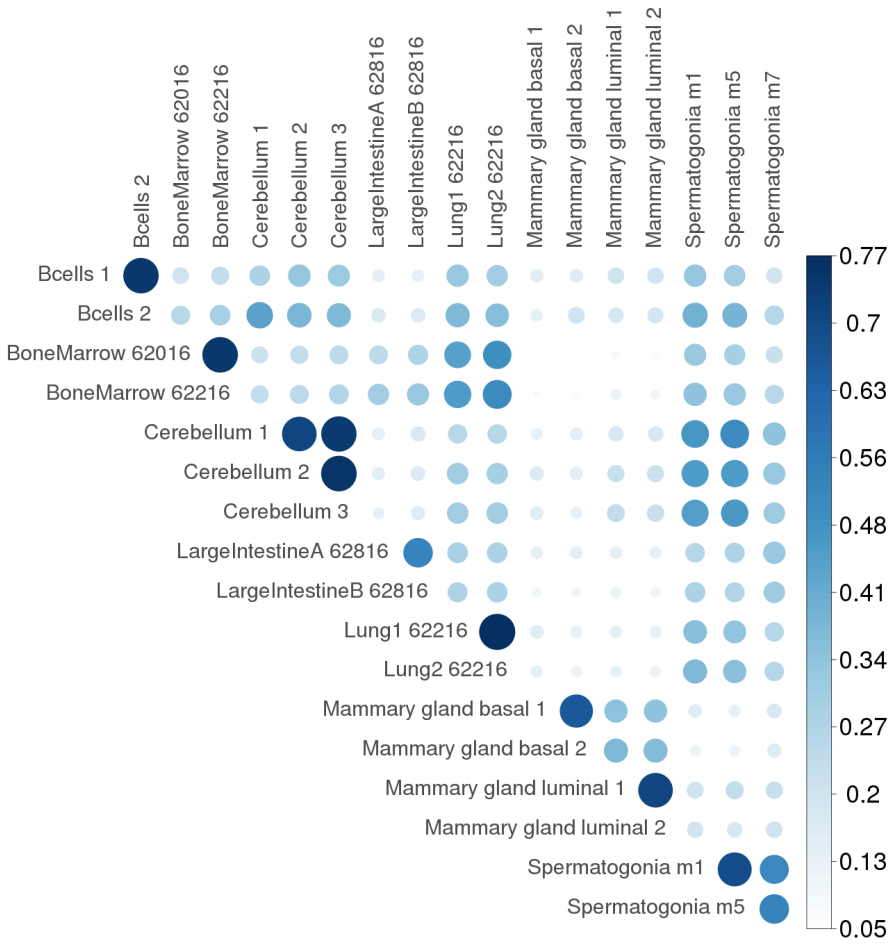


Figure 2.16: Matrix representing R^2 scores from linear regression of the ***mouse*** *AF peaks called from each of the individual datasets (only technical replicates combined, not biological replicates). Only scores for peaks that were detected in both compared datasets were used.*

Source tissue	Total SF peak number	Common SF peaks	Specific SF peaks	Common	Specific
Mammary Gland basal	162,550	21,351	46,245	13.14%	28.45%
Cerebellum	159,879	18,787	40,678	11.75%	25.44%
Mammary Gland luminal	146,245	19,222	20,811	13.14%	14.23%
B cells	126,413	18,391	40,717	14.55%	32.21%
Lung	113,355	17,882	51,444	15.78%	45.38%
Spermatogonia	112,460	17,919	6,664	15.93%	5.93%
Large Intestine	73,948	16,231	21,684	21.95%	29.32%
Bone Marrow	71,937	15,958	19,347	22.18%	26.89%

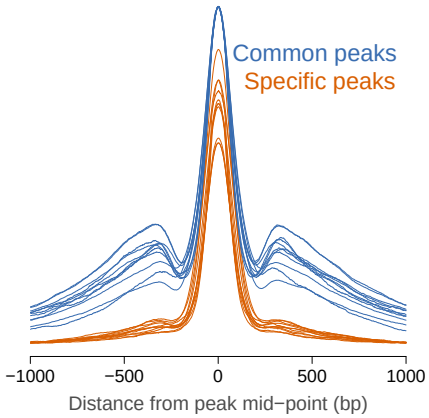
Table 2.9: Numbers of ***mouse*** *SF (short fragment) peaks identified in each of the tissue types with counts and proportions of peaks that fall into 'common' or 'tissue-type specific' category based on intersections with AF peaks (all fragment peaks) of the corresponding category.*

2.3.5 *'Common'* peaks are more proximal to transcription start sites

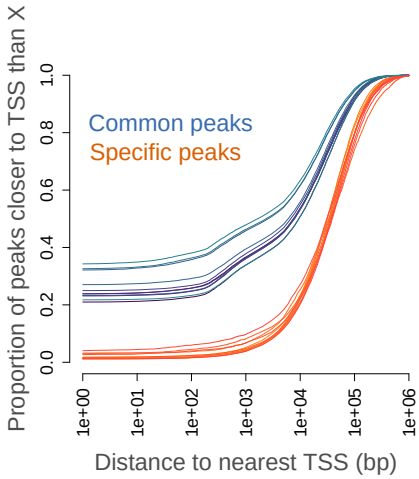
Numbers of SF and AF peaks defined as *'tissue-specific'* or *'common'* can be seen in the Tables 2.6 and 2.8 (human); Tables 2.7 and 2.9 (mouse). While I used the scores from MACS2 output to match the *'tissue-specific'* and *'common'* peaks, those pairs might represent sites that are located in distinct genomic contexts, such as with different proximities to other open regions or genes. Scores output by MACS2 `callpeak` command represent the measure of fold enrichment of the fragment coverage in the peak area relative to that observed in local region (5-10Kb). Thus, the score of the peak that is located in a generally more open region (such as at promoters) could be assigned a lower score than a peak of the same magnitude, but located in a region that is generally less accessible. Figure 2.17a illustrates a difference in absolute fragment coverage between the same number of *'tissue-specific'* and *'common'* sites in the same dataset. *'Common'* peaks have what looks like a 'dip' next to the peaks that are being compared (which probably marks the +1 nucleosome) in general more open regions with additional clustering of adjacent signals. Total number of fragments at the summit is also higher for *'common'* peaks. This suggests that more *'common'* peaks might be located at promoter regions, while *'tissue-specific'* peaks are more distant, stand-alone binding sites. Indeed, when measuring the distance from a peak to the nearest TSS, in all the datasets examined, *'common'* peaks seem to be located markedly closer to TSSs than the *'tissue-specific'* ones (Figure 2.17b). Figure 2.17c shows that 40%-70% of all the *'common'* peaks defined in all the tissue types are overlapping transcription start sites (RefSeq-defined), while less than 10% do so in the *'tissue-specific'* peak category.

This is an important observation to keep in mind when comparing those two

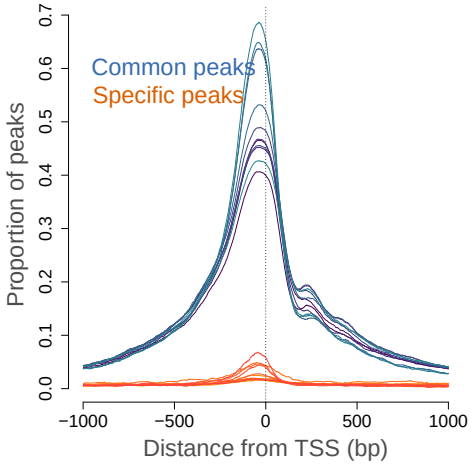
Figure 2.17: Differences in distributions and scores of the 'common' and 'tissue-specific' peaks. Each individual line represent data from one SF peak dataset. In 2.17a, for each 'common' and 'tissue-specific' pair the maximal point of the plot is the maximal value in that pair, hence the y-axis would differ between pair sets.



(a) Absolute coverage of fragments in sets of 'common' and 'tissue-specific' peaks that have been matched by peak scores.



(b) Distance to the nearest Ref-seq transcription start site. Note log-scale x-axis.

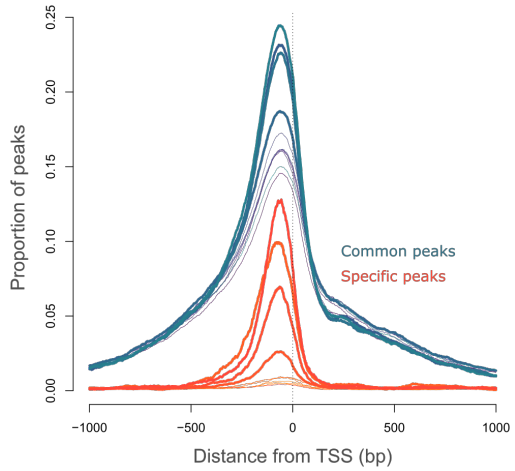


(c) 'Common' peaks are enriched around Ref-seq transcription start sites in comparison to the 'tissue-specific' peaks.

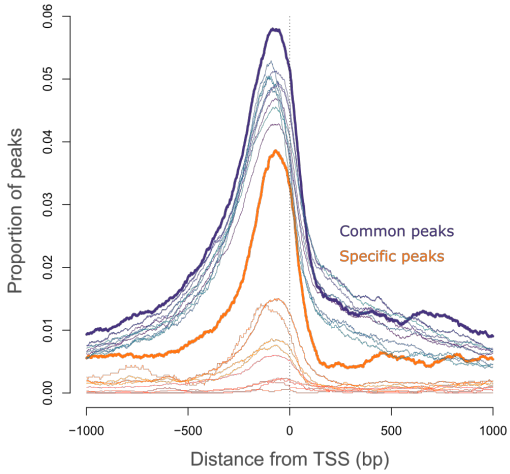
categories of sites, as we are looking at peaks of slightly different magnitude, and therefore possibly varied binding frequency/strength, and also diverse types of sites located in distinct genomic contexts.

2.3.6 *'Tissue-specific peaks'* are enriched over tissue-biased promoters

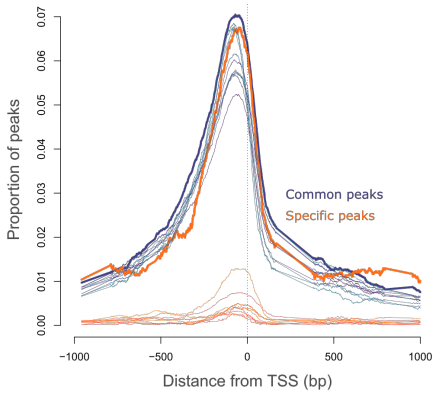
To see if the distribution of *'common'* and *'tissue-specific'* peaks differs between cell types, I looked at the enrichment of each of the peak categories over promoter regions that have been classified as tissue-biased using CAGE data of 52 tissues from Young et al. (2015). Figure 2.18 shows what proportion of peaks overlaps with the testis-biased (2.18a), skeletal muscle-biased (2.18b) and pancreas-biased (2.18c) promoters. While *'common'* peaks from all of the datasets show enrichment over the tissue-biased promoter regions, when looking at the *'tissue-specific'* peaks, they show greater enrichment over the matched tissue type-biased promoter.



(a) Testis-biased promoters and spermatogonia peaks. Spermatogonia-specific peaks (bold orange lines) are enriched at the testis-biased promoters in contrast to other peaks specific to other tissues (thin orange lines).



(b) Skeletal muscle-biased promoters and leg muscle peaks. Leg muscle-specific peaks (bold orange line) are enriched at the skeletal muscle-biased promoters to a larger extent than peaks specific to other tissues (thin orange lines).



(c) Pancreas-biased promoters and pancreas peaks. Pancreas-specific peaks (bold orange lines) are enriched at the pancreas-biased promoters to a larger extent than peaks specific to other tissues (thin orange lines).

Figure 2.18: Enrichment of common and tissue-type specific peaks over the tissue-biased promoter regions. Tissue-biased promoter regions have been defined by the enrichment of the CAGE-tags in different tissues (Young et al., 2015). Thin lines represent peak enrichment for individual tissues, with the bold lines showing enrichment of peaks from the tissue that matches the tissue to which the promoter is biased. Promoters have been oriented in a way so that the gene would be located on the right side of the plot. Enrichment of the tissue-specific peaks at the promoter regions biased to a matched tissue type indicates successful identification and classification of cell-specific regulatory elements.

2.3.7 Novel method identifies edges of protein binding sites

Using the alternative method that I have developed and implemented to find the single-nucleotide-precision edges of the bound proteins (see description of the FLOP method in Subsection 2.2.8), I have identified 400,000-2,000,000 genomic positions in human data that could be classified to represent either right or left boundary of an interacting protein (see Table 2.10 for the total numbers for each of the datasets). For each of the tissue types, I intersected the identified edges with the *'tissue-specific'* and *'common'* peaks described earlier in this Chapter (Table 2.8).

To assess whether the separation of the edges with the FLOP method is effective, I plotted the distribution of edges around a number of motifs of the sequence-specific TFs found under ChIP-seq peaks in the matched cell type (Figure 2.19). Classification of the edges by FLOP method appears to work effectively, with most of the edges falling at the expected side relative to the motif. Different TFs appear to show distinct patterns of the edges around them. For example, the pattern around CTCF is reminiscent of the protein bound in promoter-distal regions (Chen et al., 2012a), with few other edges found in the vicinity; whereas CREB1 and ZNF143, which are known to bind around promoter regions (Conkright et al., 2003; Bailey et al., 2015), show a noisier patterns that are consistent with more TFs being bound in the immediate proximity.

The advantage of this method is that more precise boundaries can be identified, but it is also more prone to identification of false-positive binding sites due to the more sporadic nature of the insertion peak incidence, as reflected by a large numbers of identified peaks. Low levels of 'smoothing' effect derived from the extension of single-nucleotide insertion positions by 9 bp act as a trade-off for the higher degree of precision and effectively leads to more frequent formation of smaller peaks and higher degree of sensitivity to the contribution of Tn5 enzyme preferential insertion bias. While some papers have discussed the Tn5 preferential insertion bias phenomenon in ATAC-seq data (Buenrostro et al., 2013; Madrigal, 2015) and there have been some studies that tried to quantify and account for this bias when analysing the data (Wang et al., 2017; Martins et al., 2018), to date there is no method or software that has been successful at completely eliminating it. Looking at the patterns of Tn5 insertion around binding sites of some of the sequence-specific DNA-interacting proteins, TFs would be expected

Source tissue	Total edges	Common edges	Specific edges	Common	Specific
Spermatogonia H5.25	2,525,015	77,339	11,909	3.06%	0.47%
Spermatogonia H5.14	1,546,774	79,348	18,671	5.13%	1.21%
Spermatogonia H107	1,441,764	104,326	66,330	7.24%	4.60%
SigmoidColon	1,200,572	123,158	79,515	10.26%	6.62%
AdrenalGland	1,031,095	147,267	86,178	14.28%	8.36%
ThyroidGland	946,665	152,324	134,246	16.09%	14.18%
LegMuscle	728,823	129,508	91,465	17.77%	12.55%
BreastEpithelium	661,451	120,665	47,202	18.24%	7.14%
Colon	623,907	124,398	65,954	19.94%	10.57%
Pancreas	608,928	123,099	17,318	20.22%	2.84%
Spermatogonia (Guo et al)	413,807	83,470	14,276	20.17%	3.45%

Table 2.10: Numbers of increased Tn5 insertion frequency peaks identified that represent protein binding edges in each of the human tissue types with counts and proportions of those that have been classified as 'common' or 'tissue-type specific' based on intersections with SF peaks (short fragment peaks) of corresponding category (Table 2.8).

to be affected to a varied degree. In particular, a TF REST appears to have a binding motif that closely resembles the Tn5 preferential insertion sequence. Therefore, binding sites for some of the TFs might not be effectively captured with this method and a relatively high false positive rate in overall binding edge identification is anticipated. However, by combining the FLOP method with intersection of binding edges with different categories of peaks described in the previous subsections, this can be mitigated.

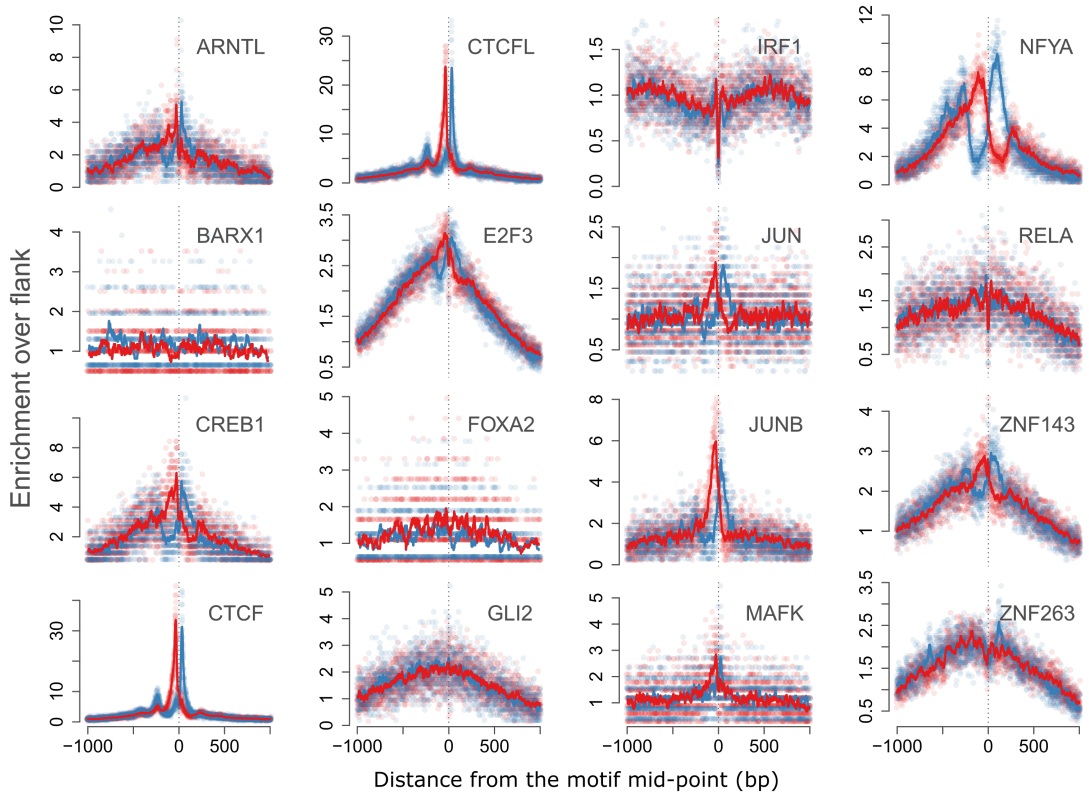


Figure 2.19: Distribution of the putative protein binding edges (identified by FLOP method from sigmoid colon ATAC-seq data). Edges that were classified as marking right side of the bound protein are in blue, while those marking left side are in red. Individual plots are centered on the midpoint of aggregate TF binding motifs found under the ChIP-seq peaks (LoVo cells, colon adenocarcinoma; from Yan et al. (2013)). Points represent an enrichment of the number of edges at each of the position relative to the average of -1000:-800 and +800:+1000 regions away from the motif mid-point, while the lines represent a rolling average of 20bp. FLOP method edge separation is working well, as evidenced by an enrichment of right-classified edges to the right of the putative binding sites, and left-classified edges to the left around most of the TF motifs.

2.4 Discussion

2.4.1 Novel chromatin accessibility primary data for spermatogonial cells

In order to be able to compare the differential mutational pressures acting at the protein binding sites of the somatic and germline cells, it is necessary to have aggregate information about the protein-binding landscape of multiple tissues. There is already abundant information from different types of assays such as DNase-seq, ChIP-seq/exo and its variations, and ATAC-seq available through large data collections such as ENCODE (Dunham et al., 2012) and Roadmap Epigenomics (Roadmap Epigenomics Consortium et al., 2015) Consortia as well as other individual studies. While some of those methods, such as ChIP-exo, are ideal for high-confidence and high-resolution binding site identification, they also require a high number of cells as input material, along with protein-specific antibodies, and performing those assays on large numbers of TFs in multiple tissues is labour-intensive. One would also need sufficiently good knowledge of the types of TFs that are likely to be bound in different cells. Therefore, there is a rather limited range of proteins for which those assays have been performed, let alone in multiple types of tissues or looking at tissue-specific TFs. For the type of analysis proposed here we require an aggregate of a relatively large number of tissue-specific binding sites. In this respect, assays such as DNase-seq and ATAC-seq are ideal, as they can provide a 'snapshot' of the accessibility landscape of the whole cell in one experiment, and the whole protein-binding landscape can be inferred from that.

We propose that large numbers of mutations occur at germline active protein-binding sites as a result of TF interference with normal processes of replication and repair. Therefore, we expect the highest numbers of mutations to accumulate at the protein-binding sites that are active in the most replicative cells of the germline, which are spermatogonial cells. In this work, we have generated novel, primary ATAC-seq data obtained from the FAC-sorted spermatogonial cell populations from both mouse and human testicular tissues. We used the FGFR3 surface marker to enrich for human spermatogonial cells. At the time of our first cell separation experiments FGFR3 has not been utilized for spermatogonial cell enrichment. Since then it has been confirmed

to be suitable for this purpose by another study that used the FGFR3 antibody to perform magnetic cell isolation on cell populations obtained from fragments of testis tissue (Von Kopylow et al., 2016).

To confirm that we are working with the right cell type, I looked at the promoters of genes that we expect to be expressed in human spermatogonial cells based on RNA-seq evidence from Guo et al. (2017). In most of our FAC-sorted cells there was some degree of openness at promoter regions of germ-specific genes *DAZL* and *DDX4*, which is encouraging. Promoter regions for most of the pluripotency TFs that have been confirmed to be expressed in both unipotent and differentiating spermatogonia (Guo et al., 2017) show a high level of accessibility in most of the samples. Surprisingly, the promoter region of *FGFR3*, which is the marker that was used for cell sorting does not show evidence of being accessible in our datasets. While *FGFR3* promoter does show some evidence of openness in some of the other (somatic) datasets, the chromatin state of the region seems to be rather poorly captured by ATAC-seq based on visual inspection of the fragment pileups. It is fair to note that the open state of a promoter region of a gene does not necessarily imply gene expression, and transcript abundance (hence, protein abundance) is affected by other factors such as post-transcriptional regulation.

I assert that the number of replicates of our primary data are of good quality and sufficient depth to use for protein-binding site identification. There were also 8 samples whereby the quality did not appear to be suitable for analysis. Those datasets were exhibiting a high level of 'noisiness', which could conceivably be a reflection of the perturbed chromatin state of cells due to the stress of desegregation procedures and FACS or, alternatively, as consequence of the over-digestion with the Tn5 enzyme. While some of the datasets appeared to show a better result after an adjustment of the Tn5 amounts to the numbers of cells, others were still exceedingly 'noisy'. This could be due to the fact that the alteration of the Tn5 amounts was approximate, or to widespread rupture of the nuclear membrane in the sample and fragmentation of genomic DNA with loss of native chromatin conformation. While the exact nature of this phenomenon currently remains unclear, I provide some measures that could help identify such datasets (in ways distinct from the visual inspection of the coverages), by quantifying overlap of fragments with DNase-seq footprints. When narrowing analysis to the good quality datasets, the fragment coverage distributions shows that our data

is suitable for identification of promoter regions as well as protein binding sites.

Overall, these data provide a valuable resource for exploration of the spermatogonial cells outlined in this work, or any of the future studies that might be undertaken to further investigate the chromatin landscape of spermatogonial cells in both humans and mice.

2.4.2 Identification of separate categories of protein-binding sites

In this work, I have applied the same analysis to data from one type of assay (ATAC-seq) performed on a number of different tissue types. I used some publicly available raw data from ATAC-seq performed on various human and mouse somatic tissues from GTEX and other studies; SSCs from Guo et al. (2017); and our in-house generated human and mouse spermatogonial cell ATAC-seq. I developed a data-processing pipeline for identification of protein-binding sites in each of the analysed type of tissue, and also separate categories of sites that are either *'tissue-specific'* or *'common'* across all of the tissues. I devised a method that identifies potential edges of the bound proteins at near-nucleotide resolution and assigns which side the edge represents.

I applied the aforementioned analyses to spermatogonial cells, and human tissues of pancreas, colon, thyroid and adrenal glands, breast epithelium and skeletal muscle; along with mouse spermatogonial cells, B cells, cerebellum tissue, bone marrow, lung, intestine and mammary glands. I identified a number of sites that I conclude to be either specific to each of those tissues, or common to all. I find that *'common'* and *'tissue-specific'* sites exemplify slightly distinct categories of regions, with the former potentially being more representative of sites located at promoters, while the latter showing patterns more reminiscent of 'stand-alone' protein binding regions, *e.g.* enhancers. This might be a reflection of the fact that sites such as enhancers differ more between tissues than promoters do (Villar et al., 2015). For further analysis of those sites, I propose using pairs of sets of *'common'* and *'tissue-specific'* sites with matched distributions of peak scores, which here I use as a proxy for the strength or frequency of the protein binding.

One of the shortcomings relating to the use of ATAC-seq for nucleotide-resolution protein-binding site identification is the phenomenon of the Tn5 preferential insertion bias. Tn5 has previously been reported to have a preferential sequence that is enriched around its insertion sites, a feature common to greater or lesser degrees

to all DNA cleavage methods (Martins et al., 2018). Lately, there have been multiple attempts to characterize and account for this bias (Wang et al., 2017; Martins et al., 2018), however none of those are able to fully eliminate it. I was not successful at removing the bias either by my own methods or by utilizing recently available software (Martins et al., 2018). While Tn5 preferential insertion sequence being present at the aggregated edges of identified binding sites might distort the pattern of mutation and variation we are looking at, due to the uniform nature of the bias, it is unlikely to affect the comparisons of the sites between different tissues.

There have been a number of methods developed previously for near nucleotide-resolution protein-binding site identification from the chromatin accessibility data, such as footprinting (reviewed in Gusmao et al. (2016)), mainly for DNase-seq data analysis, less so ATAC-seq specific methods. Most of those methods still rely on pre-defined motifs of the TFs that one would expect to find. Here, our primary goal is to identify protein binding sites specific to the poorly-characterised population of spermatogonial cells. Therefore we expect that there might be a number of TFs for which defined motifs are not available. The 'footprints' detection method as such has also been shown to be widely affected by the ragged digestion/insertion patterns within the binding motif, speculated to result from the DNA shape (He et al., 2014; Sung et al., 2014) or protein binding conformation (Neph et al., 2012), with estimated 80% of TFs being affected by this (Baek et al., 2017). The advantage of the FLOP method described here is that it does not rely on finding regions of insertion depletion, but rather an increase in insertion frequency in accessible DNA proximal to bound TFs. It is also built on the ATAC-seq data specific observation about the distributions of the fragments around the binding sites. This has potential in identification of the binding edges and their spatial classification, but could be improved on. One possible line of improvement would be to integrate this method together with machine-learning approaches, such as the *Segway* software designed for the identification of epigenetic states (Roberts et al., 2016). One can envisage an implementation of Hidden Markov or dynamic Bayesian network (as used in *Segway*) models with several observations, such as insertion frequencies within the sample and control (*e.g.* 'naked' DNA) on either strand, counts of short and long fragments overlapping each genomic position, and DNA sequence. One could define a number of chromatin states, such as 'closed' (1), 'open accessible' (2), 'open nucleosome-occupied' (3), 'open left side of TF' (4), 'TF occupied' (5) , 'open

right side of TF' (6), and some state transition probabilities. Knowing that along the DNA sequence states can change in 1→ 2→ 3→ 4→ 5→ 6→ 3→ 2→ 1 manner, but not in 1→ 4→ 3 sequence, the model could be trained on a set of known binding sites in a well-studied cell type or line, where both ATAC-seq and ChIP-seq or equivalent are available.

CHAPTER 3

Elevated germline mutation rates at protein-binding sites

3.1 Introduction

3.1.1 Patterns of mutation rate and selection are intermixed, and not uniform across the genome

Differences in DNA sequences between species (*e.g.* human and mouse), or between individuals of the same species, originate as mutations that have occurred in a cell of germline lineage. The frequency with which mutations occur is defined as the *mutation rate*. The mutation rate is not the only thing shaping the pattern of germline variation that we observe. Some mutations that occur might be deleterious for the organism, or even incompatible with life. Therefore, selection shapes the landscape of the genome on top of the mutational pressures in observed nucleotide diversity data. In a simplistic view, the absence of variation is often taken as evidence of purifying selection at functionally important underlying locus, while high variability is often interpreted as a lack of function. Increased sequence variability can also be evidence of diversifying selection, where a change in sequence if favoured and its retention is driven by natural selection and adaptation (Gittelman et al., 2015).

This, however, is complicated by the fact that mutation rates are not uniform, but vary greatly across the genome on different scales (Wolfe et al., 1989; Makova and Hardison, 2015). In order to be able to determine whether sequence is significantly conserved to a higher or lesser degree than expected, one needs to have an estimate of the nucleotide conservation in the absence of selection, under "neutrality". Due to the variability in the mutation rates across the genome, the estimation of conservation under neutrality would differ accordingly. Non-uniformity of mutation rates gives rise to the notion of *regional mutation rates*. A variety of different factors have been implicated as having an effect on the regional mutation rate. Those factors can affect different types of mutations in various ways - for example there can be an inverse reverse relationship between the rates of single nucleotide variants (SNVs) and indels (Semple and Taylor, 2009). In this work, I look exclusively on SNVs.

3.1.2 Regional mutations rates vary at different scales

Each genomic region is affected by multiple categories of features that determine its mutation rate. Replication timing is one of the large-scale correlates of the variability in mutation rates, with more mutations occurring in late-replicating regions, which is proposed to be due to depletion of free nucleotides (Watt et al., 2015), accumulation of single-stranded DNA (Stamatoyannopoulos et al., 2009) and lower level of mismatch repair (Supek and Lehner, 2015).

Sequence composition can affect the rate at which mutations occur. CpGs exhibit an elevated mutation rate (with a distinct high frequency of C$\longrightarrow$T change) due to the high incidence of methylated cytosine deamination (Youssoufian et al., 1986; Hodgkinson and Eyre-Walker, 2011). In turn, most regulatory regions, such as promoters, in the human genome tend to be GC-rich (Gardiner-Garden and Frommer, 1987) and unmethylated (Thurman et al., 2012), leading to lower mutation rates than sequences of similar composition but located outside of regulatory regions.

There is a complex association between the mutation rates and transcription. On one hand, transcribed regions tend to spend more time in an open state and be more accessible to mutagens, in addition to a non-template strand spending some time in a sensitive single-stranded conformation (Jinks-Robertson and Bhagwat, 2014). On the other hand, that is counterbalanced by the presence of transcription-coupled repair (Hanawalt and Spivak, 2008). Exons are generally more conserved than introns, but they are also have been shown to exhibit lower mutation rates due to the preferential targeting of mismatch repair (Frigola et al., 2017). This could also be counted as a dependence of mutation rates on histone marks, as H3K36me3, that is enriched in exons of actively transcribed genes, is hypothesised to be responsible for recruitment of MMR (Frigola et al., 2017). DNA accessibility shows a positive correlation with the incidence of SNVs. At a low resolution (several kilobases), promoter regions, important for binding transcription machinery and other TFs, have been found to exhibit increased mutation rates (Young et al., 2015; Taylor et al., 2008, 2006), while there is more heterogeneity at higher resolution around individual protein binding sites. The way that certain bits of DNA are packaged can influence the likelihood of them being mutated - there is a higher level of substitutions observed at nucleosomal dyads when compared to linker sequences (Sasaki et al., 2009; Semple and Taylor, 2009; Tolstorukov et al.,

2011).

3.1.3 Increased germline variation around protein binding sites can be shaped by variable mutation rate or selection

Increased germline variation is observed surrounding some sequence-specific TF binding sites, that rises above the estimated level of neutrality (Figure 1.8). There are several conceivable explanations for how this pattern has occurred, represented in Figure 3.1. The *'diversifying selection'* model explains the observed pattern purely as a reflection of selection, with diversifying selection favouring variation near TF binding motifs (Figure 3.1, left). The '*reduced constraint*' model similarly explains the observed pattern of single nucleotide substitutions between species as shaped by selection (Figure 3.1, right). This model assumes that the increased numbers of substitutions in fact matches the level of neutrality (dashed green line), and that the proximal sites are under some selective constraint due to a higher probability of other binding motifs occurring in the vicinity (Boyle et al., 2011). The '*increased mutation rate*' model describes the observed shape of substitutions as a composite imprint of patterns laid down by both purifying selection that is acting to preserve a functionally important protein-binding motif, and mutation rate, that is increased all across the binding site spanning the motif and the flanking sequences that are also physically occupied by protein.

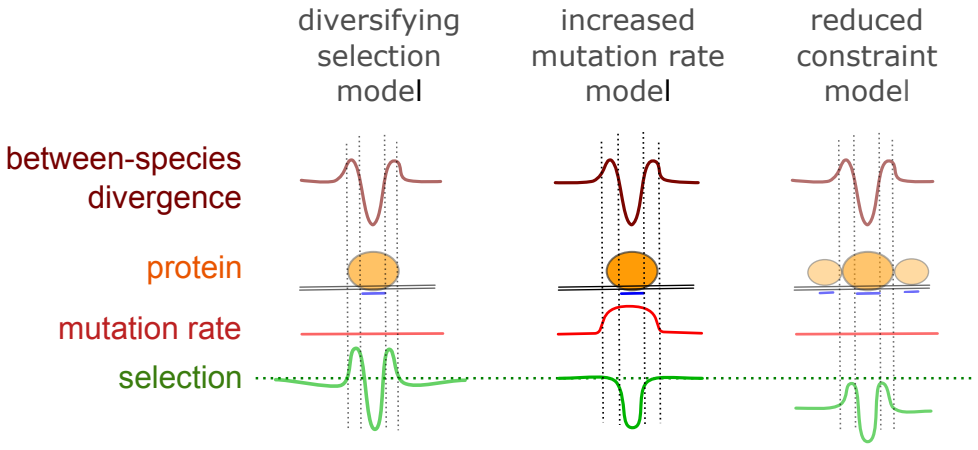


Figure 3.1: Possible explanations for the observed pattern on increased divergence near TF-binding motifs (burgundy lines on the top). Green "selection" curves at the bottom show the influence of selection pressure on nucleotide diversity, with dotted lines denoting level of neutrality. For detailed explanation of all models see Section 3.1

As with most statistical analyses, the shape of these substitution rate profiles

is more accurately measured with more data. Many sequence changes are observed between species, for example between human and mouse there are typically 0.57 nucleotide substitutions per site in four-fold degenerate positions of protein coding genes (Forrest et al., 2014), where as there 0.002 substitutions per site in the same four-fold degenerate sites between any two distantly related humans (Gibbs et al., 2015). In a simple rate-comparison analysis one would need alignments of 285 human sequences to equal the statistical power of one human to mouse pairwise alignment. However, not all sequences do align between human and mouse, and biology has diverged such that many protein binding sites in human are not present at orthologous positions in mouse (Schmidt et al., 2010), problems that are exacerbated by comparisons over greater evolutionary distance. It is also the case that selection pressures and mutational processes may have shifted over large evolutionary times. For these reasons, the calculation of substitution rate profiles on within species variation is a useful complement to between-species analysis. There is another major advantage to utilising within species variation to investigate patterns in substitution rate profiles. That is the leveraging of ancestral state and allele frequency information to deconvolve the contributions of mutation from selection - the central theme of this chapter.

3.1.4 Selection can be inferred through derived allele frequency distribution

It is possible to infer the selective pressures through analysis of the frequency distribution of the derived alleles in a population (Fay et al., 2001). Derived alleles initiate as mutations, which result in differences between ancestral sequence and the one that is present in derived state (Figure 1.6). The frequency of the derived allele would then depend only on subsequent evolutionary pressures acting on the site. Alleles that have a deleterious effect on organism fitness result in lower frequency within a population. In contrast, alleles that confer an advantage to the organisms fitness would increase in frequency up to the point where they become fixed. Alleles that have no effect on the fitness of an organism would only by subject to the stochastic sampling of genetic drift and neutral evolution. In the case of neutral evolution, most derived alleles are expected to be rare, assuming a relatively large population size (under neutral evolution, the probability of an allele to become fixed is inversely proportional to the population size) (Kimura, 1968, 1991). In case of purifying selection acting upon the category of

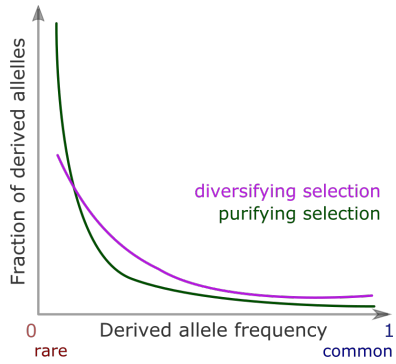


Figure 3.2: Schematic of derived allele distributions under different selectional pressures. Purple line shows expected derived allele frequency distribution under diversifying selection, while green one under purifying.

sites being studied, the frequency distribution of those derived alleles would be expected to be shifted to the left even rarer (Figure 3.2). Conversely, diversifying selection would favour variation and drive alleles to shift to the right, moving into the *common* category leading them towards fixation. By performing a derived allele frequency (DAF) test - comparing the ratios of the *rare*/*common* alleles at regions of interest (*e.g.* protein-binding sites) with the ratio of *rare*/*common* alleles in regions that are thought to be neutrally evolving, it is possible to infer the predominant selection pressure acting on categories of sites selected for analysis. Such analyses are usually performed relative to a reference category of sites that is expected *a priori* to be under no or minimal selection pressure.

3.1.5 *De novo* mutations are a most direct way of measuring mutation rate

The most direct and practical way of measuring mutational pressures is to look at the accumulation of *de novo* mutations (Veltman and Brunner, 2012). *De novo* mutations are variants (alleles) that are present in all of the cells in the offspring, but were not inherited by either of the parents, *i.e.* they must have arisen during the lifetime of the parental germ cell (in some cases those mutations might represent low-level parental mosaicism, or have occurred in the zygote during the first few cell divisions) (Acuna-Hidalgo et al., 2015). Detection of *de novo* mutations is possible though sequencing of the family trios, where both parents and one, or occasionally more, proband are sequenced and their genomes compared (Kong et al., 2012). Sequencing of large numbers of families is costly, so the datasets with *de novo* variants are not very large (when comparing the numbers of variants with numbers we get from the comparisons of sequences within populations), with around 70 *de novo* mutations per proband detected

from trio sequencing (Jónsson et al., 2017).

3.1.6 Questions addressed in the current Chapter

In the current Chapter I test the hypothesis that DNA-binding proteins, such as TFs, are causing occurrence of mutations at their binding sites. To this end, I utilize protein-binding sites that I have defined in Chapter 2 to (1.) test if mutation rate is elevated within the binding sites. I aim to (2.) discriminate the selection *versus* mutation models (Figure 3.1). I also aim to (3.) test the model of elevated germline mutation by comparison to somatic-only binding sites defined in Chapter 2.

3.2 Methods

3.2.1 Between-species conservation measures

Genomic Evolutionary Rate Profiling (GERP) provides a per-nucleotide measure of constraint across the genome (Davydov et al., 2010). Neutrality is estimated genome-wide from the number of substitutions observed at fourfold degenerate sites and normalized to zero. By subtracting the number of substitutions at every position of the genome from the number of those expected under neutrality, the resulting scores represent the amount of 'rejected substitutions' (RS). In this manner, positive scores represent constraint (purifying selection), while negative scores represent an excess of substitutions (positive selection). An underlying assumption of GERP is that the mutation rate does not vary across the genome. As previously discussed in Section 3.1, that is an invalid assumption, so with this in mind positive scores could imply purifying selection or reduced mutation rate and negative scores could be either positive selection or elevated mutation rate.

I used GERP ++ software to create an RS score for every position of the multiple species alignment of the human genome (*hg38*) to the genomes of 30 other mammalian species (27 primate) from UCSC (`http://hgdownload.soe.ucsc.edu/goldenPath/hg38/multiz30way/`). Pairwise alignments were generated using `lastz` and then linked into chains using a dynamic programming algorithm (Kent et al., 2003). The best-in-genome pairwise alignments were progressively aligned using `multiz` (Blanchette et al., 2004) to produce multiple alignments. Multiple species alignment files for each of the human chromosomes, along with the file that describes the evolutionary tree of the species for which the alignments were present have been supplied to the `gerpcol` component of the GERP++ software. Mouse RS scores (for *mm9* assembly) were downloaded form UCSC (`http://hgdownload.cse.ucsc.edu/gbdb/mm9/bbi/All_mm9_RS.bw`)

The advantage of using GERP over other available methods, such as PhastCons (Siepel et al., 2005), is its ability to detect the above neutral substitution rates, and to logically deal with alignment gaps, which reduces the assignment of erroneous scores. Detection of above neutral substitution rates is of particular value, as it allows

for detection of elevated mutation rate and positive selection.

3.2.2 Within-species human variation measures

To estimate the evolutionary forces acting upon the region of interest, I used the frequency of the derived alleles in the population. While there have been multiple large-scale studies looking at within-species variation, such as 1K Genomes (Durbin et al., 2010), many of those studies only include variant calls from exome or targeted sequencing. In this work, I am interested in identifying the variation that occurs at protein-binding sites that are mostly located within the non-coding portion of the genome, therefore requiring variants obtained from the whole-genome sequencing.

To measure variation within the human population I used data from the whole-genome sequencing of the isolated Icelandic population from the deCODE cohort (Jónsson et al., 2017). This data contains information about the alleles found within the Icelandic population and their frequency. To resolve the state of the allele (ancestral or derived), I used human ancestral reconstructed sequence based on the 12-way mammalian EPO alignments (version 86) from Ensembl. Variants where the ancestral state could not be resolved, or where the change did not represent a single-nucleotide substitution, were discarded. Variants were then split into *rare* (<1.5%) and *common* (>5%), based on the frequency of the derived allele. The 1.5% and 5%-thresholds were defined by Young et al. (2015) based on maximizing the odds ratio while minimizing the confidence interval for the comparison of (1.) the second codon position that are assumed to be constrained and (2.) fourfold-degenerate sites (which are often used as proxy for neutral evolution) in protein-coding sequence.

3.2.3 Within-species mouse variation measures

While human within-species and within-population variation has been relatively widely investigated, there is much less information available for the mouse within-species or within-strain variation. Here I used data from a study where 10 wild house mice individuals from North-West India have been whole-genome sequenced to a mean depth of $\approx$ 30-fold (Halligan et al., 2013). Since there were few individuals sequenced in this study, I have discriminated between *rare* and *common* alleles by assigning any variants seen once as heterozygous as *rare* and rest as *common*.

3.2.4 Human *de novo* mutations

One of the largest collections of data with the human *de novo* mutations is the MSSNG cohort (C Yuen et al., 2017). MSSNG contains variants inferred from whole-genome sequencing of family trios, where at least one of the probands has been diagnosed with autism, and these data are expected to be enriched in variants that are autism-causing. Therefore, these data are particularly well suited for use in the work described here, as it has been previously reported that numbers of cases of autism in children show a linear increase with paternal age, hypothesising that this is a reflection of the large numbers of divisions that spermatogonial cells undergo before forming mature sperm (Goriely et al., 2013). We, in turn, speculate that at least some of those mutations could be occurring at protein-binding sites.

Currently MSSNG cohort holds whole-genome sequences of 1,740 probands. In total, there were 121,181 *de novo* mutations, averaging at $\approx$ 70 *de novo* mutations per proband. The latest release (acquired December 2017) of data used here has been obtained from the MSSNG portal API and *de novo* mutation coordinates have been taken from the file `Annotated_de_novo_variants.xlsx`. Any variants marked as having failed filters have not been included in the analysis.

Another large source of the *de novo* mutations is the deCODE study (Jónsson et al., 2017) described in the Subsection 3.2.2. The deCODE cohort holds 108,778 *de novo* mutations from whole-genome sequencing of families with 1,548 probands. Some of the variants are phased, which means that it is inferred whether the mutation occurred in the maternal or paternal germline.

Data from a few other smaller-scale studies were included in the analysis described here – 36,441 *de novo* mutations from the Goldmann et al. (2016) study, which is of a particular interest, as it is enriched for older fathers; 11,020 mutations from the Francioli et al. (2015); and 251 mutations from the Wang et al. (2009). All coordinates were converted to correspond to the *hg38* genome assembly using `liftOver` utility from USCS (*v326*) and chain files from USCS database. The final aggregated dataset contained 268,759 mutations.

3.2.5 Mouse *de novo* mutations

While there have been multiple studies generating whole-genome sequencing data for human family trios, there have not been many studies looking to identify *de novo* mutations in mice. Here, I used the data from whole-genome sequencing of more than 20 generations of wild-type C57BL/6 and mutator mice, which have high DNA replication error rates (Uchimura et al., 2015). A total of 7 mice were whole-genome sequenced (2 wild-type and 5 mutator mice) to >40X average coverage, and the final dataset contains 7,007 mutations.

3.2.6 Measuring variation at protein-binding sites

Peaks in each category (from Chapter 2) have been aggregated together and centred either on the peak mid-point (in case of SF peaks, Subsection 2.3.5, Table 2.8) or the identified edge of the bound protein (Subsection 2.3.7 , Table 2.10). In the latter case, corresponding aggregated scores centred on the edges identifies as right have been reversed and added to the left-edge scores.

When looking at human population variation, the proportions of *rare* and *common* variants from deCODE relative to the average over the region excluding 400bp around the centre point were plotted using R (3.3.2), as shown in Figure 3.3. Odds ratios and p-values were calculated using the Fisher exact test for count data (`fisher.test` function in R), as shown in Figure 3.4.

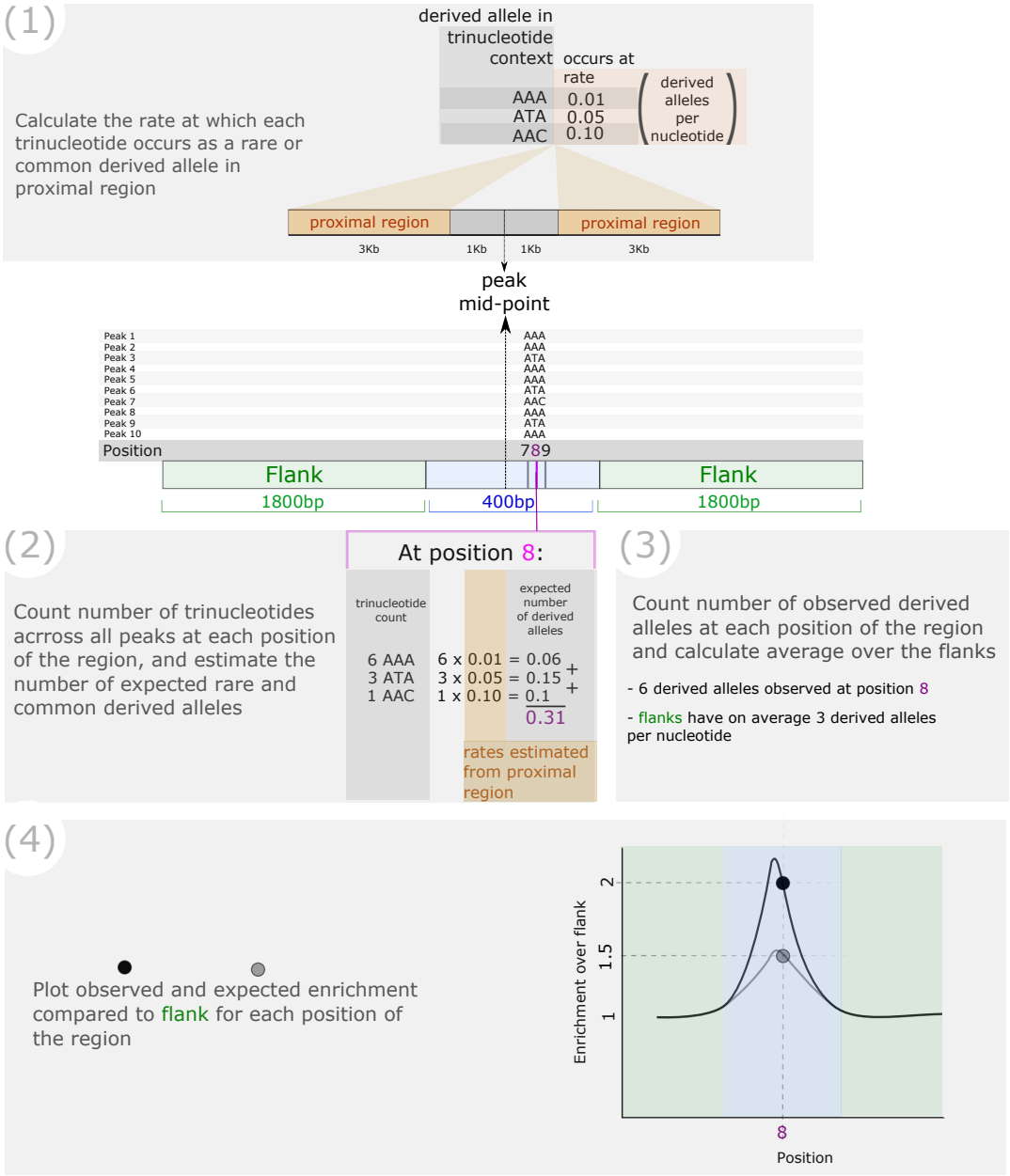


Figure 3.3: Expected numbers of derived alleles over the region were estimated by (1.) first calculating the frequency of common and rare allele occurrence in the context of each trinucleotide within proximal regions on either side of the peak mid-point. Then, (2.) for each position within the region, this calculated frequency of each trinucleotide rare and common derived allele occurrence was multiplied by the number of instances in which the corresponding trinucleotide occurred. Observed (3.) and expected derived alleles (4.) were plotted relative to the average across flanks.

When looking at between-species divergence, the same set of sites as described above were used. Mean GERP scores were calculated for every position of the region,

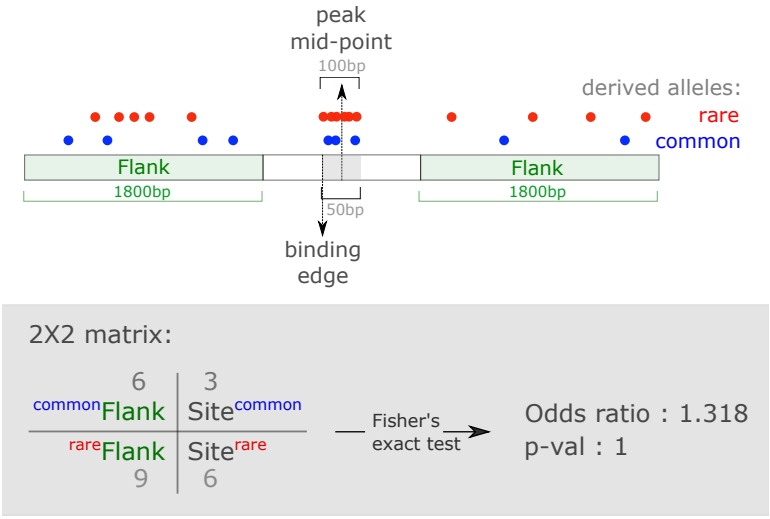


Figure 3.4: Binding sites were defined as either `0bp:+50bp` *window relative to the identified left edge (by FLOP method; Section 2.2.8, Box 2.2), or as* `100bp` *window centred on the peak mid-point. 2x2 contingency matrix containing the counts of the rare and common derived alleles in the binding site and flank was used to perform Fisher's exact test to obtain odds ratio value and associated p-value.*

and plotted with lower values signifying higher divergence going up the y-axis, and higher values, signifying higher constraint, going lower down the y-axis. While counter-intuitive, this upside-down plotting strategy enforces correspondence of the within- and between-species variation plots. Expected GERP scores over the region were estimated by first calculating the mean GERP score for each trinucleotide within the 1kB:4kB flank regions on either side of the mid-point. Similar to within-species variation plots, for each position of the region, this calculated mean was multiplied by the number of instances in which the corresponding nucleotide occurred.

Beanplots with *de novo* mutations were plotted using the `beanplot` R (3.4.1) package. Flanks were defined as regions -1000:-800 and 800:1000bp away from the peak boundary or edge. In cases where the binding site was defined as a peak, its area was taken as a binding site. In the case of edges, binding sites were defined as region 0:100bp relative to the left edge, and -100:0bp relative to the right edge.

Due to spareness of the data, distributions with the numbers of *de novo* mutation per flank/site were bootstrapped, sampling 10,000 times with replacement. Each point in the plotted distribution represents the mean number of mutations in flank/site set from each individual bootstrapped sample. Empirical p-values were calculated as a proportion of the instances where the mean *de novo* rate in the flank matched or was

higher than *de novo* rate in the binding site/peak.

3.3 Results

3.3.1 Germline protein-binding sites are hotspots for functionally consequential mutations

To see if protein-binding sites show an increase of variation in germline cells, I looked at the sites that I defined as 'active' in spermatogonial cells based on the ATAC-seq signal. Figure 3.5 shows the patterns of observed and expected GERP scores and amounts of rare and common derived alleles over aggregate of those sites in humans. GERP scores (top plot) show a high level of constraint around the midpoint of the binding site, as would be expected for the region containing a functionally important binding motif. The middle plot of the Figure 3.5 shows the enrichment of rare and common derived alleles over the region relative to the flanks. Both rare and common derived alleles show accumulation over the binding sites, indicating that there is increased variation all over the region. While the proportion of rare and common derived allele numbers stays relatively the same in the flanks, there is a higher proportion of rare derived alleles compared to common when moving into the binding site (a leftward shift of the DAF, Section 3.1), indicating that there is purifying selection acting all across the region covered by the peak (DAF OR : 1.188; p-val : 2 x 10^{-46}). This is described by the green line in the bottom plot of the Figure 3.5, where the proportion of common derived alleles is used to represent the pattern of selection unaffected by mutation rate.

The region showing the enrichment of derived alleles extends ≈50bp either direction from the defined peak mid-point. While a single bound protein would probably be expected to cover a region that is smaller than this, identified peaks frequently represent clusters of bound proteins, for example, promoter and enhancer regions. The extent of binding site clustering indicated by the span of peak coverage which closely mirrors the distribution of increased polymorphism rate (Figure 3.6).

Figure 3.9 shows similar plots with variation measures over the mouse spermatogonial binding sites. Similarly to the human data, there is a striking increase in rare variants across the binding site.

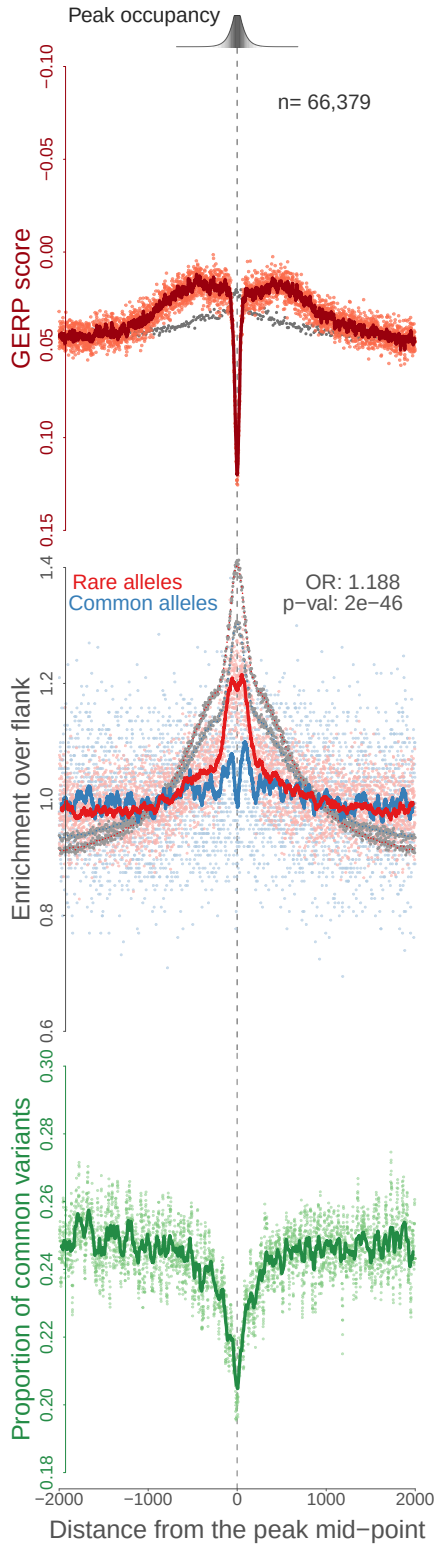


Figure 3.5: Germline variation at aggregate of all identified ***human*** *spermatogonial binding sites (n=66,379). All plots are centered at aggregate peaks midpoint. Individual points represent measures at single nucleotide resolution (y-axis), while lines are plotted through rolling average in 10bp (top plot), or 50bp (middle and bottom plots) sliding window. Grey points and lines represent an expectation from trinucleotide sequence context. Top plot shows sequence between-species divergence as measured by GERP scores. Higher GERP scores (lower down the y-axis) represent sequence constraint. Middle plot shows enrichment of rare and common deCODE derived alleles relative to flanks (see Figure 3.3 for details of how this was plotted). 'OR' denotes odds ratio of the enrichment of the rare derived alleles at the site relative to the flanks, with associated p-value. Bottom plot shows proportion of common derived alleles, representing pattern of selection unaffected by mutation rate, with lower values indicating stronger purifying selection pressure. Protein binding sites explored here show increased numbers of rare derived alleles relative to the putatively neutrally evolving flanking regions which is not driven by diversifying selection.*

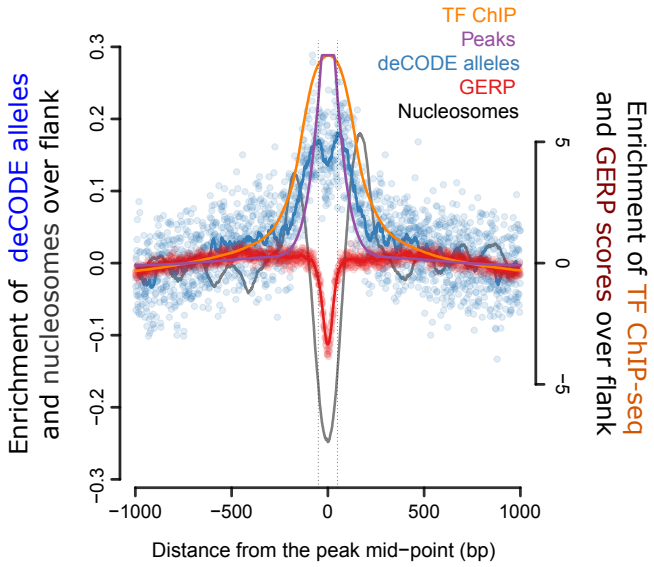


Figure 3.6: Spacial distributions of the peaks (purple), nucleosomes (grey), TF ChIP-seq (orange), enrichment of deCODE alleles (blue), and GERP scores (red). Plot is centered on the aggregate peak midpoint with a putative binding sites in the middle.

3.3.2 Increase in expected pattern of derived alleles over the binding sites is driven by methylated state of CpGs in the flanks

Along with the pattern of observed increased variation over the binding sites, there is also a very distinct pattern of increased *expected* variation over the same region. While this intuitively suggests that the variation that we see is purely a reflection of the sequence composition, I investigated this further.

The regions that we are looking at (ATAC-seq peaks) appear to have a distinctly different sequence composition (with a high frequency of CpG dinucleotides) from what one can observe in the flanks (more AT-rich) (Figure 3.7). In addition to that, most of the sites in question would also be expected to be unmethylated, as has been reported before for the TF binding sites and accessible regulatory regions of the genome (Groudine and Conkin, 1985; Thurman et al., 2012). That is indeed the case, as demonstrated in Figure 3.7. This shows the pattern of methylation measured by whole-genome shotgun bisulfite sequencing in the pancreatic tissue (ENCODE accession ENCSR344YUA) around the aggregate of pancreas-defined peaks, along with the

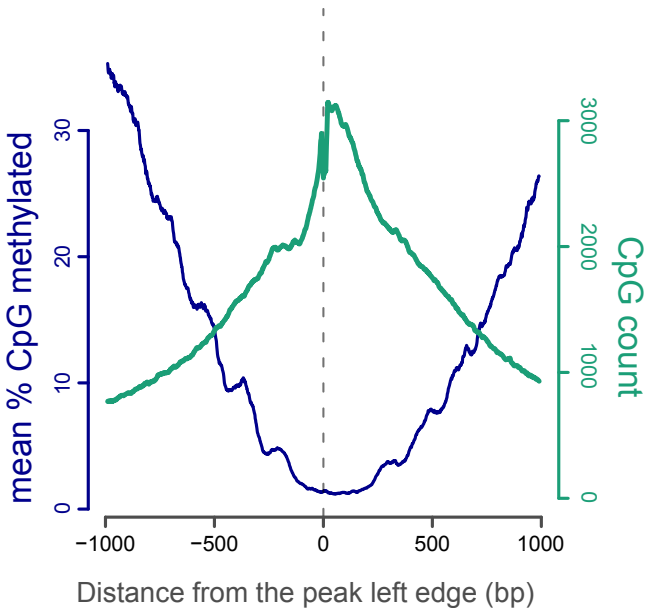


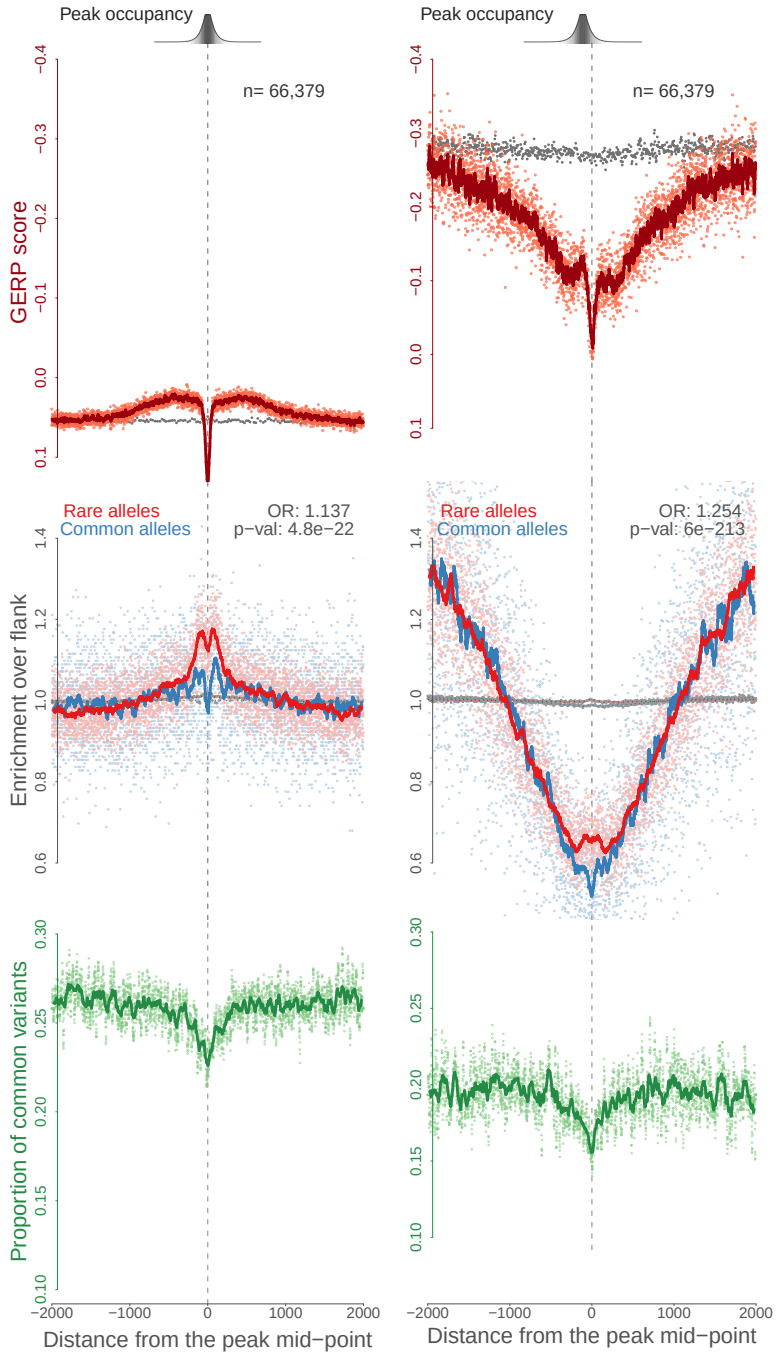
Figure 3.7: CpG counts (cyan) and methylation (blue) around the aggregate set of binding sites. Regions around protein binding sites have higher number of CpGs which are not methylated relative to more distant surrounding sequences.

numbers of CpG dinucleotides. The expected measures of polymorphism frequencies (for Figure 3.5, middle) in each trinucleotide context are taken from the flanks that are 1-4kB away from the binding sites (as described in Subsection 3.2.6). Those flank regions are more likely to have CpGs methylated and therefore would increase the expected frequency of polymorphisms at any CpG-containing trinucleotides over what actually would be expected at the unmethylated binding sites.

To disentangle the contributions of CpG and non-CpG effects, I have produced similar plots, but either excluding all of the alleles in the CpG context, or conversely only leaving the alleles that were found to have occurred in the context of CpG (N$\underline{\text{C}}$G or C$\underline{\text{G}}$N). As the numbers of sites at each position would then vary (due to each of the positions having different numbers of CpG and non-CpG context trinucleotides), I plotted the relative proportions of rates, rather than observed counts (Figures 3.8). In both of those cases the expected rate stays uniform across the region, as it is estimated once again from the flank regions. The pattern of non-CpG polymorphism rate still shows an enrichment in the binding site region (Figure 3.8a). Rates of CpG-context polymorphisms shows a decrease towards the middle (Figure 3.8b), closely reflecting

the methylation pattern (Figure 3.7). In the area of the putative binding sites, however, there is an increase in the rare derived allele frequency, despite the lower methylation status of that region, suggesting that a different mutational process is responsible from this increase, in agreement with our hypothesis.

Through separation of individual contributions of mutation rate and selection, data presented here demonstrate that former, rather than the latter is the predominant force increasing nucleotide diversity at the protein binding sites. Protein binding sites are shown to be under purifying selection pressure, and there is no evidence of diversifying selection driving increase in sequence variation at or near those regions. Germline mutation rate is elevated both adjacent to and within the sequence specific binding site. Thus, germline protein binding sites would be expected to act as hotspots for potentially functionally consequential mutations.



(a) Without CpGs (b) Only CpGs

Figure 3.8: GERP scores (top), rates of common (blue) and rare (red) derived alleles (middle), and pattern of selection (bottom) in the context of trinucleotides that don't contain (a), or contain only (b) CpG dinucleotides over the spermatogonia binding sites. Individual points represent measures at single nucleotide resolution (y-axis), while the lines are plotted through rolling average in 10bp (GERP), or 50bp (deCODE) sliding window. Grey points and lines represent an expectation from trinucleotide context. 'OR' denotes odds ratio of the enrichment of the rare derived alleles at the site relative to the flanks, with associated p-value.

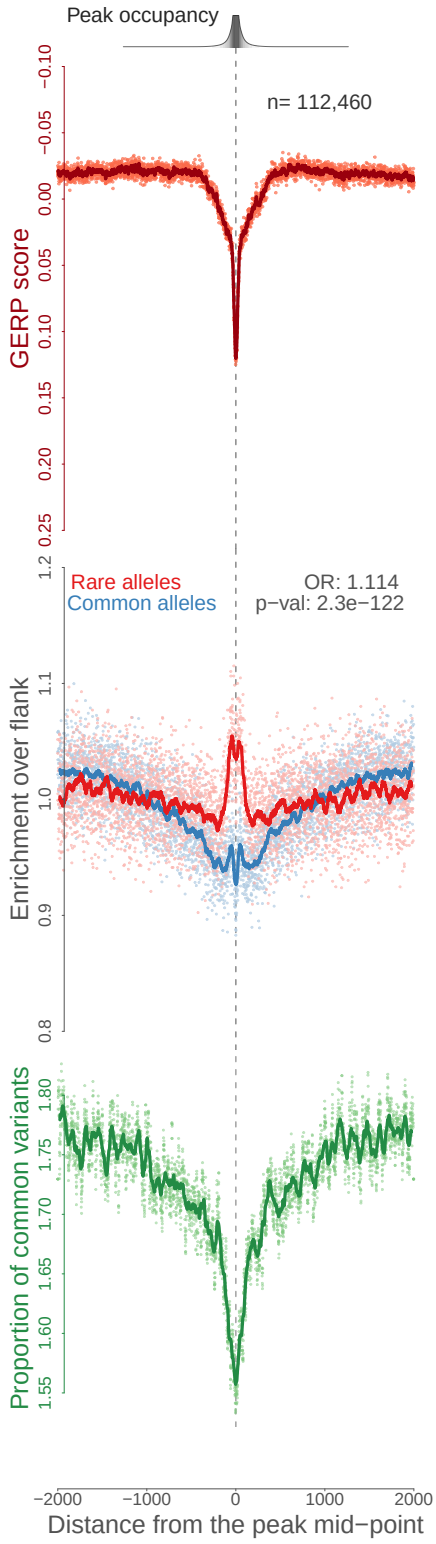


Figure 3.9: Germline variation at aggregate of all identified ***mouse*** *spermatogonial binding sites (n=12,460). All plots are centered at aggregate peaks midpoint. Individual points represent measures at single nucleotide resolution (y-axis), while lines are plotted through rolling average in 10bp (top plot), or 50bp (middle and bottom plots) sliding window. Top plot shows sequence between-species divergence as measured by GERP scores. Higher GERP scores (lower down the y-axis) represent sequence constraint. Middle plot shows enrichment of rare and common wild mice alleles relative to flanks (see Figure 3.3 for details of how this was plotted). 'OR' denotes odds ratio of the enrichment of the rare alleles at the site relative to the flanks, with associated p-value. Bottom plot shows proportion of common alleles, representing pattern of selection unaffected by mutation rate, with lower values indicating stronger purifying selection pressure. Protein binding sites explored here show increased numbers of rare alleles relative to the putatively neutrally evolving flanking regions which is not driven by diversifying selection..*

3.3.3 Protein binding sites active in germline, but not somatic-specific ones show enrichment of germline mutations

My analysis of mutation and selection around spermatogonial binding sites has so far considered all binding sites collectively. However, if protein binding is causally related to the elevated mutation rates as the results thus far suggest, then one would expect to see the elevated mutation rate at ubiquitously bound sites and sites bound in germ lineage cells but not somatic cells. We would not expect to see evidence of elevated mutation at binding sites occupied exclusively in somatic cells. In this section I set out to test these expectations, accepting that I only have partial measurements of binding across all cells of the germ and somatic lineages, so cannot exclude binding site occupancy in some unmeasured cell type.

Chapter 2 describes the separation of the different categories of sites for several tissue types based on the ATAC-seq data. Here, I looked at the germline variation measures over those separated categories of regions. Figure 3.10 shows patterns of between- and within- species variation over the protein-binding sites that have been defined to be active in both human spermatogonial and somatic cells ('housekeeping' or *'common'* sites), those specific to somatic cells, and those specific to spermatogonial cells. Corresponding plots with mouse protein-binding sites are in Figure 3.11. Two categories of *'common'* sites presented here include a set of *'ultimate common'* sites, and a set of sites occupied across all cell types in regions that were defined as consistently open between tissues (see Figure 2.4 for classification of those two categories).

In humans, there appears to be an increase in variation over the both *'common'* categories of binding sites (Figures 3.10a and 3.10b), but not over either germline or somatic-specific ones (Figures 3.10d and 3.10c). Highest increase is observed across the set of *'ultimate common'* sites (Figure 3.10a). While this category contains smallest number of sites (n=14,574), those are regions that are most consistently occupied between all of the cell types analysed, and therefore likely to represent the strongest binders with highest occupancy time. *'Common'* sites found in regions that are consistently open across cell types (n=52,017) also show enrichment of rare derived alleles (Figure 3.10b), while the somatic-specific ones (n=68,273) do not (Figure 3.10c). Spermatogonia-specific set has relatively small number of sites (n=24,330), and does not show a convincing enrichment of rare derived alleles (Figure 3.10d). It does, how-

ever, show a decrease in GERP scores (increase in sequence divergence) proximal to the putative binding sites.

In mice, there is a clear enrichment of rare variants across the *'ultimate common'* set of binding sites (Figure 3.11a), but not a very evident enrichment over the more general category of *'common'* sites (Figure 3.11b). However, there is a clear depletion of variation over the *'somatic-specific'* category (Figure 3.11c), similar to the human binding sites. There appears to be a more striking enrichment of rare variants over the *'spermatogonia-specific'* set of binding sites (Figure 3.11d), albeit a much lower levels of evolutionary constraint and purifying selection across them.

For human set of binding sites, to perform a more fair type of comparison, I attempted to account for the differential numbers of regions and variable protein-binding strength and occupancy between *'tissue-specific'* and *'common'* sites that are being compared. For each of the individual datasets I compared the sets of *'common'* and *'tissue-specific'* regions that contained the same numbers of sites with the matched distributions of peak scores (see Chapter 2, Subsection 2.2.7 for details of how this was done). As seen in Figure 3.12, with the matched categories of regions, sites in *'common'* category show the enrichment of variants, while ones in *'somatic-specific'* category do not. For one set of the germline binding sites (H5.25; Figure 3.13), there appears to be an increase in the *'spermatogonia-specific'* category. While with a modest increase compared to a *'common'* category, this set of *'spermatogonia-specific'* sites shows an enrichment of the rare derived alleles in contrast to the absence of variation across *'somatic-specific'* set. This increase is not present in other sets of *'spermatogonia-specific'* sites, potentially indicating that H5.25 cells represent the subset of cells where majority of mutations at binding sites are occurring.

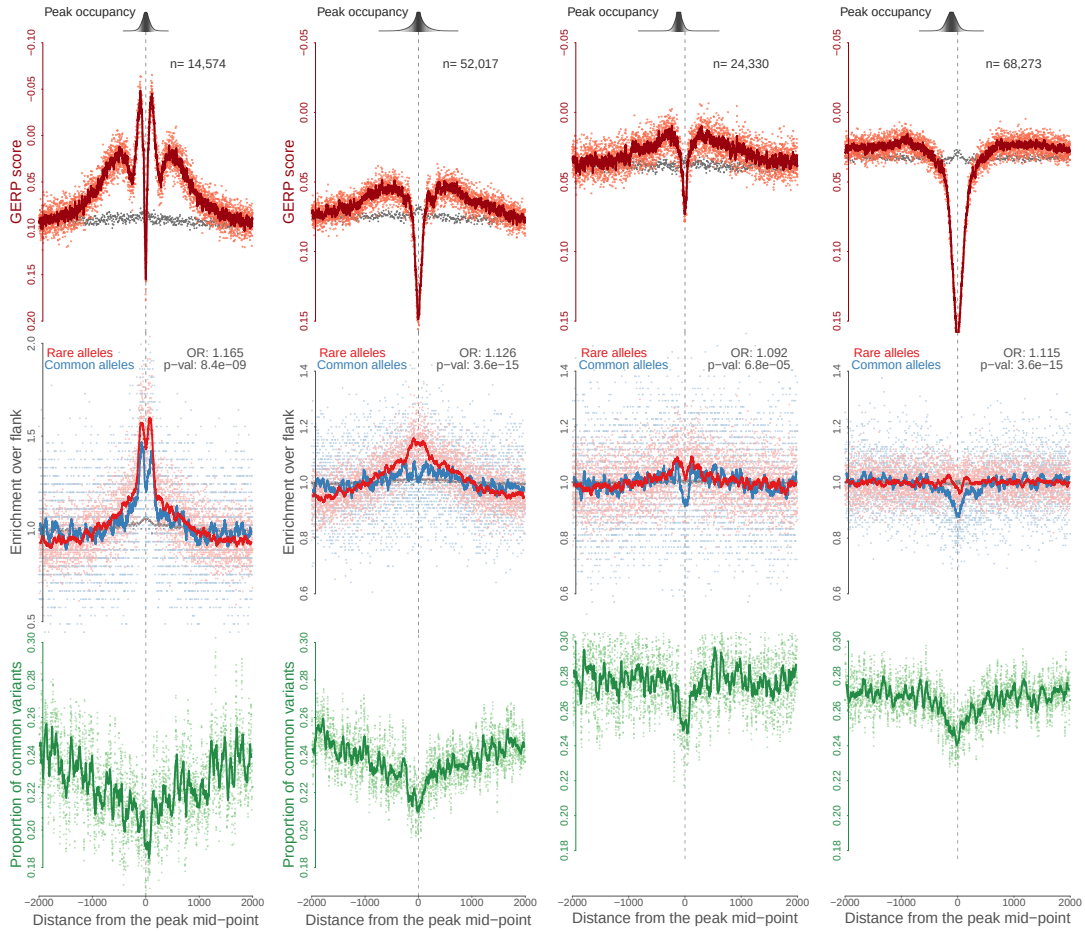


Figure 3.10: Variation over (a) 'ultimate' set of common, (b) common ,(c) spermatogonia-specific, and (d) somatic-specific ***human*** *protein-binding sites. Individual points represent measures at single nucleotide resolution (y-axis), while the lines are plotted through rolling average in 10bp (GERP), or 50bp (deCODE) sliding window. Grey points and lines represent an expectation from trinucleotide context. 'OR' denotes odds ratio of the enrichment of the rare derived alleles over common at the site relative to the flanks, with associated p-value. There is an enrichment of the rare derived alleles across binding sites that are bound by proteins in spermatogonial cells (a,b and c), but at those that are not bound in spermatogonial cells (d).*

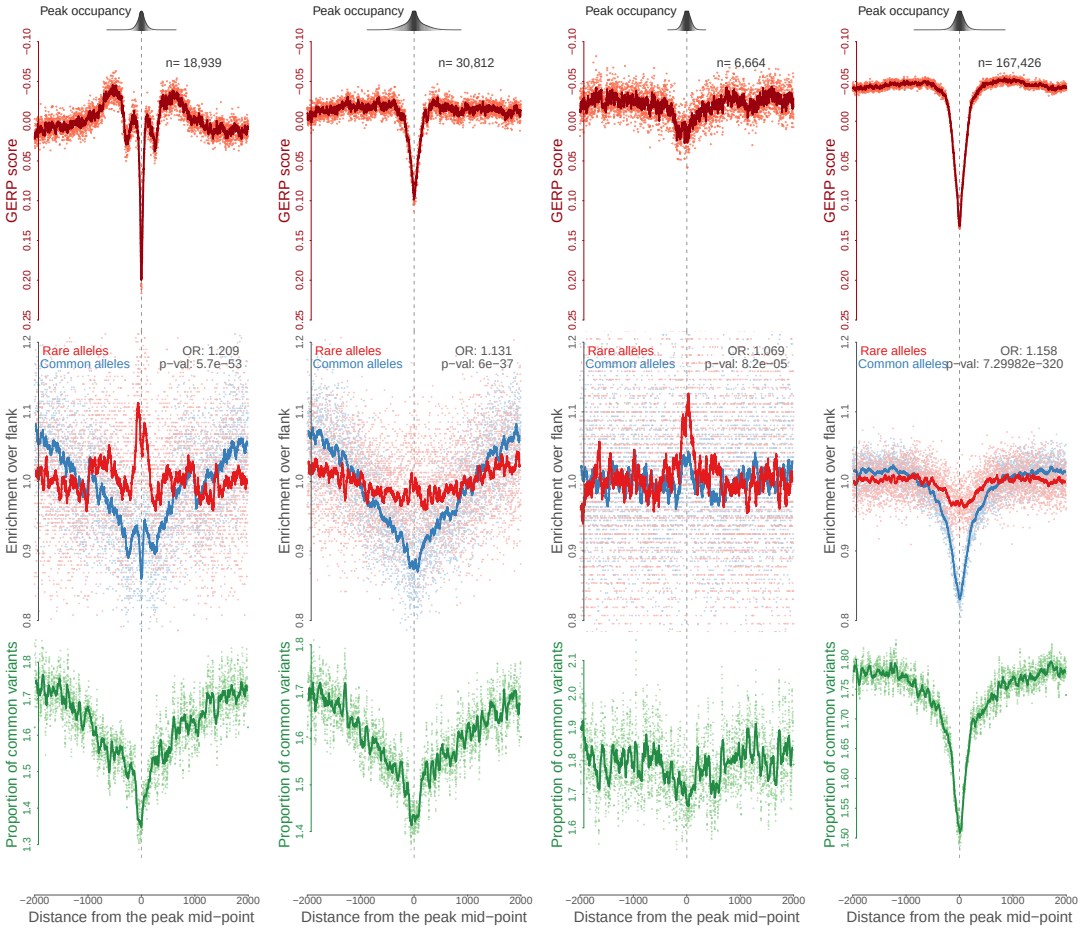
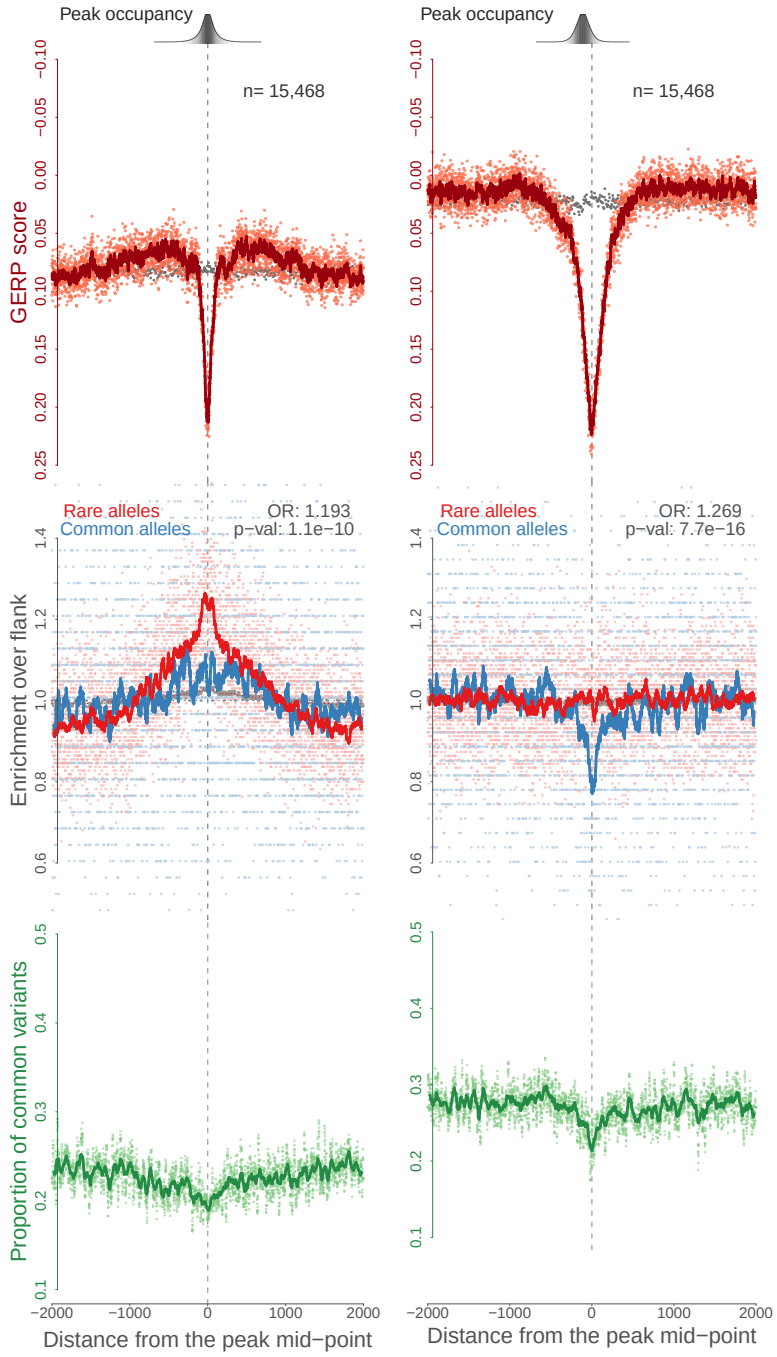
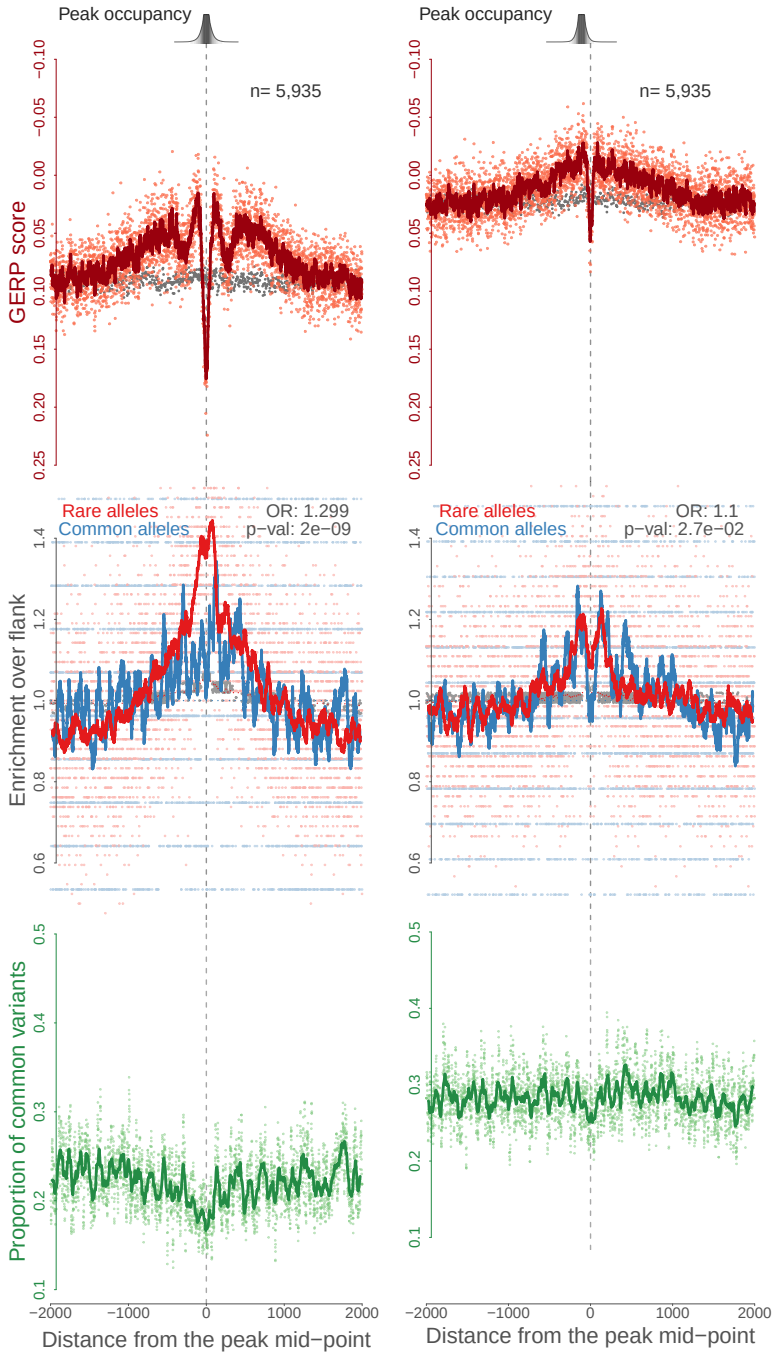


Figure 3.11: Variation over (a) 'ultimate' set of common, (b) common ,(c) spermatogonia-specific, and (d) somatic-specific ***mouse*** *protein-binding sites. Individual points represent measures at single nucleotide resolution (y-axis), while the lines are plotted through rolling average in 10bp (GERP), or 50bp (wild mice alleles) sliding window.'OR' denotes odds ratio of the enrichment of the rare alleles over common at the site relative to the flanks, with associated p-value. There is an enrichment of the rare alleles across binding sites that are bound by proteins in spermatogonial cells (a,b and c), but at those that are not bound in spermatogonial cells (d).*



(a) (b)

Figure 3.12: Variation over human colon (a) common and (b) somatic-specific binding sites sites. Both sets of sites have been matched in number and peak score distribution. Individual points represent measures at single nucleotide resolution (y-axis), while the lines are plotted through rolling average in 10bp (GERP), or 50bp (deCODE) sliding window. Grey points and lines represent an expectation from trinucleotide context. 'OR' denotes odds ratio of the enrichment of the rare derived alleles at the site relative to the flanks, with associated p-value.



(a) *(b)*

Figure 3.13: Variation over human spermatogonia (H525) (a) common and (b) germline-specific sites. Both sets of sites have been matched in number and peak score distribution. Individual points represent measures at single nucleotide resolution (y-axis), while the lines are plotted through rolling average in 10bp (GERP), or 50bp (deCODE) sliding window. Grey points and lines represent an expectation from trinucleotide context. 'OR' denotes odds ratio of the enrichment of the rare derived alleles at the site relative to the flanks, with associated p-value.

3.3.4 Edges of protein binding sites active in germline, but not somatic-specific ones show enrichment of germline mutations

I have also looked at the same variation measures around the protein-binding edges identified by the FLOP method (see Chapter 2, Subsection 2.3.7 for details of the method). It is more advantageous to aggregate the protein binding sites by centring them on the edge of the bound TF, rather than the mid-point, as we are particularly interested in the patterns of variation at the edges of the bound proteins. When looking at the aggregate set of multiple different proteins, their binding sites are going to be of varied widths. When aggregating different-widths sites centred on the mid-point, the pattern over the edges will be consequently diffuse. When aggregating multiple sites and centring on one of the edges, the signal over the boundary is going to be more defined.

Figures 3.14 and 3.15 show patterns of between- and within-species variation over the protein-binding edges (CpGs excluded). A similar pattern as seen before can be observed with an increase in the between-species divergence near the edge of the defined *'common'* binding sites, followed by region of constraint when looking at GERP scores, and increased number of derived alleles from deCODE. There is an observable and highly localized spike of increased *expected* enrichment of both derived alleles and between-species substitutions right at the edge, which is mirrored by the *observed* pattern in the case of GERP scores, but not alleles within the human population. This peak is most likely caused by the sequence context of the Tn5 preferential insertion bias.

An increase in the proportional abundance of the rare derived alleles *versus* common relative to the flanking regions is observed at both binding site and over the binding site edge. This indicates that there is purifying selection acting all across the region. Consistent with the results from the previous subsection, this enrichment is absent from the somatic-specific edges category. There is a subtle increase of variation surrounding the spermatogonia-specific edges and a pronounced difference in GERP profiles between spermatogonia-specific and somatic-specific binding sites.

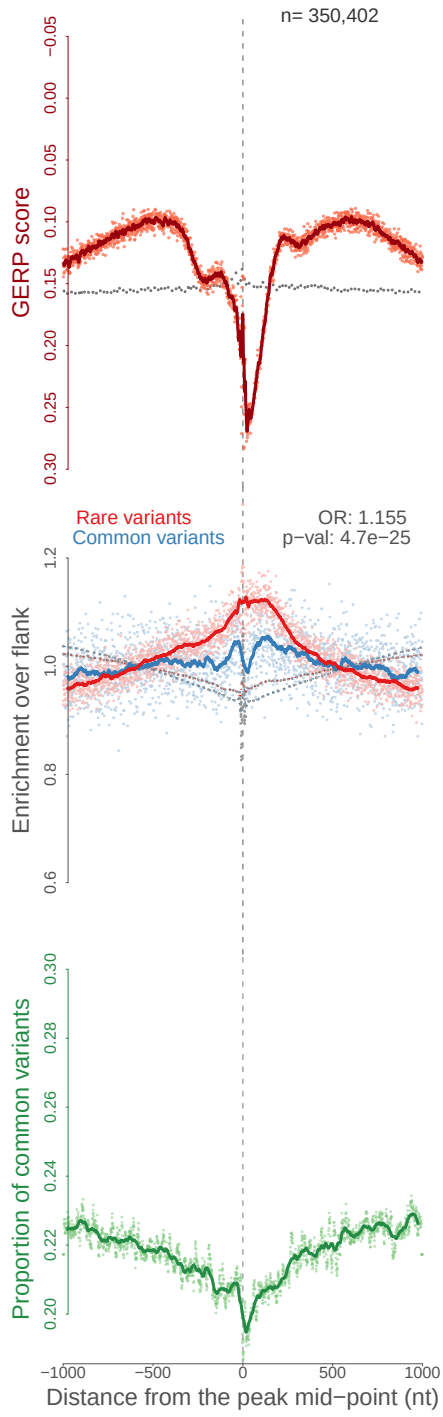
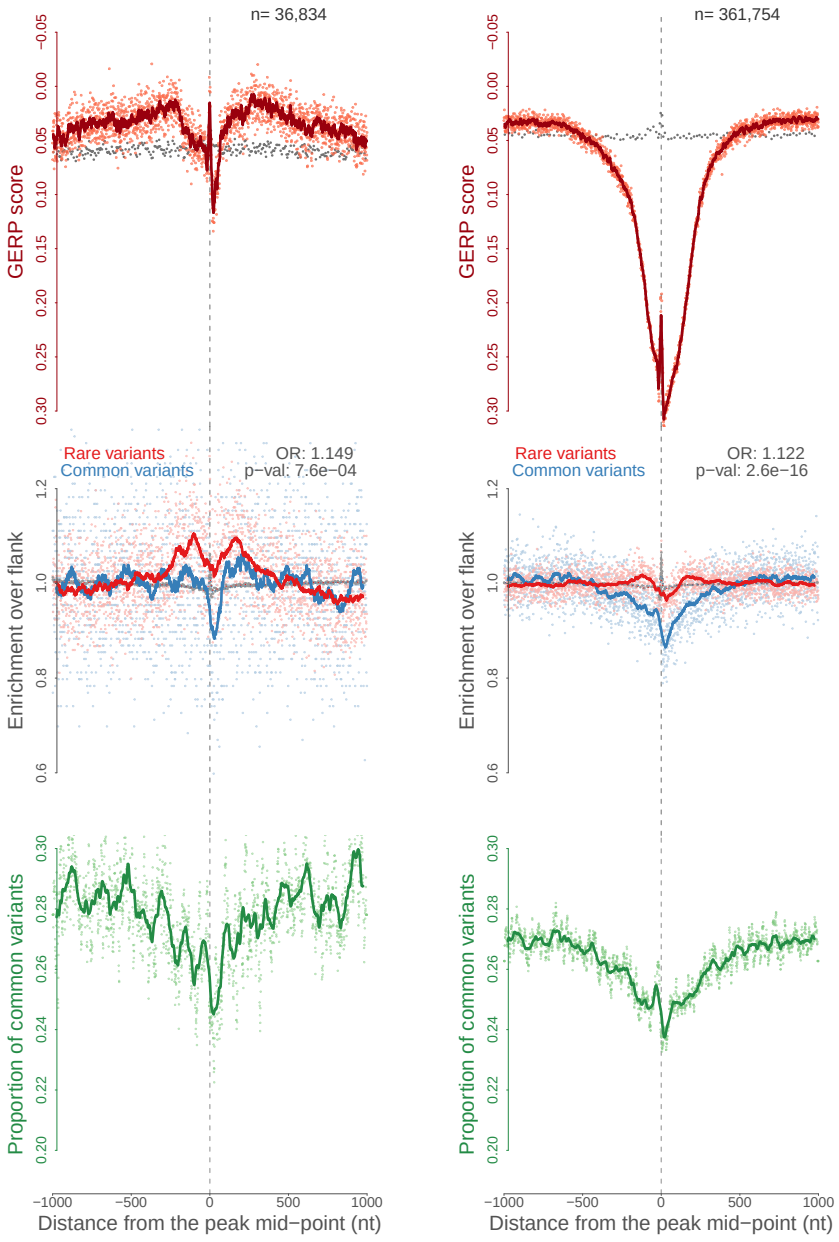


Figure 3.14: Germline variation at aggregate of the 'common' set of human binding edges (CpGs excluded). All binding edges have been oriented so that the putative binding sites proceed to the right of plot midpoint. Individual points represent measures at single nucleotide resolution (y-axis), while lines are plotted through rolling average in 10bp (top plot), or 50bp (middle and bottom plots) sliding window. Grey points and lines represent an expectation from trinucleotide sequence context. Top plot shows sequence between-species divergence as measured by GERP scores. Higher GERP scores (lower down the y-axis) represent sequence constraint. Middle plot shows enrichment of rare and common deCODE derived alleles relative to flanks (see Figure 3.3 for details of how this was plotted). 'OR' denotes odds ratio of the enrichment of the rare derived alleles at the -5:+5bp region around the binding edge relative to the flanks, with associated p-value. Bottom plot shows proportion of common derived alleles, representing pattern of selection unaffected by mutation rate, with lower values indicating stronger purifying selection pressure.



(a) Spermatogonia-specific

(b) Somatic-specific

Figure 3.15: Variation over the human spermatogonia-specific (a) and somatic-specific (b) edges (CpGs excluded). Individual points represent measures at single nucleotide resolution (y-axis), while the lines are plotted through rolling average in 10bp (GERP), or 50bp (deCODE) sliding window. Grey points and lines represent an expectation from trinucleotide sequence context. 'OR' denotes odds ratio of the enrichment of the rare derived alleles at the -5:+5bp region around the binding edge relative to the flanks, with associated p-value.

3.3.5 There is an enrichment of *de novo* mutations at 'housekeeping' protein-binding sites

As mentioned in Subsection 3.1, *de novo* mutations are the most direct way of measuring germline mutation rate, but currently available datasets have less power compared to derived alleles from population variation measures. There are two ways in which statistical power could be increased – by increasing the number *de novo* mutations within the dataset, or by looking at larger number of sites. We are rather limited as far as the latter point, as there is a finite number of spermatogonia-specific sites that one can look at. However, there have been increasing numbers of studies looking at *de novo* mutations, and those are likely to be growing in the future, at least as far as human data is concerned. Prior to starting these analyses, I estimated the size of the human *de novo* variant dataset that would be required to be able to detect the enrichment of mutations within the binding site (in this case the FLOP method-derived binding sites defined from ATAC-seq from GM12878 cell line from Buenrostro et al. (2013) were used). That was done by using rare derived alleles from deCODE as proxy for *de novo* mutations. From down-sampling of this dataset, I estimated that a cohort with $\approx 200,000$ *de novo* mutations would have a 97% power to detect a significant increase in mutational burden of the complete set of human binding sites (based on 1000 bootstraps, Figure 3.16).

There is currently information on 268,759 human *de novo* mutations available from multiple different studies (see Subsection 3.2.4 for details). While the size of the datasets is not sufficiently large to be able to detect the enrichment of variants across individual categories of the binding sites, for example, spermatogonia-restricted or somatic tissue-restricted, I have compared bootstrapped average *de novo* mutation rates in the peak region *versus* flanks in all defined human binding sites across all tissues (Figure 3.17), as well as *'tissue-specific'* (Figure 3.18) and *'common'* (Figure 3.19) binding sites in all tissues types separately. While there is a clear and significant increase in the *de novo* mutation rate at the binding site *versus* flank in the combined set of sites and in the *'common'* category, there is no definitive difference in the *'tissue-specific'* one.

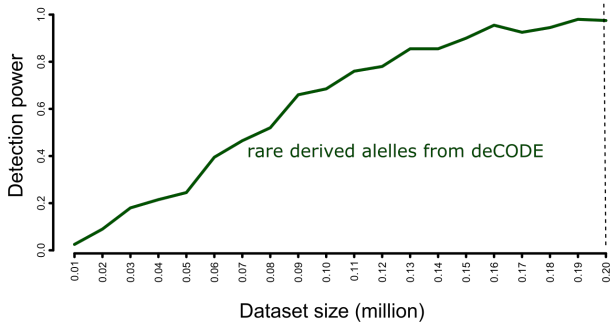


Figure 3.16: Estimation the required size of human de novo mutation dataset. Rare derived alleles from deCODE were used as proxy for de novo mutations. Increasing number of variants were randomly sampled until the significant increase in numbers of mutations in binding sites (GM12878 cell line binding sites defined by FLOP method) versus flanks could be detected. From this, a de novo mutation dataset with 200 000 variants is sufficient to detect elevated mutational burden at binding sites.

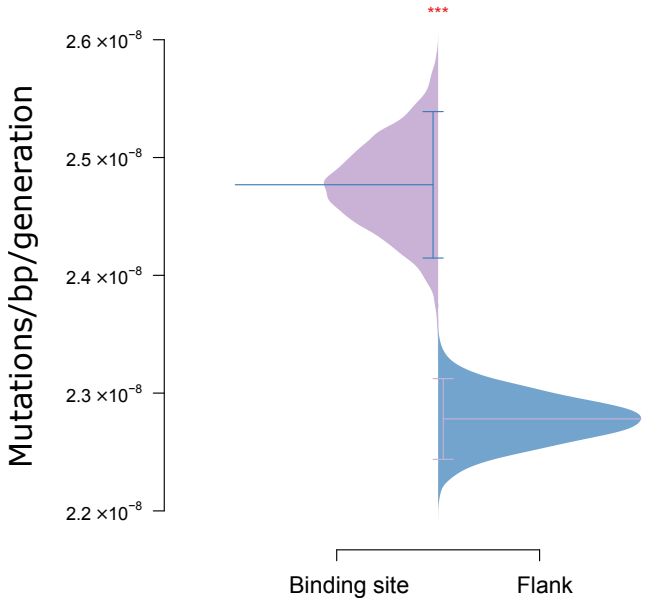


Figure 3.17: Differences in the numbers of human de novo mutations (with bootstrapping) between binding sites (-100:100bp relative to the middle of the peak) and flanks (-1000:-800,800:1000bp away from the peak boundaries). Inverse-coloured line represents an observed mean (without bootstrapping) and bars indicate 95% confidence intervals (estimated from bootstrapping). P-val (***) < 0.0005

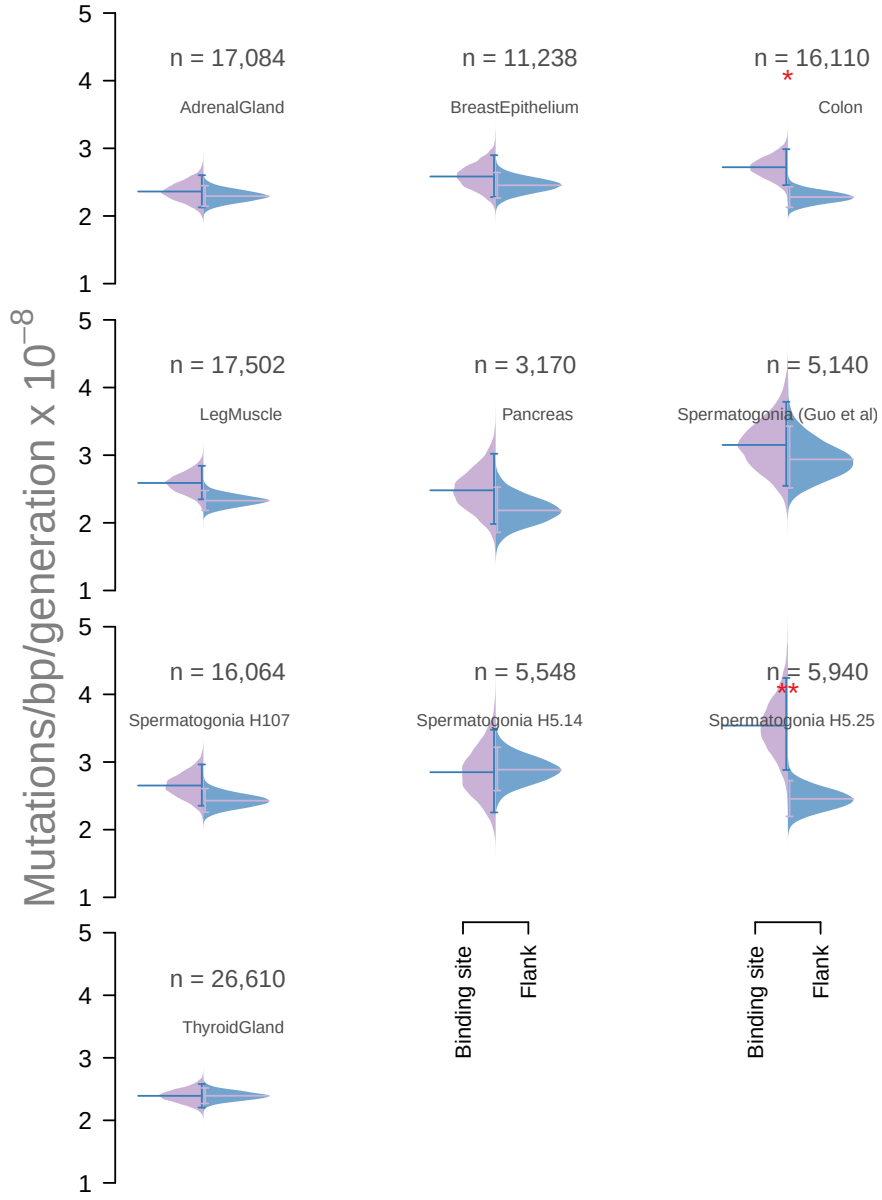


Figure 3.18: Differences in the numbers of human de novo mutations (with bootstrapping) between tissue-specific peaks (purple) and flanks (blue; -1000:-800,800:1000bp relative to the peak midpoint. Inverse-coloured line represents an observed mean (without bootstrapping) and bars indicate 95% confidence intervals (estimated from bootstrapping). P-val (**) < 0.005 < p-val (*) < 0.05

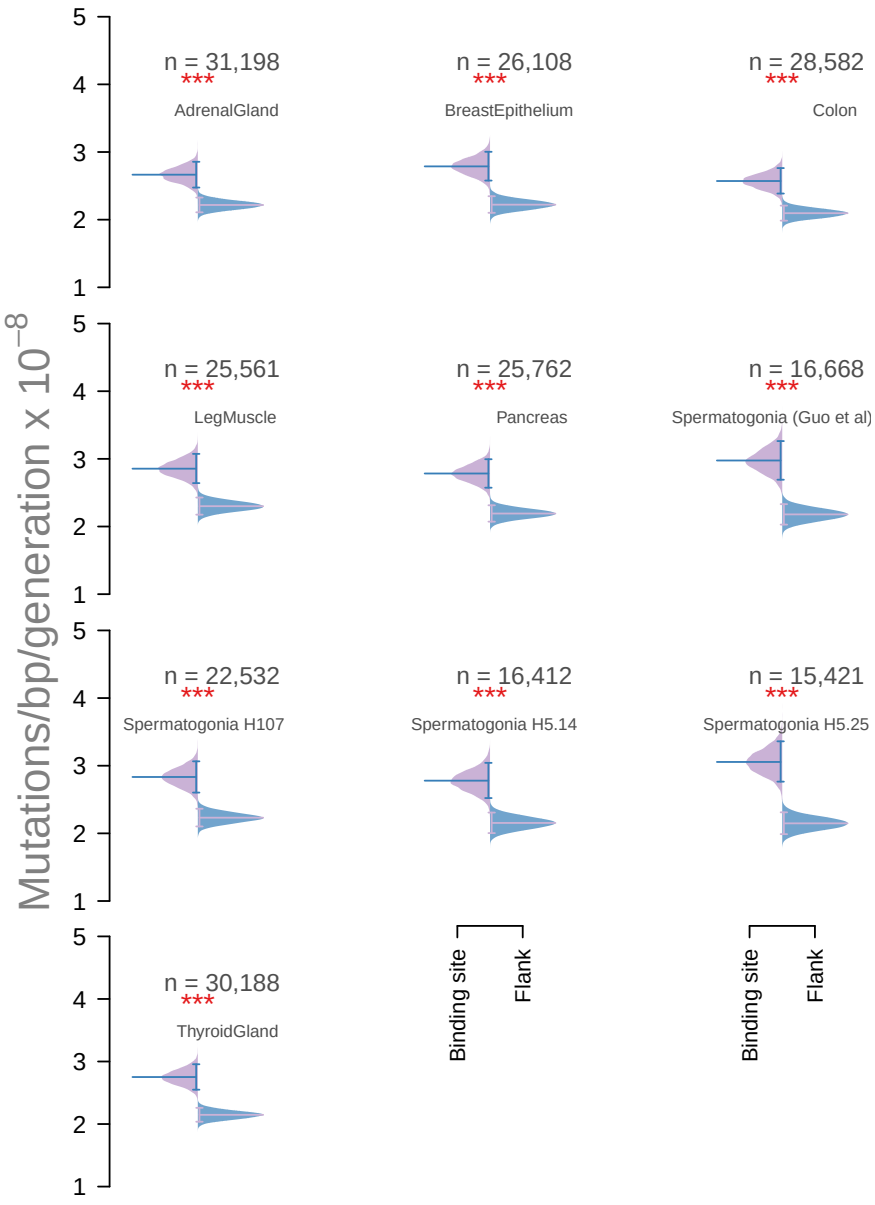


Figure 3.19: Differences in the numbers of human de novo mutations (with bootstrapping) between common peaks (purple) and flanks (blue; -1000:-800,800:1000bp relative to the middle of the peak). Inverse-coloured line represents an observed mean (without bootstrapping) and bars indicate 95% confidence intervals (estimated from bootstrapping). P-val (**) < 0.005 < p-val (*) < 0.05

The number of potentially deleterious human *de novo* mutations per birth (generation) can be calculated using set of formulas below, where:

ω - number of deleterious mutations per birth

μ - *de novo* mutations per bp

ϑ - common derived alleles per bp

B - binding site

F - flank

δ - *de novo* mutations per proband

$$\frac{\mu_B}{\mu_F} = \frac{1.099 \times 10^{-4}}{9.955 \times 10^{-5}} = 1.104$$

$\longrightarrow$ 10.4% excess in *de novo* mutation in the binding site (*f*:3.3.1)

$$\frac{\vartheta_B}{\vartheta_F} = \frac{1.996 \times 10^{-3}}{2.106 \times 10^{-3}} = 0.947$$

$\longrightarrow$ 5.3% variants lost to purifying selection under naïve assumption of uniform mutation rate (*f*:3.3.2)

$$\frac{\mu_B \div \mu_F}{\vartheta_B \div \vartheta_F} = \frac{1.104}{0.947} = 1.165$$

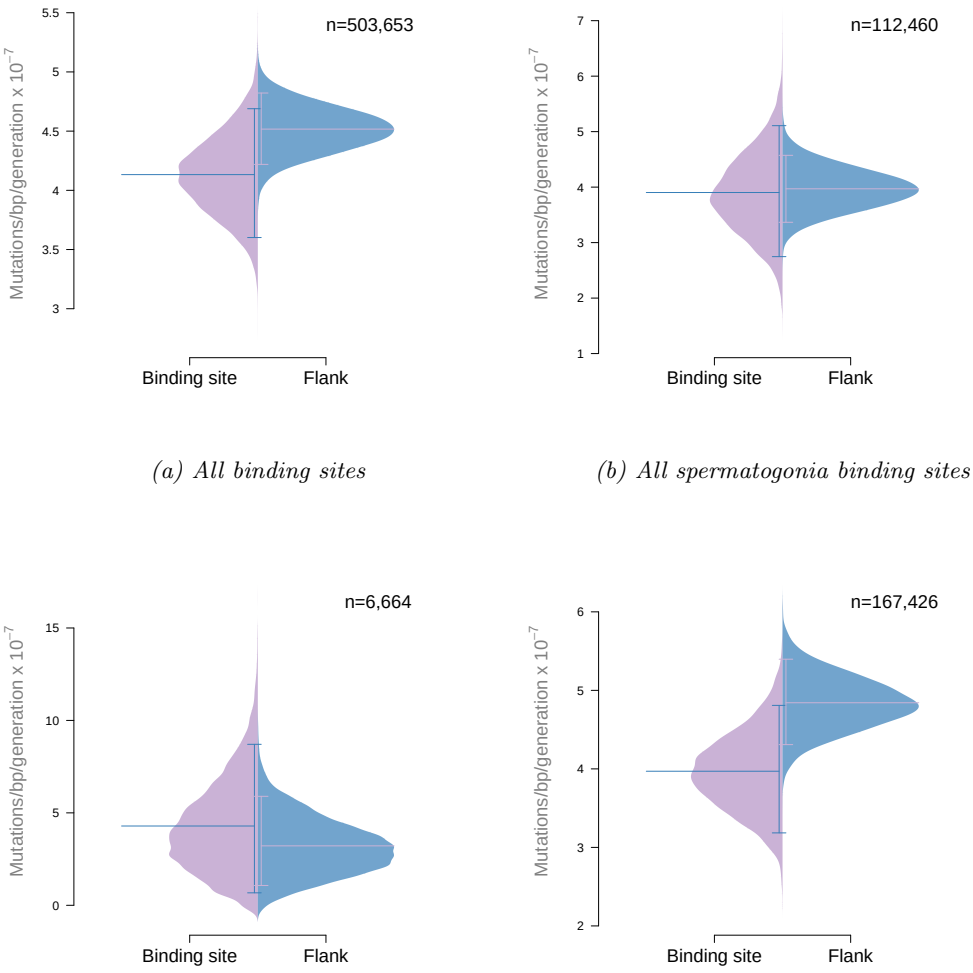
$\longrightarrow$ 16.5% of *de novo* mutations are deleterious (*f*:3.3.3)

$$\delta_B = \frac{6601}{4370} = 1.51 \quad (f{:}3.3.4)$$

$$\omega = \text{proportion of deleterious mutations} \times \delta_B =$$

$$= 0.165 \times 1.51 = 0.249 \text{ (deleterious mutations/birth)} \quad (f{:}3.3.5)$$

Mouse *de novo* mutation dataset only contains 7,007 substitutions, and bootstrapped data for mouse binding sites is presented in Figure 3.20. As this dataset contained few mutations, unsurprisingly, there is no significant difference between numbers of *de novo* substitutions between the binding site and the flanks. However, there is a slight trend in somatic-specific binding sites being depleted of *de novo* mutations, while spermatogonia-specific ones showing a small increase.



(a) All binding sites

(b) All spermatogonia binding sites

(c) Spermatogonia-specific binding sites

(d) Somatic-specific binding sites

Figure 3.20: Differences in the numbers of ***mouse*** *de novo mutations (with bootstrapping) between peaks (purple) and flanks (blue; -1000:-800,800:1000bp relative to the middle of the peak). Inverse-coloured line represents an observed mean (without bootstrapping) and bars indicate 95% confidence intervals (estimated from bootstrapping). P-val (***) < 0.0005 < p-val (**) < 0.005 < p-val (*) < 0.05*

3.4 Discussion

The aim of the work described in this Chapter was to test that protein binding sites in the germline carry an increased mutational burden and if this results in an accumulation of deleterious mutations at those potentially functionally important regions. In addition to that, I wanted to examine whether the protein binding is causal of this increase in mutation rate. I explored variation over sets of binding sites that are active in multiple cell types, and tried to disentangle patterns of mutation and selection that contribute to the observed variation.

Here, the particular interest was to investigate the increased numbers of germline mutations that could be potential causes of heritable disease. Evidence of the possible elevated mutation rates in germline has been presented previously by Reijns et al. (2015), from between-species divergence measures (Figure 1.8). They have at the same time proposed a potential causal replication-associated mutational mechanism (Figure 1.9). Following on from that study, I was especially interested in looking at cells where most of the germline replications is thought to be occurring - spermatogonial cells (Figure 2.1; Box 2.1).

The aggregate set of putative binding sites identified in spermatogonial cells showed an increased in polymorphic sites compared to the flanking regions (Figure 3.5 and 3.9). Those variants have arisen as mutations at some point in germline lineage cells. This enrichment either demonstrates a differential mutation rate between sites and flanks, or can be evidence of diversifying selection. Those binding sites at the same time show high level of between-species conservation, consistent with the action of purifying selection. Through separation of those variants based on the frequency of the derived allele, I was able to perform a derived allele frequency test and infer the predominant selectional pressure acting on the region. The proportion of the rare alleles within the binding site is higher than it is in the presumably more neutrally evolving flanking regions. This suggests that the binding sites are under purifying selection, and therefore an elevated mutation rate is likely responsible for increased variability. The observation of the increase in the between-species divergence at the edges of the protein-binding sites is similarly driven by the increased mutation rate, rather than diversifying selection (Figure 3.15a).

This means that many of those mutations are potentially harmful, as they are being purged from the population, making protein binding sites in human and mouse germlines hotspots for deleterious mutations that could lead to heritable disease. Based on the distributions of rare and common derived alleles over the spermatogonia-active binding sites, about 20% of mutations occurring within them are likely to be deleterious. From analysis of *de novo* mutation and common derived allele frequencies across all of the identified human binding sites, I calculate that ≈15% of all *de novo* mutations at those sites are likely to be deleterious, corresponding to ≈0.245 mutations per generation (Formula *f:*3.3.5). This means that nearly 1 in 4 births is likely to harbour one of those deleterious mutations within a binding site.

I hypothesise that the physical presence of the protein at the binding site is necessary for this mutational pressure to be exerted. In order to test this, I looked at the germline variation at two different categories of regions – those that are commonly bound between all different tissue types (so called "housekeepers") and sites that are preferentially occupied in only one tissue type. I was particularly interested in looking at sites that are preferentially bound only in the cells of the germline and those bound only in the somatic cells. To demonstrate the dependence of the elevated mutation rate on protein binding, one would expect to see the increase in germline variation at the sites occupied in germ cells (this category includes both housekeeping and germline-specific binding sites), but not at somatic-specific binding sites. It is worth emphasizing that we do not necessarily expect all of the binding sites for all the TFs to exhibit an increase in mutation rate, but possibly only a subset.

There is an evident enrichment of human population polymorphisms over the *'common'* category of binding sites (≈ 25% increase). Those measures are made using the variants in non-CpG context, to circumvent our inability to get a correct estimation of the *expected* numbers of rare and common derived alleles from the trinucleotide context, due to a differential methylation state of the regulatory sites and their flanks. Measure of derived alleles found in a CpG context indicates that addition of those variants would make the mutational burden even bigger. Somatic-specific binding sites do not appear to exhibit increased germline variation. This is supported by the variation patterns observed across mouse protein-binding sites. The fact that the somatic-specific sites do not exhibit increase in germline variation means that the elevated mutation rate is not just an inherent feature of the binding sites as such. Taken together, this sup-

ports the hypothesis of protein binding being associated with the increased mutational burden.

There is an enrichment of rare alleles across the *'spermatogonia-specific'* category of binding sites in mouse, further supporting this hypothesis. However, complete sets of human spermatogonia-specific binding sites does not show an increase in germline variation to the same extent as germline sites in mouse, or human housekeeping binders do. There does appear to be a modest enrichment over what is observed for the somatic-specific binders, mainly at the edges of the putative binding sites. The more prominent increase in mutation rate over the spermatogonia-specific protein-binding site category in mice rather than humans is surprising. Mouse variants were obtained from only 10 sequenced individual, and therefore this dataset was expected to have less power than human one to detect an increase in mutation rate. However, with the advantage of the genetic lineage-tracing method for cell isolation, cell populations obtained from mice might be a better representative of the highly dividing spermatogonial cells. Therefore, possibly, more specific and relevant protein binding sites were identified in mice.

Furthermore, the *'common'* category does have more regions than the *'spermatogonia-specific'* ones, so insufficient power cannot be ruled out as a potential reason. To test this, I looked at the paired sets of common and tissue-type specific human protein-binding sites that have been matched in number and score peak distribution (a joint proxy for binding strength and occupancy). When comparing those two matched sets, *'common'* binding sites still exhibit more variation than the *'tissue-specific'* category ones. However, there is some evidence for elevated mutation rate in a subset of spermatogonia-specific binding sites (H5.25), even though it is a modest effect compared to *'common'* binders and is not compellingly seen across all replicates.

Those results are recapitulated by the comparisons of the numbers of human *de novo* mutations at binding sites *versus* flanks. Although these comparisons were performed with bootstrapped data, I observe the mutation rate to be significantly increased in human protein-binding sites in the *'common'* subset, while neither of the *'tissue-specific'* subsets are significantly enriched for mutations. *'Spermatogonia-specific'* H5.25 subset of protein-binding sites shows the largest enrichment of mutations within the site, consistent with previous results.

There are number of potential reasons why we are not seeing such clear evidence of elevated mutation at human germline-specific binding sites:

- The types of sites we are comparing might be different. As described in Subsection 2.3.5, the *'common'* and *'tissue-specific'* sites are found in different genomic contexts and might even represent binding sites of variable level of interaction strength and occupancy. It is worth noting that this point would equally be applicable to the somatic-specific category of sites as well.

- In addition to that, we do not necessarily expect all of the binding sites for all the TFs to induce an increase in mutation rate. It might be that it is primarily the 'housekeeping' TFs that are effecting increased mutational pressures, and therefore a higher proportion of 'mutable' binding sites happen to be in the *'common'* set.

- It could be that we are actually looking at the binding sites in the wrong cell type. The main advantage of a mouse model is the ability to more precisely target a specific cell population for isolation through use of genetic lineage tracing methods. Therefore, cells isolated from mice are more likely to represent the most highly dividing subset of spermatogonial cells, while human ones might represent a more diverse population.

- Potentially, ATAC-seq might not have captured the most representative protein-binding landscape of the cells that we have isolated, or possibly the binding site identification and separation method used here does not give the best differentiation between the different categories of binders.

Taken together, this analysis shows that protein-binding sites that are active within the human and mouse germlines are hotspots for likely deleterious mutations that could potentially lead to heritable disease, with 'housekeeping' category of sites being most highly affected. Physical presence of the protein on DNA appears to be associated with induction of those mutations, as demonstrated by the lack of germline polymorphism enrichment over the somatic-specific binding sites. The affected sites are active in the population of potentially most highly dividing cells of male germline, with the numbers of division increasing linearly with age of the male. This implies association of this mutational property and the increasing numbers of *de novo* mutations in offspring of older fathers, if the mechanism is indeed replication-dependent.

CHAPTER 4

Transcription factors as a biased mask of mutagenic lesions

134

4.1 Introduction

4.1.1 Somatic mutations can lead to cancer and drive further mutagenesis

Mutations occurring within the cells of the germline are going to be propagated to the next generation and could become a cause of hereditary disease, while those that occur within the somatic cells are not heritable and would only affect the individual in question, and would perish with the death of the organism. The consequences of these mutations, however, are much more important from the perspective of an individual, as they would potentially lead to diseases that would affect an organism's fitness or survival.

Tumour growth and cancer are some of the most pronounced and medically important consequences of somatic cell mutations. Tumour growth is characterised by dysregulated control of cell division and survival, and if coupled to invasiveness of other tissues, another aspect of dysregulated genetic control, is considered a cancer. These dysregulations are often a consequence of specific genetic mutations that drive proliferation or inhibit apoptosis (Hanahan and Weinberg, 2011). As multiple mutations are generally required to transform a normal somatic cell into a dis-regulated tumour cell (Zarnack et al., 2013), cancers are often associated with mutator phenotypes - genetic dis-regulations that lead to the increased rate of somatic mutagenesis or arise as a consequence of exposure to mutagenic agents that increase the mutational load. Consequently making the other mutational hits necessary for oncogenesis more likely. As noted previously (Subsection 3.1), the statistical power of local mutation pattern analysis is dependent in part on the number of mutations available. The high mutation load and large number of cancers sequenced along with non-cancerous somatic cells from the same individuals, make this an attractive system in which to explore the details of transcription factor binding associated mutagenesis. Comparisons between the genome sequence of cancer and normal samples of the same individuals provide a well controlled analysis system to identify somatic mutations with confidence. This is also helped by the nature of cancers to grow by clonal expansion, making it easy to call those mutations that occurred early in the progression of the cancer in contrast to

calling somatic mutations from clonally diverse tissue samples.

4.1.2 Different cancers are driven by various processes and can exhibit distinct mutations and lesions

Different cancer types are known to exhibit distinct mutational signatures - when complete set of mutations acquired during cancer progression is dominated by particular types of changes (Alexandrov, 2018), often evaluated in the context of immediately neighbouring bases (trinucleotide context) (Blake et al., 1992). For example, malignant melanoma is dominated by the C⟶T mutations (Alexandrov et al., 2013). Mutational signatures are often a reflection of particular processes that are responsible for initiation and progression of the disease. Since across the genome most of the changes are *passenger* mutations that do not confer an advantage to the cell, they are not strongly affected by selection. The origins of some signatures are known, for example, the aforementioned signature 5 dominated by C⟶T changes is known to be caused by ultraviolet radiation, and signature 4 (C⟶A) by tobacco smoke, while others remain elusive (Alexandrov et al., 2016). Endogenous sources of damage, frequently originating from the failures of DNA maintenance processes, also leave distinct imprints, such as signature 6 due to the defective mismatch repair, or signatures 2 and 13 as a hallmark of the activity of the APOBEC enzymes (Alexandrov et al., 2013).

Large cohorts of cancer mutation data, such as TCGA (`https://cancergenome.nih.gov`) and ICGC (`https://dcc.icgc.org/`), have made it possible to study the mutational landscapes of different tumours and infer the processes that contribute to them. Because of the different mutational processes attributed to different cancers, and the different mutational loads in cancer cohorts, one might expect variation between cancer types. Typically, cohorts contain large numbers of mutations that allow for exploration of the differences in the types of mutations and the rates at which they occur at various sites across the genome, such as protein binding sites. Because there are variable defects and sources of mutagenic lesions that can be inferred from these cancer genomes, it provides an avenue with which to explore the mechanistic basis of particular mutation patterns, an approach exemplified by Supek and Lehner (2015, 2017).

4.1.3 Paradoxical observations of mutation retention by transcription factors at protein binding sites in cancer

In Chapter 3, I showed that mutations are elevated at the protein binding sites. The mechanistic basis of this mutational heterogeneity is not clear, but binding of the proteins appears to be necessary, and potentially causal for increase in mutation rate.

As mentioned in Section 1.2.2, most DNA-interacting proteins, and in particular TFs, tend to have a sequence-specific preference in where they bind. Signalling networks that control proper cell function are reliant on this specificity to turn on and off expression of certain genes at particular times. This sequence preference can be defined by the methods of ChIP-seq and its variations, followed by search for recurrence of specific sequences, or motifs (*de novo* motif discovery methods reviewed in Lihu and Holban (2015)). From those, the position weight matrices (PWMs) can be constructed, which describe how often a particular type of nucleotide occurs within each position of the motif. Alternatively, systematic evolution of ligands by exponential enrichment (SELEX) method can be used, where protein affinity to a pool of oligonucleotides with random sequences is tested to select the best binders (Klug and Famulok, 1994).

Protein binding sites have been previously observed to be mutated in cancers (Kaiser et al., 2016; Sabarinathan et al., 2016). One of the striking examples is an increase cancer mutations within the binding motifs of CTCF. This TF is a ubiquitously expressed DNA-binding protein with a variety of different roles as a transcription factor, insulator or repressor, modulator of chromatin organisation and so on, which is largely conserved in eukaryotes (Kim et al., 2015). A particular position harbours excess of single nucleotide substitutions within CTCF binding sites (Katainen et al., 2015; Kaiser et al., 2016). This specific mutation is expected to lead to abolition of the protein binding ability (Umer et al., 2016).

Models that aim to explain the increase of mutations at TF binding sites mostly propose retention of lesions by DNA-interacting proteins, be that either though impeding procession of high fidelity polymerase (Reijns et al., 2015) or through prevention of access by repair machinery (Sabarinathan et al., 2015). With CTCF being a sequence-specific binding TF, this seems paradoxical. It implies that while mutation within CTCF motif impacts the binding, a lesion at the same position does not, as to occlude it from repair processes, the protein has to still be able to bind to its target

sequence. Potential differences in effects on protein binding ability of a lesion *versus* a mutation are to date largely unexplored. This, however, provides an avenue for investigation of causes for elevated rates of harmful mutations at the TF binding sites observed in Chapter 3.

Differences in tolerance of a TF to various types of changes within its binding motif would naturally depend on the structural properties of the protein-DNA contacts and its plasticity. CTCF is classified as a member of C2H2 zinc finger (ZnF) proteins, as it interacts with DNA with different combinations of its central domain tandem array of $Cys_2 - His_2$ zinc fingers (Filippova et al., 1996). Crystal structure of CTCF-DNA complex has recently been described (Hashimoto et al., 2017), allowing for investigation of the differences in binding in context of physical interactions between protein residues and DNA strands. Investigation of mutational patterns within motifs of other TFs could provide an insight into whether the CTCF motif-observed pattern is generalisable across a particular type of binding factor (C2H2 ZnF) and beyond. Separation of the substitutions observed in different types of cancers could allow for speculations about types of lesions and changes that are likely to be causal.

4.1.4 Questions addressed in the current Chapter

In this Chapter I (1.) propose a model that could explain the mutational pattern at the CTCF binding site, (2.) explore other examples of TFs showing an increase in mutations within their motifs, and (3.) consider how those proteins interact with DNA to glean any insights into the mechanistic basis of the protein binding-induced mutagenesis. I extend the analysis done by Kaiser et al. (2016) by looking at the mutational patterns at the sequence-specific binding sites in distinct cancer cohorts that exhibit different predominant mutation signatures and types of lesions, focusing on TFs for which motifs and ChIP-seq data is available. I also (4.) touch on the possibility of discovering the highly mutable binding sites without relying on the predefined information about the specific motifs.

4.2 Methods

4.2.1 Cancer mutation data

Most of the cancer mutation data used here was obtained through the ICGC (**I**nternational **C**ancer **G**enome **C**onsortium). Files containing simple somatic mutations were downloaded for each cancer and study type separately from the ICGC portal (`https://dcc.icgc.org/releases/release_26/Projects`; release 26; *hg19* assembly). Variants were then filtered to only include single nucleotide substitutions from whole genome sequencing. Some studies (LINC-JP; LIRI-JP; PRAD-CA) had entries for mutations without base identity change (e.g. C⟶C), and those have been excluded from the analysis (total of 6,179 that kind of mutations were excluded). The filtered set of single nucleotide substitutions contained a total of **65,103,967** mutations from **5,249** donors. Table 4.1 shows the numbers of variants from donors obtained from each of the study and cancer type.

4.2.2 Mismatch repair-deficient cancer data

Mutations from the mismatch repair-deficient (MMRd) tumours were obtained from both ICGC and Wang et al. (2014). From ICGC, mismatch repair-deficient breast cancer (BRCA) tumours were identifies same as in Supek and Lehner (2017), where any tumours with $>$ 3000 indels were classed as MMRd (**9** donors; **352,554** single nucleotide substitutions). Mutations from the MMRd stomach cancers were obtained from Wang et al. (2014) study (**10** donors; **678,464** single nucleotide substitutions).

4.2.3 ChIP-seq data and motif scanning

ChIP-seq data for multiple TFs was downloaded from the ENCODE database (`https://www.encodeproject.org`; data obtained on 22.01.2018). All the ChIP-seq files which did not pass at least one of the audit filters set by ENCODE (such as insufficient read depth/length, severe bottlenecking, missing control alignments, or poor library complexity) were excluded from the analysis. All of the coordinates were converted to correspond with *hg19* assembly. Multiple ChIP-seq files for the same TF were intersected with each other (`bedtools multiinter` command) and for each TF, any of

Study	Donors	Mutations	Cancer type
MELA-AU	183	22,744,345	Skin Cancer - AU
ESAD-UK	301	9,631,784	Esophageal Adenocarcinoma - UK
SKCA-BR	100	7,510,534	Skin Adenocarcinoma - BR
BRCA-EU	569	3,588,578	Breast ER+ and HER2- Cancer - EU/UK
LIRI-JP	258	3,530,231	Liver Cancer - RIKEN, JP
MALY-DE	241	2,970,635	Malignant Lymphoma - DE
LUSC-KR	30	1,693,299	Lung Cancer - KR
PACA-CA	209	1,665,047	Pancreatic Cancer - CA
PACA-AU	235	1,216,876	Pancreatic Cancer - AU
LICA-FR	39	1,212,543	Liver Cancer - FR
OV-AU	93	1,000,155	Ovarian Cancer - AU
PBCA-DE	453	891,640	Pediatric Brain Cancer - DE
PRAD-UK	215	746,044	Prostate Adenocarcinoma - UK
RECA-EU	95	683,947	Renal Cell Cancer - EU/FR
PRAD-CA	236	656,225	Prostate Adenocarcinoma - CA
BRCA-FR	72	621,805	Breast Cancer - FR
LMS-FR	67	577,784	Soft tissue cancer - Leiomyosarcoma - FR
EOPC-DE	202	491,360	Early Onset Prostate Cancer - DE
GACA-CN	37	480,269	Gastric Cancer - CN
LINC-JP	31	399,770	Liver Cancer - NCC, JP
CLLE-ES	151	393,081	Chronic Lymphocytic Leukemia - ES
BRCA-UK	45	356,631	Breast Triple Negative/Lobular Cancer - UK
BTCA-SG	71	341,889	Biliary Tract Cancer - SG
COCA-CN	26	247,326	Colorectal Cancer - CN
NKTL-SG	23	229,580	Blood Cancer - T-cell and NK-cell lymphoma - SG
ORCA-IN	25	226,614	Oral Cancer - IN
BOCA-UK	64	198,992	Bone Cancer - UK
PAEN-AU	50	160,224	Pancreatic Cancer Endocrine neoplasms - AU
PAEN-IT	37	147,017	Pancreatic Endocrine Neoplasms - IT
PRAD-FR	25	123,894	Prostate Cancer - Adenocarcinoma - FR

Table 4.1: Numbers of donors and counts of mutations for each cancer study from ICGC used in this analysis (continued on the next page)

Study	Donors	Mutations	Cancer type
PRAD-FR	25	123,894	Prostate Cancer - Adenocarcinoma - FR
COAD-US	44	52,826	Colon Adenocarcinoma - TCGA, US
CMDI-UK	30	49,160	Chronic Myeloid Disorders - UK
LIAD-FR	5	40,649	Benign Liver Tumour - FR
BOCA-FR	98	36,337	Soft Tissue cancer - Ewing sarcoma - FR
SKCM-US	37	36,084	Skin Cutaneous melanoma - TCGA, US
READ-US	16	24,244	Rectum Adenocarcinoma - TCGA, US
STAD-US	38	17,421	Gastric Adenocarcinoma - TCGA, US
LUSC-US	48	17,005	Lung Squamous Cell Carcinoma - TCGA, US
THCA-SA	128	14,589	Thyroid Cancer - SA
LAML-KR	8	12,767	Acute Myeloid Leukemia - KR
LUAD-US	38	11,741	Lung Adenocarcinoma - TCGA, US
BRCA-US	91	9,098	Breast Cancer - TCGA, US
HNSC-US	44	7,920	Head and Neck Squamous Cell Carcinoma - TCGA, US
GBM-US	41	6,522	Brain Glioblastoma Multiforme - TCGA, US
BLCA-US	23	6,232	Bladder Urothelial Cancer - TCGA, US
LIHC-US	54	5,659	Liver Hepatocellular carcinoma - TCGA, US
OV-US	42	3,581	Ovarian Serous Cystadenocarcinoma - TCGA, US
KIRP-US	33	2,351	Kidney Renal Papillary Cell Carcinoma - TCGA, US
KIRC-US	37	2,326	Kidney Renal Clear Cell Carcinoma - TCGA, US
SARC-US	34	1,965	Sarcoma - TCGA, US
KICH-US	45	1,900	Kidney Chromophobe - TCGA, US
CESC-US	20	1,533	Cervical Squamous Cell Carcinoma - TCGA, US
DLBC-US	7	1,375	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma - TCGA, US
ESCA-CN	17	1,064	Esophageal Cancer - CN
LGG-US	18	518	Brain Lower Grade Glioma - TCGA, US
PRAD-US	19	422	Prostate Adenocarcinoma - TCGA, US
LUSC-CN	10	421	Lung Cancer - CN
NBL-US	40	134	Neuroblastoma - TARGET, US
ALL-US	1	4	Acute Lymphoblastic Leukemia - TARGET, US

Table 4.1: Numbers of donors and counts of mutations for each cancer study from ICGC used in this analysis (continued)

the regions that were found to be present in at least half of the files were kept. Genomic sequences in regions of interest were extracted using `bedtools getfasta` command.

FIMO software (part of MEME suite version 4.11.1) (Grant et al., 2011) was then used to look for the occurrence of motifs for the sequence-specific TFs in regions identified by ChIP-seq with the default options and default p-value threshold of 10^{-4}. PWMs for all the motifs were downloaded from the Jaspar database (`http://jaspar.genereg.net/downloads/`), keeping the redundant motifs for each of the TFs (multiple PWMs for the same TF from different studies). Only matrices obtained from human studies were retained. Table 4.2 details TFs that had both ChIP-seq and PWMs available, numbers of ChIP-seq replicates, and counts of motifs identified.

Factor	PWM	Number of ChIP experiments	Cell type	Motifs on +	Motifs on -
ARNT	MA0004.1	10	K562	0	0
ATF1	MA0604.1	10	K562	832	807
BCL6B	MA0731.1	10	HEK293	307	265
CEBPB	MA0466.1	10	K562	6928	7094
CEBPB	MA0466.2	10	K562	5351	5263
CEBPG	MA0838.1	10	K562	3009	2964
CREB1	MA0018.1	30	HepG2,MCF-7	686	695
CREB1	MA0018.2	30	HepG2,MCF-7	1245	1230
CREB1	MA0018.3	30	HepG2,MCF-7	2802	2802
CTCF	MA0139.1	41	foreskin keratinocyte, HEK293,LNCaP clone FGC	11638	11838
EGR1	MA0162.2	2	H1-ESC	3885	3630
EGR1	MA0162.3	2	H1-ESC	1300	1288
FOS	MA0476.1	10	MCF-7	6270	6286
FOXA1	MA0148.1	10	HepG2	463	483
FOXA1	MA0148.2	10	HepG2	471	488
FOXA1	MA0148.3	10	HepG2	436	473
FOXA2	MA0047.1	10	HepG2	2498	2435
FOXA2	MA0047.2	10	HepG2	4051	3906
FOXK2	MA1103.1	10	HepG2	262	224
GABPA	MA0062.1	90	GM12878,HeLa-S3,HepG2,HepG2,HL-60,K562,liver,MCF-7	706	760
GABPA	MA0062.2	90	GM12878,HeLa-S3,HepG2,HepG2,HL-60,K562,liver,MCF-7	931	988
GFI1B	MA0483.1	10	HEK293	456	429
GLIS1	MA0735.1	10	HEK293	1151	1134
GLIS2	MA0736.1	10	HEK293	4338	4355
HES1	MA1099.1	10	MCF-7	107	101
HIC1	MA0739.1	10	HEK293	1259	1370
IRF9	MA0653.1	10	K562	175	198
KLF1	MA0493.1	20	HEK293,K562	1289	1392
KLF16	MA0741.1	10	HEK293	2114	2221
KLF4	MA0039.3	1	hESC	5697	2594
KLF4	MA0039.3	1	hESC	5697	2594
KLF4	MA0039.2	1	hESC	6393	20474
KLF4	MA0039.1	1	hESC	4935	9585
MAFG	MA0659.1	10	K562	2925	2945
MNT	MA0825.1	30	HepG2,MCF-7	119	130
NFE2L2	MA0150.1	39	A549,HeLa-S3,HepG2,IMR-90	786	836
NFE2L2	MA0150.2	39	A549,HeLa-S3,HepG2,IMR-90	866	899
NRF1	MA0506.1	10	MCF-7	966	970
PKNOX1	MA0782.1	30	HEK293T,K562,MCF-7	367	369
PRDM1	MA0508.1	20	HEK293,HEK293	8163	7811
PRDM1	MA0508.2	20	HEK293,HEK293	2799	2964
REST	MA0138.1	10	HEK293	1074	1141
REST	MA0138.2	10	HEK293	1018	961
RFX1	MA0509.1	10	HepG2	540	529
SCRT1	MA0743.1	10	HEK293	2214	2159
SCRT2	MA0744.1	10	HEK293	4993	4977
SP1	MA0079.1	30	HEK293T,HepG2,MCF-7	32	39
SP1	MA0079.2	30	HEK293T,HepG2,MCF-7	123	104
SP1	MA0079.3	30	HEK293T,HepG2,MCF-7	140	110
SP2	MA0516.1	9	HEK293	43515	44124
SP3	MA0746.1	10	HEK293	5067	5464
YY2	MA0748.1	10	HEK293	271	335
ZBED1	MA0749.1	10	K562	45	51
ZBTB7B	MA0694.1	10	MCF-7	248	255
ZEB1	MA0103.1	10	HEK293	0	0
ZEB1	MA0103.2	10	HEK293	1347	1503
ZEB1	MA0103.3	10	HEK293	1624	1698
ZNF24	MA1124.1	20	HEK293,K562	1335	1431
ZNF384	MA1125.1	10	HEK293T	23817	21074
ZNF423	MA0116.1	10	HEK293	401	398
ZSCAN4	MA1155.1	10	HEK293	18892	18333

Table 4.2: List of transcription factors with available ChIP-seq and motif data.

4.2.4 Mutation rate calculation and plots

Genome-wide mutation rates were calculated for each of 64 possible trinucleotide contexts (where the mutated base was in the middle of the triplet) recording the change to each of the 3 alternate bases. Rates were calculated by dividing the number of observed mutations in each trinucleotide context by the total number of the trinucleotide occurring in the genome (excluding the low mappability regions). For plotting purposes, 64 trinucleotide categories were folded into 32 categories, where each of the trinucleotides was matched with the corresponding reverse-complement trinucleotide (*e.g.* A$\underline{\text{G}}$C $\longrightarrow$ A$\underline{\text{T}}$C change was matched with G$\underline{\text{C}}$T $\longrightarrow$ G$\underline{\text{A}}$T change).

For the aggregate set of motifs (binding sites) for individual TFs, the *observed* mutation rate at each of the individual positions of the motif was calculated by dividing the number of observed mutations (change to specific type of base) in particular trinucleotide context by the number of occurrences of that trinucleotide at the position. The *expected* mutation rate (change to a specific type of base) at each position was calculated by multiplying the genome-wide calculated rate for each trinucleotide by the instances of that trinucleotide observed at particular position. The ratio of the *observed* to *expected* mutation rate was then plotted to look for the enrichment of particular mutations at certain positions of the motif. Trinucleotides that did not constitute a reasonable enough proportion (10%) of all the trinucleotides at an individual position were excluded from plotting, for clarity of visualization. All the graphs were plotted using R (3.3.2).

4.2.5 Pentanucleotide mutational frequencies

Mutation rates for each pentamer (5-nucleotide sequence with the mutated base at the middle position 3) were calculated for each of the donors in the same way as trinucleotide mutation rates. Rates were either compared within a trinucleotide category (all pentanucleotides with the same trinucleotide in the middle) genome-wide, or between genomic contexts (different regions in the genome). The rates observed in different genomic contexts were compared by dividing mutation rate observed in one region by the observed mutation rate in another region. Those were then plotted as either the ratios in the context of each trinucleotide (32 possible categories when folded), or as

individual points in Region 1 *versus* Region 2 mutation rate plots.

4.3 Results

4.3.1 Biased mask model

Mutations mostly initiate as lesions that occur on one of the Watson or Crick strands of double-stranded DNA. If the lesion is not repaired, after the next round of replication one of the daughter cells will receive a copy with original base, while the other will receive a copy with a lesion that will have been either fixed or converted into a mutation (Figure 4.1).

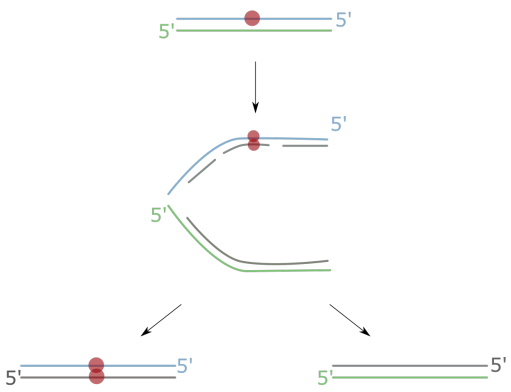


Figure 4.1: Consequences of replicating an unrepaired lesion on a single strand of DNA. A lesion (red circle) has formed on a single strand of DNA (blue), and has not been repaired prior to replication commencing. After synthesis of new strands (grey) with affected (blue)) and unaffected (green) strands as templates, one of the daughter cells receives a non-mutated copy (bottom right), while the other daughter cell receives a copy where lesion has turned into mutation with change in nucleotide identity on both of the strands (bottom left).

While TFs can be thought of as sequence-specific binders, the range of contacts that they make with either of DNA strands can differ. The importance of position within a motif is typically measured in a strand-agnostic nature - *e.g.* we do not necessarily know if one strand, both strands or neither are making contact with protein residues, unless we have some structural evidence, such as a crystal structure of the protein-DNA complex. Even if there is a structure available, we can't readily predict how change at any position of a motif sequence on a strand that is contacted or not contacted by the TF can influence the conformation of the DNA. While molecular dynamic simulations may offer some insights in this case (Liu and Heermann, 2015; Blanco et al., 2018) there is almost no grounding experimental data on the interactions of sequence specific binding TFs interacting with mismatched or modified bases beyond

base methylations that don't alter base-pairing (Hashimoto et al., 2017). While binding motifs are defined primarily by the nature of the nucleotides that proteins bind over, the shape of the DNA, rather than sequence as such, has a major contribution to binding site recognition by TFs (Mathelier et al., 2016; Yang and Ramsey, 2015). A lesion or a mismatch can potentially distort DNA conformation in a way that could affect binding. Similarly, we do not possess knowledge of the strand (Watson or Crick) that mutations have occurred on.

The unusual pattern of mutations observed at CTCF binding sites could be explained by a model that we propose, and term here the *biased mask model*. This model explores the possibility of some of the sequence-specific binding proteins exhibiting a higher tolerance to occurrence of a single-stranded lesion, in contrast to a mutation, at a particular position of the motif. Tolerance of those lesions by proteins, and thus their retention through them being masked from normal cellular maintenance processes by binders, could lead to consequent occurrence of mutations post-replication. This model explains the observed mutational patterns (such as observed at CTCF binding sites) as the consequence of several possible scenarios, as illustrated in Figure 4.2.

In the first scenario (Figure 4.2, left), change of the nucleotide identity on either of the strands has no effect on protein binding. Thus, the TF is likely to frequently bind across and occlude a lesion and one would predict an excess of mutations at that position. However, those changes would not be expected to have any phenotypic consequence, unless the mutation at that position affects binding more than individual mismatches on either of the strands separately. In the another scenario (Figure 4.2, middle), change of the nucleotide identity at either of the strands results in impairment of the protein binding. In that case, upon formation of a lesion, a protein is unable to act as a barrier to the procession of high-fidelity polymerase, or obscure it from detection by repair machinery. Hence, that lesion has a higher probability of being fixed and not be observed to be mutated very frequently. In the third scenario, a lesion on one of the strands affects protein binding ability, while a lesion on the other strand does not. As a result, occurrence of a lesion on the permissive strand would allow for protein binding and lead to its occlusion. After the next round of replication, one of the daughter cells would end up with a binding site that is disrupted on both of the strands, including the non-permissive one and binding would be lost. Similarly, protein tolerance to lesions on both strands, and not to the mutation would lead to a similar

outcome. The mutational excess one would expect to see in scenario 3 is exactly what is observed at position 9 of the CTCF motif with a nucleotide that is important for binding, but at the same time highly mutated.

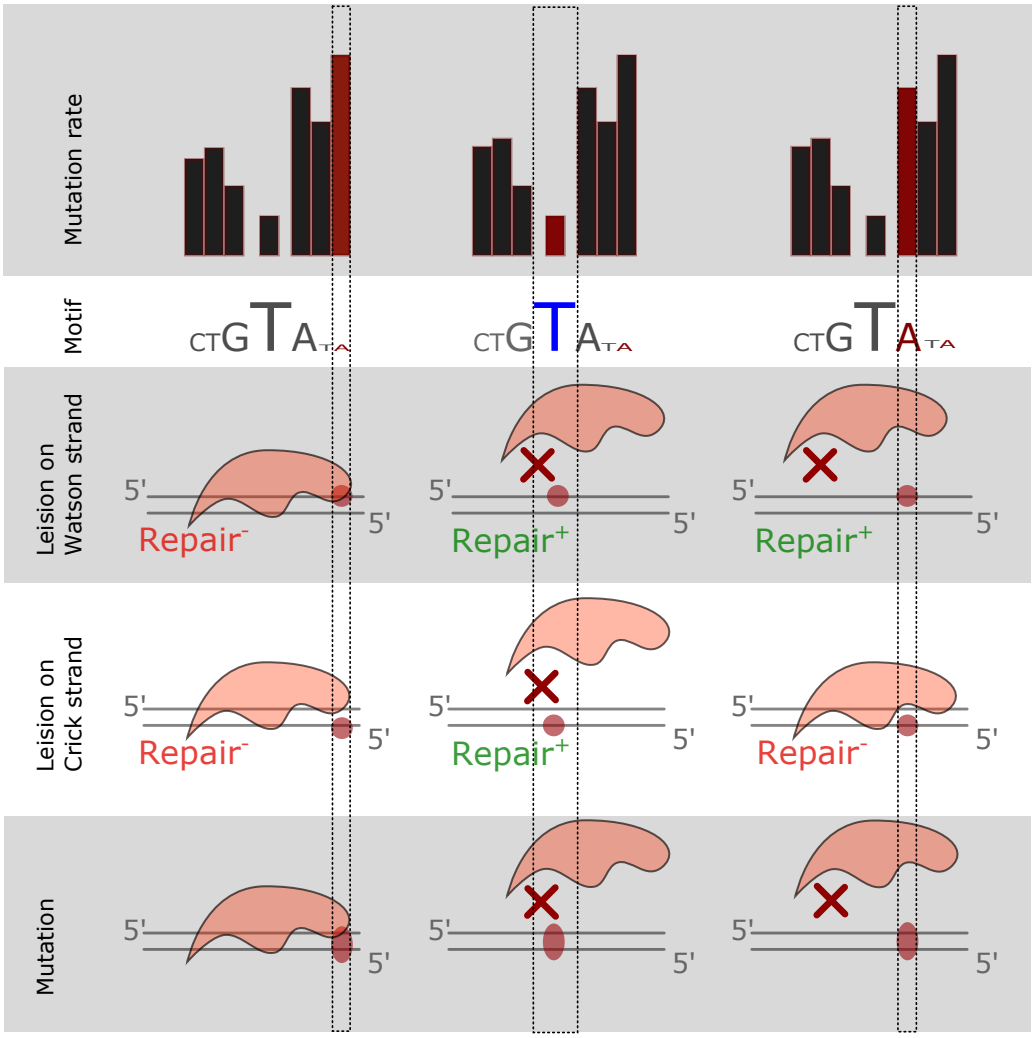


Figure 4.2: Biased mask model. This model aims to explain the preferential occurrence of specific types of mutations at particular sites of protein-binding motifs as a consequence of protein binding dynamics. Mutations originate as lesions on either Watson or Crick strands, and there are different consequences for protein binding depending on whether lesion is tolerated by the protein. Higher mutation rates at sites bound by a protein in the presence of a lesion assumes protection of lesions from repair machinery (Repair$^-$). Described fully in Section 4.1

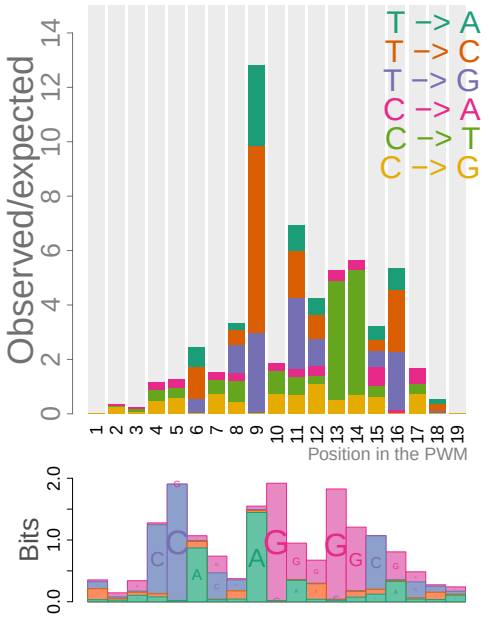
4.3.2 Other zinc finger protein motifs have positions with increased mutation load similar to CTCF

In light of the observation that position within the motif of one of the C2H2 ZnFs DNA-binding TFs, CTCF, exhibits an unusually high number of mutations in cancers despite its importance in binding, I wanted to investigate whether there are other TFs in that family that exhibit similar mutational patterns. Two proteins were of particular interest - KLF4 and EGR1, as crystal structures in complex with DNA are available for those (Schuetz et al., 2011; Hashimoto et al., 2014, 2016). KLF4 and EGR1 are C2H2 ZnF proteins each containing 3 zinc fingers that they use to interact with target DNA sequences. KLF4 belongs to the category of 'pioneer' TFs that are able to bind closed chromatin and initiate its opening (Zaret and Carroll, 2011), and is one of the four Yamanaka factors (Takahashi and Yamanaka, 2006), that have an ability to reprogram a cell to induce a pluripotent state. It is expressed in a wide range of tissues, with one of the main roles being promoting cell survival (Ghaleb and Yang, 2017). EGR1 (also known as ZIF268) is an early growth response protein involved in signal transduction, which is rapidly induced by various signals such as growth factors, stress, injury and oxygen deprivation (reviewed in Pagel and Deindl (2011)).

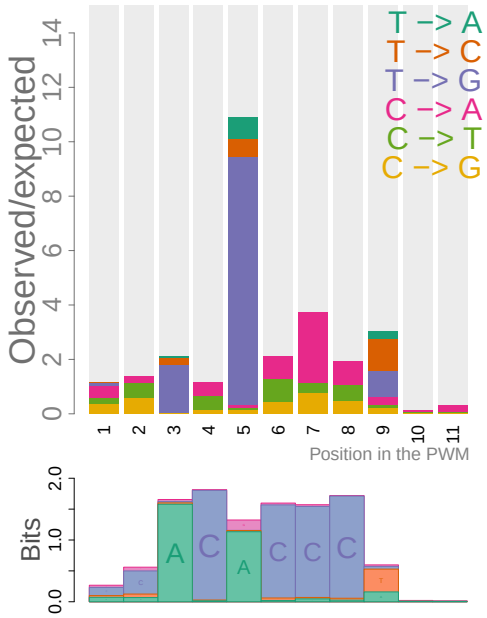
Figure 4.3 shows the excess of particular types of mutations at each position of the motifs over what would be expected from the genome-wide mutation spectrum in that trinucleotide context. All CTCF, KLF4 and EGR1 appear to have positions within their binding motifs that exhibit a site-specific excess of mutations in the pan-cancer dataset.

In agreement with previous observations (Kaiser et al., 2016; Umer et al., 2016; Katainen et al., 2015), position number 9 within the CTCF motif shows largest excess of the T$\longrightarrow$C (A$\longrightarrow$G) changes, followed by T$\longrightarrow$A/G (A$\longrightarrow$T/C) single nucleotide substitutions at position number 9, where adenine (thymine on opposite strand) is a preferred base (Figure 4.3a). Umer et al. (2016) has previously shown that T$\longrightarrow$C and T$\longrightarrow$A mutations at that position disrupt CTCF binding.

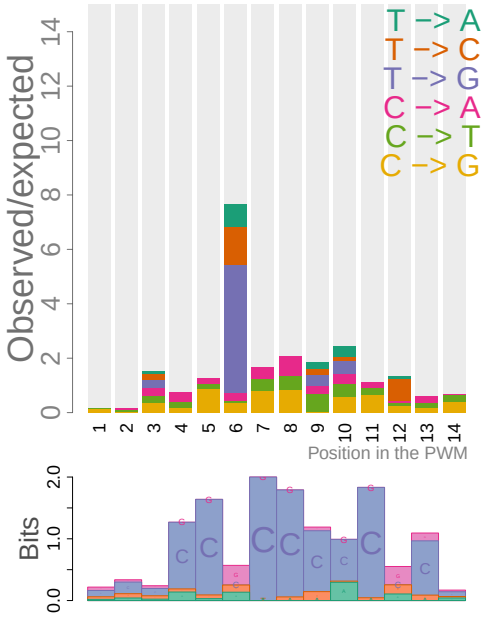
There are several PWMs available for KLF4 protein in the Jaspar database (`http://jaspar.genereg.net`; accessed July 2018), two of which are annotated as mouse-derived (MA0039.1 and MA0039.2), and one as human-derived (MA0039.3). MA0039.1 is a SELEX-derived matrix and does not appear to resemble most of the



(a) CTCF mutational spectrum



(b) KLF4 mutational spectrum



(c) EGR1 mutational spectrum

Figure 4.3: Mutational spectra for KLF4, CTCF and EGR1. Top set of bars in each subplot represent the ratio of the observed mutation rate at each position versus the expected mutation rate, based on the genome-wide rate of substitutions to each alternate nucleotide in each trinucleotide context, essentially showing the excess of mutations when accounted for the tumour-specific mutational pattern. Bottom set of bars represents the PWM for TF in question. Positions number 9 in CTCF motif, number 5 in KLF4 motif, and number 6 in EGR1 motif show peaks of unusually elevated mutation rates.

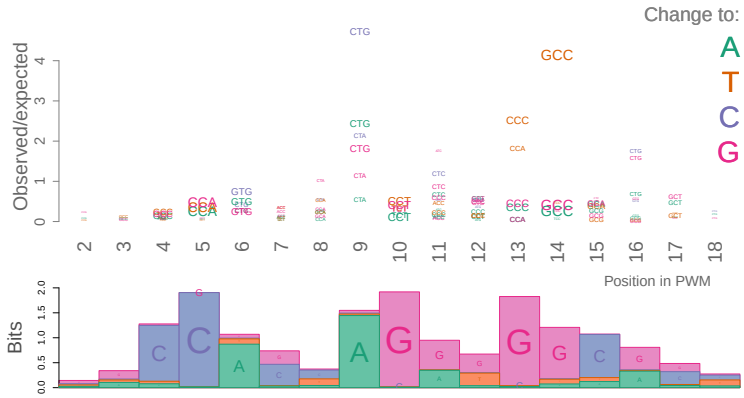
published KLF4 binding motifs. MA0039.2 is marked as mouse-derived, but was validated in the Chen et al. (2008) study. There *de novo* hESC ChIP-seq derived matrix appears similar, but with a slightly higher weight of thymine at the position number 5. This is in accordance with the human MA0039.3 matrix, derived from the reanalysis of multiple human cell line/tissue KLF4 ChIP-seq datasets (Chèneby et al., 2018), where position number 5 adenine (reverse complement of position 5 thymine in MA0039.2)

is an overrepresented base. Therefore, here MA0039.3 was used as a representative of the KLF4 binding sites. KLF4 TF normally exhibits a preference for adenine or guanine at the position number 5 (Watson strand) of its binding motif (MA0039.3), while the most frequently encountered mutation is potentially disrupting T⟶G (A⟶C), rather than T⟶C (A⟶G) or C⟶T (G⟶A) that would preserve the binding motif (Figure 4.3b). C⟶T is generally most abundant type of change at binding sites and genome-wide, but as here we are observing an excess of changes when adjusted for the genome-wide trinucleotide mutational rare, this particular mutation class does not dominate. EGR1 exhibits and abundance of the T⟶G mutations at the position that does not carry as much importance as positions numbers 9 and 5 according to the PWMs in CTCF and KLF4, respectively, but the level raises above what we can see in the flanks.

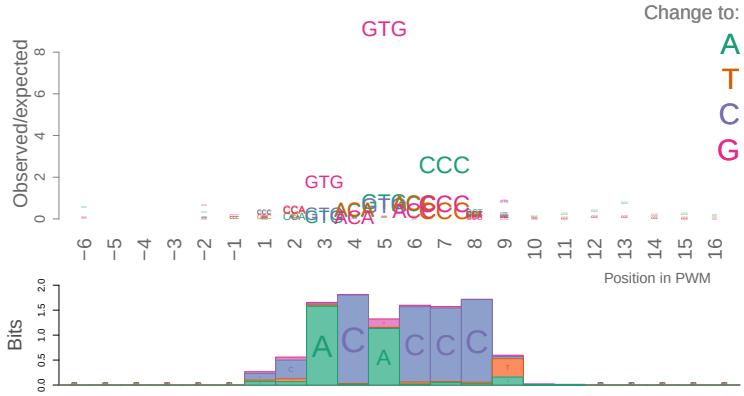
When exploring the particular types of change in a trinucleotide context, we can see that G<u>T</u>G⟶G<u>G</u>G mutation is observed most frequently at position number 5 in KLF4 (Figure 4.4b), which is the most abundant trinucleotide at that position, while for EGR1, T<u>C</u>C⟶T<u>G</u>C is the most elevated change, while not the most frequent trinucleotide at the position (Figure 4.4c).

The extent of the contacts that those proteins make with DNA are largely known and represented in Figures 4.5 and 4.6. Position number 5 within the KLF4 motif makes direct hydrogen bond and van der Waals contact with the second zinc finger of KLF4 protein (Figure 4.5). Thymine (or cytosine) at the position number 5, rather than adenine (or guanine) on opposing strand, appears to be important for making the hydrogen bond, while van der Waals contact is made by the adenine/guanine. If there were to be a strand asymmetry in protein tolerance for a lesion at position number 5, one might expect modified adenine/guanine to be tolerated better than altered thymine/cytosine. The excess of T⟶G mutations would therefore be expected to arise from occlusion of, for example, mismatch involving A⟶C change. In case of EGR1, level of preference towards thymine or adenine at position number 6 of the motif is similar, but lesion on the C-rich strand would be expected to be tolerated better (Figure 4.6).

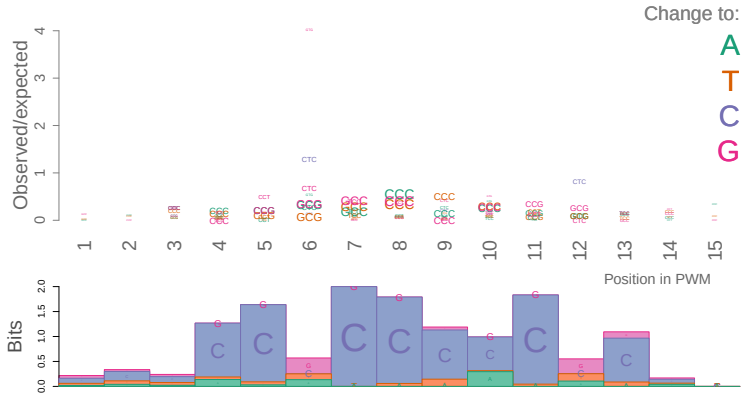
Figure 4.7 shows the pan-cancer mutation pattern of several other TFs analysed. All of those, with the exception of FOXA1 and GABPA, are C2H2 ZnF-containing proteins and have a certain position within their binding motifs that appear to exhibit



(a) CTCF trinucleotide mutational spectrum



(b) KLF4 trinucleotide mutational spectrum



(c) EGR1 trinucleotide mutational spectrum

Figure 4.4: Trinucleotide mutational patterns TF over motifs. Each triplet on the plot represents the middle nucleotide change in a particular trinucleotide context. The size of the triplet is proportional to the relative abundance of that trinucleotide at the position in question. Only triplets that constitute at lest 10% of all trinucleotides at a position have been plotted.

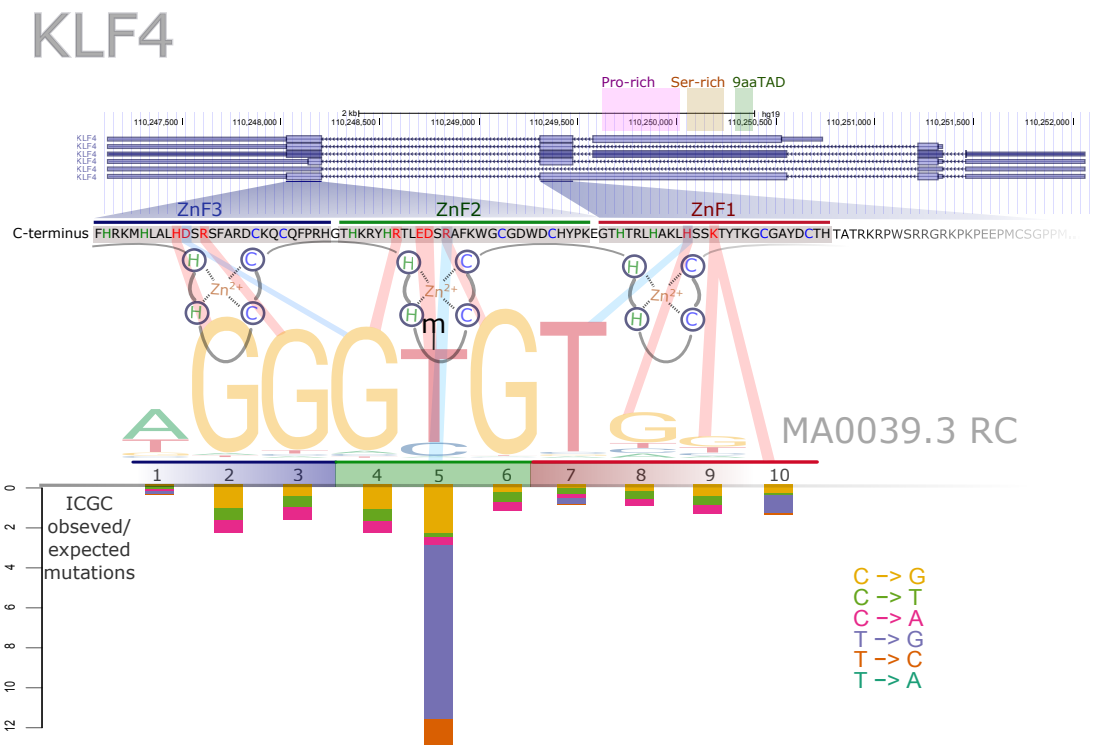


Figure 4.5: KLF4 protein Ref-seq transcripts from UCSC (top) and zoom in at the amino acid sequence that forms zinc fingers. Positions within amino acid sequence important for DNA contacts are indicated in red, and semi-transparent lines represent contacts made with motif sequence (red lines – direct hydrogen bonds, blue lines – van der Waals contacts). Connection of the semi-transparent line with the top of base letter means that contact occurs with nucleotide on represented strand, while connection on the bottom means contact with the complement base. Below motif are mutational profiles KLF4 binding sites as measured by single nucleotide substitutions from ICGC cohort (same as Figure 4.3b). Principal source Hashimoto et al. (2016)

an excess of mutations similar to CTCF, KLF4, and EGR1, suggesting that the observed phenomenon of highly mutated positions within a motif is likely be a general feature of DNA-binding TFs, and might extend beyond the ZnF family.

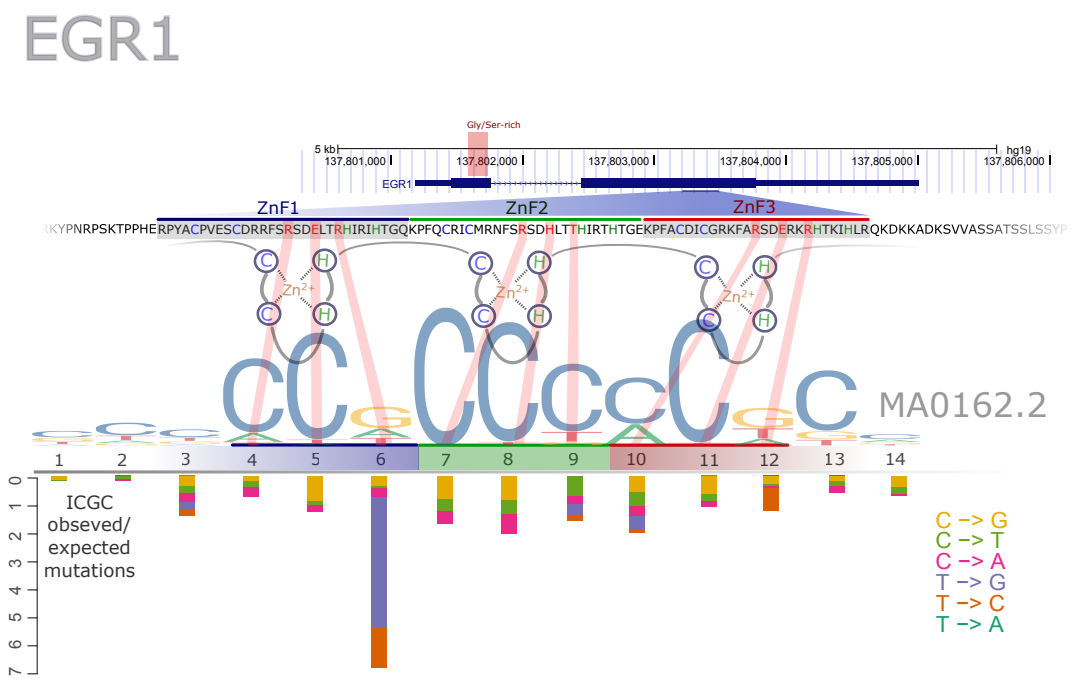
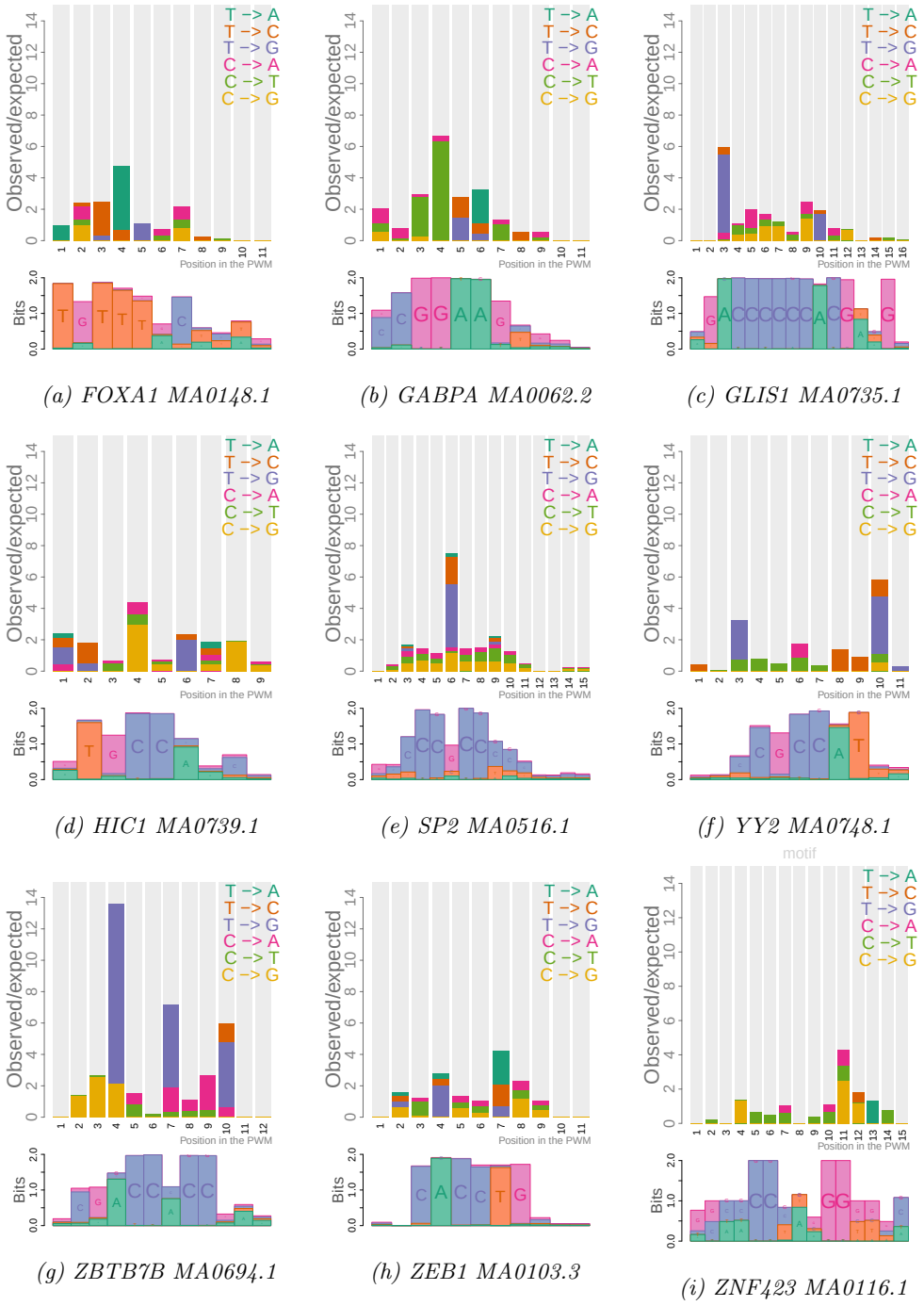


Figure 4.6: EGR1 protein Ref-seq transcripts from UCSC (top) and zoom in at the amino acid sequence that forms zinc fingers. Positions within amino acid sequence important for DNA contacts are indicated in red, and semi-transparent lines represent contacts made with motif sequence (red lines – direct hydrogen bonds, blue lines – van der Waals contacts). Connection of the semi-transparent line with the top of base letter means that contact occurs with nucleotide on represented strand, while connection on the bottom means contact with the complement base. Below motif are mutational profiles EGR1 binding sites as measured by single nucleotide substitutions from ICGC cohort (same as Figure 4.3c). Principal source Hashimoto et al. (2014)



(a) FOXA1 MA0148.1 *(b) GABPA MA0062.2* *(c) GLIS1 MA0735.1*

(d) HIC1 MA0739.1 *(e) SP2 MA0516.1* *(f) YY2 MA0748.1*

(g) ZBTB7B MA0694.1 *(h) ZEB1 MA0103.3* *(i) ZNF423 MA0116.1*

Figure 4.7: Excess of mutations over expectation (based on the trinucleotide context mutability genome-wide) from the ICGC (pan-cancer) for several TFs binding sites where both motifs and ChIP-seq were available. These TF motifs also ascribe to similar property as TF motifs for CTCF, KLF4 and EGR1, with an important positions within binding motifs exhibiting excess of substitutions.

4.3.3 Mutational patterns at binding motifs vary across cancer types

While pan-cancer analysis is beneficial in a sense that it provides us with great statistical power, at the same time it contains data from multiple different cancers, each with their individual spectra contributing to the observed variation. Each type of cancer would be expected to be driven, or to exhibit, a different combination of predominant mutational processes (Helleday et al., 2014). While the patterns of the observed/expected rate plotted here are meant to account for the mutational spectra dominating each particular cancer type/tumour, the excess of mutations that we observe could be shaped by processes that do not leave genome-wide signature, but might still be limited to one type or subset of cancers. Hence, I have conducted similar analysis of TF motif mutation rates, but for each of the available cancer types (Table 4.1) separately for all binding TFs listed in Table 4.2 and show results for the KLF4 and CTCF binding sites in Figures 4.8 and 4.9.

Figures 4.8 and 4.9 show patterns over KLF4 and CTCF binding sites, respectively, which can be observed in the subset of cancers from ICGC with the highest numbers of mutations. While most of the cancer types seem to exhibit an increase in the mutation rate at positions numbers 5 (KLF4) and 9 (CTCF), they do so to a varying degree and differ between the two TFs. Skin adenocarcinoma (SKCA), lung cancer (LUSC), ovarian cancer (OV), paediatric brain cancer (PBCA), early onset prostate cancer (EOPC), chronic lymphocytic leukaemia (CLLE) and pancreatic cancer endocrine neoplasm (PAEN) appear to exhibit the highest excess of KLF4 mutations, while esophageal adenocarcinoma (ESAD), breast cancer (BRCA), and liver cancer (LIRI and LICA) show highest peaks within CTCF motif.

The T$\longrightarrow$G change does not seem to be a particularly abundant type of change in any of the cancers genome-wide (Figure 4.10 shows the relative abundances of substitution types across all the cancer datasets within ICGC). Therefore this particular change is unlikely to arise purely due to the specific cancer-driving process.

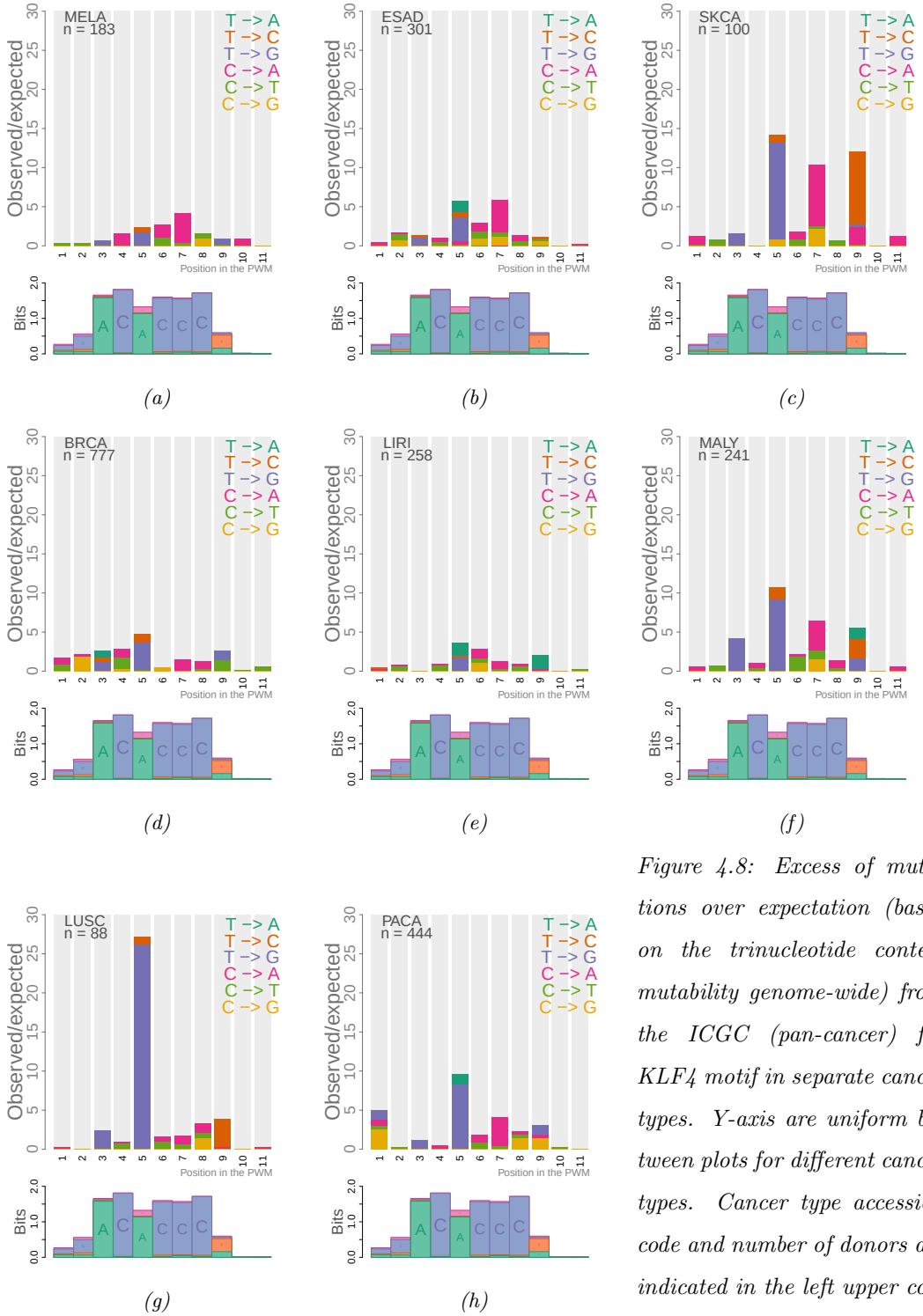


Figure 4.8: Excess of mutations over expectation (based on the trinucleotide context mutability genome-wide) from the ICGC (pan-cancer) for KLF4 motif in separate cancer types. Y-axis are uniform between plots for different cancer types. Cancer type accession code and number of donors are indicated in the left upper corner.

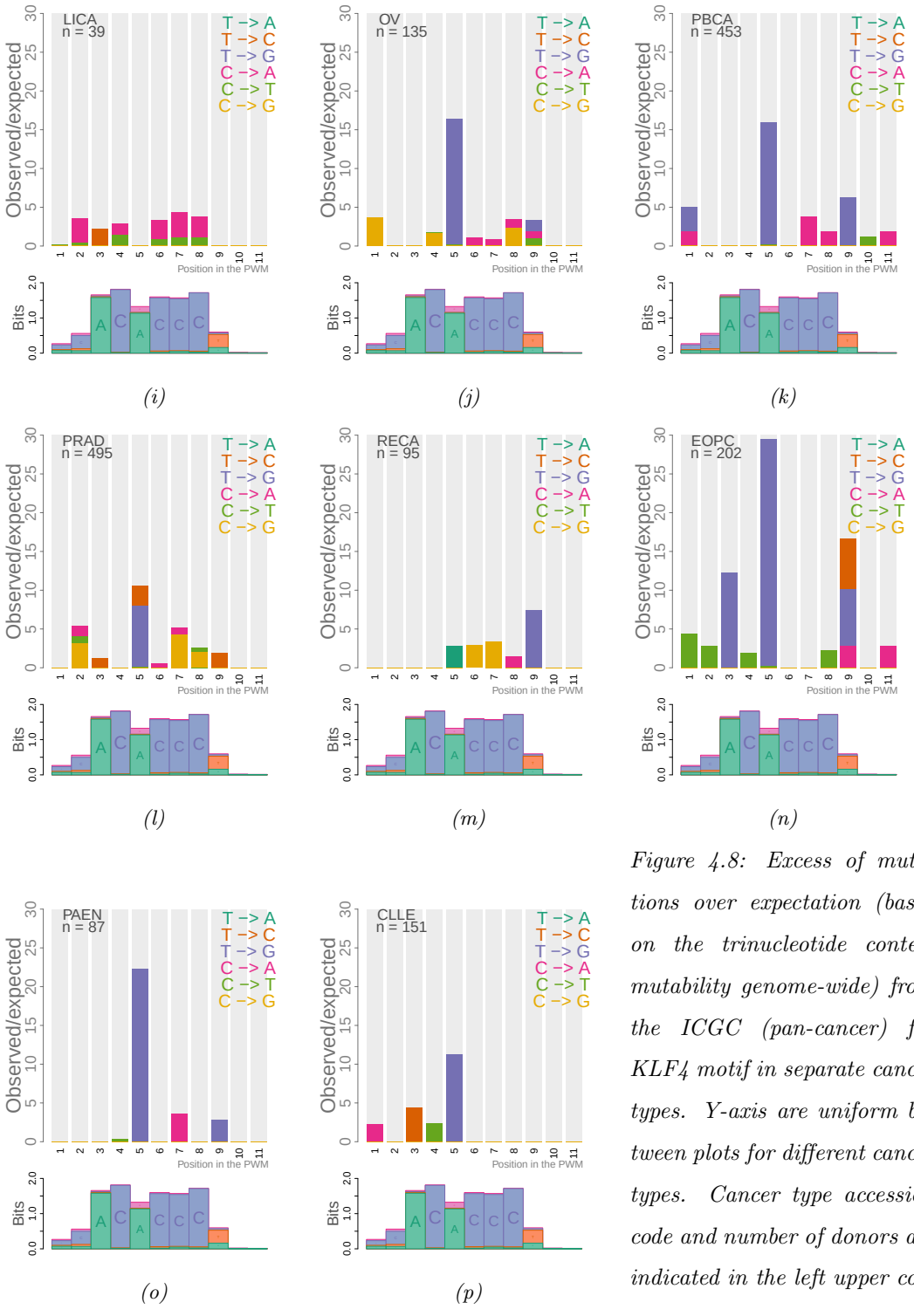


Figure 4.8: Excess of mutations over expectation (based on the trinucleotide context mutability genome-wide) from the ICGC (pan-cancer) for KLF4 motif in separate cancer types. Y-axis are uniform between plots for different cancer types. Cancer type accession code and number of donors are indicated in the left upper corner.

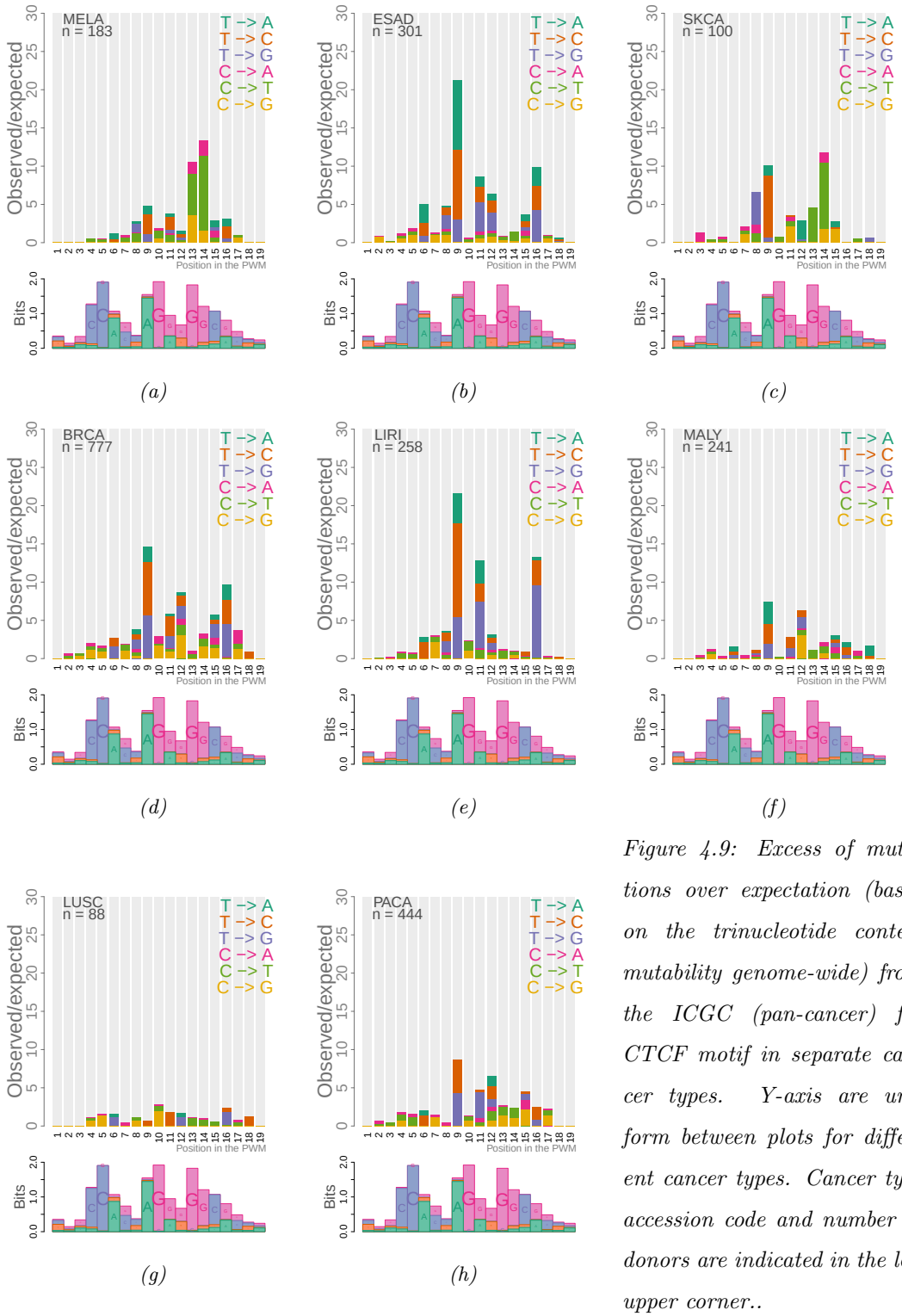


Figure 4.9: Excess of mutations over expectation (based on the trinucleotide context mutability genome-wide) from the ICGC (pan-cancer) for CTCF motif in separate cancer types. Y-axis are uniform between plots for different cancer types. Cancer type accession code and number of donors are indicated in the left upper corner..

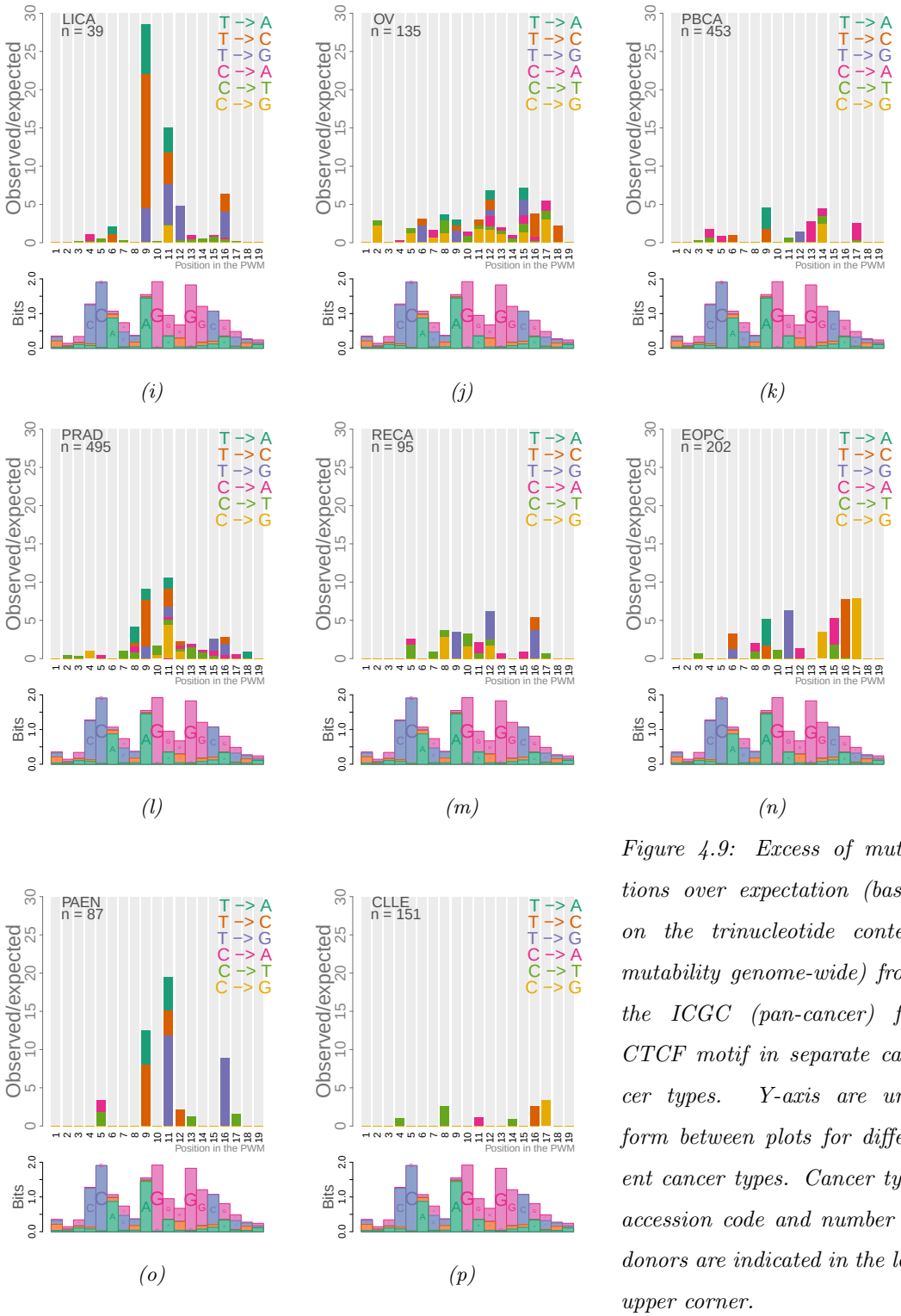


Figure 4.9: Excess of mutations over expectation (based on the trinucleotide context mutability genome-wide) from the ICGC (pan-cancer) for CTCF motif in separate cancer types. Y-axis are uniform between plots for different cancer types. Cancer type accession code and number of donors are indicated in the left upper corner.

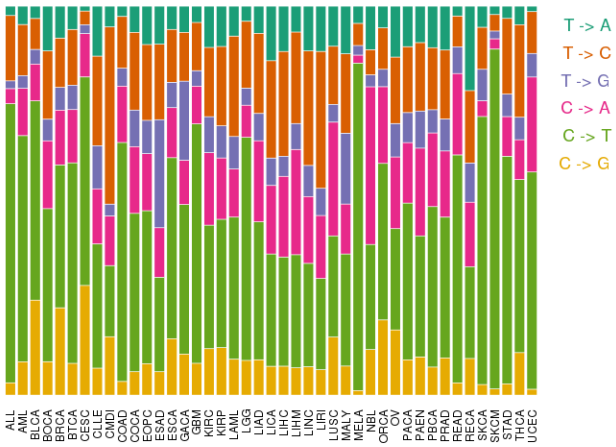


Figure 4.10: Relative proportions of different types of substitutions for various cancer types from ICGC

4.3.4 Increased mutation rates within transcription factor motifs are consistent with protection from mismatch repair

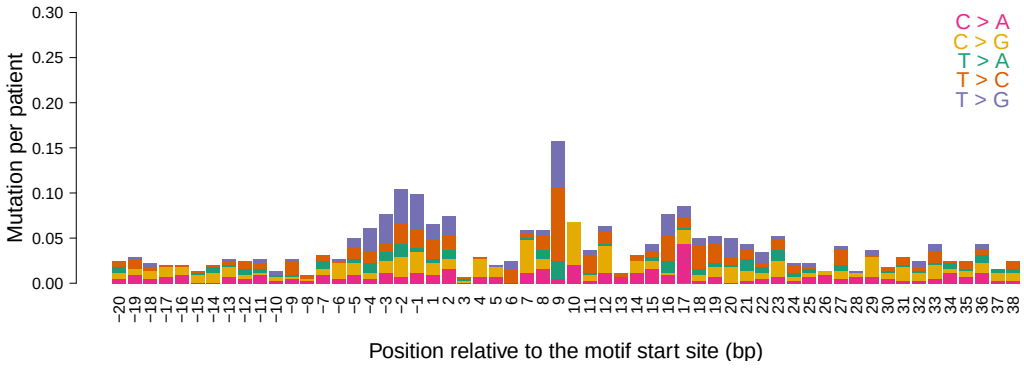
Here I propose that the increase in the mutation rates at certain positions within the protein-binding motifs can be attributed to the protein binding interfering with normal DNA surveillance and repair processes, primarily by protecting the underlying sequence from processes such as mismatch repair (MMR). As described in the Section 4.1, MMR plays crucial part during replication to perform a significant role in to preserving fidelity of newly synthesised DNA. Loss or reduction of the MMR activity leads to the miscrosatellite instability phenotype (MSI), as many microsatellite regions are primarily repaired by MMR and in its absence exhibit high rates of short indels. In addition to high numbers of indels at the microsatellite regions, MMR-deficient (MMRd) cells also exhibit high levels of single nucleotide substitutions.

Higher mutation rates at the positions numbers 9 and 5 within binding motifs for CTCF and KLF4, respectively, would then be a result of lower levels of MMR activity at those positions in comparison to the flanking sequences that are not occupied by proteins. In this case, one would expect to see equally high numbers of mutations in the flanks, as in the site, when looking at the mutations from the tumours that are MMRd.

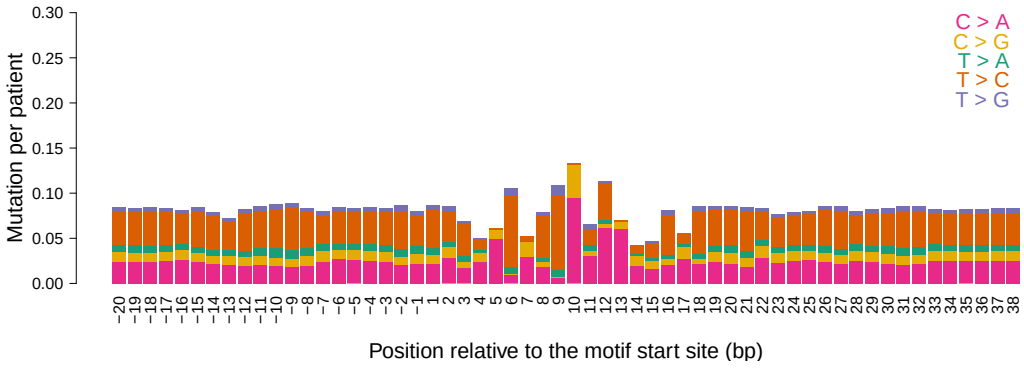
Sizes of the datasets containing single nucleotide substitution data from the MMRd tumours are not large enough to be able to see if that is actually the case across a relatively small subset of protein-binding sites for a specific TF. One can, however, model the *expected* pattern of substitutions at the aggregate of protein-binding motifs just from the observed genome-wide trinucleotide mutational frequencies of the MMRd tumours. That allows one to see if the *observed* levels and patterns of mutations at the candidate regions that we *propose* are deprived of MMR match what we *expect* to see in cells that we *know* are deprived of MMR.

Breast cancer (BRCA) is one of the tumour types where I can see increase in mutations at position number 9 of the CTCF motif, and also have single nucleotide substitution data for MMRd tumours. Figure 4.11a shows the patterns that we *observe* at the aggregate of CTCF binding sites (with MMR intact) next to the pattern that we *expect* to see at the same set of sites in the MMR-deficient tumours given the sufficiently powered dataset (Figure 4.11b). First, one can see that, as expected, the overall mutation rate is increased in the MMRd tumours. The level of increase with MMRd appears to match the level observed at the highly mutated positions in cells with intact MMR, and while types of changes appear to be somewhat different, this is consistent with the observed pattern being shaped by protection from MMR. The levels of mutations in flanks are of the same magnitude to what one would expect from tumours with MMR intact (Figure 4.11c). It is worth to note that breast cancers are rarely subject to MMR deficiency. The MMRd status of the tumours here was inferred from the large numbers of the indels (as in Supek and Lehner (2017)), however breast cancers can be subject to other repair defects that might produce large numbers of indels. Therefore some fo these samples might not in fact be MMRd.

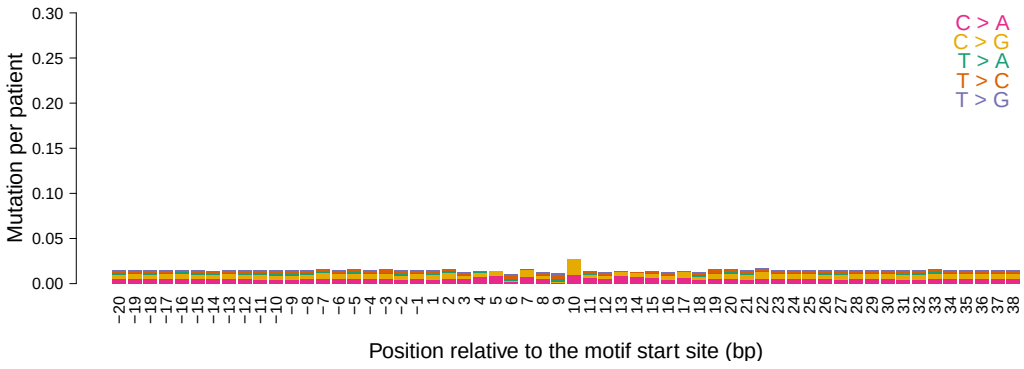
Figures 4.12 and 4.13 shows similar plots for CTCF and KLF4, using pan-cancer ICGC MMR intact tumours and MMRd stomach cancer tumours from Wang et al. (2014). I used pan-cancer ICGC data here, as the numbers of mutations in the cancer types that are similar to stomach cancer are too few to be able to observe the excess of mutations within the binding site. Even though the MMRd and non-MMRd cancer types are not matched here, it appears that the level of mutations within the binding sites is again of the same magnitude to what might be expected from the MMRd tumour, albeit the mutational processes will likely differ and only specific types of lesions might be tolerated by the TF.



(a) BRCA observed mutation pattern over CTCF sites

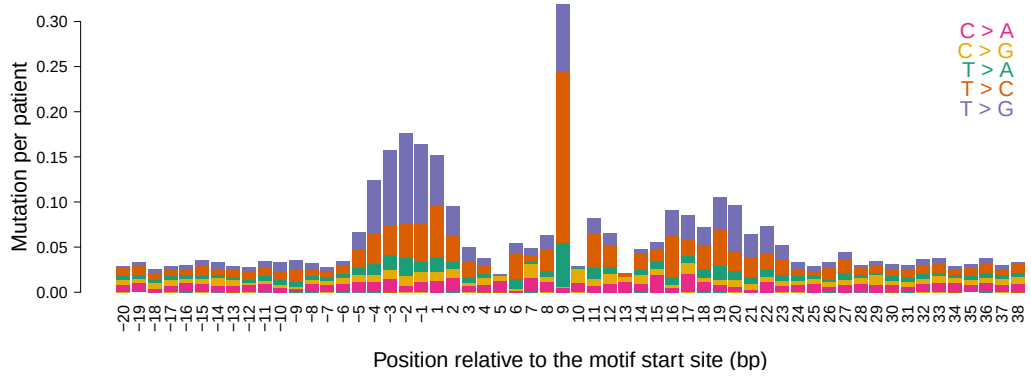


(b) Expected mutations over CTCF binding sites applying genome-wide rates of MMRd BRCA

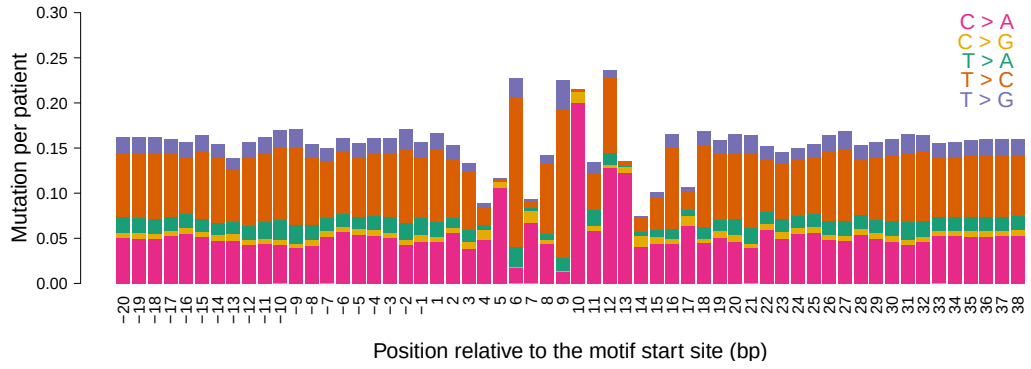


(c) BRCA expected mutation pattern given genome-wide rates of non-MMRd tumours

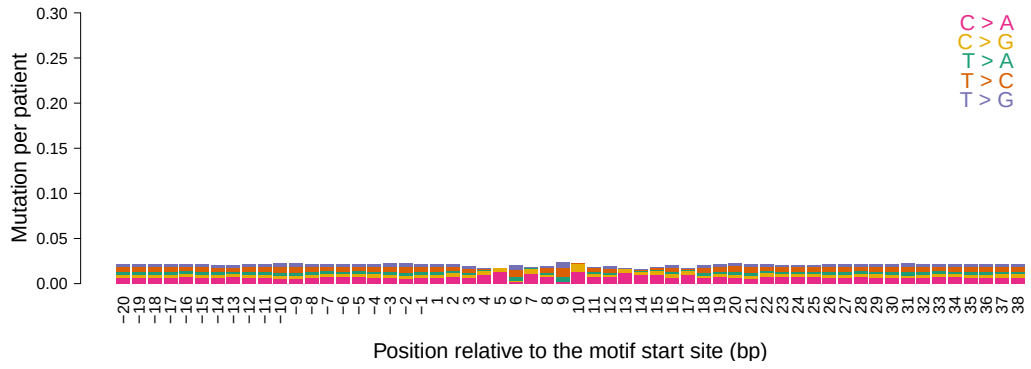
Figure 4.11: Observed (a) and expected (c) BRCA (breast cancer) mutations over the CTCF binding motif (MA0139.1) in comparison with the expected mutational pattern over the same sites based on genome-wide trinucleotide mutational frequencies in BRCA MMRd tumours (b). Level of mutations observed at the position 9 of the motif (a) is similar to expected level of mutations when MMR is suppressed (b).



(a) CTCF; observed in ICGC pan-cancer

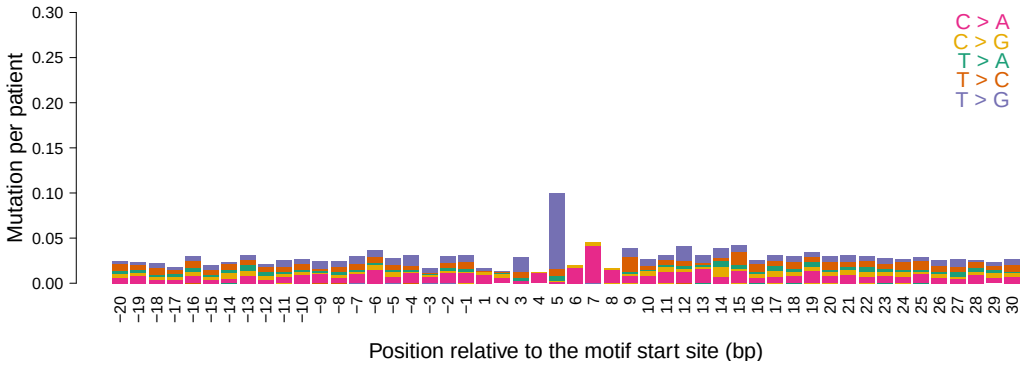


(b) CTCF; expected in MMRd stomach cancer

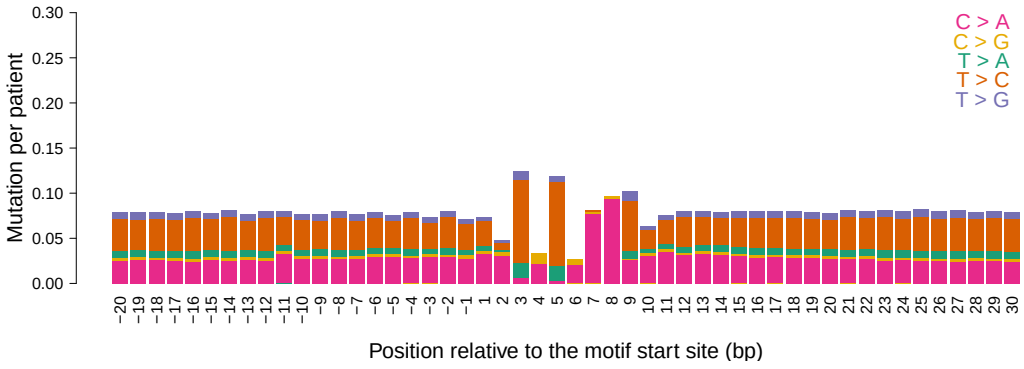


(c) CTCF; expected from ICGC pan-cancer

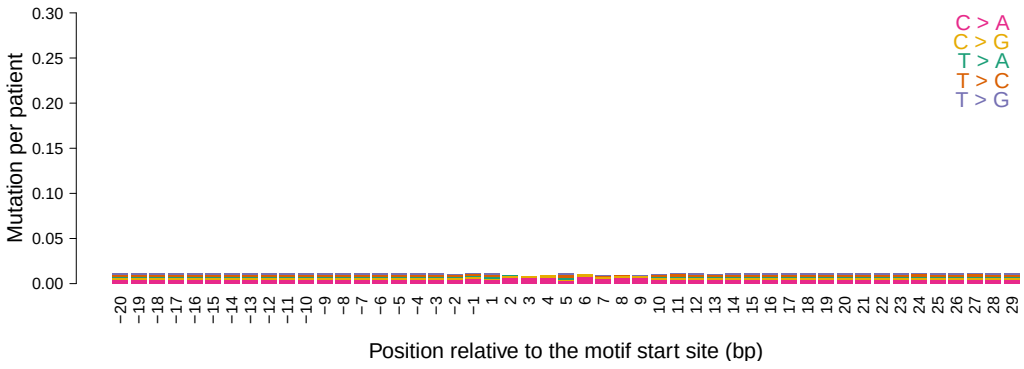
Figure 4.12: Observed (a) and expected (c) pan-cancer (from ICGC) mutations over the CTCF binding motif (MA0139.1) in comparison with the expected mutational pattern over the same sites based on genome-wide trinucleotide mutational frequencies in stomach cancer MMRd tumours (b). Level of mutations observed at the position 9 of the motif (a) is similar to expected level of mutations when MMR is suppressed (b).



(a) KLF4; observed in ICGC pan-cancer



(b) KLF4; expected in MMRd stomach cancer



(c) KLF4; expected from ICGC pan-cancer

Figure 4.13: Observed (a) and expected (c) pan-cancer (from ICGC) mutations over the KLF4 binding motif (MA0139.1) in comparison with the expected mutational pattern over the same sites based on genome-wide trinucleotide mutational frequencies in stomach cancer MMRd tumours (b). Level of mutations observed at the position 9 of the motif (a) is similar to expected level of mutations when MMR is suppressed (b).

4.3.5 Analysis of mutational frequencies beyond trinucleotide sequence can reveal highly mutated motifs

Investigation of unusually highly mutated positions within the protein binding motifs requires several types of input information - such as motif PWM for a specific TFs, and preferably ChIP-seq or an equivalent type of data to know which instances of the motif are bound by the TF genome-wide. That limits the possibility of such analysis only to the subset of reasonably well-studied TFs for which such information is available, and also can make an integration of such inputs from various sources labour-intensive. Limited availability of ChIP-seq data from cell matched with mutation origin leads to the lack of knowledge of whether the binding site is actually active in that particular cell type. Therefore, I prototyped novel approaches to extend this analysis to a more general, genome-wide level and escape the need for the additional inputs.

Mutation rate of a particular position is determined in big part by its trinucleotide context (*i.e.* by the identity of the surrounding base on either side) (Blake et al., 1992), but has also been extended to longer tracks of sequences (penta- and septa-nucleotides) serving as predictors for mutation rate (Aggarwala and Voight, 2016). In the motif based analyses already presented (Subsections 4.3.3 and 4.3.4) I considered the mutation rate of trinucleotides within a specific motif compared to the genome wide rate for those motifs calculated from the same cancer samples. Here I generalise that approach by considering the mutation rate of each penta-nucleotide relative to the expectation for the central tri-nucleotide. The rationalisation being that generally the petanucleotide rate will be well estimated by the central trinucleotide rate. Where it is not, it indicates the wider sequence context contributes to defining outlier mutation rate properties, for example by being part of a wider sequence specific binding motif. Figure 4.14 shows the natural log ratio of the observed mutation rate of each pentanucleotide to rate of middle trinucleotide. Any pentanucleotides that are above 1 on the y-axis are mutated more frequently genome-wide than expected, while the ones below - less frequently, indicating that this difference is driven by the identity of first and last positions. There are some outliers which exhibit mutation rate which is strikingly different (increased) from expectation based on the context of middle trinucleotide, such as TT$\underline{\text{T}}$AA$\longrightarrow$TT$\underline{\text{A}}$AA, TC$\underline{\text{C}}$TT$\longrightarrow$TC$\underline{\text{T}}$TT, and TT$\underline{\text{T}}$CC$\longrightarrow$TT$\underline{\text{C}}$CC, which resemble mutations occurring in repetitive sequences, but also elevated rate of such changes

as TCCAT⟶TCTAT and GGTTG⟶GGCTG. Interestingly, GGTGN⟶GGGGN changes, which correspond to the increased mutation rate at position 5 of KLF4 motif, all show mutation rate above trinucleotide-based expectation across the genome. This genome-wide approach is unlikely to reveal the mutational pattern shaped by the interactions of the TFs with specific sequences, unless the TF can be assumed to be bound universally across the genome.

A more informative approach is to compare the mutational frequencies of the pentanucleotides between two types of genomic states - such as accessible/active and non-accessible regions of the genome. Variation in the mutation rate of a particular pentanucleotide between what is observed genome-wide versus what is observed in the region of interest can then be attributed to the nature of the region. Figure 4.15 shows the natural log ratios of substitution rates in the subset of binding sites that were found to be commonly occupied in multiple tissues by analysis of ATAC-seq data, as described in Chapter 2, to the rates that are observed in the rest of the genome. Any pentanucleotides that are above 0 on the y-axis are found to be more mutated in the regions defined as binding sites commonly occupied in multiple tissues, while the ones below are more mutated in the rest of the genome. Figure 4.16 shows similar data, but as mutation rates correlation plot.

C⟶T change in the CpG context is unsurprisingly the most pronounced mutational pattern exhibiting the highest mutation rate in the genome-wide context. All of the CpG-containing pentanucleotides, however, cluster around 1 on y-axis and closer to left side of the plot, with not much spread (except for CCG trinucleotide context), indicating that deamination of the methylated CpG is likely to be such an influential driver of C⟶T change that any of the surrounding sequence does not add much effect. It also shows a large change in the mutation rate in the sites that I have identified as *common* binding sites *versus* rest of the genome, clustering into a separate group (Figures 4.14 and 4.16 ⟶T change). This is an expected observations, as regulatory sites tend to exhibit lower levels of methylation, as discussed in Chapter 3.

Interestingly, the sequence that resembles the KLF4 motif (GGTGG/T) is exhibiting the largest fold change to G between the *common* binding sites and the rest of the genome, which follows my previous observation for KLF4 binding sites (Figure 4.3b).

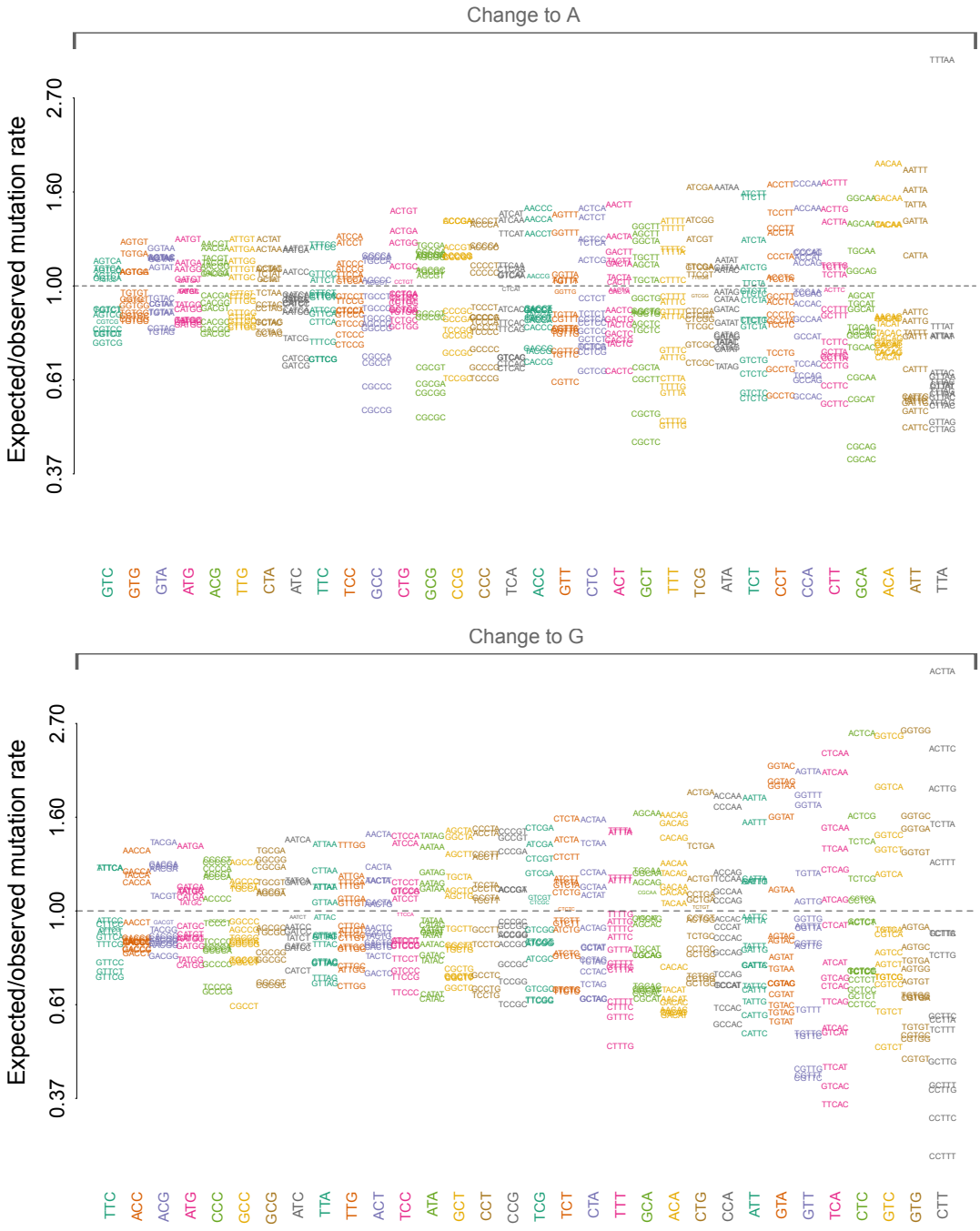


Figure 4.14: Natural log of ratios of pentanucleotide pan-cancer mutation rates to trinucleotide mutation rates genome-wide, where the central base is mutated to an A (top) or a G (bottom). Note logarithmic y-axis.

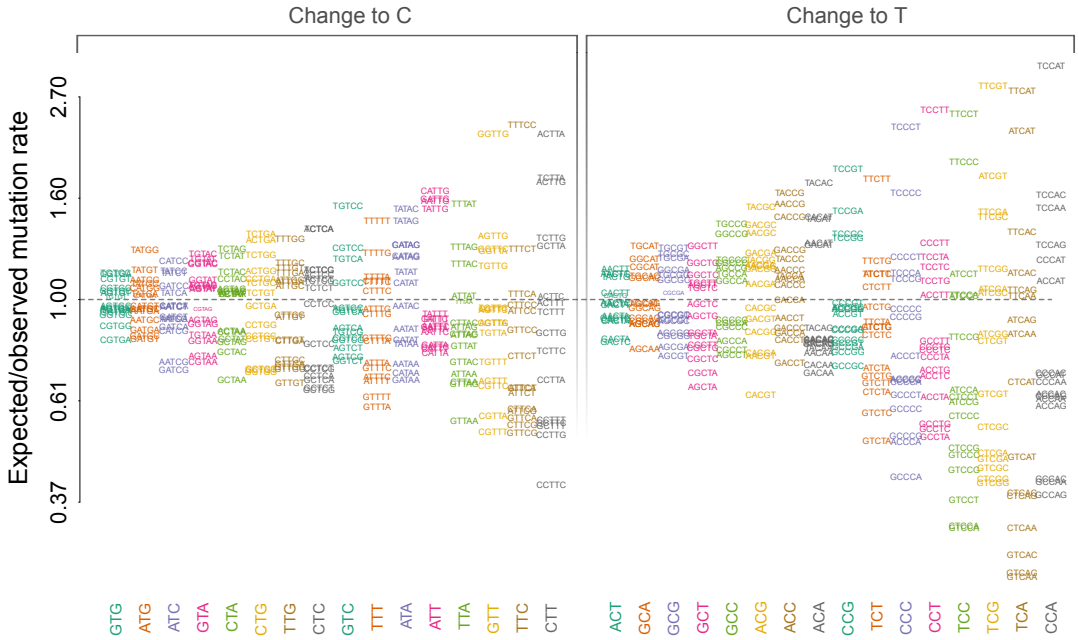


Figure 4.14: Natural log of ratios of pentanucleotide pan-cancer mutation rates to trinucleotide mutation rates genome-wide, where the central base is mutated to an C (left) or a T (right) (continued). Note logarithmic y-axis.

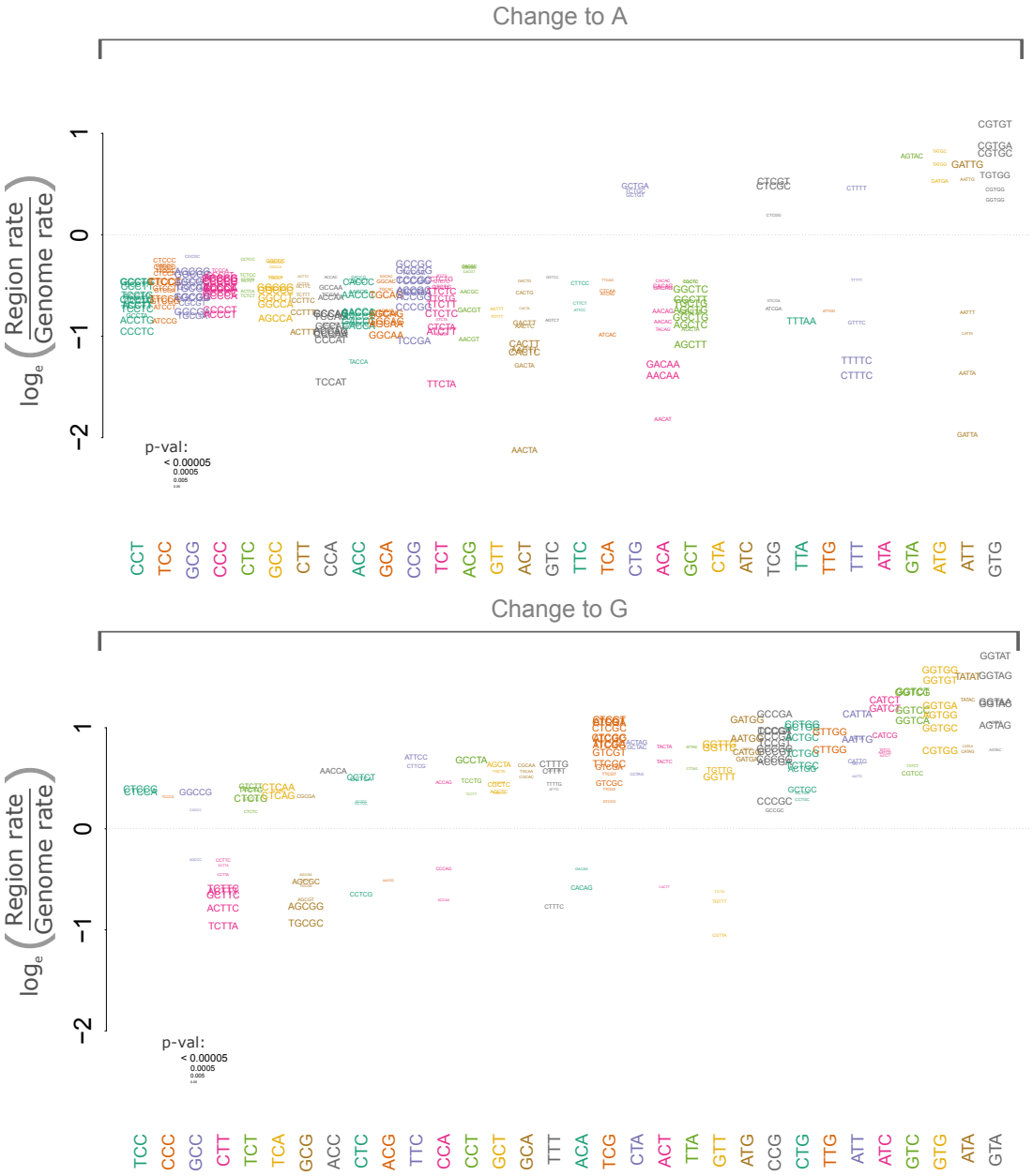


Figure 4.15: Ratios of pentanucleotide pan-cancer mutation rates in the common binding sites to the rest of the genome, where the central base is mutated to an A (top) or a G (bottom). Size of each pentanucleotide corresponds to the $-\log_{10}$ p-value of the Fishers exact test, and each pentanucleotide group is ordered according to the level of the mutation rate spread within the group. Note logarithmic y-axis.

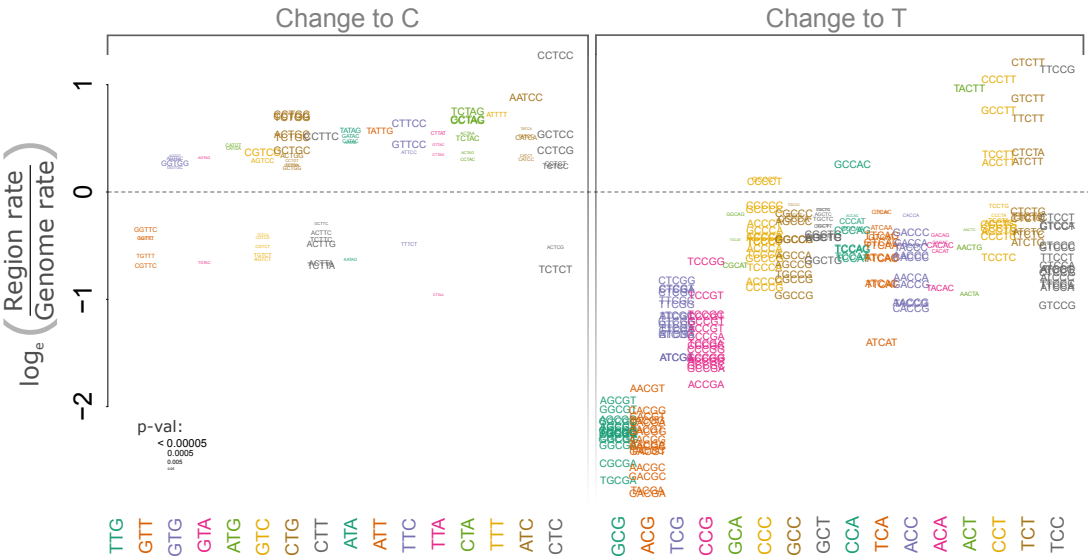


Figure 4.15: Ratios of pentanucleotide pan-cancer mutation rates in the common binding sites to the rest of the genome, where the central base is mutated to an C (left) or a T (right) (continued). Size of each pentanucleotide corresponds to the $-\log_{10}$ p-value of the Fishers exact test, and each pentanucleotide group is ordered according to the level of the mutation rate spread within the group. Note logarithmic y-axis.

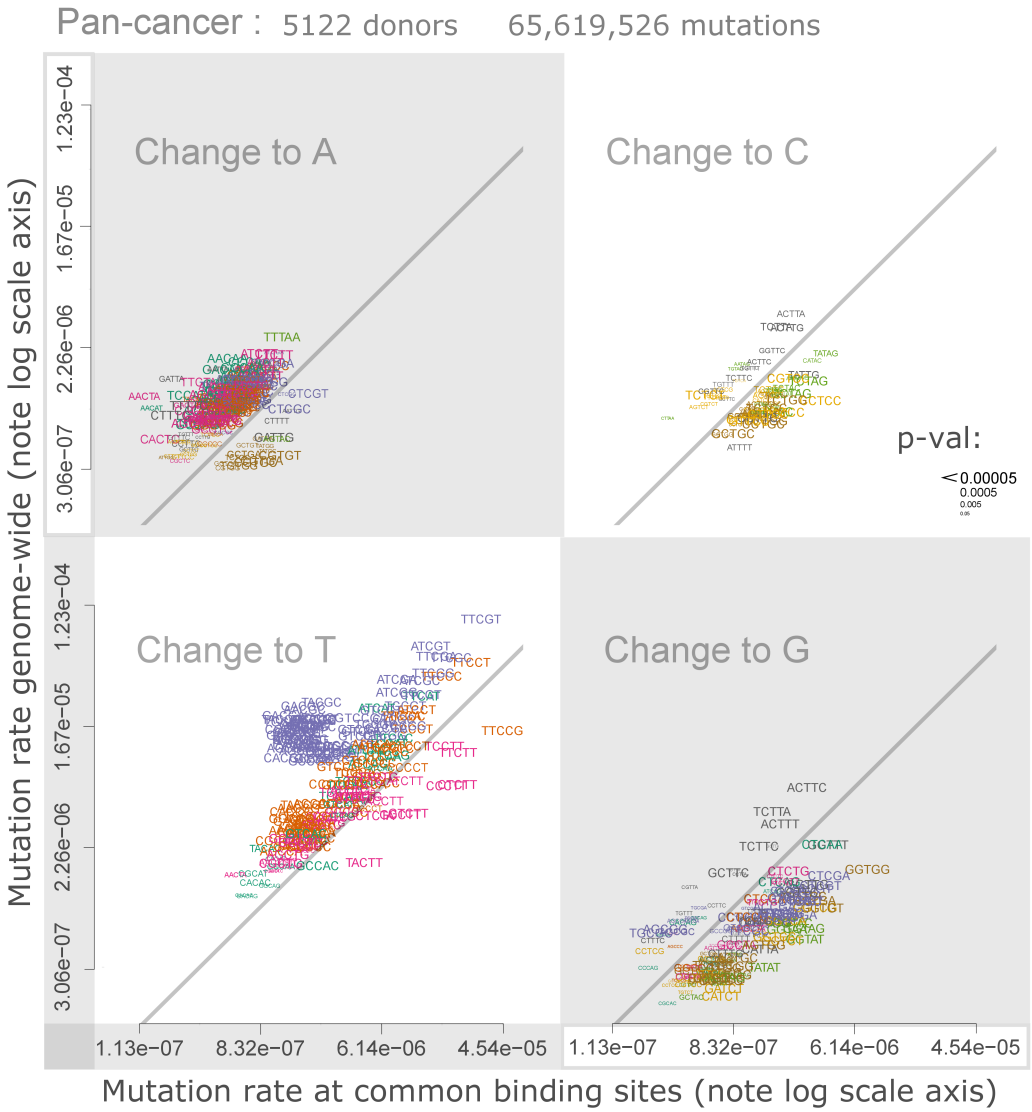


Figure 4.16: Petanucleotide pan-cancer mutation rates at common binding sites compared to those occurring across the rest of the genome. The size of each pentanucleotide corresponds to the $-\log_{10}$ *p-value of the Fishers exact test.*

4.4 Discussion

In this Chapter, I investigated the patterns of the mutations at the motifs of the specific binding factors and introduced a *biased mask model* that could explain those observations. I show that the striking pattern of cancer mutation abundance at the positions in the CTCF motif that is important for binding is shared by wider range of TFs, such as KLF4, EGR1 and others.

The unusual spike in mutations observed at position 9 of the CTCF motif here corresponds with what has been previously observed by Kaiser et al. (2016) in pan-cancer analysis and by Katainen et al. (2015) in colorectal cancer. Katainen et al. (2015) has also shown that the binding of CTCF is necessary for induction of this mutational increase, consistent with our hypothesis of protein binding causing retention of mutations. There is a peak of mutation excess at the position number 5 of the KLF4 and position number 6 of the EGR1 motif, which are both C2H2 ZnF binding TFs, like CTCF. KLF4 shows increase in mutation rate of G$\underline{\text{T}}$G triplet, which is the most abundant trinucleotide at that position, to potentially motif-disrupting G$\underline{\text{G}}$G. The predominant mutation peak within the EGR1 binding motif occurs at the position that does not carry substantial amount of importance. Change occurs in the similar G$\underline{\text{T}}$G $\longrightarrow$ G$\underline{\text{G}}$G context as with KLF4 , but is not the most frequent trinucleotide at this position for EGR1. The mutation spikes in the pan-cancer analysis are not dominated by a particular cancer type, but some types of tumours do tend to show a more distinct pattern than others, and those differ between CTCF and KLF4. Variability in the mutational pattern that can be seen across the cancers is not all that surprising, different types of tumours are known to be driven or to exhibit distinct mutational processes and signatures. This just highlights the level of complexity that we are faced with when trying to disentangle contributions of different mutational mechanisms in various contexts. The characteristic T$\longrightarrow$G change dominating at KLF4 and EGR1 motifs does not appear to be a type of change that could be attributed to a particular cancer type or mutational process. It might, therefore, represent a type of change that is preferentially retained specifically at those sites, in accordance with the *biased mask* model.

Analysis of the bulk of binding sites of other TFs, for which PWMs and ChIP-

seq were available showed that there are more instances of similar mutational patterns for both C2H2 ZnF and other types of TFs (Figure 4.7).

As I expect this to be more of a global feature of many of the binding sites, I have implemented a more general, genome-wide approach for finding sequences that exhibit mutational rate distinct from what would be expected. I trialed the analysis of sequences limited to 5-nucleotide length and compared their mutation rates between sites bound by the protein and the rest of the genome. However, this type of approach could be taken further to investigate longer sequence lengths (presuming sufficiently powered datasets with mutations are available), and to test differences between different types of genomic regions. Typical length of the protein binding motif is $\approx 6 - 12$. Conceivably, by taking a set of sequences of approximate protein-binding motif length, say 11 bases, with every possible combination of nucleotides (4,194,304 possible combinations), and counting the frequencies of substitutions of every position to each alternate nucleotide (138,412,032 possible types of changes), and comparing rates of those between accessible and non-accessible regions, one could find outliers that could reveal the highly mutated positions within potential motifs. This is currently now being developed as a follow-on project in the group.

One of the things to note is that the calculations of the genome-wide mutation rates (both trinucleotides and pentanucleotides) are based on the whole of the genome. At the same time there might be certain regions of the genome where one is unlikely to find any cancer somatic mutations, not due to the biological reason, but rather because of the technical bias in the way that ICGC data (or somatic mutation data from other cohorts) has been processed. Therefore, I suspect that the genome-wide mutation rates that I calculate here might be underestimated. That would mean that the fold-enrichment in the mutation rates for some, or all, of the positions and changes would be lower.

It would be of a great advantage if we could associate the magnitude of the mutational effect within the motifs with the levels of the protein expression in the same cells. That would be possible to do providing that there is WGS mutation data and RNA-seq data available from the same donors. Within the ICGC cohort, there are not enough samples for different types of tumours where both WGS mutation and RNA-seq data is available. In addition, in relation to the KLF4 motif, there are multiple members of the KLF family that bind similar motifs and those could be expressed at variable

levels on different types of cells. That would then make it difficult to disentangle the mutational effect contributed by each of the KLF family members.

I hypothesise that the abundance of mutations within protein binding motifs in excess of what can be observed genome-wide is due to the TFs interfering with the normal processes of replication and repair, in particular MMR. One of the ways to test this is to compare the mutational pattern of tumours that have intact mismatch repair with those that are MMRd. While mismatch deficiency does mean that any individual tumour is going to harbour a larger number of mutations, MMRd tumours occur relatively rarely. In the absence of a dataset with enough mutations from the MMRd tumours to be able to observe the mutational patterns at the relatively small number of binding sites, I instead looked at the mutational patterns that we would *expect* to see at the motifs based on the mutational spectrum of tumours lacking mismatch repair. I find that the level of mutation excess at the highly mutated positions within KLF4 and CTCF motifs match the expectation from MMRd tumours. Relative proportions of types of mutations that we expect to observe over the motif with MMRd differ from the proportions we observe at the highly mutated positions. This can be explained by only certain types of lesions being tolerated by the TFs, and therefore over-representation of a particular types of changes.

CHAPTER 5

Strand-specific tolerance to mismatches by the KLF4 transcription factor

5.1 Introduction

5.1.1 Viability of the 'biased mask' model in the context of mismatch lesions

In the previous Chapter, I introduced the *'biased mask'* model that predicts that some positions within the motifs of some of the sequence-specific binding TFs might become highly mutated due to strand-specific tolerance to mismatches and other lesions. In the current Chapter, I will describe the experimental *in vitro* validation of this model, using KLF4 protein as an example, which I have shown before to exhibit an unusually high mutation rate at position number 5 within its binding motif (Figure 4.3b).

There are multiple types of lesions, whose occurrence and failure to repair could result in a mutated sequence following a single round of replication. Those include abasic sites, modified bases (such as thymine dimers), interstrand crosslinks, and mismatches (Chatterjee and Walker, 2017). Mismatch repair, together with the exonuclease activity of replication polymerases, is responsible for fixing wrongly incorporated bases during replication. Failure of the mismatch repair machinery to do that would result in the formation of mismatches - duplex DNA sequences, where non-complementary bases are paired opposite each other. Unfixed, such mis-pairing will lead to mutation in one of the daughter cells after the next round of replication.

In the previous Chapter, I showed that an unusually high level of mutations at some positions within the KLF4 and CTCF motifs is consistent with a lack of protection from mismatch repair. Therefore, the analysis presented here involves testing the ability of the KLF4 protein to bind over sequences containing mismatches, although I expect that this model is not restricted to this particular type of lesion and is likely to extend to a wider variety of DNA base changes. To my knowledge, there have not been any studies asking whether sequence-specific binding TFs are able to bind over a mismatch lesion. Closest studies have investigated the role of naturally occurring DNA base, such as methylation influencing TF binding affinity in site and strand-specific manner (Hashimoto et al., 2016, 2017).

5.1.2 Methods for measuring affinity of a protein to target DNA sequence

It is possible to test how well protein is able to bind a specific DNA sequence with a range of different assays using synthetic oligonucleotide duplexes. These oligonucleotides can be artificially synthesised with fluorophores attached to the 5' end of the desired sequence. Binding affinity estimation methods involve mixing protein and target oligonucleotide duplexes in solution in the presence of other necessary components, such as zinc in case of ZnF proteins.

Detection and estimation of the bound and unbound fractions of oligonucleotide duplexes differ between assays. One of the methods is electrophoretic mobility shift assay (EMSA) (Hellman and Fried, 2007), which involves loading the binding reaction onto a polyacrylamide gel and passing a voltage to separate species by size. By using fluorescently-labelled oligonucleotides, it is possible to detect where on the gel they have migrated to. That allows one to detect an 'unbound' fraction that has migrated faster, and a 'bound' fraction, that has migrated slower due to an increased size of DNA-protein complex. By measuring the amount of fluorescence in each of the fractions, it is then possible to estimate the proportion of the total DNA that has been bound by the protein.

Another method used here is fluorescence anisotropy. This assay estimates the proportion of bound fluorescently-labelled oligonucleotide by measuring the rotational mobility of the labelled DNA. This is achieved by measurement of the fluorophore emitted light parallel and perpendicular with respect to the plane of polarized light excitation. In the time between excitation and emission, small molecules that rotate fast due to Brownian motion and achieve a randomised orientational distribution, resulting in a low ratio of parallel:perpendicular emitted light compared to larger species that rotate relatively slowly, so a greater fraction of emitted light is parallel to the plane of excitation (reviewed in Hall et al. (2017)).

5.1.3 Questions addressed in the current Chapter

In the current Chapter, I aim to experimentally test the viability of the *'biased mask'* model. I do this by measuring the affinity of the KLF4 protein to its target motif containing DNA sequences in the presence or absence of a mismatch or mutation primarily

at the position that shows an increase of substitutions in cancers, but also at other sites across the motif.

5.2 Methods

5.2.1 Fluorescently-labelled oligonucleotides

Here I test the affinity of the KLF4 protein to double-stranded oligonucleotides containing the KLF4 preferential binding sequence with or without a mismatch or mutation at one position within the motif. As a 'perfect' sequence (CORE; positive control) that would be expected to bind KLF4, I have chosen an actual 37 base-pair genomic site (`chr14:23,340,913-23,340,949`; *hg19*) that was found to be present under a KLF4 ChIP-seq peak within the core promoter of the *LRP10* gene, and showed an occurrence of 4 T⟶G mutations at position number 5 of the motif in the ICGC skin adenocarcinoma (SKCA) dataset (≈ ×3000 increase relative to the genome wide expectation in SKCA for GTG trinucleotide context). A complete set of oligo sequences can be found in Table 5.1. Oligonucleotides containing changes at position number 5 on the KLF4 motif (T⟶G change in the *forward* oligo and A⟶C change in the *reverse* oligo) were used to create duplexes with mismatched base on the forward strand (FMM - forward mismatch) or reverse strand (RMM - reverse mismatch) when annealed to the 'perfect' sequence *reverse* or *forward* oligo, respectively (Figure 5.1). Annealing of the two altered oligos together produced a sequence without a mismatch, but containing a mutation at position number 5 (MUT).

Additional oligos containing changes T⟶A on forward strand and A⟶T on the reverse strand were obtained. To test changes at other positions of the motif, we obtained oligos with T⟶A change on forward and A⟶T change on reverse at position number 7 (Jaspar motifs MA0039.2). This position is similar to position number 5 in PWM, but is not found to be highly mutated in cancers. Also we obtained oligos with G⟶A and C⟶T changes at position number 4 (Jaspar motifs MA0039.2), that would be strongly expected to abolish binding due to the high information content of this position in all KLF4 PWMs.

We have also obtained an alternative set of 35 base-pair sequences that contain the KLF4 binding site in the promoter region of the *LEFTY2* gene (`chr1:226,128,999 - 226,129,033`; *hg19*), same as in Soufi et al. (2015), including a set of oligonucleotides with incorporated uracil at the position number 5 of the KLF4 motif, which should allow

Name	Oligonucleotide sequence	Orientation	Position changed	Type of change	Mismatch
ktKlfLrpFor	5-/Cy5/CGTGCGTTGATTGGCAGGGGTGGGACCGGGCTTGTCA-3	Forward	-	-	CORE
ktKlfLrpRev	5-/Cy5/TGACAAGCCCGGTCCCACCCCTGCCAATCAACGCACG-3	Reverse	-	-	CORE
ktKlfLrp5agFor	5-/Cy5/CGTGCGTTGATTGGCAGGGG(G)GGGACCGGGCTTGTCA-3	Forward	5	T > G	G:A (FMM)
ktKlfLrp5agRev	5-/Cy5/TGACAAGCCCGGTCCC(C)CCCCTGCCAATCAACGCACG-3	Reverse	5	A > C	T:C (RMM)
ktKlfLrp5aaFor	5-/Cy5/CGTGCGTTGATTGGCAGGGG(A)GGGACCGGGCTTGTCA-3	Forward	5	T > A	A:A (FMM)
ktKlfLrp5aaRev	5-/Cy5/TGACAAGCCCGGTCCC(T)CCCCTGCCAATCAACGCACG-3	Reverse	5	A > T	T:T (RMM)
ktKlfLrp4caFor	5-/Cy5/CGTGCGTTGATTGGCAGGG(A)TGGGACCGGGCTTGTCA-3	Forward	4	G > A	A:C (FMM)
ktKlfLrp4caRev	5-/Cy5/TGACAAGCCCGGTCCCA(T)CCCTGCCAATCAACGCACG-3	Reverse	4	C > T	G:T (RMM)
ktKlfLrp7caFor	5-/Cy5/CGTGCGTTGATTGGCAGGGGTG(A)GACCGGGCTTGTCA-3	Forward	7	G > A	A:C (FMM)
ktKlfLrp7caRev	5-/Cy5/TGACAAGCCCGGTC(T)CACCCCTGCCAATCAACGCACG-3	Reverse	7	C > T	G:T (RMM)
Nanog	5-/Cy5/CTTACAGCTTCTTTTGCATTACAATGTCCATGGTGGA-3	Forward	-	-	-
Nanog	5-/Cy5/TCCACCATGGACATTGTAATGCAAAAGAAGCTGTAAG-3	Reverse	-	-	-

Table 5.1: KLF4 motif-containing Cy5-labelled synthetic oligonucleotide sequences. Positions that have been changed relative to 'perfect' CORE sequence (first two rows) are in brackets. ***'Orientation'*** *denotes the direction of the KLF4 motif (Jaspar matrix MA0039.2; Figure 5.1), and* ***'position'*** *denotes changed position according to the same motif PWM.* ***'Type of change'*** *indicates change relative to the CORE sequence.* ***'Mismatch'*** *indicated a type of mismatch that will form upon annealing that oligo with CORE oligo of opposite orientation.*

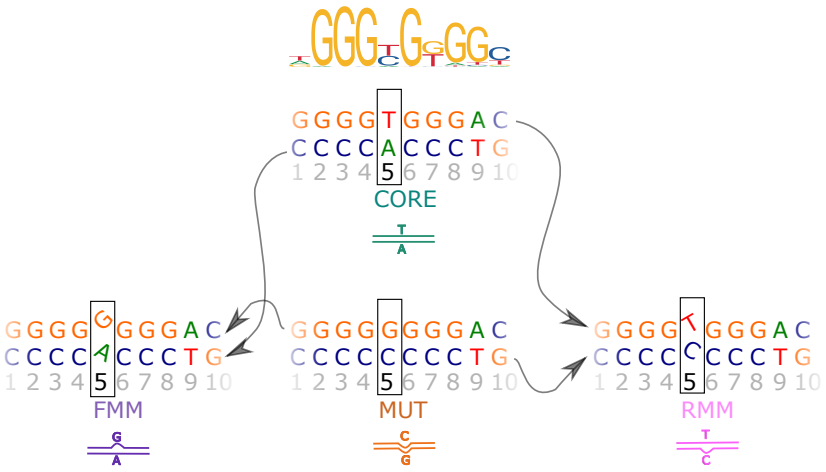


Figure 5.1: Formation of oligonucleotide duplexes with mismatches and mutation at position number 5 of the KLF4 motif (Jaspar PWM MA.0039.2; top). Forward and reverse oligos that form a 'perfect' CORE double-stranded motif can be combined with oligos that form a mutations (MUT) to get forward (FMM) and reverse (RMM) mismatched duplex oligos. Icons further used to represent those duplexes are drawn under each of the sequences.

for production of abasic sites by uracil-DNA glycosylase (no results from use of those oligos are presented in this work).

As a non-specific negative control I used oligonucleotides containing binding motif for the NANOG protein (where no sequence resembling the KLF4 binding site is present) from Soufi et al. (2015).

All oligonucleotides were ordered from IDT with 5' Cy5 label and HPLC purification. Oligo annealing was performed in annealing buffer that contained 20μM TrisHCl (pH7.6), 50μM NaCl, 0.1μM DTT (added shortly before use), 1μM EDTA. Annealing reaction was carried out for 15 minutes at ≈80°C in a water-bath, and then left to cool gradually overnight.

5.2.2 KLF4 protein

Full-length refolded histidine-tagged KLF4 protein (in 2M urea) was supplied by Dr Abdenour Soufi, and previously published (Soufi et al., 2015). Concentration of the protein stock was determined by running a range of KLF4 stock dilutions alongside known concentrations of bovine serum albumin (BSA) on a pre-cast SDS-PAGE gel (Figure 5.2). Stock concentration was calculated to be ≈ 0.7mg/ml.

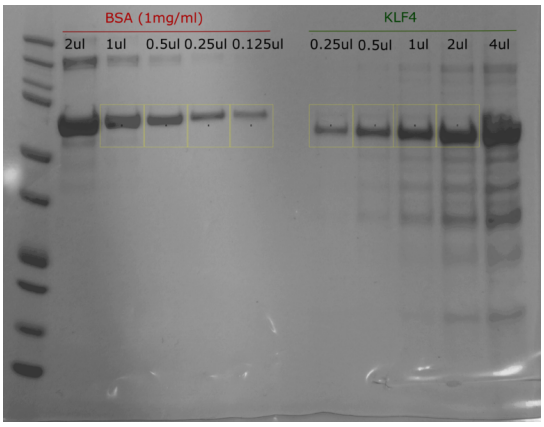


Figure 5.2: KLF4 protein stock quantification. Different volumes of KLF4 protein dilution were compared to a range of volumes of BSA dilutions with of known concentration. Protein amount per unit of fluorescence was calculated from BSA range (except at 2µl, as signal was over-saturated at in that lane), and protein amount per ml of KLF4 protein loaded was calculated accordingly.

5.2.3 EMSA experimental set-up

5% acrylamide gel was prepared by mixing 2.5ml of 10X TBE (0.9M Tris, 0.9M Borate, 0.02M EDTA; pH 8.3), 8.3ml acrylamide (30% Acrylamide Gel Solution Bis-Acrylamide Ratio 29:1; *Severn Biotech* 20-2600-05), 38.8ml H_2O, 350μl ammonium persulfate (10% W/V ; *Thermo Fisher Acrōs Organics, 98%+, 7727-54-0*) and 50μl TEMED (*Sigma-Aldrich, T9281*). After pouring, the gels were left to polymerise either overnight at

+4°C or for 1 hour at room temperature, then pre-run for 1 hour at 90V in 0.5X TBE prior to loading.

5X binding buffer was pre-made with 50mM Tris-HCl (pH 7.5), 5mM $MgCl_2$, 50μM DTT, 25% glycerol and 2.5 mg/ml BSA, and stored at -20°C. The binding reaction was performed in 1X binding buffer (made up from 5X binding buffer with H_2O). Mixtures with ranges of protein concentrations were prepared by serial dilutions of protein stock in 1X binding buffer (on ice).

Non-specific inhibitor poly(deoxyinosinic-deoxycytidylic) acid sodium salt (poly(dIdC); *Sigma-Aldrich, P4929*) was added to the oligo diluted to the desired concentration prior to the binding reaction. An optimal concentration of poly(dI-dC) was estimated by performing a titration (5-30ng/μl) in reaction with 1nM of either the perfect binding sequence (CORE) or non-specific (Nanog) oligo, and 5nM of KLF4 protein (Figure 5.3). As low as 5ng/μl of poly(dI-dC) did not completely prevent binding of all CORE oligo by the KLF4 protein, while fully abolishing non-specific binding of Nanog oligo. In results presented here either 2.5ng/μl or 1.25ng/μl of poly(dI-dC) was used, as indicated.

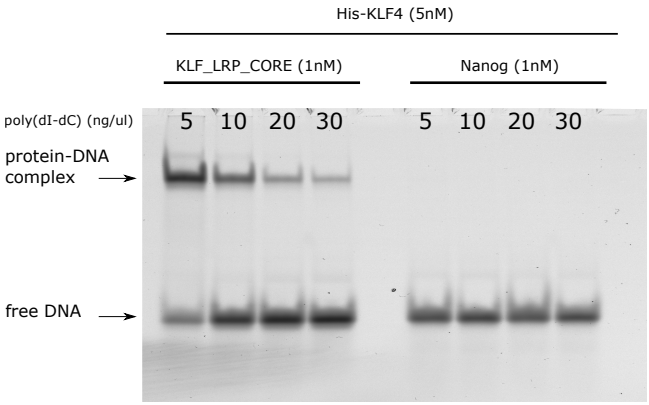


Figure 5.3: Titration of poly(dIdC) amounts. Range of poly(dIdC) concentrations were added to the binding reactions containing 5nM KLF4 protein and 1nM oligo duplex, in order to find concentration that would abolish non-specific binding (Nanog), while still retaining specific binding (KLF_LRP_CORE).

40μl reactions containing Cy5-labelled annealed DNA probes and a range of protein concentrations were made up by mixing equal volumes of protein mix and DNA mix, and the binding reaction was allowed to proceed for 1 hour at room temperature in the dark. 30μl of sample was then loaded into each well, alongside 2μl of loading dye (6X, *Thermo Fisher*, cat.num: R0611) (to monitor the progression of the gel, as the binding reactions did not contain any dye, so could not be tracked), while the gel was run at 150V. At least 2 free wells were left between dye and the sample, as blue dye gives out strong fluorescence at Cy5 wavelength). Gels were run at 90V at room

temperature for ≈3.5 hours. Gels were imaged on Fujifilm FLA-5100 fluorescent image analyser (Cy5 filter; 635 wavelength; 400V).

5.2.4 Quantification of binding affinity from EMSA gel images

Fluorescence intensity was quantified using MultiGauge (*Ver 3.0*) software. Background correction was applied to each individual gel image by subtracting the fluorescence value from an image area outside the lanes. Equal-sized windows were used to measure region intensity (measured in LAU - Linear Arbitrary Units) of each band. LAU values that became negative after background correction were set to 0. Following Soufi et al. (2015), the fraction of bound DNA for each lane was calculated using the following equation:

$$\text{Fraction } Oligo_{bound} = \frac{(Oligo_{bound})}{(Oligo_{total})} \qquad (f{:}5.2.1)$$

The LAU value that corresponded to the $Oligo_{total}$ was measured in one of two ways:

$$Oligo_{total} = Oligo_{unbound\ at\ 0nM\ protein} \qquad (f{:}5.2.2)$$

or

$$Oligo_{total} = Oligo_{bound} + Oligo_{unbound} \qquad (f{:}5.2.3)$$

The resulting value from Formula *f:*5.2.2 was used for a set of lanes with a range of protein concentrations and same oligo, while Formula *f:*5.2.3 value was calculated for each of the lanes individually. While the resulting estimation of *total* in case of Formula *f:*5.2.2 is more affected by the variability in sample loading between lanes, the result from Formula *f:*5.2.3 would be more affected by the presence of 'smeared' signal between measured *bound* and *unbound* fraction that can result from dynamic association and dissociation of the protein-oligo complex during electrophoresis (Dr A.Soufi, personal communication). Therefore, in the latter case, the proportion of total amount of oligo could be underestimated, which would lead to overestimation of the bound proportion. Another option is to measure *total* amount of oligo as the florescence value obtained from the whole lane, however that would lead to the introduction of a larger error due to high levels of background noise. While all of the described methods could

potentially introduce an error in estimation of the bound proportion, this error is likely to be uniform between all experiments, so would not be expected to affect comparison between these.

Apparent dissociation constant (K_d) was used to measure the affinity of the protein to each oligo. K_d was measured by fitting a non-linear model to the data using R (3.4.1). Binding curves describing the fraction of bound DNA as a function of the protein concentration (*Prot*) from multiple separate experiments for each oligo type were fitted to the data using either nonlinear least squares `nlsLM` function from the `minpack.lm` (v1.2-1) package, or `nls` function in R. The K_d and Hill coefficient (*HC*) for each oligo was calculated using the following equation:

$$\text{Fraction}_{\text{DNA bound}} = \frac{B_{max} \times Prot^{HC}}{K_d^{HC} + Prot^{HC}} \qquad (f\text{:}5.2.4)$$

Where *Prot* denotes protein concentration; B_{max} (maximal fraction of oligo bound) was kept constant and equal to 1; and the supplied starting values were 1 and 0.1 for K_d and *HC*, respectively.

5.2.5 Anisotropy experimental setup

The anisotropy assay was performed in black non-transparent 384-well plates (*Corning, product number: 10109202*). The binding reaction was carried out in PBS with addition of 50μM $ZnSO_4$ (*Sigma Aldrich, 83265*). 5' 6-FAM labelled oligos were used for this assay. 20μl of oligo mix was added to a range of protein dilutions (20μ of each dilution). There were 3 replicate binding reactions for each protein concentration. Incubation reactions were carried out for 1 hour at room temperature in the dark. Fluorescence polarization was then measured on a PerkinElmer **Wallac 1420 Victor**^{2} microplate reader, with 3 replicate readings for each well (F485 CW-lamp excitation filter; F535 emission filter; measurement time 0.2sec; CW-lamp energy 65,535 (instrument arbitrary units); constant voltage; G-factor 1).

Fluorescence polarization values (mP) were background-corrected by subtracting the polarization value obtained at 0nM protein for each replicate and each reading separately. Resulting values were then used to fit non-linear least squares model using `nls` function in `R` (3.4.1), similar to what was done for EMSAs, where Formula *f*:5.2.4 was used, but with HC value fixed to 1, and a non-fixed B_{max} value. Confidence intervals were calculated using the `quantile` function in `R` (3.4.1).

5.3 Results

5.3.1 KLF4 shows strand-specific affinity to sequences with mismatches by electrophoretic mobility shift assay

To test the tolerance of KLF4 protein to the T⟶G mutation at position number 5 of the motif (Figure 4.3b), along with a G:A (FMM - forward mismatch) and T:C (RMM - reverse mismatch) mismatches, I performed EMSAs for a range of protein concentrations with fixed amounts of oligo, constructed binding curves for each oligo type, and estimated a dissociation constant value (K_d) using Formula *f:*5.2.3 for estimation of total amount of oligo in a lane. Here the value of K_d essentially represents a protein concentration at which half of the ligand (fluorescently-labelled oligonucleotide duplex) present in solution is free, and half is bound. Lower K_d values represent higher affinity. K_d values reported here are *apparent* dissociation constant value and do not reflect the physiological concentrations of protein required to achieve half-maximal binding.

I initially tested affinity of a range of KLF4 protein concentrations (0-10nM) to 2nM oligonucleotide duplexes in the absence of any non-specific inhibitor. Fitted binding curves with estimated K_d values and images of the gels can be seen on Figure 5.4. In the absence of non-specific inhibition KLF4 protein shows highest affinity (K_d = 2.42nM) for the 'perfect' binding sequence (CORE). Sequences containing mismatches at position number 5 of the motif (FMM and RMM), while reach almost full binding at 10nM protein, show a lower affinity in the intermediate protein concentrations, increasing K_d to ≈4.9nM ,with no significant difference between the two. The oligo duplex containing a mutation at position number 5 of the motif shows the lowest affinity amongst all, failing to reach half maximal binding at the highest protein concentration tested, with a calculated K_d of 63.6nM. This oligo, however, was only tested with one replicate and therefore this result lacks a measure of significance. As can be seen from the gel images on Figure 5.4, there is a high degree of 'smeared' signal over the lanes in between the bound and un-bound fractions, particularly with increasing protein concentrations, indicative of non-specific binding, as those initial experiments were done in absence of any non-specific inhibitor. Therefore in the subsequent experiments I used the non-specific inhibitor poly(dI-dC).

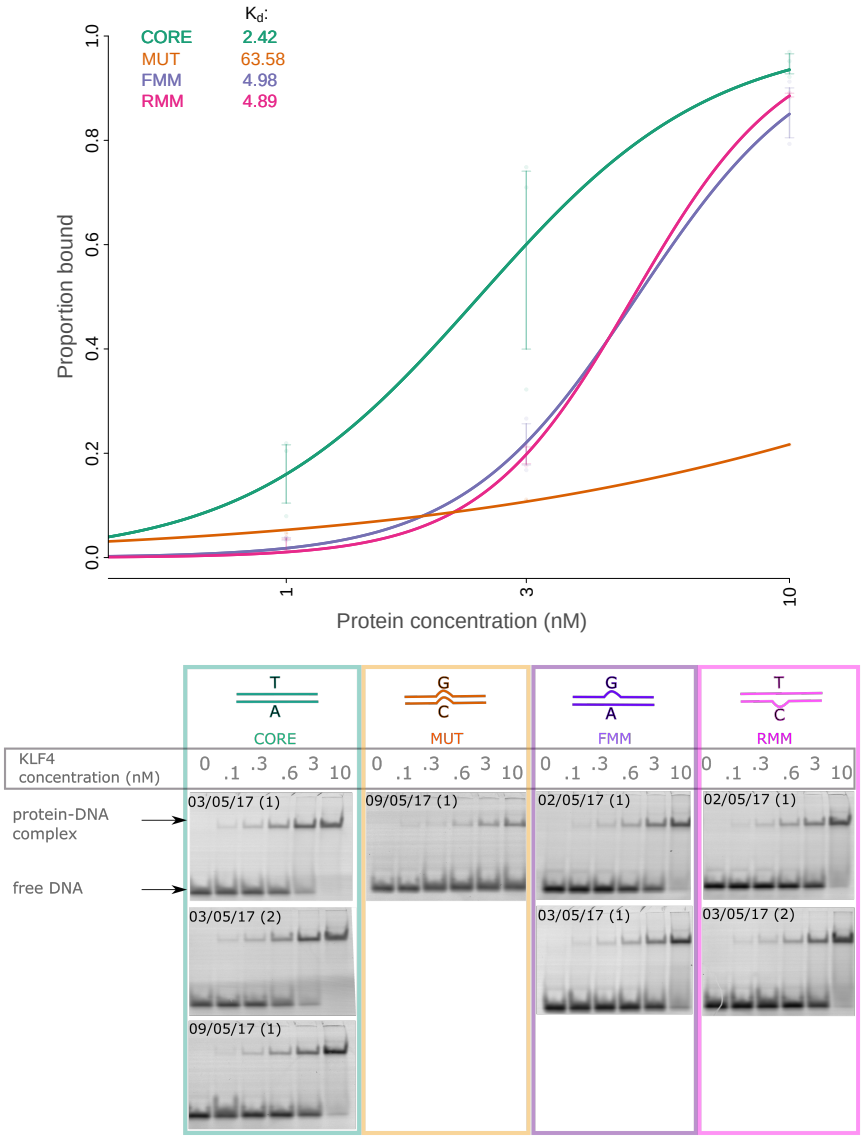


Figure 5.4: KLF4 binding affinity in the 0-10nM protein concentration range for CORE (n=3), MUT (n=1), FMM (n=2), and RMM (n=2) oligonucleotide duplexes (2nM) in absence of non-specific inhibitor. Dissociation constant (K_d) and Hill coefficient (HC) obtained from fitted model. Error bars represent 90% confidence interval. Note log-scale x-axis. Gel images from each separate replicate are shown below, and bands corresponding to both free DNA and protein-DNA complex indicated.

Figure 5.5 shows fitted binding curves with estimated K_d and gel images of binding assays where a range of KLF4 protein concentrations (0-10nM) were tested for the affinity to different types of duplex oligonucleotides (1nM) in the presence of 2.5ng/μl poly(dI-dC). K_d values became higher with addition of poly(dIdC), as expected. The protein still shows highest affinity to the oligo duplex containing 'perfect'

binding motif (CORE), with K_d = 10.72nM. Similar to what has been seen without non-specific inhibition, oligo duplex containing a guanine instead of thymine at position number 5 of the motif on the forward strand (FMM) is bound by protein worse than CORE, but better than oligo duplex with mutation at the same position (MUT). Interestingly, the oligo duplex with a mismatch on the reverse strand (adenine replaced by cytosine on the reverse strand) is now bound by the protein with similar affinity to MUT with K_d of 25.29nM. None of the oligonucleotide duplexes, including CORE, reach full or near full binding at the highest protein concentration, with CORE only just about reaching the half-maximal binding.

I have therefore increased protein concentration to get a fuller picture of binding dynamics. Figure 5.6 shows KLF4 binding to oligonucleotide duplexes (1nM) in 0-50nM protein concentration range in presence of 1.25ng/μl poly(dI-dC). Higher affinity of KLF4 to FMM over RMM hold true at 20nM protein concentration, after which point there is no significant difference between two mismatch types. Difference between FMM and MUT, however holds true up to 35nM protein (Figure 5.7, RMM values have been removed for better discrimination between FMM and MUT), after which point there is also an increase in non-specific binding, as measured by affinity of the protein to the non-specific NANOG oligonucleotide duplex.

Importantly, these results show that KLF4 protein is able to bind over the mismatched lesion.

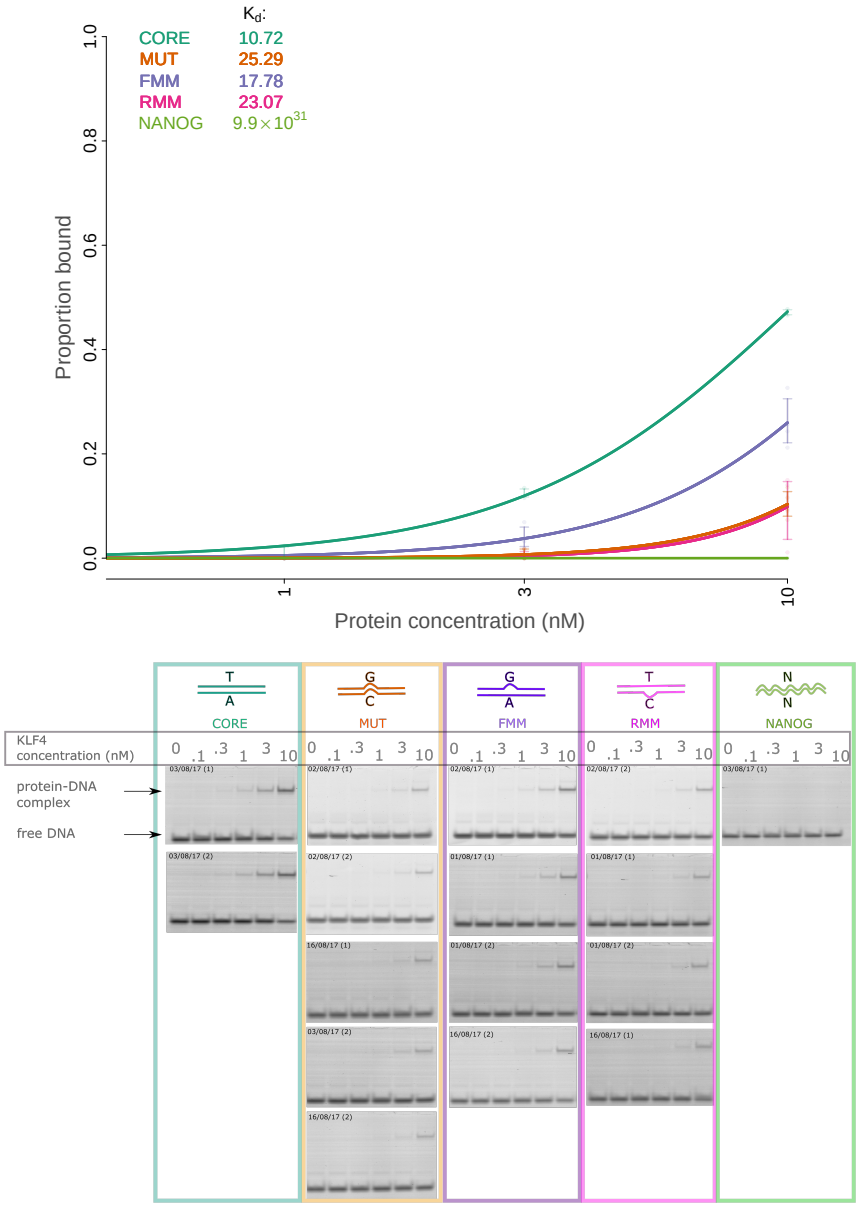


Figure 5.5: KLF4 binding affinity in the 0-10nM protein concentration range for CORE (n=2), MUT (n=5), FMM (n=4), RMM (n=4), and NANOG (n=1) oligonucleotide duplexes (1nM) in presence of 2.5ng/µl non-specific inhibitor poly(dIdC). Dissociation constants (K_d) and Hill coefficients (HC) were obtained from fitted model. Error bars represent the 90% confidence interval. Note log-scale x-axis. Gel images from each separate replicate are shown below, and bands corresponding to both free DNA and protein-DNA complex indicated. Dates are indicated in upper left corner.

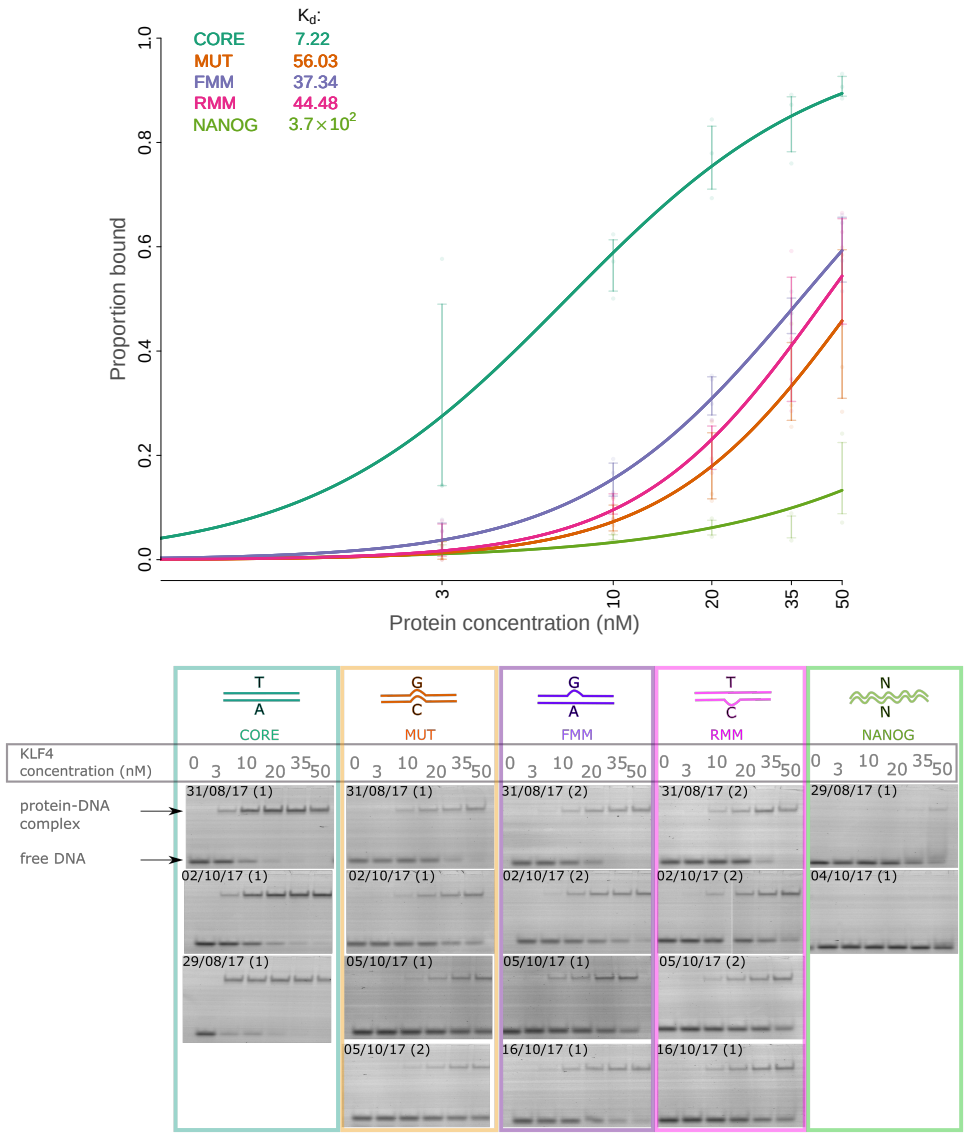


Figure 5.6: KLF4 binding affinity in the 0-50nM protein concentration range for CORE (n=3), MUT (n=4), FMM (n=4), RMM (n=4), and NANOG (n=2) oligonucleotide duplexes (1nM) in presence of 1.25ng/µl non-specific inhibitor poly(dIdC). Dissociation constants (K_d) and Hill coefficients (HC) were obtained from fitted model. Error bars represent the 90% confidence interval. Note log-scale x-axis. Gel images from each separate replicate are shown below, and bands corresponding to both free DNA and protein-DNA complex indicated.

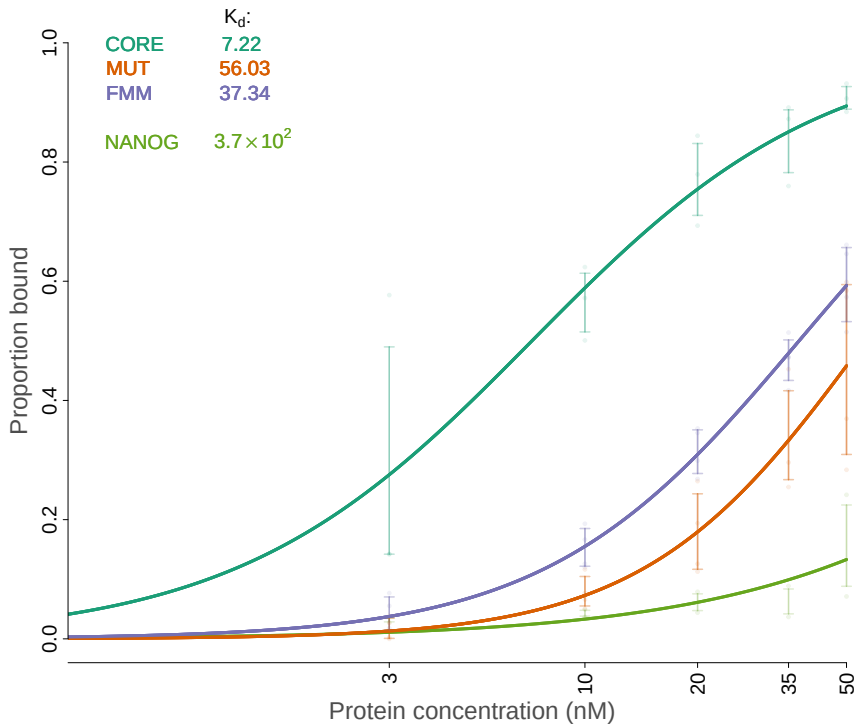


Figure 5.7: KLF4 binding affinity in the 0-50nM protein concentration range for CORE (n=3), MUT (n=4), FMM (n=4), and NANOG (n=2) oligonucleotide duplexes (1nM) in presence of 1.25ng/µl non-specific inhibitor poly(dIdC). Dissociation constants (K_d) and Hill coefficients (HC) were obtained from fitted model. Error bars represent 90% confidence interval. Note log-scale x-axis.

5.3.2 KLF4 binding affinity to a winder range of mismatches

I have also tested the affinity of the KLF4 protein to a wider range of changes at three different positions (n=1; Figure 5.8). Similar to what has been observed in the binding assays described above, at lower protein concentration (3nM) there is considerable variability that is not necessarily consistent with values at higher protein concentration (10nM). This is not surprising in light of the fact that the largest amount of poly(dIdC) (3ng/μl) was added here. At 10nM protein concentration, in accordance with the previous assays, the G:C mutation at position number 5 is bound worse than the reverse strand mismatch (T:C), which in turn shows lower affinity for the protein than the forward mismatch (G:A). Other types of change at position number 5 (forward mismatch A:A ; reverse mismatch T:T ; or mutation A:T) appear to be better tolerated by the KLF4 protein, and unlike the previous type of change, mutation is tolerated better than either of the mismatches. Interestingly, at position number 4 of the motif, which is strongly biased towards G:C within the PWM, mutated A:T is tolerated almost just as well as 'perfect' sequence, while either of the mismatches (G:T or A:C) show much lower affinity. Position 7, at which KLF4 protein exhibits a nucleotide preference most similar to position number 5 (except for G being expected to be more tolerated), is least tolerated with the mismatch on the reverse strand (T:G) and mutation (C:G), while mismatch on the forward strand (C:A) binds KLF4 better. This binding assay has only been performed in one replicate, therefore any conclusions made here are not statistically robust.

5.3.3 KLF4 binding affinity measured by fluorescence anisotropy

I also performed a binding assay that allows detection of the protein affinity by fluorescence polarization/anisotropy. In principle, this method allows for the faster, higher-throughput measurement of the affinity of the protein to its ligand, and can be done with a number of different fluorescently-labelled oligonucleotide duplexes in one 396-well plate. I tested 0-200nM range of KLF4 protein concentrations with 1nM of CORE, MUT (position number 5 G:C), FMM (position number 5 G:A), and RMM (position number 5 T:C) oligos in the presence of 0.625ng/μl poly(dIdC) (Figure 5.9). There is a lot of variability in the range of polarization values between replicates, which might

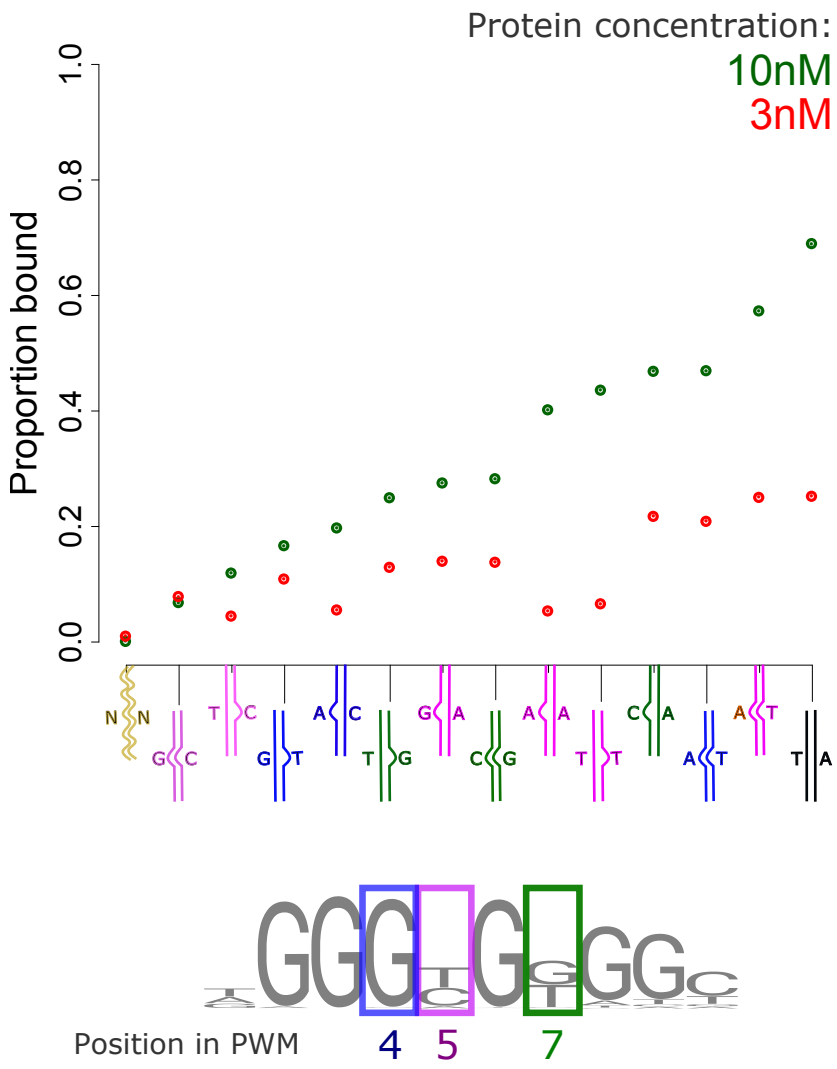


Figure 5.8: KLF4 binding affinity to mismatches and mutations at positions numbers 4,5 and 7 of the KLF4 motif (MA0039.2). The proportion of oligo (1nM) duplex bound at 10nM and 3nM (green and red, respectively) in presence of 3ng/µl poly(dIdC) are plotted in ascending order of 10nM protein binding affinity. Simplified illustrations of lesions are represented below the graph with the forward (G-rich) strand to the left of the image. Colours represent different position within the motif (blue - 4; pink - 5; green - 7), as indicated at PWM at the bottom. The non-altered motif is in black, and non-specific sequence (Nanog) in yellow.

be due to failure to define the most optimal binding conditions or measure-instrument settings. Perfect motif-containing CORE sequence has shown the lowest binding affinity, contrary to what would be expected. The rest of the oligonucleotide duplexes show patterns similar to what was observed before with EMSAs, with FMM binding better than RMM or MUT. However, none of these show differences between each other that

could be deemed significant. More assay optimization would be needed to apply this method in the future.

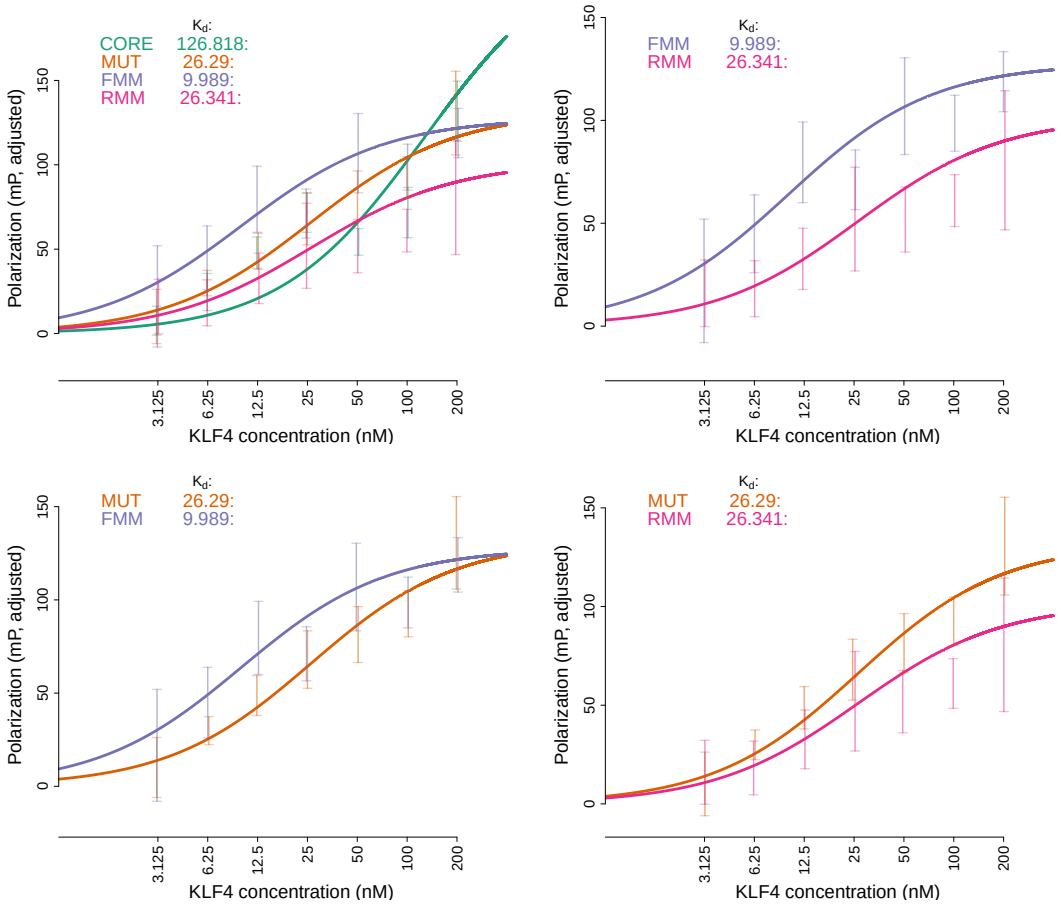


Figure 5.9: KLF4 (0-200nM) binding affinity to CORE, MUT, FMM and RMM oligo duplexes (1nM) measured by fluorescence anisotropy (3 replicates for each oligo type were read 3 times) in presence of 0.625ng/µl poly(dIdC). Polarization values (mP) were adjusted by subtracting the mP value at 0nM protein. K_d values for each oligo type are indicated on the plots. Error bars represent 90% confidence intervals. All four plots represent the same data, individual comparisons included for better visual discrimination between pairs of oligonucleotide duplexes.

5.4 Discussion

In the current Chapter, I explored the experimental plausibility of the *biased mask* model in relation to the binding site of the KLF4 transcription factor. As described in the previous chapter, KLF4 binding sites show an enrichment of potentially binding-abolishing mutations in cancers (in particular T:A⟶G:C) at position number 5 of the motif (Figure 4.3b). By assessing the affinity of KLF4 protein to duplex oligonucleotides containing a KLF4 binding motif *in vitro*, I show that the T:A⟶G:C mutation at position number 5 of the motif is indeed likely to lead to a significant reduction in binding affinity (Figure 5.6).

The *biased mask* model proposed here predicts that the high level of mutations at position number 5 of KLF4 motif arises due to the protein being able to bind over a mismatch, or other types of lesions, and thereby protect those from being fixed by repair machinery. This then assumes that protein is able to bind over a lesion at a position within a motif which might be important for protein-DNA interaction. Here I show that KLF4 can indeed bind over the G:A mismatch at the position number 5 of the motif (Figure 5.6). While the binding affinity is reduced to below what can be seen for the 'perfect' motif sequence, it is higher than the protein's affinity to the mutated position. There is also a general trend for KLF4 being able to tolerate a T:G mismatch somewhat better than A:C mismatch, which is consistent with the proposed idea of the strand-specific lesion tolerance by sequence-specific binding factors. The binding dynamics of the RMM (A:C at position number 5 of the motif) appears to differ with various protein concentrations, lagging behind FMM (T:G at position number 5 of the motif) at lower concentrations, but reaching a similar level of binding at higher concentrations.

It is important to note that those conclusions come from binding assays performed *in vitro* with the concentrations of protein and oligonucleotides that do not necessarily represent the true physiological conditions. The K_d values estimated here are also not absolute, but *apparent* K_d values, specific only to the conditions of assay and differ with the amounts of added non-specific inhibitor. In addition to that, the proportion of the total protein in the solution that is catalytically active is not known. While seen consistently, the differences in binding affinity of KLF4 protein to mutated

position number 5 *versus* T:G mismatch as the same position is not very big. The effect of this difference *in vivo* physiological conditions on the binding affinity is impossible to quantify.

A brief analysis of the wider range of changes including different positions within the motif shows that at the highly important position number 4 of the MA0039.1 KLF4 motif, a mismatch on either of the strands substantially reduces binding affinity, while the G:C⟶A:T mutation has a smaller effect. This is consistent with our model and the low mutation rate at that position, as an inability of protein to bind over a lesion of this type means it would get fixed more efficiently. T:A⟶A:T mutation at position number 5 of the motif appears to be affecting KLF4 binding the least amongst all the changes tested here. Corresponding mismatches (A:A and T:T) appear to reduce binding by ≈40% (at 10nM protein ; by ≈80% at 3nM), which is somewhat consistent with a lack of increase in T:A⟶A:T mutations at that position. Position 7 of the motif, however, shows a ≈35% decrease in KLF4 binding affinity with C:A mismatch, while either T:G mismatch or T:A⟶C:G mutation show ≈60% decrease. This leads to an anticipation of a higher rate of T:A⟶C:G mutation at this position, which we do not observe in cancer data. Due to the fact that this assay was not performed in replicate, further experiments would be necessary to confirm any of the findings and conclusions made here.

When trying to replicate the results obtained by EMSAs using the fluorescence anisotropy assay, KLF4 binding affinity to mutated position number 5 and both types of mismatches showed similar trends. However, the binding affinity to oligonucleotide duplex containing a 'perfect' binding motif, which should represent a positive control and would be expected to show highest binding affinity, did not do so. This, in conjunction with high variability in anisotropy values for replicate binding reactions, suggests that a better optimization of the assay conditions and measurement setting is needed in the future. Fluorescence anisotropy provides a potential effective high-throughput way to test the binding affinity of a protein to a variety of different oligonucleotide duplexes in a more time-efficient manner. It also provides an opportunity to perform competition assays, where affinity of the protein to different oligonucleotide duplexes labelled with fluorophores with non-overlapping emission spectra could be tested in the same reaction.

The results presented here are supportive of the *biased mask model*, and act

as a preliminary indication of the plausibility of this model as a mechanism that drives retention of certain types of lesions. However, further implementation of, ideally, more high-throughput methods, such as anisotropy, would be necessary to confirm the current results, and to show generalizability of this model to a wider range of TFs.

A further idea is to implement a larger-scale assay that can detect the binding affinity of a protein to a complete range of mismatches and mutations at all of the positions within the motif. That would involve use of 'doped' oligonucleotides, where during oligonucleotide production, whenever each position gets synthesised, instead of providing the polymerase with 100% nucleotides of correct identity, a certain percentage of alternate nucleotides are introduced. That means that in the final pool of oligonucleotides every position is going to have a specific percentage of alternate nucleotide. Combining the pool of 'doped' forward or reverse strand oligos with reverse or forward strand oligos containing an unaltered binding motif, respectively, would result in pools of oligonucleotide duplexes where every position would be expected to have every type of mismatch at each position that has been 'doped' on both strands. A GST-tagged protein can then be used to perform binding assays with both of these pools. The protein-DNA complexes can then be attached to a magnetic bead with glutathione, and pulled down with a magnet. Both input and pulled-down fractions can then be sequenced to determine the relative proportion of every type of mismatch that has formed a complex.

I have already attempted to implement the pull-down approach using two types of oligonucleotide duplexes (either two from perfect motif-containing, mismatch-containing at highly mutated position on either strand, or mutation-containing at the same position) that have been labelled with different fluorophores with non-overlapping excitation spectra and commercially obtained GST-tagged CTCF protein. The aim was to estimate the proportions of each type of oligonucleotide duplex that has formed a protein-DNA complex by comparing the amounts of each type of fluorescence in both input and bound fractions. Unfortunately, the GST-tagged CTCF protein did not appear to be functional for DNA binding in this assay, and neither did it show binding to its consensus binding motif in EMSA. The ongoing work by Susan Campbell is aimed at expression and purification of the CTCF, BORIS, KLF4 and EGR1 proteins for use in this assay.

Overall, the results presented here are consistent with the *biased mask* model

proposed in the previous chapter, but more experiments would be necessary to confirm findings presented here, as well as to generalise to a broader range of proteins, and positions to show that this model applies to more than just one TF and a single type of change. In particular, it would be interesting to implement a these assays with the CTCF protein, due to its highly diverse and important roles within the cell.

CHAPTER 6

Conclusions and general discussion

6.1 Protein binding sites are mutational hotspots

Work presented here was specifically motivated by the observation of increased sequence variation proximal to some sequence-specific binding sites by Reijns et al. (2015). Several possible models could have explained the observed pattern, with either selection pressures or differential mutation rate as a cause. Elevated mutation model was previously the favoured hypothesis, but it had not been explicitly tested prior to this work. By utilizing measures of human and mouse population variation and performing derived allele frequency tests, I have separated patterns of selection and mutation in order to differentiate between those possibilities across protein binding sites that are active in spermatogonial cells. Those sites were defined through analysis of in-house generated ATAC-seq data.

This analysis has revealed that there is an increase in germline variation across and at the edges of the binding sites (Figure 3.5). I have shown for the first time that this variability is not driven by diversifying selection, but rather by differential mutation rate, which is elevated at regions occupied by proteins. This means that protein binding sites in the human and mouse populations are mutational hotspots for deleterious mutations. While with less power, differences in rates of *de novo* mutations between binding sites and proximal regions further support this observation (Figure 3.17). I estimate that almost 1 in 4 births is likely to harbour one of those deleterious *de novo* mutations across the protein binding sites.

This finding has implications for human health - many protein binding sites harbour sequences that are important for recruiting TFs which are part of gene regulatory networks, and dysregulation of those could lead to disease. Here I observe an increase in germline mutations at these sites, which means that they are also likely to be inherited. This is not to say that this mutational mechanism is active exclusively in the germline cells, as we anticipate that protein binding sites in somatic cells show the same mutational property.

The importance of increased rate of mutations at protein binding sites is also reflected in an expansion of our knowledge about the neutral expectation of single nucleotide substitutions across the genome. Estimations of neutrality are important, as they are utilized as a baseline to understand whether a sequence is mutated above or

below expectation and used to define functionality. While the work described here has been limited to investigation of properties of single nucleotide substitutions, exploration of other types of DNA sequence changes, such as indels, would be a useful complement to this analysis.

Investigation of transcriptional profiles of our isolated cells would also be of great interest. For one, this could allow us to better characterise the isolated cell populations, but also to associate the mutagenic effects at the protein binding sites with transcriptional activity. In addition, it is possible to perform single-cell ATAC-seq on the isolated spermatogonial cells, which would allow for more detailed exploration of the individual cell populations.

6.2 Identified binding sites in the human germline can allow for the prioritization of disease-causing variants

While the link between an increased mutational burden of protein binding sites and replication has not been conclusively demonstrated in the current work, general association between mutagenesis and replication has previously been shown through the observation of increased numbers of *de novo* mutations in the children of older fathers. The incidence of conditions such as autism and schizophrenia in offspring have also been linked to the age of father at time of conception (Kong et al., 2012). Additionally, a large proportion of cases of such disorders currently can't be explained by mutations found within the protein-coding regions, implying that non-coding mutations, particularly those located within the regulatory sites are likely to be contributing to the incidence of the disease (Zhou et al., 2018).

The main focus of the current work was to explore the germline mutational burden of the non-coding part of the genome that contains functionally active regulatory regions in highly dividing cells of the male germline (spermatogonial cells). Therefore, a map of the protein binding sites in spermatogonial cells can be a place to look for the disease-causing variants, such as hits from genome-wide association studies (GWAS). Being able to narrow search space to regions that are germline-active regulatory sites might aid with the identification of candidate variants for prioritisation in assessment of their functional roles both in eliciting molecular and organismal phenotypes.

Future work could also involve the further development of the FLOP method to be used for the protein binding site identification. As mentioned in Subsection 2.4.2 , use of the machine learning methods and utilization of the *Segway* software for this purpose can lead to a new way of ATAC-seq data utilization for protein-binding landscape mapping.

While the current work was focused on identification of mutational landscape at the protein binding sites in spermatogonial cell, it could be extended to other germline cell populations. For example, arrested meiosis in oocytes, oxidative stress in sperm cells, as well as rapid divisions during initial stages of embryogenesis, all could probably lead to distinct heritable mutational patterns.

6.3 Germline-active, but not somatic-specific binding sites show increased germline variation supporting a link between physical DNA-protein interaction and mutagenesis

The causal relationship between an increased rate of mutation and physical interaction of DNA by TFs has been suggested before and attributed to the occlusion and retention of DNA lesions by bound proteins (Reijns et al., 2015; Sabarinathan et al., 2016; Perera et al., 2016). Here I tested this association through comparison of different categories of binding sites - those bound in both germline and somatic cells, or exclusively in either. Identification of these categories of sites was achieved through generation and analysis of primary ATAC-seq data from spermatogonial cells that we have isolated from testicular tissues, complemented by analysis of ATAC-seq data from a range of somatic cell types from publicly available sources.

I have shown that there is an association between protein binding and an increase in mutation rate, as only the sites that are active in germline cells, and not somatic-specific ones, show increased rates of germline mutations (Figure 3.10). The strongest mutational effect was observed in the 'housekeeping' category of sites, those that are bound most consistently across all the tissues analysed (Figure 3.10a). The fact that those sites were detected to be occupied in all of the tissues might also reflect a stronger interaction of these proteins with DNA, with longer residency time, which makes them more likely to be detected by ATAC-seq. Most 'housekeeping' sites appear to be located close to the transcription start sites and are proximal or located within the promoters (Figures 2.17c and 2.17b).

Rapid binding of those regions by TFs post-replication (Smith and Whitehouse, 2012) might be necessary for establishment of the proper gene expression patterns in the cell and priming of pluripotency-associated gene promoters in germline cells (Guo et al., 2017). To explore whether the increase in mutation rate is replication-associated, I looked at the protein binding sites that are specific to the highly dividing spermatogonial cells. Spermatogonia-specific binding sites showed a modest increase in mutation rates, although not consistently across all replicates tested (Figure 3.13b).

While in the current work I have considered the protein-binding landscape map as a whole, only making a distinction between tissue-specific and housekeeping protein binding sites, in future it would be useful to more specifically interrogate the nature of these binders.

The mutation measures used here were germline mutations derived or inferred from large cohort studies. Exploring where those mutations are most likely to occur and have functional consequences is a question directly relevant to human health. However, other avenues could be explored to test the dependency of this mutational effect on replication. Comparisons of numbers, types and locations of mutations acquired at protein binding sites through multiple generations of slow and fast replicating cells in culture, for example. Sequencing of individual sperm cell and acquisition of information about where mutations have occurred is also one of the potential avenues that could provide enough power to explore the mutation rate heterogeneity in germline cells. But with current technology it is not economically feasible to attain a sufficient discrimination of real mutations from sequencing errors. Future work could involve investigation of distribution of the *de novo* mutations that have been phased (ideally from the a range of the paternal ages), to explore whether most of mutations at the protein binding sites indeed come from the male germline. This is currently impeded by the insufficient available data about these mutations.

6.4 Biased mask model - mechanistic basis for retention of mutations by proteins at transcription factor binding sites

An association of the increased mutation rate and DNA-protein interactions leads to a question about the mechanistic basis of this relationship. The sequence-specific nature of the interaction of transcriptions factors with DNA makes it counter-intuitive to assume that a lesion can be occluded or protected from repair at a position that is crucial for binding. However, as exemplified by the binding motif of CTCF and cancer-derived mutations (Katainen et al., 2015; Kaiser et al., 2016), there is an increase in mutation load at positions within the motif that are likely to lead to an abolition of binding ability (Umer et al., 2016).

There is an open question whether the occurrence of a lesion on a single strand, such as a mismatch, would similarly lead to loss of binding, and to knowledge, has not been tested previously beyond base modifications such as methylation (Hashimoto et al., 2016, 2017). Assuming there is a different outcome for mutation or a single lesion, the mutational pattern over the CTCF binding motif could be explained by the 'biased mask model' that is proposed here. This model explains specific types of changes at particular positions within binding motifs as a consequence of differential tolerance of various types of lesions by a TF, possibly in a strand-specific manner. This is complemented by the level of mutagenesis at positions of interest being consistent with protection from mismatch repair (Figures 4.11, 4.12, and 4.13), supporting the idea that a particular type of change can escape surveillance and repair through occlusion by bound TF. Future work aimed at exploration of larger datasets of mutations from tumours with impaired mismatch repair would be advantageous, especially if there was enough power to look at the observed, rather than expected, level of non-repaired mutations over and next to the binding sites. Also, as only certain subsets of cancer cohorts appear to exhibit elevated mutations within the motifs, matched cancer type MMRd data would allow one to explore whether it is certain types of lesions that are normally protected from repair. Such datasets have been generated by studies such as Katainen et al. (2015), and procedures toward obtaining these data for further analyses

have already been initiated.

The presence of unusually highly mutated positions within the motifs of other TFs, such as KLF4 and EGR1, shows that this is likely to be a generalisable mutational mechanism which is not specific to just CTCF. The incidence of those mutations are cancer-type dependent and might be driven by particular processes or occurrences of distinct types of lesions. In future, it would also be interesting to explore data, such as recently published study (Corces et al., 2018), where chromatin accessibility and mutation data is available from the same cancer samples.

6.5 The biased mask model is supported by the DNA binding properties of the KLF4 protein

I have experimentally validated the viability of the 'biased mask' model through testing binding affinity of KLF4 protein to synthetic DNA sequences with mismatch lesions. I show that presence of a mismatch, while weakening, does not abolish KLF4 binding to its target motif (Figure 5.6). There is also evidence of the hypothesised strand asymmetry in this tolerance. The presence of double-stranded mutation at the tested position significantly lowers protein binding affinity even further. An important finding here is that the KLF4 is able to bind over a mismatch, which means a protein could protect a lesion from being fixed and thereby induce formation of mutation after a single round of replication, something that has not been demonstrated before for any DNA-interacting TF.

In future, more high-throughput method for testing other positions and other TFs could be achieved by optimizing the condition for the anisotropy assay. Testing of a wider variety of the binders will be necessary for generalization of the *biased mask* model to DNA-interacting protein families beyond C_2H_2 family. In addition to that, further testing of more diverse types of lesions should be carried out in the future. Work aimed at this, and purification of CTCF and EGR1 proteins is currently underway within the group, carried out by Susan Campbell. Further complementation of this with the pull-down method described in Section 5.4, will provide another orthogonal approach towards measuring of the tolerance of the proteins to strand-specific lesions.

In addition, future work on structural and molecular dynamic modelling of KLF4, CTCF and other zinc finger TFs in the context of mismatches and other lesions within their binding motifs could provide useful insights and further support for the *biased mask* model.

6.6 Final remarks

This work explores protein-binding sites as a feature that influences germline regional mutation rates, showing that those sites are mutational hotspots that could potentially lead to occurrence of variant that would result in hereditary disease. It also provides a map of protein-binding sites within germline cells that are potentially most prone to accumulate those mutations, which could narrow the search space for disease-causing variants. This elevation in mutation rate at protein-binding sites is expected to be equally applicable to mutations in somatic cells, that could lead to cancer and other non-heritable diseases. In addition, this work proposes and experimentally tests a mechanistic model through which those mutation escape the normal processes of surveillance and repair, and are therefore retained in the genome. The results presented in this thesis will improve our ability to detect genetic mutations and the mechanisms driving these events which lead to both heritable disease and cancer, which is crucial for widening our understanding of disease aetiology and future development of targeted therapies.

Bibliography

Acuna-Hidalgo, R., Bo, T., Kwint, M. P., Vorst, M. V. D., Pinelli, M., Veltman, J. A., Hoischen, A., Vissers, L. E. L. M., and Gilissen, C. (2015). Post-zygotic Point Mutations Are an Underrecognized Source of De Novo Genomic Variation. *The American Journal of Human Genetics*, 97(1):67–74.

Acuna-Hidalgo, R., Veltman, J. A., and Hoischen, A. (2016). New insights into the generation and role of de novo mutations in health and disease. *Genome Biology*, 17(1):241.

Adey, A., Morrison, H. G., Asan, Xun, X., Kitzman, J. O., Turner, E. H., Stackhouse, B., MacKenzie, A. P., Caruccio, N. C., Zhang, X., and Shendure, J. (2010). Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome biology*, 11(12):R119.

Aggarwala, V. and Voight, B. F. (2016). An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nature Genetics*, 48(4):349–355.

Alexandrov, L. B. (2018). The Repertoire of Mutational Signatures in Human Cancer. *bioRxiv*.

Alexandrov, L. B., Jones, P. H., Wedge, D. C., Sale, J. E., Campbell, P. J., Nik-Zainal, S., and Stratton, M. R. (2015). Clock-like mutational processes in human somatic cells. *Nature Genetics*, 47(12):1402–1407.

Alexandrov, L. B., Ju, Y. S., Haase, K., Van Loo, P., Martincorena, I., Nik-Zainal, S., Totoki, Y., Fujimoto, A., Nakagawa, H., Shibata, T., Campbell, P. J., Vineis, P., Phillips, D. H., and Stratton, M. R. (2016). Mutational signatures associated with tobacco smoking in human cancer. *Science (New York, N.Y.)*, 354(6312):618–622.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., Eyfjörd, J. E., Foekens, J. A., Greaves, M., Hosoda, F., Hutter, B., Ilicic, T., Imbeaud, S., Imielinski, M., Imielinsk, M., Jäger, N., Jones, D. T. W., Jones, D., Knappskog, S., Kool, M., Lakhani, S. R., López-Otín, C., Martin, S., Munshi, N. C., Nakamura, H., Northcott, P. A., Pajic, M., Papaemmanuil, E., Paradiso, A., Pearson, J. V., Puente, X. S., Raine, K., Ramakrishna, M., Richardson, A. L., Richter, J., Rosenstiel, P., Schlesner, M., Schumacher, T. N., Span, P. N., Teague, J. W., Totoki, Y., Tutt, A. N. J., Valdés-Mas, R., van Buuren, M. M., van 't Veer, L., Vincent-Salomon, A., Waddell, N., Yates, L. R., Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, Zucman-Rossi, J., Futreal, P. A., McDermott, U., Lichter, P., Meyerson, M., Grimmond, S. M., Siebert, R., Campo, E., Shibata, T., Pfister, S. M., Campbell, P. J., and Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–21.

Altman, E., Yango, P., Moustafa, R., Smith, J. F., Klatsky, P. C., and Tran, N. D. (2014). Characterization of human spermatogonial stem cell markers in fetal, pediatric, and adult testicular tissues. *Reproduction*, 148(4):417–427.

Andrianova, M. A., Bazykin, G. A., Nikolaev, S. I., and Seplyarskiy, V. B. (2017). Human mismatch repair system balances mutation rates between strands by removing more mismatches from the lagging strand. *Genome Research*, 27(8):1336–1343.

Baek, S., Goldstein, I., and Hager, G. L. (2017). Bivariate Genomic Footprinting Detects Changes in Transcription Factor Activity. *Cell Reports*, 19(8):1710–1722.

Bailey, S. D., Zhang, X., Desai, K., Aid, M., Corradin, O., Cowper-Sallari, R., Akhtar-Zaidi, B., Scacheri, P. C., Haibe-Kains, B., and Lupien, M. (2015). ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nature Communications*, 2.

Berk, A. J. and Sharp, P. A. (1977). Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease-digested hybrids. *Cell*, 12(3):721–732.

Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J.,

Fry, B., Meissner, A., Wernig, M., Plath, K., Jaenisch, R., Wagschal, A., Feil, R., Schreiber, S. L., and Lander, E. S. (2006). A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell*, 125(2):315–326.

Blake, R. D., Hess, S. T., and Nicholson-Tuell, J. (1992). The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. *Journal of Molecular Evolution*, 34(3):189–200.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smith, A. F., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., Haussler, D., and Miller, W. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research*, 14(4):708–715.

Blanco, J. D., Radusky, L., Climente-González, H., and Serrano, L. (2018). FoldX accurate structural protein–DNA binding prediction using PADA1 (Protein Assisted DNA Assembly 1). *Nucleic Acids Research*, 46(8):3852–3863.

Boyle, A. P., Song, L., Lee, B.-K., London, D., Keefe, D., Birney, E., Iyer, V. R., Crawford, G. E., and Furey, T. S. (2011). High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome research*, 21(3):456–64.

Brenner, S., Jacob, F., and Meselson, M. (1961). An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature*, 190(4776):576–581.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods*, 10(12):1213–8.

Bulger, M. and Groudine, M. (2011). Functional and mechanistic diversity of distal transcription enhancers. *Cell*, 144(3):327–339.

Burgers, P. M. J. and Kunkel, T. A. (2017). Eukaryotic DNA Replication Fork. *Annual review of biochemistry*, 86(1):417–438.

C Yuen, R. K., Merico, D., Bookman, M., L Howe, J., Thiruvahindrapuram, B., Patel, R. V., Whitney, J., Deflaux, N., Bingham, J., Wang, Z., Pellecchia, G., Buchanan, J. A., Walker, S., Marshall, C. R., Uddin, M., Zarrei, M., Deneault, E., D'Abate,

L., Chan, A. J. S., Koyanagi, S., Paton, T., Pereira, S. L., Hoang, N., Engchuan, W., Higginbotham, E. J., Ho, K., Lamoureux, S., Li, W., MacDonald, J. R., Nalpathamkalam, T., Sung, W. W. L., Tsoi, F. J., Wei, J., Xu, L., Tasse, A.-M., Kirby, E., Van Etten, W., Twigger, S., Roberts, W., Drmic, I., Jilderda, S., Modi, B. M., Kellam, B., Szego, M., Cytrynbaum, C., Weksberg, R., Zwaigenbaum, L., Woodbury-Smith, M., Brian, J., Senman, L., Iaboni, A., Doyle-Thomas, K., Thompson, A., Chrysler, C., Leef, J., Savion-Lemieux, T., Smith, I. M., Liu, X., Nicolson, R., Seifer, V., Fedele, A., Cook, E. H., Dager, S., Estes, A., Gallagher, L., Malow, B. A., Parr, J. R., Spence, S. J., Vorstman, J., Frey, B. J., Robinson, J. T., Strug, L. J., Fernandez, B. A., Elsabbagh, M., Carter, M. T., Hallmayer, J., Knoppers, B. M., Anagnostou, E., Szatmari, P., Ring, R. H., Glazer, D., Pletcher, M. T., and Scherer, S. W. (2017). Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nature neuroscience*, 20(4):602–611.

Calo, E. and Wysocka, J. (2013). Modification of Enhancer Chromatin: What, How, and Why? *Molecular Cell*, 49(5):825–837.

Calviello, A. K., Hirsekorn, A., Wurmus, R., Yusuf, D., and Ohler, U. (2018). Reproducible inference of transcription factor footprints in ATAC-seq and DNase-seq datasets via protocol-specific bias modeling. *bioRxiv*.

Campbell, C. D. and Eichler, E. E. (2013). Properties and rates of germline mutations in humans. *Trends in Genetics*, 29(10):575–584.

Campbell, I. M., Shaw, C. A., Stankiewicz, P., and Lupski, J. R. (2015). Somatic mosaicism : implications for disease and transmission genetics. *Trends in Genetics*, 31(7):382–392.

Campbell, P. J. (2016). Somatic mutation in cancer and normal cells. *Science*, 349(6255):961–968.

Cech, T. R. and Steitz, J. A. (2014). The noncoding RNA revolution - Trashing old rules to forge new ones. *Cell*, 157(1):77–94.

Chan, K. and Gordenin, D. A. (2016). Clusters of Multiple Mutations: Incidence and Molecular Mechanisms. *Annual review of genetics*, (12):243–267.

Chatterjee, N. and Walker, G. C. (2017). Mechanisms of DNA damage, repair, and mutagenesis. *Environmental and Molecular Mutagenesis*, 58(5):235–263.

Chen, H., Tian, Y., Shu, W., Bo, X., and Wang, S. (2012a). Comprehensive identification and annotation of cell type-specific and ubiquitous CTCF-binding sites in the human genome. *PLoS ONE*, 7(7).

Chen, X., Chen, Z., Chen, H., Su, Z., Yang, J., Lin, F., Shi, S., and He, X. (2012b). Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes. *Science (New York, N.Y.)*, 335(6073):1235–8.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., Loh, Y. H., Yeo, H. C., Yeo, Z. X., Narang, V., Govindarajan, K. R., Leong, B., Shahab, A., Ruan, Y., Bourque, G., Sung, W. K., Clarke, N. D., Wei, C. L., and Ng, H. H. (2008). Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. *Cell*, 133(6):1106–1117.

Chèneby, J., Gheorghe, M., Artufel, M., Mathelier, A., and Ballester, B. (2018). ReMap 2018: An updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Research*, 46(D1):D267–D275.

Chow, L. T., Roberts, J. M., Lewis, J. B., and Broker, T. R. (1977). A map of cytoplasmic RNA transcripts from lytic adenovirus type 2, determined by electron microscopy of RNA:DNA hybrids. *Cell*, 11(4):819–836.

Ciccia, A. and Elledge, S. J. (2010). The DNA Damage Response: Making It Safe to Play with Knives. *Molecular Cell*, 40(2):179–204.

Cleaver, J. E. (1968). Defective Repair Replication of DNA in Xermoderma Pigmentosum. *Nature*, 218:652–656.

Cohen, S. A. and Leininger, A. (2014). Cohen, Leininger - 2014 - The genetic basis of Lynch syndrome and its implications for clinical practice and risk management. pages 147–158.

Conkright, M. D., Guzmán, E., Flechner, L., Su, A. I., Hogenesch, J. B., and Montminy, M. (2003). Genome-wide analysis of CREB target genes reveals a core promoter requirement for cAMP responsiveness. *Molecular Cell*, 11(4):1101–1108.

Corces, M. R., Granja, J. M., Shams, S., Louie, B. H., Seoane, J. A., Zhou, W., Silva, T. C., Groeneveld, C., Wong, C. K., Cho, S. W., Satpathy, A. T., Mumbach, M. R., Hoadley, K. A., Robertson, A. G., Sheffield, N. C., Felau, I., Castro, M. A., Berman, B. P., Staudt, L. M., Zenklusen, J. C., Laird, P. W., Curtis, C., Greenleaf, W. J., and Chang, H. Y. (2018). The chromatin accessibility landscape of primary human cancers. *Science*, 362(6413).

Cremer, T., Cremer, C., Not at Dartmouth/Dhmclibraries, and on interlibrary loan, R. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Reviews Genetics*, 2(4):292–301.

Crick, F. H. (1958). On protein synthesis. *Symposia of the Society for Experimental Biology*, 12:138–163.

Cusanovich, D. A., Hill, A. J., Aghamirzaie, D., Daza, R. M., Pliner, H. A., Berletch, J. B., Filippova, G. N., Huang, X., Christiansen, L., DeWitt, W. S., Lee, C., Regalado, S. G., Read, D. F., Steemers, F. J., Disteche, C. M., Trapnell, C., and Shendure, J. (2018). A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell*, pages 1–16.

Cutter, A. and Hayes, J. J. (2015). A Brief Review of Nucleosome Structure. *FEBS Lett*, 589:2914–2922.

Daigaku, Y., Keszthelyi, A., Müller, C. A., Miyabe, I., Brooks, T., Retkute, R., Hubank, M., Nieduszynski, C. A., and Carr, A. M. (2015). A global profile of replicative polymerase usage. *Nature Structural & Molecular Biology*, 22(3):192–198.

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., and Batzoglou, S. (2010). Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Computational Biology*, 6(12):e1001025.

De, S. (2011). Somatic mosaicism in healthy human tissues. *Trends in Genetics*, 27(6):217–223.

Deaton, A. M. and Bird, A. (2011). CpG islands and the regulation of transcription. *Genes & development*, 25(10):1010–22.

Di Persio, S., Saracino, R., Fera, S., Muciaccia, B., Esposito, V., Boitani, C., Berloco,

B. P., Nudo, F., Spadetta, G., Stefanini, M., de Rooij, D. G., and Vicini, E. (2017). Spermatogonial kinetics in humans. *Development*, 144(19):3430–3439.

Dixon, J. R., Gorkin, D. U., and Ren, B. (2016). Chromatin Domains: The Unit of Chromosome Organization. *Molecular Cell*, 62(5):668–680.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380.

Dravis, C., Chung, C.-Y., Lytle, N. K., Herrera-Valdez, J., Luna, G., Trejo, C. L., Reya, T., and Wahl, G. M. (2018). Epigenetic and Transcriptomic Profiling of Mammary Gland Development and Tumor Models Disclose Regulators of Cell State Plasticity. *SSRN Electronic Journal*.

Drost, J. B. and Lee, W. R. (1995). Biological basis of germline mutation: comparisons of spontaneous germline mutation rates among drosophila, mouse, and human. *Environmental and molecular mutagenesis*, 25 Suppl 2(1 995):48–64.

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B. K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., Shoresh, N., Simon, J. M., Song, L., Trinklein, N. D., Altshuler, R. C., Birney, E., Brown, J. B., Cheng, C., Djebali, S., Dong, X., Ernst, J., Furey, T. S., Gerstein, M., Giardine, B., Greven, M., Hardison, R. C., Harris, R. S., Herrero, J., Hoffman, M. M., Iyer, S., Kellis, M., Kheradpour, P., Lassmann, T., Li, Q., Lin, X., Marinov, G. K., Merkel, A., Mortazavi, A., Parker, S. C., Reddy, T. E., Rozowsky, J., Schlesinger, F., Thurman, R. E., Wang, J., Ward, L. D., Whitfield, T. W., Wilder, S. P., Wu, W., Xi, H. S., Yip, K. Y., Zhuang, J., Bernstein, B. E., Green, E. D., Gunter, C., Snyder, M., Pazin, M. J., Lowdon, R. F., Dillon, L. A., Adams, L. B., Kelly, C. J., Zhang, J., Wexler, J. R., Good, P. J., Feingold, E. A., Crawford, G. E., Dekker, J., Elnitski, L., Farnham, P. J., Giddings, M. C., Gingeras, T. R., Guigó, R., Hubbard, T. J., Kent, W. J., Lieb, J. D., Margulies, E. H., Myers, R. M., Stamatoyannopoulos, J. A., Tenenbaum, S. A., Weng, Z., White, K. P., Wold, B., Yu, Y., Wrobel, J., Risk, B. A., Gunawardena, H. P., Kuiper, H. C., Maier, C. W., Xie, L., Chen, X., Mikkelsen, T. S., Gillespie, S., Goren, A., Ram, O., Zhang, X., Wang, L., Issner, R.,

Coyne, M. J., Durham, T., Ku, M., Truong, T., Eaton, M. L., Dobin, A., Tanzer, A., Lagarde, J., Lin, W., Xue, C., Williams, B. A., Zaleski, C., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Batut, P., Bell, I., Bell, K., Chakrabortty, S., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Li, G., Luo, O. J., Park, E., Preall, J. B., Presaud, K., Ribeca, P., Robyr, D., Ruan, X., Sammeth, M., Sandhu, K. S., Schaeffer, L., See, L. H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Hayashizaki, Y., Reymond, A., Antonarakis, S. E., Hannon, G. J., Ruan, Y., Carninci, P., Sloan, C. A., Learned, K., Malladi, V. S., Wong, M. C., Barber, G. P., Cline, M. S., Dreszer, T. R., Heitner, S. G., Karolchik, D., Kirkup, V. M., Meyer, L. R., Long, J. C., Maddren, M., Raney, B. J., Grasfeder, L. L., Giresi, P. G., Battenhouse, A., Sheffield, N. C., Showers, K. A., London, D., Bhinge, A. A., Shestak, C., Schaner, M. R., Kim, S. K., Zhang, Z. Z., Mieczkowski, P. A., Mieczkowska, J. O., Liu, Z., McDaniell, R. M., Ni, Y., Rashid, N. U., Kim, M. J., Adar, S., Zhang, Z., Wang, T., Winter, D., Keefe, D., Iyer, V. R., Zheng, M., Wang, P., Gertz, J., Vielmetter, J., Partridge, E. C., Varley, K. E., Gasper, C., Bansal, A., Pepke, S., Jain, P., Amrhein, H., Bowling, K. M., Anaya, M., Cross, M. K., Muratet, M. A., Newberry, K. M., McCue, K., Nesmith, A. S., Fisher-Aylor, K. I., Pusey, B., DeSalvo, G., Parker, S. L., Balasubramanian, S., Davis, N. S., Meadows, S. K., Eggleston, T., Newberry, J. S., Levy, S. E., Absher, D. M., Wong, W. H., Blow, M. J., Visel, A., Pennachio, L. A., Petrykowska, H. M., Abyzov, A., Aken, B., Barrell, D., Barson, G., Berry, A., Bignell, A., Boychenko, V., Bussotti, G., Davidson, C., Despacio-Reyes, G., Diekhans, M., Ezkurdia, I., Frankish, A., Gilbert, J., Gonzalez, J. M., Griffiths, E., Harte, R., Hendrix, D. A., Hunt, T., Jungreis, I., Kay, M., Khurana, E., Leng, J., Lin, M. F., Loveland, J., Lu, Z., Manthravadi, D., Mariotti, M., Mudge, J., Mukherjee, G., Notredame, C., Pei, B., Rodriguez, J. M., Saunders, G., Sboner, A., Searle, S., Sisu, C., Snow, C., Steward, C., Tapanari, E., Tress, M. L., Van Baren, M. J., Washietl, S., Wilming, L., Zadissa, A., Zhang, Z., Brent, M., Haussler, D., Valencia, A., Addleman, N., Alexander, R. P., Auerbach, R. K., Balasubramanian, S., Bettinger, K., Bhardwaj, N., Boyle, A. P., Cao, A. R., Cayting, P., Charos, A., Cheng, Y., Eastman, C., Euskirchen, G., Fleming, J. D.,

Grubert, F., Habegger, L., Hariharan, M., Harmanci, A., Iyengar, S., Jin, V. X., Karczewski, K. J., Kasowski, M., Lacroute, P., Lam, H., Lamarre-Vincent, N., Lian, J., Lindahl-Allen, M., Min, R., Miotto, B., Monahan, H., Moqtaderi, Z., Mu, X. J., O'Geen, H., Ouyang, Z., Patacsil, D., Raha, D., Ramirez, L., Reed, B., Shi, M., Slifer, T., Witt, H., Wu, L., Xu, X., Yan, K. K., Yang, X., Struhl, K., Weissman, S. M., Penalva, L. O., Karmakar, S., Bhanvadia, R. R., Choudhury, A., Domanus, M., Ma, L., Moran, J., Victorsen, A., Auer, T., Centanin, L., Eichenlaub, M., Gruhl, F., Heermann, S., Hoeckendorf, B., Inoue, D., Kellner, T., Kirchmaier, S., Mueller, C., Reinhardt, R., Schertel, L., Schneider, S., Sinn, R., Wittbrodt, B., Wittbrodt, J., Jain, G., Balasundaram, G., Bates, D. L., Byron, R., Canfi (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

Durbin, R. M., Altshuler, D. L., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., and Others, A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073.

Eijk, P. V., Nandi, S. P., Yu, S., Bennett, M., Leadbitter, M., Teng, Y., and Reed, S. H. (2018). Nucleosome remodelling at origins of Global Genome-Nucleotide Excision Repair occurs at the boundaries of higher-order chromatin structure. *bioRxiv*, 62716(74219):1–45.

Eisenberg, E. and Levanon, E. Y. (2013). Human housekeeping genes, revisited. *Trends in Genetics*, 29(10):569–574.

Elkon, R. and Agami, R. (2017). Characterization of noncoding regulatory DNA in the human genome. *Nature Biotechnology*, 35(8):732–746.

Erdel, F. and Rippe, K. (2018). Formation of Chromatin Subcompartments by Phase Separation. *Biophysical Journal*, 114(10):2262–2270.

Ewen, K. A., Olesen, I. A., Winge, S. B., Nielsen, A. R., Nielsen, J. E., Graem, N., Juul, A., and Rajpert-De Meyts, E. (2013). Expression of FGFR3 during human testis development and in germ cell-derived tumours of young adults. *International Journal of Developmental Biology*, 57(2-4):141–151.

Faili, A. and Gue, Q. (2009). Competitive repair pathways in immunoglobulin gene hypermutation. *Philisophical Transactions Royal Society*, (November 2008):613–619.

Fay, J. C., Wyckoff, G. J., and Wu, C. I. (2001). Positive and negative selection on the human genome. *Genetics*, 158(3):1227–34.

Fei, J. and Ha, T. (2013). Watching DNA breath one molecule at a time. *Proceedings of the National Academy of Sciences of the United States of America*, 110(43):17173–4.

Feng, W., Kawauchi, D., Körkel-Qu, H., Deng, H., Serger, E., Sieber, L., Lieberman, J. A., Jimeno-González, S., Lambo, S., Hanna, B. S., Harim, Y., Jansen, M., Neuerburg, A., Friesen, O., Zuckermann, M., Rajendran, V., Gronych, J., Ayrault, O., Korshunov, A., Jones, D. T. W., Kool, M., Northcott, P. A., Lichter, P., Cortés-Ledesma, F., Pfister, S. M., and Liu, H.-K. (2017). Chd7 is indispensable for mammalian brain development through activation of a neuronal differentiation programme. *Nature communications*, 8:14758.

Filippova, G. N., Fagerlie, S., Klenova, E. M., Myers, C., Dehner, Y., Goodwin, G., Neiman, P. E., Collins, S. J., and Lobanenkov, V. V. (1996). An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Molecular and Cellular Biology*, 16(6):2802–2813.

Flemming, W. (1879). Contributions to the knowledge of the cell and its life phenomena. *Archiv für Mikroskopische Anatomie*, 16(1):302–436.

Forbes, C. M., Flannigan, R., and Schlegel, P. N. (2018). Spermatogonial stem cell transplantation and male infertility: Current status and future directions. *Arab Journal of Urology*, 16(1):171–180.

Forrest, A. R. R., Kawaji, H., Rehli, M., Baillie, J. K., de Hoon, M. J. L., Haberle, V., Lassmann, T., Kulakovskiy, I. V., Lizio, M., Itoh, M., Andersson, R., Mungall, C. J., Meehan, T. F., Schmeier, S., Bertin, N., Jørgensen, M., Dimont, E., Arner, E., Schmidl, C., Schaefer, U., Medvedeva, Y. A., Plessy, C., Vitezic, M., Severin, J., Semple, C. A., Ishizu, Y., Young, R. S., Francescatto, M., Alam, I., Albanese, D., Altschuler, G. M., Arakawa, T., Archer, J. A. C., Arner, P., Babina, M., Rennie, S., Balwierz, P. J., Beckhouse, A. G., Pradhan-Bhatt, S., Blake, J. A., Blumenthal, A., Bodega, B., Bonetti, A., Briggs, J., Brombacher, F., Burroughs, A. M., Califano, A., Cannistraci, C. V., Carbajo, D., Chen, Y., Chierici, M., Ciani, Y., Clevers, H. C., Dalla, E., Davis, C. A., Detmar, M., Diehl, A. D., Dohi, T., Drabløs, F., Edge, A.

S. B., Edinger, M., Ekwall, K., Endoh, M., Enomoto, H., Fagiolini, M., Fairbairn, L., Fang, H., Farach-Carson, M. C., Faulkner, G. J., Favorov, A. V., Fisher, M. E., Frith, M. C., Fujita, R., Fukuda, S., Furlanello, C., Furino, M., Furusawa, J.-i., Geijtenbeek, T. B., Gibson, A. P., Gingeras, T., Goldowitz, D., Gough, J., Guhl, S., Guler, R., Gustincich, S., Ha, T. J., Hamaguchi, M., Hara, M., Harbers, M., Harshbarger, J., Hasegawa, A., Hasegawa, Y., Hashimoto, T., Herlyn, M., Hitchens, K. J., Ho Sui, S. J., Hofmann, O. M., Hoof, I., Hori, F., Huminiecki, L., Iida, K., Ikawa, T., Jankovic, B. R., Jia, H., Joshi, A., Jurman, G., Kaczkowski, B., Kai, C., Kaida, K., Kaiho, A., Kajiyama, K., Kanamori-Katayama, M., Kasianov, A. S., Kasukawa, T., Katayama, S., Kato, S., Kawaguchi, S., Kawamoto, H., Kawamura, Y. I., Kawashima, T., Kempfle, J. S., Kenna, T. J., Kere, J., Khachigian, L. M., Kitamura, T., Klinken, S. P., Knox, A. J., Kojima, M., Kojima, S., Kondo, N., Koseki, H., Koyasu, S., Krampitz, S., Kubosaki, A., Kwon, A. T., Laros, J. F. J., Lee, W., Lennartsson, A., Li, K., Lilje, B., Lipovich, L., Mackay-Sim, A., Manabe, R.-i., Mar, J. C., Marchand, B., Mathelier, A., Mejhert, N., Meynert, A., Mizuno, Y., de Lima Morais, D. A., Morikawa, H., Morimoto, M., Moro, K., Motakis, E., Motohashi, H., Mummery, C. L., Murata, M., Nagao-Sato, S., Nakachi, Y., Nakahara, F., Nakamura, T., Nakamura, Y., Nakazato, K., van Nimwegen, E., Ninomiya, N., Nishiyori, H., Noma, S., Noma, S., Noazaki, T., Ogishima, S., Ohkura, N., Ohimiya, H., Ohno, H., Ohshima, M., Okada-Hatakeyama, M., Okazaki, Y., Orlando, V., Ovchinnikov, D. A., Pain, A., Passier, R., Patrikakis, M., Persson, H., Piazza, S., Prendergast, J. G. D., Rackham, O. J. L., Ramilowski, J. A., Rashid, M., Ravasi, T., Rizzu, P., Roncador, M., Roy, S., Rye, M. B., Saijyo, E., Sajantila, A., Saka, A., Sakaguchi, S., Sakai, M., Sato, H., Savvi, S., Saxena, A., Schneider, C., Schultes, E. A., Schulze-Tanzil, G. G., Schwegmann, A., Sengstag, T., Sheng, G., Shimoji, H., Shimoni, Y., Shin, J. W., Simon, C., Sugiyama, D., Sugiyama, T., Suzuki, M., Suzuki, N., Swoboda, R. K., 't Hoen, P. A. C., Tagami, M., Takahashi, N., Takai, J., Tanaka, H., Tatsukawa, H., Tatum, Z., Thompson, M., Toyodo, H., Toyoda, T., Valen, E., van de Wetering, M., van den Berg, L. M., Verado, R., Vijayan, D., Vorontsov, I. E., Wasserman, W. W., Watanabe, S., Wells, C. A., Winteringham, L. N., Wolvetang, E., Wood, E. J., Yamaguchi, Y., Yamamoto, M., Yoneda, M., Yonekura, Y., Yoshida, S., Zabierowski, S. E., Zhang, P. G., Zhao, X., Zucchelli, S., Summers, K. M., Suzuki, H., Daub, C. O., Kawai, J., Heutink, P., Hide, W., Freeman, T. C., Lenhard, B.,

Bajic, V. B., Taylor, M. S., Makeev, V. J., Sandelin, A., Hume, D. A., Carninci, P., and Hayashizaki, Y. (2014). A promoter-level mammalian expression atlas. *Nature*, 507(7493):462–70.

Francioli, L. C., Polak, P. P., Koren, A., Menelaou, A., Chun, S., Renkens, I., van Duijn, C. M., Swertz, M., Wijmenga, C., van Ommen, G., Slagboom, P. E., Boomsma, D. I., Ye, K., Guryev, V., Arndt, P. F., Kloosterman, W. P., de Bakker, P. I. W., and Sunyaev, S. R. (2015). Genome-wide patterns and properties of de novo mutations in humans. *Nature Genetics*, (October 2014).

Frigola, J., Sabarinathan, R., Mularoni, L., Muiños, F., Gonzalez-Perez, A., and López-Bigas, N. (2017). Reduced mutation rate in exons due to differential mismatch repair. *Nature Genetics*, (September 2016).

Fulton, D. L., Sundararajan, S., Badis, G., Hughes, T. R., Wasserman, W. W., Roach, J. C., and Sladek, R. (2009). TFCat: The curated catalog of mouse and human transcription factors. *Genome Biology*, 10(3).

Gallardo, T., Shirley, L., John, G. B., and Castrillon, D. H. (2007). Generation of a germ cell-specific mouse transgenic Cre line, Vasa-Cre. *Genesis (New York, N.Y. : 2000)*, 45(6):413–7.

Ganai, R. A. and Johansson, E. (2016). DNA Replication - a Matter of Fidelity. *Molecular Cell*, 62(5):745–755.

Gao, Z., Moorjani, P., Amster, G., and Przeworski, M. (2018). Overlooked roles of DNA damage and maternal age in generating human germline mutations. *bioRxiv*.

Garbacz, M. A., Lujan, S. A., Burkholder, A. B., Cox, P. B., Wu, Q., Zhou, Z.-x., Haber, J. E., and Kunkel, T. A. (2018). Evidence that DNA polymerase δ contributes to initiating leading strand DNA replication in Saccharomyces cerevisiae Marta. *Nature Communications*, (9):1–11.

Gardiner-Garden, M. and Frommer, M. (1987). CpG islands in vertebrate genomes. *Journal of molecular biology*, 196(2):261–82.

Garinis, G. A., van der Horst, G. T., Vijg, J., and H.J. Hoeijmakers, J. (2008). DNA damage and ageing: new-age ideas for an age-old problem. *Nature Cell Biology*, 10(11):1241–1247.

Ghaleb, A. M. and Yang, V. W. (2017). Krüppel-like factor 4 (KLF4): What we currently know. *Gene*, 611(1):27–37.

Gibbs, R. A., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J. G., Zhu, Y., Wang, J., Chang, Y., Feng, Q., Fang, X., Guo, X., Jian, M., Jiang, H., Jin, X., Lan, T., Li, G., Li, J., Li, Y., Liu, S., Liu, X., Lu, Y., Ma, X., Tang, M., Wang, B., Wang, G., Wu, H., Wu, R., Xu, X., Yin, Y., Zhang, D., Zhang, W., Zhao, J., Zhao, M., Zheng, X., Lander, E. S., Altshuler, D. M., Gabriel, S. B., Gupta, N., Gharani, N., Toji, L. H., Gerry, N. P., Resch, A. M., Flicek, P., Barker, J., Clarke, L., Gil, L., Hunt, S. E., Kelman, G., Kulesha, E., Leinonen, R., McLaren, W. M., Radhakrishnan, R., Roa, A., Smirnov, D., Smith, R. E., Streeter, I., Thormann, A., Toneva, I., Vaughan, B., Zheng-Bradley, X., Bentley, D. R., Grocock, R., Humphray, S., James, T., Kingsbury, Z., Lehrach, H., Sudbrak, R., Albrecht, M. W., Amstislavskiy, V. S., Borodina, T. A., Lienhard, M., Mertes, F., Sultan, M., Timmermann, B., Yaspo, M.-L., Mardis, E. R., Wilson, R. K., Fulton, L., Fulton, R., Sherry, S. T., Ananiev, V., Belaia, Z., Beloslyudtsev, D., Bouk, N., Chen, C., Church, D., Cohen, R., Cook, C., Garner, J., Hefferon, T., Kimelman, M., Liu, C., Lopez, J., Meric, P., O'Sullivan, C., Ostapchuk, Y., Phan, L., Ponomarov, S., Schneider, V., Shekhtman, E., Sirotkin, K., Slotta, D., Zhang, H., McVean, G. A., Durbin, R. M., Balasubramaniam, S., Burton, J., Danecek, P., Keane, T. M., Kolb-Kokocinski, A., McCarthy, S., Stalker, J., Quail, M., Schmidt, J. P., Davies, C. J., Gollub, J., Webster, T., Wong, B., Zhan, Y., Auton, A., Campbell, C. L., Kong, Y., Marcketta, A., Gibbs, R. A., Yu, F., Antunes, L., Bainbridge, M., Muzny, D., Sabo, A., Huang, Z., Wang, J., Coin, L. J. M., Fang, L., Guo, X., Jin, X., Li, G., Li, Q., Li, Y., Li, Z., Lin, H., Liu, B., Luo, R., Shao, H., Xie, Y., Ye, C., Yu, C., Zhang, F., Zheng, H., Zhu, H., Alkan, C., Dal, E., Kahveci, F., Marth, G. T., Garrison, E. P., Kural, D., Lee, W.-P., Fung Leong, W., Stromberg, M., Ward, A. N., Wu, J., Zhang, M., Daly, M. J., DePristo, M. A., Handsaker, R. E., Altshuler, D. M., Banks, E., Bhatia, G., del Angel, G., Gabriel, S. B., Genovese, G., Gupta, N., Li, H., Kashin, S., Lander, E. S., McCarroll, S. A., Nemesh, J. C., Poplin, R. E., Yoon, S. C., Lihm, J., Makarov, V., Clark, A. G., Gottipati, S., Keinan, A., Rodriguez-Flores, J. L., Korbel, J. O., Rausch, T., Fritz, M. H., Stütz, A. M., Flicek, P., Beal, K., Clarke, L., Datta, A., Herrero, J., McLaren, W. M., Ritchie, G. R. S., Smith, R. E., Zerbino, D., Zheng-Bradley, X., Sabeti, P. C., Shlyakhter, I., Schaffner, S. F., Vitti,

J., Cooper, D. N., Ball, E. V., Stenson, P. D., Bentley, D. R., Barnes, B., Bauer, M., Keira Cheetham, R., Cox, A., Eberle, M., Humphray, S., Kahn, S., Murray, L., Peden, J., Shaw, R., Kenny, E. E., Batzer, M. A., Konkel, M. K., Walker, J. A., MacArthur, D. G., Lek, M., Sudbrak, R., Amstislavskiy, V. S., Herwig, R., Mardis, E. R., Ding, L., Koboldt, D. C., Larson, D., Ye, K., Gravel, S., Swaroop, A., Chew, E., Lappalainen, T., Erlich, Y., Gymrek, M., Frederick Willems, T., Simpson, J. T., Shriver, M. D., Rosenfeld, J. A., Bustamante, C. D., Montgomery, S. B., De La Vega, F. M., Byrnes, J. K., Carroll, A. W., DeGorter, M. K., Lacroute, P., Maples, B. K., Martin, A. R., Moreno-Estrada, A., Shringarpure, S. S., Zakharia, F., Halperin, E., Baran, Y., Lee, C., Cerveira, E., Hwang, J., Malhotra, A., Plewczynski, D., Radew, K., Romanovitch, M., Zhang, C., Hyland, F. C. L., Craig, D. W., Christoforides, A., Homer, N., Izatt, T., Kurdoglu, A. A., Sinari, S. A., Squire, K., Sherry, S. T., Xiao, C., Sebat, J., Antaki, D., Gujral, M., Noor, A., Ye, K., Burchard, E. G., Hernandez, R. D., Gignoux, C. R., Haussler, D., Katzman, S. J., James Kent, W., Howie, B., Ruiz-Linares, A., Dermitzakis, E. T., Devine, S. E., Abecasis, G. R., Min Kang, H., Kidd, J. M., Blackwell, T., Caron, S., Chen, W., Emery, S., Fritsche, L., Fuchsberger, C., Jun, G., Li, B., Lyons, R., Scheller, C., Sidore, C., Song, S., Sliwerska, E., Taliun, D., Tan, A., Welch, R., Kate Wing, M., Zhan, X., Awadalla, P., Hodgkinson, A., Li, Y., Shi, X., Quitadamo, A., Lunter, G., McVean, G. A., Marchini, J. L., Myers, S., Churchhouse, C., Delaneau, O., Gupta-Hinch, A., Kretzschmar, W., Iqbal, Z., Mathieson, I., Menelaou, A., Rimmer, A., Xifara, D. K., Oleksyk, T. K., Fu, Y., Liu, X., Xiong, M., Jorde, L., Witherspoon, D., Xing, J., Eichler, E. E., Browning, B. L., Browning, S. R., Hormozdiari, F., Sudmant, P. H., Khurana, E., Durbin, R. M., Hurles, M. E., Tyler-Smith, C., Albers, C. A., Ayub, Q., Balasubramaniam, S., Chen, Y., Colonna, V., Danecek, P., Jostins, L., Keane, T. M., McCarthy, S., Walter, K., Xue, Y., Gerstein, M. B., Abyzov, A., Balasubramanian, S., Chen, J., Clarke, D., Fu, Y., Harmanci, A. O., Jin, M., Lee, D., Liu, J., Jasmine Mu, X., Zhang, J., Zhang, Y., Li, Y., Luo, R., Zhu, H., Alkan, C., Dal, E., Kahveci, F., Marth, G. T., Garrison, E. P., Kural, D., Lee, W.-P., Ward, A. N., Wu, J., (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

Gittelman, R. M., Hun, E., Ay, F., Madeoy, J., Pennacchio, L., Noble, W. S., Hawkins, R. D., and Akey, J. M. (2015). Comprehensive identification and analysis of human accelerated regulatory DNA. *Genome research*, 25(9):1245–55.

Goldmann, J. M., Seplyarskiy, V. B., Wong, W. S. W., Vilboux, T., Neerincx, P. B., Bodian, D. L., Solomon, B. D., Veltman, J. A., and Deeken, J. F. (2018). Germline de novo mutation clusters arise during oocyte aging in genomic regions with high double-strand-break incidence. *Nature Genetics*, 50(April).

Goldmann, J. M., Wong, W. S., Pinelli, M., Farrah, T., Bodian, D., Stittrich, A. B., Glusman, G., Vissers, L. E., Hoischen, A., Roach, J. C., Vockley, J. G., Veltman, J. A., Solomon, B. D., Gilissen, C., and Niederhuber, J. E. (2016). Parent-of-origin-specific signatures of de novo mutations. *Nature Genetics*, 48(8):935–939.

Goriely, A., McGrath, J. J., Hultman, C. M., Wilkie, A. O. M., and Malaspina, D. (2013). "Selfish spermatogonial selection": a novel mechanism for the association between advanced paternal age and neurodevelopmental disorders. *The American journal of psychiatry*, 170(6):599–608.

Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: Scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018.

Groudine, M. and Conkin, K. F. (1985). Chromatin structure and de novo methylation of sperm DNA: implications for activation of the paternal genome. *Science (New York, N.Y.)*, 228(4703):1061–8.

Guo, F., Yan, L., Guo, H., Li, L., Hu, B., Zhao, Y., Yong, J., Hu, Y., Wang, X., Wei, Y., Wang, W., Li, R., Yan, J., Zhi, X., Zhang, Y., Jin, H., Zhang, W., Hou, Y., Zhu, P., Li, J., Zhang, L., Liu, S., Ren, Y., Zhu, X., Wen, L., Gao, Y. Q., Tang, F., and Qiao, J. (2015). The Transcriptome and DNA Methylome Landscapes of Human Primordial Germ Cells. *Cell*, 161(6):1437–1452.

Guo, J., Grow, E. J., Yi, C., Mlcochova, H., Maher, G. J., Lindskog, C., Murphy, P. J., Wike, C. L., Carrell, D. T., Goriely, A., Hotaling, J. M., and Cairns, B. R. (2017). Chromatin and Single-Cell RNA-Seq Profiling Reveal Dynamic Signaling and Metabolic Transitions during Human Spermatogonial Stem Cell Development. *Cell Stem Cell*, 21(4):533–546.e6.

Gusmao, E. G., Allhoff, M., Zenke, M., and Costa, I. G. (2016). Analysis of computational footprinting methods for DNase sequencing experiments. *Nature methods*, 13(4):303–9.

Haberle, V. and Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. *Nature reviews. Molecular cell biology*, page 1.

Hall, M. D., Yasgar, A., Peryea, T., Braisted, J. C., Jadhav, A., and Coussens, N. P. (2017). Fluorescence polarization assays in high-throughput screening and drug discovery: a review. *Methods and Applications in Fluorescence*, 4(2):1–41.

Halligan, D. L., Kousathanas, A., Ness, R. W., Harr, B., Eöry, L., Keane, T. M., Adams, D. J., and Keightley, P. D. (2013). Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. *PLoS genetics*, 9(12):e1003995.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of Cancer: The Next Generation. *Cell*, 144(5):646–674.

Hanawalt, P. C. and Spivak, G. (2008). Transcription-coupled DNA repair: two decades of progress and surprises. *Nature reviews. Molecular cell biology*, 9(12):958–970.

Hashimoto, H., Olanrewaju, Y. O., Zheng, Y., Wilson, G. G., and Zhang, X. (2014). Wilms tumor protein recognizes 5-carboxylcytosine within a specific DNA sequence. 4:2304–2313.

Hashimoto, H., Wang, D., Horton, J. R., Zhang, X., Corces, V. G., and Cheng, X. (2017). Structural Basis for the Versatile and Methylation-Dependent Binding of CTCF to DNA. *Molecular Cell*, pages 1–10.

Hashimoto, H., Wang, D., Steves, A. N., Jin, P., Blumenthal, R. M., Zhang, X., and Cheng, X. (2016). Distinctive Klf4 mutants determine preference for DNA methylation status. *Nucleic Acids Research*, 44(21):gkw774.

He, H. H., Meyer, C. a., Hu, S. S., Chen, M.-W., Zang, C., Liu, Y., Rao, P. K., Fei, T., Xu, H., Long, H., Liu, X. S., and Brown, M. (2014). Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nature methods*, 11(1):73–8.

He, Z., Kokkinaki, M., Jiang, J., Dobrinski, I., and Dym, M. (2010). Isolation, Characterization, and Culture of Human Spermatogonia. *Biology of Reproduction*, 82(2):363–372.

Helleday, T., Eshtad, S., and Nik-Zainal, S. (2014). Mechanisms underlying mutational signatures in human cancers. *Nature Reviews Genetics*, 15(9):585–598.

Hellman, L. M. and Fried, M. G. (2007). Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nature protocols*, 2(8):1849–61.

Herbert, M., Kalleas, D., Cooney, D., Lamb, M., and Lister, L. (2015). Meiosis and maternal aging: insights from aneuploid oocytes and trisomy births. *Cold Spring Harbor perspectives in biology*, 7(4):a017970.

Hodgkinson, A. and Eyre-Walker, A. (2011). Variation in the mutation rate across mammalian genomes. *Nature Reviews Genetics*, 12(11):756–766.

House, N. C. M., Koch, M. R., and Freudenreich, C. H. (2014). Chromatin modifications and DNA repair: beyond double-strand breaks. *Frontiers in Genetics*, 5(September):1–18.

Hughes, T. R. (2011). Introduction to "a handbook of transcription factors". *Subcellular biochemistry*, 52:1–6.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(February).

Jacob, F., Brenner, S., and Cuzin, F. (1963). On the Regulation of DNA Replication in Bacteria. *Cold Spring Harbor Symposia on Quantitative Biology*, 28:329–348.

Janion, C. (2008). Inducible SOS Response System of DNA Repair and Mutagenesis in Escherichia coli. *International Journal of Biological Sciences*, 4(6):338–344.

Jin, C., Zang, C., Wei, G., Cui, K., Peng, W., Zhao, K., and Felsenfeld, G. (2009). H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. *Nature Genetics*, 41(8):941–945.

Jinks-Robertson, S. and Bhagwat, A. S. (2014). Transcription-associated mutagenesis. *Annual review of genetics*, 48:341–59.

Johansson, E. and Dixon, N. (2013). Replicative DNA Polymerases. *Cold Spring Harbor Perspectives in Biology*, 5(6):a012799–a012799.

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science*, 316(5830):1497–1502.

Jónsson, H., Sulem, P., Kehr, B., Kristmundsdottir, S., Zink, F., Hjartarson, E., Gudjonsson, S. A., Ward, L. D., Hardarson, M. T., Hjorleifsson, K. E., Hannes, P., Rafnar, T., Frigge, M., Stacey, S. N., Magnusson, O. T., Thorsteinsdottir, U., Masson, G., Helgason, A., Gudbjartsson, D. F., Stefansson, K., Kong, A., and Bjarni, V. (2017). Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature Publishing Group*, 549(7673):519–522.

Kaiser, V. B., Taylor, M. S., and Semple, C. A. (2016). Mutational Biases Drive Elevated Rates of Substitution at Regulatory Sites across Cancer Types. *PLOS Genetics*, 12(8):e1006207.

Katainen, R., Dave, K., Pitkänen, E., Palin, K., Kivioja, T., Välimäki, N., Gylfe, A. E., Ristolainen, H., Hänninen, U. a., Cajuso, T., Kondelin, J., Tanskanen, T., Mecklin, J.-P., Järvinen, H., Renkonen-Sinisalo, L., Lepistö, A., Kaasinen, E., Kilpivaara, O., Tuupanen, S., Enge, M., Taipale, J., and Aaltonen, L. a. (2015). CTCF/cohesin-binding sites are frequently mutated in cancer. *Nature Genetics*, (June):8–13.

Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. (2003). Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 100(20):11484–9.

Kim, S., Yu, N. K., and Kaang, B. K. (2015). CTCF as a multifunctional protein in genome regulation and gene expression. *Experimental & molecular medicine*, 47(6):e166.

Kim, T. K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P. F., Kreiman, G., and Greenberg, M. E. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295):182–187.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217(5129):624–6.

Kimura, M. (1991). The neutral theory of molecular evolution: a review of recent evidence. *Idengaku zasshi*, 66(4):367–86.

King, M.-c. and Wilson, A. (1975). Evolution at two levels in humans and chimpanzees. *Science*, 188(4184):107–116.

Klug, S. J. and Famulok, M. (1994). All you wanted to know about SELEX. *Molecular Biology Reports*, 20(2):97–107.

Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S. a., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., Wong, W. S. W., Sigurdsson, G., Walters, G. B., Steinberg, S., Helgason, H., Thorleifsson, G., Gudbjartsson, D. F., Helgason, A., Magnusson, O. T., Thorsteinsdottir, U., and Stefansson, K. (2012). Rate of de novo mutations and the importance of father's age to disease risk. *Nature*, 488(7412):471–475.

Kopylow, K. V. and Spiess, A.-n. (2017). Human spermatogonial markers. *Stem Cell Research*, 25:300–309.

Koshland Jr., D. E. (2002). SPECIAL ESSAY: The Seven Pillars of Life. *Science*, 295(5563):2215–2216.

Kossack, N., Terwort, N., Wistuba, J., Ehmcke, J., Schlatt, S., Schöler, H., Kliesch, S., and Gromoll, J. (2013). A combined approach facilitates the reliable detection of human spermatogonia in vitro. *Human Reproduction*, 28(11):3012–3025.

Kunkel, T. a. (2009). Evolving views of DNA replication (in)fidelity. *Cold Spring Harbor symposia on quantitative biology*, 74:91–101.

Kunkel, T. A. and Erie, D. A. (2015). Eukaryotic Mismatch Repair in Relation to DNA Replication. *Annual Review of Genetics*, 49(1):291–313.

Lai, W. K. M. and Pugh, B. F. (2017). Understanding nucleosome dynamics and their links to gene expression and DNA replication. *Nature Publishing Group*, 18(9):548–562.

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., and Weirauch, M. T. (2018). The Human Transcription Factors. *Cell*, 172(4):650–665.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–9.

Lans, H., Marteijn, J. A., and Vermeulen, W. (2012). ATP-dependent chromatin remodeling in the DNA-damage response. *Epigenetics & Chromatin*, 5(1):4.

Lehmann, A. R., Mcgibbon, D., and Stefanini, M. (2011). Xeroderma pigmentosum. *Orphanet Journal of Rare Diseases*, 6(1):70.

Li, G. and Reinberg, D. (2011). Chromatin higher-order structures and gene regulation. *Current opinion in genetics & development*, 21(2):175–86.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–9.

Lihu, A. and Holban, t. (2015). A review of ensemble methods for de novo motif discovery in ChIP-Seq data. *Briefings in Bioinformatics*, 16(6):964–973.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., Ward, L. D., Lowe, C. B., Beal, K., Chang, J., Clawson, H., Holloway, A. K., Clamp, M., Gnerre, S., Alfo, J., Cuff, J., Palma, F. D., Fitzgerald, S., Flicek, P., Guttman, M., Hubisz, M. J., Jaffe, D. B., Jungreis, I., Kent, W. J., Kostka, D., Lara, M., Martins, A. L., Massingham, T., Moltke, I., Raney, B. J., Rasmussen, M. D., Robinson, J., Stark, A., Vilella, A. J., Wen, J., and Xie, X. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370):476–481.

Liu, L. and Heermann, D. W. (2015). The interaction of DNA with multi-Cys ₂ His ₂ zinc finger proteins. *Journal of Physics: Condensed Matter*, 27(6):064107.

Lodato, M. A., Woodworth, M. B., Lee, S., Evrony, G. D., Mehta, B. K., Karger, A., Lee, S., Chittenden, T. W., D'Gama, A. M., Cai, X., Luquette, L. J., Lee, E., Park, P. J., and Walsh, C. A. (2015). Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science*, 350(6256):94–98.

Lujan, S. A., Clausen, A. R., Clark, A. B., Macalpine, H. K., Macalpine, D. M., Malc, E. P., Mieczkowski, P. A., Burkholder, A. B., Fargo, D. C., Gordenin, D. A., and

Kunkel, T. A. (2014). Heterogeneous polymerase fidelity and mismatch repair bias genome variation and composition. *Genome Research*, pages 1751–1764.

Madrigal, P. (2015). On Accounting for Sequence-Specific Bias in Genome-Wide Chromatin Accessibility Experiments: Recent Advances and Contradictions. *Frontiers in Bioengineering and Biotechnology*, 3(September):1–4.

Makova, K. D. and Hardison, R. C. (2015). The effects of chromatin organization on variation in mutation rates in the genome. *Nature Reviews Genetics*, advance on(March).

Marteijn, J. A., Lans, H., Vermeulen, W., and Hoeijmakers, J. H. (2014). Understanding nucleotide excision repair and its roles in cancer and ageing. *Nature Reviews Molecular Cell Biology*, 15(7):465–481.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10.

Martins, A. L., Walavalkar, N. M., Anderson, W. D., Zang, C., and Guertin, M. J. (2018). Universal correction of enzymatic sequence bias reveals molecular signatures of protein/DNA interactions. *Nucleic Acids Research*, 46(2):1–12.

Mathelier, A., Xin, B., Chiu, T.-p., Yang, L., Rohs, R., Wasserman, W. W., Mathelier, A., Xin, B., Chiu, T.-p., Yang, L., Rohs, R., and Wasserman, W. W. (2016). Article DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo Article DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. *Cell Systems*, 3(3):278–286.e4.

Meselson, M. and Stahl, F. W. (1958). The Replication of E Coli. *Proceedings of the National Academy of Sciences*, 44:671–682.

Michaelson, J. J., Shi, Y., Gujral, M., Zheng, H., Malhotra, D., Jin, X., Jian, M., Liu, G., Greer, D., Bhandari, A., Wu, W., Corominas, R., Peoples, Á., Koren, A., Gore, A., Kang, S., Lin, G. N., Estabillo, J., Gadomski, T., Singh, B., Zhang, K., Akshoomoff, N., Corsello, C., McCarroll, S., Iakoucheva, L. M., Li, Y., Wang, J., and Sebat, J. (2012). Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell*, 151(7):1431–1442.

Minnich, M., Tagoh, H., Bönelt, P., Axelsson, E., Fischer, M., Cebolla, B., Tarakhovsky, A., Nutt, S. L., Jaritz, M., and Busslinger, M. (2016). Multifunctional role of the transcription factor Blimp-1 in coordinating plasma cell differentiation. *Nature immunology*, 17(3):331–43.

Moyer, S. E., Lewis, P. W., and Botchan, M. R. (2006). Isolation of the Cdc45/Mcm2-7/GINS (CMG) complex, a candidate for the eukaryotic DNA replication fork helicase. *Proceedings of the National Academy of Sciences of the United States of America*, 103(27):10236–10241.

Neph, S., Vierstra, J., Stergachis, A. B., Reynolds, A. P., Haugen, E., Vernot, B., Thurman, R. E., John, S., Sandstrom, R., Johnson, A. K., Maurano, M. T., Humbert, R., Rynes, E., Wang, H., Vong, S., Lee, K., Bates, D., Diegel, M., Roach, V., Dunn, D., Neri, J., Schafer, A., Hansen, R. S., Kutyavin, T., Giste, E., Weaver, M., Canfield, T., Sabo, P., Zhang, M., Balasundaram, G., Byron, R., MacCoss, M. J., Akey, J. M., Bender, M. A., Groudine, M., Kaul, R., and Stamatoyannopoulos, J. A. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489(7414):83–90.

Nirenberg, M., Caskey, T., Marshall, R., Brimacombe, R., Kellogg, D., Doctor, B., Hatfield, D., Levin, J., Rottman, F., Pestka, S., Wilcox, M., and Anderson, F. (1966). The RNA code and protein synthesis. *Cold Spring Harbor symposia on quantitative biology*, 31(Table 1):11–24.

Okazaki, R., Okazaki, T., Sakabe, K., Sugimoto, K., Kainuma, R., Sugino, A., and Iwatsuki, N. (1968). In Vivo Mechanism of DNA Chain Growth. *Cold Spring Harbor Symposia on Quantitative Biology*, 33(0):129–143.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A.,

Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D., and Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745.

O'Sullivan, R. J. and Karlseder, J. (2010). Telomeres: protecting chromosomes against genome instability. *Nature reviews. Molecular cell biology*, 11(3):171–81.

Pagel, J.-I. and Deindl, E. (2011). Early growth response 1–a transcription factor in the crossfire of signal transduction cascades. *Indian journal of biochemistry & biophysics*, 48(4):226–35.

Palade, G. E. (1955). A SMALL PARTICULATE COMPONENT OF THE CYTOPLASM. *The Journal of Cell Biology*, 1(1):59–68.

Peltomäki, P. (2001). DNA mismatch repair and cancer. *Mutation Research - Reviews in Mutation Research*, 488(1):77–85.

Perera, D., Poulos, R. C., Shah, A., Beck, D., Pimanda, J. E., and Wong, J. W. H. (2016). Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. *Nature*, 532(7598):259–263.

Perera, R. L., Torella, R., Klinge, S., Kilkenny, M. L., Maman, J. D., and Pellegrini, L. (2013). Mechanism for priming DNA synthesis by yeast DNA polymerase α. *eLife*, 2(2):e00482.

Petryk, N., Dalby, M., Wenger, A., Stromme, C. B., Strandsby, A., Andersson, R., and Groth, A. (2018). MCM2 promotes symmetric inheritance of modified histones during DNA replication. *Science*, 13(3):eaau0294.

Polak, P., Karlić, R., Koren, A., Thurman, R., Sandstrom, R., Lawrence, M., Reynolds, A., Rynes, E., Vlahoviček, K., Stamatoyannopoulos, J. A., and Sunyaev, S. R. (2015). Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature*, 518(7539):360–364.

Polak, P., Lawrence, M. S., Haugen, E., Stoletzki, N., Stojanov, P., Thurman, R. E., Garraway, L. A., Mirkin, S., Getz, G., Stamatoyannopoulos, J. A., and Sunyaev, S. R. (2014). Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nature Biotechnology*, 32(1):71–75.

Pott, S. and Lieb, J. D. (2015). What are super-enhancers? *Nature Genetics*, 47(1):8–12.

Prendergast, J. G. D. and Semple, C. A. M. (2011). Widespread signatures of recent selection linked to nucleosome positioning in the human lineage. *Genome Research*, pages 1777–1787.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6):841–2.

Ramachandran, S. and Henikoff, S. (2016). Transcriptional Regulators Compete with Nucleosomes Post-replication. *Cell*, 165(3):580–592.

Randall, S. K., Eritja, R., Kaplan, B. E., Petruska, J., and Goodman, M. F. (1987). Nucleotide insertion kinetics opposite abasic lesions in DNA. *Journal of Biological Chemistry*, 262(14):6864–6870.

Rao, S. S., Huang, S.-C., Glenn St Hilaire, B., Engreitz, J. M., Perez, E. M., Kieffer-Kwon, K.-R., Sanborn, A. L., Johnstone, S. E., Bascom, G. D., Bochkov, I. D., Huang, X., Shamim, M. S., Shin, J., Turner, D., Ye, Z., Omer, A. D., Robinson, J. T., Schlick, T., Bernstein, B. E., Casellas, R., Lander, E. S., and Aiden, E. L. (2017). Cohesin Loss Eliminates All Loop Domains. *Cell*, 171(2):305–320.e24.

Reijns, M. A. M., Kemp, H., Ding, J., Marion de Procé, S., Jackson, A. P., and Taylor, M. S. (2015). Lagging-strand replication shapes the mutational landscape of the genome. *Nature*, 518(7540):502–506.

Rhee, H. S. and Pugh, B. F. (2011). Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, 147(6):1408–1419.

Richardson, G. M., Lannigan, J., and Macara, I. G. (2015). Does FACS perturb gene expression? *Cytometry Part A*, 87(2):166–175.

Roach, J. C., Glusman, G., Smit, A. F. A., Huff, C. D., Shannon, P. T., Rowen, L., Pant, K. P., Goodman, N., Shendure, J., Drmanac, R., Jorde, L. B., Hood, L., and Galas, J. (2011). Analysis of Genetic Inheritance in a Family Quartet by Whole Genome Sequencing. *Science*, 328(5978):636–639.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T., and Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–30.

Roberts, E. G., Mendez, M., Viner, C., Karimzadeh, M., Chan, R. C. W., Ancar, R., Chicco, D., Hesselberth, J. R., Kundaje, A., and Hoffman, M. M. (2016). Semi-automated genome annotation using epigenomic data and Segway. *bioRxiv*, page 080382.

Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A., and Lopez-Bigas, N. (2015). Nucleotide excision repair is impaired by binding of transcription factors to DNA. *bioRxiv*, 532(7598):028886.

Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A., and López-Bigas, N. (2016). Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature*, 532(7598):264–7.

Sasaki, S., Mello, C. C., Shimada, A., Nakatani, Y., Hashimoto, S.-I., Ogawa, M., Matsushima, K., Gu, S. G., Kasahara, M., Ahsan, B., Sasaki, A., Saito, T., Suzuki, Y., Sugano, S., Kohara, Y., Takeda, H., Fire, A., and Morishita, S. (2009). Chromatin-

associated periodicity in genetic variation downstream of transcriptional start sites. *Science (New York, N.Y.)*, 323(5912):401–404.

Saxonov, S., Berg, P., and Brutlag, D. L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences of the United States of America*, 103(5):1412–7.

Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C. P., Mackay, S., Talianidis, I., Flicek, P., and Odom, D. T. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science (New York, N.Y.)*, 328(5981):1036–40.

Schmitges, F. W., Radovani, E., Najafabadi, H. S., Barazandeh, M., Campitelli, L. F., Yin, Y., Jolma, A., Zhong, G., Guo, H., Kanagalingam, T., Dai, W. F., Taipale, J., Emili, A., Greenblatt, J. F., and Hughes, T. R. (2016). Multiparameter functional diversity of human C2H2 zinc finger proteins. *Genome research*, 26(12):1742–1752.

Schuetz, A., Nana, D., Rose, C., Zocher, G., Milanovic, M., Koenigsmann, J., Blasig, R., Heinemann, U., and Carstanjen, D. (2011). The structure of the Klf4 DNA-binding domain links to self-renewal and macrophage differentiation. *Cellular and Molecular Life Sciences*, 68(18):3121–3131.

Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N. A., Huber, W., Haering, C., Mirny, L., and Spitz, F. (2017). Two independent modes of chromatin organization revealed by cohesin removal. *Nature*, 542(7641):377–380.

Ségurel, L., Wyman, M. J., and Przeworski, M. (2014). Determinants of Mutation Rate Variation in the Human Germline. *Annual review of genomics and human genetics*, (May):1–24.

Seitan, V. C., Faure, A. J., Zhan, Y., McCord, R. P., Lajoie, B. R., Ing-Simmons, E., Lenhard, B., Giorgetti, L., Heard, E., Fisher, A. G., Flicek, P., Dekker, J., and Merkenschlager, M. (2013). Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. *Genome Research*, 23(12):2066–2077.

Semple, C. a. M. and Taylor, M. S. (2009). The Structure of Change. *Science*, 323(5912):347–348.

Sherman, S. L., Petersen, M. B., Freeman, S. B., Hersey, J., Pettay, D., Taft, L., Frantzen, M., Mikkelsen, M., and Hassold, T. J. (1994). Non-disjunction of chromosome 21 in maternal meiosis I: evidence for a maternal age-dependent mechanism involving reduced recombination. *Human molecular genetics*, 3(9):1529–35.

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., Fukuda, S., Sasaki, D., Podhajska, A., Harbers, M., Kawai, J., Carninci, P., and Hayashizaki, Y. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences*, 100(26):15776–15781.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. D. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., and Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050.

Simon, J. M., Giresi, P. G., Davis, I. J., and Lieb, J. D. (2012). Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. *Nature Protocols*, 7(2):256–267.

Smith, D. J. and Whitehouse, I. (2012). Intrinsic coupling of lagging-strand synthesis to chromatin assembly. *Nature*, 483(7390):434–438.

Song, L. and Crawford, G. E. (2010). DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor protocols*, 2010(2):pdb.prot5384.

Soufi, A., Garcia, M. F., Jaroszewicz, A., Osman, N., Pellegrini, M., and Zaret, K. S. (2015). Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell*, 161(3):555–568.

Srinivas, S., Watanabe, T., Lin, C. S., William, C. M., Tanabe, Y., Jessell, T. M., and

Costantini, F. (2001). Cre reporter strains produced by targeted insertion of EYFP and ECFP into the ROSA26 locus. *BMC Developmental Biology*, 1:1–8.

Stamatoyannopoulos, J. a., Adzhubei, I., Thurman, R. E., Kryukov, G. V., Mirkin, S. M., and Sunyaev, S. R. (2009). Human mutation rate associated with DNA replication timing. *Nature genetics*, 41(4):393–395.

Stith, C. M., Sterling, J., Resnick, M. A., Gordenin, D. A., and Burgers, P. M. (2008). Flexibility of eukaryotic Okazaki fragment maturation through regulated strand displacement synthesis. *The Journal of biological chemistry*, 283(49):34129–40.

Strahl, B. D. and Allis, C. D. (2000). The language of covalent histone modifications. *Nature*, 403(6765):41–45.

Sung, M. H., Guertin, M. J., Baek, S., and Hager, G. L. (2014). DNase footprint signatures are dictated by factor dynamics and DNA sequence. *Molecular Cell*, 56(2):275–285.

Supek, F. and Lehner, B. (2015). Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature*, 521(7550):81–84.

Supek, F. and Lehner, B. (2017). Clustered Mutation Signatures Reveal that Error-Prone DNA Repair Targets Mutations to Active Genes. *Cell*, 170(3):534–547.e23.

Takahashi, K. and Yamanaka, S. (2006). Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell*, 126(4):663–676.

Taylor, M. S., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., and Semple, C. a. M. (2006). Heterotachy in mammalian promoter evolution. *PLoS Genetics*, 2(4):627–639.

Taylor, M. S., Massingham, T., Hayashizaki, Y., Carninci, P., Goldman, N., and Semple, C. a. M. (2008). Rapidly evolving human promoter regions. *Nature genetics*, 40(11):1262–1263; author reply 1263–1264.

Thul, P. J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., Alm, T., Asplund, A., Björk, L., Breckels, L. M., Bäckström, A., Danielsson, F., Fagerberg, L., Fall, J., Gatto, L., Gnann, C., Hober, S., Hjelmare, M., Johansson, F., Lee,

S., Lindskog, C., Mulder, J., Mulvey, C. M., Nilsson, P., Oksvold, P., Rockberg, J., Schutten, R., Schwenk, J. M., Sivertsson, Å., Sjöstedt, E., Skogs, M., Stadler, C., Sullivan, D. P., Tegel, H., Winsnes, C., Zhang, C., Zwahlen, M., Mardinoglu, A., Pontén, F., von Feilitzen, K., Lilley, K. S., Uhlén, M., and Lundberg, E. (2017). A subcellular map of the human proteome. *Science (New York, N.Y.)*, 356(6340):eaal3321.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutyavin, T., Lajoie, B., Lee, B.-K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. a., Stamatoyannopoulos, G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E., and Stamatoyannopoulos, J. a. (2012). The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82.

Tiemann-Boege, I., Schwarz, T., Striedner, Y., Heissl, A., and Tiemann-boege, I. (2017). The consequences of sequence erosion in the evolution of recombination hotspots. *Philisophical Transactions Royal Society B*, 372.

Tolstorukov, M. Y., Volfovsky, N., Stephens, R. M., and Park, P. J. (2011). Impact of chromatin structure on sequence variability in the human genome. *Nature structural & molecular biology*, 18(4):510–515.

Tomasetti, C. and Vogelstein, B. (2015). Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, 347(6217).

Tsompana, M. and Buck, M. J. (2014). Chromatin accessibility: a window into the genome. *Epigenetics & Chromatin*, 7(1):1–16.

Uchimura, A., Higuchi, M., Minakuchi, Y., Ohno, M., Toyoda, A., Fujiyama, A., Miura, I., Wakana, S., Nishino, J., and Yagi, T. (2015). Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. *Genome research*, 25:1–10.

Umer, H. M., Cavalli, M., Dabrowski, M. J., Diamanti, K., Kruczyk, M., Pan, G., Komorowski, J., and Wadelius, C. (2016). A Significant Regulatory Mutation Burden at a High-Affinity Position of the CTCF Motif in Gastrointestinal Cancers. *Human Mutation*, 37(9):904–913.

Vaisman, A. and Woodgate, R. (2017). Translesion DNA polymerases in eukaryotes: what makes them tick? *Critical Reviews in Biochemistry and Molecular Biology*, 52(3):274–303.

Valli, H., Sukhwani, M., Dovey, S. L., Peters, K. A., Donohue, J., Castro, C. A., Chu, T., Marshall, G. R., and Orwig, K. E. (2014). Fluorescence- and magnetic-activated cell sorting strategies to isolate and enrich human spermatogonial stem cells. *Fertility and Sterility*, 102(2):566–580.e7.

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., and Luscombe, N. M. (2009). A census of human transcription factors: Function, expression and evolution. *Nature Reviews Genetics*, 10(4):252–263.

Veltman, J. A. and Brunner, H. G. (2012). De novo mutations in human genetic disease. *Nature Publishing Group*, 13(8):565–575.

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., Park, T. J., Deaville, R., Erichsen, J. T., Jasinska, A. J., Turner, J. M., Bertelsen, M. F., Murchison, E. P., Flicek, P., and Odom, D. T. (2015). Enhancer evolution across 20 mammalian species. *Cell*, 160(3):554–566.

Vinagre, J., Almeida, A., Pópulo, H., Batista, R., Lyra, J., Pinto, V., Coelho, R., Celestino, R., Prazeres, H., Lima, L., Melo, M., da Rocha, A. G., Preto, A., Castro, P., Castro, L., Pardal, F., Lopes, J. M., Santos, L. L., Reis, R. M., Cameselle-Teijeiro, J., Sobrinho-Simões, M., Lima, J., Máximo, V., and Soares, P. (2013). Frequency of TERT promoter mutations in human cancers. *Nature communications*, 4:2185.

Von Kopylow, K., Kirchhoff, C., Jezek, D., Schulze, W., Feig, C., Primig, M., Steinkraus, V., and Spiess, A. N. (2010). Screening for biomarkers of spermatogonia within the human testis: A whole genome approach. *Human Reproduction*, 25(5):1104–1112.

Von Kopylow, K., Schulze, W., Salzbrunn, A., and Spiess, A.-N. (2016). Isolation and gene expression analysis of single potential human spermatogonial stem cells. *Molecular Human Reproduction at Gdanski Uniwersytet Medyczny on*, 22(4):229–239.

Waheeb, R. and Hofmann, M.-C. (2011). Human spermatogonial stem cells: a possible origin for spermatocytic seminoma. *International Journal of Andrology*, 4(164):1–17.

Wang, H., Zhai, L., Xu, J., Joo, H.-y., Jackson, S., Erdjument-bromage, H., Tempst, P., Hill, C., and Carolina, N. (2006). Histone H3 and H4 Ubiquitylation by the CUL4-DDB-ROC1 Ubiquitin Ligase Facilitates Cellular Response to DNA Damage. *Molecular Cell*, pages 383–394.

Wang, J., Jia, S. T., and Jia, S. (2016). New Insights into the Regulation of Heterochromatin. *Trends in genetics : TIG*, 32(5):284–294.

Wang, J., Zibetti, C., Shang, P., Sripathi, S. R., Zhang, P., Cano, M., Hoang, T., Xia, S., Ji, H., Merbs, S. L., Zack, D. J., Handa, J. T., Sinha, D., Blackshaw, S., and Qian, J. (2018). ATAC-Seq analysis reveals a widespread decrease of chromatin accessibility in age-related macular degeneration. *Nature communications*, 9(1):1364.

Wang, J. R., Quach, B., and Furey, T. S. (2017). Correcting nucleotide-specific biases in high-throughput sequencing data. *BMC Bioinformatics*, 18(1):357.

Wang, K., Yuen, S. T., Xu, J., Lee, S. P., Yan, H. H., Shi, S. T., Siu, H. C., Deng, S., Chu, K. M., Law, S., Chan, K. H., Chan, A. S., Tsui, W. Y., Ho, S. L., Chan, A. K., Man, J. L., Foglizzo, V., Ng, M. K., Chan, A. S., Ching, Y. P., Cheng, G. H., Xie, T., Fernandez, J., Li, V. S., Clevers, H., Rejto, P. A., Mao, M., and Leung, S. Y. (2014). Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nature Genetics*, 46(6):573–582.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63.

Watson, J. D. and Crick, F. H. D. (1953). A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738.

Watt, D. L., Buckland, R. J., Lujan, S. A., Kunkel, T. A., and Chabes, A. (2015).

Genome-wide analysis of the specificity and mechanisms of replication infidelity driven by imbalanced dNTP pools. *Nucleic Acids Research*, 44(4):1669–1680.

Whitehouse, I. and Smith, D. J. (2013). Chromatin dynamics at the replication fork: there's more to life than histones. *Current Opinion in Genetics & Development*, 23(2):140–146.

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I., and Young, R. A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2):307–319.

Wittkopp, P. J. and Kalay, G. (2012). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature reviews. Genetics*, 13(1):59–69.

Wolfe, K. H., Sharp, P. M., and Li, W. H. (1989). Mutation rates differ among regions of the mammalian genome. *Nature*, 337(6204):283–285.

Wu, X., Schmidt, J. a., Avarbock, M. R., Tobias, J. W., Carlson, C. a., Kolon, T. F., Ginsberg, J. P., and Brinster, R. L. (2009). Prepubertal human spermatogonia and mouse gonocytes share conserved gene expression of germline stem cell regulatory molecules. *Proceedings of the National Academy of Sciences of the United States of America*, 106(51):21672–21677.

Yan, J., Enge, M., Whitington, T., Dave, K., Liu, J., Sur, I., Schmierer, B., Jolma, A., Kivioja, T., Taipale, M., and Taipale, J. (2013). Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell*, 154(4):801–813.

Yang, J. and Ramsey, S. A. (2015). Sequence analysis A DNA shape-based regulatory score improves position-weight matrix-based recognition of transcription factor binding sites. 31(June):3445–3450.

Yazdi, P. G., Pedersen, B. A., Taylor, J. F., and Khattab, O. S. (2015). Increasing Nucleosome Occupancy Is Correlated with an Increasing Mutation Rate so Long as DNA Repair Machinery Is Intact. *PLoS ONE*, pages 1–16.

Young, R. S., Hayashizaki, Y., Andersson, R., Sandelin, A., Kawaji, H., Itoh, M., Lassmann, T., Carninci, P., FANTOM Consortium, Bickmore, W. a., Forrest, A. R.,

and Taylor, M. S. (2015). The frequent evolutionary birth and death of functional promoters in mouse and human. *Genome research*, 25(10):1546–57.

Young, R. S., Kumar, Y., Bickmore, W. A., and Taylor, M. S. (2017). Bidirectional transcription initiation marks accessible chromatin and is not specific to enhancers. *Genome Biology*, 18(1):1–11.

Youssoufian, H., Kazazian, H. H., Phillips, D. G., Aronis, S., Tsiftis, G., Brown, V. A., and Antonarakis, S. E. (1986). Recurrent mutations in haemophilia a give evidence for CpG mutation hotspots. *Nature*, 324(6095):380–382.

Zaret, K. S. and Carroll, J. S. (2011). Pioneer transcription factors : establishing competence for gene expression Parameters affecting transcription factor access to target sites in chromatin Initiating events in chromatin : pioneer factors bind first. *Genes and Development*, pages 2227–2241.

Zarnack, K., König, J., Tajnik, M., Martincorena, I., Eustermann, S., Stévant, I., Reyes, A., Anders, S., Luscombe, N. M., and Ule, J. (2013). Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell*, 152(3):453–66.

Zhang, Y., Liu, T., Meyer, C. a., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome biology*, 9(9):R137.

Zhao, Y. and Garcia, B. A. (2015). Comprehensive Catalog of Currently Documented Histone Modifications. *Cold Spring Harbor perspectives in biology*, 7(9):a025064.

Zhou, J., Park, C. Y., Theesfeld, C. L., Yuan, Y., Sawicka, K., Darnell, C., Scheckel, C., Fak, J. J., Tajima, Y., Darnell, R. B., and Olga, G. (2018). Whole-genome deep learning analysis reveals causal role of noncoding mutations in autism. *bioRxiv*, pages 1–29.