Title	Analysis of putative somatic mutations in 200,000 human exomes
Author(s)	Bennett, Declan
Publication Date	2023-09-19
Publisher	NUI Galway
Item record	http://hdl.handle.net/10379/17914

Analysis of putative somatic mutations in 200,000 human exomes

A thesis submitted

by

Declan Bennett

to

The Discipline of Bioinformatics,
School of Mathematical & Statistical Sciences,
University *of* Galway

in partial fulfilment of the requirements for the degree of

Ph.D. in Bioinformatics

August 3, 2023

Thesis Supervisor: Prof. Cathal Seoighe

Declaration of Authorship

I, Declan Bennett, declare that this thesis titled, ‘Analysis of putative somatic mutations in 200,000 human exomes’ submitted to the Discipline of Bioinformatics, School of Mathematical & Statistical Science, University of Galway in partial fulfilment of the requirements for the degree of Doctor of Philosophy (PhD.) in Bioinformatics is entirely my own work and I have acknowledged any assistance or contributions and cited the published work of others where applicable. The research contained within this thesis has been conducted with the financial support of Science Foundation Ireland under Grant Number 16/IA/4612 and the Irish Research Council under Grant Number GOIPG/2018/1291. This work has not been submitted, in whole or in part, by me or another person, for the purpose of obtaining any other degree.

I agree freely that the library may lend or copy this thesis upon request.

June 22nd, 2023

Declan Bennett

Date

Acknowledgements

To my parents, Monica and Kevin.

I am deeply indebted to my supervisor Prof. Cathal Seoighe for his never-ending patience, deep insights, and guidance over the course of this PhD. project. Cathal was always available to discuss any issue or idea that arose and he was always on hand to pull me out of research rabbit holes. It has been a privilege to work with Cathal and I have learned a huge amount both professionally and personally from my time in his research group. A special thank you to Dr Pilib Ó Broin and Prof. Aaron Golden for always having an open door whether they wanted it or not!

During my PhD., I have been lucky to work in a community that has too many names to mention. They fostered learning, creativity, and most importantly coffee breaks as priority number one. With that, a huge thanks to the 1018 office family, Siobhán, Noor, Brian, Laura, and Sumaira. The breadth of expertise within the office meant I never had to travel far to discuss some of the finer problems that arose over the course of this PhD. A big thank you to Barry and Stephen for their sympathetic ears while I complained about that week's problems to the backdrop of an open fire in the 'bunch of grapes'.

Over the course of this project, there have been many people who have impacted the trajectory of this thesis. To my sister, Maryann, brothers, Kevin & Stephen along with friends old and new for their unquantifiable contribution to this work and for their ongoing love, support, and encouragement over my time in Galway, I will always be grateful.

List of Figures

List of Tables

Abstract

Somatic mutations accumulate throughout life and contribute significantly to disease risk. While research into somatic mutation is well established in cancer, it is only in recent years that investigations into the implications of somatic mutations in healthy tissues have begun to be feasible, due to advances in sequencing technologies and protocols. The requirement of specialist techniques has, however, limited the study of somatic mutations in healthy tissues to small sample sizes, which do not allow for assessment of the impact of somatic mutations on human health on a population scale. We posited that it may be possible to study variation in the somatic mutation rate between individuals and across the genome through analysis of low-depth sequencing data, by developing strategies to distinguish the contribution of somatic mutations to the mismatches (relative to the reference genome) observed in these data from sequencing errors, DNA damage and other artefacts.

Using somatic mutation rates obtained from the literature, we estimated that 0.4% of the mismatches between the UK Biobank exome sequencing reads and the reference genome were due to somatic mutations. We demonstrated that this proportion was sufficient to induce a relationship between the abundance of mismatches and age, when individuals were grouped by integer age. We then searched for additional sample properties that are correlated with the mismatch burden and found positive correlations with cancer diagnosis and smoking status. However, by carefully examining the UK Biobank exome sequencing data, we uncovered previously unreported batch effects relating to sequencing run. The observed associations with cancer diagnosis and smoking status were lost when we corrected for this batch effect. However, the batch correction improved the correlation between age and mismatch load.

Individuals diagnosed with Lynch syndrome have increased somatic mutation loads due to deficiencies in mismatch repair genes and we investigated whether this effect could be detected in the exome sequencing data. In the UK Biobank, we identified 160 individuals with pathogenic variants associated with Lynch syndrome. Using the COSMIC signatures associated with mismatch repair, we compared the contribution of mismatch repair mutational signatures between the Lynch syndrome samples and the remaining samples. We detected a marginally statistically significant difference between the contribution of SBS18 between the two sample groups; however, this result did not survive multiple correction testing.

Somatic and germline mutations show transcription-strand asymmetry, arising from transcription-associated DNA damage and repair. We postulated that the strength of transcription-strand asymmetry could provide insights into the contribution of somatic mutations to the exome sequencing data, because technical sources of mismatches, such as DNA damage and sequencing error, should not be directly affected by transcription. We indeed observed substantial transcription-strand asymmetry; however, this was far stronger than we expected, given the inferred proportion of somatic mutations in the data. This result led us to identify a technical effect that resulted in transcription-strand asymmetry, arising from the use of single-stranded probes targeting the coding strand in the exome capture kit used by the UK Biobank. Surprisingly, this has not previously been published and it has important implications for NGS quality control and rare variant analyses.

The large sample size of the UK Biobank also raised the possibility of testing for genetic variation affecting the somatic mutation rate. Treating the normalized number of mismatches per sample as a quantitative phenotype, we performed a GWAS and discovered a genome-wide significant hit in linkage with an eQTL for *ERCC8*, an integral component of the transcription-coupled repair machinery. Although promising, this candidate GWAS locus turned out to be a false-positive association, resulting from an unusual genetic variant that our germline filter had not removed. In the course of this work, we also proposed a methodological innovation in GWAS that consists of including background genetic variation as a fixed effect in the linear mixed models used in GWAS. We demonstrated that this can improve the power of GWAS when combined with state-of-the-art polygenic scoring methodologies. Our method substantially improved the estimation of effect sizes and power. However, the improvement depended on heritability and polygenicity and consequently, the mismatch data, which showed low heritability, did not benefit from our method.

We then pivoted our focus from understanding the variation in mismatch load acting across samples to understanding variation across the genome. We again found evidence that variation in the somatic mutation rate across the genome can be detected in the exome sequencing data, observing correlations in the expected directions for known mutation rate modifiers, such as gene expression, replication timing and chromatin structure. Interestingly, we recovered a complex relationship between mismatch recurrence and gene expression, consistent with the literature. The recurrence of potentially functional mismatches also provides a means to infer positive selection acting on somatic mutations and we found that several genes associated with clonal haematopoiesis of indeterminate potential showed strong evidence of positive selection.

Contents

Chapter 1

Introduction

1.1 DNA

1.1.1 A brief history of the DNA Sequence

DNA was first isolated in 1869 by Friedrich Miescher at the University of T*ü*bingen while attempting to purify protein from the nucleus of leukocytes coining the unknown substance, nuclein [1]. It would be 75 years before the Avery-MacLeod-McCarty experiment provided evidence that DNA was, in fact, the molecule in which the genetic information was encoded [2] and another nine years before Watson and Crick published their seminal 1953 paper on the structure of DNA [3]. In the succeeding years, the mechanism of DNA replication was resolved to be semi-conservative [4], Francis Crick proposed the central dogma of molecular biology [5] and the first cancer proto-oncogene was identified as *c-SRC* in the Rous sarcoma virus transforming our understanding of cancer [6].

Perhaps the development with the biggest impact on our understanding of evolution and health relating to DNA is not the modes of inheritance defined by Gregor Mendel but that of Frederick Sanger, a two-time Nobel prize-winning British biochemist. In 1977 Sanger developed a chromatography-based method to read the nucleic acid sequence directly [7]. The advancement of nucleic acid sequencing quickly replaced comparative biochemistry and peptide sequencing in understanding molecular evolution by uncovering the role of mutation outside of the coding region.

Early DNA sequence studies investigated heritable differences between nucleic acid sequences across species. This led to the development of many mathematical models of DNA evolution. These models exploited the molecular clock, which was largely in agreement with estimates from the fossil record and comparative proteomics [8, 9, 10, 11]. Building on the established molecular evolutionary theory from the twenty-first century, the evolution of clonally expanding tissues could be modelled using somatic mutation data to understand cancer aetiology and progression [12, 13].

In addition to understanding the evolutionary dynamics of cancer, by sequencing an individual's genome, genomic medicine has enormous potential to stratify patients into treatment groups but also for use in cancer prevention. Individuals with familial histories of disease can be screened for Mendelian diseases, such as hereditary cancers due to *BRCA1/2* variants or Lynch syndrome, removing the need for invasive mastectomies or guiding clinical action towards better treatments by genotype stratification [14]. In recent years, researchers have been able to aggregate small effect variants into clinically relevant polygenic risk scores (PGS/PRS). PGS can contribute to quantifying the risk of developing diseases such as cardiovascular disease and cancer [15].

1.1.2 DNA chemistry and structure

DNA exists in many possible conformations, with B-DNA being the most abundant across life [17]. The nucleic acid sequence is a linear series of the nitrogenous bases adenine, thymine, cytosine, and guanine bound to a sugar-phosphate backbone. DNA is directional (anisotropic), with the carbon position on the 5-carbon deoxyribose sugar annotating the sequence's direction. As the linear strand is unstable, a second complementary strand is bound in the opposite direction and the two are held together by hydrogen bonding. The complementary strands form a major and minor groove within the DNA structure. The major groove is formed where the anti-parallel phosphate backbones are furthest apart due to the base pair orientation. The major groove is a substrate for DNA binding proteins that regulate gene activity within a cell [18]. The outer phosphate backbone of the DNA strand has a negative charge giving the DNA molecule a slightly negative charge. The electrical charge allows for the control of chromatin conformation through epigenetic modifications (nucleosome charge modulation).

The negatively charged DNA is packaged within the cell nucleus by wrapping 1.65 times (equivalent to $\approx$ 147 base pairs) around a positively charged octamer protein complex called a nucleosome, (Fig. 1.1). This structure, sometimes referred to as

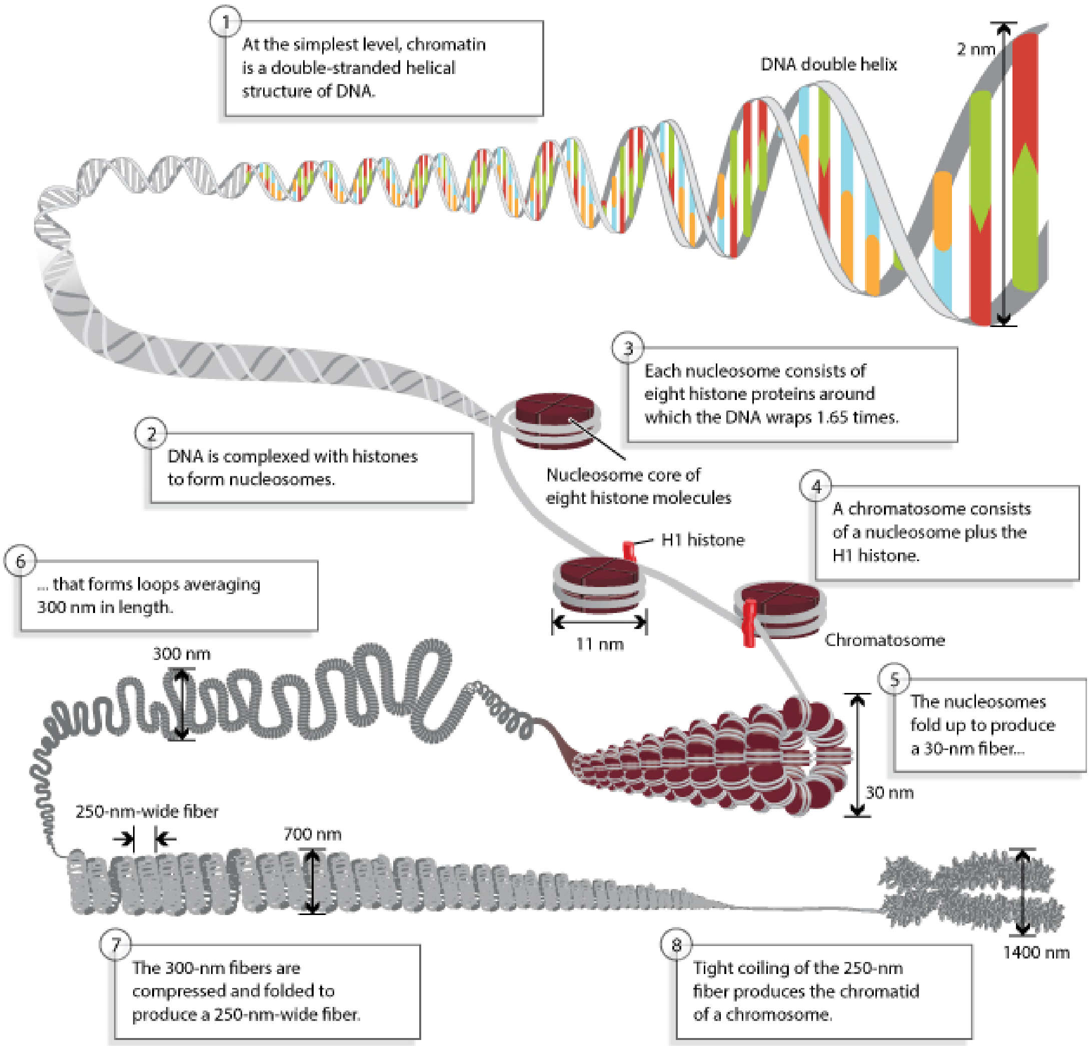
1
At the simplest level, chromatin is a double-stranded helical structure of DNA.
DNA double helix
2 nm
3
Each nucleosome consists of eight histone proteins around which the DNA wraps 1.65 times.
2
DNA is complexed with histones to form nucleosomes.
Nucleosome core of eight histone molecules
4
A chromatosome consists of a nucleosome plus the H1 histone.
H1 histone
6
... that forms loops averaging 300 nm in length.
11 nm
Chromatosome
300 nm
5
The nucleosomes fold up to produce a 30-nm fiber...
30 nm
250-nm-wide fiber
700 nm
1400 nm
7
The 300-nm fibers are compressed and folded to produce a 250-nm-wide fiber.
8
Tight coiling of the 250-nm fiber produces the chromatid of a chromosome.

Figure 1.1: The basic structure of DNA within the nucleus of a cell. The secondary structure of DNA (top right) is wound around protein histone complexes to form nucleosomes. Nucleosomes are coiled to form the 30-nm chromatin fibre in turn coiling to form the 250-nm chromatin fibre. Super coiling of the chromatin fibre forms the quaternary condensed chromatid. Figure reused under creative commons license [16].

‘beads on a string’, allows the secondary 2nm DNA structure to be coiled into a tertiary 30nm 3-dimensional chromatin fibre. The 30nm chromatin fibre is then supercoiled into a higher-order chromatid structure by forming loops anchored by the proteins cohesin and CTCF [19]. The looping structure has two main functions. Firstly, it allows the six billion base pairs in the nucleus to be compactly organised. Secondly, it allows for the formation of functional topologically associated domains (TADs) [19]. TADs are megabase structures with common genomic features such as gene expression levels, lamina interaction, histone chromatin interactions and replication timing [19].

1.1.3 DNA replication

The ability of a cell to divide and replicate is a fundamental process in almost all tissues, apart from denucleated blood cells and terminally differentiated cells [20]. Despite recent work on assessing the variation in replication timing across individuals the mutation rate variation due to replication error, the inter-individual mutation rate variation remains poorly understood [21]. However, the advent of sequencing data has allowed researchers to uncover mutational processes such as DNA replication mutation signatures and replication-associated mutational asymmetry (R-Asymmetry) [22]. The lifetime number of stem cell divisions in a tissue has also been proposed as a risk factor for cancer [23]. The mutational load data derived from cancer datasets supports this proposal, but the correlation between mutation burden and cancer risk does not reflect the strength of the correlation between stem cell divisions and cancer risk [23, 24].

DNA is replicated semi-conservatively, giving rise to two daughter cells, each with 50% of the parent DNA and 50% newly synthesised via DNA polymerase [4]. The cell cycle stage where DNA replication occurs is called the synthesis phase (S-phase). As DNA is anisotropic, and replication is bidirectional [25], DNA polymerase synthesises the new strand in the 5’ to 3’ direction. This poses a challenge for the replication fork running in a 3’ to 5’ direction. To address this, the replication machinery forms 200bp single-strand loops and synthesises what are known as Okazaki fragments in the 5’ to 3’ direction [26]. DNA ligase joins the discontinuous fragments together. This process of ‘leading’ and ‘lagging’ strand synthesis gives rise to an asymmetry in replication-associated mutation profile with errors most likely to occur on the lagging strand, owing to the extended period the lagging strand is single-stranded. As maintaining genomic integrity is a fundamental process within the cell cycle, *Goulian et al.* proposed in 1968, that the nuclease activity of the DNA polymerase they had observed may play a role in error correction [27]. In 1974, John Hopfield proposed the biochemical mechanism from which the kinetics of proofreading was derived [28]. When the wrong nucleotide is inserted into the DNA strand, the helical structure is distorted, activating the 3' to 5' exonuclease activity [29].

The fate of a cell, i.e progression to division or to programmed cell death, is determined at several checkpoints throughout the cell cycle. Failure to progress correctly through each cell cycle checkpoint may lead to severe genomic instability, called replication catastrophe [30]. When DNA lesions or adducts impede DNA replication, error-prone polymerases (DNA pol family Y) may be used to ensure cellular integrity at the cost of introducing somatic mutations. The Y-family of polymerases typically does not have proofreading capabilities [31]. However, severe DNA damage may direct the cell into senescence via apoptosis or other programmed cell death pathways. To safeguard against excessive DNA damage repair during the S-phase, several other mechanisms maintain the integrity of the genomic sequence throughout the cell cycle.

1.1.4 Maintaining genomic integrity via the DNA damage response

Genomic instability is a hallmark of cancer and ageing [32, 33]. A human cell can experience up to 10,000 lesions per day [34], with UV-exposed epidermal cells experiencing up to 100,000 UV-induced adducts per hour [35]. In the absence of repair mechanisms, cells would quickly be overwhelmed by DNA damage and become unviable within a few replications. To maintain genomic stability, eukaryotic organisms have evolved several distinct DNA damage repair pathways, all grouped together under the DNA damage response (DDR) [35]. DDR includes cell cycle checkpoints [36], DNA damage identification and signalling [37], apoptosis [38], replication fork stalling [39], telomere contraction [40], and repair fidelity [41]. Deficiencies in DDR mechanisms have been well-established driving factors in cancer, underpinning the importance of maintaining genomic stability across the soma [42]. On the other hand, the inhibition of the DDR has also transformed cancer therapeutics by exploiting genomic instability in cancer. Perhaps the best-known example of inhibiting the DDR in cancer therapeutics is PARP inhibition in *BRCA1/2* deficient breast and ovarian cancers [43]. PARP plays a key role in BER and NHEJ signalling by recognizing single-strand breaks (SSBs). As *BRCA1 & BRCA2* deficient tumours ineffectively repair DSBs, inhibiting PARP forms excessive DSBs as the SSBs are not repaired by PARP within the tumours, driving the tumour cells into cell death.

The DDR response has evolved several mechanisms for signalling and repairing DNA damage. DNA repair can be categorised by five main pathways depending on the type of DNA damage accrued. These are base excision repair (BER), nucleotide excision repair (NER), mismatch repair (MMR), homologous recombination (HR), and non-homologous end joining (NHEJ), (Fig. 1.2). BER, NER, and MMR are all excision repair pathways. However, they function on different types of damage. BER identifies and repairs non-bulky adducts or DNA damage that does not disrupt the double helix structure, e.g., 8-oxoguanine and AP-sites. NER rectifies bulky adducts such as UV-light-induced thymine dimers and 6,4-photoproducts. The two main NER mechanisms are transcription-coupled NER and global NER giving rise to variation in repair rates across the genome [43]. MMR identifies and repairs the misincorporation of bases during replication and recombination [44]. HR-mediated double-strand break (DSB) repair is restricted to the S and G2 phases of the cell cycle, as the sister chromatid is used as the homologous template. In contrast, in NHEJ, the blunt ends of a DSB are joined together in the absence of a template strand, often leading to deletions within the nucleic acid sequence.

Given the diversity of the DNA damage response to types of DNA damage, it follows that mutations that slip the leash of repair tend to have a negative effect on the overall fitness of an organism [45]. The next section will highlight sources of SNV somatic mutations drawing attention to some key sources of DNA damage that drive mutagenesis and its impact on human health. As this body of work aims to recover signals of somatic mutation within a population-scale dataset, understanding how somatic mutations may arise and some key concepts such as somatic evolution and somatic mutation in health as well as how somatic mutations may differ from sources of technical artefacts within the data is paramount.

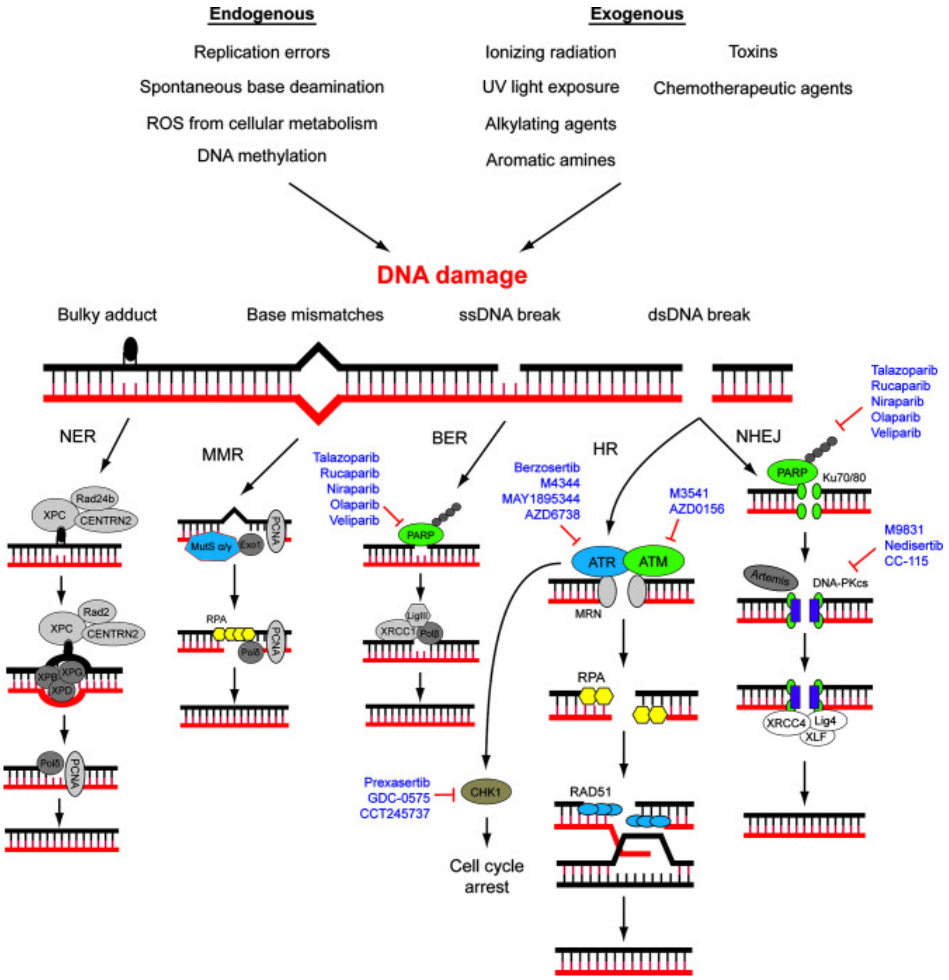

Figure 1.2: Schematic representing the 5 main DDR pathways (NER, MMR, BER, HR & NHEJ). Clinically approved inhibitors that target DNA damage response pathways are illustrated. Figure reused under creative commons license [46].

1.2 Somatic mutation

A somatic mutation is a change in the DNA sequence acquired post-fertilization in non-germline tissues. Somatic mutations accumulate throughout the lifetime of an organism. The single base pair substitution (point mutation, single nucleotide variant [SNV]) is the most frequent somatic mutation. However, any change to the nucleotide sequence, such as insertion/deletions (indels), copy number variants (CNV) and structural variants (SVs) are all classes of somatic mutation. In 2020, the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium reported that, across 2,658 whole-cancer genomes, approximately 95% of all somatic mutations were SNVs [47]. Although purported to be 'hiding in plain sight' [48], somatic mutation has long been proposed as playing a causal role in ageing and malignancy [49].

A subset of somatic mutations may confer a selective advantage to a cell leading to a clonal expansion that carries the mutation to high frequency. Although these mutations may be beneficial to the cell, in the context of multicellular organisms, clonal expansions often convey an increased disease risk to the organism. Mutations that give a selective advantage to a cell are called driver mutations, these typically occur in tumour suppressor genes and proto-oncogenes. However, these terms have been used primarily in the study of cancer, and it was assumed that all driver mutations were pathogenic. Recent work from *Martincorena et al.* found that for some canonical cancer driver genes, such as *NOTCH1*, the frequency of driver mutations is higher in normal, healthy oesophageal tissue than that in cancerous oesophageal tissues [50]. Higa & DeGregori have proposed that *NOTCH1* mutants may have a lower probability of carcinogenesis through selective competition favouring a decoy selective fitness peak [51]. Although the study of somatic mutations in healthy tissues is still in its infancy, over the last five years it has often redefined how cancer aetiology is viewed.

1.2.1 Sources of somatic mutation

The sources of somatic mutation can be categorised as intrinsic and extrinsic. Extrinsic sources arise from exposure to mutagens like radiation such as ultraviolet light, cooked meats and chemical carcinogens such as nicotine smoking [52, 53, 54]. The intrinsic sources of DNA damage arise from by-products of metabolism, such as reactive oxygen species (ROS) and biological processes, such as replication and the innate immune system e.g., the Activation-induced cytidine deaminase (AID)/ Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like (APOBEC) response to viral genomes. Highlighted below are some of the main sources of intrinsic and extrinsic factors that give rise to somatic mutations.

1.2.1.1 Intrinsic sources of mutation

Intrinsic sources of somatic mutations are internal factors that can cause mutations in the DNA of an organism's somatic cells. These mutations can arise during normal cellular processes and intrinsic factors can often accumulate in a clock-like or rate-like fashion over time [55].

1.2.1.1.1 DNA replication errors

DNA replication is a complex and highly accurate process. However, errors can occur during DNA replication, despite its accuracy. As mentioned above, several mechanisms have evolved to identify and correct errors, reducing the mutation rate of DNA replication to the range of 2.8×10^{-7} mutations per base pair [56]. The wrong nucleotide may be incorporated during polymerase elongation, resulting in a mismatched Watson and Crick pairing. In E.coli, the MMR response is directed to the incorrect base using the absence of methylated DNA on the nascent strand [57, 58]. In humans, the precise mechanism of MMR strand discrimination remains unknown but the asymmetric loading of proliferating cell nuclear antigen (*PCNA*) by strand nicks within the daughter strand is a leading hypothesis [59]. During DNA replication, the newly synthesized DNA strand can slip or loop out, leading to the addition or deletion of nucleotides. This can occur particularly in repetitive DNA sequences, where the repeated motifs can cause the replication machinery to slip, resulting in DNA strand length variations and potentially leading to mutations. Depending on the structure and conformation of the DNA loop a copy number variant can arise or the mutational rate around the loop can be increased by strand breaks and the subsequent repair.

DNA polymerases can replicate through DNA lesions via a process called translesion synthesis (TLS) [60]. TLS allows for a cell to progress through the cell cycle, rectifying damaged DNA in the following cell cycle. When DNA polymerase β (or family X member) encounters a thymidine dimer it may be replaced by polymerase η

(or family Y, such as ι & κ), that can correctly insert AA dinucleotides across from the bonded thymidine dimers [61]. This allows the cells to progress through division but increases the error rate as DNA pol η, ι and κ do not have proofreading capabilities [61]. During replication of repeating motifs slipped-strand mispairing (SSM) may occur. In SSM, the template and newly synthesized strands denature, and the synthesised strand then misaligns with a similar motif, leading to the insertion or deletion of nucleotides via NER [62]. Although both expansion and contractions are possible, experimental evidence suggests that repeat expansions occur more frequently [63].

1.2.1.1.2 DNA repair defects

Mutations in genes that are involved in DNA repair pathways can result in defective DNA repair, leading to the accumulation of mutations in somatic cells. As highlighted previously, DNA repair is a key process within a cell to ensure that genomic integrity is maintained. Defective DNA repair machinery, either through inherited variation or acquired somatic mutation, can lead to accelerated ageing and the development of several cancer syndromes such as Cockayne syndrome and Lynch syndrome [64]. The impact of DNA damage repair on human health will be discussed further in the section ‘Somatic mutation and health’.

1.2.1.1.3 Oxidative processes

Oxidative damage is a significant contributor to DNA damage [65]. Oxidative DNA damage can arise from various sources, including reactive oxygen species (ROS). ROS are by-products of normal cellular metabolism, particularly in the mitochondria during cellular respiration [65]. ROS can include metabolised molecules such as superoxide anion (O^{2-}), hydrogen peroxide ($2H_2O_2$), and hydroxyl radical ($OH^{\cdot}$) [66]. ROS can induce modifications of DNA by reacting with the nitrogenous base of the DNA sequence, for example the oxidation of guanine (G) to 8-oxoguanine and other oxidized forms of DNA bases. Other intrinsic processes, such as chronic inflammation, can give rise to the generation of ROS [67]. The methylation of cytosine to 5-methyl-cytosine results in a state that is highly susceptible to deamination to thymine in the presence of hydroxyl radicals [68].

Environmental factors or extrinsic factors can also lead to ROS generation. Exposure to environmental factors such as ionizing radiation, pollution, and toxins can also generate ROS that can cause oxidative DNA damage [69, 70, 71, 72]. Ionizing radiation, such as exposure to gamma rays or X-rays, can increase ROS production in addition to directly altering the DNA sequence through strand breaks, abasic sites and oxidative damage [69]. Ionizing radiation may be intentionally administered as cancer therapy but may have undesirable long-term effects on the health of the patient [73, 74]. Pollutants from excessive traffic congestion, such as particulate matter and polyaromatic hydrocarbons, have also been linked to increased DNA instability and oxidative damage in traffic conductors in Tapei City, Taiwan [70]. Aflatoxin B1 (AFB1) is a potent mycotoxin targeting the liver [72]. AFB1 is produced by the Aspergillus species and is commonly found in grains and feedstuffs [75]. AFB1 is metabolised in the liver into a genotoxic intermediate state, AFB1-exo-8,9-epoxide, by cytochrome P450 enzymes [76]. AFB1-exo-8,9-epoxide reacts with guanine residues through S_N2 substitution. In AFB1-exo-8,9-epoxide exposed cell lines, the mutational spectra were predominantly found on G nucleotides with 50-68% of G-to-N mutations resulting in G-to-T mutations [77, 78]. AFB1 exposure has been attributed to 0.7% of hepatocellular carcinomas (HCC) in north America and 16% of HCC cases in Hong Kong [78]. In addition to exposure to environmental factors, lifestyle factors can also induce ROS production. Poor diet, excessive alcohol consumption, over-cooked meats, and tobacco use can contribute to oxidative DNA damage [52, 53, 54, 79]. Several antioxidant defence mechanisms have evolved to neutralise ROS and repair oxidative DNA damage, including enzymes such as superoxide dismutase (SOD), peroxisomal catalases, and the pre-mentioned DNA repair pathways [80].

1.2.1.1.4 DNA strand breaks and transposable elements

Transposons, or 'jumping genes' are mobile genetic elements that can move within the genome and cause mutations by inserting themselves into the DNA sequence. In a 3000 bp window around transposable elements active in rice and grasses, the mutation rate is ten times higher than the genomic background [81]. While transposons do not directly cause single base mutations, they do form DSBs which are repaired by using a complex system of protein complexes and replication machinery. An example of transposon activity creating DSBs and recruitment of DNA repair complexes within human cells is long interspersed element-1 (L1) [82]. DSB-induced replication or break-induced replication (BIR) has a low fidelity around strand breaks due to increased instability of the replication fork [83]. In somatic mutation data, kataegis is commonly observed around sites of somatic rearrangement [84]. The same BIR mechanism is hypothesised to play a role in the increased mutation burden around chromothripsis events.

1.2.1.1.5 APOBEC & AID activity

AID and APOBEC are enzymes that can cause somatic mutations through nucleotide deamination [85]. Both AID and APOBEC are involved in the immune system in humans, but they can also lead to mutations when they function outside of their normal context. AID is primarily expressed in antibody-producing B cells [86]. AID deaminates cytosine to uracil that pairs with adenine to introduce a C-to-T mutation. AID is responsible for mutating antibody genes during VDJ recombination through a process called somatic hypermutation [85]. Somatic hypermutation is essential for the generation of diversity (GOD) in antibody production that allows the recognition of a wide range of pathogens [87]. However, AID can also have off-target effects when DNA is in a single-strand state during repair, leading to mutations outside of antibody genes. This can result in the development of cancer-causing mutations in certain cases, particularly in B cell lymphomas [88]. Similarly, the APOBEC family of enzymes are involved in the immune response as they protect against retroviruses and other mobile genetic elements [85]. APOBEC enzymes can deaminate cytosine residues in single-stranded DNA, leading to the formation of uracil. If not repaired properly, uracil can pair with adenine during DNA replication, resulting in C-to-T or G-to-A mutations in the newly synthesized DNA strand. APOBEC has increased activity on DNA loops or mesoscale genomic features which are enriched for ‘driver genes’. This observation has cast doubt on the role of some canonical driver genes in cancer [89].

Both AID and APOBEC are tightly regulated in normal physiological conditions to prevent excessive mutations [90, 91]. However, when their regulation is disrupted, such as in certain cancer cells, they can lead to increased mutation rates and contribute to the accumulation of somatic mutations, which can drive tumour development and evolution. AID/APOBEC activity may lead to localised regions of somatic hypermutation called kataegis [22]. As DSBs are being repaired via resection, AID/APOBEC is recruited and deaminates the exposed single-stranded DNA [92].

1.2.1.2 Extrinsic sources of mutation

While intrinsic sources of somatic mutations are inherent to normal cellular processes, they can be influenced by external factors, such as environmental exposures and lifestyle, affecting the rate and types of mutations that accumulate in somatic cells. For example, although oxidative damage can be a source of intrinsic DNA damage, UV, ionising radiation, and inflammation can also generate oxidative DNA damage.

1.2.1.2.1 UV-induced damage

UV light is a form of radiation that is a leading source of DNA damage in sun-exposed skin [54, 93]. In addition to the formation of ROS, UV light exposure leads to two predominant types of DNA damage, the formation of pyrimidine dimers and DNA strand breaks [54]. Pyrimidine dimers are the most common type of DNA damage caused by UV light. UV radiation can induce the formation of covalent bonds between adjacent pyrimidine bases (thymine or cytosine) on the same DNA strand, creating a dimer. This distorts the normal structure of the DNA helix, leading to disruptions in DNA replication and transcription. Single covalently bonded pyrimidine dimers are called 6-4 photoproducts, while the formation of double covalent bonds forms cyclobutane pyrimidine dimers [54]. UV radiation can induce SSBs and DSBs, where the DNA molecule is physically severed into two or more pieces. DNA strand breaks have a highly deleterious effect on the fitness of a cell [94].

1.2.1.2.2 Chemical mutagens and carcinogens

Exposure to certain chemicals or toxins can cause somatic mutations. Chemical mutagens can interact with DNA and cause changes in the DNA sequence. For example, polycyclic aromatic hydrocarbons and nitrosamines found in tobacco smoke [95], aflatoxins produced by fungi [72], and various industrial chemicals have been shown to be mutagenic and can cause somatic mutations [71]. While chemicals may lead to oxidative damage, they also may have other mechanisms of mutation, such as intercalating with the DNA bases (Aflatoxin) and forming bulky adducts through chemical reactions (e.g. alkylation in the case of mustard gas).

Carcinogens are sometimes defined as chemicals that can cause cancer; however, not all carcinogens are chemicals and, of course, not all chemicals are carcinogens. Tobacco smoking is a leading cause of cancer worldwide and as of 2015, there were 933 million tobacco smokers [96, 97]. Cigarettes have been found to contain up to 72 known carcinogens, including, but are not limited to, nitrosamines, polycyclic aromatic hydrocarbons, toxic metals and aldehydes [98]. Nitrosamines (NNN, NNK & NNAL) are alkylating agents that add methyl or ethyl groups to the DNA base or phosphate backbone [99]. Alkylators can cause base mispairing (alkylated-guanine pairs with thymine) or they can lead to the formation of cross-strand links [99].

1.2.1.2.3 Viral mutagenesis

Some viral infections can lead to somatic mutations through multiple mechanisms. As mentioned previously, viral infections can stimulate an inflammatory response which creates ROS. Viral infections can also affect the mutational burden within a cell through direct mechanisms, such as interaction and disruption of the DDR [99] and the integration of viral DNA at fragile sites [100]. Of the 14 million cancer cases in 2014, approximately 9.7% have been attributed to viral infections [101]. Epstein-Barr virus (EBV) and human papillomavirus (HPV) accounted for 760,000 new cancer cases in North America alone [100]. Viruses may recruit host cell DNA repair and replication machinery, depleting the cell of repair proteins resulting in prolonged periods of single-strand DNA and, consequently, increasing the mutation rate at these sites [102]. By reducing the availability of proteins related to repair and replication, viral infections increase the global genomic mutation rate within a cell. In addition to the depletion of DDR resources, viral sequences can also integrate into the host genome at fragile sites. Viral integration causes DSBs, which increase the local mutation rate through BIR. In the ICGC dataset, HPV-positive cancers had approximately 2.9 times more mutations within 1,000bp of fragile sites than HPV-negative cancers [100].

1.2.2 Mutational signatures

1.2.2.1 Definition of mutational signatures

Mutational signatures or patterns are the result of various underlying mutagenic processes or DNA repair mechanisms that lead to specific types of DNA mutations [84, 103]. Mutational signatures can provide valuable insights into the causes and mechanisms of DNA damage and mutation accumulation in various biological contexts, including cancer research, environmental exposure assessment, and evolutionary studies [104, 105, 106].

1.2.2.2 Signature discovery methods

Mutational signatures are typically represented as numerical matrices or graphical plots that summarize the frequencies and patterns of mutations observed in a set of samples. Statistical methods, machine learning, and computational algorithms are often used to identify and characterize mutational signatures from large-scale genomic data. The most widely used technique for mutation signature discovery is non-negative matrix factorisation (NMF). NMF works by factorising the mutation data matrix into two non-negative matrices, one representing the signatures and the other representing the contributions of the signatures to each sample. This assumes that the mutations in a sample are the result of the activity of a number of distinct processes, each with a characteristic mutational signature. The factorization is performed by iteratively estimating the signatures and their contributions to the samples until convergence.

Other statistical and machine-learning approaches are also commonly used. Examples include Bayesian NMF, PCA, ICA and convolutional neural networks, [107, 108, 109]. Mutational signature extraction can be computationally intensive, leading to the development of sequential coordinate-wise descent NMF methods, such as the R package NNLM, which was developed for factorisation of large datasets [110, 111]. In small sample sets with insufficient power to decompose the true set of linear mutational processes, non-negative least squares regression can be used to estimate the contribution of known signatures to a dataset [112]. Mutational signatures have been widely applied to the study of cancer genomes leading to the identification of over 70 verified mutational signatures [113].

1.2.2.3 Databases of mutation signatures

The Catalogue of Somatic Mutations in Cancer (COSMIC) is a publicly available and widely used mutational signature database that contains comprehensive information about the mutational patterns observed in various types of cancer. The current set of cancer signatures has been extracted using data from 23,000 cancer patients [113]. The COSMIC database allows researchers to compare novel mutational signatures against a curated reference with the potential to uncover new cancer biology [114].

1.2.2.4 Relating mutation signatures to sources of mutation

Mutational signature extraction is a purely mathematical approach to uncovering mutation types that covary across samples in a dataset. Relating the resultant signatures to the underlying biology has remained challenging. For example, single base substitution (SBS) signature 5 has been frequently found in cancer and normal tissues [113, 55, 115]. SBS5 exhibits clock-like accumulation but the underlying biology remains unknown, although it is believed to be a result of intrinsic biological processes given its accumulation with age [55].

By considering the clinical metadata of the dataset, the aetiology for each signature can be investigated. However, experimental replication is often required for a mutational signature to be definitively linked to an underlying process, as the exposure history of a sample may not be captured by the sample metadata. *Meier et al.* used MMR knockouts in *C. elegans* to experimentally link the MMR phenotype to the MMR cosmic signatures [116]. Extrinsic factors also contribute to the mutational signature activity within a sample for example, SBS4 and SBS92 have been associated with tobacco smoking and validated in animal studies [117]. Since the 1700's physicians have linked environmental exposures to cancer risk [118]. Recent mutational signature work has examined signatures through an invitro approach, using controlled mutagenesis to directly link mutagen exposure to mutational signatures and making mutational signatures of known carcinogens and mutagens available to researchers [106].

1.2.3 Variation in the somatic mutation rate across the genome

1.2.3.1 Gene expression

Gene expression can impact somatic mutation rate variation across the genome through two main opposing processes, transcription-coupled NER (TC-NER) and transcription-associated mutagenesis (TAM) [119, 120]. As mentioned in the section on DNA repair, NER surveys and repairs the coding region through TC-NER. The process of transcription is mutagenic in itself, due to the unwinding of DNA to a single-stranded state and polymerase misincorporation. The unwinding of DNA causes torsional strain on the nucleic acid sequence and exposes the ssDNA to endogenous and exogenous sources of DNA damage.

Interestingly, the mutation rate is not a linear function of gene expression. Recent work from *Chen et al.* has identified a non-linear relationship between the somatic mutation rate and gene expression [119]. In genes with no transcriptional activity, the mutation rate is high. As expression increases, the mutation rate begins to decrease until it reaches an inflection point and begins to increase once more as a consequence of DNA damage accrued via high levels of transcription. Although not experimentally validated, a likely reason for this increase in mutation at high gene expression levels is that the availability of repair enzymes limits repair efficiency. As gene expression varies from tissue to tissue it is expected that the mutation rate within a gene is not consistent across the soma.

Transcription is a strand-specific process with half of the coding genes on the forward strand and the remaining on the reverse strand. As DNA is preferentially damaged on the non-template strand and repaired on the template strand, a mutational asymmetry is observed in the rates of mutation type. Transcription strand asymmetry has been described in bacteria, mammalian evolution and somatic mutation within cancer [121, 122, 123]. In genetic variation and liver cancers, a strong A-to-G asymmetry is observed over T-to-C, while C-to-A over G-to-T is the predominant asymmetry in cancer samples. Interestingly, in liver cancer, there is an excess of A-to-G mutations on the non-transcribed strand, indicating that the asymmetry is associated not with preferential repair but with preferential damage to the non-transcribed strand [123].

1.2.3.2 Sequence content and chromatin structure

Sequence context drives variation in the somatic mutation rate across the genome [124]. The mutation rate in a 5' upstream and 3' downstream nucleotide context is one of the strongest predictors of the variation across the genome. For example, the sequence context TpCpN shows preferential mutation of C-to-T and C-to-G mutations. Through mutational signature analysis, this preference has been explained by APOBEC activity (SBS2) and is active in 60.7% of cancers [103]. The local sequence context is an important covariate in the analysis of selection acting on genes. Early studies assumed that the substitution rates were constant across the genome [125]. However, the substitution rate is heterogeneous across the genome and accounting for differences in the local mutation rate reduces the number of false positive cases of positive selection using dNdS methods [126].

The GC content of a sequence affects the somatic mutation rate [127]. The local GC content is the proportion of G and C nucleotides within the region. Across the genome, the mutation rate is positively correlated with GC content [127]. However, this relationship is contentious and it is not entirely linear. The level of DNA damage is reduced in regions with high GC content, due to the open and active nature of high GC content sites [128]. Extremely high GC content can result in the formation of stable DNA and G-quadruplex structures that can inhibit DNA repair mechanisms, also increasing the mutation rate [129, 130]. In addition to stoichiometric and DNA stability influences, high GC content regions are also enriched for CpG dinucleotides. In the presence of ROS, such as the hydroxyl radical, the deamination of methylated cytosine to thymine occurs at a rate twice that than unmethylated cytosine [131].

The organisation of chromatin determines the 3D structure of the genome [132]. The 3D structure of the genome plays a crucial role in regulating gene expression and other mutation-associated cellular processes [133]. For example, regions of the genome with tightly packed chromatin structure may be less accessible to DNA repair machinery, increasing the likelihood of mutations [133]. Additionally, regions of the genome that are physically close together in the 3D structure may be more likely to interact, leading to increased rates of mutations through mechanisms such as DNA translocations and copy number changes [134]. Breakpoints introduced during translocations and copy number changes increase the somatic mutation rate through BIR.

Chromatin accessibility can be measured via specialised NGS protocols such as

ChIP-seq or ATAC-seq [135, 136]. ATAC-seq enriches transposase-accessible chromatin. Transposases, such as the hyperactive Tn5 transposase, can cleave and tag open chromatin and ATAC-seq has a number of advantages over ChIP-seq such as dramatically reduced sample preparation time [137]. No prior information is required removing the need for specific protein-binding antibodies and the sensitivity of the assay is increased for open chromatin enrichment [138, 136].

1.2.3.3 Replication timing

The 3D structure of the genome is closely related to replication timing [139]. Replication timing is the temporal order in which genomic regions are replicated during the S-phase of the cell cycle. Replication timing is tightly regulated and varies across different cell types and developmental stages. Differing rates of replication timing between active and X-inactivated chromosomes were first observed in 1960, implicating chromatin organisation as a major factor in the variation in replication timing [140]. Replication timing has a complex relationship with the somatic mutation rate. The inaccessibility of genomic regions to the replication and DNA repair mechanisms increases the somatic mutation load in regions of late replication [141]. Nucleotide pool depletion occurs at the later stages of genome replication. As the number of available nucleotides diminishes the replication machinery stalls increasing the exposure of ss-DNA to mutagens and DSBs [142]. Oxidative damage also occurs in the nucleotide pool. This leads to the incorporation of DNA damage into the nascent lagging strand [143].

1.2.3.4 Use of mutation signatures to study variation across the genome

Mutational processes are not consistently active across the genome. This can be explained via the correlation with genomic features such as sequence content, gene expression, chromatin structure and replication timing features. Mutational signatures can provide insight into the mutagenic processes active within a given genomic region [144]. For example, in germline variation, the activities of inferred mutational signatures closely track with genomic features such as strand-dependent repair and replication timing [144]. Strand-specific mutational signature activity attributed to the replication fork direction has also been observed in cancer somatic mutation data, with an increased mutation burden on the lagging strand [143].

1.2.4 Mutation rate evolution

Although energy is expended on maintaining genomic stability, mutation is the substrate on which evolution acts [145, 146]. However, DDR in response to somatic mutation is inefficient at removing all DNA damage lesions. The inability of selection to reduce the mutation rate to zero has been explored in an evolutionary theory called the drift-barrier hypothesis, in which the effective population size controls the efficiency to optimise a trait [147]. The rates of DNA repair in the germline and the soma are in stark contrast, with the latter having up to two orders of magnitude higher mutation rate [148]. Several possible reasons exist for this observed difference, the number of cell divisions, decreased expression of repair proteins, and elevated levels of mutagenic metabolomic by-products. Nevertheless, what is clear is that the observed mutation rate differences are consistent across species [149, 150].

1.2.4.1 Measuring the somatic mutation rate

The somatic mutation rate can be defined using several different representations. Typically, the number of somatic mutations per base pair is normalised by the number of cell divisions or per unit of time. For somatic mutation data, the rate at which a cell type divides is heterogeneous across cell types and indeed, between different ages within the organism's life cycle [21, 151]. Given this heterogeneity within cell types, representing the somatic mutation rate as an expected number of mutations per base pair per unit of time for a given cell type may not be accurate for the comparison of somatic mutation rates across different studies of the same cell type. Mathematically, the somatic mutation rate can be represented as:

$$\mu = \frac{m}{G \times C \times T}$$

Where μ is the somatic mutation rate, m is the number of somatic mutations, G is the genome length, C is the average sequencing coverage from the NGS data and T is the unit of time, typically, per years of age. The advent of single-cell whole-genome sequencing has allowed for accurate estimations of the total number of somatic mutations expected per cell for a given sample age [152]. This somatic mutational load per cell can be easily transformed to a per base pair per year estimate. The somatic mutation rate is an important parameter in studies of cancer, ageing, and other diseases, as it provides quantitative information about the frequency and dynamics of somatic mutations in different biological contexts. Accurate estimation of the somatic mutation rate provides insights into the underlying mechanisms of mutagenesis, DNA repair, and genome stability, as well as the contribution of somatic mutations to disease development and progression [153].

1.2.4.2 Evidence of selection on somatic mutations

While somatic mutations were once thought to accumulate randomly and without the influence of natural selection, there is increasing evidence that selection acts on somatic mutations even in healthy tissues [13, 154]. Genes under positive selection highlight key cellular processes that can drive clonal expansions and malignancy [13, 154, 155]. Selection acting upon a genic region is measured using molecular evolutionary techniques such as the dN/dS method (sometimes also referred to as Ka/Ks or ω). dNdS is a widely used metric that consists of the ratio of the number of non-synonymous substitutions per non-synonymous site (dN) to the number of synonymous substitutions per synonymous site (dS) between two or more DNA or protein sequences. A key assumption of the dNdS method is that synonymous mutations accumulate neutrally [156]. This assumption however has been contested in the literature where a subset of synonymous mutations in *Saccharomyces cerevisiae* were found to evolve non-neutrality via disruption of mRNA levels within a cell, thus, decreasing the fitness [157]. However, the work of *Shen et al.* has drawn criticism due to technical confounding and incorrect experimental design [158]. Moreover, there are specific biases that can be introduced by using simple models of DNA substitutions, for example the Jukes and Cantor model assumes that all substitutions occur at an equal frequency these simpler models are also biased by the nucleotide sequence around the DNA substitutions [10, 125]. To model sequence context and varying rates of nucleotide substitution *Martincorena et al.* have developed a 192 substitution rate model with a transition/transversion ratio correction factor similar to *Goldman and Yang.* [13, 159]. dNdS ω values greater than one imply that a subset of the non-synonymous mutations have accumulated under the influence of positive selection. In contrast, dNdS values less than one imply that the non-synonymous mutations were removed via negative selection.

Perhaps the most surprising result from dNdS analysis in cancer is that there is very little negative selection compared to the germline. *Martincorena et al.* found that across all cancer types in ICGC, the vast majority of genes (98%) are evolving neutrally, with only 0.2-0.5% under negative selection [13]. Of the 179 cancer genes that showed evidence of positive selection, 54% were known canonical cancer genes. Applying the dNdS methodology to somatic mutation data derived from healthy tissues has provided fascinating insights into the role of somatic mutation in tumorigenesis [160, 93, 161]. In early cancer studies the recurrence of a mutation across cancer was a tell-tale sign of pathogenicity, leading to the term 'driver mutation' to define a mutation that is intrinsic to tumorigenesis.

The application of dNdS to normal healthy tissues identified that some driver mutations, such as *Notch1* mutations, are more frequent in healthy tissue than pathological tissues. Cell clones in healthy tissues have been shown to expand and contract in response to environmental changes [162, 163]. To date, little research has been conducted on the role of somatic mutation in a protective context. One study conducted by *Wang et al.* found that somatic mutations in non-alcoholic steatohepatitis (NASH) disease genes associated with lipotoxicity suppression were under strong selection suggesting novel therapeutic targets [164]. The findings of *Wang et al.* are consistent with evolutionary decoy models, which aim to explain the higher frequency of *NOTCH1* mutants in healthy tissues compared to tumour samples [51].

1.2.5 Somatic mutation and Health

1.2.5.1 Cancer

Cancer is a disease of the genome and the best-studied example of the impact of somatic mutations for human health [165, 166, 167]. Cancer itself is not a singular pathology but a collection of diverse and complex pathologies grouped by cell or tissue type, often with shared causal aetiologies [168]. It remains a leading cause of death worldwide, with over 10 million deaths in 2020 [169]. In 2021, 6.4 billion dollars was appropriated to the US National Cancer Institute (NCI) across various projects, such as the cancer moonshot program that aims to facilitate scientific discovery, foster collaboration between institutes and increase the sharing of cancer data [170, 171].

Aggregated cancer data has been actively curated and stored in databases for nearly

20 years [172, 173]. Interestingly, the majority of cancer-associated genes were identified before the boom in cancer NGS data. Several early studies identified specific somatic mutations in cancer-associated genes, such as *TP53* coined ‘guardian of the genome’, *KRAS*, *PI3K* and *BRAF* [174, 175, 176, 177]. NGS data gave an unprecedented insight into the molecular mechanisms of cancer and the evolutionary process active within tumours.

An interesting observation that has arisen in cancer evolution is that the class of mutation type acting on tumour suppressor genes (TSGs) and oncogenes differ. TSGs show strong selection for truncating mutations, while oncogenes show preferential positive selection in missense mutations [13]. These mutations can confer increased cell survival, proliferation, and resistance to treatment, providing a selective advantage to the cancer cells carrying these mutations. Additionally, studies have shown that somatic mutations in cancer cells can be subject to negative selection as well, albeit a small proportion, where certain mutations are removed from the tumour population due to their detrimental effects on fitness. Mutations that trigger an immune response (neoantigens) have been reported to be under negative selection [178]. This result, however, is contested within the literature [179].

Normal physiological mutagenic processes and environmental exposures do not solely drive cancer aetiology. Genetic variation acting on DNA repair proteins can influence an individual’s risk of cancer development [180]. About 5 to 10% of cancers have been attributed to inherited genetic variation [181]. *BRCA1* and *BRCA2* defects in breast and ovarian cancers are amongst the most studied genes in hereditary cancer or cancer syndrome [182, 183, 184]. *BRCA1* and *BRCA2* play a role in DSB repair but have different functions within the DDR response [185]. Although it is well-established that genetic modifications can dramatically affect the mutation rate, it is not clear whether variants with a weak effect on the efficiency of DNA polymerases and repair enzymes have an impact on the mutation rate. As more data on somatic mutation becomes available, the power of studies to identify weak modifiers of the mutation rate will be uncovered, shedding light on the variation driving somatic mutation.

1.2.5.1.1 Mismatch repair mutational signatures in Lynch syndrome samples

Hereditary non-polyposis colorectal cancer (HNPCC), also known as Lynch syndrome, is an inherited cancer syndrome caused by mutations in genes involved in DNA MMR, primarily *MLH1, MSH2, MSH6*, and *PMS2* [186]. Lynch syndrome accounts for 2 to 3% of all colorectal cancers, with population-wide prevalence estimates ranging from 0.4% in Iceland to 0.05-0.27% across diverse populations [187, 188]. In addition to colorectal cancer, the risk of developing other cancers is increased, these include endometrium, ovary, stomach, small intestine, pancreas, biliary tract, urinary tract, and brain [189].

In Lynch syndrome, the loss of MMR function leads to a high frequency of somatic mutations and instability of microsatellites, short DNA sequences that are prone to replication errors. Microsatellite instability (MSI) is a hallmark of Lynch syndrome. In immuno-oncology tumours are annotated as hot or cold, reflecting MSI high/low status. MSI and the high mutation burden in Lynch syndrome have important clinical implications for targeted immunotherapies. They can be used as diagnostic and prognostic markers, as well as predictors of response to certain treatments. For example, tumours with high levels of MSI in Lynch syndrome have been shown to be more responsive to immunotherapy with immune checkpoint inhibitors, which can activate the immune system to attack the tumour cells [190].

1.2.5.1.2 Repair defects in nucleotide excision repair

An inherited pathogenic mutation in NER genes can lead to severe disorders and syndromes. Loss of function in *ERCC8* (CSA) or *ERCC6* (CSB), two genes integral to the TC-NER response, lead to Cockayne syndrome (CS) [191]. *ERCC6* also plays a role in DSB repair [192]. The life expectancy of individuals with Cockayne syndrome is, on average, 12 years [193]. Individuals with Cockayne syndrome exhibit a range of phenotypes, including, photosensitivity, 'failure to thrive', microcephaly and progeria [193, 191].

Xeroderma pigmentosa (XP) is a rare photosensitive disease caused by mutations in any of the seven *XP* genes (*XPA through G*) and, rarely, *ERCC1* [194]. The XP genes are critical for the removal of UV-induced adducts during NER. Patients diagnosed with XP have a 10,000 times increased risk of basal cell carcinomas and 2,000 times increase in melanoma risk [195]. XP is a progeria syndrome with individuals experiencing accelerated ageing phenotypes, shorter lifespans and in 20-30% of cases neurological problems and intellectual deficiencies [194]. In leukaemia, patients with XP mutations have 25 times higher mutation burdens than non-XP leukaemia [196].

Although CS and XP have shared aetiology via loss of NER, they have remarkably different somatic mutation rates and consequently risks of cancer progression [197]. The difference between XP and CS arises in the NER repair process, in XP the global NER response is deficient leading to the accumulation of UV-induced adducts leading to high mutation rates. In stark contrast, CS has deficiencies in TC-NER, resulting in a mutation rate consistent with cancers of unknown aetiology [197].

1.2.5.2 Neurological and psychiatric disorders

Somatic mutation has been widely implicated in neurological and psychiatric conditions. As mentioned above, a substantial proportion of XP patients experience neurological symptoms that can develop in childhood up to the third decade of life [198]. The resulting neurological conditions can range from mild to severe ataxia, deafness and intellectual disability [199]. In the most severe form of XP, De Sanctis-Cacchione, patients exhibit several other neurological symptoms such as hyperreflexia and altered speech [198].

What remains unclear is whether the increased mutation burden in XP is causative or if the pathogenic mutation in the NER genes acts in another pathway resulting in the observed neurological symptoms. The non-availability of brain tissue inhibits the large-scale analysis typically required to understand the role of somatic mutation. Somatic alterations are well-studied in diseases arising from genetic anticipation [200]. In anticipation, an expansion, usually in a triplet repeat, leads to somatic instability. The phenotype is dependent on the location of the expanded repeat, for example, in Huntington's disease, the CAG repeat expands with age, leading to increased genomic instability and the age-dependent onset of the disease [201]. The expansion of the CAG repeat in Huntington's disease leads to the loss of the regulatory promoter region. Fragile X syndrome is in stark contrast to Huntington's disease as the onset of neurological problems occurs during development while the physical symptoms do not develop until early adolescence [202].

Understanding the role of acquired single-base somatic mutations and their impact on neurological health is poorly understood. The vast majority of studies of somatic mutation in neurological disease arise from a phenomenon called somatic mosaicism [203]. Somatic mosaicism occurs early in development, with the propagation of the mutation depending on the location in the early tissue and the time at which the mutation occurred in differentiation [204, 203]. Somatic mosaicism is conceptually similar to the clonal expansion of cells later in life. As previously described clonal expansions may not only contribute to pathogenesis but also to improving the cellular fitness in tissues in the presence of disease resulting in a less severe disease phenotype [205, 163]. The clonal expansion of blood stem cells is an important risk factor for haematological malignancies and a range of diseases, such as CAD and cerebral infarction [206, 207].

1.2.5.3 Clonal haematopoiesis of indeterminate potential

Clonal haematopoiesis of indeterminate potential (CHIP) is the expansion of a single hematopoietic stem cell that has acquired a somatic mutation typically in epigenetic regulators, DNA repair genes and splicing factors. The most commonly mutated genes in CHIP are the epigenetic regulators *DMNT3A, TET2* and *ASXL1*; DNA repair genes *TP53* and *PPM1D*; and splicing factors *SF3B1*, and *SRSF2* [208]. CHIP typically develops late in life as the prevalence of CHIP is 1% in those under 40, increasing to 10 to 20% in those over 70 [209, 210, 211].

CHIP is a particularly interesting phenomenon for a number of reasons. CHIP is not a malignancy in itself, and while there is a shared aetiology between the drivers of CHIP and leukaemia, only 4% of individuals with CHIP develop blood cancer [211]. The number of haematopoietic stem cells (HSCs) is estimated to be in the tens of thousands this number of stem cells promoted diversity and redundancy in the stem cell population [212]. In a 115-year-old healthy woman, all of the nuclear DNA was derived from a single HSC clone [213]. No haematological malignancies were present, indicating that normal blood production can continue despite loss of diversity in the HSC pool. However, the telomere lengths were found to be shorted in blood cells compared to other tissues indicating a finite lifespan of HSC. An important issue with samples in CHIP is that somatic mutations can be carried to high frequency, this is problematic as distinguishing between germline and somatic mutations requires a second sample to accurately call high-frequency somatic mutations.

The clonal dynamic of CHIP shows that for different mutations, the age at which the mutation occurs also impacts the trajectory of the clone [214, 215]. For example, *DMNT3A* mutated clones preferentially expand quicker when arising in early life compared to later in life, whereas mutations in splicing genes expanded later in life [214]. At the outset, the differing fitness effects as a function of age are surprising as the loss of function in genes should not alter the cellular landscape. This can be reconciled by considering the biology of the main genes mutated in CHIP. Epigenetic regulators are the most common mutated genes in CHIP, and epigenetics is a well-established contributor to ageing [216]. Deregulation of epigenetic regulators leads to the loss of nucleosomes and heterochromatin [217]. Demethylation of histone mark H3K4me3 can also impact the longevity of organisms [218]. By understanding the relationship between age and mutagenesis, we can begin to uncover the variation in risk across lifespan due to age-related changes in the cellular landscape.

1.2.5.4 Ageing

Under normal physiological conditions, somatic mutations accumulate approximately linearly throughout life [219, 152]. Somatic mutations have been implicated in the ageing process for more than 60 years by Leslie Orgel and Leo Szilard in two separate theories [220, 221]. The Orgel model implicates aberrant ribosomal proteins leading to an ‘error catastrophe’ as the reduced efficacy of the ribosome to produce functioning proteins creates a feedback loop accelerating the ageing process. While no experimental evidence has supported this theory it has not been refuted [222]. The Szillard model, however, assumes a ‘two-hit‘ process with somatic mutations accumulating linearly with the ploidy of the organism. Building on the redundancy of the second copy of the chromosome in each cell, once the second ‘hit‘ occurs in the second copy of a gene, the non-linearity of the ageing phenotype in older individuals is explained.

As highlighted by *Millholland et al.*, this model breaks down when ploidy is taken into account across different species, while it is not inconceivable for a three- or four-hit model to exist, triploid flies do not outlive their diploid counterparts nor do haploid wasps have a reduced lifespan compared to diploid members of the same species [223, 224]. *Millholland et al.* have proposed the ‘somatic mutation catastrophe theory of ageing’, which borrows elements from the Orgel and Szilard models [222]. The Millholland model, suggests that the altered gene expression, in addition to altered protein sequences buffered by ploidy, creates the ageing phenotype. This may be overly simplistic to explain the complexities of ageing. Moreover, as we have described previously in this review, the integrity and stability of the genome are controlled by many genes (repair genes, epigenetic regulators etc...) and functional regions (telomeres, and regulatory sequences).

1.3 High-throughput short read sequencing

1.3.1 Background

High-throughput sequencing (HTS) is a relatively new technology that revolutionized the field of genomics. The history of HTS can be traced back to the 1990s when the first massively parallel DNA sequencing method, known as pyrosequencing, was developed by Pål Nyrén and Mostafa Ronaghi at the Royal Institute of Technology in Stockholm, Sweden [225, 226]. Similar to methods developed by Fred Sanger, pyrosequencing used sequencing-by-synthesis, i.e, the incorporation of dNTPs by a DNA polymerase. The key improvements in pyrosequencing over Sanger sequencing were both in the throughput of sequencing reads produced and as the dNTPs are incorporated into the new strand, luciferase is cleaved, allowing the direct reading of the DNA molecule to avoid the laborious and time-consuming electrophoresis stage.

The first commercially available HTS platform was 454 Life Sciences' Genome Sequencer 20, which was released in 2005. Using the pyrosequencing technology, the Genome Sequencer 20, used enzymes to produce light when nucleotides were added by DNA polymerase to a template DNA strand [227]. This light was then detected and used to determine the sequence of the DNA. Other NGS platforms quickly followed, including the Illumina Genome Analyzer, based on the Solexa technology and the SOLiD sequencing system from Applied Biosystems (now Thermo Fisher Scientific). These platforms used different sequencing technologies, such as reversible terminators (Illumina) and sequencing by ligation (SOLiD) and produced much higher throughput than the Genome Sequencer 20 [227].

Sequencing-by-synthesis relies on detecting fluorescent signals emitted during the synthesis of new DNA strands. Sequencing-by-synthesis technologies, such as the Illumina-acquired Solexa system, have dominated the short-read market. Illumina uses a technique called 'bridge amplification' to amplify the DNA fragments and generate clusters of identical sequences that can be sequenced in parallel. This simplifies the library preparation stage as the DNA is amplified on the flowcell directly. As the data used in this body of work is generated using the Illumina system, the sequencing-by-synthesis protocol used by Illumina is detailed below.

The genomic DNA is fragmented, and adapters and unique barcodes are ligated to the ends of the fragments. These adapters contain sequences that are complementary to the primers used in the sequencing reaction. The adapters are annealed to a solid surface, such as a glass slide of the flowcell, forming a lawn of DNA fragments, (Fig. 1.3). A primer is annealed to the adapters, and DNA polymerase is added to extend the primer, creating a new DNA strand. As the new strand grows, it eventually dissociates from the template, leaving a single-stranded 'bridge' between the new and old strands. Another primer is annealed to the new strand, and the process is repeated, resulting in a cluster of identical DNA strands that are each attached to the surface by one end. The clusters are then amplified by bridge PCR, which involves adding a second primer that anneals to the opposite end of the DNA fragment and extends the DNA strands across the bridge. The process is repeated multiple times, resulting in millions of clusters of identical DNA sequences that are densely packed together on the surface. A fluorescently labelled reversible terminator nucleotide is added to the reaction, and the nucleotide is incorporated by DNA polymerase and identified based on the fluorescent signal. The fluorescent group is cleaved from the nucleotide, allowing the next nucleotide to be added, and the process is repeated for many cycles. The sequence of the DNA is determined by analysing the pattern of fluorescence in each cluster. The reversible terminators allow for control of how many bases are added, resulting in a fixed read length.

Whole genome sequencing (WGS) captures the full genetic sequence of an individual. This, however, may be prohibitive in terms of cost and computational storage. Targeted sequencing allows for the sequencing of specific regions of the genome, such as the exome in the case of WES and in oncology gene panels, where only genes clinically relevant to cancer are sequenced. Targeted capture kits can use four different types of bait or probe sequences to capture the target region, dsDNA (e.g., TWIST), ssDNA (e.g., IDT xGEN Exome Research Panel v1.0), ssRNA (e.g., Agilent) and dsRNA (e.g., Dynegen). The target capture used in the UK Biobank WES was the IDT's xGEN Exome Research Panel v1.0, a ssDNA bait library protocol that captures 39 MB of exome sequence.

The Illumina technology can be used for WGS, WES and targeted sequencing. The library preparation step is a critical part of Illumina sequencing that involves the construction of DNA or RNA libraries from the original sample, which is then used for downstream sequencing. The library is a set of adapters which are ligated to the 5' and 3' ends of the sheared DNA fragments. The adapters contain a number of distinct se-

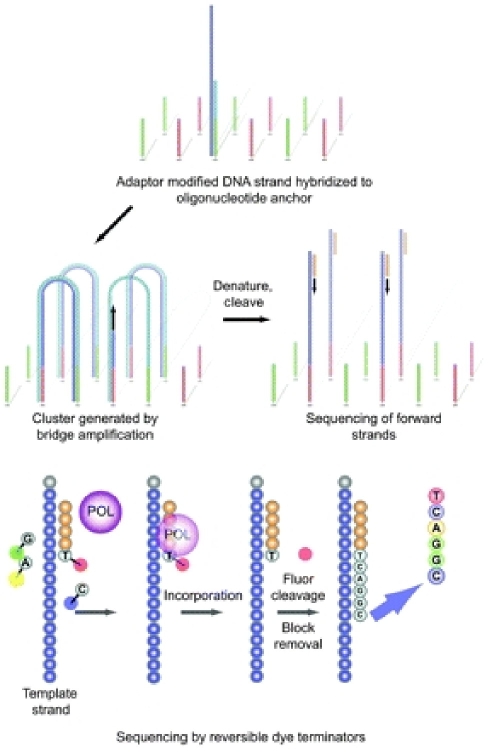
Adaptor modified DNA strand hybridized to oligonucleotide anchor
Denature, cleave
Cluster generated by bridge amplification
Sequencing of forward strands
POL
G
A
T
C
Template strand
POL
Incorporation
Fluor cleavage
Block removal
T
C
A
G
G
C
Sequencing by reversible dye terminators

Figure 1.3: Overview of the Solexa-Illumina sequencing-by-synthesis protocol. Permission granted for reuse from Oxford academic under license 5540830021859 [228]

quences, such as the P5 and P7 sequences which bind to the complementary oligos on the flow cell. The adapters also contain a unique tagging sequence identifying reads belonging to a specific sample in pooled (multiplexed sequencing) runs along with unique molecular indices (UMI) that allow for identifying specific DNA fragments. There are different types of Illumina library preparation protocols available, each tailored for specific applications. Commonly used Illumina libraries are TruSeq, NextEra and AmpliSeq. In the UK Biobank WES dataset, the IDT xGEN research capture kit was used to capture the coding region of the genome. The captured DNA was PCR amplified using the KAPA HiFi polymerases, before sequencing on an Illumina NovaSeq 6000 machine at the Regeneron sequencing facility [229].

1.3.2 Types of nucleotide mismatches in NGS data

The objective of this thesis is to investigate sources of variation in somatic mutation in low-depth sequencing data from large numbers of sequenced individuals. To understand somatic mutation within single-sample sequencing datasets, one must understand sources of error within NGS data. We have previously detailed the sources of DNA damage that may lead to somatic mutations. DNA is constantly damaged and repaired in normal physiological conditions; however, when samples are taken for sequencing, multiple factors can lead to increased DNA damage, the damaged DNA is then sequenced, leading to mismatches in the data which are indistinguishable from somatic mutations. Typically, mismatches derived from damaged DNA occur only on one NGS read. For germline calling, this is not a substantial issue as genetic variation should be present on 0%, 50%, and 100% of reads up to sampling bias and in the absence of copy number variation. Next, we will detail some of the major sources of error in NGS data.

The first major contributor to noise in NGS datasets is sequencing errors. Although the accuracy of base calls in modern sequencers has significantly improved in recent years there are still approximately 0.1-1% of erroneous bases due to sequencing errors [230]. In paired-end sequencing of DNA fragments shorter than twice the read length a proportion of the DNA fragment is sequenced twice. By sequencing the molecule twice and assuming that sequencing errors are independent, the sequencing error on overlapping read pairs can be reduced by the square of the sequencing error rate at those sites [231]. Sequencing errors arise during the incorporation of the dNTPs during the synthesis stage of sequencing and may be due to polymerase bias or the wrong base being incorporated [232]. As with endogenous polymerase activity, sequencing-by-synthesis has high error rates at repetitive sites due to polymerase slippage [233, 234]. Chemical modifications from DNA damage during library preparation, such as 8-oxoguanine, can give rise to a G-to-T mismatch during sequencing. To offset DNA damage, DNA sequencing library preparation steps typically include DNA repair enzymes, however, this repair process is limited by the exposure of the sample to the source of DNA damage, I.e. exposure to heat such as room temperature or by the use of expired reagents. During the clinical investigation of suspected pathological tissues, the biopsied material may be formalin-fixed to preserve the tissue for histological analysis. Subsequent sequencing of formalin-fixed paraffin-embedded (FFPE) tissues can introduce DNA lesions that are difficult to distinguish from somatic mutations [235]. Although sequencing of fresh specimens remains the best option for reducing false positive mutation calls, some computational approaches have been developed specifically for FFPE-sequenced tissues [236].

As we detailed previously, DNA polymerases have high fidelity. During amplification, PCR may introduce two main error types [237]. Firstly, duplicate reads may arise, these can be identified using the optical location within the flow cell and the unique barcode. The second type of error introduced by PCR is the incorporation of the wrong nucleotide. Identification of PCR errors can pose a challenge because if the error occurs in an early PCR cycle the number of reads containing the mismatch will grow exponentially. This may be problematic in some cases as the base quality score may indicate that the base call is accurate.

The alignment of the sequencing read to the reference genome may also give rise to mapping artefacts. Mapping artefacts are prevalent around recombinant sites, such as structural variants and repetitive regions such as microsatellites, centromeres and telomeres [238]. Typically, regions where multi-mapping is likely to occur, are removed from downstream analysis. Although substantial research has been conducted on genome stability in repetitive regions, for example, in genetic anticipation of Parkinson's disease, the exclusion of repetitive regions in somatic mutation has led to a blind spot in our understanding of mutagenesis.

1.3.3 How somatic mutations are called

An important stage in calling somatic mutations is removing germline variants. The nomenclature surrounding a DNA mutation can be ambiguous and often depends on the context in which the work is presented or the frequency of a mutation within a database. Broadly speaking, cells within an organism such as Homo sapiens can be split into somatic and germline cells. The DNA in germline cells is passed onto the progeny, while somatic cells make up all other tissues and are not passed on to the progeny. Mutations that arise in the germline are present in populations of individuals at a frequency influenced by parameters such as their selective coefficient and the population size [239]. These mutations are grouped by the size of the DNA alteration and often vague frequency thresholds. For example, single nucleotide variants that appear on more than 1% of chromosomes in a population are termed single nucleotide polymorphisms (SNP), whereas germline variants that occur at a frequency less than 1% are called single nucleotide variants (SNVs).

Large population sequencing projects such as 1000 Genomes Project and GNOMAD have characterised the most common variants across multiple populations [240, 241]. Over 1 billion SNPs are characterised within the dbSNP database (v155), and approximately 4-5 million SNPs are in each individual's genome [242]. Identification of germline variants is carried out against the backdrop of a reference genome. At homozygous sites, we would expect to see all DNA reads containing the reference or alternative allele. For heterozygous sites, 50% of DNA reads would be expected to contain the reference allele and 50% the alternative allele.

In contrast to genetic variant detection, the recovery of somatic mutations in cancer samples relies on using a normal tissue sample from the same individual. Variants observed in the tumour sample are compared against those found in the 'healthy' tissue and differences that pass quality control are called as somatic mutations. For the majority of somatic mutations, we expect the frequency of reads to be less than 50%. In principle, mutations can be present in more than 50% of reads if mitotic recombination events have occurred, resulting in 'loss of heterozygosity' (LOH) or if the ploidy is affected by copy number changes. The purity of the tumour samples and the clonal structure of the somatic mutation all impact the variant allele frequency.

1.4 GWAS

1.4.1 Objectives of GWAS

Genome-Wide Association Studies (GWAS) is a forward genetic approach used to identify the genetic variants associated with complex traits or diseases. The power of GWAS relies on large sample sizes to uncover the contribution of genetic variation to a disease or trait. SNPs typically contribute a small proportion of the heritable component of trait variation [243]. GWAS leverage technological advances in DNA genotyping (via microarrays and imputation, NGS) and computationally efficient algorithms [244, 245, 246, 247]. Some of the key objectives of GWAS are to identify novel genes associated with disease [248], estimate the genetic contribution to a trait [249] and unravel the complex nature of some complex polygenic diseases, such as CAD and T2D [15].

1.4.2 History

The first GWAS was published in 2005 when researchers identified an association between a genetic variant on chromosome 8 and age-related macular degeneration (AMD), a common cause of blindness in elderly individuals [250]. This study included 96 cases and 50 controls, genotyped at 116,204 sites across the genome. The next advancement of GWAS came in 2007 when the Wellcome trust case control consortium (WTCCC) published 14,000 cases and 3,000 shared controls across seven common diseases, (Fig. 1.4) [251]. The WTCCC identified 24 independent association signals across the seven disease types, some in known risk loci. A key methodological approach of the WTCCC GWAS is the use of imputation. All 17,000 samples were genotyped with the Affymetrix GeneChip 500K Mapping Array Set and imputed with the HapMap reference set to infer approximately 2 million genetic variants across the sample set.

A central tenet of early GWAS studies was ’common disease rare variant’. With the release of the 1000 genomes reference panels in 2013, greater genetic variation and rare variants could be accurately imputed [240, 252]. These advances led to what become known as the poster boy of GWAS, the 2014 GWAS of 36,989 cases and 113,075 controls by the psychiatric genetics consortium (PGC) [253].

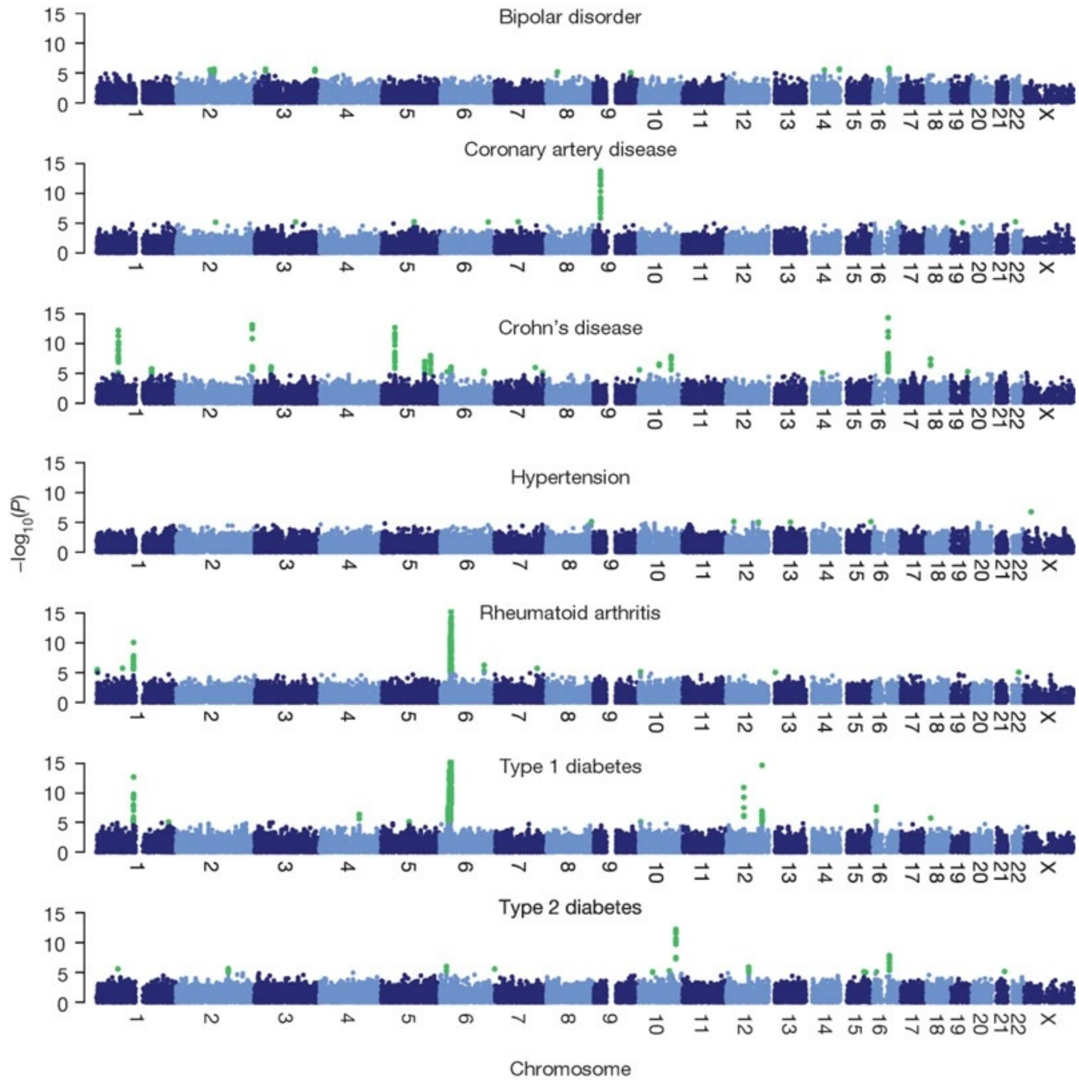

Figure 1.4: Manhattan plots of the first large-scale GWAS performed on seven common diseases. Figure reproduced with permission under the Creative Commons licence (5540960502953) from [251]

As of May 2023, *Yengo et al.* have performed a GWAS on height in 5.4 million individuals [254]. This study is the largest published GWAS and has identified 7,209 independent loci associated with height explaining approximately 40% of the phenotypic variance in European ancestry. Although larger sample sizes are unlikely to contribute new loci to height in Europeans, the work of *Yengo et al.* does highlight an important issue in quantitative genetics, namely the underrepresentation of non-Europeans in datasets. Of the 5.4 million samples, approximately 75% were of European ancestry. Although the effect sizes of the tested variants were consistent across ancestries the significant loci explained only 10-20% of the variation in other ancestries. This has major implications for genomic medicine that use the contribution of genetic effects to stratify patients based on risk and therapy [255].

1.4.3 Design and validation

GWAS typically involve large-scale analysis of genetic data from thousands to millions of individuals to identify genetic variants associated with a particular trait or disease. Study design considerations include defining the phenotype of interest, selecting appropriate controls, ancestry considerations, and determining the sample size through power calculations [256, 257].

Selection of the genotyping platform or methodology is important for obtaining accurate and reliable genetic data. Different genotyping platforms have different strengths and weaknesses, and their selection depends on factors such as the research question, budget, and available resources. For example, WGS will give the most information per sample, but the size and storage cost would not be feasible for datasets in the orders of millions [258]. For this reason, most published GWAS have been performed on microarray data that has been imputed with the latest haplotype reference datasets.

QC measures are applied to genotyping data to ensure data accuracy, reliability, and reproducibility. This may include sample quality assessment, genotype calling, SNP quality control, sample-level QC, batch effect assessment and correction, and imputation and phasing QC [259].

Statistical methods are used to identify genetic variants associated with the phenotype of interest. This may involve various statistical tests such as the chi-square test, logistic regression, linear regression, or more advanced methods like linear mixed models (LMMs) or machine learning algorithms [260, 249, 261, 262, 247, 263]. The most commonly used LMMs are detailed in full below. Multiple testing correction methods are applied to control for false positives but in practice, the genome-wide significance p-value threshold of 5×10^{-8} is often used [264].

Replication and validation of GWAS findings in independent datasets or populations are crucial to confirm the robustness and generalizability of the results. Replication studies involve repeating the GWAS analysis in a separate dataset, while validation studies test the replicated findings in additional datasets or populations [256, 257]. Functional annotation of the identified genetic variants can help elucidate their potential biological mechanisms and provide insights into the underlying biology of the phenotype. Comprehensive browser-based pipelines have been developed for downstream analyses to investigate the functional relevance of the identified variants [265].

Meta-analysis is often performed to combine GWAS summary statistics from multiple studies to increase statistical power and identify additional genetic variants associated with the phenotype of interest. Meta-analysis involves combining summary statistics or individual-level data from multiple studies, and appropriate meta-analysis methods are applied to account for heterogeneity and other sources of variability across studies [266]. For example, *Yengo et al.* performed a meta-analysis on height using 281 studies from the GIANT consortium and 23andme to achieve a sample size of 5.4 million individuals [254].

1.4.4 Linear mixed models

Linear mixed models (LMMs) are the most frequently applied statistical model in GWAS [267, 268, 269, 270, 271, 272, 273, 261, 246]. A key strength of LMMs over traditional linear modelling is their ability to account for population structure, cryptic relatedness, and other sources of genetic and environmental variation. LMMs are an extension of the traditional linear regression models used in GWAS, but in addition, they incorporate random effects to account for correlations among individuals due to genetic relatedness, via a covariance matrix proportional to the kinship matrix [274]. LMMs are particularly useful in GWAS because they can effectively control for false positive associations and increase statistical power by properly accounting for these sources of variation, which can lead to spurious associations and inflated type I error rates [275]. The specific form of the LMM will depend on the software or statistical model used, but it generally includes binary or continuous trait values as the dependent variable, the genotype data as the independent variable, a kinship matrix as the random effect and fixed-effect covariates (e.g., age, sex, PCs) to account for other sources of variation [268, 269, 270, 271, 272, 273, 261, 246, 276]. The resultant p-value and effect size are used to evaluate the strength of the association between the tested SNP and trait of interest.

Calculating the variance components of the random effect contributes a significant burden to the computational complexity of the LMM as the number of rows and columns in the kinship matrix is equal to the number of samples. This quickly becomes intractable for GWAS sample sizes approaching millions of individuals and millions of SNPs. Early GWAS models had a running time of $\mathscr{O}(MN^2)$ or $\mathscr{O}(M^2N)$, with M the number of genotypes used to estimate genetic kinship coefficients [261]. Major reductions in memory and running time have been achieved by using LD structure to subset the number of SNPs used in the random effect and efficient spectral decompositions during Monte Carlo REML variance component estimation [272].

BOLT-LMM advanced GWAS methodologies in terms of power and efficient runtime by incorporating some key algorithmic and Bayesian statistical approaches to achieve a runtime $\mathscr{O}(MN^{1.5})$. As the spectral decomposition of the random effect is cubic in nature, BOLT-LMM uses a conjugate gradient iterative method to reduce the runtime of the variance components of the LMM to linear with respect to N and M. The memory footprint of BOLT-LMM is also dramatically reduced as the random effect matrix is only computed in factorised form, meaning that the N by N random effect matrix only ever exists in M by N vectorised form [261]. The power of BOLT-LMM over traditional methodologies is due to two features. The infinitesimal model of polygenicity is relaxed and modelled as a mixture of Gaussian effect sizes allowing for large effect loci to be modelled as well as modelling the background genome-wide effects due to population structure. BOLT-LMM also adopts a leave-one-chromosome-out (LOCO) scheme to model the polygenic background and avoid proximal contamination [277]. Incorporating the polygenic background in the SNP effect size estimation has been developed further in the REGENIE GWAS model and in PGS-LMM, which provides a flexible framework for incorporating LMMs and the polygenic background [247, 278].

Although BOLT-LMM remains the gold standard in terms of accuracy, more efficient methods with comparable accuracy have been developed [246]. FastGWA uses a grid-search-based REML algorithm to estimate the variance components on a sparse kinship matrix, this advancement reduces the overall runtime to $\mathscr{O}(MN)$. In a comparison on 400,000 individuals in the UK Biobank, the runtime of fastGWA was approximately 1.1% (20 minutes) that of BOLT-LMM and required approximately 5 GB of memory compared to 55 GB for BOLT-LMM [246].

Early GWAS studies focused on the strength of the association of a particular locus to a trait. However, as the field of GWAS moves towards clinical utility, the magnitude of the effect of each locus has become increasingly important. Accounting for external sources of phenotypic variation can often lead to a reduction in the standard error, resulting in greater precision in the effect size estimate. By accounting for the polygenic background, the effect sizes can be more accurately estimated, allowing for greater precision in estimating the genetic risk [278].

1.4.5 Phenotype prediction

Polygenic scores (PGS) are a quantitative measure of an individual's genetic risk for a particular trait or disease, based on the cumulative effect of multiple genetic variants across the genome. The simplest form of PGS is the weighted sum of the genotype (encoded 0, 1 & 2) and effect size of the SNPs estimated from GWAS. PGS developed out of the field of animal and plant breeding, where an estimated breeding value (EBV) per animal or crop was derived using phenotypic data from the individual and phenotypes of its relatives [279]. The first use of genotypic markers dates back to the end of the 20th century [280, 281]. A key distinction between EBV and PGS is that EBV aims to provide a group mean, i.e a trait prediction in the offspring, while PGS aims to predict risk within the sample. This disparity leads to a large difference in the predicted values as prediction accuracy within a sample is low compared to prediction accuracy within a group. The heritability of the trait in question bounds the PGS accuracy for sufficiently large sample sizes, meaning that the utility of PGS in diseases with little to no heritability will be marginal. While not applicable to all traits, PGS have several clinical use cases, such as predicting disease risk [280], the potential stratification of individuals in clinical trials [282] and the realisation of personalised medicine [283, 284].

Type II diabetes (T2D) [285] and cardiovascular disease (CAD) [285] are the most studied human health traits using PGS. For example, In CAD, 8% of the population tested had a greater than 3-fold increased risk of CAD. For individuals with a T2D diagnosis, a 1 standard deviation decrease in the PGS is statistically associated with a reduction of 1.3 years in the age of onset [285]. PGS in several psychiatric phenotypes, such as schizophrenia [286], bipolar [287] and depression [288], has also shown clinical utility.

1.4.6 PGS construction

In its simplest form, a PGS can be constructed by summing over the product of the genotype dosage and effect size. There are some caveats to this method, firstly an assumption is that the SNPs are independent. To account for the LD structure in the genome, independent tagging SNPs are used. The addition of SNPs in LD with the tagging SNP will inflate the PGS. The PRSice framework uses the clumping and thresholding (C+T) algorithm to greedily remove 'LD friends' [289]. For a given genome window, C+T identifies the SNP with the strongest association to the trait in question based on a p-value and removes SNPs in LD above a user-defined R^2 threshold. This removal of SNPs within a window is problematic as the SNPs in LD with the tagging SNP may contribute to the variation in the phenotype, thus the derived PGS may be underpowered. If the genotypic data are available from the dataset in which the effect sizes are estimated, methods such as best linear unbiased predictors (BLUPs) can be used to avoid the issue of LD structure. This in practice is not the case as summary data only is usually used, derived from genotypic data from hundreds of thousands of individuals.

Recent work on PGS has used Bayesian modelling to condition the posterior mean effects from summary data. These methods model the estimated effect sizes conditioning on the heritability (estimated from the summary data) and genomic structure (LD maps derived from 1000 genomes). The LDpred model uses two parameters to infer a PGS score, heritability and the fraction of causal SNPs [290]. A major drawback of this approach was that the optimisation of the fraction of causal SNPs is tuned in a validation dataset. To avoid subsetting the dataset and improve computational efficiency, LDpred2 has implemented a grid search which estimates the fraction of causal markers from the data directly [291]. Other methodologies that implement Bayesian multiple regression such as the summary Bayesian alphabet (SBayesR) have also been developed [292].

1.5 UK Biobank

1.5.1 Background

The UK Biobank is a large-scale, long-term research project that aims to investigate the complex interplay between genetic and environmental factors in human health and disease [293]. It is a unique resource that provides data and samples from half a million individuals across multiple ancestries in the United Kingdom, making it one of the largest and most detailed biobanks in the world. The project was conceived as a means to facilitate large-scale genetic research by creating a comprehensive database of health-related information and biological samples from a large cohort of individuals. The goal was to understand the causes and risk factors for various diseases, including cancer, heart disease, diabetes, and dementia, and to promote the development of new diagnostic tools and treatments. The UK Biobank has enabled numerous ground-breaking research studies in diverse fields, including genetics, epidemiology, and public health [293, 294]. To date over 6,000 papers have been published using the UK Biobank data [295].

The UK Biobank project received funding from various sources, including the Wellcome Trust, the Medical Research Council, the Department of Health, and the Scottish Government. Numerous research institutions, universities, and healthcare organizations across the United Kingdom also supported the project. In 2002, the UK Biobank received initial funding of 62 million pounds sterling from the Wellcome Trust, and subsequent funding has been provided through a combination of public and private sources [296, 297].

The recruitment of participants for the UK Biobank began in 2006 and lasted until 2010 [293]. A total of 502,505 individuals between the ages of 40 and 69 years were enrolled from across the United Kingdom, capturing genetic variation and health data from a diverse set of ethnic backgrounds. Participants underwent a comprehensive baseline assessment, which included detailed questionnaires about their health, lifestyle, and medical history, as well as physical measurements such as height, weight, blood pressure, and lung function. Participants also provided blood, urine, and saliva samples for long-term storage and future analysis.

The UK Biobank continues to evolve and expand its research efforts. In addition to the extensive data and samples already collected, the UK Biobank plans to continue following up with participants over time to collect additional health-related information and imaging samples. The project also aims to enhance its data holdings by integrating genetic, environmental, and clinical data from other sources, such as the cancer and death registries and air pollution data modelled at the participant home addresses using the ESCAPE consortium data [298]. The UK Biobank remains a valuable resource for researchers and is expected to contribute significantly to our understanding of human health and disease in the future. The UK Biobank has also partnered with AWS and DNAnexus to provide a cloud computing infrastructure to circumvent the costly process of downloading and storing 100Tbs of data per application on local servers.

1.5.2 Data protection and ethics

The UK Biobank places a strong emphasis on ethical considerations and has obtained informed consent from all participants. Participants are provided with detailed information about the project and have the right to withdraw their participation at any time. The UK Biobank follows strict data protection protocols and ensures that data is anonymized and securely stored to protect participants' privacy. Access to the data and samples is provided to approved researchers through a data access committee (DAC). Since its public release the UK Biobank has granted access to over 30,000 researchers across the world [295]

1.5.3 Data types

In addition to biomarkers and baseline questionnaires, the UK Biobank has also conducted imaging studies on a subset of participants, including brain imaging (MRI), bone imaging (DXA), and retinal imaging. These imaging data provide additional information on participants' health, including brain structure and function, bone health, and eye health. The UK Biobank has collected extensive environmental data, including geographic and socioeconomic data, air pollution data, and neighbourhood characteristics. The UK Biobank has generated extensive genotyping and sequencing of participants' DNA to generate genetic data. The genetic data from the UK Biobank allows researchers to investigate the role of genetics in human health and disease, including identifying genetic risk factors, studying gene-environment interactions, and conducting polygenic risk score analyses.

1.5.4 Exome sequencing

The whole exome sequencing (WES) component of the UK Biobank has had a significant impact on advancing our understanding of genetics and its role in human health and disease. WES data from the UK Biobank has enabled the identification of rare genetic variants that may have a significant impact on disease risk or other health-related outcomes. To fully understand their role in human health large sample sizes are required to capture rare variants. Nonetheless, rare variants can have large effect sizes and potentially provide important insights into basic biological pathways and the genetic basis of diseases that are difficult to study using common variation [299]. Identifying rare genetic variants using WES data has helped uncover new genetic associations with diseases and traits, providing valuable information for understanding disease mechanisms, developing diagnostic tools, and identifying potential therapeutic targets [300, 301].

WES data has also been used to generate PGS which are calculated based on the combined effects of multiple genetic variants associated with a particular trait or disease. The WES data from the UK Biobank has been used to develop and validate PGS for various health outcomes, including cardiovascular diseases, cancer, neurodegenerative diseases, and mental health disorders, among others [302, 303].WES has enabled researchers to study gene-environment interactions, which are the complex interplay between genetic factors and environmental exposures in influencing health outcomes. By combining genetic data from WES with environmental data available in the UK Biobank, researchers have been able to investigate how rare genetic variants that are not captured by microarray genotyping and imputation may modify the effects of environmental factors on health outcomes, such as the interaction between genetic variants related to smoking and the risk of lung cancer, or the interaction between genetic variants related to BMI and obesity [304].

WES data also captures information on somatic mutations [305]. Although attempts to call somatic mutations have usually exploited the fact that WES data is derived from whole blood samples meaning CHIP genes can be exploited to identify variants which show a linear increase with age. This excludes the majority of samples with no detectable CHIP [305]. What remains unclear is if other mutational scores can be identified in the exome-seq data.

1.6 Thesis outline

The central aim of this thesis was to infer signals of somatic mutation within large sequencing datasets. Specifically, we asked whether given sufficient sample numbers it was possible to distinguish somatic mutations from the noise contained in the 200,000-sample release of the UK Biobank whole exome sequencing dataset. Given that there was indeed a signal of somatic mutation recoverable from the sequencing data, we then asked what we could learn from this signal. Chapters 2-4 specifically address this research question. In our final research question, described in Chapter 5, we asked whether it was possible to improve the power and computational efficiency of genome-wide association studies by incorporating polygenic scoring methodologies into a linear mixed model framework.

In Chapter 2, we define a computationally efficient pipeline for calling mismatches to the reference genome (excluding germline variants) in over 200,000 WES samples. By aggregating the mismatches into mismatch loads we can test the central aim of this thesis. We also developed a normalisation method to address sources of technical variation that can give rise to differing profiles of technical noise across samples. The expected number of mutations can be calculated for a given cell type from empirical estimates of the somatic mutation rate, allowing an estimate of the total somatic mutational load within a sample to be obtained for the WES data. We then validate the plausibility of the central aim of the thesis through simulations and an analysis of the relationship between mismatch loads and sample age.

In the third chapter, we set out to investigate sources of inter-individual variation in mismatch loads across the dataset. We tested for differences in the mismatch load associated with cancer and smoking status. We partitioned the mismatch loads on whether the mismatch occurred in a gene that was expressed in whole blood in GTEx allowing the recovery of signals of transcription-associated asymmetry across samples. The contribution of genetic variation influencing the somatic mutation rate remains largely unknown outside of monogenic rare diseases. We asked whether we could detect signals of genetic variation contributing to the mismatch load variation. Building on the intuition that genetic variation can influence somatic mutation rates, the final question in Chapter 3 aimed to address whether there was a significant difference in the contribution of mismatch repair mutational signatures between samples containing Lynch syndrome variants and the remaining samples.

Chapter 4 pivoted to an analysis of variation in mismatch loads along the genome, in order to mitigate the effects of substantial sources of technical variation that impeded the analysis of variation across samples. Using the gene-level data, we tested the association with known mutation rate modifiers such as GC content, replication timing, recombination hotspots, chromatin accessibility, and gene expression. Gene expression has a complex relationship with mutation rate [148]. To further explore this, we fitted non-linear models to describe the mismatch load as a function of gene expression. In Chapter 4, we also asked whether we could detect signals of positive selection acting on the mismatch loads and investigated the potential of this approach to reveal genes involved in clonal haematopoiesis of indeterminate potential (which affects upwards of 6% of UK Biobank samples) [306].

Our final research question is addressed in Chapter 5. In recent years, several sophisticated statistical models for GWAS have been developed. A major drawback of these models, is their complexity and the substantial computational resources and time that are required to fit them. We asked whether we could use state-of-the-art software for polygenic trait prediction to, one, improve the power of GWAS and, two, improve the computational efficiency of a GWAS pipeline. We developed and published a framework that addresses this question [278].

Chapter 2

Methodology and Validation Study

2.1 Abstract

In addition to genetic variation, NGS data contains information on somatic mutation. The rate at which somatic mutation arises is often much lower than the background error rates making the high-confidence calling of somatic mutations difficult at the sequencing depths typically used to identify germline variation. Here we developed a pipeline to recover mismatches to the reference in NGS data and filter known and likely germline and technical artefacts. Using simulated data and age, we can infer and validate information about somatic mutation in the UK Biobank using the median recurrence of mismatches to the reference genome across samples in the UK Biobank. We estimate that 0.4 - 1% of mismatches we identify are probable somatic mutations. This allows us to use NGS data experimentally designed to analyse germline variation in investigating processes associated with somatic mutation.

2.2 Introduction

Exome sequencing is routinely used to identify germline variants in the coding regions of the genome. By restricting to the coding regions of the genome, the variants identified are enriched for functional effects [299]. While the cost per exome sequenced is higher than that of genotyping variants using microarrays, exome sequencing allows for de novo variant discovery while microarrays require a priori variant knowledge. Exome sequencing relies on an enrichment library preparation step in which oligonucleotide probes are used to pull down the exonic targets, omitting the non-coding portion of

DNA.

A heterozygous germline variant sequenced at sufficient sequencing depth will be captured on approximately 50% of sequencing reads. This variant proportion is called the variant allele fraction (VAF). Somatic mutations typically occur at low VAFs and are difficult to distinguish in sequencing data from sequencing errors and other artefacts. DNA sequencing resembles the random sampling of segments of DNA from a population of cells. Suppose there are too few cells containing a somatic mutation. In that case, it is unlikely to be captured in the NGS data or will be indistinguishable from the background error. If the population of cells has undergone a clonal expansion, then the number of reads supporting the somatic variant may approach a VAF of 50%, making differentiation from germline events difficult. To circumvent this, cancer sequencing studies rely on a tumour-normal aired protocol to remove sites called in the normal tissue as germline, leaving only high-accuracy somatic mutation calls in the cancer sample.

Errors in sequencing data arise from multiple sources such as sequencing errors such as base position on the read, errors introduced in DNA amplification via PCR, DNA damage accrued during library preparation and erroneously mapped reads [307]. Given sufficient sequencing depth, the recurrence of a given mismatch to the reference genome within a sample is the basis for the identification of somatic mutations. A key shortcoming of this strategy is that it can only be used to identify somatic mutations with a VAF greater than the sequencing error rate. QC metrics calculated during base calling and read alignment can control the false positive rate, but these may miss changes in the DNA sequence that arise during the sample preparation steps.

The role of somatic mutation within healthy tissues and its implication in disease and ageing remains poorly understood. Recent efforts to characterise somatic mutation in sun-exposed skin, oesophageal tissue and colon crypts have given an unprecedented insight into the overlap between the background somatic mutational processes within healthy tissues and cancer [161, 160, 93]. Although the link between somatic mutation accumulation and ageing and cancer is well established [167, 40], estimating the total number of somatic mutations per cell in different tissue types remains challenging owing to the heterogeneous rate of background mutational processes across cell types [52, 53, 54]. Recent work in single-cell sequencing of B cell lymphocytes across the human lifespan has provided an empirical estimate of the somatic load per cell [307]. As

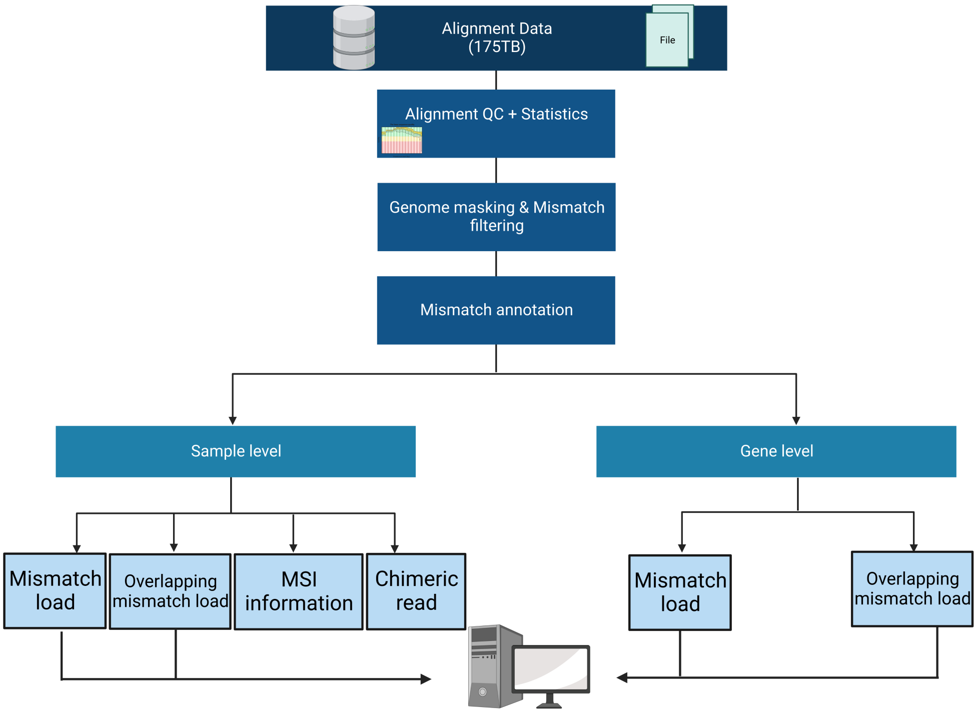

Figure 2.1: Analysis pipeline overview. Schematic illustration highlighting key pipeline processes from alignment data storage through data QC and mismatch load generation to downstream applications.

lymphocytes constitute the majority of nucleated cells in whole blood, this provides a basis to estimate the total number of somatic mutations in an exome sequencing dataset derived from whole blood samples, such as those in the UK Biobank 200k release.

2.3 Pipeline

2.3.1 Pipeline overview

We developed a pipeline to extract mismatches to the reference genome from sequencing data (Fig. 2.1). Whole exome sequencing data generated from 200,632 samples of the UK Biobank was downloaded in CRAM format and stored on a local high-performance BeeGFS filesystem as part of the bioinformatics high-performance compute (HPC) infrastructure at the University of Galway over one month. The UK Biobank 200k release consisted of 175TB of alignment files. The average sample

CRAM file was 0.75 GB. All analyses were performed on an HPC server with 264 Intel(R) Xeon(R) CPU D-1541 physical cores and 2016 GB of random-access memory (RAM).

Although this work deals explicitly with the investigation and recovery of single-base substitution somatic mutation signals, data on other biological phenomena were also generated as part of the pipeline. This included counting the number of interchromosomal chimeric reads per sample, extraction of reads that span known microsatellite instability loci and alignment statistics using the Samtools stats functionality.

We investigated two different single-base mismatches datasets. All single-base mismatches to the reference genome were called using Samtools (v1.12) mpileup [308], using the filtering parameters outlined in Table 2.1. The first dataset was obtained by restricting to mismatches that occurred on the overlapping portions of read pairs, i.e., sites sequenced by both read one and read two in which there is agreement on a mismatch to the reference genome. The restriction to mismatches consistently identified on both reads was performed to reduce the proportion of mismatches due to sequencing error (we refer to data generated using this approach as overlapping read data). In theory, the probability of a sequence call error for these overlapping bases is the product of the individual call error probabilities (which can be calculated from the Phred scores). In the non-overlapping restricted mismatch calls, the default behaviour of the Samtools software stack was used to handle the overlapping read portions, i.e., the sum of the two base qualities was used if both reads agreed on the call and 80% of the highest quality was used if the reads disagreed on the call. To extract mismatches in the overlapping data, the base quality on read one and read two are treated independently during filtering.

A stringent filtering pipeline was devised to remove germline contamination and technical artefacts. Sites called germline SNPs in the UKB 200k release were excluded, and SNPs contained in the dbSNP v155 (Jun 2021) release were excluded. Mismatches were confined to regions included in the exome capture target bed file, and sites that overlapped repeat regions and regions where a 75bp kmer had a probability of mapping more than twice (Gem > 0.5) were excluded from downstream analyses. The Encode blocklist filters sites in the problematic areas of the genome. The genome masks from *Abascal et al.* (lifted over from hg19 to GRCh38) were also used to remove loci that had higher than average error rates or were prone to mapping artefacts [309]. The

Table 2.1: Samtools (HTSLIB) read and base filtering parameters.

Filter	Value
Sequencing depth	20
baseQ	37
MapQ	60
Cigar	No gaps; exact match (75M)
Supplementary reads	removed
Secondary reads	removed
Duplicate	removed
QCFail	removed
Unmapped reads	removed

mismatches for each sample were stored as VCF-style files containing the number of reads at the locus, the variant allele fraction, whether the mismatch has been called as a SNP and whether the mismatch was included in the overlapping mismatch set.

To add information to each site containing a mismatch, the data was annotated using Ensembl's variant effect predictor (VEP). As the total number of mismatches was in the order of 35 billion we developed an efficient file format to aggregate and store the mismatch data prior to annotation. The gene name, codon change, transcription strand, amino acid change, up and downstream nucleotides and transcriptionally aligned mismatch contexts were added to the summarised data. The sequencing batch ID and flowcell lane ID were retrieved from the alignment files. Over 1500 phenotypes were downloaded from the UK Biobank data repository. This set included age at the time of assessment, assessment centre, genotyping batch, cancer status and smoking status. The imputed genetic data for the entire 500,000 individuals were downloaded and used as the basis of the genetic variation studies in Chapters 3 and 5.

The mismatches were summarised on two levels; firstly, we analysed the mismatches on the sample level, i.e., each sample has a value corresponding to a mismatch load (analysed in Chapters 2 and 3). This dataset was intended to be used to test whether there was a detectable contribution of somatic mutations to the mismatches observed in the UKB exome sequencing dataset (by searching for a correlation between mismatch load and age) and to investigate sources of inter-individual variation in the mismatch (and putative somatic mutation) load (see Chapter 3). Secondly, the mismatches were summarised on the gene level. This dataset enabled us to assess the correlation between mismatch load and known mutation rate modifiers such as gene expression and replication timing, as well as the identification of genes showing evidence of somatic selection in the form of a higher-than-expected proportion of functional mutations (see Chapter 4).

2.3.2 Data generated from pipeline

The total runtime to download and run the pipeline was approximately 12 weeks, spread over 16 weeks, to allow fair usage of the bioinformatics HPC system. The mean base quality across the sample set was 36.34 (Table 2.2 & Fig. 2.2 A). On average, 63,899,754 reads per sample were mapped to the GRCh38 reference genome, with 44,487,029 reads remaining after alignment QC (Table 2.2 & Fig. 2.2 B). The average number of bases sequenced was 3,381,014,204, while the mean number of mismatches per sample was 628,346. Applying the filtering protocol to remove germline and problematic loci reduced this to 173,327 mismatches per sample (Table 2.2 & Fig. 2.2 C & D). The total number of mismatches across all samples after filtering was 34,748,848,164. Restricting to the overlapping portions of read pairs resulted in an average of 534,069,937 sequenced bases per sample and 11,688,422,475 mismatches. This corresponded to 58,299 mismatches per sample, on average, reducing to 10,131 after the filtering protocol (Table 2.2 & Fig. 2.2 E & F). Approximately 30% of reads were filtered due to mapping quality filters or complex CIGAR entries. The mismatches are stored in single-sample VCF format. To reduce the computational complexity of annotating each site across all samples with VEP, samples were collapsed into a single file containing variant and sample information.

2.3.3 Mismatch load derivation

As the data was derived from exome sequencing, we determined the mismatch load with respect to the transcribed strand. Variants within genes transcribed on the non-reference strand were reverse-transcribed to align with the coding strand. To exclude complex common germline events, the disproportionate effect of germline contamination is highlighted in Chapter 3, a conservative threshold of variants found within 1000 samples were removed (Fig. 2.3). This filter removed approximately 60% of the remaining mismatches. Substantial variation in the number of mismatches per sample

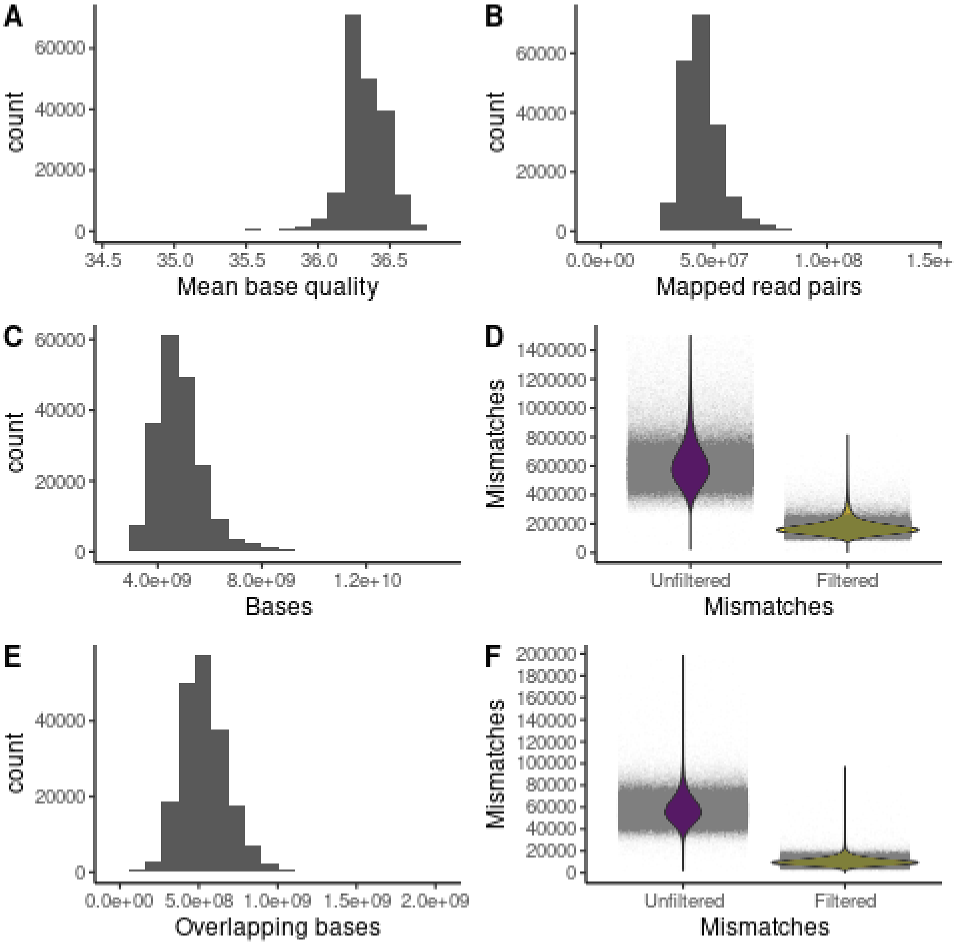

Figure 2.2: Histograms illustrating variation in properties of the mismatch data across samples are shown in panels A-D. A) The average base quality across samples. B) The total number of mapped read-pairs per sample. C) The total number of sequenced bases per sample. D) The total number of mismatching bases before and after filtering. E) The number of sequenced bases within the overlapping portions of read-pairs. F) Total number of mismatching bases within overlapping portions of read-pairs.

Table 2.2: Summary of sample level data and statistics derived from pipeline.

	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Mismatch	2.49E+04	5.13E+05	6.01E+05	6.29E+05	7.05E+05	3.16E+06
Mismatch filtered	6.34E+03	1.41E+05	1.65E+05	1.73E+05	1.93E+05	8.11E+05
Ovp. mismatch	2.06E+03	4.94E+04	5.72E+04	5.83E+04	6.58E+04	3.55E+05
Ovp. mismatch filtered	3.16E+02	7.78E+03	9.73E+03	1.01E+04	1.20E+04	9.68E+04
Sequenced bases	2.60E+09	4.21E+09	4.72E+09	4.84E+09	5.31E+09	1.47E+10
Sequenced ovp. bases	6.21E+05	4.32E+08	5.19E+08	5.34E+08	6.20E+08	2.00E+09
Mean base quality	3.46E+01	3.63E+01	3.64E+01	3.63E+01	3.65E+01	3.68E+01
Total read pairs	3.43E+07	5.56E+07	6.22E+07	6.39E+07	7.00E+07	1.94E+08
Mapped read pairs	1.52E+06	3.87E+07	4.33E+07	4.45E+07	4.87E+07	1.41E+08
Mean cigar error	1.19E-03	1.64E-03	1.81E-03	1.92E-03	2.03E-03	8.73E-03
Tri-allelic	8.60E+01	1.96E+03	2.62E+03	3.05E+03	3.48E+03	3.71E+04
Chimeric reads	1.82E+03	1.10E+05	1.27E+05	1.31E+05	1.47E+05	8.84E+05
Age	3.80E+01	5.00E+01	5.80E+01	5.65E+01	6.30E+01	7.20E+01
Mismatch rate	1.20E-06	3.11E-05	3.46E-05	3.58E-05	3.83E-05	1.18E-04
Ovp mismatch rate	5.80E-07	1.59E-05	1.83E-05	1.97E-05	2.15E-05	1.75E-02
Est. Somatic load	8.77E+02	1.07E+03	1.23E+03	1.21E+03	1.34E+03	1.56E+03

was found across sequencing runs, evidenced by the difference in slopes in the number of mismatches as a function of the number of sequenced bases per batch (Fig. 2.4). As the sequencing batch IDs were unavailable from the UK Biobank repository, the flow cell ID contained within the read name tag of the alignment data was used to group samples by sequencing run. On average, 245 samples were sequenced across 817 flow cells. The derived mismatch loads were approximately normally distributed (Fig. 2.5 A & B). To normalize the mismatch phenotype contexts across sequencing runs, the total number of mismatches in each mutation context and flow cell run was regressed against the total number of autosomal reads within a sample. The rank of residuals of this model was then divided by the number of samples in the flow cell batch, resulting in a fractional rank bounded by 0 and 1.

Interestingly, batch effects were much less evident in the mismatches in the overlapping regions (Fig. 2.4 B). The lack of batch structure within the overlapping read mismatch data may be due to differences between batches resulting from oxidative damage during the library preparation stage. Oxidative damage occurs at greater frequencies around DNA breakpoints. As the overlapping portion of paired-end reads tends to be within the centre of the DNA fragment, oxidative damage is likely to have made a lower contribution to the mismatches in these regions.

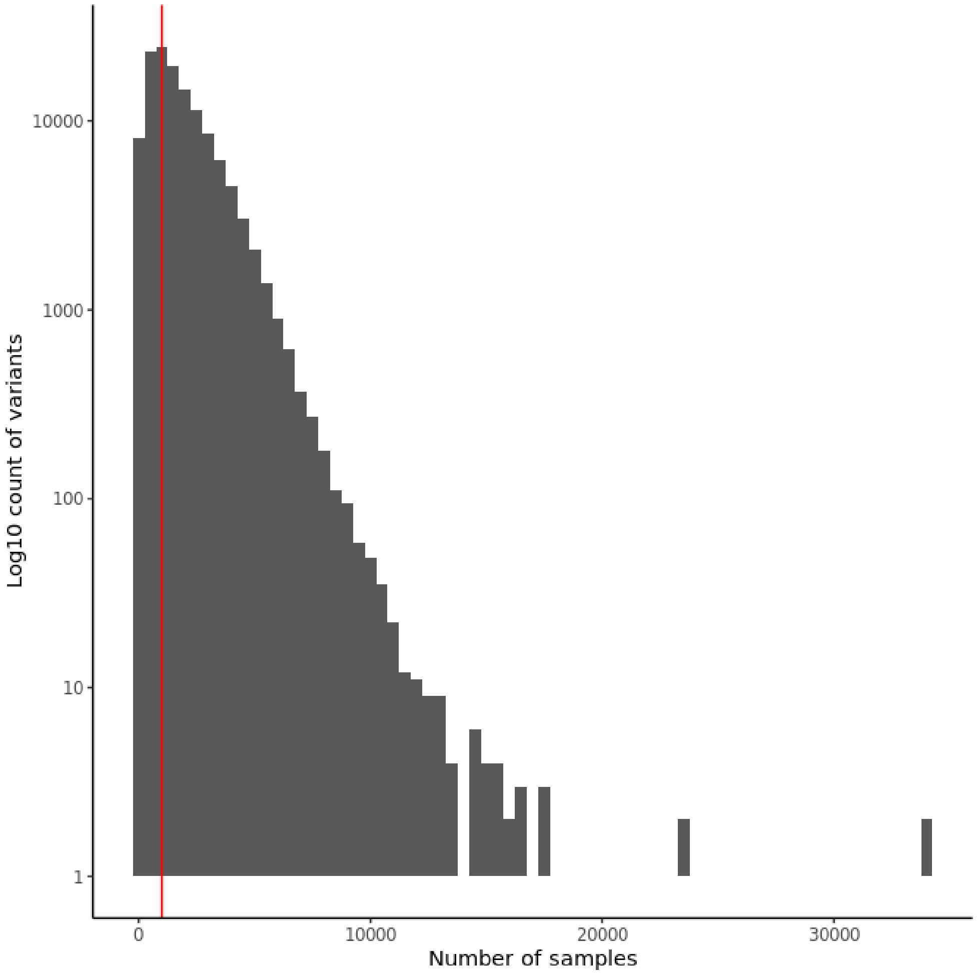

Figure 2.3: Histogram of log 10 counts of mismatches by recurrence within each sample. The red line indicates the stringent filtering threshold for germline variant exclusion.

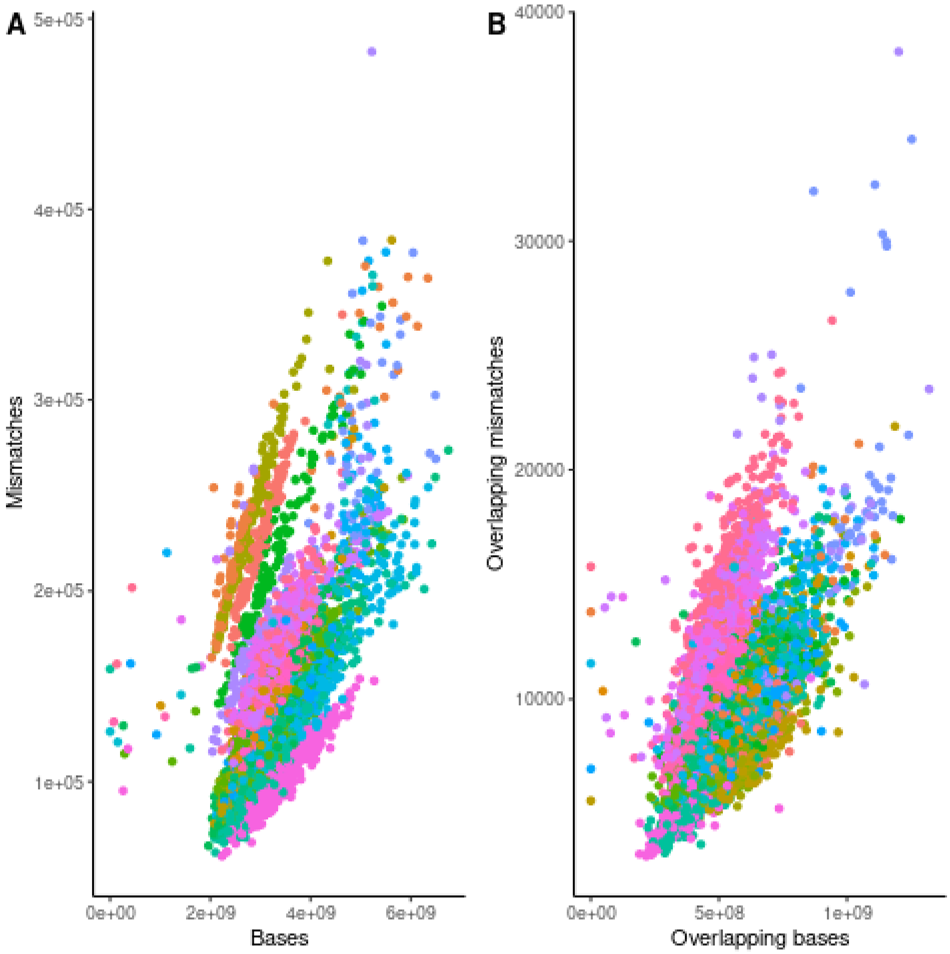

Figure 2.4: Number of mismatches as a function of sequenced bases for 25 sequencing batches.

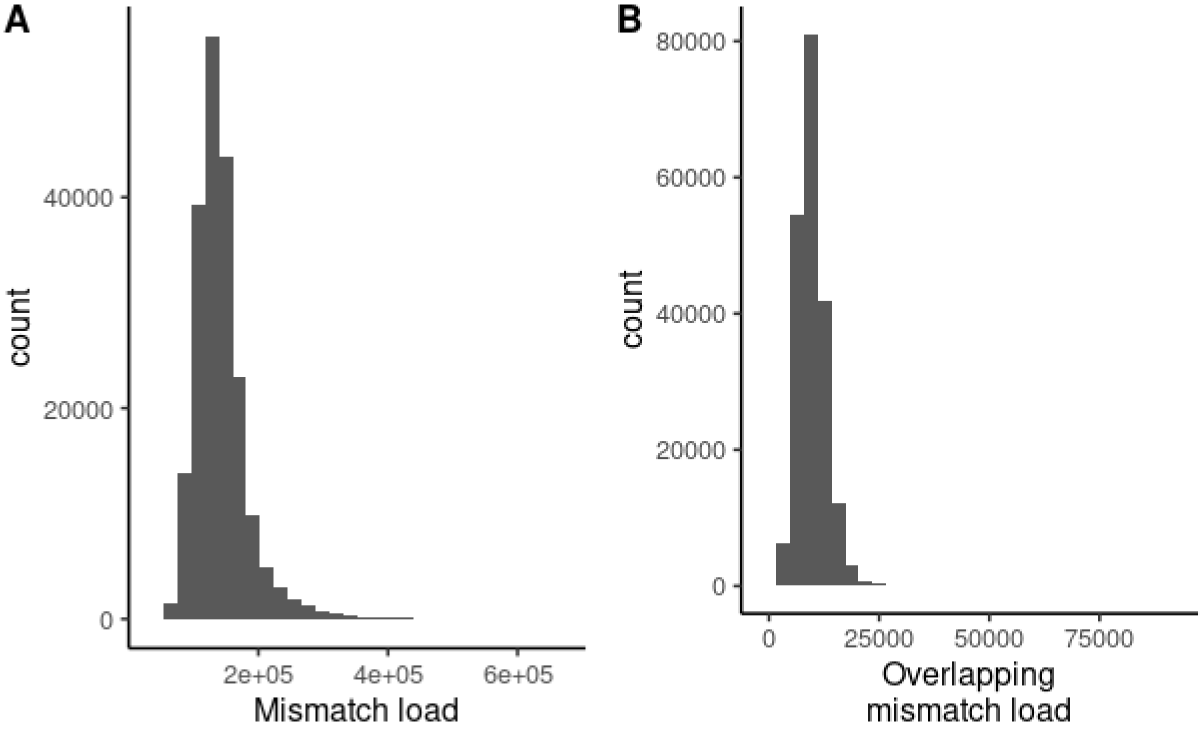

Figure 2.5: Distribution of mismatch loads per sample. A) Histogram of the mismatch load derived from single-read data B) Histogram plot of the mismatch load in the overlapping read data.

2.3.4 Data Validation

Using the equation empirically derived in Fig.1 of *Zhang et al.* [152], the average number of sequenced bases and 56 as the average age of UK Biobank participants, we can infer an average somatic load of 2101 substitutions per cell. Using the average number of mismatches and sequenced bases per sample, we estimate that approximately 0.4% of mismatches are true somatic mutations, (Eq. 2.1). About 1% of overlapping mismatches are likely somatic mutations for the overlapping read data. Simulated mismatch loads with a known proportion of somatic mutations were generated for 200,000 samples using the formula derived from *Zhang et al.* [152]. We sampled 200,000 ages from the UK Biobank to estimate an expected count of somatic mutations. We can then add noise matching the distribution of mismatches in the UK Biobank data by shuffling the actual mismatch counts in the UK Biobank data and adding the estimated somatic load per cell. We regressed the median number of mismatches against age (Fig. 2.6 A, B, D & E). The model fit for the simulated data closely matches the observed model fit after accounting for the sequencing batch. In the overlapping mismatch load simulations, we found that the variance explained across the median mismatch loads was less than that of the single-read mismatch calls (Fig. 2.6 C & F).

$$\frac{Somatic\,Mutations\,per\,bp\,\times\,N\,sequenced\,bases}{Mean\,M\,mismatches} \tag{2.1}$$

2.4 Discussion

In this chapter, we have developed a pipeline for the recovery of signals of somatic mutation in noisy exome sequencing data. We developed a stringent filtering process to remove germline variants, blocklisted genomic regions and technical artefacts. The filtering process removed approximately 70% of the UK Biobank 200K release mismatches. We uncovered significant unreported sequencing batch effects that contribute to variability in the number of mismatches per sequenced base across samples. Using information within the sequence read name tag, we grouped samples by flow cell to remove the variation contributed by sequencing batch effects. The batch effects may be due to several different factors. Different temperatures, distances travelled to the sequencing centre, and different technicians handling each sample can introduce sys-

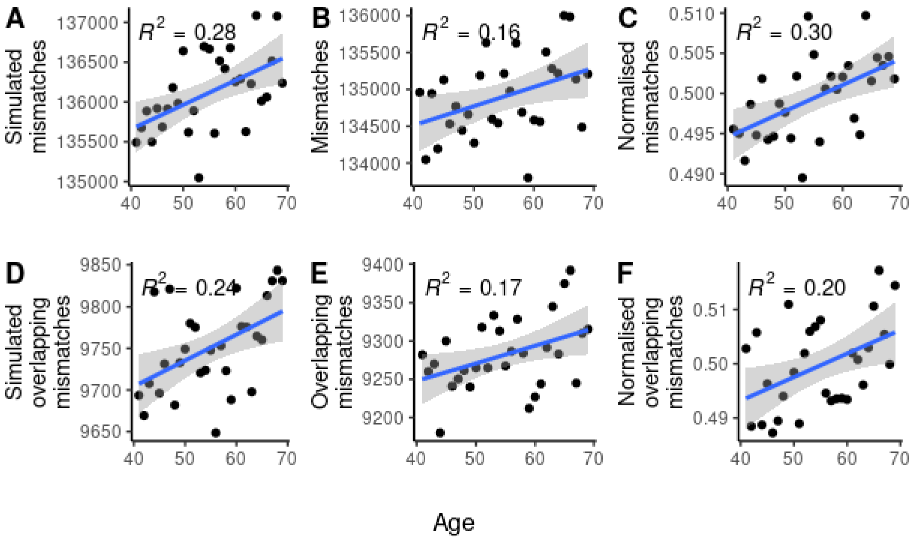

Figure 2.6: Median count of mismatches per integer age. A) Simulated data for 200,000 samples, B) Observed median mismatch load. C) Observed median mismatch load after normalizing by batch. D) Simulated overlapping data for 200,000 samples. E) Observed median mismatch load in overlapping regions. F) Observed median mismatch load in overlapping regions after normalizing for batch structure.

tematic differences between batches. Restricting to mismatches on overlapping reads reduced the differences between batches, potentially reflecting a lower rate of oxidative DNA damage in the interior portions of the DNA fragments. This observation is consistent with findings from *Chan et al.* where higher rates of DNA damage have been reported at the start and end of the read fragments [307].

To test whether the data generated from our pipeline could plausibly capture somatic mutation, we tested for a correlation between median mismatch load and participant age. As somatic mutations accumulate with age, a positive linear relationship is expected between age and the median somatic mutation load. A correlation was detected, the strength of which was consistent with what we expected, given the relatively low proportions of somatic mutations among the mismatches. We can use an expectation of somatic mutation load and the variance from the mismatch data to simulate the relationship between age and the median load of somatic mutation across simulated data

Our simulation framework is consistent with what we observe in the UK Biobank mismatch data. However, we hypothesised that by restricting mismatches sequenced on overlapping reads, we would enrich for a signal of somatic mutation by reducing the contribution of DNA damage-induced mismatches. We found the opposite in the simulated and observed UK Biobank data when testing for a correlation with age. We posit that restricting to overlapping bases has induced a stochastic effect as the number of somatic mutations is small in the overlapping data. As a result, in the overlapping data, the model cannot recover the same strength of signal of somatic mutation as within the single read data. We also note that the model fit of the observed overlapping data doesn't capture the same amount of variation as the simulated data. In data derived from high-confidence somatic mutation calls, the observed number of base substitutions is substantially lower in the centre of exons due to increased transcription-coupled repair efficiency or protection from damage driven by the greater nucleosome occupancy in exons relative to introns [310, 311]. Therefore, when restricting to overlapping reads, we are enriching for regions within exons with a reduced mutation rate, whereas in the simulated data, the mutation rate is constant across exons.

Although we estimate that a relatively small proportion of mismatches (0.4% and 1% in the full and overlapping data, respectively) to be true somatic mutations, by aggregating the recurrence of mismatches we can recover a signal of somatic mutation in the UK Biobank as a function of increasing age. This suggests that the mismatch data could be used to identify other (strong) sources of inter-individual variation in the somatic mutation burden. Given the rich phenotypic and genotypic data associated with the UK Biobank, we can use the derived mismatch loads to test for genetic variation driving somatic mutational processes and their association with phenotypic exposure data.

2.5 Data access acknowledgement

This research has been conducted using the UK Biobank Resource under application number 23739. All other sources of data used in this work have been open-access and referenced where applicable.

Chapter 3

Inter-individual variation in the somatic mutation rate

Declaration: Harrison Anthony identified 160 samples within the UK Biobank that contained pathogenic Lynch syndrome variants using the pathogenic Lynch variants as outlined by *Patel et al.* [312].

3.1 Abstract

The somatic mutation rate varies across individuals and between cell types. Through an analysis of exome sequencing data of 200,000 individuals from the UK Biobank, we investigated sources of variation in somatic mutation rate using the number of single base mismatches to the reference genome as a proxy measure. When we considered the 5' and 3' nucleotide contexts, we identified a stronger correlation between the somatic mutation rate and age than in Chapter 2. We report and detail significant batch effects within the UK Biobank exome sequencing dataset that led to spurious phenotypic associations with cancer and smoking status. We report strong transcription strand-associated asymmetry in the data and show, using gene expression data from GTEx, that it is not due to transcriptional processes but rather it is likely to be due to the use of exome capture probes designed against the coding strand. A novel finding with significant implications for rare variant testing and NGS quality metrics. We next sought to investigate the contribution of genetic variation to the variation in mismatch load across samples by performing GWAS and inferring the activity of mutational signatures. We initially discovered a strong GWAS signal on chromosome 5, close to *ERCC8*, a gene involved in transcription-coupled repair that is the basis of Cockayne syndrome. How-

ever, a careful analysis of the locus revealed that this result was artefactual, highlighting the importance of filtering highly recurrent mismatches. Despite previous reports of an elevated somatic mutation rate in Lynch syndrome patients, an analysis of the mean contribution of mismatch repair signatures in Lynch syndrome patients compared with the remaining samples showed no statistical association. However, we do report for the first time outside of paediatric oncology a significant positive association between the contribution of SBS3 and age.

3.2 Introduction

Somatic mutations are acquired sequence alterations that arise in any somatic tissue after fertilisation [313, 314]. The most common type of somatic mutation is a single base substitutions (SBS) and these may arise through spontaneous processes in response to oxidative stress or through errors in biological processes [167] Studying the mutational burden across samples can give an insight into the intrinsic biological processes that shape the mutational landscape, such as aberrant transcription-coupled nucleotide excision repair, transcription-associated mutagenesis (TC-NER/TAM) and replication-associated errors [123, 315]. In addition to intrinsic biological mutational processes, environmental exposures such as UV light and chemical exposure increase the somatic mutation rate through the generation of DNA lesions which require low-fidelity repair processes to rectify the lesion before cell-cycle progression [56, 316].

Somatic mutations accumulate throughout the genome in the natural ageing process. The positive linear relationship between the number of somatic mutations and age has been reported across a number of tissue types [317]. This relationship has been extensively demonstrated in cancer and, recently, in healthy tissues [93, 160, 209, 318]. Somatic mutation rates vary greatly across cell types and samples [317, 319]. The somatic mutation rate has been estimated using laser capture microdissection to be as low as 2.4 mutations per cell per year in the testis to 56 mutations per year per cell in the appendix [115].

To understand genetic variation in large groups of samples, sequencing is performed on whole blood or peripheral blood samples due to the minimal invasiveness of drawing blood [240, 320]. However, the accurate identification of somatic mutations in a single tissue remains difficult for a number of reasons. Firstly, the clonal nature of blood expansions allows somatic mutations to occur at relatively high frequency that can in some cases be mistaken for germline variants. Secondly, in the absence of high sequencing depths accurately calling somatic mutations without a second reference tissue remains cost prohibitive for a large number of samples. Several experimental methodologies have been developed to accurately identify somatic mutations with high confidence by reducing the sequencing and PCR error rate to below that of the somatic mutation rate, but scaling these methods to understand the variation across large numbers of samples remains a challenging problem owing to difficulties in sample preparation [309, 321, 322].

The heritable component of the mutation rate remains poorly understood for both germline and soma. Genetic variation affecting mutation rates has major implications for the clinical application of genetic risk predictions for cancer and ageing disorders. Genetic variation contributing to a phenotype can be uncovered using genome-wide association studies (GWAS). Currently, GWAS performed on somatic mutation have been restricted to large cancer cohorts [323] or in genes associated with clonal haematopoiesis of indeterminate potential (CHIP). A major caveat of measuring the somatic mutation rate in cancer samples is that the somatic mutation burden across cancer samples may not reflect the true variation in the background mutational processes and may enrich for genetic variation that drives malignancy rather than inherent background mutational processes. No GWAS on somatic mutation burden in a dataset representative of the general population has been performed. However, the contribution of background mutational processes to cancer samples has been inferred via decomposition techniques such as non-negative matrix factorisation (NMF), Independent component analysis (ICA) and variational autoencoders (VAE). The mutational signatures were then used to test for genetic variation across cancer datasets [324, 325].

Mutational signatures are generated from the factorisation of the matrix of somatic mutation counts into lower-dimensional matrices that capture the individual mutational processes contributing to the mutational profile and the relative contributions of these processes across samples. The set of basis vectors, termed mutational signatures, gives an insight into the mechanisms that contribute to the mutation profile of an individual. These signature analyses have largely been restricted to cancer data, but certain mutational signatures are common across all samples and accumulate in a clock-like way. In the Cosmic signature dataset, these are called SBS1 and SBS5 [113]. The proposed aetiology of SBS1 is the spontaneous deamination of cytosine while the mechanism of

SBS5 remain unknown [113, 84, 115, 326]. Some signatures are overactive in specific types of cancer, for example, the aberrant mismatch repair mutational signature SBS44 significantly contributes to the mutational landscape of Lynch syndrome patients [327].

Lynch syndrome is a disease conveying an increased risk of colorectal cancer due to inherited pathogenic mutations in the mismatch repair (MMR) system. It is the most common cause of hereditary colorectal cancer and accounts for 3% of all colorectal cancer cases [327]. The MMR system identifies and rectifies mismatches in the DNA sequence. It has a major role in maintaining genomic integrity. Loss of function in the MMR pathway leads to somatic hypermutation and microsatellite instability [328]. Typically, diagnosis of Lynch syndrome relies on pathogenic mutations in *MLH1, MSH2, MSH6, TATGAA* or *PMS2*. [312].

As the single-read data showed better evidence of capturing variation across samples in the number of somatic mutations than the overlapping portions of paired-end reads (Chapter 2 Fig. 2.6). The remaining analyses were restricted to the single-read mismatch data.

We recovered a relationship between the mismatch burden and age, reflecting a contribution of somatic mutations to the observed mismatches in the UK Biobank data (Chapter 2). In this Chapter, we investigate variation in the mismatch burden across samples with a view to uncovering additional variables (including genetic variation) that may contribute to variation in the somatic mutation rate. We highlight the importance of accounting for batch structure by investigating the effect of unnormalized and batch normalised mismatch loads on cancer and smoking status. We investigate batch structure by measuring the asymmetry between mismatch classes on the transcribed and non-transcribed strands. As mutagenic processes are context-specific [329], we further generate mismatch loads in 192 triplet contexts, testing for age associations and the role of genetic variation in explaining the variation across samples. To further extract true mutational processes, we decompose mismatch loads into mismatch signatures and test for stronger associations with age and for genetic variations driving variation in the contribution of each signature. Lynch-like syndromes exhibit aberrant mismatch repair. Samples with Lynch-like diagnosis are known to have increased somatic mutation burdens [330]. Using mutational signatures of mismatch repair deficiency, we test the relative contribution of each signature to Lynch-like samples and the remaining UK Biobank samples. We also examine the MMR signatures for association with age within the Lynch and remaining samples.

3.3 Results

3.3.1 Phenotypic associations with mismatch load

In Chapter 2, we demonstrated the contribution of somatic mutation to the variation in mismatches across samples. This chapter tests whether the contributions of other variables to the somatic mutation rate may also be detectable through analysis of the mismatch burden. With this idea in mind, we first tested for the association between the mismatch burden and self-reported cancer and smoking status. We found an increase in the number of mismatches in individuals with self-reported cancer relative to the non-cancer group (Fig. 3.1 A & B). The median number of mismatches per sample was 45,850 in the non-cancer group (n= 182,679). In contrast, we found 438 extra mismatches per sample in the group with self-reported cancer with a median of 46,288 (n= 17,886; Wilcoxon rank sum test $P = 0.06$; Table 3.1). This effect is clearer when the mismatch loads are decomposed into transcriptome-aligned mutational contexts (Fig. 3.1 B). Across all 192 mismatch contexts, 128 remained significant after the Bonferroni correction for multiple testing. The difference between means in the CTA-to-CGA mismatch load has the strongest statistical support. The median in the non-reported group was 39, while the self-reported group was 43 (Wilcoxon rank sum test $P = 4.07 \times 10^{-28}$)). Smoking status showed a weak association with mismatch load, with individuals who have never smoked having, on average, 125 more mismatches than the smoker group (45,844.5 compared to 45,969.5 ; Fig. 3.1 C). Of the 200,000 sample set, 21,009 reported that they were non-smokers, and 30,880 reported that they were smokers. We observed a marginally significant association in the mismatch context CCC-to-CAC (Wilcoxon rank sum test $P = 0.02$). However, given the number of statistical tests (n=193) this relationship did not survive correction for multiple testing (Fig. 3.1 D).

Given the significance of the mismatch data reflecting plausible phenotypic associations, we normalised the mismatch data by regressing the sequencing depth as a function mismatch count for each sequencing batch and transformed the data to a uniform distribution. In Chapter 2, we found significant structure within the mismatch data

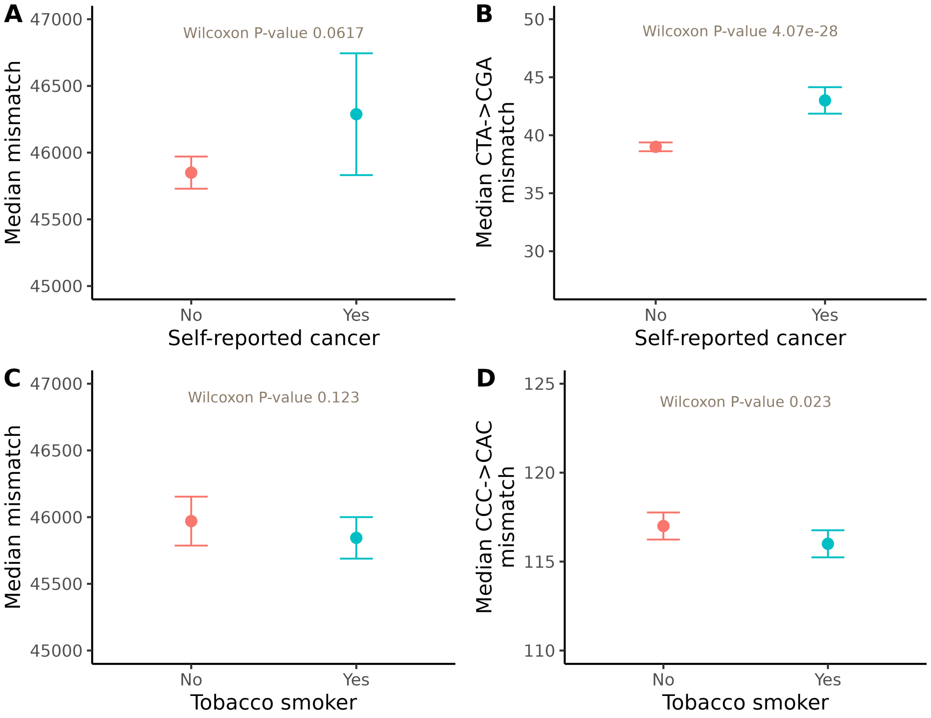

Figure 3.1: Median ± 95% CI plots of mismatch loads and self-reported cancer and smoking status. Triplet contexts were chosen by the strongest statistical associations in the unnormalized data. The median mismatch load per sample (y-axes) versus cancer or smoking status (x-axes).

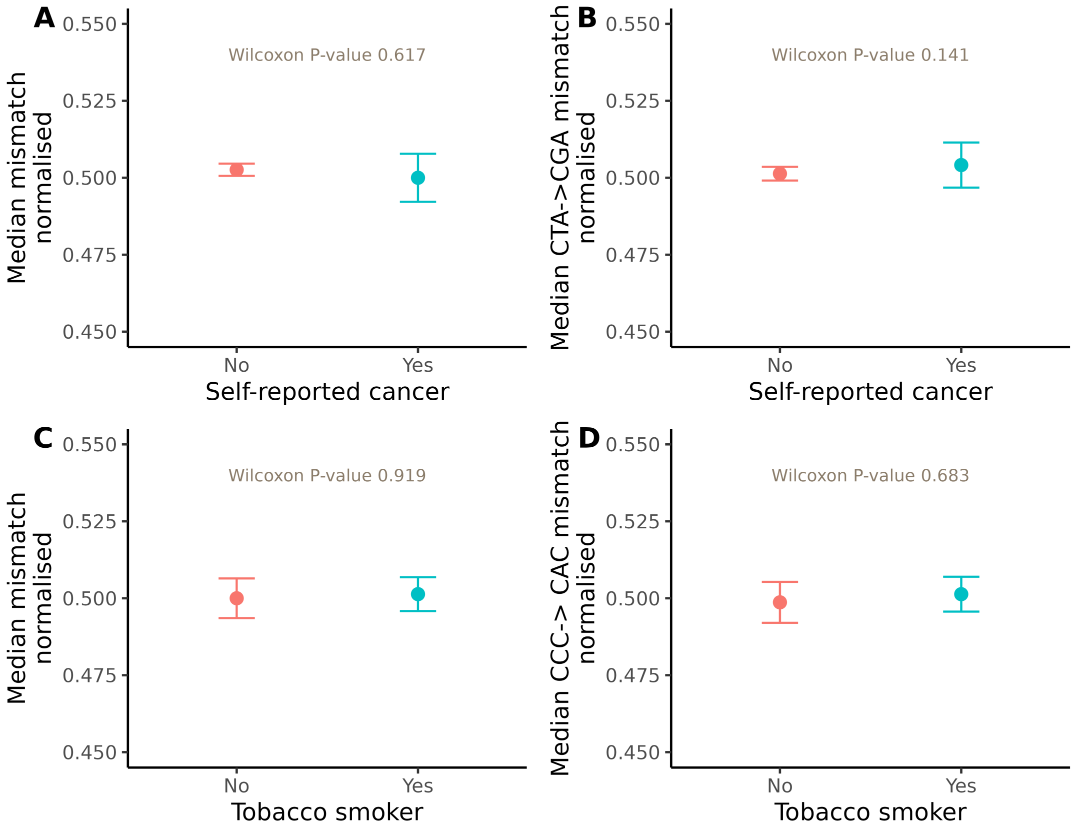

Figure 3.2: Median ± 95% CI plots of mismatch loads for the CTA-to-CGA context grouped in self-reported cancer and the CCC-to-CAC mismatch load grouped by smoking status. Triplet contexts were chosen by the strongest statistical associations in the unnormalized data. The median mismatch load within each context per sample (y-axes) versus self-reported cancer or smoking status (x-axes).

attributable to the sequencing batch. To remove the possibility of the sequencing batch giving false phenotypic associations we regressed the mismatch data on sequencing batch. This enabled us to exclude sources of technical variation across batches, such as different reagents and storage temperatures between sequencing runs. However, after accounting for batch variation, we found no discernible statistical difference between the self-identified smoking and cancer groups (Fig. 3.2). Spurious associations may be driven by differences in smoking habits and cancer rates between regions and non-random allocation of the samples from different regions to sample batches. By plotting the distribution of Wilcoxon signed rank test p-values (Table A1) for all mismatches, the effect of batch structure generating false statistical signals is highlighted (Fig. 3.3).

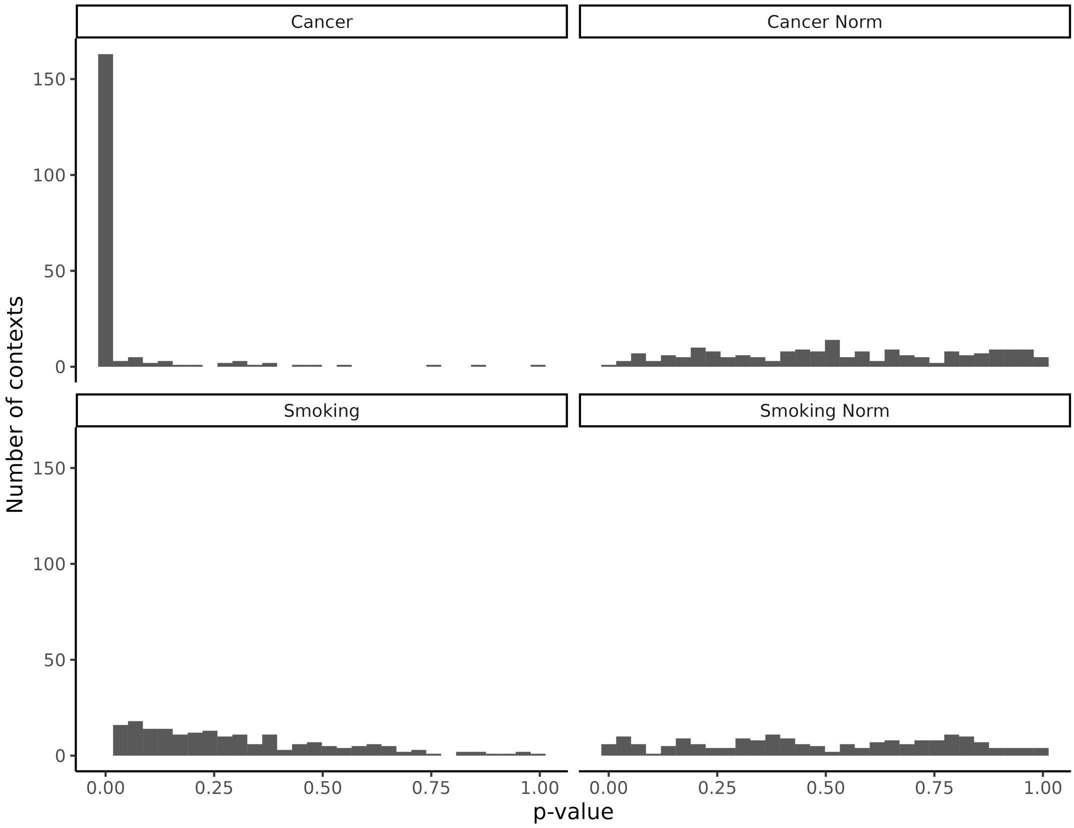

Figure 3.3: Distribution of p-values across all triplet contexts. The y-axis contains the number of Wilcoxon rank sum test p-values in each p-value bin (x-axis). Uncorrected batch structure leads to spurious statistical relationships between cancer status mismatch load.

3.3.2 Investigating age-related accumulation of mismatches in NGS data

In Chapter 2, the median mismatch load for samples grouped according to their integer age was found to have the expected linear relationship with age. Here we posit that some triplet contexts may show a stronger association with age, particularly in the cases of mutations types that accumulate with age but are not commonly associated with sequencing errors or other sources of mismatches that do not correspond to somatic mutations. The median transcription-aligned triplet mismatch loads were regressed against age. Four mismatch contexts that represent the variation across the 192 regressions are shown in Fig. 3.4, and the age-associated phenotypes that remained significant after Bonferroni correction are reported in Table 3.1 (full results are reported in Table A2). Eight of the 192 mismatch loads remained significant after Bonferroni multiple test correction (Table 3.1). The AGA-to-ATA mismatch load was strongly associated with age ($R^2 = 0.63$, $P = 9.73 \times 10^{-8}$). The reverse complement of AGA-to-ATA, TCT-to-TAT was also significantly associated with age, albeit weaker ($R^2 = 0.41$, $P = 1.02 \times 10^{-4}$). The majority of the triplet contexts did not show a strong linear relationship with age (Fig. 3.5).

Table 3.1: Significant correlations between median mismatch load and age after Bonferroni correction. R^2 and p-value from linear regression are reported.

Triplet	R^2	P
AGA-to-ATA	0.63	9.74e-08
GGA-to-GTA	0.55	1.98e-06
TCA-to-TGA	0.49	1.28e-05
CGA-to-CTA	0.44	5.04e-05
GGT-to-GAT	0.42	8.34e-05
TCT-to-TAT	0.41	1.02e-04
GCT-to-GAT	0.4	1.40e-04
TGC-to-TTC	0.39	1.60e-04

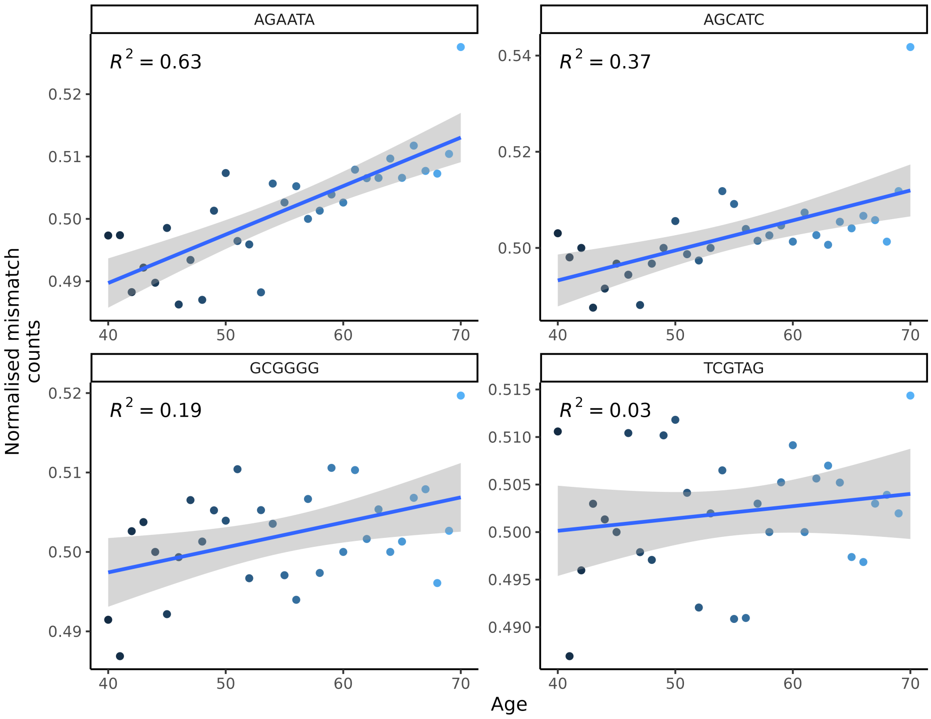

Figure 3.4: Median mismatch load as a function of integer age for four mismatch contexts AGA-to-ATA, AGC-to-ATC, GCG-to-GGG and TCG-to-TAG. Each point represents the median of the normalised mismatch count per integer age.

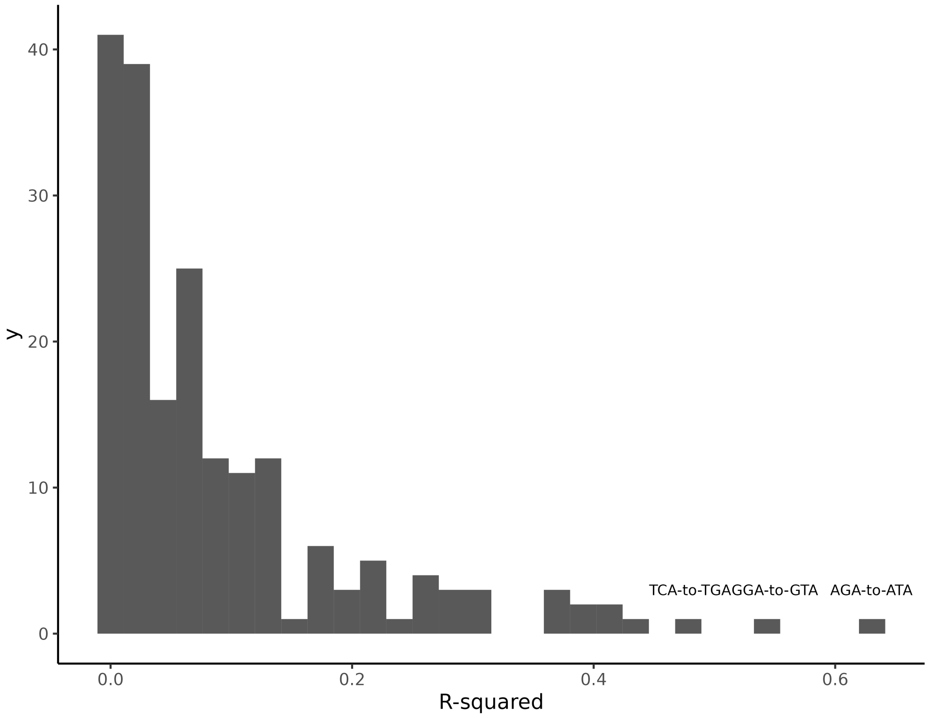

Figure 3.5: Histogram of R^2 values derived from linear models of the median mismatch load for each triplet context as a function of integer age. Statistically significant results are reported in Table 3.1 and in full in Table A2.

3.3.3 Strand asymmetry

Transcriptional processes impart strand specificity to the distribution of somatic mutations. To test for signals of transcription-associated asymmetry within the UK Biobank exome sequencing dataset, the $\log_2$ ratio of each mismatch type and its reverse complement were calculated for the C-to-{A, G, T} and T-to-{A, C, G} mismatch types (Fig. 3.6). As the exome sequencing data are aligned relative to the forward strand of the reference genome, mismatch classes within genes where the coding sequence is on the negative strand are reverse transcribed, i.e., G-to-T becomes C-to-A. We found significant levels of transcription strand asymmetry between the C-to-A and G-to-T mismatch classes with an average $\log_2$ ratio of 1.85 (3.6 times more C-to-A than G-to-T on the coding strand). The sample with the highest level of asymmetry in the C-to-A type had a $\log_2$ ratio of 5.4 or approximately 43 times more C-to-As than G-to-Ts. By inspecting the sequencing batch containing the sample with the strongest asymmetry, batch ID HF273DSXX, a subset of samples showed significant levels of asymmetry compared to most samples within the sequencing run (Fig. 3.7). The level of asymmetry within this batch was bimodal. Grouping each sample within HF273DSXX by assessment centre showed no evidence of intra-sequencing batch variation due to sample handling at different assessment centres.

Building on the intuition that DNA damage occurs randomly across transcriptional strands and given our simulations in Chapter 2, we expect a modest contribution from somatic mutations within the data. The level of asymmetry was inconsistent with the result arising from mutagenic transcriptional processes. To further investigate the possibility of true mutational processes driving the asymmetry, we partitioned the data based on gene expression level in whole blood in the GTEx dataset. The triplet mismatch counts were further normalised to account for discrepancies in the number of opportunities a triplet mismatch can occur within each group, to account for differences in nucleotide content of the coding and non-coding strands. After normalisation of the triplet counts for sequence content, we found a higher mismatch load in the non-expressed genes compared to the expressed gene sets. An elevated mismatch load was observed across all mismatch types in the triplet contexts containing CpGs in the reference sequence in both the expressed and non-expressed genes (Fig. 3.8).

Under the assumption that the mismatch loads are derived from sequencing and technical errors only, aligning each mismatch context to the transcribed strand (Tx+)

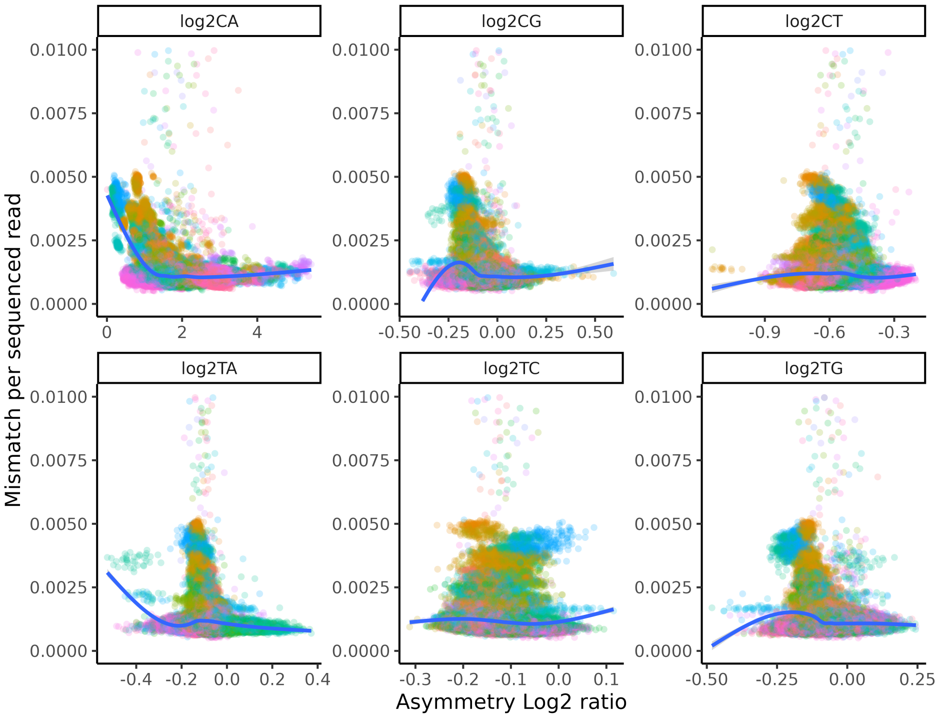

Figure 3.6: Transcription strand asymmetry. Log_2 ratio for each of the 6 mismatch types. Log_2 ratios of 0 indicate no transcriptional strand asymmetry. Samples are coloured by sequencing batch.

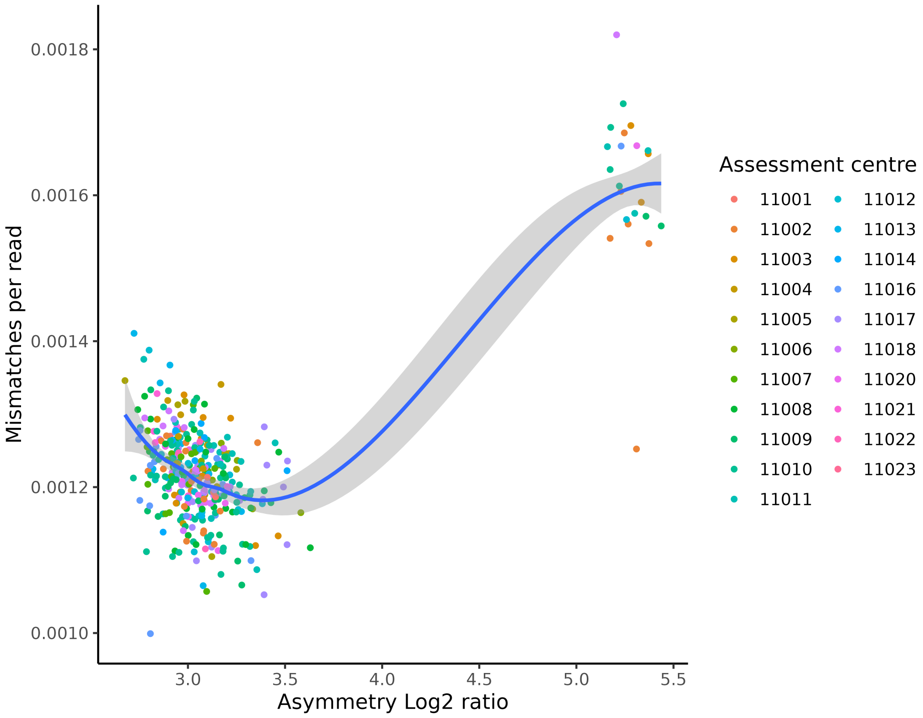

Figure 3.7: C-to-A versus G-to-T asymmetry for sequencing batch HF273DSXX. Samples are coloured by UK Biobank assessment centre. The number of mismatches per read passing QC (y-axis) is plotted against the $\log_2$ ratio of the C-to-A to G-to-T mismatch counts (x-axis).

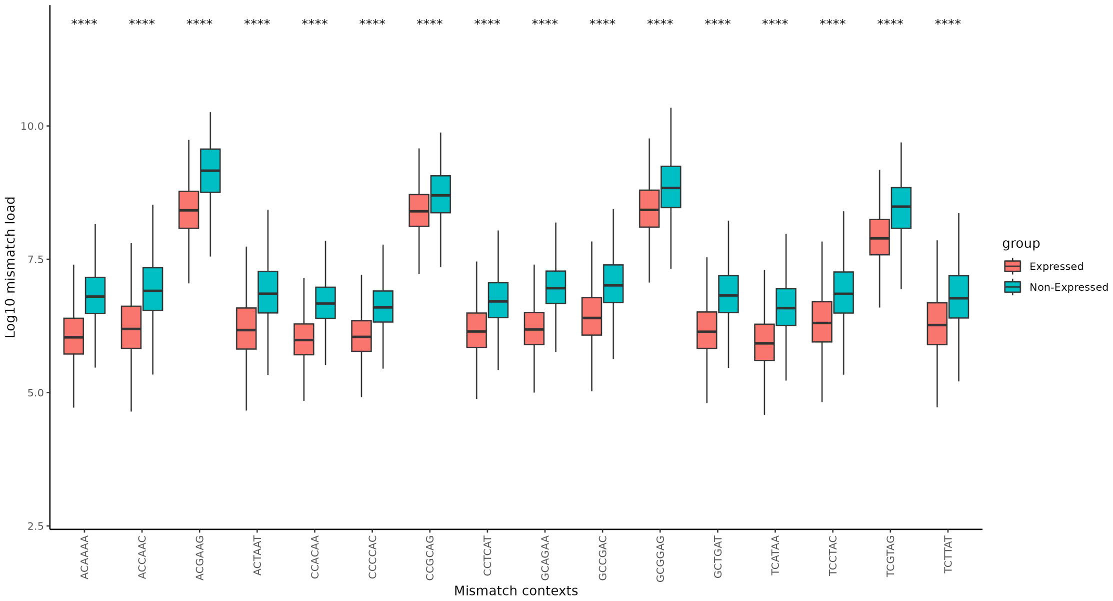

Figure 3.8: The difference in normalised mismatch loads grouped by expression status. All triplet contexts containing the C-to-A mismatch type are shown. The mismatch loads have been normalised by the frequency of each triplet context in expressed and non-expressed genes. The normalised mismatch loads are minus log transformed.

should result in a symmetric mismatch load across the mismatch types (when the data are normalized to account for differences in nucleotide opportunities between the coding and non-coding strand). We found, however, a strong asymmetry between certain classes of mismatch types and their relative reverse complements. The mismatch contexts with the largest asymmetry show enrichment for C-to-A and G-to-T transversions (Wilcoxon rank sum test $P = 5.81 \times 10^{-5}$), consistent with the asymmetry observed when the mismatch data is aggregated into mismatch types (Fig. 3.6). We find that for three mismatch loads, the $\log_2$ ratio is greater than two, indicating that there are at least four times more of one mismatch context relative to its reverse complement. The ACT-to-AAT context had a $\log_2$ fold change of 2.83, indicating that, on average, there were almost seven times more ACT-to-AAT mutations than its reverse complement AGT-to-ATT mutations. However, the difference between the $\log_2$ ratios in the expressed and non-expressed group was -0.08 (Table A3).

We also found that mismatch contexts statistically associated with age were more likely to exhibit strong asymmetry (Fig. 3.9). However, we did not see any increase in the $\log_2$ ratio as a function of sample age (data not shown). To further investigate the

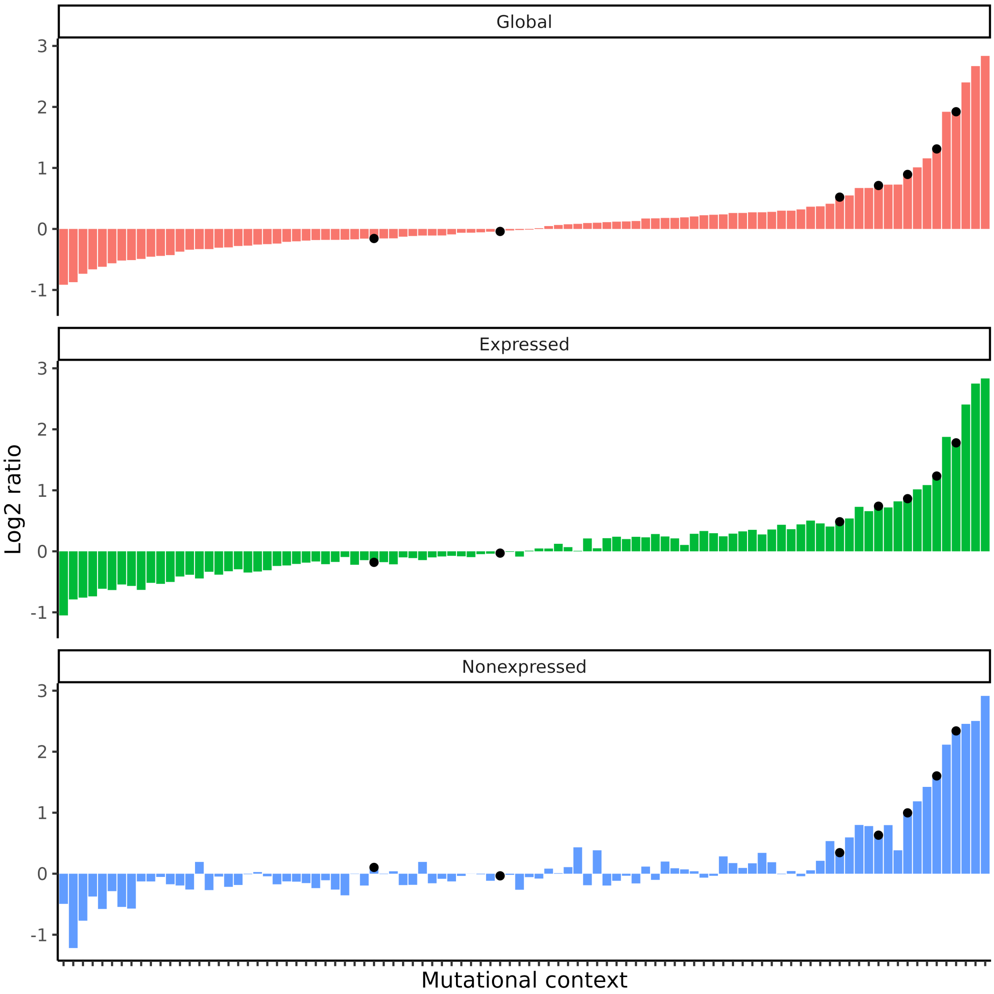

Figure 3.9: Mismatch context asymmetry. The $\log_2$ ratio of each 96-mismatch context with its reverse complement. Global is the mutational asymmetry across all genes, while expressed and non-expressed refer to gene expression group membership. The triplet contexts are ordered by increasing global asymmetry. The black dots indicate whether the context or its reverse complements is statistically associated with age (Table A3).

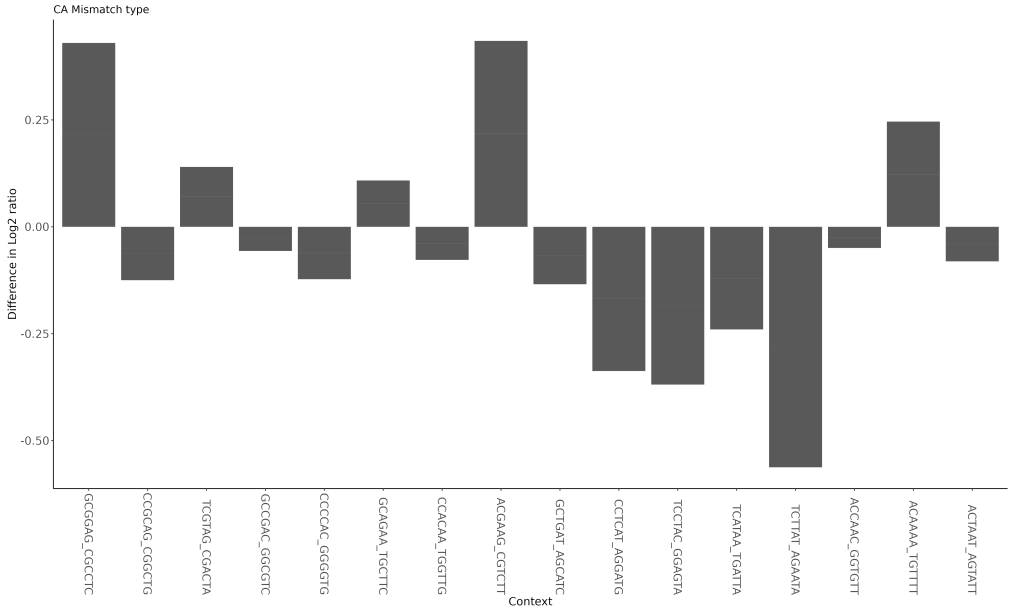

Figure 3.10: Difference in $\log_2$ ratios between expressed genes and non-expressed genes. The mismatch contexts containing the C-to-A mismatch type are only shown. Positive values indicate that the $\log_2$ ratio of the expressed genes is larger than that of the non-expressed.

asymmetry between mismatch contexts, we analysed the mismatch asymmetry in genes expressed in GTEx whole blood samples and from genes that are not expressed, (see Methods). Transcription-strand asymmetry should be more pronounced in genes that are expressed relative to the non-expressed group if the effect is due to the transcription process. The difference in $\log_2$ ratios between the expressed and non-expressed gene sets for the C-to-A mismatch type showed variation across the triplet contexts, indicative of a sequence-dependent bias in mismatch accumulation (Fig. 3.10). The lack of remaining asymmetry in the C-to-A type was consistent with our previous findings; transcriptional processes were not driving the observed asymmetry.

Consistent with our expectation from Fig. 3.6 & 3.7 that asymmetry is strongly influenced with the sequencing batch, regressing the $\log_2$ ratios against the sequencing batch and assessment centre showed that the sequencing batch had a high variance in the median $\log_2$ ratio across sequencing batches compared to assessment centre (Fig. 3.11 & 3.12; Statistical support Fig. A1). We found that for contexts with high levels of asymmetry there is variation across the sequencing batches, again, consistent with the result reported in Fig. 3.7. In contrast, the assessment centre shows no relationship

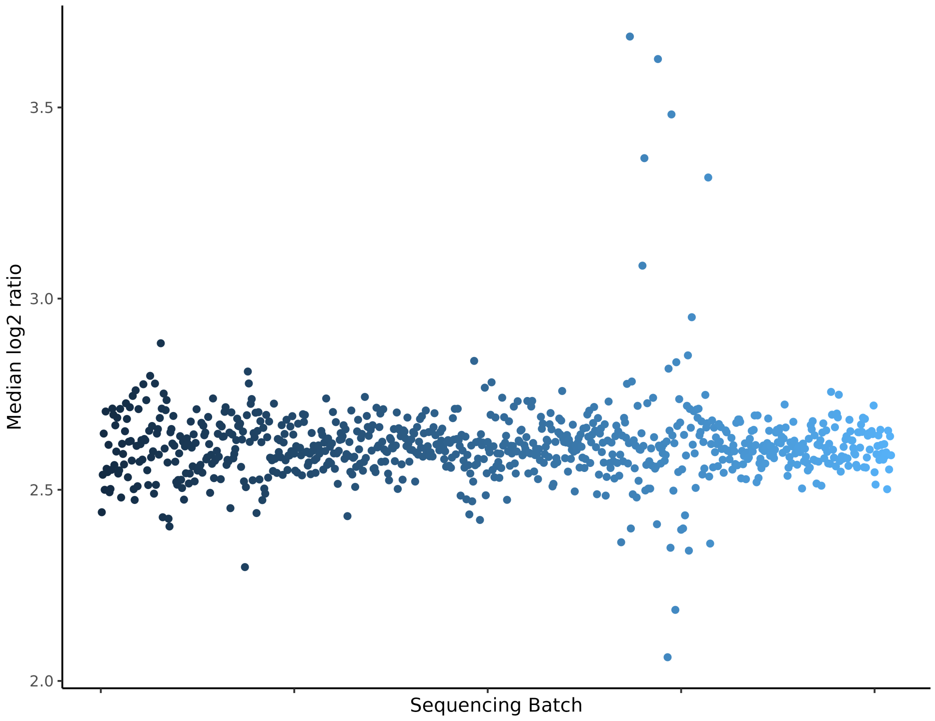

Figure 3.11: Median Log$_2$ ratio of the ACT-to-ATT and it reverse complement AGT-to-AAT mutation types. A scatter plot of the median log$_2$ ratio of the ACT-to-ATT mismatch load (y-axis) is reported as a function of sequencing batch (x-axis).

with the level of asymmetry (Fig. 3.12 & Fig A1).

3.3.4 GWAS

We performed GWAS using the 192 mismatch loads as phenotypes and the white-British samples within the UK Biobank 200,000 release (158,760). Four mismatch loads (GCG-to-GGG, TTC-to-TGC, CTC-to-CGC & CCG-to-CGG) had peaks that reached genome-wide significance (Table. 3.2). For each of the four mismatch loads, all had only one locus with a significant association. rs17537237, an intronic variant in *TMEM117*, a gene involved in mitochondrial transport, was significant for the TTC-to-TGC mismatch load. For the CTC-to-CGC mismatch load, the most significant variant was rs9906359, an intronic variant in *FMNL1*, which is involved in cell morphology and cell polarity. rs12481160, an intron variant in an uncharacterised non-coding RNA, was significant for the CCG-to-CGG mismatch load.

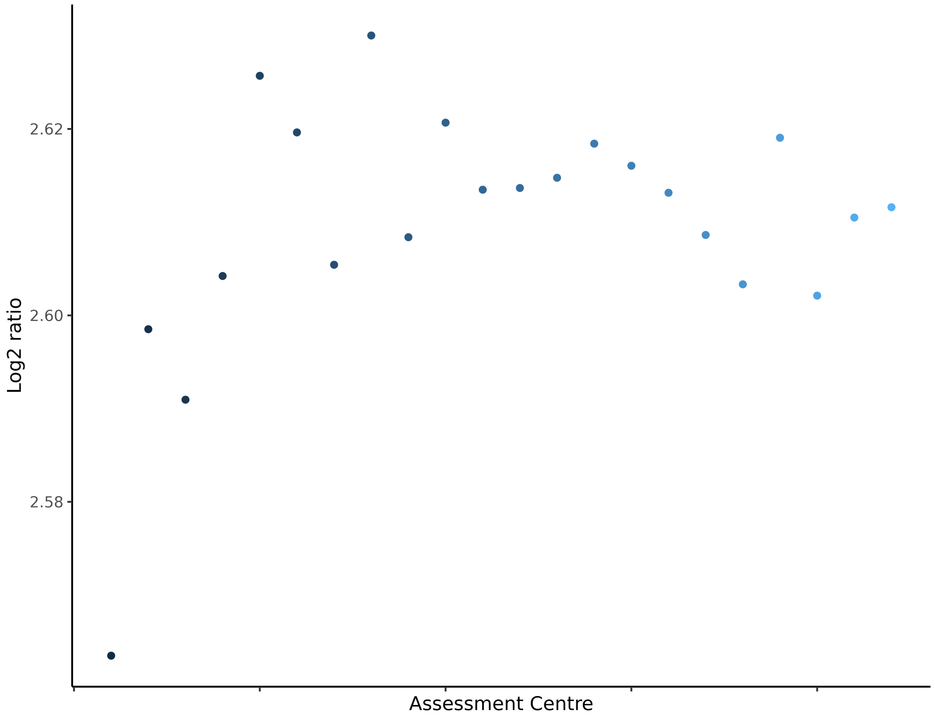

Figure 3.12: Median Log_2 ratio of the ACT-to-ATT and it reverse complement AGT-to-AAT mutation types. A scatter plot of the median $\log_2$ ratio of the ACT-to-ATT mismatch load (y-axis) is reported as a function of assessment centre (x-axis).

The strongest statistically associated locus, tagged by rs12332549 ($P = 1.49 \times 10^{-79}$), was identified for the GCG-to-GGG mismatch load (Fig. 3.13 & 3.14). rs12332549 is an intron variant within the *ZSWIM6* gene, a zinc-finger binding protein. As the effects of SNPs may be distal to their location or the statistically tagged SNP may be in linkage with the causative SNP, we analysed this locus using LDlink [331]. Strikingly, we found that rs6861729 (the second most significant SNP for GCG-to-GGG; $P = 2.45 \times 10^{-79}$) showed strong evidence of linkage (effect size= -0.343; $P = 2.32 \times 10^{-21}$; Table. 3.3) with a cis-eQTL that decreases the expression of *ERCC8*. Also known as Cockayne Syndrome A, this is an essential gene for TC-NER (Table 3.3) and, thus, a highly plausible candidate locus for association with the somatic mutation rate.

Table 3.2: Significant GWAS loci across 192 mismatch contexts

Phenotype	tagging SNP	P	Gene Symbol	Consequence
GCG→GGG	rs12332549	1.49e-79	ZSWIM6	Intron variant
TTC→TGC	rs17537237	2.09e-26	TMEM117	Intron variant
CTC→CGC	rs9906359	3.86e-18	FMNL1	Intron variant
CCG→CGG	rs12481160	3.05e-49	LOC107985448	Intron variant

Table 3.3: Ten most significant LD eQTLs in linkage with rs6861729

	Query	RSid	R^2	Gene Symbol	Tissue	Frequency	Effect Frequency	Effect Size	P
1	rs6861729	rs34902701	0.1	ERCC8	Cultured fibroblasts	T=0.247	TGTTA=0.753	-0.343	2.326e-21
2	rs6861729	rs248686	0.1	ERCC8	Cultured fibroblasts	C=0.247	T=0.753	-0.343	3.255e-21
3	rs6861729	rs35212229	0.1	ERCC8	Cultured fibroblasts	CA=0.247	C=0.753	-0.346	1.233e-20
4	rs6861729	rs35453042	0.197	ELOVL7	Esophagus - Mucosa	T=0.842	C=0.158	-0.515	6.393e-19
5	rs6861729	rs17331746	0.199	ELOVL7	Esophagus - Mucosa	T=0.843	C=0.157	-0.515	1.742e-18
6	rs6861729	rs62372074	0.199	ELOVL7	Esophagus - Mucosa	A=0.843	T=0.157	-0.515	1.742e-18
7	rs6861729	rs12516552	0.203	ELOVL7	Esophagus - Mucosa	C=0.845	G=0.155	-0.511	3.096e-18
8	rs6861729	rs35078341	0.203	ELOVL7	Esophagus - Mucosa	T=0.845	A=0.155	-0.511	3.096e-18
9	rs6861729	rs35957723	0.203	ELOVL7	Esophagus - Mucosa	G=0.845	A=0.155	-0.511	3.096e-18
10	rs6861729	rs62367880	0.203	ELOVL7	Esophagus - Mucosa	G=0.845	T=0.155	-0.511	3.096e-18

We sought to rule out technical artefacts that could give rise to this result. Although, great care was taken to remove every germline variant within the UK Biobank data. There was a possibility of high-frequency germline events remaining after our stringent SNP filtering. We devised an additional test for loci that showed a putative association with the somatic mutation rate that consisted of removing mismatches on the same chromosome as the test SNP and regenerated the mismatch loads. If there were unfiltered germline variants in LD with the test SNP driving the association, this would remove the effect. We also regenerated the 192 phenotypes imposing a per-site recurrence filter of 1,000, i.e., all mutations that occurred in more than 1,000 samples were

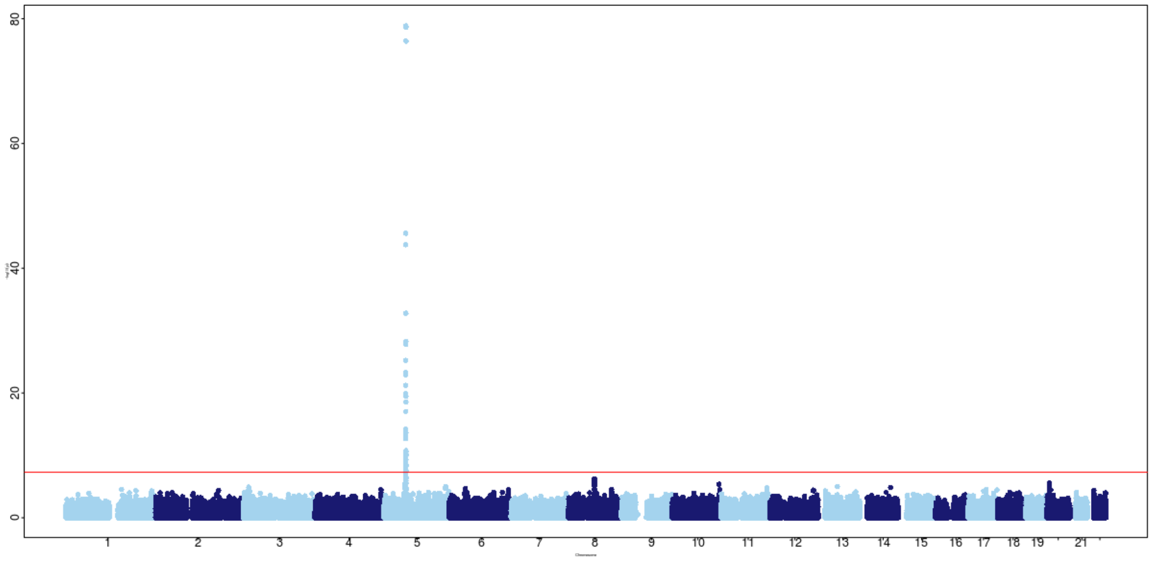

Figure 3.13: Manhattan plot of the GCG-to-GGG mismatch load. The x-axis displays the chromosome and base pair positions. The strength of association as the -$\log_{10}$ p-value is measured on the y-axis.

excluded from the mismatch load. We investigated the rs12332549 locus and discovered a mismatch within 1kb of the test SNP that was present in 25,000 samples (Chapter 2, Fig. 2.3). Our exploratory analysis of this locus showed that the mismatch occurred in fewer than 3 reads across each of the 25,000 samples; however, it showed strong evidence of LD with the associated locus, suggesting that it is a germline variant and that the association of the mutation load with this locus was an artefact. Indeed, when we placed a threshold on the maximum number of samples carrying the mismatch, as described above, the associations at this and all other significant loci described above were no longer significant (Fig. 3.15). This demonstrated that, although our method of removal of uncalled 'germline events' was stringent, some germline variants were retained, with the potential, in extreme cases, to create false signals of association between linked germline loci and the mismatch load.

3.3.5 Mutational signature analyses

Pathogenic genetic variants underlying the conditions such as Lynch syndrome cause higher rates of somatic mutations [328]. Matrix factorisation has identified several mis-

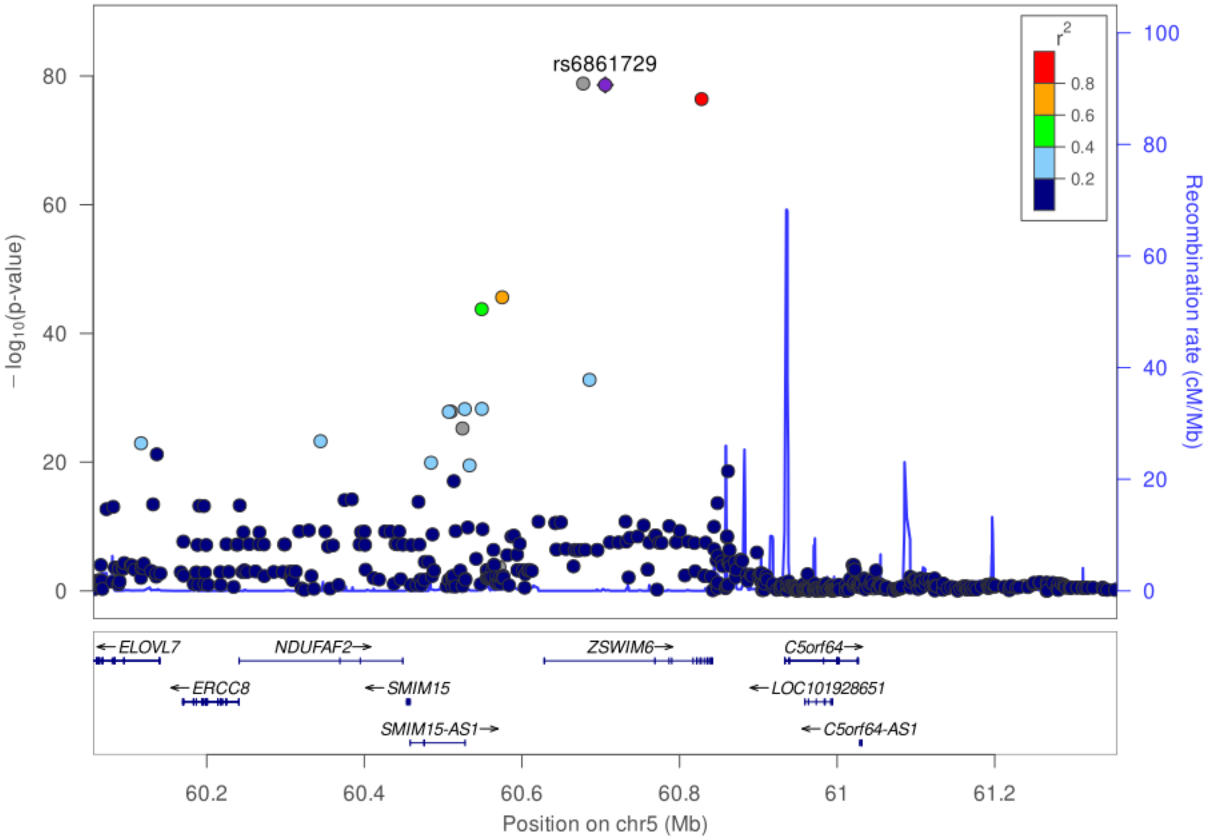

Figure 3.14: GWAS performed on the unfiltered GCG-to-GGG mismatch load. The Locus-zoom plot of tag SNP rs6861729 in the *ERCC8* & *ZSWIM6* locus. SNPs are coloured based on linkage R^2 with the tagging SNP. The y-axis (left) shows the strength of association between the test SNPs and the GCG-to-GGG mismatch load as measured by the $-log_{10}$ p-value. The y-axis (right) contains the recombination rate in centimorgans per megabase.

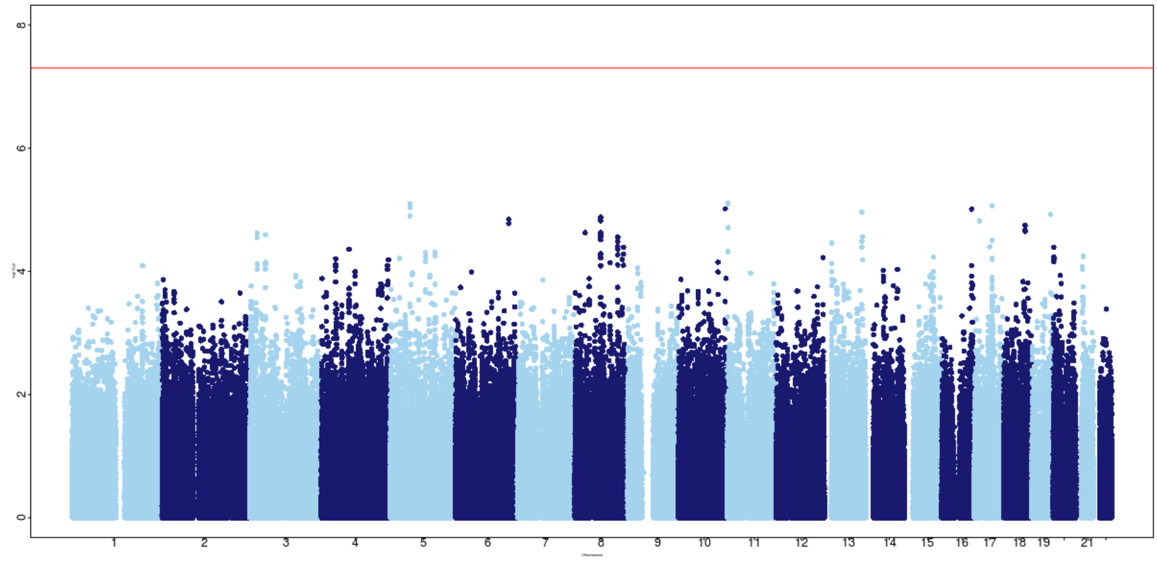

Figure 3.15: Manhattan plot for the GCG-to-GGG mismatch load following the removal of mismatches found in more than 1,000 samples.

match repair (MMR) mutational signatures that are over-active in Lynch-like samples. Of the 200,000 UK Biobank exomes, we identified 160 samples with Lynch-associated germline variants using the criteria of *Patel et al.* [312]. The contribution of 13 COSMIC mutational single base-pair signatures (Table 3.4), including two background clock-like mutational signatures SBS1 and SBS5, to samples within the UK Biobank, was estimated using non-negative least squares regression (methods). Of the 11 DNA repair signatures tested, SBS3, SBS9, SBS18 and SBS21 returned non-zero estimated contributions. The background mutational signatures, SBS1 and SBS5, showed no difference between the normal and Lynch-like group (Fig. 3.16). SBS18 had a marginally larger contribution in the normal group (Wilcoxon rank sum test $P = 0.023$). However, this result was not statistically significant after correction for multiple hypothesis testing. SBS1, SBS5 and SBS18 accumulate in a clock-like fashion over the human lifespan [332]. In the Lynch samples, no signal of somatic mutation across the mean contribution of the four mismatch repair mutational signatures tested was observed. When the signature contribution was regressed against age, however, we did find a positive significant association between the contribution of mutational signature SBS3 and age (Fig. 3.17).

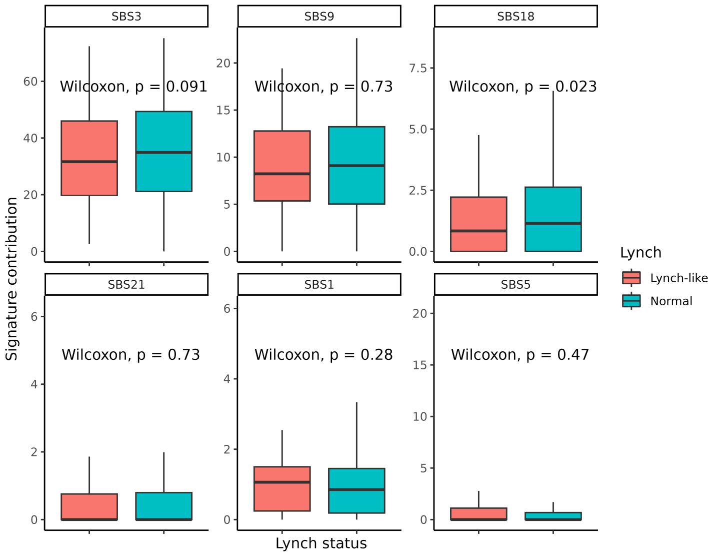

Figure 3.16: Boxplot of four MMR mutational signature contributions and two clock-like background signatures. Samples are grouped on whether they contain Lynch-like pathogenic variants or not. Wilcoxon rank sum test was performed to assess statistical differences between the groups.

3.4 Discussion

In this chapter, we set out to investigate sources of variation in somatic mutation through an analysis of the burden of mismatches of different types across 200,000 exomes in the UK Biobank. In Chapter 2, we estimated that only a small proportion of somatic mutations contribute to the mismatch loads. Yet, when we test for association with age, we find we can recover the contribution of the somatic mutations to the mismatch loads. This chapter decomposes the mismatch loads into transcriptionally-aligned triplet contexts to enrich the signals of somatic mutation. In doing this we reasoned that DNA damage and sequencing error are unlikely to accrue across all nucleotide triplets at a constant rate and transcriptional processes impart biological strand information into the profile of mismatches. This allowed us to investigate the age-related signal we recovered in Chapter 2 and to test for further phenotypic associations within the UK Biobank dataset, such as cancer and smoking status. We quantified the mismatch asymmetry within the data using transcription strand information and gene expression data from the whole blood GTEx dataset. Next, we shifted our focus from environmental factors associated with the variation in mismatch counts across samples in the UK Biobank to

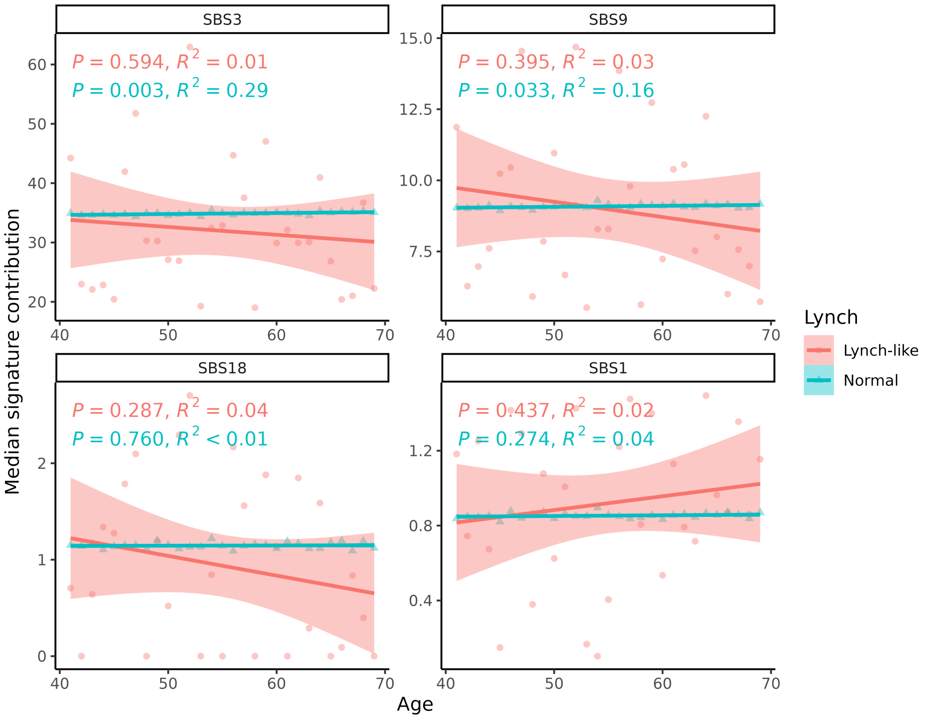

Figure 3.17: Median mismatch load as a function of integer age for four MMR mutational signatures Each point represents the median of the normalised mismatch count among individuals with the same integer age.

intrinsic processes driving somatic mutation rate variation. Using a GWAS approach we identified loci that are associated with variation in the triplet context-specific mutation loads. GWAS on somatic mutation is poorly researched owing to the difficulty in assessing the level of somatic mutation across a large set of samples. Developing on the idea of genetic variation impacting somatic mutation load, we identified samples with known pathogenic Lynch-like genetic variants. Lynch-like syndromes have been empirically shown to have higher somatic mutation rates [328]. We identified a set of DNA repair COSMIC signatures and estimated their contribution to each sample. We then tested for differences in the contribution of the mutational signatures between Lynch-like samples and the remaining individuals.

Several factors may influence the somatic mutational load within a sample. We reasoned that individuals diagnosed with cancer or individuals who are current or past smokers might have a higher somatic mutation load in blood cells and that this could be reflected in higher mismatch loads. This effect has not been previously been investigated in a cohort as large as the UK Biobank but it could have important implications for quantifying cancer risk. We found a statistically significant relationship between mismatch load and cancer and smoking status. Given the significance of this result in furthering the understanding of the role of somatic mutation and cancer risk, we searched for potential confounding effects that may have led to this result. In Chapter 2, a structure within the mismatch data associated with the sequencing batch runs was clear when we visually inspected the data. We reasoned that this technical variance might be driving the statistical associations between cancer and smoking status. Many technical challenges arise in the whole exome sequencing of 200,000 exomes, from different storage conditions across assessment centres to the trans-Atlantic shipping of frozen blood samples to the Regeneron sequencing facility in New York. At each time point in the sample handling chain, batch effects can be introduced. We identified two possible sources of large-scale batch effects, the assessment centre that reflects the initial handling of the blood samples and the sequencing batch. We found no significant structure related to the assessment centre in the mismatch data. This suggests two insights into the data. First, a standardised methodology in sample handling prior to long-term storage has limited the variance in DNA damage, and secondly, that fine-scale population structure across the assessment centres is not driving differences within the sample mismatch loads.

Although not available as supplementary data from the UK Biobank repository, sequencing batch information can be extracted from the read ID in the CRAM alignment files. We found that, on average, 800 samples were sequenced per run on the Illumina Nova-seq 6000 at the Regeneron sequencing facility. By grouping samples by their sequencing batch, we saw a clear structure within the mismatch data. The sequencing batch reflected different rates of mismatch accumulation per sequencing run. Although many technical factors may influence different error rates per sequencing run, we also discovered significant intra-batch variation within the mismatch loads. This result may have implications for researchers using the UK Biobank exome data. A best practice methodology has been developed for performing GWAS on the UK Biobank imputed genotype data this includes using genotype batch as a covariate within the regression analysis. Given the differences we observed within the rate of mismatch accumulation, GWAS studies performed on genotypes called from the exome sequencing data that fail to account for sequencing batch may lead to spurious phenotypic associations. Accounting for the sequencing structure may also decrease the false positive rate in rare variant burden tests.

Given the observed sequencing batch structure, we sought to remove as much of the sequencing batch effect as possible. As described in more detail in Chapter 2, we adjusted our data by modelling the total number of mismatches within a sample per sequencing batch as a function of the number of autosomal reads. We then determined the within-batch fractional rank of the residuals of this model for each sample. Following this process, each sample has a value in the range [0,1] corresponding to the relative position of the model residual of that sample among all samples of the same batch. When we re-examined the associations we had found between mutation load and cancer and smoking status, we found that the sequencing batch was a significant source of confounding. For example, none of the 128 mismatch contexts passed multiple testing correction for association with cancer status after adjusting for the sequencing batch. Although disappointing, by understanding the sources of confounding within the data, we can begin to refine the mismatch loads. Accounting for as many external sources of variation, we can build further confidence in the results contained within this study.

As we age, we accumulate somatic mutations - this phenomenon has been studied extensively in small cohorts of healthy individuals [93, 160]. In Chapter 2, where we found a linear relationship between the total number of mismatches within a sample and age. As DNA damage and sequence error accumulate at higher rates for some mutation types, we expected the relationship with age to depend on the triplet mutation context. We found this to be the case, with some mutations, such as AGA-to-ATA having a strong association with age (R^2 of 0.63) and others showing no age association. For most contexts, a weak relationship with age was observed. It is important to note that the relationship between mutations and age is obscured by technical factors such as sequencing error and DNA damage, which may also depend on the nucleotide context. Thus, our study has limited potential to determine the somatic mutation types that are most strongly correlated with age, due to technical confounding. A point worth noting is that the variance across each bin, in this case, each integer age, is not shown. The reported model fit is on a median across a large number of samples.

Given that we can detect an age-associated increase in the median mismatch loads, we sought to investigate if there was a difference in the rates of mismatch accumulation on the transcribed strand relative to their reverse complements. DNA damage and sequencing errors accumulate randomly across the Watson and Crick strands. The rate of DNA damage at a specific nucleotide on the Watson strand should be approximately equal to that of the same nucleotide on the Crick strand. By measuring the rate of mismatch per substitution type on the transcribed strand, i.e accumulation of all C-to-As on the transcribed strand compared with all the G-to-T substitutions on the non-transcribed, we expected to recover a signal of asymmetry resulting from somatic mutations, the accumulation of which is influenced by transcription-associated repair and damage processes. For the C-to-A class we found significant levels of asymmetry, which was much stronger than was observed for the remaining 5 mismatch types (Fig. 3.6). This class of substitution is well established in the literature as being associated with transcription associated mutagenesis [123, 148].

As the exome sequencing data was generated from whole blood, using the whole blood GTEx dataset, we sought to understand if gene expression might be driving the observed asymmetry by determining the mismatch loads separately for expressed genes and non-expressed genes. We found a significant increase in the mismatch loads for the non-expressed derived data. For somatic mutations, this direction of effect has been previously reported [148]. In addition to higher mismatch rates at low expression levels *Chen et al.* found that increased expression resulted in increased strand asymmetry between all mutation types. As sequencing errors are agnostic to biological strands, any deviation from symmetry is expected to result from true somatic mutations incurred by transcriptionally associated processes. We report a strong level of asymmetry across most mismatch contexts, with the C-to-A mismatch types showing the largest levels of transcription-strand-associated asymmetry once again consistent with our previous findings.

Although expected for a mutational phenotype inferred from high-confidence somatic mutation calls, the strength of this effect is not what we expect to see, given our findings and simulations in Chapter 2, as most mismatches are expected to be from sources of technical variation. Under the assumption that technical errors are the sole contributor to the mismatch load, we should observe strand symmetry due to Chargaff's second parity rule. We can test this hypothesis by inferring the asymmetry from genes not expressed in whole blood. Significant asymmetry was found within the non-expressed group, indicating that expression did not the cause of the transcription strand asymmetry. The level of asymmetry as a function of mismatch load also showed structure related to the sequencing batch (Fig. 3.6 & 3.11). We tested the level of asymmetry as a function of age. As transcriptional-associated mutagenesis is predicted to accumulate with age, we expected a linear increase in asymmetry. However, this was not the case. Together our results implicate 8-oxoguanine DNA damage as the primary source of this transcription strand asymmetry. Oxidation of guanine to 8-oxoguanine is the most abundant source of DNA damage, DNA polymerase inserts an adenine opposite the oxidated guanine nucleotide [65]. High levels of 8-oxoguanine arise when DNA samples are stored incorrectly [307].

Given that the transcription strand asymmetry we observed is more consistent with technical sources rather than somatic mutation, we sought reasons to explain what technical factors might cause transcription strand asymmetry. Differences in the contribution of DNA damage to read one and read two sequences in paired-end sequencing data have previously been reported [307]. Therefore, an imbalance between the number of read one and two sequences mapping to the transcribed strand could give rise to the strand asymmetry we observed. We tested for read one vs read two strand mapping imbalances for a subset of the 200,000 samples. However, we did not find a systematic imbalance in read orientation that could plausibly give rise to the asymmetry effect we observed. The most plausible explanation for the asymmetry results that we could find relates to the design of the probes used in the capture kits for the exome sequencing.

If the exome capture kit used single-strand probes, consistently designed to target the coding strand sequence rather than the reference sequence, this could result in strand asymmetry in DNA damage artefacts. Because such single-strand probes would consistently capture DNA molecules from one strand, the damaged DNA, which is strongly associated with C-to-A and G-to-T mismatches due to the frequency of 8-oxoguanine damage, will make a strand-asymmetric contribution to the mismatch profile (Fig. 3.18 schematic below). The UK Biobank library preparation used the xGen Exome Research Panel v1.0, a single-strand DNA capture library with the probe nucleotide sequence oriented to the transcribed strand [333]. This observation may impact other exome sequencing data sets where a single-strand DNA molecule is consistently pulled down during exome capture. In addition to batch membership driving spurious associations in studies using the exome genotypes, by providing a measure of the degree of DNA damage, the C-to-A relative to G-to-T asymmetry may also be useful as a quality-control metric. Although this could represent the primary explanation of the asymmetry we observe, it does not rule out the possibility that some of the variation in the $\log_2$ ratios is explained by true somatic mutations.

Next, we investigated the heritable component of the somatic mutation rate. Understanding the impact of genetic variation on somatic mutation has implications for the aetiology of cancer and ageing in addition to increasing our understanding of the biological mechanisms that drive mutagenesis. In the first iteration of our analysis, four mismatch types had loci achieving genome-wide significance. One locus, within the *ZSWIM6* gene, was in LD with an eQTL for the *ERCC8* gene that encodes the Cockayne syndrome A (CSA) protein, a component of the transcription-coupled repair pathway. Cockayne syndrome is an ageing disorder in which the cells of the affected individual cannot identify or rectify pyrimidine dimers and bulky adducts in transcribed regions, consistent with the failure of the TC-NER response [334, 335]. This signal was consistent with our expectations and would be a remarkable result. Genetic variation acting on mutational burdens outside of rare, highly pathogenic mutations has not previously been identified [336]. Given the significance of this result, we tested the possibility that there may have been technical artefacts or germline variants that had made it through our SNP and genome masks.

When we removed the mismatches within the region containing the associated locus from the mismatch load, we found that the association signal was lost. The asso-

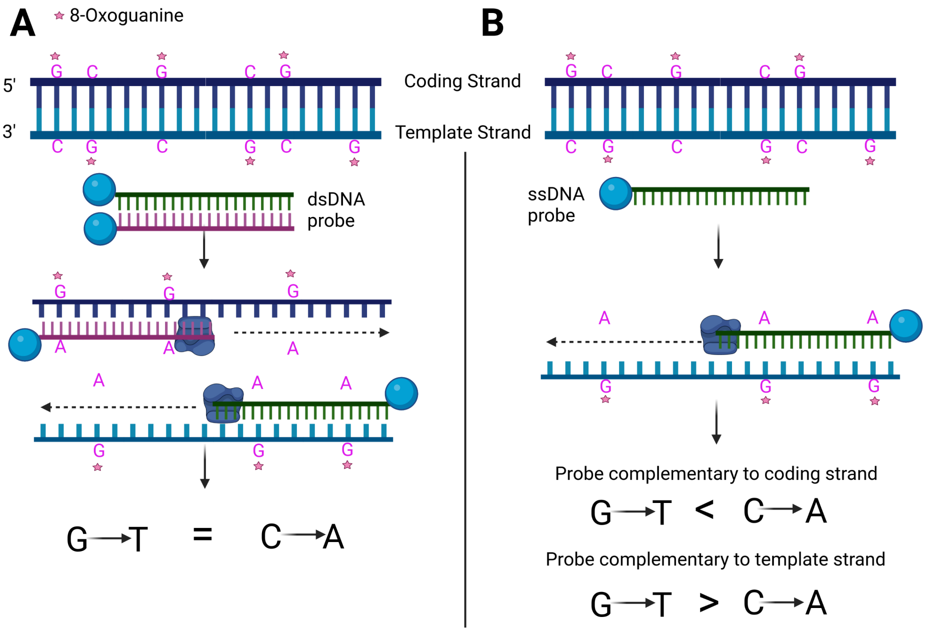

Figure 3.18: Schematic explaining how DNA damage can result in transcription-strand asymmetry. A. Double-stranded probes capture DNA damage on both the transcribed and nontranscribed strands. The mismatch profile in the resulting data, when aligned to the transcription strand, is symmetric. B. ssDNA/ssRNA probes capture DNA damage asymmetrically. Specifically, damaged G nucleotides (the most common type of DNA damage that arises after sample collection) base-pair with A, resulting in a C-to-A mutation on the coding strand and a G-to-T mutation on the template strand. When the mismatch profile is aligned to the transcription strand an imbalance in G-to-T relative to C-to-A is observed. Image created with BioRender.com

ciation signal appears to have resulted from an unusual germline variant that occurred in more than 25,000 samples (Chapter 2, Fig. 2.3). Interestingly, this mismatch only occurred on a maximum of 3 reads across all the 25,000 samples. What remains unclear is whether the variant was a true biological variant or a mapping artefact, potentially overlapping a large structural variant. Mismatches that arise from mapping errors within high-frequency structural variants correlate with the SNPs tested in the GWAS (pseudo-LD), leading to strong association signals [337]. To limit the impact of loci such as this, we applied a recurrence threshold to the mismatch load to remove mismatches that occur in more than 1000 samples. Future work could include the use of local ancestry information around candidate high-frequency mismatches to determine whether a site is likely to be germline (if a highly recurrent mismatch is due to a germline variant that has escaped filtering, we would expect that individuals containing the minor allele should be more closely related to one another within the genomic region in LD with the locus).

Since we did not find any signals of genetic variation impacting the mismatch rate, we stratified samples into those that were predicted to be Lynch-like using the genotypic data associated with the UK Biobank to determine whether the elevated rate of somatic mutation we expected to observe in those samples is detectable from the mismatch data. This could help to shed light on whether the mismatch data has the capacity to reveal germline variants associated with the somatic mutation rate. Lynch-like syndromes are defined by heritable mutations within the mismatch repair genes *MLH1, MSH2, MSH6, TATGAA* or *PMS2* and are the leading cause of heritable colorectal cancer [189]. These syndromes are characterised by microsatellite instability and increased somatic mutation burden [338]. We identified 160 individuals with Lynch-like syndromes based on the genotype data. Using non-negative least square regression, we estimated the contribution of the COSMIC MMR mutational signatures to the mismatch count data. We did not see any clear differences between the mean contribution of each mismatch repair signature to the Lynch compared to the non-Lynch groups. This may be due to several possibilities, such as a large variance across the Lynch group resulting from the small sample size (Fig. 3.16) or mutational signature bleeding, where the linear contribution of similar signatures is averaged, losing the true contribution of the relevant signature to each sample. The samples we identified as Lynch-like may also not exhibit a higher-than-normal mutation rate. Recent work has identified that some Lynch-like samples do not have an MMR deficiency phenotype [330]. As our lynch-like annotation is based on genotype data, a significant proportion of the predicted Lynch-like samples may not have a higher somatic mutation rate. It is therefore difficult to draw definitive conclusions from this result due to the small sample size of the Lynch-like participants and the possibility that we may not accurately stratify samples based on expected higher background mutation rates. We did however recover an age-dependent effect in SBS3 in the normal sample group indicating that SBS3 may accumulate with age. While SBS3 is associated with homology repair previous studies have shown that SBS3 accumulates linearly with age in paediatric tumours [339].This result would be particularly relevant in models of ageing as there is increasing evidence that non-homologous end-joining is preferred over homology-based repair pathways in replicatively aged cells [340, 341].

3.5 Conclusion

Here we investigated sources of variation in the mismatch load across 200,000 exome sequencing samples from the UK Biobank. We uncovered significant unreported batch effects in the sequence data. By decomposing the mismatch load into 192 transcriptionally aligned triplet loads, we found that for some triplet contexts, there was an increased linear relationship with age, suggesting a larger contribution of somatic mutations to these mismatch types. We also highlight the importance of addressing unreported sequencing batch effects within the UK Biobank by testing the mismatch loads between samples with self-reported cancer and smoking. We detected strong signals of asymmetry between mismatches on the transcribed vs untranscribed strand. The level of asymmetry was inconsistent with the expected contribution of somatic mutations to the observed mismatches. For the first time in the literature, we report a transcription strand asymmetry arising from a bias in the exome library preparation step.

We tested for signals of a heritable contribution to levels of mismatch within the UK Biobank by performing a GWAS on the mismatch loads. We find evidence of complex germline events which have not been reported in the dbSNP database or called in approximately 25,000 samples containing the low variant allele fraction variant. Lynch-like syndromes are characterised by aberrant mismatch repair processes we identified a set of 160 samples with pathogenic variants likely to cause Lynch-like syndromes. Using MMR mutational signatures, we estimated the relative contribution of 4 MMR signatures and two clocklike background signatures to predicted Lynch syndrome samples and the remaining normal group. No significant difference between the groups was observed. However, we did recover a strong signal of SBS3 accumulating with age in the normal group.

Overall, this chapter demonstrated that high levels of technical variation across samples make it difficult to obtain insights into inter-individual variation in somatic mutation processes through an analysis of the observed mismatches. Instead, in the next chapter we pivot the data, aggregating across samples to investigate variation in mismatches across the genome. By doing this, the impact of samples with high technical variance can be reduced. We will further analyse the mismatch data on the gene level by investigating the relationship between known somatic mutation rate modifiers and the median mismatch burden. We also test for signals of somatic selection and *de novo* mutational signatures active across the genome.

3.6 Methods

The context-specific transcriptionally-aligned mismatch loads were generated by reverse complementing mismatches that arose in genes annotated by VEP as being on the 'minus' strand. The data were normalised by the same regression-based procedure outlined in Chapter 2. Cancer and smoking data were downloaded from the UK Biobank repository using the field ID, 20001 and 22506, respectively.

Phenotypic correlations were performed using R (version 4.1.3) [342] and visualised using ggplot2 (version 3.4.0) [343]. Median TPM whole blood expression data were downloaded from the GTEx portal (version 8) [344]. The mismatch loads were regenerated using gene coordinates from genes non-zero expression levels in the GTEx whole blood and for genes without expression levels. The log base two ratios were then calculated for the median counts per mismatch context and its reverse complement. For mismatch loads derived from expressed genes vs non-expressed genes, the data was further normalised by the reference triplet frequency within each expression group.

Information and white papers for the IDT xGen Exome Research Panel v1.0 used by the UK Biobank for exome capture can be downloaded from the Wayback machine.

https://web.archive.org/web/20180403022641/http://eu.idtdna.com/pages/products/next-generation-sequencing/hybridization-capture/lockdown-panels/xgen-exome-research-panel

The coordinates and probe strand orientation can be downloaded from the following link.

http://sfvideo.blob.core.windows.net/sitefinity/docs/default-source/supplementary-product-info/xgen-exome-research-panel-probesbe255a1532796e2eaa53ff00001c1b3c.bed?sfvrsn=425c3407_7&download=true

Genome-wide association analyses were performed on 21 million genotypes filtered using PLINK2 with a minor allele frequency (MAF) threshold of 0.0001 and Hardy-Weinberg equilibrium (HWE) p-value cut-off of 1e-10 [345]. A pruned set of 1.3 million variants was generated by LD-pruning (PLINK2 –indep-pairwise params=1000,100,0.5) and a MAF of 0.01. This reduced set was then used to calculate principal components (PCs) and to generate a genetic relationship matrix (GRM). GWAS was performed using fastGWA (version = 1.93.2) with 10 PCs, age, sex, genotype batch and assessment centre as fixed effects [246]. The eQTL analyses of GWAS results were performed using the web-based LDlink resource [331].

The genotype call data was used to identify Lynch-like samples. Following the protocol outlined in *Patel et al.* [312], samples containing one of 52 likely pathogenic variants across six genes, *MLH1, MSH2, MSH6, PCSK9, PMS2* and *TATGAA* associated with Lynch syndromes were grouped as Lynch-like.

To identify the contribution of the mutational signatures to the UK Biobank mismatch data, we performed non-negative least squares regression using a set of 13 COSMIC signatures associated with DNA repair deficiencies or signatures that accumulate in a clock-like fashion over the human lifespan. Non-negative least squares regression was performed using the FitMS function of the signature.tools.lib R package (version 2.2.0) [112].

Table 3.4: List of COSMIC signatures related to DNA damage repair or increasing age. The proposed aetiology is provided.

Cosmic Signature	Proposed aetiology
SBS3	Defective homologous recombination-based DNA damage repair
SBS6	Defective DNA mismatch repair.
SBS9	Deficient DNA repair via translesion synthesis
SBS13	AID/APOBEC activity
SBS15	Defective DNA mismatch repair and microsatellite instability (MSI)
SBS18	Damage by reactive oxygen species - Clock-like
SBS20	Defective DNA mismatch repair.
SBS21	DNA mismatch repair deficiency
SBS26	Defective DNA mismatch repair.
SBS44	Defective DNA mismatch repair.
SBS84	AID activity
SBS1	Spontaneous deamination of 5-methylcytosine to thymine - Clock-like
SBS5	Aetiology unknown - Clock-like

Chapter 4

Variation in somatic mutation rate across the genome

4.1 Abstract

Intrinsic and extrinsic processes drive mutation rate variation. Through an analysis of the frequency with which mismatches to the reference genome occur across UK Biobank exome samples, we investigated the contribution of intrinsic properties such as GC content, replication timing, and recombination hotspots to variation in the rate of somatic mutation across the genome. Chromatin accessibility had a negative correlation linear relationship with mismatch frequency, whereas gene expression has a complex non-linear relationship with mismatch frequency. Many of the relationships between genomic properties and mismatch frequency are consistent with what has been reported for germline or somatic mutation data. Inferring mutational signatures using NMF, 5 of the 12 signatures has similarity to COSMIC signatures. We find evidence of positive selection in genes associated with clonal haematopoiesis of indeterminate potential. Overall, our results suggest that the exome sequencing data of 200,000 UK Biobank individuals is informative about variation in somatic mutation rate across the genome.

4.2 Introduction

The rate of somatic mutation across regions of the genome has been reported to be correlated with GC content, replication timing, chromatin accessibility and gene expression [127, 141, 133, 119]. Regions with high GC content have an increased somatic mutation burden due to the ineffective repair of spontaneously deaminated cytosines [346]. The GC content around single base errors has been shown to have a negligible effect on the rate of single base errors introduced during PCR amplification [347], suggesting that the observed correlations are derived from sources, other than sequencing error, such as somatic mutation, sequencing errors and DNA damage at guanine nucleotides during library preparation. GC-rich regions of the genome tend to be in open chromatin, confounding the relationship between mutation and GC content with DNA accessibility for replication and repair [127].

The genome can be broadly divided into regions of early and late DNA replication [140]. Open or active chromatin replicates earliest as the super-condensed chromatin in late-replicating DNA requires chromatin remodelling to allow replication initiation. As the nucleotide pool becomes depleted, the late replicating regions of DNA experience instability due to being in a ssDNA state for a prolonged period of time [142]. The relationship between ssDNA and DNA mutation rate is due to exposure to APOBEC activity [92] and other endogenous DNA damage factors, including alkylation, oxidation and deamination [65]. DNA damage on ssDNA also leads to DSBs which have been empirically shown to increase the somatic mutation rate around break sites [142]. In contrast to GC content, the mutational rate variation associated with replication timing shows a tissue-specific effect due to differences in chromatin conformation differences across cell types [143].

Genetic recombination during meiosis is a process found across vertebrates [348]. Sequences that have a high rate of recombination are referred to as recombination hotspots [349]. In addition to the recruitment of low-fidelity repair-associated polymerases to DSBs, there is an increase in APOBEC activity at regions of ssDNA [350] and consequently, recombination hotspots are associated with an elevated rate of germline mutations [351]. Recombination can also occur during mitosis, but the study and implications of mitotic recombination beyond loss of heterozygosity (LOH) remains poorly understood. In the TCGA dataset, approximately 15% of LOH events are attributed to recombination between homologous chromosomes [352].

The accessibility of DNA repair proteins to sites of DNA damage is also a known to influence the somatic mutation rate [133]. DNA that is in an accessible or in the 'open-chromatin' state is easily accessible by repair machinery, while DNA damage within an inaccessible or tightly wound state is less easily detectable by repair processes [133].

The chromatin conformation varies across cell types and stages of the cell cycle. Using sequencing-based methods such as ATAC-seq [136], the accessibility of DNA can be quantified and used as a measure of chromatin accessibility. As the chromatin accessibility increases, the sequence becomes more open to enzymatic repair, thus decreasing the mutation burden [133].

Gene expression and somatic mutation rate have a complex non-linear relationship [148]. Lowly expressed genes exhibit a higher mutation rate but as gene expression increases, the efficiency of the transcription-coupled repair machinery also increases until a point of inflection where genes that exhibit high gene expression lose the ability to repair mutated DNA base due to the mutagenic properties of transcription and the mutation rate once again increases. The complex relationship between mutation rate and expression has been previously observed in the soma and germline [119]. Transcription-coupled repair and transcription-associated damage are strand-specific processes resulting in an asymmetric mutational profile between the substitution rate on the transcribed and non-transcribed strands [143]. This bias is not clear when analysing mutation data aligned to the reference strand. A strong transcriptional-associated asymmetry has been previously described in mammalian evolution and cancer [122].

An increase in the rate of functional somatic mutations relative to the neutral expectation can be used to infer selection acting on somatic mutations, via methods derived originally from molecular evolution theory. These methods (referred to as dNdS or KaKs) determine if there is an excess or absence of functionally consequential mutations given the rate of synonymous mutations, which are assumed to be neutral [8]. An excess in the rate of functional mutations relative to the rate of non-synonymous mutations is indicative of positive selection. In contrast, an absence of functional mutations relative to the proportion of synonymous mutations is indicative of negative or purifying selection. *Martincorena et al.* have developed a 96-context and 192-context somatic mutation dNdS model (dndscv) which accounts for background genomic covariates including chromatin state, gene expression and for the differences in mutation rate across genes [13]. In expanding clones, mutations that influence clonal growth along with non-functional passenger mutations will be found at high frequencies within a tumour. By aggregating across samples only the functional mutations will be highly recurrent across cancer samples, for example in the TCGA dataset driver mutations in *BRAF* were found in 848 samples with 64% of *BRAF* driver mutations corresponding to V600E [353]. Haematological clonal expansions have been previously shown to increase with age and to be due to a relatively small number of positively selected mutations with varying effects on the expansion of haematological cells [354].

As we age, clonal expansions begin to dominate the haematopoietic stem cell (HSC) pool, resulting in a loss of stem cell diversity [209]. Clonal haematopoiesis of indeterminate potential (CHIP) results from the clonal expansion of one or more lineages of HSCs. CHIP is a prognostic marker for numerous diseases, such as cardiovascular disease and inflammation and it increases the risk of leukaemia by 0.5-1.0% per year [208]. CHIP is most commonly associated with mutations in epigenetic regulators (*DNMT3A, TET2, ASXL1*) but also in genes involved in mRNA splicing (*PRPF8, SF3B1, SRSF2, U2AF1*) [355]. A single base substitution, *DMNT3A* R882, is the most common mutation found in individuals with CHIP [354]. Between 10 and 20% of individuals over the age of 70 have clones containing established CHIP expansion mutations [355]. Although CHIP is common in older individuals, the clonal expansions can begin much earlier in life. Interestingly, mutations have different fitness effects depending on the age in life at which the mutation occurs [214].

In a dataset containing somatic mutations across samples, the underlying mutational processes are aggregated into a single spectrum of mutation frequencies. Matrix decomposition techniques, such as non-negative factorisation, have been used to factorise a matrix of triplet mutation counts into two matrices, one of which contains the mutational spectrum of each mutational process and the other containing the relative contribution of each process to the mutational burden in each sample [84, 103]. Using controlled mutagenesis studies, the mutational signatures can be mapped to certain mutagenic processes, such as APOBEC activity and clock-like mutational signatures [106]. The COSMIC database contains a curated list of mutational signatures inferred from cancer datasets, including proposed aetiologies [113].

In the Chapter 3, we found that the proportion of somatic mutations among the mismatches in the UK Biobank exome sequencing was likely to be too low to allow these data to be used to study sources of inter-individual variation in somatic mutation load. In this chapter, instead of looking at variation between individuals, we consider variation across the genome. Although the proportion of true somatic mutations in the data is likely to be low, leveraging information across many individuals has the potential to identify genomic features that correlate with mutation burden and to highlight loci that are recurrently mutated across samples, providing the possibility of identifying evidence of somatic selection. By learning information from recurrent mutations across the genome we can begin to recover true sources of somatic signal from the noise of whole exome sequencing data. By pivoting to intra-genomic variation rather than inter-individual variation we can mitigate technical bias. Leveraging the sources of known mutation rate covariation, and the power of the sample size in the UK Biobank we can extract information on the somatic mutation burden across genes even when there are large sources of technical noise in the data.

4.3 Results

4.3.1 Genomic covariates

4.3.1.1 GC content

We used the number of samples in which a mismatch was observed at a locus in the UK Biobank data to study variation in somatic mutation rate across loci, taking account of the sequence context of the locus via the upstream and downstream nucleotides. The mismatch rate was strongly correlated with gene GC content, a well-established source of mutation rate variation in the genome [127]. Across 192 sequence contexts, 190 had a significant Spearman correlation coefficients after multiple testing correction ($\alpha < 0.00026$), with Spearman correlation coefficients ranging from 0.03-0.35 (Table B1). Nine of the 192 mutational contexts had significant Spearman correlation coefficients below machine precision ($P < 2.2 \times 10^{-308}$; Table B1 & Fig. 4.1). There was no obvious pattern to the sequence contexts most strongly correlated with GC content.

To aggregate the 192 sequence contexts into a single value, we computed the mean across the ranks of the recurrence values (see Methods for details). Although a strong correlation was found between the ranked median score and GC content ($P < 2.2 \times 10^{-308}$, $\rho = 0.51$), the aggregation of median recurrences into a single score may induce bias in genes with missing mutational contexts (typically shorter genes), for this reason, we focus our results primarily on the results obtained for the individual sequence contexts. The relationship between GC content and mismatch recurrence underscores the importance of accounting for this potential confounding as GC content is a known to covary with gene expression and replication timing.

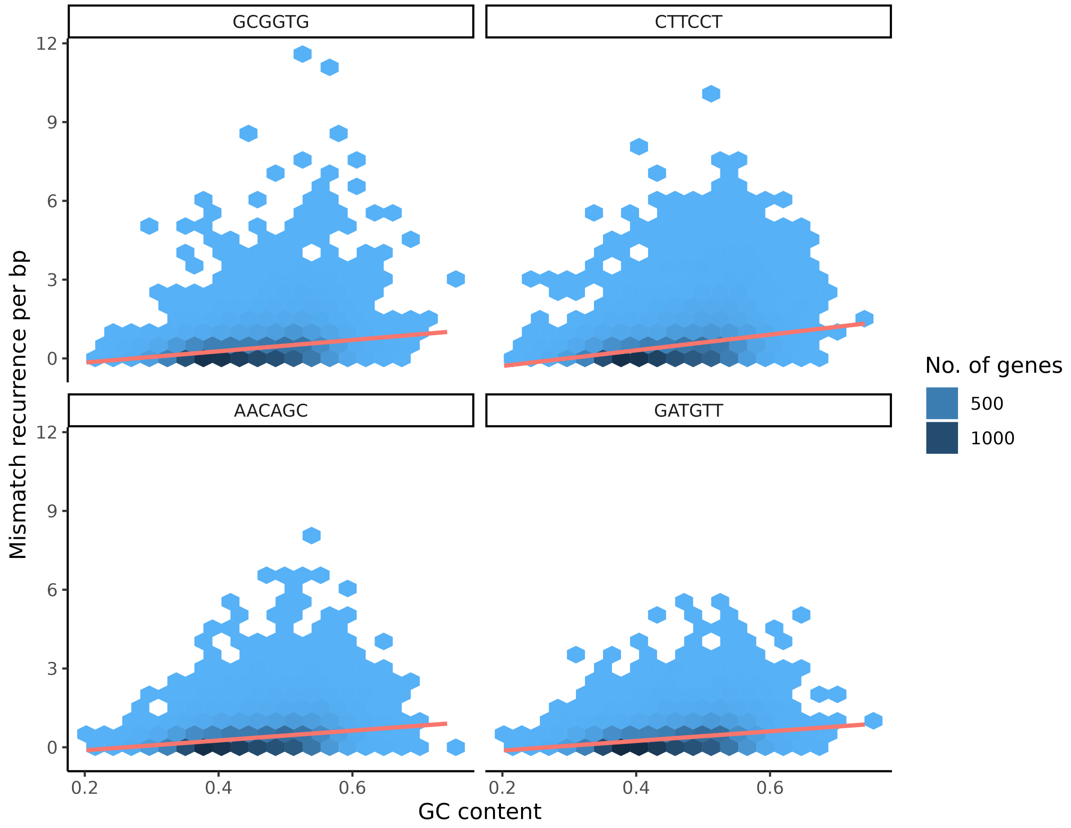

Figure 4.1: Hexbin plot of the median mismatch recurrence as a function of GC content for the GAG-to-GTG, TCT-to-TGT, AAC-to-AGC and GAT-to-GTT mismatch contexts. The red line shows the line fit from a linear model fitting the median mismatch recurrence per bp against the genes GC-content.

4.3.1.2 Replication timing

The relationship between GC content and replication timing is well established (Fig. 4.2) [356]. After accounting for confounding between GC content and replication timing, no linear relationship between replication timing and ranked aggregated score (across triplet sequence contexts) was observed ($P = 0.17$). This, however, was biased by short genes, filtering genes less than 500bp in length indicated a positive relationship between the aggregated score and replication timing (Spearman $P = 8.007 \times 10^{-45}$, $\rho = 0.11$; Fig. B1). In some sequence contexts, strong linear signals after accounting for GC content were identified (Fig. 4.3). For example, GAG-to-GTG (linear model $P = 1.5 \times 10^{-18}$; Spearman $P = 3.2 \times 10^{-182}$, $\rho = 0.22$). Of the 193 mutational loads (192 contexts + ranked aggregated score) tested 143 mismatch loads had significant spearman correlations with replication timing after multiple testing correction. Of those, only 16 had significant negative Spearman correlations after correction for multiple testing. TCT-to-TGT, the mismatch context with the second highest Spearman correlation (Table B1; $P = 4.9 \times 10^{-143}$; Spearman $\rho = 0.19$) is a high-weight context in cosmic signature SBS13. The activity of this mutation signature has been reported to correlate positively with replication timing across 17 cancer types in the COSMIC database.

4.3.1.3 Open chromatin

To understand the relationship between variation in somatic mutation rate and the accessibility of the DNA sequence, we used ATAC-seq data from a normal B-lymphoblastoid cell line (see Methods). The ATAC-seq data measures the accessibility of chromatin and is used as a proxy for sequences that are accessible to repair machinery. We hypothesized that genes that are accessible to repair machinery will have a reduced somatic mutation rate, which would be detectable as a reduction in the mismatch recurrence in open chromatin regions [141]. Indeed, all mutational contexts had a negative Spearman correlation coefficient with chromatin accessibility, with the aggregated rank score having the strongest linear relationship (Fig. 4.4).

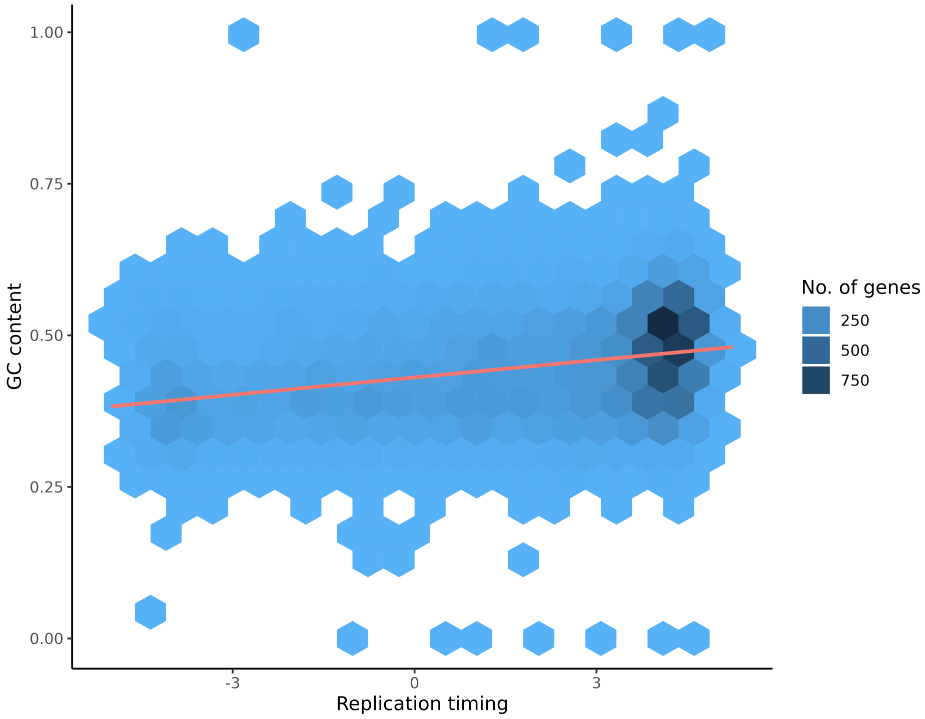

Figure 4.2: Hexbin plot highlighting the relationship between GC content and replication timing. The red line shows the line fit from a linear model between GC content and replication timing.

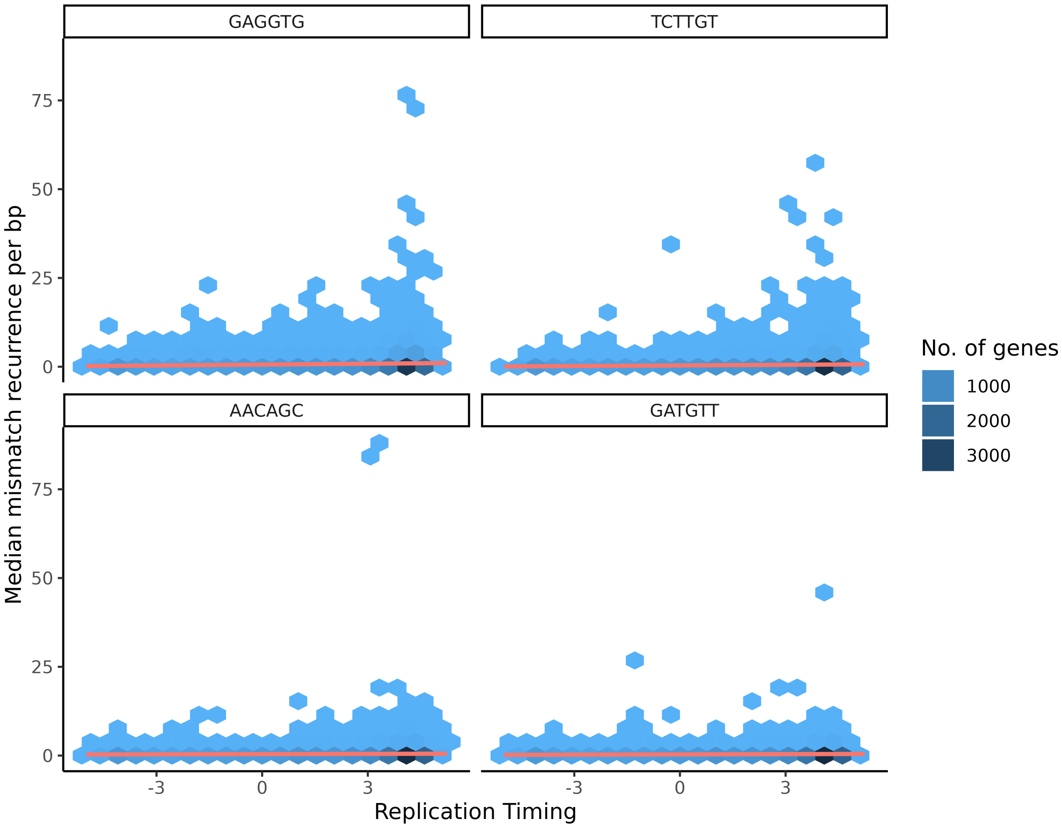

Figure 4.3: Hexbin plot of the median mutation recurrence as a function of replication timing for the GAG-to-GTG, TCT-to-TGT, AAC-to-AGC and GAT-to-GTT mismatch contexts. The red line shows the line fit from a linear model between each genes median mismatch recurrence per bp as a function of replication timing.

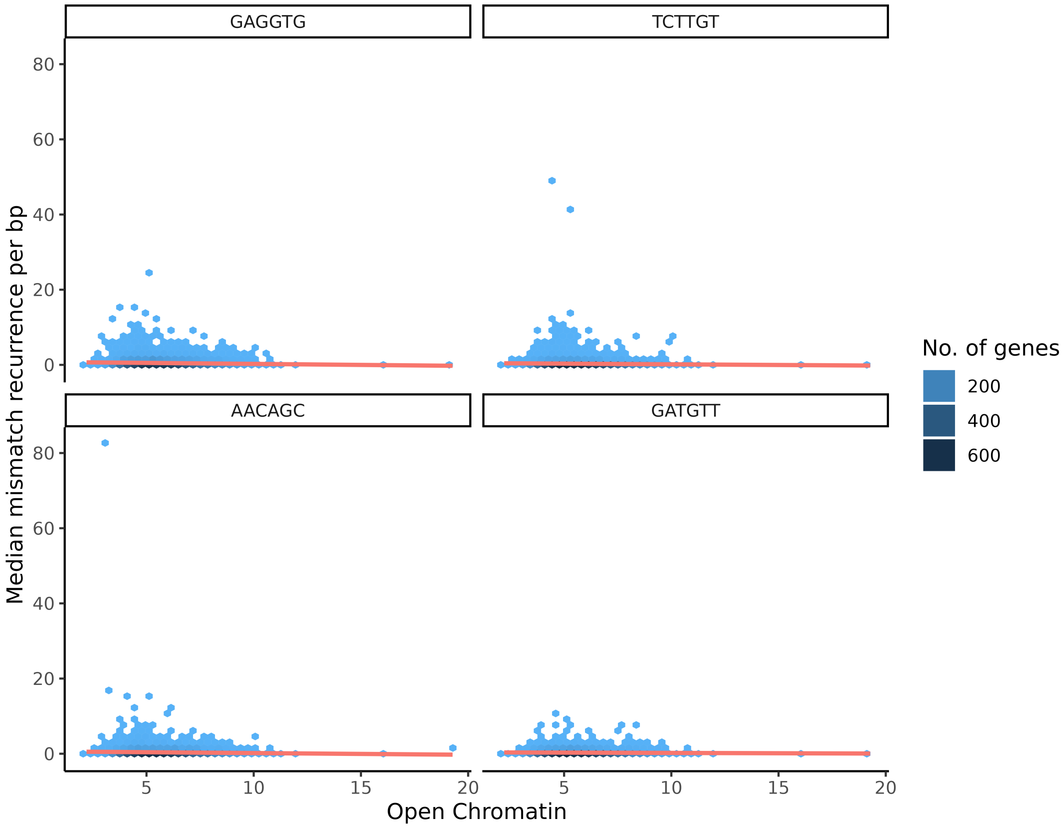

Figure 4.4: Hexbin plots. The median mutation recurrence as a function of chromatin openness for the GAG-to-GTG, TCT-to-TGT, AAC-to-AGC and GAT-to-GTT mismatch contexts. The red line shows the line fit from a linear model between a genes median mismatch recurrence per bp and chromatin accessibility.

4.3.1.4 Recombination hotspots

Genetic recombination may influence somatic mutation. To investigate the effect of recombination on the mismatch data, genes were classified according to whether they overlap recombination hotspots. The mismatch proportion of mismatches overlapping recombination hotspots (hot genes) was significantly greater than genes not overlapping recombination sites (cold genes) (t-test $P = 1.373 \times 10^{-12}$). To remove confounding from GC content we calculated genome- and exome-wide estimates of GC content and the GC content in recombination hotspots (Table B2). The GC content is significantly lower in recombination hotspots, compared to the remainder of the genome. To account for this sequence content differences between hot and cold genes we can model the mutational load as a function of recombination 'hotness', chromatin openness and GC content. In total 87 mutational contexts had a significant model fit after multiple testing correction (Table B3). A positive linear relationship between the proportion of sites within a gene overlapping a recombination hotspot and the total mutational load per gene was observed, consistent with recombination increasing the number of mismatches (Fig. 4.5).

4.3.1.5 Gene expression

4.3.1.5.1 Linear models

Gene expression is a well-established modifier of mutational rate. In a linear model of mismatch recurrence as a function of median gene expression (from GTEx Whole Blood) and other correlates of somatic mutation (replication timing and GC content), a significant negative trend was observed (Table B4). A negative relationship was observed for most mutational contexts (Fig. 4.6). The aggregated ranked mutation score per gene showed the strongest association with expression (effect size = -0.01, $P = 3.06 \times 10^{-23}$; R^2= 0.16). We also observed a strong negative linear relationship between gene expression and the individual mutational contexts. In particular, the GCT-to-GAT mutation context had the strongest statistical significance (effect size = -0.07, $P = 1.3 \times 10^{-10}$; $R^2 = 0.004$, Fig. 4.6). In total 18 mutational contexts had significant coefficients for gene expression after multiple testing correction, with all 18 having negative model coefficients.

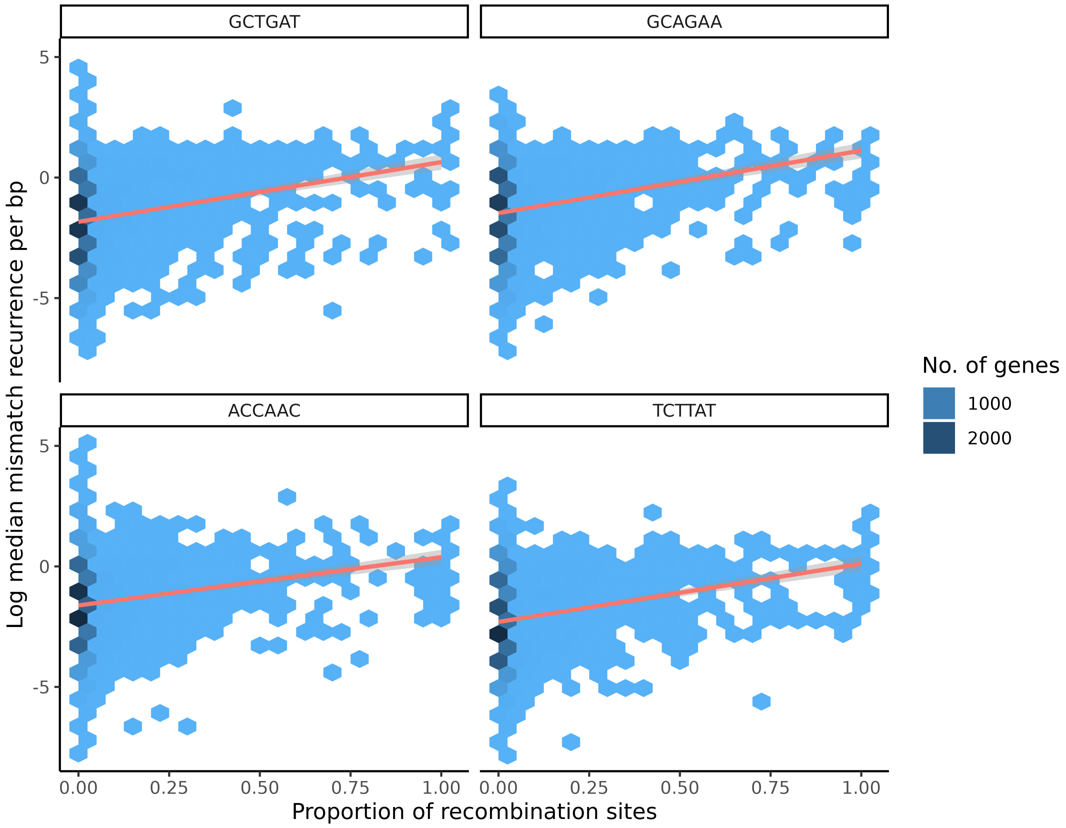

Figure 4.5: Hexbin plots of median mutation recurrence vs proportion of sites overlapping a recombination hotspot. The GCT-to-GAT, GCA-to-GAA, ACC-to-AAC and TCT-to-TAT mismatch contexts are shown as representative of the 192 mismatch contexts. The red line shows the line fit from a linear model between a genes median mismatch recurrence per bp and the proportion of a genes bases that lay within a recombination hotspot.

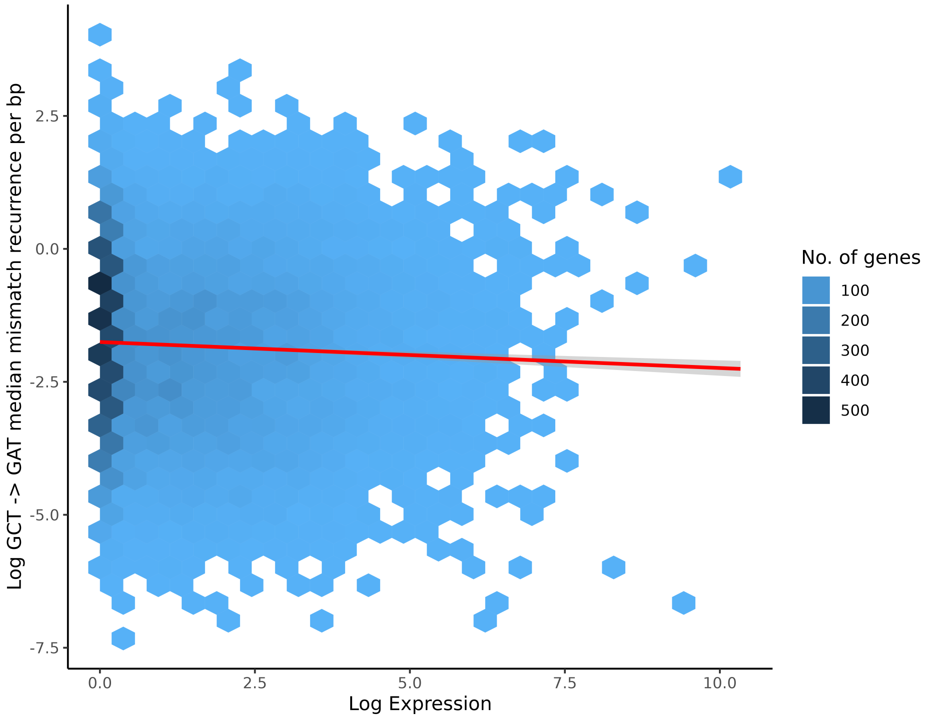

Figure 4.6: Hexbin plot of the *log* GCT-to-GAT mutation recurrence and *log* gene expression. The red line shows the line fit from a linear model of a genes median mismatch recurrence per bp and *log* gene expression.

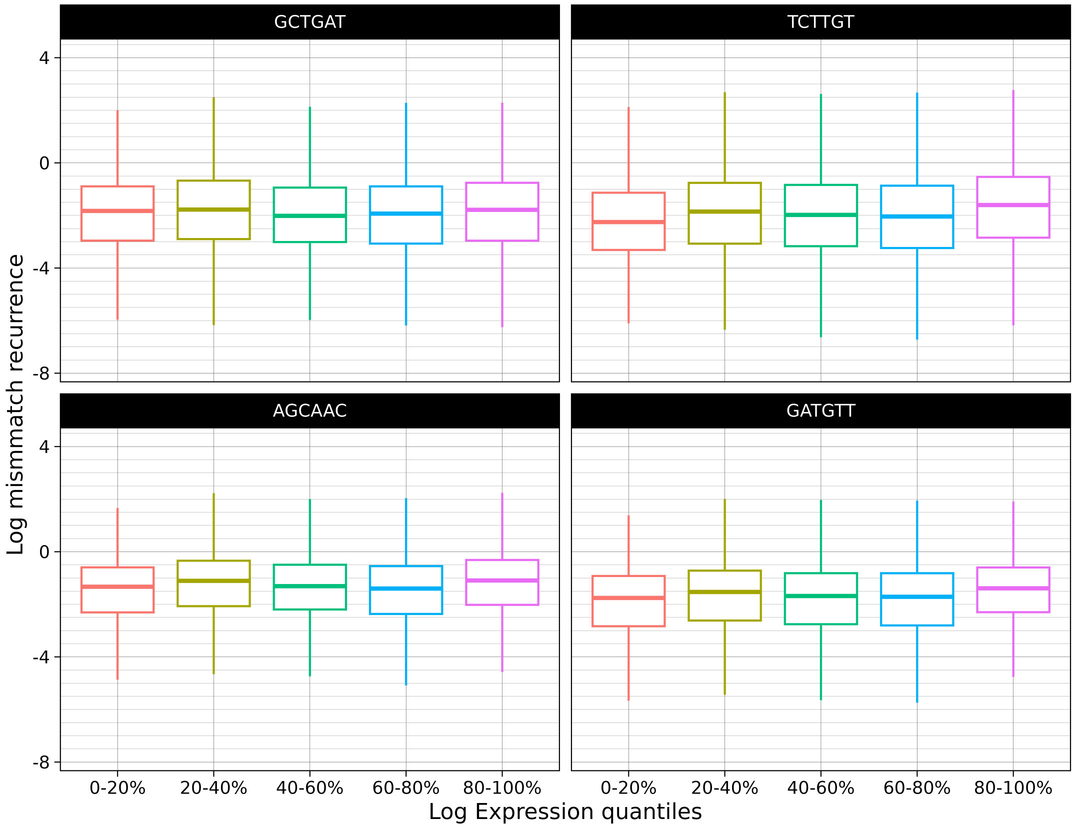

Figure 4.7: Boxplots of *log* median mismatch recurrence and *log* gene expression quintiles for the GCT-to-GAT, TCT-to-TGT, AGC-to-AAC and GAT-to-GTT mismatch contexts.

By analysing the relationship between mismatch recurrence and gene expression quantiles, a more nuanced relationship emerged. There was a weak parabolic relationship between the median recurrence as a function of gene expression quintiles, with the upper and lower quantiles having an increased rate of putative somatic mutations (as measured by the median number of individuals in which the site contained a putative somatic mutation) and the lowest rate at intermediate expression levels. This is consistent with the opposing mutational forces of transcription coupled repair, which does not act effectively on the lowest expressed genes, and transcription associated mutagenesis, leading to an elevated burden of putative somatic mutation for the highest expressed genes (Fig. 4.7) [148].

4.3.1.5.2 GAMs (Generalised additive models)

As opposing processes underlying transcriptional associated mutagenesis and repair are likely to result in a non-linear relationship between somatic mutation and gene expression, we used generalised additive models (GAMs) and smoothing splines to visualize the relationship (Fig. 4.8). Given large sample sizes such as that of the UK Biobank, GAMs can reveal the non-linear relationships between variables. The GAM model shows a higher rate of mismatches for lowly expressed genes. A similar trend was seen in the GAM and expression quantile analysis. As expression increases the rate of mismatches decreases until a certain expression point where the recurrence begins to increase once more, until the data becomes sparse and dominated by uncertainty. The aggregated rank score shows a similar trend.

4.3.2 Analysis of selection

The selective pressure acting on a gene over evolutionary time is frequently assessed through a comparison of the relative rate of non-synonymous substitutions per nonsynonymous site to synonymous substitutions per synonymous site (referred to as dNdS) [156]. A similar approach has been applied to infer evidence of somatic selection in cancer and normal cells [13]. Optimising an approach developed for this purpose, (see Methods section), we found a strong correlation between the strength of the evidence for selection and CDS length, ($P < 2.2 \times 10^{-16}$, Pearson's ρ= -0.77). This is not surprising, as evidence of selection (purifying or negative selection) can accumulate over the length of the gene. To account for this bias the *log* of the p-value was divided by the CDS length. To test whether genes that drive blood malignancies are enriched for positive selection in the UK Biobank, we ordered genes according to the evidence of positive selection. Using a list of 75 genes from the Archer variantPlex 75 Myeloid gene panel we produced a receiver operator characteristic (ROC) curve illustrating evidence for positive selection to identify genes in the myeloid panel (Fig. 4.9; Table B5). A ROC curve represents the true positive rate as a function of the false negative rate. The area under the ROC curve (AUC) is a measure of the models performance in discriminating between haematological and non-haematological genes. An AUC of 0.5 reflects the model has a 50% chance of a correct prediction while a AUC of 1 re-

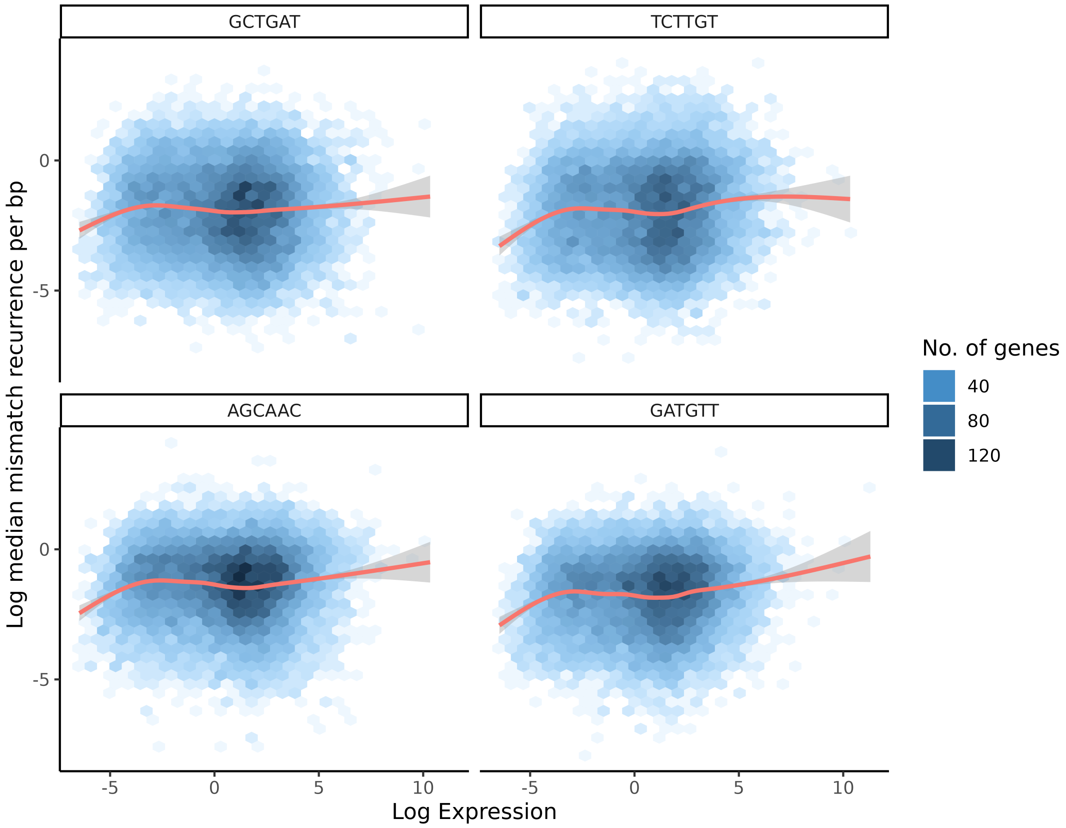

Figure 4.8: Generalised additive model (GAM) of the *log* mismatch median recurrence and *log* gene expression. The datapoints are shown as hexbins to highlight the trends revealed by the GAM and density of the mismatch recurrence.

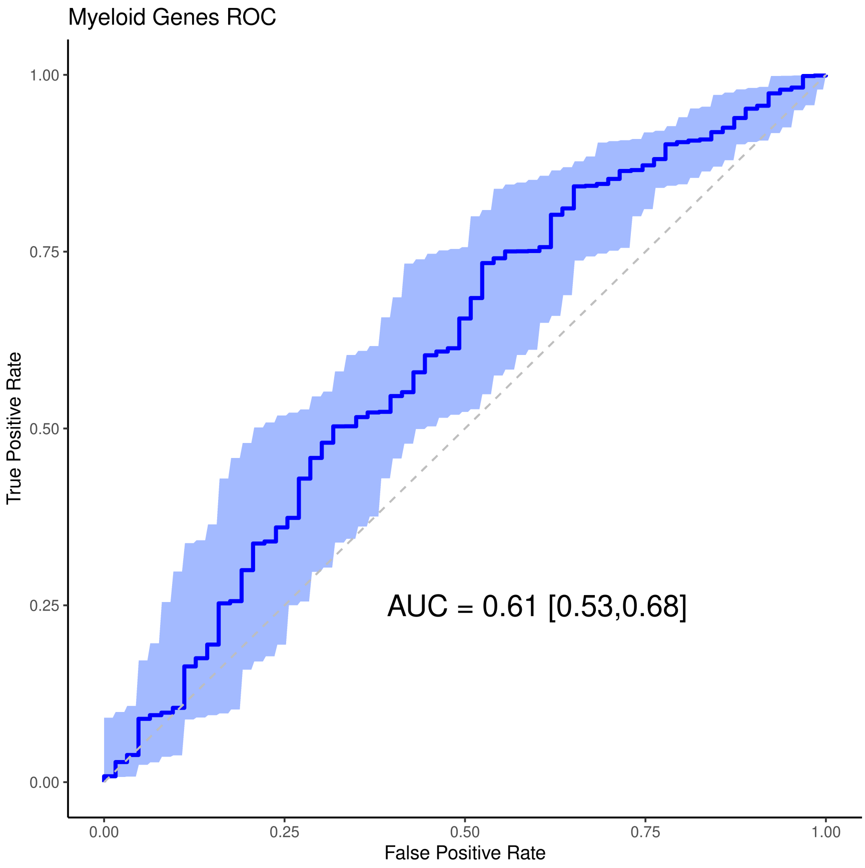

Figure 4.9: ROC curve showing the enrichment of myeloid genes over 1000 bootstraps. The p-value have been adjusted for myeloid gene length.

flects 100% predictions. The AUC of the dnds model was 0.61 (CI95%= 0.53-0.68; confidence interval generated from 1000 bootstraps), suggesting that positive selection had weak (but significantly better than random) power to identify genes in the myeloid panel.

The top 20 genes ordered on adjusted p-value were strongly enriched for histone genes and genes involved in cytoskeleton pathways, Table 4.1. There was considerable evidence of positive selection acting on *SRSF2*, a gene that is known to contribute to CHIP [355] and is critical in driving myelodysplastic syndromes [357]. *SRSF2* had over twice the expected rate of non-synonymous mutations, given the observed number of synonymous mutations (omega = 2.323, Table 4.1) [358].

Table 4.1: Top 20 positively selected genes ranked on adjusted p-value.

Gene name	N-syn[1]	N-mis[2]	ω-mis	P-value[3]	CDS length	P-adj[4]
TUBB4B	32942	428496	3.732637	-38852.777	1338	-29.03795
ACTG1	12413	247709	5.911971	-32725.678	1128	-29.01213
HIST1H2AC	5677	68983	4.737266	-9848.482	393	-25.05975
H3F3B	10573	95847	3.379037	-9710.978	411	-23.62768
ACTB	27712	313482	3.278052	-25026.961	1128	-22.18702
CFL1	22536	202790	2.752010	-13572.942	615	-22.06982
HIST1H4E	2873	39088	5.792993	-6866.701	312	-22.00866
EEF1A2	46941	460770	2.815484	-30195.347	1392	-21.69206
HIST1H2AE	9801	81509	3.252676	-8219.057	393	-20.91363
TXNL4A	16552	164724	2.562580	-8606.809	429	-20.06249
ATP6V0C	22376	167452	2.375905	-9172.920	468	-19.60026
ARF1	28020	214048	2.301080	-10612.815	546	-19.43739
AC011530.1	8794	84857	2.853683	-5811.733	321	-18.10509
RHOB	18769	178607	2.667427	-10640.051	591	-18.00347
EEF2	72711	685477	2.790245	-45525.398	2577	-17.66605
ACTA1	29894	313731	2.816091	-19460.442	1134	-17.16088
ARL4C	20639	186946	2.541988	-10359.423	606	-17.09476
SRSF2	30146	189634	2.322876	-11244.955	666	-16.88432
RHOG	32160	206185	2.144864	-9702.280	576	-16.84424
ARF6	23010	168462	2.313264	-8765.738	528	-16.60178

[1] Number of synonymous mismatches

[2] Number of missense mismatches

[3] Natural logarithm of p-value

[4] CDS adjusted p-value

4.3.3 Mutational signatures

Matrix factorization techniques, such as NMF, can be used to decompose a matrix of mutation counts into a latent feature space capturing the underlying mutational processes and technical artefacts and their relative weights across samples. Given the size of the UK Biobank dataset, the RcppML implementation of NMF was used for speed and memory efficiency [359]. Our count matrix consisted of 96 triplet contexts on the columns and 17,435 genes on the rows. While the body of this work deals with 192 contexts in this analysis we have opted for 96 contexts to easily compare with the publicly available COSMIC mutational signatures. The rank of the decomposition determines the number of latent mutational processes. To determine the rank that best captures the latent processes we use the elbow point method on mean squared error over a range of ranks, 1-60. As the plot of the ranks against the MSE did not show an obvious elbow point, to use in the matrix factorisation, the unit invariant knee (UIK) method of the inflection R package was used to infer an 'optimal' rank [360]. The UIK method indicated an optimal rank of 12 for the factorization (Fig. 4.10).

To aid interpretation, the 12 estimated mutational signatures were mapped to known mutational signatures derived from version 3.2 of catalogue of somatic mutations in cancer (COSMIC), (Fig 4.11). Five of the 12 estimated signatures had cosine similarities of greater than 0.75. These were COSMIC signatures SBS1, SBS10b, SBS15, SBS49 and SBS90.

SBS1 is a clock-like mutational signature which is intrinsic to normal and cancer cells, with a proposed aetiology of spontaneous deamination of 5-methylcytosine. It has been proposed as a cell division/mitotic clock. SBS10b is found in hypermutated cancers resulting from POLE deficiency. SBS15 is associated with defective DNA repair and MSI instability and typically occurs in samples with other signatures of MSI instability. SBS49 is proposed to be a sequencing artefact. SBS90 has been proposed to arise from exposure to the alkylating agent duocarmycin. Duocarmycin is derived from the soil-dwelling bacteria streptomyces, found in the human microbiome [361]. For estimated signatures, SBS1 and SBS15, the relationship to the cosmic signatures are influenced by a single mutational context that may over-inflate the cosine similarity (Fig. 4.12).

Over multiple runs of the same factorisation rank the NMF algorithm may not converge on the same latent vectors (or the order of the latent features that appear in the

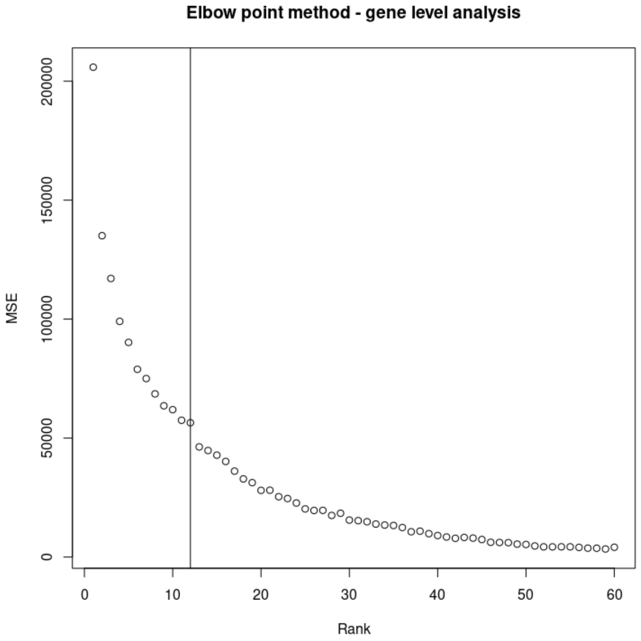

Figure 4.10: Elbow point plot using the UIK method to determine the optimum factorization rank among all ranks in the range 1 to 60. A vertical line indicates the optimal factorization rank (12) inferred using this approach. The mean squared error between the input counts matrix and the dot product of the inferred basis and coefficient matrices (y-axis).

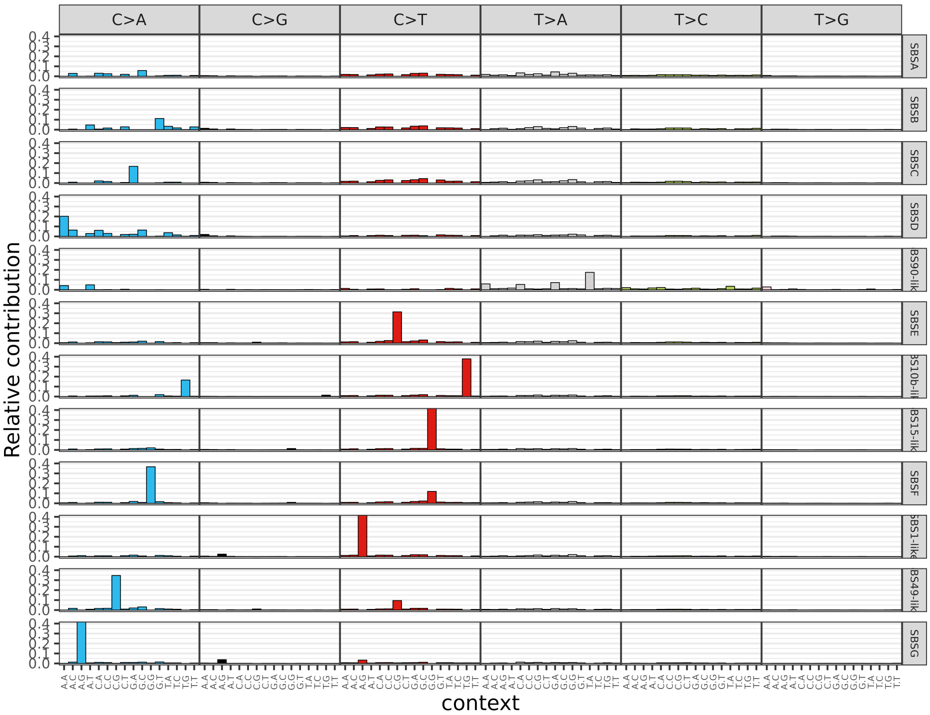

Figure 4.11: Mismatch spectra of the ‘optimum’ signatures. The bars measure the relative contribution of each triplet mismatch to the inferred signature y-axis. Signatures with similarity to COSMIC signatures have been renamed, right legend.

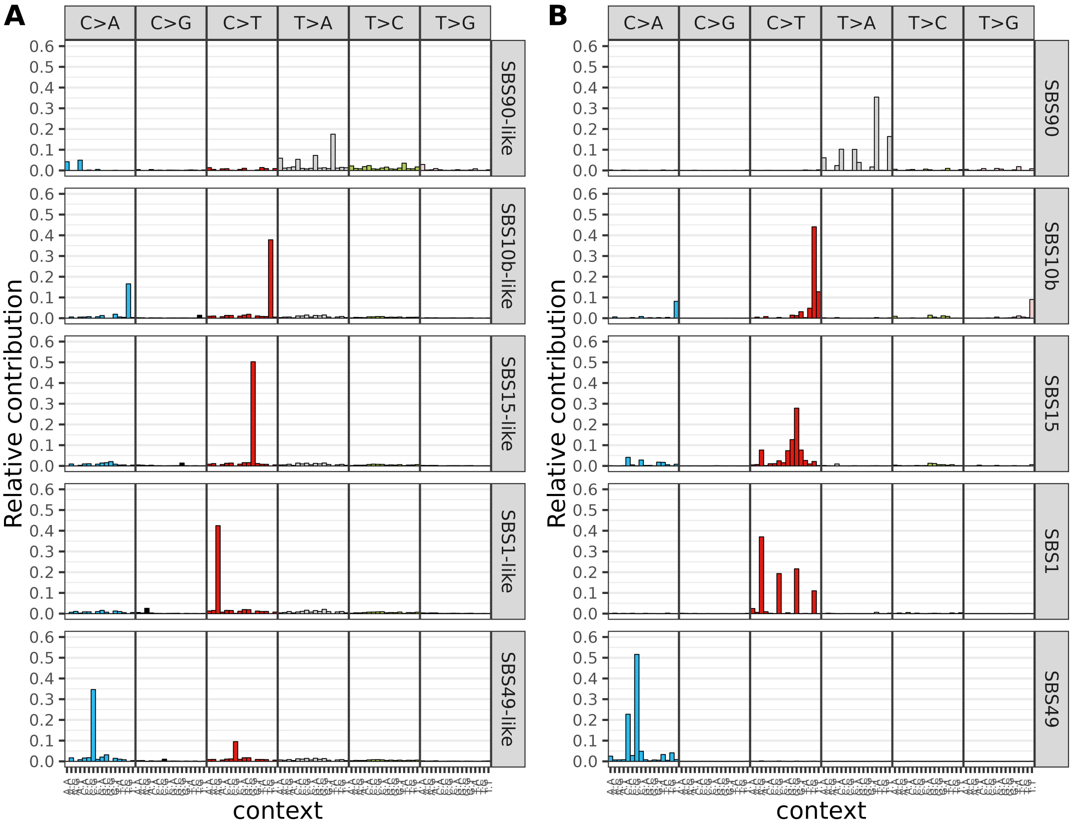

Figure 4.12: A. COSMIC-like estimated signatures from the mismatch data. B. The closest COSMIC signature to the estimated signature on the same row.

basis matrix may differ between runs). To estimate the stability of the rank 12 approximation we repeated the factorization 250 times. The variance across genes in the contribution (H) matrix of the first four estimated signatures was constant whereas for estimated signatures 5 - 12 the variance increases dramatically (Fig 4.13 A.) Taking the mean of the D matrix (diagonal scaling matrix) from the 250 NMF runs allows us to see that the first four signatures contribute the largest proportion of variation (Fig. 4.13 B.) A finding that is consistent with the variance explained (Fig 4.13 A.) The cosine pairwise similarity between each signature estimated from the 250 NMF bootstraps indicates that estimated signatures, 1, 2, 3 & 4, have the largest stability (Fig. 4.14). This reflects the shape of Fig. 4.13 A where the position of the higher number of signatures becomes increasingly unstable. The variance in the contribution of the inferred signatures and the cosine similarity between the signature replicates indicates the instability in the estimation of higher signatures (eSig5-12) which is consistent with the slight drop in the MSE from the optimum rank analyses.

The signature contributions were all significantly correlated with chromatin openness and replication timing (Table 4.2). In linear regression models that included replication timing, GC content and chromatin accessibility as fixed effect covariates, all signatures except the SBS90-like signature, had a significant association (after multiple testing correction) with replication timing and chromatin accessibility. The inclusion of gene expression to the models made only a minor difference to the total variance explained R^2 across all signatures.

4.3.4 Strand asymmetry

In Chapter 3, we quantified strand asymmetry per individual. We proposed that the asymmetry in the number of mutations observed on the transcribed versus non-transcribed strand was due to the use of ssDNA probes in the exome capture assay. The use of probes that are designed to be complementary to the coding sequence resulted in signals of DNA damage (specifically G to T mutations) being found at a higher rate on the non-coding strand. We reasoned that this effect can also be detected in the $\log_2$ ratio of the mutation recurrence between transcription strands across the genome (Fig. 4.15). While an excess of C-to-A mutations is indicative of transcription-associated mutagenesis [123], the strong asymmetry in C:G-to-A:T mutations that we observe in the

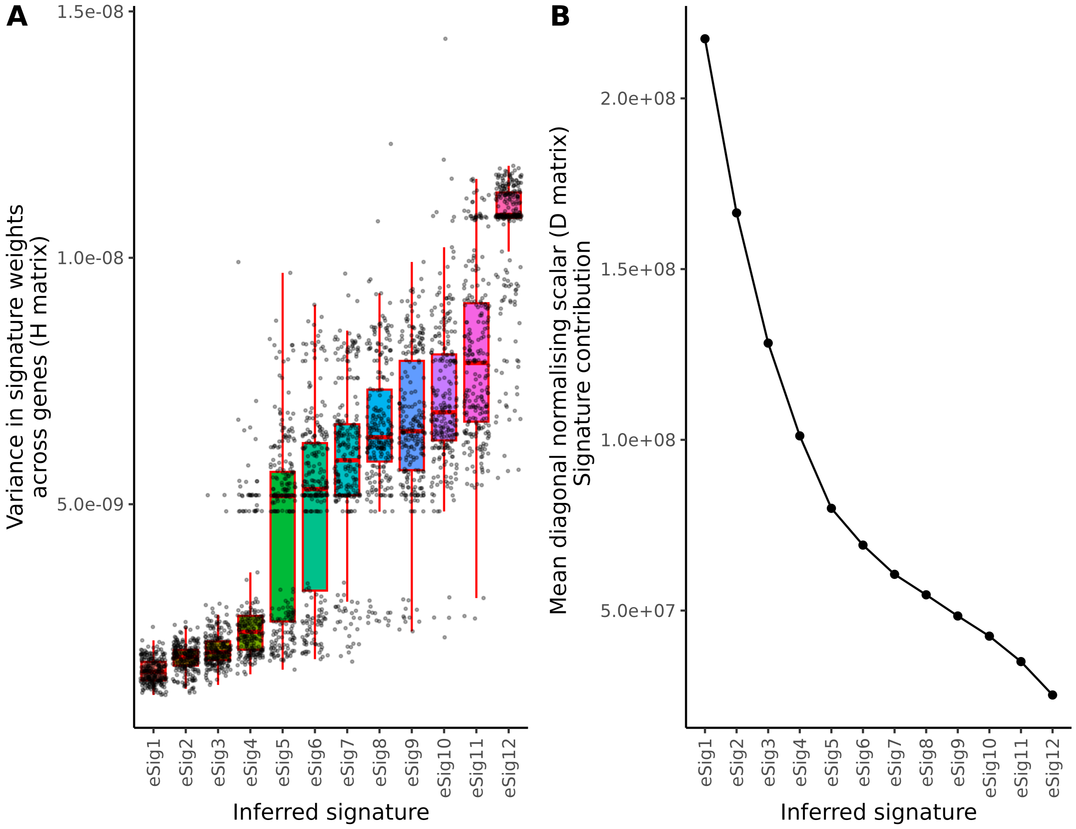

Figure 4.13: 250 NMF bootstraps of optimal rank 12 to assess the stability of the approximation. A) The variance in signature weights across the genome. B) The mean scaling factor for each estimated signature.

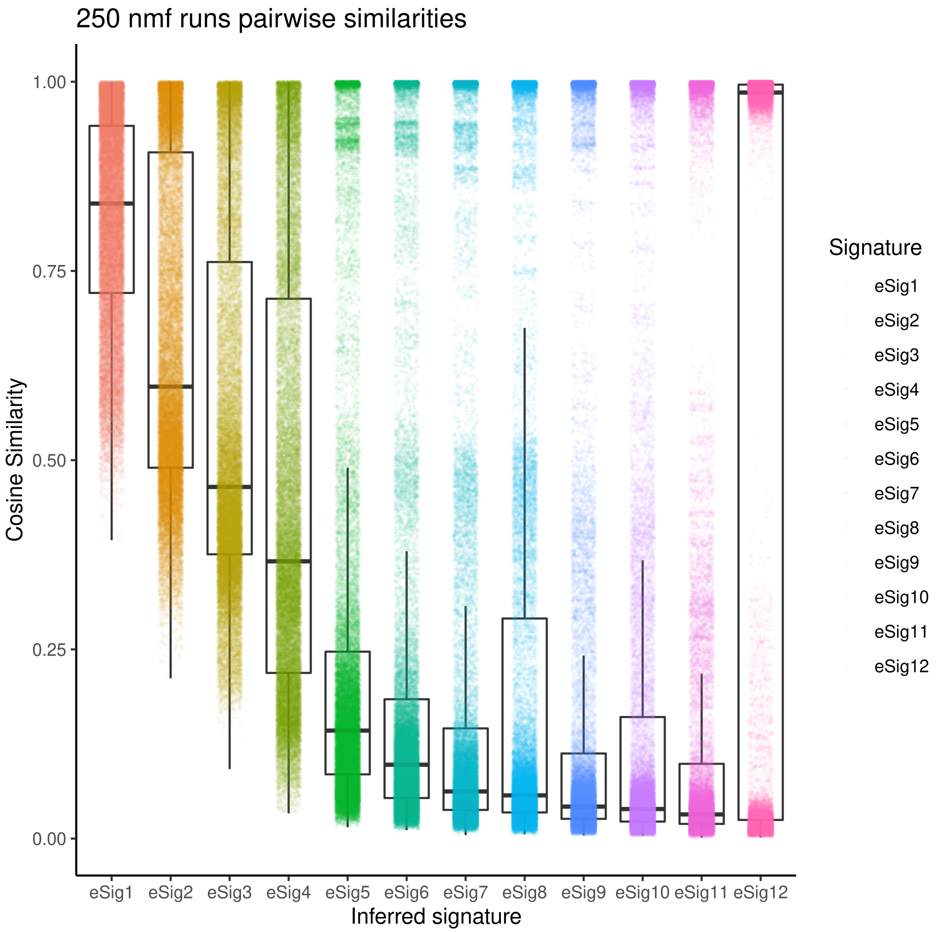

Figure 4.14: Pairwise comparisons using the cosine similarity in 250 NMF runs with factorization rank 12. Each boxplot contains 31,125 data points $\binom{250}{2}$.

Table 4.2: Spearman correlation tests between the inferred signatures and gene expression, replication timing and chromatin accessibility.

	Gene expr P	Gene expr ρ	RepT P	RepT ρ	Chromatin P	Chromatin ρ
SBSE	1.2e-98	0.16	4.6e-28	0.084	2.2e-11	-0.077
SBSF	2.3e-77	0.14	1.9e-30	0.088	4.4e-06	-0.053
SBS15-like	5.9e-74	0.14	1.3e-34	0.094	1.4e-09	-0.069
SBSB	1.3e-58	-0.12	8.9e-100	0.16	3.1e-14	-0.087
SBS49-like	2.1e-53	0.12	2.7e-13	0.056	3.3e-11	-0.076
SBS10b-like	3.1e-53	0.12	1.5e-16	0.063	7.3e-09	-0.066
SBS1-like	2.7e-52	0.12	1.1e-52	0.12	4e-09	-0.067
SBSA	9.8e-52	-0.12	1.5e-49	0.11	4.8e-09	-0.067
SBSG	9.8e-27	0.083	1.6e-08	0.043	8.7e-05	-0.045
SBSD	1.3e-15	-0.062	4e-28	0.084	1.1e-09	-0.07
SBSC	0.00025	-0.028	5e-44	0.11	2e-07	-0.06
SBS90-like	0.0067	0.021	1.9e-50	0.11	4.9e-09	-0.067

mismatch recurrence data were consistent with what we uncovered in Chapter 3 (Fig. 3.4). We proposed that most of the C:G-to-A:T asymmetry arises from 8-oxoguanine on the non-coding DNA fragments that have been pulled down in the exome capture Chapter 3 (Fig. 3.18). A $\log_2$ ratio value of 2.5 corresponds to approximately 5.6 times more C-to-A mutations on the coding strand compared to the non-coding strand. Interestingly, the asymmetry was not extreme for the other mutation types (Fig 4.15).

4.4 Discussion

Here we describe the relationship between the frequency with which mismatches to the reference genome are observed across 200,000 UK Biobank exome sequencing samples and genomic features known to covary with somatic mutation. In the previous chapter, we investigated variation across individuals in the UK Biobank in the number of mismatches in alignments derived from whole exome sequencing, inferred using the pipeline described in Chapter 2 (Fig. 2.1). In addition to somatic mutations, these mismatches include a contribution from sequencing error, DNA damage and other technical artefacts. We hypothesized that significant contributors, such as cancer status or age, to the mismatch variation across samples in the somatic mutation rate could be detected against this background of technical noise due to the large sample numbers available

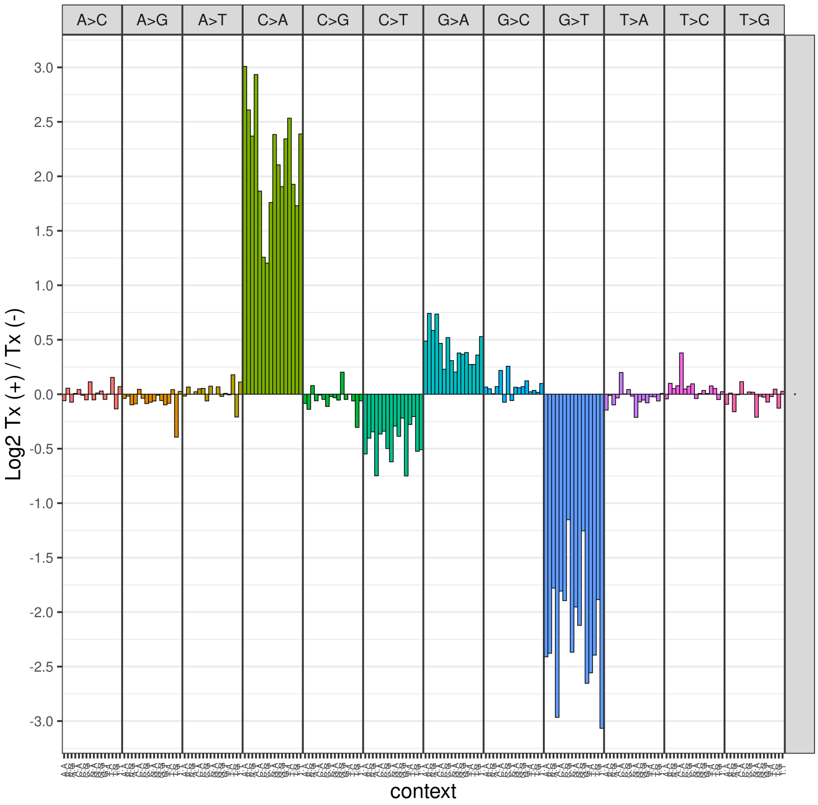

Figure 4.15: Transcription associated asymmetry for 192 mismatch contexts grouped by the 12 mismatch types. Log$_2$ ratios were calculated using the number of reference normalised mismatch contexts in the transcribed genes (forward strand) to the number of mismatches in the same context in the genes on the reverse strand.

for analysis from the UK Biobank. We found evidence to suggest that this was the case, at least to some extent for age, which showed the expected positive correlation with the number of mismatches. Still, there was no evidence of a genetic contribution to variation in the number of mismatches. Rather than variation across individuals, this chapter focuses on variation in the number of mismatches observed at sites in the exome. In much the same way as experimental techniques that have been designed to overcome technical artefacts by sequencing the same mutation multiple times (e.g., by culturing single cells), observing a mutation of a given type in more individuals than expected by chance can provide information, both on factors that influence variation in the somatic mutation across the genome and somatic selection that may increase the frequency with which a specific mutation is observed. In this chapter, we assess these effects by examining mismatch recurrence, defined as the number of individuals in which a specific mismatch is observed at a genomic position.

Here we describe variation in the median mismatch recurrence across the genome. A strong relationship between mismatch recurrence and known sources of variation in somatic mutation rate across the genome was observed, lending support to this idea, (Fig. 4.1, 4.3, 4.4, 4.5 & 4.8). There was striking statistical support for a relationship between GC content and mismatch recurrence, which likely reflects the tendency for guanine and cytosines to accumulate DNA damage [128]. Replication time is a driver of local mutation rate variation [141]. There was a positive correlation between later replication timing and mismatch recurrence. The reason for the relationship between somatic mutation and replication timing have been extensively discussed in the literature for two main reasons [141]. Firstly, as the nucleotide pool becomes depleted, the length of time that late-replicating regions remain in a mutable ssDNA state is increased and secondly, late replicating regions are in a heterochromatin state which requires chromatin remodelling that can increase strand breaks due to increase rates of replication fork stalling [362]. We found that the TCT-to-TGT mutation context had a strong linear correlation with replication timing. This specific context was dominant in the SBS13 cosmic mutational signature. In 10 out of 17 cancer types analysed by the COSMIC consortium, the activity of SBS13, which has been proposed to correspond to AID/APOBEC activity [84], and increased with increasing replication timing. The AID/APOBEC molecules are a family of highly prevalent cytosine deaminases which act as DNA mutators in the generation of diversity during T/B-cell somatic hypermutation and edit retroviral cDNA during infection, tagging the viral DNA for degradation or transformation into provirus [85]. AID/APOBEC activity was also found in 18% of cancers in the TCGA dataset and is more prevalent on the lagging strand [363]. The negative relationship between chromatin 'openness' and recurrence is also consistent with expectation, as loci that are inaccessible to global and transcriptionally associated repair mechanisms will accumulate damage at a greater rate than accessible chromatin regions [133].

We found that the median mismatch recurrence per gene is increased for genes that overlap recombination hotspots. As with replication timing, DNA at DSB sites shows increased APOBEC activity due to prolonged periods of being in a single-strand state during homologous repair [92]. During the repair of DSBs, error-prone polymerases are recruited to repair strand breaks. Given the increase in DSBs at recombination hotspots we hypothesise that the somatic mutation rate in genes in proximity to recombination hotspots will have a higher mutation rate than genes that are not located within the hotspots. Using proportions of sites within a gene that overlap with a recombination hotspot as a proxy for distance to a recombination hotspot we, indeed, see a linear increase in the mutational burden as a function of increasing hotspot overlap. Note, however, that the recombination hotspot data that was used for this analysis relates to meiotic recombination, which will not affect somatic mutations. Recombination also occurs during mitosis, and it is this mitotic recombination that has the capacity to induce somatic mutations; however, mitotic recombination, although believed to occur frequently [364] is more difficult to study than meiotic recombination and this analysis, therefore, depends on the assumption that there is some degree of sharing between meiotic and mitotic recombination hotspots. This appears likely given that recombination hotspots are determined to a large extent by sequence content and there is a high degree of sharing of meiotic recombination hotspots between the female and male germlines.

We recovered a complex relationship between median gene expression in the GTEx Whole Blood dataset and mismatch recurrence, both in the total mismatch recurrence and in specific contexts, with increased mismatch recurrence in the tails of the gene expression distribution. A linear model relating gene expression and mismatch recurrence indicated a weak negative relationship between them, corresponding to slightly higher mismatch recurrence in low to non expressed genes, a finding that is consistent with our findings in Chapter 3, and with what has been described previously by multiple groups [148, 365]. However, the relationship between expression and the mismatch rate appeared non-linear (Fig. 4.8). By fitting generalised additive models (GAMs), we found a U-shaped relationship, similar to what has been previously described, between gene expression and both the putative somatic and the germline mutation rate [148]. This was consistent with a detectable contribution of somatic mutations to the variation in the mismatch recurrence across the WES data. The aetiology of this relationship, as proposed by *Chen et al.* [148], is that when gene expression is low, there is a reduced effect of transcription-coupled repair, whereas, at high expression levels, the effects of transcription-associated mutagenesis begin to dominate over transcription-coupled repair, thus, increasing the mutation rate for the most highly expressed genes.

In addition to considering potential sources of variation in the somatic mutation rate, we can leverage the recurrence of mismatches to better understand how selection may be acting across genes in somatic cells. Consistent with the literature on somatic selection, we find an absence of negative selection. This may also reflect the fact that many mismatches are artefacts as to detect negative selection most mismatches must be true somatic mutations. Across genes critical to myeloid malignancies a signal of positive selection was found. Given that a substantial proportion of the genes on the myeloid panel are involved in clonal haematopoiesis of indeterminate potential (CHIP) this suggests that the mismatch recurrence contains a signal of true somatic mutation. *SRSF2* (Table 4.1), which is critical for haematopoiesis, and mutated in a large fraction of individuals with CHIP, showed a strong signal of positive selection. We find enrichment for signals of positive selection for histone genes and for genes involved in cytoskeleton pathways. There is increasing evidence for the role of histone mutation in cancer with recurrent somatic mutations well documented in some histone genes [366]. For example, H2AX forms at double-strand breaks and signals the DNA damage response. Reduction in the capacity to recruit repair mechanisms during damage directly impacts the mutation rate within the cell [367].

There was also an absence of evidence of selection acting on some of the canonical CHIP genes. Although CHIP is diagnosed in about 10 to 20% of individuals over the age of 70 [209, 210, 211], CHIP has only been identified in a small proportion (5%) of the UK Biobank 200k cohort [305]. This makes detecting a signal of CHIP in 200,000 individuals challenging. Our hypothesis, however, was that the mutations that drive clonal expansions arise much earlier in life and may show signals of selection across the

UK Biobank cohort before CHIP can be identified. A future direction of work would be to try to reintroduce high-frequency recurrent variants. A possible methodology to do this is by using local ancestry around the filtered variant of interest, given the structure and large sample sizes of the UK Biobank germline variants would on average share more variants in a linkage window than unrelated samples. This could potentially remove germline variants while retaining high-frequency recurrent mismatches.

Decomposing the matrix of mismatch counts (in their triplet nucleotide contexts) into latent factors has the potential to separate the various artefacts and mutational processes that contribute to the data. Here we use non-negative matrix factorization to decompose our matrix of reference mismatches into a set of linear components reflective of the mutational or artefactual processes. The top inferred mutational signatures showed relative robustness while the lower ranked signatures showed evidence of instability in Fig. 4.13 and 4.14. A well-known problem of NMF is the difficulty of inferring the rank of the factorization. There was no obvious elbow point in the diagnostic plot, Fig. 4.10. The twelve estimated signatures showed a strong association with known sources of variation in mutation rates. No single inferred mutational signature showed a signal of being specific to any genomic covariate, which may be indicative of the effect of signature bleeding or inter-sample bleeding, i.e. mutational processes active in a subset of samples being erroneously attributed to all samples due to the decomposition algorithm assuming a similar mutational landscape across all samples [368].

The effects of transcription on somatic mutation can also be studied by investigating the asymmetry in mismatch recurrence on the coding and non-coding strands. There was significant asymmetry for the C-to-A relative to G-to-T mutation types. In Chapter 3, we postulated that this was due to the exome capture step pulling down DNA fragments that are complementary to the coding strand (Fig. 3.18). Consistent with the sample-level variation in mismatch type asymmetry discussed in Chapter 3, we do not see strong asymmetry across other mutation types indicating that the primary source of DNA damage arise from 8-oxoguanine.

4.5 Conclusion

The exome sequencing data of the UK Biobank includes many somatic variants [305]. Here we provide evidence that by investigating the recurrence of mismatches to the reference we can extract a signal of somatic mutation. Although the data has large numbers of mismatches resulting from technical artefacts, the large number of samples in the UK Biobank allows the effects of somatic mutation to be detected even without identifying individual somatic mutations with high confidence. We detected associations between mutation recurrence and GC content, replication timing, recombination hotspots, chromatin structure and gene expression. Mutation recurrence can also be used to understand selective pressures acting on somatic cells. An enrichment for positive selection acting on haematological genes was seen in the UK Biobank data. An observation that is in line with earlier studies in the UK Biobank and the fact that the data was derived from whole blood [305]. Decomposing the recurrence values into mutational signatures collapsed the mutational contexts onto a much smaller number of putative sources. Again, we found a strong signal of association with gene expression except for one signal (SBS90-like). Analysing the effects of transcription strand asymmetry on the genome, we find significant strand asymmetry for the G-to-T relative to C-to-A mutational types consistent with DNA damage during library preparation steps.

Here we set out to probe the effects of somatic mutation across the genome. By rotating the data to perform a gene-level analysis that combined data from all individuals, we avoid technical batch effects that are introduced through sample collection and batched NGS sequencing. We find that for all genomic covariates of somatic mutation that we investigated (gene expression, replication-timing, chromatin structure and recombination hotspots) the expected association was also detectable in our analysis of all-source mismatches. In addition, we found evidence of somatic selection in the mismatch data. A set of genes that contribute to CHIP showed greater evidence of positive selection in these blood-derived data than the remaining genes.

4.6 Methods

4.6.1 Mutation data

A counts matrix, genes by triplet context recurrence, was generated from the filtered annotated dataset outlined in Chapter 2 by calculating the median recurrence of each mismatch context across samples for each gene. Measuring the median mutation triplet context is conceptually like normalising a raw mismatch context count per gene by the equivalent gene length. To summarize the mutation recurrence across all mutation contexts within a specific gene, the mean rank of each context within a specific gene is calculated.

4.6.2 Genomic covariate analysis

Median gene expression in whole blood for healthy individuals was downloaded from the GTEx website [369]. Replication timing data for an Epstein-Barr virus-transformed healthy blood sample (GM12878_B-Lymphocyte_Int61574576) was downloaded from the https://www2.replicationdomain.com/database.php database. Bedtools intersect was used to map replication timing data to gene IDs [370]. The average replication time across each region within a gene was taken to calculate a per-gene measure of replication timing. ATAC seq data was downloaded from ATACdb for the GM12878 cell line (http://www.licpathway.net/ATACdb/Download.php). Chromatin accessibility was defined using the fold enrichment value calculated using the ratio of read counts against the genomic background [371]. Recombination hotspots for the European population of the 1000 genomes project were generated from the supplementary data from *Li et al.* [372]. The proportion of mismatches overlapping a recombination hotspot per gene was then calculated and used as a genomic covariate.

All analyses were performed using R (v4.2) and all visualizations were performed using ggplot2 [343, 342]. GAMs were fitted using the geom_smooth function in ggplot2. GC content, chromatin openness and replication timing were fitted as fixed effect covariates in all models fitting gene expression against mutation recurrence.

4.6.3 Analysis of selection

To generate the selection results on the UK Biobank data, we used a modified version of the dndscv R package [13]. dndscv has been designed to work with sparse somatic mutation data. To process many mismatches and samples, several modifications to the dndscv package were made. The main changes to the algorithm included increasing memory efficiency and prevention of p-value underflow, no changes were made to the mathematical models except for transforming the p-values to the natural logarithm and returning the lower tail on the probability distribution rather than 1 - pchisq(llog_ll, df). The dndscv tool requires 5 inputs; sample ID, chromosome ID, base position, reference allele and mutated allele. The first change was to remove redundant information, as we are analysing hundreds of thousands of samples and billions of variants, the R implementation quickly exhausts RAM and a more efficient method of creating the internal N matrices was needed. If a mismatch to the reference was found in 1000 samples, then this site would need to be annotated 1000 times which is highly inefficient given the size of the dataset with each annotated site being stored in memory, unnecessarily. We removed the need for a sample ID and replaced it with a sample count. When the site is annotated, this count is appended to the N matrices. We have also removed warnings that show if two mutations are beside each other and instruct the model to return $-log$ p-values. To reduce the bias of gene length having inflated signals of selection we further normalise by CDS length to give a per-nucleotide adjusted measure of selection significance. The input data we used for the dNdS is aligned relative to the reference strand. To reduce the effect of complex germline confounding, as highlighted in our GWAS study in Chapter 3, mismatches that have occurred in more than 1000 samples have been excluded.

Once the selection omega values and p-values were calculated, a set of genes that drive haematological malignancies were obtained from the Archer variantPlex 75 Myeloid gene panel [215]. To assess whether there was evidence for selection acting on haematological genes we used a binary classification analysis, i.e are haematological genes enriched for positive selection given the statistics derived from the dnds model. We used receiver operator curves (ROC) and bootstrapping to assess the enrichment of selection and the confidence of the ROC curve. The fbroc package was used to generate ROC plots and AUC with 1000 bootstraps [373]. The genes in the Archer VariantPlex 75 Myeloid gene panel are listed below.

ABL1, CBLC, DNMT3A, IDH2, MYC, RAD21, STAG2, ANKRD26, CCND2, ETNK1, IKZF1, MYD88, RBBP6, STAT3, ASXL1, CDC25C, ETV6, JAK2, NF1, RPS14, TET2, ATRX, CDKN2A, EZH2, JAK3, NOTCH1, RUNX1, TP53, BCOR, CEBPA, FBXW7, KDM6A, NPM1, SETBP1, U2AF1, BCORL1, CSF3R, FLT3, KIT, NRAS, SF3B1, U2AF2, BRAF, CUX1, GATA1, KMT2A, PDGFRA, SH2B3, WT1, BTK, CXCR4, GATA2, KRAS, PHF6, SLC29A1, XPO1, CALR, DCK, GNAS, LUC7L2, PPM1D, SMC1A, ZRSR2, CBL, DDX41, HRAS, MAP2K1, PTEN, SMC3, CBLB, DHX15, IDH1, MPL, PTPN11, SRSF2

4.6.4 Mutational signatures

RccpML was used to decompose a mutational counts matrix via the non-negative factorisation [359]. RcppML factorises a matrix, A_{mn} of counts. The matrix consisted of 96 mismatch contexts (on the rows) and genes on the columns. The factorisation results in two latent matrices (W_{mk} & H_{kn}) and a scaling matrix, d. W_{mk} is a matrix containing the contribution of each of the 96 contexts to each of k signatures, H_{kn} contains the contribution of each signature to the mutational count within each gene. As highlighted in the GWAS study in Chapter 3, some common germline variants remain in our data. To limit the impact of any remaining germline variants, sites at which a mismatch was seen in more than 3,000 individuals were excluded. The unit invariant knee method from the inflection package was used to estimate the optimal rank to use in the factorisation [360]. To visualise the results of the factorisation the plotting functions from the MutationalPatterns Bioconductor package was used [374].

4.6.5 Strand asymmetry

To calculate strand asymmetry the genome was split into genes annotated by Ensembl as transcribed on the forward strand of the reference (TX+) and genes transcribed on the reverse strand of reference (TX-). The mean count of each of 96 mismatch contexts across genes in each group was then obtained and the $\log_2$ ratios of these means in the forward strand genes compared to reverse strand genes were calculated.

Chapter 5

Controlling for background genetic effects improves the power of genome-wide association studies

The content of this chapter has been published as:

D. Bennett, D. O'Shea, J. Ferguson, D. Morris, and C. Seoighe, 'Controlling for background genetic effects using polygenic scores improves the power of genome-wide association studies,' Sci Rep, vol. 11, p. 19571, Oct 2021.

C. Seoighe carried out difference in the AUC analyses. D. O'Shea performed case control simulations. J. Ferguson. provided the mathematical justification of the method outlined in Appendix C. All authors proofed the submitted manuscript

5.1 Abstract

Ongoing increases in the size of human genotype and phenotype collections offer the promise of improved understanding of the genetics of complex diseases. In addition to the biological insights that can be gained from the nature of the variants that contribute to the genetic component of complex trait variability, these data bring forward the prospect of predicting complex traits and the risk of complex genetic diseases from genotype data. Here we show that advances in phenotype prediction can be applied to improve the power of genome-wide association studies. We demonstrate a simple and efficient method to model genetic background effects using polygenic scores derived from SNPs that are not on the same chromosome as the target SNP. Using simulated and real data we found that this can result in a substantial increase in the number of variants passing genome-wide significance thresholds. This increase in power to detect trait-associated variants also translates into an increase in the accuracy with which the resulting polygenic score predicts the phenotype from genotype data. Our results suggest that advances in methods for phenotype prediction can be exploited to improve the control of background genetic effects, leading to more accurate GWAS results and further improvements in phenotype prediction.

5.2 Introduction

Linear mixed effects models (LMMs) are routinely applied to detect associations between SNPs and phenotypes in genome-wide association studies (GWAS) and many methods have been developed that enable these models to be applied efficiently to the large scale datasets that are typically now encountered in studies of complex traits [267, 268, 269, 270, 271, 272, 273, 261, 246]. Compared to fixed effects models for GWAS [260], LMMs can be designed that have the advantage of being applicable to samples that include related individuals [375, 267, 274]. LMMs for this purpose typically include a random effect with covariance proportional to the kinship matrix that indicates the degree of relatedness between pairs of individuals in the sample [274]. The relatedness of individuals in the sample may be known *a priori* or may be derived from the genotype data by constructing a genetic relationship matrix (GRM), with entries corresponding to the genotypic covariance between pairs of individuals. When the entries of the GRM below a specified threshold are set to zero, the GRM is approximately equivalent to a family kinship matrix, with the degree of relatedness that the matrix captures controlled by this threshold. Thresholding the matrix to capture close family relationships (or cryptic relatedness [376]) allows specialized computational methods for sparse matrices to be applied so that model fitting remains tractable for studies that include large numbers of individuals [246]. This is the approach taken by fastGWA [246], a recently developed tool that has been shown to generate correctly calibrated statistical results efficiently for biobank-scale GWAS.

In addition to enabling application to samples containing related individuals, LMMs can also account for genetic background effects [277, 375]. When a statistical model is used to test for a relationship between a given SNP (the test SNP) and a phenotype, genetic variants in the genome that are not in linkage disequilibrium with the test SNP may also make a substantial contribution to the phenotypic variation. If this contribution to phenotypic variation is not accounted for it contributes to the error term in the model. If the trait of interest is both highly polygenic and highly heritable this noise may be substantial. Failure to account for sources of variance in the response in a statistical model can reduce the power to detect a relationship of interest [377, 378]. A LMM with a full GRM (i.e. derived from all SNPs in the data and with no threshold applied on the level of genetic correlation between individuals) is equivalent to a model in which all variants are assumed to have a causal effect on the phenotype, with effect sizes consisting of independent samples from a Gaussian distribution [261]. This is typically not a good fit to the true effect size distribution, and instead, the software package BOLT-LMM [261] uses a spike-and-slab Gaussian mixture for the effect size distribution, with a component (the spike) close to zero corresponding to weak genome-wide effects and accounting for family relationships, and component with larger variance (the slab) corresponding to variants with large effects [261]. Fitting this more sophisticated model requires specialist numeric methods, that are relatively computationally intensive. Consequently BOLT-LMM is much more computationally intensive than fastGWA [246].

The full GRM is an $N \times N$ matrix, where N is the number of individuals in the study. The memory requirement of BOLT-LMM is kept tractable by not explicitly evaluating the GRM but rather BOLT-LMM solves the mixed model equations by computing the product of the inverse GRM and the phenotype vectors. Nonetheless, the overall compute time and memory requirements of BOLT-LMM are a function of both N and the number of model SNPs, M, that contribute to the GRM (with $O((NM)^{1.5})$ compute time and $\frac{NM}{4}$ bytes of memory required. Various options have been explored for which SNPs to include in the calculation of the GRM [375]. Including SNPs in LD with the target SNP results in loss of power, as the effect of the target SNP is partially accounted for by the random effect through the GRM. This has been referred to as proximal contamination [277]. On the other hand, including all (or most) SNPs that are not in LD with the target SNP, e.g. using a Leave One Chromosome Out (LOCO) approach, can result in dilution of the extent to which the relevant part of the genetic background is captured by the GRM. In the latter case, SNPs that are not relevant, in that they do not capture direct genetic effects or tag relevant population structure effects, effectively add noise to the GRM [277]. Alternatively, the GRM can be built from only the SNPs that are found using a linear model to be associated with the phenotype. Although this results in an increase in statistical power [272, 375, 379], it does not fully control for population structure and is not recommended if population structure is of substantial concern [261, 375]. Methods have been developed that incorporate principal components into the GRM calculation built from significant SNPs; however, most of these methods are not suited to large biobank-scale data, without access to cloud computing or large compute farms [380, 381, 382]. Background genetic effects can also be included in the statistical model as fixed effects and this is the recommended approach when there are SNPs with large effect sizes [375]. A model fitting approach to determine the SNPs to include as fixed effects has been developed, and this also results in increased power in GWAS [277].

As the genomic architecture of complex diseases is uncovered with the help of large biobanks, there is an advancing prospect of predicting quantitative phenotypes and the risk of complex diseases from genotype data. Recent years have seen substantial success and emerging clinical utility in phenotype prediction from polygenic scores (PGS) [383, 384]. PGS are constructed from weighted sums of allele dosages, with the weights corresponding to the effects size of the variants. Risk variants (variants associated with the phenotype) are typically inferred from the largest available GWAS, generally a meta-analysis. The clinical potential of PGS has already been shown in complex diseases such as coronary artery disease (CAD), diabetes and cancer [15, 384, 385]. In CAD, the identification of individuals with similar risk to those with rare high-risk monogenic variants has been reported [15]. Similarly, in breast cancer, pathogenic variants in *BRCA1/2* account for 25% of familial risk of the disease with genome wide variants accounting for a further 18% of the risk [386, 387]. It is likely that in the future specialist machine learning methods will be developed to predict phenotype from genotype [384], potentially achieving higher accuracy by incorporating the possibility of non-additive effects.

Here, we set out an approach to GWAS that seeks to separate the model fitting at the test locus and estimation of the genetic background effect. After carrying out an initial round of GWAS using an existing method, we derive a PGS for each chromosome, using the summary statistics for SNPs on the remaining chromosomes. We refer to this as the Leave One Chromosome Out Poly Genic Score (LOCO PGS). We then perform a second round of GWAS, including the relevant LOCO PGS as a fixed effect to account for the contribution to the variation in the phenotype of SNPs that are not on the same chromosome as the test locus. We tested this approach in two ways. Firstly, using simulated data we tested for an improvement in power on the task of recovering known causal variants as a function of study size, number of causal variants and trait heritability. In addition, we applied the method to standing height data from the UK Biobank and determined the number and characteristics of additional variants that were detected. For an objective assessment of performance on real data, where the true associations are unknown, we divided the data into test and training sets and predicted the phenotype in the test set. The improvement in performance on the critical task of complex phenotype prediction illustrates the utility of the PGS as a means of accounting for off target genetic effects. This straightforward, modular approach to accounting for genetic background effects in GWAS has the advantage of leveraging advances in phenotype prediction as they become available. It also offers significant improvements in speed relative to existing methods that correct for genetic background.

5.3 Results

We incorporated the LOCO PGS as a fixed effect in a linear mixed model using the existing tools, GCTA fastGWA, BOLT-LMM and REGENIE [246, 261, 247]. We refer to the methods that result from including the LOCO PGS fixed effect by appending PGS and the name of the method used to calculate the PGS to the name of the original tool. For example, fastGWA with a LOCO PGS fixed effect, calculated using the P&T or LDpred2 methods are referred to as fastGWA-PGS-PT and fastGWA-PGS-LDPred2, respectively. We simulated data to evaluate the impact of including the LOCO PGS as a fixed effect in GWAS. The simulations consisted initially of a normally-distributed continuous trait in 100,000 individuals. The trait had a narrow-sense heritability (h^2) of 0.5 and there were 1,000 causal SNPs with normally-distributed effects on the trait (see Methods for details). To check the validity of this approach we performed simulations under the null model of no association between genotype and phenotype and found that the method was well calibrated (Fig. C1). This was the case both for the P&T method of calculating the LOCO PGS (with a fixed p-value threshold of 5×10^{-5}) and for the LDpred2 method and was in-line with our expectations, as the LOCO PGS is approximately uncorrelated with the genotype of the tested SNP (see Appendix C for a mathematical justification). The false positive rate rose slightly when we used a high p-value threshold with the P&T method to calculate the LOCO PGS (Fig. C1). In this case the majority of the variants contributing to the LOCO-PGS are likely to be false positives and the LOCO PGS may pick up some residual population structure.

In 100 simulations we found that including a LOCO PGS resulted in a substantial improvement in power to detect the known causal SNPs (Fig. 5.1). We considered two alternative methods to select the SNP effects to include in the PGS calculation: pruning and thresholding (P&T) and LDpred2 [388, 291]. When we included the PGS obtained using P&T as a fixed effect with fastGWA (which we refer to as fastGWA-PGS-PT) we recovered 82 additional causal variants, on average, below the conventional p-value threshold of $5x10^{-8}$ compared to fastGWA (corresponding to a relative increase in power of 18.4%; $p = 3.0 \times 10^{-32}$ from a paired T-test; Tables C1, C2 & C3). The performance was further improved when we used LDpred2 to calculate the LOCO PGS (referred to as fastGWA-PGS-LDpred2). This resulted in the recovery of, on average, 115 more causal variants than fastGWA alone (relative increase of 25.9%; $p = 2.3 \times 10^{-36}$). Inclusion of a LOCO PGS with an under powered BOLT-LMM (BOLT-LMM-165-PGS-PT) resulted in a large boost in power, recovering an additional 55 variants over BOLT-LMM-165 , while BOLT-LMM with a GRM derived from all variants (BOLT-LMM-664), had the second highest power to recover causal variants after fastGWA-PGS-LDpred2, recovering 112 more causal variants than fastGWA (relative increase in power of 25.3%; $p = 2.3 \times 10^{-40}$). Recently, a new fast method, REGENIE [247], has been released that also includes control of the polygenic background effect based on prediction of the phenotype from SNPs that are not on the same chromosome as the test SNP. In our simulations the performance of REGENIE was higher than fastGWA but well behind fastGWA-PGS-LDpred2. REGENIE showed no improvement when the LOCO PGS was added as a fixed effect, suggesting that it accounts adequately for the genetic background effect. We also simulated case control data for a binary traits with h^2 of 0.5 and 1,000 causal loci, with disease prevalence, k, of 0.1 and

0.3. As with the quantitative trait simulations, inclusion of a LOCO PGS fixed effect always resulted in an increase in the average number of casual loci recovered, with an average of 28 more causal loci recovered for a disease prevalence of k=0.1 (p = 0.19) while, k = 0.3 recovered on average 48 more causal loci (p= 0.03) (Fig. C2 and Table C4 & C5).

The contribution to phenotype variance of background SNPs can also be modelled as a random effect in a linear mixed model. This approach is applied by BOLT-LMM, which uses a normal mixture random effect, with a component corresponding to SNPs with large effects. The running time of BOLT-LMM is proportional to $MN^{1.5}$ and the memory requirement is approximately $MN/4$ bytes, where N is the number of individuals in the dataset and M is the number of SNPs included in the GRM [261]. When we ran BOLT-LMM with a subset of 165,683 SNPs (see Methods for how these were selected) we found that including the LOCO PGS as a fixed effect resulted in a substantial gain in power (Fig. 5.1), likely resulting from inability of the reduced GRM to account fully for genetic background. No further improvement was obtained by adding the LOCO PGS fixed effect to BOLT-LMM with a GRM consisting of all of the 664,393 directly genotyped SNPs (Fig. C3); however, the power obtained with the smaller GRM with the PGS fixed effect was close to the power obtained with the larger GRM, but with a much lower memory requirement (Table 5.1). Note that REGENIE was omitted from Table 5.1, as the simulation is based on a single phenotype and would unfairly disadvantage REGENIE, which is optimized for the task of performing association analyses on multiple phenotypes simultaneously.

Table 5.1: Pipeline computation time and memory (N=100,000, M=664k). Analyses were performed on a single HPC node with 32 Xeon(R) CPU D-1541 CPUs with 128GB of RAM.

	CPU Time (s)				
Method	GWAS	LOCO PGS	GWAS(22 chr)	Total (CPU Time)	Max Memory (GB)
fastGWA-PGS-LDpred2	501.2	58,880.0	2,953.3	62,334.5	6.3
fastGWA-PGS-PT	501.2	245.8	2,953.3	3,700.2	0.7
BOLT-LMM-664	119,202.0	0.0	0.0	119,202.0	15.5
BOLT-LMM-165-PGS-PT	92,108.0	245.8	614,514.4	706,868.2	3.9

We calculated receiver operator characteristic (ROC) curves to investigate whether the increased number of causal variants recovered when we included the LOCO PGS as a fixed effect reflected a reduction in p-values across the board for the phenotype-associated variants or also an improvement in the ordering of the variants, when the

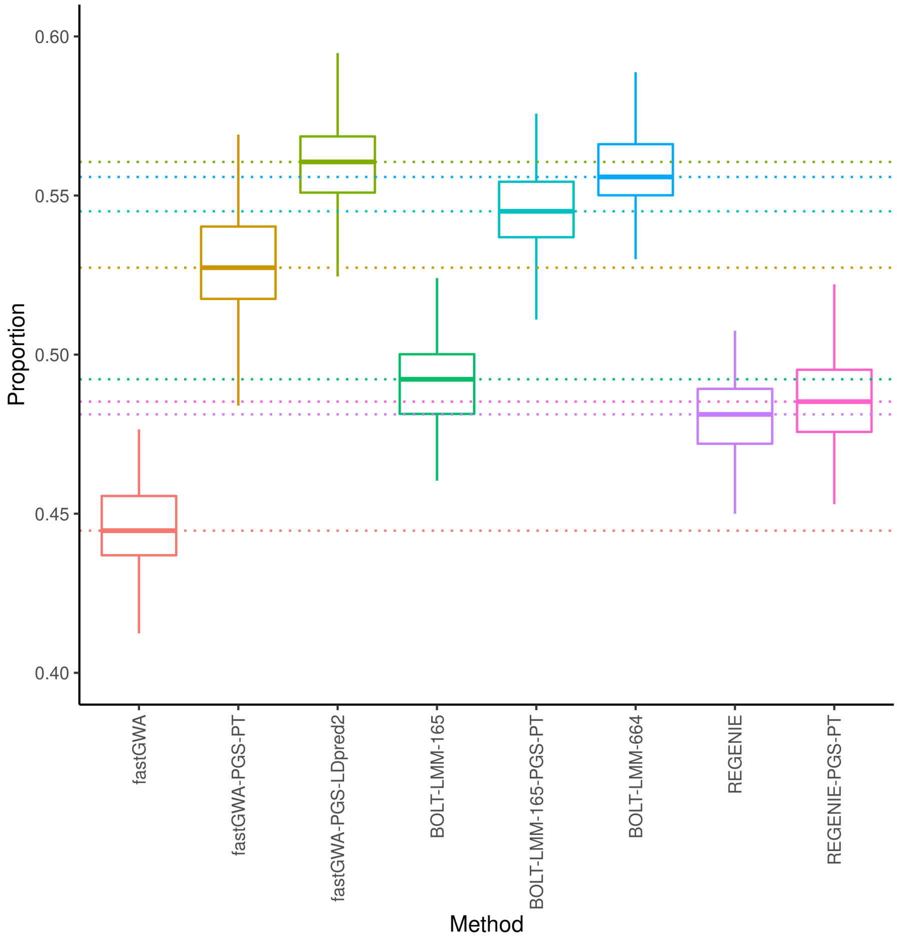

Figure 5.1: The proportion of causal variants recovered in 100 simulations. The boxplot shows the median (center line), upper and lower quartiles (hinges) and the maximum and minimum values not more than 1.5 times the interquartile range from the corresponding hinge (whiskers). The simulations consisted of 100,000 individuals and a continuous trait, with narrow-sense heritability of 0.5 and 1,000 causal variants. BOLT-LMM-165 denotes BOLT-LMM with a GRM derived from 165,684 variants resulting from strict LD-pruning. BOLT-LMM-664 refers to the use of BOLT-LMM with a GRM derived from all 664,393 variants in the simulations. Methods that include PGS in the name involved the use of a LOCO PGS fixed effect, derived either from pruning and thresholding (methods ending in PT) or using LDpred2.

variants are ordered by the evidence of an association with the phenotype. Over 100 simulations we found that the area under the ROC curve (AUC) was always higher for fastGWA-PGS-LDPred2 than for fastGWA without the LOCO PGS fixed effect (Fig. 5.2). This was also the case for 99 of the 100 simulations when we added the PGS fixed effect to BOLT-LMM-165. The difference in sensitivity as a function of specificity (Table C6) showed that the sensitivity was consistently higher at a given specificity when the LOCO PGS-LDpred2 was included as a fixed effect, indicating an improvement in the ordering of the SNPs. The increase in mean sensitivity was up to 0.073 in the case of fastGWA-PGS-LDPred2 vs fastGWA, corresponding to a relative increase of 11.6% (at a specificity of 0.9988) over fastGWA. The addition of the LOCO PGS fixed effect led to a smaller but still consistent increase in sensitivity for BOLT-LMM-165. In this case, the greatest increase in the mean sensitivity was 0.028, corresponding to a 4.2% relative increase in sensitivity (at a specificity of 0.9991)

In addition to increasing the statistical power to detect causal variants, including the PGS fixed effect also resulted in an improvement in effect size estimates (Fig. C4). We found that when a fixed effect PGS was incorporated into the association study the median squared error (MEDSE) of the effect size estimate was substantially reduced (Fig. C4, Tables C7, C7 & C9). Interestingly, the MEDSE of the effect size estimate was largest across all methods for BOLT-LMM with the reduced GRM (Fig. C4).

5.3.1 Effects of trait heritability, number of causal variants and sample size

We simulated data over a range of values of sample size, h^2 and of the number of causal SNPs to investigate how these parameters affect the impact of including the LOCO PGS as a fixed effect on GWAS power. For this analysis we used the P&T method to calculate the LOCO PGS, due to its lower computational cost (Table 5.1). For the larger sample size, a small improvement in power was obtained even for the lowest values of h^2 (0.1) simulated, with a statistically significant improvement for $h^2 \geq 0.2$ (Fig. 5.3). The improvement was not statistically significant at this value of h^2 when only 100,000 samples were used in the simulation, but even in this case the number of causal variants recovered was always at least as large and typically larger when the PGS fixed effect was included in the model (Tables C10 & C11). This was somewhat surprising, given

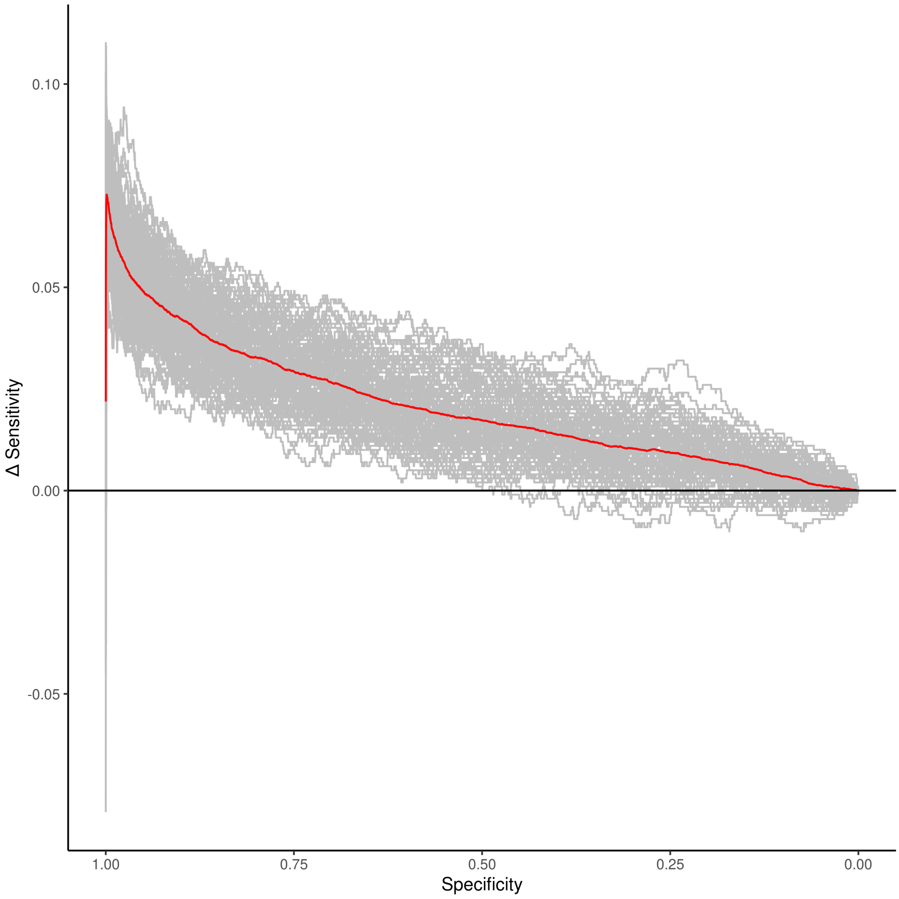

Figure 5.2: Difference in sensitivity (between fastGWA-PGS-LDpred2 and fastGWA) as a function of specificity for 100 simulations of a continuous trait with narrow-sense heritability of 0.5 and 1,000 causal variants in 100,000 individuals. The specificity (x-axis) is discretized in bins of size 0.0001. Each grey line shows the results of one simulation. The red line shows the mean difference over all simulations.

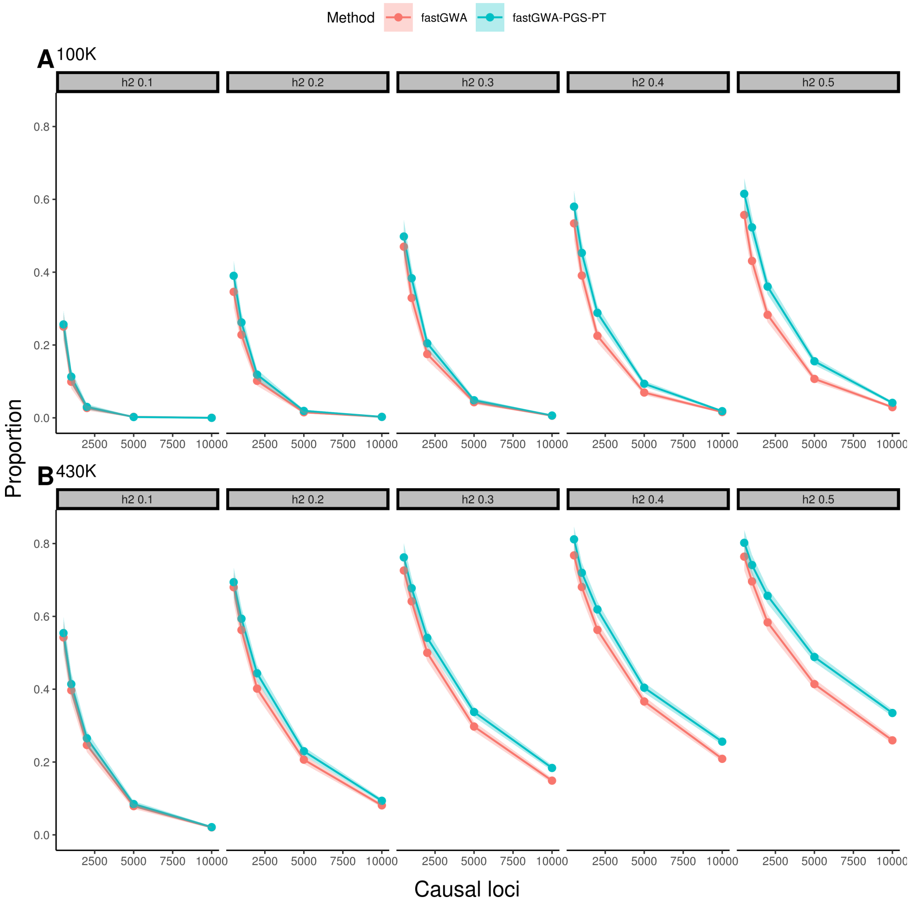

Figure 5.3: Proportion of causal variants recovered in simulations of a quantitative trait over a range of values of h^2 and the number of causal loci. Simulations on the top (A) and bottom (B) panels were based on 100,000 and 430,000 randomly sampled individuals from the UK Biobank, respectively.

that it is assumed that large sample sizes are required for accurate phenotype prediction from PGS [389].

The improvement in power resulting from the inclusion of the PGS fixed effect increased consistently with increasing numbers of causal variants in the case of the larger sample size. This was not the case for the smaller sample size, for which the improvement decreased or was lost altogether when the number of causal variants was large (Fig. 5.3). This is likely due to the loss of power to detect true causal variants and to estimate their effect sizes accurately when the genetic effect is distributed over too large a number of causal variants, resulting in the inability to correct for the genetic background using the PGS. This suggests that larger sample sizes would be required for highly polygenic traits in order to obtain a benefit from using the LOCO PGS fixed effect. However, the larger sample size simulated is comparable in scale to the UK Biobank and with a sample of this size our simulations suggest that a significant improvement in power can be obtained, even for a trait with 10,000 independent causal loci. For the case control simulation (N=100,000), a more modest increase in power was observed as heritability increased, whereas the power to recover smaller effect loci decreased dramatically compared to the quantitative simulation. However, we found that for all except three simulations the inclusion of a fixed effect LOCO PGS improved the power to detect associated loci (Fig. C5, Table C12).

5.3.2 Application to UK Biobank phenotypes

We assessed the impact of including the LOCO PGS fixed effect on the performance of fastGWA on real data using standing height, BMI, and heel bone mineral density (HBMD) in individuals of British ancestry (N_{height}=395,133, N_{BMI}=395,149 & N_{HBMD}=229,191) from the UK Biobank. The distribution of p-values obtained from fastGWA with the LOCO PGS fixed effect was lower than that obtained using fastGWA, regardless of the method used to calculate the PGS (Fig. C6, C7 & C8). At a genome-wide significance level of $5x10^{-8}$ inclusion of a LOCO PGS always increased the number of independent loci recovered, compared to fastGWA (Table 5.2). We also applied BOLT-LMM to the real data. In this case we used all 556,516 lightly pruned HAPMAP3 variants for the GRM (see Methods for details). Across height, HBMD, and BMI, BOLT-LMM identified the largest number of independent associated loci.

Including the PGS fixed effect resulted in substantial increases in the number of independent associated loci, compared to fastGWA alone for all phenotypes (Table 5.2, Table C13).

Table 5.2: Number of independent significant loci identified and resulting phenotype prediction model fit. R^2 Full is the coefficient of determination of a model that includes the PGS, sex, age & 10 PCs as covariates while R^2 PGS is the coefficient for a model that includes only the PGS. BOLT-LMM was applied with a GRM consisting of 556,516 variants.

Method	Significant loci	R^2 full	95% CI	R^2 PGS	95% CI	Spearman's ρ	Phenotype
fastGWA	1,381	0.696	0.689, 0.702	0.165	0.158, 0.170	0.382	Height
fastGWA-PGS-PT	1,583	0.701	0.694, 0.707	0.173	0.166, 0.179	0.391	
fastGWA-PGS-LDpred2	1,717	0.703	0.696, 0.709	0.176	0.170, 0.182	0.395	
BOLT-LMM	1,804	0.703	0.697, 0.709	0.170	0.164, 0.176	0.388	
fastGWA	450	0.151	0.146, 0.158	0.130	0.124, 0.135	0.351	BMI
fastGWA-PGS-PT	493	0.153	0.147, 0.159	0.130	0.125, 0.136	0.351	
fastGWA-PGS-LDpred2	500	0.151	0.146, 0.157	0.127	0.121, 0.133	0.346	
BOLT-LMM	583	0.155	0.150, 0.162	0.134	0.128, 0.139	0.356	
fastGWA	324	0.216	0.204, 0.232	0.158	0.144, 0.171	0.427	HBMD
fastGWA-PGS-PT	365	0.221	0.208, 0.238	0.164	0.152, 0.178	0.439	
fastGWA-PGS-LDpred2	385	0.225	0.210, 0.241	0.167	0.154, 0.182	0.444	
BOLT-LMM	393	0.223	0.209, 0.238	0.165	0.152, 0.178	0.437	

One way to determine objectively whether fastGWA with a LOCO PGS fixed effect outperforms fastGWA on real data is to apply the methods on the key task of phenotype prediction. We used summary statistics from the 3 analyses above to calculate PGS scores using LDpred2 and pruning and P&T (see Methods for details on defining independent training and test data). For two of the three phenotypes (height and HBMD), the PGS fixed effect resulted in an increase in the correlation between the PGS and the phenotype in the test data (Table 5.2). In both cases the highest correlation between with the phenotype was obtained using fastGWA-PGS-LDpred2, which out-performed BOLT-LMM on this task. For the remaining phenotype (BMI), the addition of the PGS fixed effected resulted in no change or a slightly worse correlation with the phenotype in the test data. In this case the highest performance was obtained by BOLT-LMM (but at a substantial cost in terms of computational cost; Table 1). However, even in this case, we found that including only the SNPs with low p-values in the polygenic score (as implemented by the P&T method) resulted in an improvement over fastGWA (Fig. C9).

5.4 Discussion

Omitting covariates that are associated with a response and independent of an effect of interest can result in a reduction in the efficiency of the estimation of the effect of interest [377, 378]. Complex traits are associated with the genotype of many loci across the genome, but the effects of genetic variants other than the variant being tested are often not fully modelled by GWAS methods. We evaluated a simple two-stage approach to accounting for this genetic background effect that consists of performing an initial GWAS and using the summary statistics to calculate a polygenic score and then including the polygenic score, derived from SNPs not on the same chromosome as the target SNP, as a fixed effect in a second round of association testing. Using simulated data, we found that this led to a substantial improvement in power of fastGWA, an efficient tool for biobank scale GWAS that does not fully control for genetic background effects. When we included the LOCO polygenic score as a fixed effect with fastGWA (which we refer to as fastGWA-PGS), the power exceeded that of REGENIE [247], a recent, computationally efficient tool for GWAS that uses ridge regression to control for genetic background effects. When BOLT-LMM [261] was used with a GRM derived from all of the simulated variants, the LOCO PGS fixed effect did not provide any boost in power (Fig. C3); however, the equivalent (or slightly improved) performance of fastGWA-PGS-LDpred2 (Fig. 5.1) was achieved at a much lower computational cost (Table 5.1). Furthermore, we note, that our simulations were favourable to BOLT-LMM because the LOCO PGS was calculated from the same set of variants that were used in the GRM of BOLT-LMM. In practice, in the case of P&T millions of variants can be included in the LOCO PGS calculations, but the number of model variants, M, that can be included in the GRM of BOLT-LMM is constrained by memory and compute time, both of which scale at least linearly with M. A further key advantage of the approach that we propose is that it is modular. Any phenotype prediction method can be used to predict the combined effect of the LOCO genetic variants on the phenotype. As methods for phenotype prediction improve, we anticipate that the performance of this approach will increase.

The increase in power using the PGS fixed effect was largest for simulated phenotypes with high heritability and a large number of causal variants (Fig. 5.3). In these cases the many background SNPs collectively explain a substantial proportion of the phenotypic variance and summarizing the contribution of these background SNPs to the phenotype via the LOCO PGS is likely to result in a better estimate of the effect of the target SNP and its standard error. The boost in performance derived from including the LOCO PGS as a fixed effect also depended on study sizes. For example, when the number of causal variants became large (10,000) there was no substantial boost in performance in the simulation that included 100,000 individuals, presumably because in this case the study size was not sufficient to identify and accurately estimate the effects of the causal variants. Even with this large number of causal variants the larger simulation (with 430,000 individuals) still showed a significant improvement arising from the LOCO PGS fixed effect (Fig. 5.3). Across all the simulation parameters we investigated, the performance of fastGWA with a LOCO PGS fixed effect (calculated using the P&T method) was never worse than fastGWA without the fixed effect included. We also note that we calculated the P&T LOCO PGS using SNPs that were selected based on a fixed p-value threshold. Further increases in power may be possible by optimizing the SNPs that are used to calculate the PGS separately for each omitted chromosome but care needs to be taken when using pruning and thresholding in order to avoid increasing the false positive rate (Fig. C1). Thresholding based on a p-value was not required for LDpred2, which may help to explain why we achieved significantly better power when the LOCO PGS was calculated using this method rather than pruning and thresholding (Fig. 5.1).

We also applied the method to real data (standing height, heel bone mineral density (HBMD) and body mass index (BMI) in individuals of British ancestry in the UK Biobank). Consistent with the simulation results, we found more independent trait-associated loci using fastGWA-PGS-LDpred2 than with fastGWA alone for all three traits (30%, 19%, & 11% more for height, HBMD, and BMI, respectively; Table 5.2). Although, BOLT-LMM recovered the largest number of independent significant loci across all UK Biobank traits, this did not always translate into better correlation between a PGS calculated from the resulting summary statistics and the phenotype in the test dataset. In fact, the highest correlation was obtained by fastGWA-PGS-LDpred2 for two of the three traits. This could be explained by a higher proportion of true positives among the loci detected using the PGS-based methods or a more accurate estimate of the effects sizes by these methods, as suggested by Fig. C4. For BMI, the correlation was in fact lower between the PGS and the phenotype in the test dataset when the LOCO PGS fixed effect was used (Table 5.2). However, even in this case a larger number of significant variants were recovered than with fastGWA.

The use of polygenic scores for phenotype prediction from genotype is an increasingly important application of the results of GWAS [390]. High polygenic scores can capture a substantial component of the risk of complex diseases [15, 391] and guide interventions that can confer health benefits to individuals and reduce the stress on health systems [392]. Performing GWAS on a subset of samples and predicting on the remainder, we observed an increase in the correlation of the PGS with the phenotype when we included the LOCO PGS as a fixed effect in two out of three traits considered, consistent with improved effect size estimates (Fig. C4). Our results suggest that a modular approach that integrates advances in phenotype prediction with efficient GWAS methods can have a significant impact on the power of GWAS and that this can, in turn, lead to more accurate phenotype prediction. A recent study showed that models that allow unequal a priori contribution of SNPs to trait heritability can lead to substantial improvements in the accuracy of trait [393]. Although not explored in this work, the incorporation of external PGS instruments from large meta-analyses in the first round of GWAS may also provide an additional gain in performance, similar to proposed in Bulik-Sullivan [394]. Indeed, our results show that GWAS summary statistics can be used to account for genetic background effects, with results matching the performance of methods such as BOLT-LMM that require individual-level data for this purpose. As new efficient methods emerge from these and further insights, they can be easily substituted for the calculation of the LOCO PGS fixed effect. The current fast pace of methodological innovation in phenotype prediction supports the use, at least for the time being, of the simple modular approach to modelling genetic background effects evaluated here.

5.5 Conclusion

The tasks of detecting trait-associated variants and predicting the trait in a new sample from the summary statistics of these variants are closely intertwined. Improved performance on the trait-association task can result in more associated variants and better estimates of their effect sizes, resulting in improvement on the prediction task. On the other hand, improved methods for phenotype prediction can help to control for background genetic effects in methods that identify the trait-associated variants and their effects. The method that we have explored here consists of incorporating a LOCO PGS as a fixed-effect covariate to control for these background genetic effects; however, any method for phenotype prediction could play this role, once its application is restricted to variants that are not linked to the target SNP. We show here that incorporating the PGS as a fixed-effect covariate results in increased power to detect trait-associated variants in GWAS. The resulting trait-associated variants and effect size estimates can lead to an improvement in the PGS, as illustrated by improved performance in the task of predicting the phenotype in a test dataset.

5.6 Methods

5.6.1 Simulations

5.6.1.1 Genotype QC

The use of the UK Biobank Materials falls within UK Biobank's generic Research Tissue Bank (RTB) approval from the NHS North West Research Ethics Committee, UK. The simulated genotype data was based on autosomal genotyped data from the UK Biobank. To limit the effects of population stratification only individuals reporting white British ancestry (data field 21000; code 1001; N=443,076) were included in these analyses. The genotype data for the simulation analysis was based on directly genotyped variants with minor allele frequency (MAF) greater than 0.05%. Variants with genotype missingness greater than 2% or that failed a test for Hardy-Weinberg equilibrium (HWE) at $\alpha = 0.0001$ were excluded, resulting in a total of 664,393 genetic variants. There were 429,359 samples remaining following filtering. The sparse GRM required by fastGWA was created by setting entries corresponding to sample pairs with an estimated relatedness of less than 0.05 to 0. To account for population structure in the association studies, principal component analysis (PCA) was performed on a set of 165,684 variants LD-pruned with an R^2 greater than 0.1 in a sliding window of size 500bp, sliding by 200bp. This set was also used as the basis of the BOLT-LMM analyses with the reduced GRM size (referred to as BOLT-LMM-165 in Results). All genotype QC was implemented in PLINK2 [345].

Based on the above genotype data, we simulated a continuous phenotype using the

GCTA software suite [249]. The initial simulation (Fig. 5.1) consisted of 100,000 individuals, 1,000 randomly sampled causal variants and $h^2 = 0.5$. This simulation was repeated 100 times with the 664,393 variants remaining after variant filtering for the GRM calculation. Power was calculated as the proportion of the causal variants recovered. Further simulations were carried out to investigate the effects of varying the number of causal SNPs (500, 1000, 2000, 5000 & 10,000), h^2 (0.1, 0.2, 0.3 0.4, 0.5) and the sample size (100,000 & 430,000) on method performance. In each case all parameters other than the ones being varied were the same as the initial simulation, and one simulation was performed per set of parameter values. The pROC R package was used to generate receiver operating characteristic (ROC) curves, variants within 1 Mb of the causal variants were removed. [395]. We applied the same simulation strategy to binary traits with two levels of disease prevalence, 0.1 & 0.3, using 1,000 causal loci with $h^2 = 0.5$, and 100,000 samples. To calculate the false positive rate we performed 30 simulations with 100,000 samples, an $h^2 = 0.5$ and 1000 causal variants restricted to the even chromosomes.

5.6.1.2 Simulation association tests

Association testing was performed using fastGWA, REGENIE and BOLT-LMM. To account for known sources of covariation (technical batch effects, population structure, biological effects) 10 PCs, sex, age, genotype batch and assessment centre were included as fixed-effect covariates in statistical models. For the PGS method we first performed GWAS (using fastGWA, REGENIE or BOLT-LMM) and calculated PGS scores on a Leave One Chromosome Out (LOCO) basis. This resulted in 22 sets of PGS values (one for each autosomal chromosome, calculated from the summary statistics of variants on all other autosomal chromosomes). Two PGS strategies were used in this study, pruning and thresholding (P+T), denoted with the suffix PGS-PT and LDpred2, denoted by the suffix PGS-LDpred2. The LOCO PGS-PT were calculated using PRSice2 (version 2.2.12 (2020-02-20))[388]. To decrease computation time and reduce the likelihood of over-fitting a p-value threshold of 5×10^{-5} was chosen, a priori, for the LOCO PGS-PT calculation. Association testing was then performed using fastGWA in a chromosome-wise manner, with the corresponding LOCO PGS included as a fixed effect. The bigsnpr R package (bigsnpr v1.6.1 [291] & R v3.6.1 [342]) was used to calculate the LOCO PGS-LDpred2 fixed effects. To reduce computation time, 22 LOCO genotype objects containing the SNP correlations were precomputed.

5.6.2 Application to the UK Biobank

5.6.2.1 UK Biobank association tests

The genotype selection, quality control and genetic relationship matrix were performed following the QC procedure in *Jiang et al.*[246]. The genetic relationship matrix used with fastGWA and BOLT-LMM was calculated for all European individuals (N=458,686), using a set of 556,516 lightly pruned HAPMAP3 variants (R^2 greater than 0.9 in a 100 variant sliding window of size 1,000 & MAF > 0.01) [246]. Association summary statistics were generated from a set of 1.1 million HAPMAP3 variants (MAF > 0.01, HWE $\alpha = 1 \times 10^{-6}$ and missingness < 0.05) [246]. Principal components were calculated using a set of 34,775 variants (LD-pruned with R^2 = 0.05 in a sliding window of size 1,000bp, sliding by 50bp)[396]. To identify white British samples with similar genetic backgrounds we clustered samples based on the first 6 principal components[396], resulting in a subset of 406,319 white-British samples. Sample pairs that had a KING kinship coefficient above 0.05, with one member of the pair within the white-British group and the other in the group self-reporting as white European were removed. This left 399,135 white British and 46,406 other European samples [397, 396]. To account for known sources of phenotype and genotype variation, 10 PCs, age, sex, genotype batch and assessment centre were included as fixed-effect covariates for the BOLT-LMM and fastGWA analyses. PRSice2 and LDpred2 were used to calculate the LOCO PGS. Independent loci were identified using the clumping algorithm in plink2 (p-value threshold = 5×10^{-9}, window size = 5Mb, and LD R^2 threshold = 0.01).

5.6.2.2 UK Biobank phenotype prediction

To test the performance of fastGWA with a LOCO PGS fixed effect on the task of predicting standing height, BMI and HBMD, the UK Biobank data was partitioned into training and test datasets. The test data consisted of white British individuals with similar genetic background described above and the polygenic score predictions were tested on the remaining independent European samples. Summary statistics were generated using fastGWA, fastGWA-PGS-PT, fastGWA-PGS-LDpred2 and BOLT-LMM. We used LDpred2 and PRSice2 to predict the phenotypic values in the test set. LDpred2 requires LD correlation data and we used a pre-computed set built on the 1.1 million HAPMAP3 variants for this purpose. The model fit was assessed for each method by fitting a linear model to the values of the phenotype in the test set as a function of their predicted values, accounting for known sources of phenotypic variation, i.e sex, age, PC's. We report both the proportion of variation explained collectively by the PGS, sex, age, the first 4 principal components and assessment centre as well as the R^2 using only the PGS in the regression model.

Chapter 6

Summary and Concluding remarks

In this thesis, the central research question was whether we could uncover sources of variation in somatic mutation from whole exome sequencing data given a large sample size. We addressed this question using the UK Biobank, a population-scale biobank containing genetic and phenotypic information on over 500,000 individuals across the United Kingdom. We first investigated sources of variation in somatic mutation between individuals and then in genic regions across the genome. Analysis of variation across individuals included the application of a genome-wide association study approach and the final research question pertained to improving power of statistical models that are used in such studies.

6.1 Summary

Somatic mutations can provide insights into both the aetiology of ageing and disease while also informing therapeutic decisions. The study of somatic mutation has been largely limited to cancer cohorts and studies on healthy individuals have been restricted to small sample sizes. The role of somatic mutation within the general population remains understudied. A major challenge in studying somatic mutation in large sample sizes is that the rate at which somatic mutations occur in NGS data is typically below the sequencing error rate. High sequencing depth and specialised NGS library preparation can be used to accurately call somatic mutations. This, however, increases the cost per sample and is not easily scalable to large cohorts.

In this thesis, we postulated that by leveraging the large number of samples in the

UK Biobank we could recover information on sources of variation in somatic mutation from the set of mismatches to the reference genome observed in noisy NGS data. While large sample sizes, improve the power of statistical tests, analysing population-scale genetic data poses several computational challenges. Genomic data is typically embarrassingly parallel; however, the scale of the UK Biobank 200K WES data quickly exhausted the resources available on the University of Galway high-performance computing cluster for Bioinformatics research. Addressing the computational burden of analysing the UK Biobank WES data relied on the use of careful data management and code optimisation. In Chapter 2, we describe a computationally efficient pipeline to extract and annotate mismatches in 175 TB of WES data across 200,632 samples.

To assess the feasibility of our central research aim and as a proof of concept, in Chapter 2 we inferred the expected proportion of somatic mutations within the observed mismatches and used simulations to determine whether we had the power to recover a known source of variation in somatic mutation load from the exome sequencing data. We estimated using empirical estimates of the somatic mutation rate in healthy blood that approximately 0.4% of mismatches arose from somatic mutation. We reasoned that by restricting to mismatches that occurred on overlapping pair-end reads, where both the sequencing reads agree on the mismatch, we could reduce the sequencing error rate to the square of the sequencing error probability, enriching the proportion of somatic mutation within the mismatch data. The use of overlapping reads to correct the sequencing error rate has been previously used in variant calling software [231]. We estimated that by using the overlapping read data, the estimated proportion of somatic mutations in the mismatch data was increased to 1%. To test whether we could plausibly recover modest signals of variation we simulated somatic mutation loads using the estimated somatic mutation rates, age and the number of mismatches to match the variance of the noise within the NGS data. Surprisingly, we found that for the overlapping reads did not provide better power to detect age-associated variation in the somatic mutation load across individuals. We hypothesised two reasons for this result. Firstly, by restricting to mismatches supported by overlapping reads we reduced the total number of mismatches within the data and that increased the variance within the estimate of somatic mutations leaving our study underpowered. Secondly, the somatic mutation rate may be reduced close to the centre of exons relative to exon boundaries and introns, due to nucleosome occupancy. Using the mismatches with no overlapping read restriction we found we could in fact recover the signals of somatic mutation with respect to age within the UK Biobank validating the feasibility of this thesis.

Given that we were able to recover somatic mutation in the UK Biobank data, in Chapter 3 we investigated whether individuals with self-reported cancer or tobacco smoking status had detectably higher mismatch loads. Outside of circulating tumour DNA, it is not known whether cancer status is associated with an increased somatic mutation in adjacent tissues, although it can be argued that cancers driven by Mendelian errors in repair genes or carcinogen exposure may indeed have elevated levels of somatic mutation in non-cancerous tissues. We found significantly higher mismatch loads in tobacco smokers and individuals who had had a self-reported cancer diagnosis, compared to the control group. However, during the verification of the results, we identified a strong batch effect, resulting in much higher numbers of mismatches per sequenced read in some groups of samples (Fig. 2.4). We reasoned that the sequencing batch might influence the mismatch loads because sequencing error can vary between sequencing runs. Although the sequencing batch IDs were not available from the UK Biobank, we retrieved the flowcell IDs from the alignment data and found that the sequencing run explained the batch structure. After adjusting the mismatch data for batch structure, we found that the association between mismatch load and cancer and smoking status was artefactual (Fig 3.1 & 3.2). An attempt to address the sequencing batch structure has been described in the UK biobank for the 450,000-sample exome release [320]; however, the samples were assigned to six clusters of oligo batch. Our results suggest that the guidelines released by the UK Biobank full exome release do not fully address the serious issue of batch structure within the sequencing data. Addressing known sources of technical confounding is an important step to ensure accurate results. The UK Biobank WES data is used to study rare coding variation and its impact on health. A high proportion of spurious rare variant calls could lead to the publication of misleading results, particularly as sequencing batch is non-random, as illustrated by the relationship it induced. Interestingly and most importantly for this thesis, we found that by addressing the batch structure within the UK Biobank exome release the association between age and mismatch load increased.

Transcription coupled-repair and mutagenesis can influence the mutation spectra in a strand-specific manner, i.e. lesions on the non-coding strand are preferentially repaired while the coding strand accumulates DNA damage. This gives rise to an asymmetry in mutation spectra between the coding and non-coding strands. We reasoned that this effect might be recoverable from the exome sequencing data of the UK Biobank. We found a high level of transcription strand-associated asymmetry in C-to-A vs G-to-T mismatches (discussed in Chapters 3 and 4). The preference for C-to-As over G-to-Ts is a well-established result of transcriptional mutagenesis found in both the soma and germline. The level of asymmetry observed in Chapters 3 and 4 was, however, in conflict with an approximate proportion of mismatches arising from somatic mutation (0.4%). DNA damage and sequencing errors accumulate randomly with respect to the Watson and Crick strands and should not be impacted by transcription-associated damage or repair. To explain the high level of transcription-strand asymmetry observed we proposed that aspects of the library preparation protocol induced transcription-strand asymmetry in the mismatch profile. Specifically, the use of a set of oligonucleotides that are complementary to the coding strand during exon capture could preferentially pull down damaged 8-oxoguanines on the coding strand, resulting in a high level of G-to-T mismatches on this strand. Indeed, we were able to confirm from the IDT website that the IDT oligonucleotides used in the UK Biobank exome sequencing data generation target the coding strand only. This is a surprisingly simple explanation, but nonetheless, a result that has not been previously documented and it has implications for NGS QC metrics and rare variant analyses.

Postulating that common genetic variants may contribute to the variation in mismatch loads across samples, we performed GWAS on the mismatch loads. We discovered a genome-wide significant locus, linked to an eQTL for *ERRC8*, a gene essential for transcription-coupled repair (Fig. 3.14). Variation acting upon the repair machinery influencing the mutation rate in a largely healthy population has not been previously discovered and would be a remarkable result. To remove the possibility of the GWAS test SNP being linked to a SNP that escaped stringent SNP filtering and thus influencing the mismatch variation we regenerated the mismatch loads excluding mismatches on the chromosome containing the GWAS hit. This, however, removed the effect, revealing that this result was an artefact, caused by an unusual genetic variant that had escaped filtering in our computational pipeline.

We also investigated whether the previously-reported impact of pathogenic Lynch syndrome variants on the somatic mutation rate could be detected in the mismatch loads observed in individuals carrying these variants. Using mismatch repair mutational signatures, we estimated the contribution of each signature in a group of 160 Lynch syndrome samples and compared to the remaining samples (Fig 3.16). We did not detect significant differences in the contribution of the mismatch repair signatures between groups. Several factors may explain the failure to observe a significantly elevated burden of mismatches in individuals with Lynch syndrome variants. Firstly, of the 200,000 samples only 160 contained pathogenic variants associated with Lynch syndrome, limiting the power to detect an effect, given the noise inherent in the mismatch data, with somatic mutations representing a relatively small proportion of the total. Secondly, recent work has shown that not all pathogenic Lynch variants result in increased somatic mutational loads [328], introducing further noise and decreasing the number of samples that may have increased somatic mutation burdens. Nevertheless, we discovered that SBS3, a signature involved in homology-based repair accumulated linearly with age across samples without Lynch syndrome variants (Fig. 3.17). This is an interesting result as it suggests a potential link between the efficiency of homologous recombination-based repair and ageing, something that has been proposed in theories of ageing but never shown outside of paediatric tumours [339].

Intrinsic biological factors are an important source of somatic mutation rate variation. To explore the effect of intrinsic somatic mutation rate modifiers on the mismatch load, in Chapter 4 we transferred our focus from interindividual variation to variation across the genome. This allowed us to explore the effects of gene expression, GC content, replication timing, chromatin accessibility and recombination hotspots on the derived mismatch loads. Across all mutation rate modifiers, we found that the direction of effect is consistent with published data on somatic mutation rates. This suggests that mismatch loads averaged over large numbers of individuals, could be useful to study variation in somatic mutation processes across the genome. The association between mismatch load and gene expression is particularly interesting as transcription and transcription-coupled repair have a complex relationship with gene expression levels (Fig 4.8). We found that for lowly expressed genes the mismatch load is high, consistent with ineffective transcription-coupled repair. As gene expression levels increase the mismatch load begins to decrease until transcription-associated mutagenesis begins to dominate over the transcription-coupled repair, thus, increasing the mismatch load once again. This effect has been previously published for somatic mutations by *Chen et al.* [148].

As the mismatch burden correlated consistently with different mutation rate modifiers, attempting to further enrich for somatic mutation, we decomposed the recurrent mismatch data into *de novo* mismatch signatures and compared the resulting signatures with validated mutational signatures from the COSMIC database. We estimated twelve mismatch signatures from the data, five of which showed similarity to COSMIC signatures. Importantly a signature similar to SBS1, a clock-like signature active in normal tissue and cancer, was inferred to be present in the mismatch data. This adds further weight to our conclusion in Chapter 2 that we can recover an age signature of somatic mutation from mismatches in the exome sequencing data.

By analysing mismatch recurrence across the genome, we could detect signals of positive selection acting on genes known to be involved in CHIP. As the UK Biobank WES data was generated from whole blood samples, positive selection inferred by analyzing mismatches in a large cohort could be informative about factors that drive clonal expansions in blood. *SRSF2* is an epigenetic modifier frequently mutated in individuals with CHIP. We detected that *SRSF2* was accumulating twice the rate of non-synonymous mutations compared what was expected by chance. An important point to reiterate is that we may have excluded high-frequency somatic mutations in other canonical CHIP genes by removing mismatches that occur in a large number of samples. Hence, additional genes known to be associated with CHIP, such as *DNMT3A*, may also contribute to CHIP within the UK Biobank mismatch data but have been filtered due to high recurrence across samples.

In Chapter 5, we address our final research question. Can we increase the power of GWAS? We present our published work on improving the computational efficiency and power of GWAS. Using state-of-the-art polygenic prediction methods, we incorporated leave-one-chromosome-out polygenic scores as fixed effects into a computationally efficient linear mixed model. Importantly, we found that using our PGS-LMM framework with traits that are sufficiently heritable and polygenic leads to better estimation of SNP effect sizes, therefore increasing the trait prediction capability of the GWAS summary statistics. Although, in the context of this thesis, the method was intended to be applied to infer genetic variation acting upon the somatic mutation rate, this method is not well-suited to traits with small narrow-sense heritability. Nonetheless, recent work has also found that by using a similar method can improve the power of gene-collapsed rare variant burden analyses, aligning with our conclusions [303].

6.2 Concluding remarks

We set out to recover information on somatic mutation within the UK Biobank. Throughout this body of work, we have presented evidence that in some cases the recovery of information on somatic mutation from noisy, low-depth sequencing data is indeed possible. We assessed the feasibility of this study through simulation and validated our results using the UK Biobank exome sequencing data and sample age. We found a linear increase in median mismatch burden with age, indicative of the fact that somatic mutations accumulate linearly throughout life, a result that. was consistent with our simulation study.

We then asked whether we could, using the mismatch data, detect signs of variation in the somatic mutation load across samples. While we found that significant sources of technical variation remained after accounting for the sequencing batch, we could use mutational signatures to uncover an increasing contribution of a COSMIC MMR mutation signature (SBS3) with sample age consistent with both our findings in Chapter 2 and theories of ageing. We next reframed the mismatch data to examine the effects of intrinsic biological mechanisms acting across the genome. Using the median recurrence of a given mismatch context we again found that the mismatch recurrence covaried with several mutation rate modifiers such as replication timing, GC content, chromatin accessibility and mitotic recombination. A key finding was the recovery of a non-linear relationship with gene expression, a result that has been shown in both germline and somatic mutation data previously. Using NMF, we inferred several mutation signatures that showed similarity COSMIC signatures. Of note, we recovered a signature with similarity to SBS1, a clock-like signature that accumulates with age across samples.

Here we present the first attempt to analyse somatic mutations in a population-scale dataset. Our methodology differs from methods designed to call somatic mutations in single-sample NGS data. By using the recurrence across many samples, we can aggregate the signal of somatic mutation to learn information about the variation across both samples and the genome without specifically calling somatic mutations at each site. We present both results that are known to be associated with somatic mutation and novel results that, upon replication, may impact the fields of ageing and rare variant analyses.

6.3 Future directions

The inter-individual and cross-genome variation in somatic mutation rates can shed light on the processes that contribute to both ageing and disease. Here we developed a method to infer somatic mutation within the UK Biobank, we have shown that we can capture variation in the somatic mutation rate across both the genome and across samples by replicating the association with known mutation covariates, detecting positive selection in genes responsible for clonal expansions and mutational signature analyses. Nevertheless, several questions remain to be addressed. For example, we have imposed stringent recurrence filters to remove germline variants. The reintroduction of high-frequency mismatches that have been removed is particularly pertinent for the detection of selection. This could be achieved through analysis of the haplotypes containing the putative high-frequency somatic variant. The implications of the transcription strand asymmetry we observed for variant calling merits further investigation, as it could be informative for pipelines used to call germline and somatic mutations.

Bibliography

[1] R. Dahm, “Friedrich miescher and the discovery of dna,” *Developmental biology*, vol. 278, no. 2, pp. 274–288, 2005.

[2] O. T. Avery, C. M. MacLeod, and M. McCarty, *Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III*. Springer, 2017.

[3] J. D. Watson and F. H. Crick, “Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid,” *Nature*, vol. 171, no. 4356, pp. 737–738, 1953.

[4] M. Meselson and F. W. Stahl, “The replication of dna in escherichia coli,” *Proceedings of the national academy of sciences*, vol. 44, no. 7, pp. 671–682, 1958.

[5] F. Crick, “Central dogma of molecular biology,” *Nature*, vol. 227, no. 5258, pp. 561–563, 1970.

[6] D. Stehelin, H. E. Varmus, J. M. Bishop, and P. K. Vogt, “Dna related to the transforming gene (s) of avian sarcoma viruses is present in normal avian dna,” *Nature*, vol. 260, no. 5547, pp. 170–173, 1976.

[7] F. Sanger, S. Nicklen, and A. R. Coulson, “Dna sequencing with chain-terminating inhibitors,” *Proceedings of the national academy of sciences*, vol. 74, no. 12, pp. 5463–5467, 1977.

[8] M. Kimura, “A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences,” *Journal of molecular evolution*, vol. 16, pp. 111–120, 1980.

[9] J. Felsenstein, “Evolutionary trees from dna sequences: a maximum likelihood approach,” *Journal of molecular evolution*, vol. 17, pp. 368–376, 1981.

[10] T. H. Jukes, C. R. Cantor, *et al.*, “Evolution of protein molecules,” *Mammalian protein metabolism*, vol. 3, pp. 21–132, 1969.

[11] E. J. Douzery, E. A. Snell, E. Bapteste, F. Delsuc, and H. Philippe, “The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils?,” *Proceedings of the National Academy of Sciences*, vol. 101, no. 43, pp. 15386–15391, 2004.

[12] M. J. Williams, B. Werner, C. P. Barnes, T. A. Graham, and A. Sottoriva, “Identification of neutral tumor evolution across cancer types,” *Nature genetics*, vol. 48, no. 3, pp. 238–244, 2016.

[13] I. Martincorena, K. M. Raine, M. Gerstung, K. J. Dawson, K. Haase, P. Van Loo, H. Davies, M. R. Stratton, and P. J. Campbell, “Universal patterns of selection in cancer and somatic tissues,” *Cell*, vol. 171, no. 5, pp. 1029–1041, 2017.

[14] L. H. Cavallari, “Tailoring drug therapy based on genotype,” *J Pharm Pract*, vol. 25, pp. 413–416, Aug 2012.

[15] A. V. Khera, M. Chaffin, K. G. Aragam, M. E. Haas, C. Roselli, S. H. Choi, P. Natarajan, E. S. Lander, S. A. Lubitz, P. T. Ellinor, *et al.*, “Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations,” *Nature genetics*, vol. 50, no. 9, pp. 1219–1224, 2018.

[16] A. Annunziato, “Dna packaging: Nucleosomes and chromatin. nature education 1: 26,” 2008.

[17] A. Ghosh and M. Bansal, “A glossary of dna structures from a to z,” *Acta Crystallographica Section D: Biological Crystallography*, vol. 59, no. 4, pp. 620–626, 2003.

[18] C. O. Pabo and R. T. Sauer, “Protein-dna recognition,” *Annual review of biochemistry*, vol. 53, no. 1, pp. 293–321, 1984.

[19] B. A. Bouwman and W. De Laat, “Getting the genome in shape: the formation of loops, domains and compartments,” *Genome biology*, vol. 16, no. 1, pp. 1–9, 2015.

[20] R. Sender and R. Milo, “The distribution of cellular turnover in the human body,” *Nature medicine*, vol. 27, no. 1, pp. 45–48, 2021.

[21] A. Koren, R. E. Handsaker, N. Kamitaki, R. Karlić, S. Ghosh, P. Polak, K. Eggan, and S. A. McCarroll, “Genetic variation in human dna replication timing,” *Cell*, vol. 159, no. 5, pp. 1015–1026, 2014.

[22] D. Mas-Ponte and F. Supek, “DNA mismatch repair promotes APOBEC3-mediated diffuse hypermutation in human cancers,” *Nat Genet*, vol. 52, pp. 958–968, Sep 2020.

[23] C. Tomasetti, L. Li, and B. Vogelstein, “Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention,” *Science*, vol. 355, no. 6331, pp. 1330–1334, 2017.

[24] P. S. Robinson, T. H. H. Coorens, C. Palles, E. Mitchell, F. Abascal, S. Olafsson, B. C. H. Lee, A. R. J. Lawson, H. Lee-Six, L. Moore, M. A. Sanders, J. Hewinson, L. Martin, C. M. A. Pinna, S. Galavotti, R. Rahbari, P. J. Campbell, I. Martincorena, I. Tomlinson, and M. R. Stratton, “Increased somatic mutation burdens in normal human cells due to defective DNA polymerases,” *Nat Genet*, vol. 53, pp. 1434–1442, Oct 2021.

[25] D. Prescott and P. Kuempel, “Bidirectional replication of the chromosome in escherichia coli,” *Proceedings of the National Academy of Sciences*, vol. 69, no. 10, pp. 2842–2845, 1972.

[26] R. Okazaki, T. Okazaki, K. Sakabe, K. Sugimoto, and A. Sugino, “Mechanism of dna chain growth. i. possible discontinuity and unusual secondary structure of newly synthesized chains.,” *Proceedings of the National Academy of Sciences*, vol. 59, no. 2, pp. 598–605, 1968.

[27] J. Josse, A. Kaiser, and A. Kornberg, “Enzymatic synthesis of deoxyribonucleic acid,” *J biol chem*, vol. 236, no. 3, pp. 864–875, 1961.

[28] J. J. Hopfield, “Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity,” *Proceedings of the National Academy of Sciences*, vol. 71, no. 10, pp. 4135–4139, 1974.

[29] A. Bebenek and I. Ziuzia-Graczyk, “Fidelity of dna replication-a matter of proofreading,” *Current Genetics*, vol. 64, no. 5, pp. 985–996, 2018.

[30] L. Toledo, K. J. Neelsen, and J. Lukas, “Replication catastrophe: when a checkpoint fails because of exhaustion,” *Molecular cell*, vol. 66, no. 6, pp. 735–749, 2017.

[31] W. Yang, “An overview of y-family dna polymerases and a case study of human dna polymerase η,” *Biochemistry*, vol. 53, no. 17, pp. 2793–2803, 2014.

[32] L. R. Ferguson, H. Chen, A. R. Collins, M. Connell, G. Damia, S. Dasgupta, M. Malhotra, A. K. Meeker, A. Amedei, A. Amin, *et al.*, “Genomic instability in human cancer: Molecular insights and opportunities for therapeutic attack and prevention through diet and nutrition,” in *Seminars in cancer biology*, vol. 35, pp. S5–S24, Elsevier, 2015.

[33] J. Vijg and Y. Suh, “Genome instability and aging,” *Annual review of physiology*, vol. 75, pp. 645–668, 2013.

[34] T. Lindahl and D. Barnes, “Repair of endogenous dna damage,” in *Cold Spring Harbor symposia on quantitative biology*, vol. 65, pp. 127–134, Cold Spring Harbor Laboratory Press, 2000.

[35] S. P. Jackson and J. Bartek, “The dna-damage response in human biology and disease,” *Nature*, vol. 461, no. 7267, pp. 1071–1078, 2009.

[36] S. Solier, Y. W Zhang, A. Ballestrero, Y. Pommier, and G. Zoppoli, “Dna damage response pathways and cell cycle checkpoints in colorectal cancer: current concepts and future perspectives for targeted treatment,” *Current cancer drug targets*, vol. 12, no. 4, pp. 356–371, 2012.

[37] L. Stojic, R. Brun, and J. Jiricny, “Mismatch repair and dna damage signalling,” *DNA repair*, vol. 3, no. 8-9, pp. 1091–1101, 2004.

[38] J. Y. Wang, “Focus: death: cell death response to dna damage,” *The Yale journal of biology and medicine*, vol. 92, no. 4, p. 771, 2019.

[39] C. Bruhn and M. Foiani, “A model of dna damage response activation at stalled replication forks by sprtn,” *Nature Communications*, vol. 10, no. 1, p. 5671, 2019.

[40] F. Rossiello, U. Herbig, M. P. Longhese, M. Fumagalli, and F. d. di Fagagna, “Irreparable telomeric dna damage and persistent ddr signalling as a shared causative mechanism of cellular senescence and ageing,” *Current opinion in genetics & development*, vol. 26, pp. 89–95, 2014.

[41] S. Kakoti, H. Sato, S. Laskar, T. Yasuhara, and A. Shibata, “Dna repair and signaling in immune-related cancer therapy,” *Frontiers in Molecular Biosciences*, vol. 7, p. 205, 2020.

[42] T. Qing, T. Jun, K. E. Lindblad, A. Lujambio, M. Marczyk, L. Pusztai, and K.-l. Huang, “Diverse immune response of dna damage repair-deficient tumors,” *Cell Reports Medicine*, vol. 2, no. 5, p. 100276, 2021.

[43] M. Rose, J. T. Burgess, K. O’Byrne, D. J. Richard, and E. Bolderson, “Parp inhibitors: clinical relevance, mechanisms of action and tumor resistance,” *Frontiers in cell and developmental biology*, vol. 8, p. 564601, 2020.

[44] A. A. Larrea, S. A. Lujan, and T. A. Kunkel, "SnapShot: DNA mismatch repair," *Cell*, vol. 141, p. 730.e1, May 2010.

[45] M. Lynch, "Mutation and human exceptionalism: our future genetic load," *Genetics*, vol. 202, no. 3, pp. 869–875, 2016.

[46] O. G. W. Wong, J. Li, and A. N. Y. Cheung, "Targeting DNA Damage Response Pathway in Ovarian Clear Cell Carcinoma," *Front Oncol*, vol. 11, p. 666815, 2021.

[47] "Pan-cancer analysis of whole genomes," *Nature*, vol. 578, no. 7793, pp. 82–93, 2020.

[48] E. Levy-Lahad and M.-C. King, "Hiding in plain sight-somatic mutation in human disease," 2020.

[49] L. Szilard, "On the nature of the aging process," *Proceedings of the National Academy of Sciences*, vol. 45, no. 1, pp. 30–45, 1959.

[50] I. Martincorena, J. C. Fowler, A. Wabik, A. R. Lawson, F. Abascal, M. W. Hall, A. Cagan, K. Murai, K. Mahbubani, M. R. Stratton, *et al.*, "Somatic mutant clones colonize the human esophagus with age," *Science*, vol. 362, no. 6417, pp. 911–917, 2018.

[51] K. C. Higa and J. DeGregori, "Decoy fitness peaks, tumor suppression, and aging," *Aging Cell*, vol. 18, p. e12938, June 2019.

[52] N. H. Yamaguchi, "Smoking, immunity, and DNA damage," *Transl Lung Cancer Res*, vol. 8, pp. S3–S6, May 2019.

[53] F. E. Imad, H. Drissi, N. Tawfiq, K. Bendahhou, A. Benider, and D. Radallah, "A case-control study on dietary risk factors for colorectal cancer in Morocco," *Pan Afr Med J*, vol. 35, p. 59, 2020.

[54] R. P. Rastogi, A. Kumar, M. B. Tyagi, R. P. Sinha, *et al.*, "Molecular mechanisms of ultraviolet radiation-induced dna damage and repair," *Journal of nucleic acids*, vol. 2010, 2010.

[55] L. B. Alexandrov, P. H. Jones, D. C. Wedge, J. E. Sale, P. J. Campbell, S. Nik-Zainal, and M. R. Stratton, "Clock-like mutational processes in human somatic cells," *Nat Genet*, vol. 47, pp. 1402–1407, Dec 2015.

[56] B. Milholland, X. Dong, L. Zhang, X. Hao, Y. Suh, and J. Vijg, "Differences between germline and somatic mutation rates in humans and mice," *Nat Commun*, vol. 8, p. 15183, May 2017.

[57] R. Wagner Jr and M. Meselson, “Repair tracts in mismatched dna heteroduplexes.,” *Proceedings of the National Academy of Sciences*, vol. 73, no. 11, pp. 4135–4139, 1976.

[58] P. Modrich, “Mechanisms in E. coli and Human Mismatch Repair (Nobel Lecture),” *Angew Chem Int Ed Engl*, vol. 55, pp. 8490–8501, Jul 2016.

[59] C. D. Putnam, “Strand discrimination in DNA mismatch repair,” *DNA Repair (Amst)*, vol. 105, p. 103161, Sep 2021.

[60] J. E. Sale, “Translesion dna synthesis and mutagenesis in eukaryotes,” *Cold Spring Harbor perspectives in biology*, vol. 5, no. 3, p. a012708, 2013.

[61] S. D. McCulloch and T. A. Kunkel, “The fidelity of dna synthesis by eukaryotic replicative and translesion synthesis polymerases,” *Cell research*, vol. 18, no. 1, pp. 148–161, 2008.

[62] G. Levinson and G. A. Gutman, “Slipped-strand mispairing: a major mechanism for DNA sequence evolution,” *Mol Biol Evol*, vol. 4, pp. 203–221, May 1987.

[63] S. T. Lovett, P. T. Drapkin, V. A. Sutera, and T. J. Gluckman-Peskind, “A sister-strand exchange mechanism for recA-independent deletion of repeated DNA sequences in Escherichia coli,” *Genetics*, vol. 135, pp. 631–642, Nov 1993.

[64] V. Tiwari and D. M. Wilson, “DNA Damage and Associated DNA Repair Defects in Disease and Premature Aging,” *Am J Hum Genet*, vol. 105, pp. 237–257, Aug 2019.

[65] J. Cadet and K. J. A. Davies, “repair: An introduction,” *Free Radic Biol Med*, vol. 107, pp. 2–12, Jun 2017.

[66] C. A. Juan, J. M. rez de la Lastra, F. J. Plou, and E. a, “The Chemistry of Reactive Oxygen Species (ROS) Revisited: Outlining Their Role in Biological Macromolecules (DNA, Lipids and Proteins) and Induced Pathologies,” *Int J Mol Sci*, vol. 22, Apr 2021.

[67] A. R. Poetsch, “The genomics of oxidative DNA damage, repair, and resulting mutagenesis,” *Comput Struct Biotechnol J*, vol. 18, pp. 207–219, 2020.

[68] G. S. Madugundu, J. Cadet, and J. R. Wagner, “Hydroxyl-radical-induced oxidation of 5-methylcytosine in isolated and cellular DNA,” *Nucleic Acids Res*, vol. 42, pp. 7450–7460, Jun 2014.

[69] G. Borrego-Soto, R. pez, and A. nez, "Ionizing radiation-induced DNA injury and damage detection in patients with breast cancer," *Genet Mol Biol*, vol. 38, pp. 420–432, Dec 2015.

[70] H. B. Huang, C. H. Lai, G. W. Chen, Y. Y. Lin, J. J. Jaakkola, S. H. Liou, and S. L. Wang, "Traffic-related air pollution and DNA damage: a longitudinal study in Taiwanese traffic conductors," *PLoS One*, vol. 7, no. 5, p. e37412, 2012.

[71] J. L. Barnes, M. Zubair, K. John, M. C. Poirier, and F. L. Martin, "Carcinogens and DNA damage," *Biochem Soc Trans*, vol. 46, pp. 1213–1224, Oct 2018.

[72] W. H. Feng, K. S. Xue, L. Tang, P. L. Williams, and J. S. Wang, "-Induced Developmental and DNA Damage in Caenorhabditis elegans," *Toxins (Basel)*, vol. 9, Dec 2016.

[73] J. Byrne, "Long-term genetic and reproductive effects of ionizing radiation and chemotherapeutic agents on cancer patients and their offspring," *Teratology*, vol. 59, pp. 210–215, Apr 1999.

[74] M. E. Lomax, L. K. Folkes, and P. O'Neill, "Biological consequences of radiation-induced DNA damage: relevance to radiotherapy," *Clin Oncol (R Coll Radiol)*, vol. 25, pp. 578–585, Oct 2013.

[75] S. Marchese, A. Polo, A. Ariano, S. Velotto, S. Costantini, and L. Severino, "Aflatoxin B1 and M1: Biological Properties and Their Involvement in Cancer Development," *Toxins (Basel)*, vol. 10, May 2018.

[76] J. D. Groopman, J. W. Smith, A. Rivera-Andrade, C. S. Alvarez, M. F. Kroker-Lobos, P. A. Egner, E. Gharzouzi, M. Dean, K. A. McGlynn, and M. rez Zea, "Aflatoxin and the Etiology of Liver Cancer and Its Implications for Guatemala," *World Mycotoxin J*, vol. 14, no. 3, pp. 305–317, 2021.

[77] W. W. Johnson and F. P. Guengerich, "Reaction of aflatoxin B1 exo-8,9-epoxide with DNA: kinetic analysis of covalent binding and DNA-induced hydrolysis," *Proc Natl Acad Sci U S A*, vol. 94, pp. 6121–6125, Jun 1997.

[78] M. N. Huang, W. Yu, W. W. Teoh, M. Ardin, A. Jusakul, A. W. T. Ng, A. Boot, B. Abedi-Ardekani, S. Villar, S. S. Myint, R. Othman, S. L. Poon, A. Heguy, M. Olivier, M. Hollstein, P. Tan, B. T. Teh, K. Sabapathy, J. Zavadil, and S. G. Rozen, "Genome-scale mutational signatures of aflatoxin in cells, mice, and human tumors," *Genome Res*, vol. 27, pp. 1475–1486, Sep 2017.

[79] X. C. Li, M. Y. Wang, M. Yang, H. J. Dai, B. F. Zhang, W. Wang, X. L. Chu, X. Wang, H. Zheng, R. F. Niu, W. Zhang, and K. X. Chen, “A mutational signature associated with alcohol consumption and prognostically significantly mutated driver genes in esophageal squamous cell carcinoma,” *Ann Oncol*, vol. 29, pp. 938–944, Apr 2018.

[80] P. Davalli, G. Marverti, A. Lauriola, and D. D’Arca, “Targeting Oxidatively Induced DNA Damage Response in Cancer: Opportunities for Novel Cancer Therapies,” *Oxid Med Cell Longev*, vol. 2018, p. 2389523, 2018.

[81] T. Wicker, Y. Yu, G. Haberer, K. F. Mayer, P. R. Marri, S. Rounsley, M. Chen, A. Zuccolo, O. Panaud, R. A. Wing, and S. Roffler, “DNA transposon activity is associated with increased mutation rates in genes of rice and other grasses,” *Nat Commun*, vol. 7, p. 12790, Sep 2016.

[82] S. L. Gasior, T. P. Wakeman, B. Xu, and P. L. Deininger, “The human LINE-1 retrotransposon creates DNA double-strand breaks,” *J Mol Biol*, vol. 357, pp. 1383–1393, Apr 2006.

[83] A. Deem, A. Keszthelyi, T. Blackgrove, A. Vayl, B. Coffey, R. Mathur, A. Chabes, and A. Malkova, “Break-induced replication is highly inaccurate,” *PLoS Biol*, vol. 9, p. e1000594, Feb 2011.

[84] S. Nik-Zainal, L. B. Alexandrov, D. C. Wedge, P. Van Loo, C. D. Greenman, K. Raine, D. Jones, J. Hinton, J. Marshall, L. A. Stebbings, A. Menzies, S. Martin, K. Leung, L. Chen, C. Leroy, M. Ramakrishna, R. Rance, K. W. Lau, L. J. Mudie, I. Varela, D. J. McBride, G. R. Bignell, S. L. Cooke, A. Shlien, J. Gamble, I. Whitmore, M. Maddison, P. S. Tarpey, H. R. Davies, E. Papaemmanuil, P. J. Stephens, S. McLaren, A. P. Butler, J. W. Teague, G. nsson, J. E. Garber, D. Silver, P. Miron, A. Fatima, S. Boyault, A. d, A. Tutt, J. W. Martens, S. A. Aparicio, Ã. Borg, A. V. Salomon, G. Thomas, A. L. rresen Dale, A. L. Richardson, M. S. Neuberger, P. A. Futreal, P. J. Campbell, and M. R. Stratton, “Mutational processes molding the genomes of 21 breast cancers,” *Cell*, vol. 149, pp. 979–993, May 2012.

[85] R. Pecori, S. Di Giorgio, J. Paulo Lorenzo, and F. Nina Papavasiliou, “Functions and consequences of AID/APOBEC-mediated DNA and RNA deamination,” *Nat Rev Genet*, vol. 23, pp. 505–518, Aug 2022.

[86] M. Muramatsu, V. Sankaranand, S. Anant, M. Sugai, K. Kinoshita, N. O. Davidson, and T. Honjo, “Specific expression of activation-induced cytidine deaminase (aid), a novel

member of the rna-editing deaminase family in germinal center b cells," *Journal of Biological Chemistry*, vol. 274, no. 26, pp. 18470–18476, 1999.

[87] D. L. French, R. Laskov, and M. D. Scharff, "The role of somatic hypermutation in the generation of antibody diversity," *Science*, vol. 244, pp. 1152–1157, Jun 1989.

[88] D. Mechtcheriakova, M. Svoboda, A. Meshcheryakova, and E. Jensen-Jarolim, "Activation-induced cytidine deaminase (AID) linking immunity, chronic inflammation, and cancer," *Cancer Immunol Immunother*, vol. 61, pp. 1591–1598, Sep 2012.

[89] R. Buisson, A. Langenbucher, D. Bowen, E. E. Kwan, C. H. Benes, L. Zou, and M. S. Lawrence, "Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features," *Science*, vol. 364, Jun 2019.

[90] Z. Xu, E. J. Pone, A. Al-Qahtani, S. R. Park, H. Zan, and P. Casali, "Regulation of aicda expression and AID activity: relevance to somatic hypermutation and class switch DNA recombination," *Crit Rev Immunol*, vol. 27, no. 4, pp. 367–397, 2007.

[91] J. Gao, H. Choudhry, and W. Cao, "Apolipoprotein B mRNA editing enzyme catalytic polypeptide-like family genes activation and regulation during tumorigenesis," *Cancer Sci*, vol. 109, pp. 2375–2382, Aug 2018.

[92] B. J. Taylor, S. Nik-Zainal, Y. L. Wu, L. A. Stebbings, K. Raine, P. J. Campbell, C. Rada, M. R. Stratton, and M. S. Neuberger, "Dna deaminases induce break-associated mutation showers with implication of apobec3b and 3a in breast cancer kataegis," *elife*, vol. 2, p. e00534, 2013.

[93] I. Martincorena, A. Roshan, M. Gerstung, P. Ellis, P. Van Loo, S. McLaren, D. C. Wedge, A. Fullam, L. B. Alexandrov, J. M. Tubio, *et al.*, "High burden and pervasive positive selection of somatic mutations in normal human skin," *Science*, vol. 348, no. 6237, pp. 880–886, 2015.

[94] W. J. Cannan and D. S. Pederson, "Mechanisms and Consequences of Double-Strand DNA Break Formation in Chromatin," *J Cell Physiol*, vol. 231, pp. 3–14, Jan 2016.

[95] G. P. Pfeifer, M. F. Denissenko, M. Olivier, N. Tretyakova, S. S. Hecht, and P. Hainaut, "Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers," *Oncogene*, vol. 21, pp. 7435–7451, Oct 2002.

[96] B. C. Bade and C. S. Dela Cruz, "Lung Cancer 2020: Epidemiology, Etiology, and Prevention," *Clin Chest Med*, vol. 41, pp. 1–24, Mar 2020.

[97] M. B. Reitsma, N. Fullman, M. Ng, J. S. Salama, A. Abajobir, K. H. Abate, C. Abbafati, S. F. Abera, B. Abraham, G. Y. Abyu, A. O. Adebiyi, Z. Al-Aly, A. V. Aleman, R. Ali, A. Al Alkerwi, P. Allebeck, R. M. Al-Raddadi, A. T. Amare, A. Amberbir, W. Ammar, S. M. Amrock, C. A. T. Antonio, H. Asayesh, N. T. Atnafu, P. Azzopardi, A. Banerjee, A. Barac, T. Barrientos-Gutierrez, A. C. Basto-Abreu, S. Bazargan-Hejazi, N. Bedi, B. Bell, A. K. Bello, I. M. Bensenor, A. S. Beyene, N. Bhala, S. Biryukov, K. Bolt, H. Brenner, Z. Butt, F. Cavalleri, K. Cercy, H. Chen, D. J. Christopher, L. G. Ciobanu, V. Colistro, M. Colomar, L. Cornaby, X. Dai, S. A. Damtew, L. Dandona, R. Dandona, E. Dansereau, K. Davletov, A. Dayama, T. T. Degfie, A. Deribew, S. D. Dharmaratne, B. D. Dimtsu, K. E. Doyle, A. Y. Endries, S. P. Ermakov, K. Estep, E. J. A. Faraon, F. Farzadfar, V. L. Feigin, A. B. Feigl, F. Fischer, J. Friedman, T. T. G/Hiwot, S. L. Gall, W. Gao, R. F. Gillum, A. L. Gold, S. V. Gopalani, C. C. Gotay, R. Gupta, R. Gupta, V. Gupta, R. R. Hamadeh, G. Hankey, H. L. Harb, S. I. Hay, M. Horino, N. Horita, H. D. Hosgood, A. Husseini, B. V. Ileanu, F. Islami, G. Jiang, Y. Jiang, J. B. Jonas, Z. Kabir, R. Kamal, A. Kasaeian, C. N. Kesavachandran, Y. S. Khader, I. Khalil, Y. H. Khang, S. Khera, J. Khubchandani, D. Kim, Y. J. Kim, R. W. Kimokoti, Y. Kinfu, L. D. Knibbs, Y. Kokubo, D. Kolte, J. Kopec, S. Kosen, G. A. Kotsakis, P. A. Koul, A. Koyanagi, K. J. Krohn, H. Krueger, B. K. Defo, B. K. Bicer, C. Kulkarni, G. A. Kumar, J. L. Leasher, A. Lee, M. Leinsalu, T. Li, S. Linn, P. Liu, S. Liu, L. T. Lo, A. D. Lopez, S. Ma, H. M. A. El Razek, A. Majeed, R. Malekzadeh, D. C. Malta, W. A. Manamo, J. Martinez-Raga, A. B. Mekonnen, W. Mendoza, T. R. Miller, K. A. Mohammad, L. Morawska, K. I. Musa, G. Nagel, S. P. Neupane, Q. Nguyen, G. Nguyen, I. H. Oh, A. S. Oyekale, M. Pa, A. Pana, E. K. Park, S. T. Patil, G. C. Patton, J. Pedro, M. Qorbani, A. Rafay, M. Rahman, R. K. Rai, U. Ram, C. L. Ranabhat, A. H. Refaat, N. Reinig, H. S. Roba, A. Rodriguez, Y. Roman, G. Roth, A. Roy, R. Sagar, J. A. Salomon, J. Sanabria, I. de Souza Santos, B. Sartorius, M. Satpathy, M. Sawhney, S. Sawyer, M. Saylan, M. P. Schaub, N. Schluger, A. E. Schutte, S. G. Sepanlou, B. Serdar, M. A. Shaikh, J. She, M. J. Shin, R. Shiri, K. Shishani, I. Shiue, I. D. Sigfusdottir, J. I. Silverberg, J. Singh, V. Singh, E. L. Slepak, S. Soneji, J. B. Soriano, S. Soshnikov, C. T. Sreeramareddy, D. J. Stein, S. Stranges, M. L. Subart, S. Swaminathan, C. E. I. Szoeke, W. M. Tefera, R. Topor-Madry, B. Tran, N. Tsilimparis, H. Tymeson, K. N. Ukwaja, R. Updike, O. A. Uthman, F. S. Violante, S. K. Vladimirov, V. Vlassov, S. E. Vollset, T. Vos, E. Weiderpass, C. P. Wen, A. Werdecker, S. Wilson, M. Wubshet, L. Xiao, B. Yakob, Y. Yano, P. Ye, N. Yonemoto, S. J. Yoon, M. Z. Younis, C. Yu, Z. Zaidi, M. El Sayed Zaki, A. L. Zhang, B. Zipkin, C. J. L. Murray, M. H. Forouzanfar, and E. Gakidou, "Smoking prevalence and attributable disease burden in

195 countries and territories, 1990-2015: a systematic analysis from the Global Burden of Disease Study 2015," *Lancet*, vol. 389, pp. 1885–1906, May 2017.

[98] S. S. Hecht, "Research opportunities related to establishing standards for tobacco products under the Family Smoking Prevention and Tobacco Control Act," *Nicotine Tob Res*, vol. 14, pp. 18–28, Jan 2012.

[99] J. Fahrer and M. Christmann, "DNA Alkylation Damage by Nitrosamines and Relevant DNA Repair Pathways," *Int J Mol Sci*, vol. 24, Feb 2023.

[100] M. Zapatka, I. Borozan, D. S. Brewer, M. Iskar, A. Grundhoff, M. Alawi, N. Desai, H. ltmann, H. Moch, C. S. Cooper, R. Eils, V. Ferretti, P. Lichter, M. Alawi, I. Borozan, D. S. Brewer, C. S. Cooper, N. Desai, R. Eils, V. Ferretti, A. Grundhoff, M. Iskar, K. Kleinheinz, P. Lichter, H. Nakagawa, A. I. Ojesina, C. S. Pedamallu, M. Schlesner, X. Su, and M. Zapatka, "The landscape of viral associations in human cancers," *Nat Genet*, vol. 52, pp. 320–330, Mar 2020.

[101] M. Plummer, C. de Martel, J. Vignat, J. Ferlay, F. Bray, and S. Franceschi, "Global burden of cancers attributable to infections in 2012: a synthetic analysis," *Lancet Glob Health*, vol. 4, pp. e609–616, Sep 2016.

[102] A. S. Mekawy, Z. Alaswad, A. A. Ibrahim, A. A. Mohamed, A. AlOkda, and M. Elserafy, "The consequences of viral infection on host DNA damage response: a focus on SARS-CoVs," *J Genet Eng Biotechnol*, vol. 20, p. 104, Jul 2022.

[103] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A. L. rresen Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. rd, J. A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinski, N. ger, D. T. Jones, D. Jones, S. Knappskog, M. Kool, S. R. Lakhani, C. n, S. Martin, N. C. Munshi, H. Nakamura, P. A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V. Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki, A. N. Tutt, R. s Mas, M. M. van Buuren, L. van 't Veer, A. Vincent-Salomon, N. Waddell, L. R. Yates, J. Zucman-Rossi, P. A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell, M. R. Stratton, A. Claviez, A. Rosenwald, A. Rosenwald, A. Borkhardt, B. Brors, B. Radlwimmer, C. Lawerenz, C. Lopez, D. Langenberger, D. Karsch, D. Lenze, D. Kube, E. Leich, G. Richter, J. Korbel, J. Hoell,

J. Eils, K. Hezaveh, L. mper, M. Rosolowski, M. Weniger, M. Rohde, M. Kreuz, M. Loeffler, M. Schilhabel, M. Dreyling, M. L. Hansmann, M. Hummel, M. Szczepanowski, O. Ammerpohl, P. F. Stadler, P. ller, R. ppers, S. Haas, S. Eberth, S. Schreiber, S. H. Bernhart, S. Hoffmann, S. Radomski, U. Kostezka, W. Klapper, C. Sotiriou, D. Larsimont, D. Vincent, M. Maetens, O. Mariani, A. M. Sieuwerts, J. W. Martens, J. G. Jonasson, I. Treilleux, E. Thomas, G. Mac Grogan, C. Mannina, L. Arnould, L. Burillier, J. L. Merlin, M. Lefebvre, F. Bibeau, B. Massemin, F. Penault-Llorca, Q. Lopez, M. C. Mathieu, P. E. Lonning, M. Schlooz-Vries, J. Tol, H. van Laarhoven, F. Sweep, and P. Bult, "Signatures of mutational processes in human cancer," *Nature*, vol. 500, pp. 415–421, Aug 2013.

[104] T. Helleday, S. Eshtad, and S. Nik-Zainal, "Mechanisms underlying mutational signatures in human cancers," *Nat Rev Genet*, vol. 15, pp. 585–598, Sep 2014.

[105] M. Gerstung, C. Jolly, I. Leshchiner, S. C. Dentro, S. Gonzalez, D. Rosebrock, T. J. Mitchell, Y. Rubanova, P. Anur, K. Yu, M. Tarabichi, A. Deshwar, J. Wintersinger, K. Kleinheinz, I. a, K. Haase, L. Jerman, S. Sengupta, G. Macintyre, S. Malikic, N. Donmez, D. G. Livitz, M. Cmero, J. Demeulemeester, S. Schumacher, Y. Fan, X. Yao, J. Lee, M. Schlesner, P. C. Boutros, D. D. Bowtell, H. Zhu, G. Getz, M. Imielinski, R. Beroukhim, S. C. Sahinalp, Y. Ji, M. Peifer, F. Markowetz, V. Mustonen, K. Yuan, W. Wang, Q. D. Morris, P. T. Spellman, D. C. Wedge, P. Van Loo, S. C. Dentro, I. Leshchiner, M. Gerstung, C. Jolly, K. Haase, M. Tarabichi, J. Wintersinger, A. G. Deshwar, K. Yu, S. Gonzalez, Y. Rubanova, G. Macintyre, D. J. Adams, P. Anur, R. Beroukhim, P. C. Boutros, D. D. Bowtell, P. J. Campbell, S. Cao, E. L. Christie, M. Cmero, Y. Cun, K. J. Dawson, J. Demeulemeester, N. Donmez, R. M. Drews, R. Eils, Y. Fan, M. Fittall, D. W. Garsed, G. Getz, G. Ha, M. Imielinski, L. Jerman, Y. Ji, K. Kleinheinz, J. Lee, H. Lee-Six, D. G. Livitz, S. Malikic, F. Markowetz, I. Martincorena, T. J. Mitchell, V. Mustonen, L. Oesper, M. Peifer, M. Peto, B. J. Raphael, D. Rosebrock, S. C. Sahinalp, A. Salcedo, M. Schlesner, S. Schumacher, S. Sengupta, R. Shi, S. J. Shin, O. Spiro, L. D. Stein, I. a, S. Vembu, D. A. Wheeler, T. P. Yang, X. Yao, K. Yuan, H. Zhu, W. Wang, Q. D. Morris, P. T. Spellman, D. C. Wedge, P. Van Loo, L. A. Aaltonen, F. Abascal, A. Abeshouse, H. Aburatani, D. J. Adams, N. Agrawal, K. S. Ahn, S. M. Ahn, H. Aikata, R. Akbani, K. C. Akdemir, H. Al-Ahmadie, S. T. Al-Sedairy, F. Al-Shahrour, M. Alawi, M. Albert, K. Aldape, L. B. Alexandrov, A. Ally, K. Alsop, E. G. Alvarez, F. Amary, S. B. Amin, B. Aminou, O. Ammerpohl, M. J. Anderson, Y. Ang, D. Antonello, P. Anur, S. Aparicio, E. L. Appelbaum, Y. Arai, A. Aretz,

K. Arihiro, S. I. Ariizumi, J. Armenia, L. Arnould, S. Asa, Y. Assenov, G. Atwal, S. Aukema, J. T. Auman, M. R. R. Aure, P. Awadalla, M. Aymerich, G. D. Bader, A. Baez-Ortega, M. H. Bailey, P. J. Bailey, M. Balasundaram, S. Balu, P. Bandopadhayay, R. E. Banks, S. Barbi, A. P. Barbour, J. Barenboim, J. Barnholtz-Sloan, H. Barr, E. Barrera, J. Bartlett, J. Bartolome, C. Bassi, O. F. Bathe, D. Baumhoer, P. Bavi, S. B. Baylin, W. Bazant, D. Beardsmore, T. A. Beck, S. Behjati, A. Behren, B. Niu, C. Bell, S. Beltran, C. Benz, A. Berchuck, A. K. Bergmann, E. N. Bergstrom, B. P. Berman, D. M. Berney, S. H. Bernhart, R. Beroukhim, M. Berrios, S. Bersani, J. Bertl, M. Betancourt, V. Bhandari, S. G. Bhosle, A. V. Biankin, M. Bieg, D. Bigner, H. Binder, E. Birney, M. Birrer, N. K. Biswas, B. Bjerkehagen, T. Bodenheimer, L. Boice, G. Bonizzato, J. S. De Bono, A. Boot, M. S. Bootwalla, A. Borg, A. Borkhardt, K. A. Boroevich, I. Borozan, C. Borst, M. Bosenberg, M. Bosio, J. Boultwood, G. Bourque, P. C. Boutros, G. S. Bova, D. T. Bowen, R. Bowlby, D. D. L. Bowtell, S. Boyault, R. Boyce, J. Boyd, A. Brazma, P. Brennan, D. S. Brewer, A. B. Brinkman, R. G. Bristow, R. R. Broaddus, J. E. Brock, M. Brock, A. Broeks, A. N. Brooks, D. Brooks, B. Brors, S. Brunak, T. J. C. Bruxner, A. L. Bruzos, A. Buchanan, I. Buchhalter, C. Buchholz, S. Bullman, H. Burke, B. Burkhardt, K. H. Burns, J. Busanovich, C. D. Bustamante, A. P. Butler, A. J. Butte, N. J. Byrne, A. L. rresen Dale, S. J. Caesar-Johnson, A. Cafferkey, D. Cahill, C. Calabrese, C. Caldas, F. Calvo, N. Camacho, P. J. Campbell, E. Campo, C. u, S. Cao, T. E. Carey, J. Carlevaro-Fita, R. Carlsen, I. Cataldo, M. Cazzola, J. Cebon, R. Cerfolio, D. E. Chadwick, D. Chakravarty, D. Chalmers, C. W. Y. Chan, K. Chan, M. Chan-Seng-Yue, V. S. Chandan, D. K. Chang, S. J. Chanock, L. A. Chantrill, A. Chateigner, N. Chatterjee, K. Chayama, H. W. Chen, J. Chen, K. Chen, Y. Chen, Z. Chen, A. D. Cherniack, J. Chien, Y. E. Chiew, S. F. Chin, J. Cho, S. Cho, J. K. Choi, W. Choi, C. Chomienne, Z. Chong, S. P. Choo, A. Chou, A. N. Christ, E. L. Christie, E. Chuah, C. Cibulskis, K. Cibulskis, S. Cingarlini, P. Clapham, A. Claviez, S. Cleary, N. Cloonan, M. Cmero, C. C. Collins, A. A. Connor, S. L. Cooke, C. S. Cooper, L. Cope, V. Corbo, M. G. Cordes, S. M. Cordner, I. s Ciriano, K. Covington, P. A. Cowin, B. Craft, D. Craft, C. J. Creighton, Y. Cun, E. Curley, I. Cutcutache, K. Czajka, B. Czerniak, R. A. Dagg, L. Danilova, M. V. Davi, N. R. Davidson, H. Davies, I. J. Davis, B. N. Davis-Dusenbery, K. J. Dawson, F. M. De La Vega, R. De Paoli-Iseppi, T. Defreitas, A. P. D. Tos, O. Delaneau, J. A. Demchok, J. Demeulemeester, G. M. Demidov, D. lu, N. M. Dennis, R. E. Denroche, S. C. Dentro, N. Desai, V. Deshpande, A. G. Deshwar, C. Desmedt, J. Deu-Pons, N. Dhalla, N. C. Dhani, P. Dhingra, R. Dhir, A. DiBiase, K. Diamanti, L. Ding, S. Ding, H. Q. Dinh, L. Dirix, H. Doddapaneni, N. Donmez, M. T. Dow, R. Drapkin, O. Drechsel, R. M.

Drews, S. Serge, T. Dudderidge, A. Dueso-Barroso, A. J. Dunford, M. Dunn, L. J. Dursi, F. R. Duthie, K. Dutton-Regester, J. Eagles, D. F. Easton, S. Edmonds, P. A. Edwards, S. E. Edwards, R. A. Eeles, A. Ehinger, J. Eils, R. Eils, A. El-Naggar, M. Eldridge, K. Ellrott, S. Erkek, G. Escaramis, S. M. G. Espiritu, X. Estivill, D. Etemadmoghadam, J. E. Eyfjord, B. M. Faltas, D. Fan, Y. Fan, W. C. Faquin, C. Farcas, M. Fassan, A. Fatima, F. Favero, N. Fayzullaev, I. Felau, S. Fereday, M. L. Ferguson, V. Ferretti, L. Feuerbach, M. A. Field, J. L. Fink, G. Finocchiaro, C. Fisher, M. W. Fittall, A. Fitzgerald, R. C. Fitzgerald, A. M. Flanagan, N. E. Fleshner, P. Flicek, J. A. Foekens, K. M. Fong, N. A. Fonseca, C. S. Foster, N. S. Fox, M. Fraser, S. Frazer, M. Frenkel-Morgenstern, W. Friedman, J. Frigola, C. C. Fronick, A. Fujimoto, M. Fujita, M. Fukayama, L. A. Fulton, R. S. Fulton, M. Furuta, P. A. Futreal, A. llgrabe, S. B. Gabriel, S. Gallinger, C. Gambacorti-Passerini, J. Gao, S. Gao, L. Garraway, Ã. Garred, E. Garrison, D. W. Garsed, N. Gehlenborg, J. L. L. Gelpi, J. George, D. S. Gerhard, C. Gerhauser, J. E. Gershenwald, M. Gerstein, M. Gerstung, G. Getz, M. Ghori, R. Ghossein, N. H. Giama, R. A. Gibbs, B. Gibson, A. J. Gill, P. Gill, D. D. Giri, D. Glodzik, V. J. Gnanapragasam, M. E. Goebler, M. J. Goldman, C. Gomez, S. Gonzalez, A. Gonzalez-Perez, D. A. Gordenin, J. Gossage, K. Gotoh, R. Govindan, D. Grabau, J. S. Graham, R. C. Grant, A. R. Green, E. Green, L. Greger, N. Grehan, S. Grimaldi, S. M. Grimmond, R. L. Grossman, A. Grundhoff, G. Gundem, Q. Guo, M. Gupta, S. Gupta, I. G. Gut, M. Gut, J. ke, G. Ha, A. Haake, D. Haan, S. Haas, K. Haase, J. E. Haber, N. Habermann, F. Hach, S. Haider, N. Hama, F. C. Hamdy, A. Hamilton, M. P. Hamilton, L. Han, G. B. Hanna, M. Hansmann, N. J. Haradhvala, O. Harismendy, I. Harliwong, A. O. Harmanci, E. Harrington, T. Hasegawa, D. Haussler, S. Hawkins, S. Hayami, S. Hayashi, D. N. Hayes, S. J. Hayes, N. K. Hayward, S. Hazell, Y. He, A. P. Heath, S. C. Heath, D. Hedley, A. M. Hegde, D. I. Heiman, M. C. Heinold, Z. Heins, L. E. Heisler, E. Hellstrom-Lindberg, M. Helmy, S. G. Heo, A. J. Hepperla, J. M. Heredia-Genestar, C. Herrmann, P. Hersey, J. M. Hess, H. Hilmarsdottir, J. Hinton, S. Hirano, N. Hiraoka, K. A. Hoadley, A. Hobolth, E. Hodzic, J. I. Hoell, S. Hoffmann, O. Hofmann, A. Holbrook, A. Z. Holik, M. A. Hollingsworth, O. Holmes, R. A. Holt, C. Hong, E. P. Hong, J. H. Hong, G. K. Hooijer, H. j, F. Hosoda, Y. Hou, V. Hovestadt, W. Howat, A. P. Hoyle, R. H. Hruban, J. Hu, T. Hu, X. Hua, K. L. Huang, M. Huang, M. N. Huang, V. Huang, Y. Huang, W. Huber, T. J. Hudson, M. Hummel, J. A. Hung, D. Huntsman, T. R. Hupp, J. Huse, M. R. Huska, B. Hutter, C. M. Hutter, D. bschmann, C. A. Iacobuzio-Donahue, C. D. Imbusch, M. Imielinski, S. Imoto, W. B. Isaacs, K. Isaev, S. Ishikawa, M. Iskar, S. M. A. Islam, M. Ittmann, S. Ivkovic, J. M. G. Izarzugaza, J. Jacquemier, V. Jakrot, N. B. Jamieson, G. H. Jang,

S. J. Jang, J. C. Jayaseelan, R. Jayasinghe, S. R. Jefferys, K. Jegalian, J. L. Jennings, S. H. Jeon, L. Jerman, Y. Ji, W. Jiao, P. A. Johansson, A. L. Johns, J. Johns, R. Johnson, T. A. Johnson, C. Jolly, Y. Joly, J. G. Jonasson, C. D. Jones, D. R. Jones, D. T. W. Jones, N. Jones, S. J. M. Jones, J. Jonkers, Y. S. Ju, H. Juhl, J. Jung, M. Juul, R. I. Juul, S. Juul, N. ger, R. Kabbe, A. Kahles, A. Kahraman, V. B. Kaiser, H. Kakavand, S. Kalimuthu, C. von Kalle, K. J. Kang, K. Karaszi, B. Karlan, R. c, D. Karsch, K. Kasaian, K. S. Kassahn, H. Katai, M. Kato, H. Katoh, Y. Kawakami, J. D. Kay, S. H. Kazakoff, M. D. Kazanov, M. Keays, E. Kebebew, R. F. Kefford, M. Kellis, J. G. Kench, C. J. Kennedy, J. N. A. Kerssemakers, D. Khoo, V. Khoo, N. Khuntikeo, E. Khurana, H. Kilpinen, H. K. Kim, H. L. Kim, H. Y. Kim, H. Kim, J. Kim, J. Kim, J. K. Kim, Y. Kim, T. A. King, W. Klapper, K. Kleinheinz, L. J. Klimczak, S. Knappskog, M. Kneba, B. M. Knoppers, Y. Koh, J. Komorowski, D. Komura, M. Komura, G. Kong, M. Kool, J. O. Korbel, V. Korchina, A. Korshunov, M. Koscher, R. Koster, Z. Kote-Jarai, A. Koures, M. Kovacevic, B. Kremeyer, H. Kretzmer, M. Kreuz, S. Krishnamurthy, D. Kube, K. Kumar, P. Kumar, S. Kumar, Y. Kumar, R. Kundra, K. bler, R. ppers, J. Lagergren, P. H. Lai, P. W. Laird, S. R. Lakhani, C. M. Lalansingh, E. Lalonde, F. C. Lamaze, A. Lambert, E. Lander, P. Landgraf, L. Landoni, A. d, A. s, D. Larsimont, E. Larsson, M. Lathrop, L. M. S. Lau, C. Lawerenz, R. T. Lawlor, M. S. Lawrence, A. J. Lazar, A. M. Lazic, X. Le, D. Lee, D. Lee, E. A. Lee, H. J. Lee, J. J. Lee, J. Y. Lee, J. Lee, M. T. M. Lee, H. Lee-Six, K. V. Lehmann, H. Lehrach, D. Lenze, C. R. Leonard, D. A. Leongamornlert, I. Leshchiner, L. Letourneau, I. Letunic, D. A. Levine, L. Lewis, T. Ley, C. Li, C. H. Li, H. I. Li, J. Li, L. Li, S. Li, S. Li, X. Li, X. Li, X. Li, Y. Li, H. Liang, S. B. Liang, P. Lichter, P. Lin, Z. Lin, W. M. Linehan, O. C. rde, D. Liu, E. M. Liu, F. F. Liu, F. Liu, J. Liu, X. Liu, J. Livingstone, D. Livitz, N. Livni, L. Lochovsky, M. Loeffler, G. V. Long, A. Lopez-Guillermo, S. Lou, D. N. Louis, L. B. Lovat, Y. Lu, Y. J. Lu, Y. Lu, C. Luchini, I. Lungu, X. Luo, H. J. Luxton, A. G. Lynch, L. Lype, C. pez, C. n, E. Z. Ma, Y. Ma, G. MacGrogan, S. MacRae, G. Macintyre, T. Madsen, K. Maejima, A. Mafficini, D. T. Maglinte, A. Maitra, P. P. Majumder, L. Malcovati, S. Malikic, G. Malleo, G. J. Mann, L. ffler, K. Marchal, G. Marchegiani, E. R. Mardis, A. A. Margolin, M. G. Marin, F. Markowetz, J. Markowski, J. Marks, T. Marques-Bonet, M. A. Marra, L. Marsden, J. W. M. Martens, S. Martin, J. I. Martin-Subero, I. Martincorena, A. Martinez-Fundichely, Y. E. Maruvka, R. J. Mashl, C. E. Massie, T. J. Matthew, L. Matthews, E. Mayer, S. Mayes, M. Mayo, F. Mbabaali, K. McCune, U. McDermott, P. D. McGillivray, M. D. McLellan, J. D. McPherson, J. R. McPherson, T. A. McPherson, S. R. Meier, A. Meng, S. Meng, A. Menzies, N. D. Merrett, S. Merson, M. Meyerson, W. Meyerson, P. A. Mieczkowski, G. L.

Mihaiescu, S. Mijalkovic, T. Mikkelsen, M. Milella, L. Mileshkin, C. A. Miller, D. K. Miller, J. K. Miller, G. B. Mills, A. Milovanovic, S. Minner, M. Miotto, G. M. Arnau, L. Mirabello, C. Mitchell, T. J. Mitchell, S. Miyano, N. Miyoshi, S. Mizuno, F. bor, M. J. Moore, R. A. Moore, S. Morganella, Q. D. Morris, C. Morrison, L. E. Mose, C. D. Moser, F. os, L. Mularoni, A. J. Mungall, K. Mungall, E. A. Musgrove, V. Mustonen, D. Mutch, F. Muyas, D. M. Muzny, A. oz, J. Myers, O. Myklebost, P. ller, G. Nagae, A. M. Nagrial, H. K. Nahal-Bose, H. Nakagama, H. Nakagawa, H. Nakamura, T. Nakamura, K. Nakano, T. Nandi, J. Nangalia, M. Nastic, A. Navarro, F. C. P. Navarro, D. E. Neal, G. Nettekoven, F. Newell, S. J. Newhouse, Y. Newton, A. W. T. Ng, A. Ng, J. Nicholson, D. Nicol, Y. Nie, G. P. Nielsen, M. M. Nielsen, S. Nik-Zainal, M. S. Noble, K. Nones, P. A. Northcott, F. Notta, B. D. O'Connor, P. O'Donnell, M. O'Donovan, S. O'Meara, B. P. O'Neill, J. R. O'Neill, D. Ocana, A. Ochoa, L. Oesper, C. Ogden, H. Ohdan, K. Ohi, L. Ohno-Machado, K. A. Oien, A. I. Ojesina, H. Ojima, T. Okusaka, L. Omberg, C. K. Ong, S. Ossowski, G. Ott, B. F. F. Ouellette, C. P'ng, M. Paczkowska, S. Paiella, C. Pairojkul, M. Pajic, Q. m, E. Papaemmanuil, I. Papatheodorou, N. Paramasivam, J. W. Park, J. W. Park, K. Park, K. Park, P. J. Park, J. S. Parker, S. L. Parsons, H. Pass, D. Pasternack, A. Pastore, A. M. Patch, I. Ã©, A. Pea, J. V. Pearson, C. S. Pedamallu, J. S. Pedersen, P. Pederzoli, M. Peifer, N. A. Pennell, C. M. Perou, M. D. Perry, G. M. Petersen, M. Peto, N. Petrelli, R. Petryszak, S. M. Pfister, M. Phillips, O. Pich, H. A. Pickett, T. D. Pihl, N. Pillay, S. Pinder, M. Pinese, A. V. Pinho, E. nen, X. Pivot, E. ez, L. Planko, C. Plass, P. Polak, T. Pons, I. Popescu, O. Potapova, A. Prasad, S. R. Preston, M. Prinz, A. L. Pritchard, S. D. Prokopec, E. Provenzano, X. S. Puente, S. Puig, M. s, S. Pulido-Tamayo, G. M. Pupo, C. A. Purdie, M. C. Quinn, R. Rabionet, J. S. Rader, B. Radlwimmer, P. Radovic, B. Raeder, K. M. Raine, M. Ramakrishna, K. Ramakrishnan, S. Ramalingam, B. J. Raphael, W. K. Rathmell, T. Rausch, G. Reifenberger, J. Reimand, J. Reis-Filho, V. Reuter, I. Reyes-Salazar, M. A. Reyna, S. M. Reynolds, E. Rheinbay, Y. Riazalhosseini, A. L. Richardson, J. Richter, M. Ringel, M. r, Y. Rino, K. Rippe, J. Roach, L. R. Roberts, N. D. Roberts, S. A. Roberts, A. G. Robertson, A. J. Robertson, J. B. Rodriguez, B. Rodriguez-Martin, F. G. lez, M. H. A. Roehrl, M. Rohde, H. Rokutan, G. Romieu, I. Rooman, T. Roques, D. Rosebrock, M. Rosenberg, P. C. Rosenstiel, A. Rosenwald, E. W. Rowe, R. Royo, S. G. Rozen, Y. Rubanova, M. A. Rubin, C. Rubio-Perez, V. A. Rudneva, B. C. Rusev, A. Ruzzenente, G. tsch, R. Sabarinathan, V. Y. Sabelnykova, S. Sadeghi, S. C. Sahinalp, N. Saini, M. Saito-Adachi, G. Saksena, A. Salcedo, R. Salgado, L. Salichos, R. Sallari, C. Saller, R. Salvia, M. Sam, J. S. Samra, F. Sanchez-Vega, C. Sander, G. Sanders, R. Sarin, I. Sarrafi, A. Sasaki-Oku,

T. Sauer, G. Sauter, R. P. M. Saw, M. Scardoni, C. J. Scarlett, A. Scarpa, G. Scelo, D. Schadendorf, J. E. Schein, M. B. Schilhabel, M. Schlesner, T. Schlomm, H. K. Schmidt, S. J. Schramm, S. Schreiber, N. Schultz, S. E. Schumacher, R. F. Schwarz, R. A. Scolyer, D. Scott, R. Scully, R. Seethala, A. V. Segre, I. Selander, C. A. Semple, Y. Senbabaoglu, S. Sengupta, E. Sereni, S. Serra, D. C. Sgroi, M. Shackleton, N. C. Shah, S. Shahabi, C. A. Shang, P. Shang, O. Shapira, T. Shelton, C. Shen, H. Shen, R. Shepherd, R. Shi, Y. Shi, Y. J. Shiah, T. Shibata, J. Shih, E. Shimizu, K. Shimizu, S. J. Shin, Y. Shiraishi, T. Shmaya, I. Shmulevich, S. I. Shorser, C. Short, R. Shrestha, S. S. Shringarpure, C. Shriver, S. Shuai, N. Sidiropoulos, R. Siebert, A. M. Sieuwerts, L. Sieverling, S. Signoretti, K. O. Sikora, M. Simbolo, R. Simon, J. V. Simons, J. T. Simpson, P. T. Simpson, S. Singer, N. Sinnott-Armstrong, P. Sipahimalani, T. J. Skelly, M. Smid, J. Smith, K. Smith-McCune, N. D. Socci, H. J. Sofia, M. G. Soloway, L. Song, A. K. Sood, S. Sothi, C. Sotiriou, C. M. Soulette, P. N. Span, P. T. Spellman, N. Sperandio, A. J. Spillane, O. Spiro, J. Spring, J. Staaf, P. F. Stadler, P. Staib, S. G. Stark, L. Stebbings, A. nsson, O. Stegle, L. D. Stein, A. Stenhouse, C. Stewart, S. Stilgenbauer, M. D. Stobbe, M. R. Stratton, J. R. Stretch, A. J. Struck, J. M. Stuart, H. G. Stunnenberg, H. Su, X. Su, R. X. Sun, S. Sungalee, H. Susak, A. Suzuki, F. Sweep, M. Szczepanowski, H. ltmann, T. Yugawa, A. Tam, D. Tamborero, B. K. T. Tan, D. Tan, P. Tan, H. Tanaka, H. Taniguchi, T. J. Tanskanen, M. Tarabichi, R. Tarnuzzer, P. Tarpey, M. L. Taschuk, K. Tatsuno, S. Ã©, D. F. Taylor, A. Taylor-Weiner, J. W. Teague, B. T. Teh, V. Tembe, J. Temes, K. Thai, S. P. Thayer, N. Thiessen, G. Thomas, S. Thomas, A. Thompson, A. M. Thompson, J. F. F. Thompson, R. H. Thompson, H. Thorne, L. B. Thorne, A. Thorogood, G. Tiao, N. Tijanic, L. E. Timms, R. Tirabosco, M. Tojo, S. Tommasi, C. W. Toon, U. H. Toprak, D. Torrents, G. Tortora, J. Tost, Y. Totoki, D. Townend, N. Traficante, I. Treilleux, J. R. Trotta, L. H. P. mper, M. Tsao, T. Tsunoda, J. M. C. Tubio, O. Tucker, R. Turkington, D. J. Turner, A. Tutt, M. Ueno, N. T. Ueno, C. Umbricht, H. M. Umer, T. J. Underwood, L. Urban, T. Urushidate, T. Ushiku, L. la Reimand, A. Valencia, D. J. Van Den Berg, S. Van Laere, P. Van Loo, E. G. Van Meir, G. G. Van den Eynden, T. Van der Kwast, N. Vasudev, M. Vazquez, R. Vedururu, U. Veluvolu, S. Vembu, L. P. C. Verbeke, P. Vermeulen, C. Verrill, A. Viari, D. Vicente, C. Vicentini, K. VijayRaghavan, J. Viksna, R. E. Vilain, I. Villasante, A. Vincent-Salomon, T. Visakorpi, D. Voet, P. Vyas, I. a, N. M. Waddell, N. Waddell, C. Wadelius, L. Wadi, R. Wagener, J. A. Wala, J. Wang, J. Wang, L. Wang, Q. Wang, W. Wang, Y. Wang, Z. Wang, P. M. Waring, H. J. Warnatz, J. Warrell, A. Y. Warren, S. M. Waszak, D. C. Wedge, D. Weichenhan, P. Weinberger, J. N. Weinstein, J. Weischenfeldt, D. J. Weisenberger, I. Welch, M. C. Wendl, J. Werner, J. P.

Whalley, D. A. Wheeler, H. C. Whitaker, D. Wigle, M. D. Wilkerson, A. Williams, J. S. Wilmott, G. W. Wilson, J. M. Wilson, R. K. Wilson, B. Winterhoff, J. A. Wintersinger, M. Wiznerowicz, S. Wolf, B. H. Wong, T. Wong, W. Wong, Y. Woo, S. Wood, B. G. Wouters, A. J. Wright, D. W. Wright, M. H. Wright, C. L. Wu, D. Y. Wu, G. Wu, J. Wu, K. Wu, Y. Wu, Z. Wu, L. Xi, T. Xia, Q. Xiang, X. Xiao, R. Xing, H. Xiong, Q. Xu, Y. Xu, H. Xue, S. Yachida, S. Yakneen, R. Yamaguchi, T. N. Yamaguchi, M. Yamamoto, S. Yamamoto, H. Yamaue, F. Yang, H. Yang, J. Y. Yang, L. Yang, L. Yang, S. Yang, T. P. Yang, Y. Yang, X. Yao, M. L. Yaspo, L. Yates, C. Yau, C. Ye, K. Ye, V. D. Yellapantula, C. J. Yoon, S. S. Yoon, F. Yousif, J. Yu, K. Yu, W. Yu, Y. Yu, K. Yuan, Y. Yuan, D. Yuen, C. K. Yung, O. Zaikova, J. Zamora, M. Zapatka, J. C. Zenklusen, T. Zenz, N. Zeps, C. Z. Zhang, F. Zhang, H. Zhang, H. Zhang, H. Zhang, J. Zhang, J. Zhang, J. Zhang, X. Zhang, X. Zhang, Y. Zhang, Z. Zhang, Z. Zhao, L. Zheng, X. Zheng, W. Zhou, Y. Zhou, B. Zhu, H. Zhu, J. Zhu, S. Zhu, L. Zou, X. Zou, A. deFazio, N. van As, C. H. M. van Deurzen, M. J. van de Vijver, L. Van't Veer, and C. von Mering, "The evolutionary history of 2,658 cancers," *Nature*, vol. 578, pp. 122–128, Feb 2020.

[106] J. E. Kucab, X. Zou, S. Morganella, M. Joel, A. S. Nanda, E. Nagy, C. Gomez, A. Degasperi, R. Harris, S. P. Jackson, V. M. Arlt, D. H. Phillips, and S. Nik-Zainal, "A Compendium of Mutational Signatures of Environmental Agents," *Cell*, vol. 177, pp. 821–836, May 2019.

[107] R. A. Rosales, R. D. Drummond, R. Valieris, E. Dias-Neto, and I. T. da Silva, "signeR: an empirical Bayesian approach to mutational signature discovery," *Bioinformatics*, vol. 33, pp. 8–16, Jan 2017.

[108] G. L. Stein-O'Brien, R. Arora, A. C. Culhane, A. V. Favorov, L. X. Garmire, C. S. Greene, L. A. Goff, Y. Li, A. Ngom, M. F. Ochs, Y. Xu, and E. J. Fertig, "Enter the Matrix: Factorization Uncovers Knowledge from Omics," *Trends Genet*, vol. 34, pp. 790–805, Oct 2018.

[109] S. M. E. Sahraeian, R. Liu, B. Lau, K. Podesta, M. Mohiyuddin, and H. Y. K. Lam, "Deep convolutional neural networks for accurate somatic mutation detection," *Nat Commun*, vol. 10, p. 1041, Mar 2019.

[110] V. Franc, V. Hlaváč, and M. Navara, "Sequential coordinate-wise algorithm for the non-negative least squares problem," in *Computer Analysis of Images and Patterns: 11th International Conference, CAIP 2005, Versailles, France, September 5-8, 2005. Proceedings 11*, pp. 407–414, Springer, 2005.

[111] X. Lin and P. C. Boutros, "Optimization and expansion of non-negative matrix factorization," *BMC Bioinformatics*, vol. 21, p. 7, Jan 2020.

[112] A. Degasperi, T. D. Amarante, J. Czarnecki, S. Shooter, X. Zou, D. Glodzik, S. Morganella, A. S. Nanda, C. Badja, G. Koh, S. E. Momen, I. Georgakopoulos-Soares, J. M. L. Dias, J. Young, Y. Memari, H. Davies, and S. Nik-Zainal, "A practical framework and online tool for mutational signature analyses show inter-tissue variation and driver dependencies," *Nat Cancer*, vol. 1, pp. 249–263, Feb 2020.

[113] L. B. Alexandrov, J. Kim, N. J. Haradhvala, M. N. Huang, A. W. Tian Ng, Y. Wu, A. Boot, K. R. Covington, D. A. Gordenin, E. N. Bergstrom, S. M. A. Islam, N. Lopez-Bigas, L. J. Klimczak, J. R. McPherson, S. Morganella, R. Sabarinathan, D. A. Wheeler, V. Mustonen, G. Getz, S. G. Rozen, M. R. Stratton, L. B. Alexandrov, E. N. Bergstrom, A. Boot, P. Boutros, K. Chan, K. R. Covington, A. Fujimoto, G. Getz, D. A. Gordenin, N. J. Haradhvala, M. N. Huang, S. M. A. Islam, M. Kazanov, J. Kim, L. J. Klimczak, N. Lopez-Bigas, M. Lawrence, I. Martincorena, J. R. McPherson, S. Morganella, V. Mustonen, H. Nakagawa, A. W. Tian Ng, P. Polak, S. Prokopec, S. A. Roberts, S. G. Rozen, R. Sabarinathan, N. Saini, T. Shibata, Y. Shiraishi, M. R. Stratton, B. T. Teh, I. a, D. A. Wheeler, Y. Wu, F. Yousif, W. Yu, L. A. Aaltonen, F. Abascal, A. Abeshouse, H. Aburatani, D. J. Adams, N. Agrawal, K. S. Ahn, S. M. Ahn, H. Aikata, R. Akbani, K. C. Akdemir, H. Al-Ahmadie, S. T. Al-Sedairy, F. Al-Shahrour, M. Alawi, M. Albert, K. Aldape, L. B. Alexandrov, A. Ally, K. Alsop, E. G. Alvarez, F. Amary, S. B. Amin, B. Aminou, O. Ammerpohl, M. J. Anderson, Y. Ang, D. Antonello, P. Anur, S. Aparicio, E. L. Appelbaum, Y. Arai, A. Aretz, K. Arihiro, S. I. Ariizumi, J. Armenia, L. Arnould, S. Asa, Y. Assenov, G. Atwal, S. Aukema, J. T. Auman, M. R. R. Aure, P. Awadalla, M. Aymerich, G. D. Bader, A. Baez-Ortega, M. H. Bailey, P. J. Bailey, M. Balasundaram, S. Balu, P. Bandopadhayay, R. E. Banks, S. Barbi, A. P. Barbour, J. Barenboim, J. Barnholtz-Sloan, H. Barr, E. Barrera, J. Bartlett, J. Bartolome, C. Bassi, O. F. Bathe, D. Baumhoer, P. Bavi, S. B. Baylin, W. Bazant, D. Beardsmore, T. A. Beck, S. Behjati, A. Behren, B. Niu, C. Bell, S. Beltran, C. Benz, A. Berchuck, A. K. Bergmann, E. N. Bergstrom, B. P. Berman, D. M. Berney, S. H. Bernhart, R. Beroukhim, M. Berrios, S. Bersani, J. Bertl, M. Betancourt, V. Bhandari, S. G. Bhosle, A. V. Biankin, M. Bieg, D. Bigner, H. Binder, E. Birney, M. Birrer, N. K. Biswas, B. Bjerkehagen, T. Bodenheimer, L. Boice, G. Bonizzato, J. S. De Bono, A. Boot, M. S. Bootwalla, A. Borg, A. Borkhardt, K. A. Boroevich, I. Borozan, C. Borst, M. Bosenberg, M. Bosio, J. Boultwood, G. Bourque, P. C. Boutros, G. S. Bova, D. T. Bowen, R. Bowlby, D. D. L. Bowtell,

S. Boyault, R. Boyce, J. Boyd, A. Brazma, P. Brennan, D. S. Brewer, A. B. Brinkman, R. G. Bristow, R. R. Broaddus, J. E. Brock, M. Brock, A. Broeks, A. N. Brooks, D. Brooks, B. Brors, S. Brunak, T. J. C. Bruxner, A. L. Bruzos, A. Buchanan, I. Buchhalter, C. Buchholz, S. Bullman, H. Burke, B. Burkhardt, K. H. Burns, J. Busanovich, C. D. Bustamante, A. P. Butler, A. J. Butte, N. J. Byrne, A. L. rresen Dale, S. J. Caesar-Johnson, A. Cafferkey, D. Cahill, C. Calabrese, C. Caldas, F. Calvo, N. Camacho, P. J. Campbell, E. Campo, C. u, S. Cao, T. E. Carey, J. Carlevaro-Fita, R. Carlsen, I. Cataldo, M. Cazzola, J. Cebon, R. Cerfolio, D. E. Chadwick, D. Chakravarty, D. Chalmers, C. W. Y. Chan, K. Chan, M. Chan-Seng-Yue, V. S. Chandan, D. K. Chang, S. J. Chanock, L. A. Chantrill, A. Chateigner, N. Chatterjee, K. Chayama, H. W. Chen, J. Chen, K. Chen, Y. Chen, Z. Chen, A. D. Cherniack, J. Chien, Y. E. Chiew, S. F. Chin, J. Cho, S. Cho, J. K. Choi, W. Choi, C. Chomienne, Z. Chong, S. P. Choo, A. Chou, A. N. Christ, E. L. Christie, E. Chuah, C. Cibulskis, K. Cibulskis, S. Cingarlini, P. Clapham, A. Claviez, S. Cleary, N. Cloonan, M. Cmero, C. C. Collins, A. A. Connor, S. L. Cooke, C. S. Cooper, L. Cope, V. Corbo, M. G. Cordes, S. M. Cordner, I. s Ciriano, K. Covington, P. A. Cowin, B. Craft, D. Craft, C. J. Creighton, Y. Cun, E. Curley, I. Cutcutache, K. Czajka, B. Czerniak, R. A. Dagg, L. Danilova, M. V. Davi, N. R. Davidson, H. Davies, I. J. Davis, B. N. Davis-Dusenbery, K. J. Dawson, F. M. De La Vega, R. De Paoli-Iseppi, T. Defreitas, A. P. D. Tos, O. Delaneau, J. A. Demchok, J. Demeulemeester, G. M. Demidov, D. lu, N. M. Dennis, R. E. Denroche, S. C. Dentro, N. Desai, V. Deshpande, A. G. Deshwar, C. Desmedt, J. Deu-Pons, N. Dhalla, N. C. Dhani, P. Dhingra, R. Dhir, A. DiBiase, K. Diamanti, L. Ding, S. Ding, H. Q. Dinh, L. Dirix, H. Doddapaneni, N. Donmez, M. T. Dow, R. Drapkin, O. Drechsel, R. M. Drews, S. Serge, T. Dudderidge, A. Dueso-Barroso, A. J. Dunford, M. Dunn, L. J. Dursi, F. R. Duthie, K. Dutton-Regester, J. Eagles, D. F. Easton, S. Edmonds, P. A. Edwards, S. E. Edwards, R. A. Eeles, A. Ehinger, J. Eils, R. Eils, A. El-Naggar, M. Eldridge, K. Ellrott, S. Erkek, G. Escaramis, S. M. G. Espiritu, X. Estivill, D. Etemadmoghadam, J. E. Eyfjord, B. M. Faltas, D. Fan, Y. Fan, W. C. Faquin, C. Farcas, M. Fassan, A. Fatima, F. Favero, N. Fayzullaev, I. Felau, S. Fereday, M. L. Ferguson, V. Ferretti, L. Feuerbach, M. A. Field, J. L. Fink, G. Finocchiaro, C. Fisher, M. W. Fittall, A. Fitzgerald, R. C. Fitzgerald, A. M. Flanagan, N. E. Fleshner, P. Flicek, J. A. Foekens, K. M. Fong, N. A. Fonseca, C. S. Foster, N. S. Fox, M. Fraser, S. Frazer, M. Frenkel-Morgenstern, W. Friedman, J. Frigola, C. C. Fronick, A. Fujimoto, M. Fujita, M. Fukayama, L. A. Fulton, R. S. Fulton, M. Furuta, P. A. Futreal, A. llgrabe, S. B. Gabriel, S. Gallinger, C. Gambacorti-Passerini, J. Gao, S. Gao, L. Garraway, Ã. Garred, E. Garrison, D. W.

Garsed, N. Gehlenborg, J. L. L. Gelpi, J. George, D. S. Gerhard, C. Gerhauser, J. E. Gershenwald, M. Gerstein, M. Gerstung, G. Getz, M. Ghori, R. Ghossein, N. H. Giama, R. A. Gibbs, B. Gibson, A. J. Gill, P. Gill, D. D. Giri, D. Glodzik, V. J. Gnanapragasam, M. E. Goebler, M. J. Goldman, C. Gomez, S. Gonzalez, A. Gonzalez-Perez, D. A. Gordenin, J. Gossage, K. Gotoh, R. Govindan, D. Grabau, J. S. Graham, R. C. Grant, A. R. Green, E. Green, L. Greger, N. Grehan, S. Grimaldi, S. M. Grimmond, R. L. Grossman, A. Grundhoff, G. Gundem, Q. Guo, M. Gupta, S. Gupta, I. G. Gut, M. Gut, J. ke, G. Ha, A. Haake, D. Haan, S. Haas, K. Haase, J. E. Haber, N. Habermann, F. Hach, S. Haider, N. Hama, F. C. Hamdy, A. Hamilton, M. P. Hamilton, L. Han, G. B. Hanna, M. Hansmann, N. J. Haradhvala, O. Harismendy, I. Harliwong, A. O. Harmanci, E. Harrington, T. Hasegawa, D. Haussler, S. Hawkins, S. Hayami, S. Hayashi, D. N. Hayes, S. J. Hayes, N. K. Hayward, S. Hazell, Y. He, A. P. Heath, S. C. Heath, D. Hedley, A. M. Hegde, D. I. Heiman, M. C. Heinold, Z. Heins, L. E. Heisler, E. Hellstrom-Lindberg, M. Helmy, S. G. Heo, A. J. Hepperla, J. M. Heredia-Genestar, C. Herrmann, P. Hersey, J. M. Hess, H. Hilmarsdottir, J. Hinton, S. Hirano, N. Hiraoka, K. A. Hoadley, A. Hobolth, E. Hodzic, J. I. Hoell, S. Hoffmann, O. Hofmann, A. Holbrook, A. Z. Holik, M. A. Hollingsworth, O. Holmes, R. A. Holt, C. Hong, E. P. Hong, J. H. Hong, G. K. Hooijer, H. j, F. Hosoda, Y. Hou, V. Hovestadt, W. Howat, A. P. Hoyle, R. H. Hruban, J. Hu, T. Hu, X. Hua, K. L. Huang, M. Huang, M. N. Huang, V. Huang, Y. Huang, W. Huber, T. J. Hudson, M. Hummel, J. A. Hung, D. Huntsman, T. R. Hupp, J. Huse, M. R. Huska, B. Hutter, C. M. Hutter, D. bschmann, C. A. Iacobuzio-Donahue, C. D. Imbusch, M. Imielinski, S. Imoto, W. B. Isaacs, K. Isaev, S. Ishikawa, M. Iskar, S. M. A. Islam, M. Ittmann, S. Ivkovic, J. M. G. Izarzugaza, J. Jacquemier, V. Jakrot, N. B. Jamieson, G. H. Jang, S. J. Jang, J. C. Jayaseelan, R. Jayasinghe, S. R. Jefferys, K. Jegalian, J. L. Jennings, S. H. Jeon, L. Jerman, Y. Ji, W. Jiao, P. A. Johansson, A. L. Johns, J. Johns, R. Johnson, T. A. Johnson, C. Jolly, Y. Joly, J. G. Jonasson, C. D. Jones, D. R. Jones, D. T. W. Jones, N. Jones, S. J. M. Jones, J. Jonkers, Y. S. Ju, H. Juhl, J. Jung, M. Juul, R. I. Juul, S. Juul, N. ger, R. Kabbe, A. Kahles, A. Kahraman, V. B. Kaiser, H. Kakavand, S. Kalimuthu, C. von Kalle, K. J. Kang, K. Karaszi, B. Karlan, R. c, D. Karsch, K. Kasaian, K. S. Kassahn, H. Katai, M. Kato, H. Katoh, Y. Kawakami, J. D. Kay, S. H. Kazakoff, M. D. Kazanov, M. Keays, E. Kebebew, R. F. Kefford, M. Kellis, J. G. Kench, C. J. Kennedy, J. N. A. Kerssemakers, D. Khoo, V. Khoo, N. Khuntikeo, E. Khurana, H. Kilpinen, H. K. Kim, H. L. Kim, H. Y. Kim, H. Kim, J. Kim, J. Kim, J. K. Kim, Y. Kim, T. A. King, W. Klapper, K. Kleinheinz, L. J. Klimczak, S. Knappskog, M. Kneba, B. M. Knoppers, Y. Koh, J. Komorowski, D. Komura, M. Komura, G. Kong, M. Kool, J. O. Korbel, V. Korchina, A. Korshunov, M. Koscher, R. Koster, Z. Kote-Jarai, A. Koures, M. Kovacevic, B. Kremeyer, H. Kretzmer, M. Kreuz, S. Krishnamurthy, D. Kube, K. Kumar, P. Kumar, S. Kumar, Y. Kumar, R. Kundra, K. bler, R. ppers, J. Lagergren, P. H. Lai, P. W. Laird, S. R. Lakhani, C. M. Lalansingh, E. Lalonde, F. C. Lamaze, A. Lambert, E. Lander, P. Landgraf, L. Landoni, A. d, A. s, D. Larsimont, E. Larsson, M. Lathrop, L. M. S. Lau, C. Lawerenz, R. T. Lawlor, M. S. Lawrence, A. J. Lazar, A. M. Lazic, X. Le, D. Lee, D. Lee, E. A. Lee, H. J. Lee, J. J. Lee, J. Y. Lee, J. Lee, M. T. M. Lee, H. Lee-Six, K. V. Lehmann, H. Lehrach, D. Lenze, C. R. Leonard, D. A. Leongamornlert, I. Leshchiner, L. Letourneau, I. Letunic, D. A. Levine, L. Lewis, T. Ley, C. Li, C. H. Li, H. I. Li, J. Li, L. Li, S. Li, S. Li, X. Li, X. Li, X. Li, Y. Li, H. Liang, S. B. Liang, P. Lichter, P. Lin, Z. Lin, W. M. Linehan, O. C. rde, D. Liu, E. M. Liu, F. F. Liu, F. Liu, J. Liu, X. Liu, J. Livingstone, D. Livitz, N. Livni, L. Lochovsky, M. Loeffler, G. V. Long, A. Lopez-Guillermo, S. Lou, D. N. Louis, L. B. Lovat, Y. Lu, Y. J. Lu, Y. Lu, C. Luchini, I. Lungu, X. Luo, H. J. Luxton, A. G. Lynch, L. Lype, C. pez, C. n, E. Z. Ma, Y. Ma, G. MacGrogan, S. MacRae, G. Macintyre, T. Madsen, K. Maejima, A. Mafficini, D. T. Maglinte, A. Maitra, P. P. Majumder, L. Malcovati, S. Malikic, G. Malleo, G. J. Mann, L. ffler, K. Marchal, G. Marchegiani, E. R. Mardis, A. A. Margolin, M. G. Marin, F. Markowetz, J. Markowski, J. Marks, T. Marques-Bonet, M. A. Marra, L. Marsden, J. W. M. Martens, S. Martin, J. I. Martin-Subero, I. Martincorena, A. Martinez-Fundichely, Y. E. Maruvka, R. J. Mashl, C. E. Massie, T. J. Matthew, L. Matthews, E. Mayer, S. Mayes, M. Mayo, F. Mbabaali, K. McCune, U. McDermott, P. D. McGillivray, M. D. McLellan, J. D. McPherson, J. R. McPherson, T. A. McPherson, S. R. Meier, A. Meng, S. Meng, A. Menzies, N. D. Merrett, S. Merson, M. Meyerson, W. Meyerson, P. A. Mieczkowski, G. L. Mihaiescu, S. Mijalkovic, T. Mikkelsen, M. Milella, L. Mileshkin, C. A. Miller, D. K. Miller, J. K. Miller, G. B. Mills, A. Milovanovic, S. Minner, M. Miotto, G. M. Arnau, L. Mirabello, C. Mitchell, T. J. Mitchell, S. Miyano, N. Miyoshi, S. Mizuno, F. bor, M. J. Moore, R. A. Moore, S. Morganella, Q. D. Morris, C. Morrison, L. E. Mose, C. D. Moser, F. os, L. Mularoni, A. J. Mungall, K. Mungall, E. A. Musgrove, V. Mustonen, D. Mutch, F. Muyas, D. M. Muzny, A. oz, J. Myers, O. Myklebost, P. ller, G. Nagae, A. M. Nagrial, H. K. Nahal-Bose, H. Nakagama, H. Nakagawa, H. Nakamura, T. Nakamura, K. Nakano, T. Nandi, J. Nangalia, M. Nastic, A. Navarro, F. C. P. Navarro, D. E. Neal, G. Nettekoven, F. Newell, S. J. Newhouse, Y. Newton, A. W. T. Ng, A. Ng, J. Nicholson, D. Nicol, Y. Nie, G. P. Nielsen, M. M. Nielsen, S. Nik-Zainal, M. S. Noble, K. Nones, P. A. Northcott, F. Notta, B. D. O'Connor, P. O'Donnell, M. O'Donovan, S. O'Meara, B. P. O'Neill, J. R. O'Neill, D. Ocana, A. Ochoa, L. Oesper, C. Ogden, H. Ohdan, K. Ohi, L. Ohno-

Machado, K. A. Oien, A. I. Ojesina, H. Ojima, T. Okusaka, L. Omberg, C. K. Ong, S. Ossowski, G. Ott, B. F. F. Ouellette, C. P'ng, M. Paczkowska, S. Paiella, C. Pairojkul, M. Pajic, Q. m, E. Papaemmanuil, I. Papatheodorou, N. Paramasivam, J. W. Park, J. W. Park, K. Park, K. Park, P. J. Park, J. S. Parker, S. L. Parsons, H. Pass, D. Pasternack, A. Pastore, A. M. Patch, I. e, A. Pea, J. V. Pearson, C. S. Pedamallu, J. S. Pedersen, P. Pederzoli, M. Peifer, N. A. Pennell, C. M. Perou, M. D. Perry, G. M. Petersen, M. Peto, N. Petrelli, R. Petryszak, S. M. Pfister, M. Phillips, O. Pich, H. A. Pickett, T. D. Pihl, N. Pillay, S. Pinder, M. Pinese, A. V. Pinho, E. nen, X. Pivot, E. ez, L. Planko, C. Plass, P. Polak, T. Pons, I. Popescu, O. Potapova, A. Prasad, S. R. Preston, M. Prinz, A. L. Pritchard, S. D. Prokopec, E. Provenzano, X. S. Puente, S. Puig, M. s, S. Pulido-Tamayo, G. M. Pupo, C. A. Purdie, M. C. Quinn, R. Rabionet, J. S. Rader, B. Radlwimmer, P. Radovic, B. Raeder, K. M. Raine, M. Ramakrishna, K. Ramakrishnan, S. Ramalingam, B. J. Raphael, W. K. Rathmell, T. Rausch, G. Reifenberger, J. Reimand, J. Reis-Filho, V. Reuter, I. Reyes-Salazar, M. A. Reyna, S. M. Reynolds, E. Rheinbay, Y. Riazalhosseini, A. L. Richardson, J. Richter, M. Ringel, M. r, Y. Rino, K. Rippe, J. Roach, L. R. Roberts, N. D. Roberts, S. A. Roberts, A. G. Robertson, A. J. Robertson, J. B. Rodriguez, B. Rodriguez-Martin, F. G. lez, M. H. A. Roehrl, M. Rohde, H. Rokutan, G. Romieu, I. Rooman, T. Roques, D. Rosebrock, M. Rosenberg, P. C. Rosenstiel, A. Rosenwald, E. W. Rowe, R. Royo, S. G. Rozen, Y. Rubanova, M. A. Rubin, C. Rubio-Perez, V. A. Rudneva, B. C. Rusev, A. Ruzzenente, G. tsch, R. Sabarinathan, V. Y. Sabelnykova, S. Sadeghi, S. C. Sahinalp, N. Saini, M. Saito-Adachi, G. Saksena, A. Salcedo, R. Salgado, L. Salichos, R. Sallari, C. Saller, R. Salvia, M. Sam, J. S. Samra, F. Sanchez-Vega, C. Sander, G. Sanders, R. Sarin, I. Sarrafi, A. Sasaki-Oku, T. Sauer, G. Sauter, R. P. M. Saw, M. Scardoni, C. J. Scarlett, A. Scarpa, G. Scelo, D. Schadendorf, J. E. Schein, M. B. Schilhabel, M. Schlesner, T. Schlomm, H. K. Schmidt, S. J. Schramm, S. Schreiber, N. Schultz, S. E. Schumacher, R. F. Schwarz, R. A. Scolyer, D. Scott, R. Scully, R. Seethala, A. V. Segre, I. Selander, C. A. Semple, Y. Senbabaoglu, S. Sengupta, E. Sereni, S. Serra, D. C. Sgroi, M. Shackleton, N. C. Shah, S. Shahabi, C. A. Shang, P. Shang, O. Shapira, T. Shelton, C. Shen, H. Shen, R. Shepherd, R. Shi, Y. Shi, Y. J. Shiah, T. Shibata, J. Shih, E. Shimizu, K. Shimizu, S. J. Shin, Y. Shiraishi, T. Shmaya, I. Shmulevich, S. I. Shorser, C. Short, R. Shrestha, S. S. Shringarpure, C. Shriver, S. Shuai, N. Sidiropoulos, R. Siebert, A. M. Sieuwerts, L. Sieverling, S. Signoretti, K. O. Sikora, M. Simbolo, R. Simon, J. V. Simons, J. T. Simpson, P. T. Simpson, S. Singer, N. Sinnott-Armstrong, P. Sipahimalani, T. J. Skelly, M. Smid, J. Smith, K. Smith-McCune, N. D. Socci, H. J. Sofia, M. G. Soloway, L. Song, A. K.

Sood, S. Sothi, C. Sotiriou, C. M. Soulette, P. N. Span, P. T. Spellman, N. Sperandio, A. J. Spillane, O. Spiro, J. Spring, J. Staaf, P. F. Stadler, P. Staib, S. G. Stark, L. Stebbings, A. nsson, O. Stegle, L. D. Stein, A. Stenhouse, C. Stewart, S. Stilgenbauer, M. D. Stobbe, M. R. Stratton, J. R. Stretch, A. J. Struck, J. M. Stuart, H. G. Stunnenberg, H. Su, X. Su, R. X. Sun, S. Sungalee, H. Susak, A. Suzuki, F. Sweep, M. Szczepanowski, H. ltmann, T. Yugawa, A. Tam, D. Tamborero, B. K. T. Tan, D. Tan, P. Tan, H. Tanaka, H. Taniguchi, T. J. Tanskanen, M. Tarabichi, R. Tarnuzzer, P. Tarpey, M. L. Taschuk, K. Tatsuno, S. Ã©, D. F. Taylor, A. Taylor-Weiner, J. W. Teague, B. T. Teh, V. Tembe, J. Temes, K. Thai, S. P. Thayer, N. Thiessen, G. Thomas, S. Thomas, A. Thompson, A. M. Thompson, J. F. F. Thompson, R. H. Thompson, H. Thorne, L. B. Thorne, A. Thorogood, G. Tiao, N. Tijanic, L. E. Timms, R. Tirabosco, M. Tojo, S. Tommasi, C. W. Toon, U. H. Toprak, D. Torrents, G. Tortora, J. Tost, Y. Totoki, D. Townend, N. Traficante, I. Treilleux, J. R. Trotta, L. H. P. mper, M. Tsao, T. Tsunoda, J. M. C. Tubio, O. Tucker, R. Turkington, D. J. Turner, A. Tutt, M. Ueno, N. T. Ueno, C. Umbricht, H. M. Umer, T. J. Underwood, L. Urban, T. Urushidate, T. Ushiku, L. la Reimand, A. Valencia, D. J. Van Den Berg, S. Van Laere, P. Van Loo, E. G. Van Meir, G. G. Van den Eynden, T. Van der Kwast, N. Vasudev, M. Vazquez, R. Vedururu, U. Veluvolu, S. Vembu, L. P. C. Verbeke, P. Vermeulen, C. Verrill, A. Viari, D. Vicente, C. Vicentini, K. VijayRaghavan, J. Viksna, R. E. Vilain, I. Villasante, A. Vincent-Salomon, T. Visakorpi, D. Voet, P. Vyas, I. a, N. M. Waddell, N. Waddell, C. Wadelius, L. Wadi, R. Wagener, J. A. Wala, J. Wang, J. Wang, L. Wang, Q. Wang, W. Wang, Y. Wang, Z. Wang, P. M. Waring, H. J. Warnatz, J. Warrell, A. Y. Warren, S. M. Waszak, D. C. Wedge, D. Weichenhan, P. Weinberger, J. N. Weinstein, J. Weischenfeldt, D. J. Weisenberger, I. Welch, M. C. Wendl, J. Werner, J. P. Whalley, D. A. Wheeler, H. C. Whitaker, D. Wigle, M. D. Wilkerson, A. Williams, J. S. Wilmott, G. W. Wilson, J. M. Wilson, R. K. Wilson, B. Winterhoff, J. A. Wintersinger, M. Wiznerowicz, S. Wolf, B. H. Wong, T. Wong, W. Wong, Y. Woo, S. Wood, B. G. Wouters, A. J. Wright, D. W. Wright, M. H. Wright, C. L. Wu, D. Y. Wu, G. Wu, J. Wu, K. Wu, Y. Wu, Z. Wu, L. Xi, T. Xia, Q. Xiang, X. Xiao, R. Xing, H. Xiong, Q. Xu, Y. Xu, H. Xue, S. Yachida, S. Yakneen, R. Yamaguchi, T. N. Yamaguchi, M. Yamamoto, S. Yamamoto, H. Yamaue, F. Yang, H. Yang, J. Y. Yang, L. Yang, L. Yang, S. Yang, T. P. Yang, Y. Yang, X. Yao, M. L. Yaspo, L. Yates, C. Yau, C. Ye, K. Ye, V. D. Yellapantula, C. J. Yoon, S. S. Yoon, F. Yousif, J. Yu, K. Yu, W. Yu, Y. Yu, K. Yuan, Y. Yuan, D. Yuen, C. K. Yung, O. Zaikova, J. Zamora, M. Zapatka, J. C. Zenklusen, T. Zenz, N. Zeps, C. Z. Zhang, F. Zhang, H. Zhang, H. Zhang, H. Zhang, J. Zhang, J. Zhang, J. Zhang, X. Zhang, X. Zhang, Y. Zhang, Z. Zhang, Z. Zhao, L. Zheng, X. Zheng, W. Zhou, Y. Zhou, B. Zhu,

H. Zhu, J. Zhu, S. Zhu, L. Zou, X. Zou, A. deFazio, N. van As, C. H. M. van Deurzen, M. J. van de Vijver, L. Van't Veer, and C. von Mering, "The repertoire of mutational signatures in human cancer," *Nature*, vol. 578, pp. 94–101, Feb 2020.

[114] J. Ma, J. Setton, N. Y. Lee, N. Riaz, and S. N. Powell, "The therapeutic significance of mutational signatures from DNA repair deficiency in cancer," *Nat Commun*, vol. 9, p. 3292, Aug 2018.

[115] L. Moore, A. Cagan, T. H. H. Coorens, M. D. C. Neville, R. Sanghvi, M. A. Sanders, T. R. W. Oliver, D. Leongamornlert, P. Ellis, A. Noorani, T. J. Mitchell, T. M. Butler, Y. Hooks, A. Y. Warren, M. Jorgensen, K. J. Dawson, A. Menzies, L. O'Neill, C. Latimer, M. Teng, R. van Boxtel, C. A. Iacobuzio-Donahue, I. Martincorena, R. Heer, P. J. Campbell, R. C. Fitzgerald, M. R. Stratton, and R. Rahbari, "The mutational landscape of human somatic and germline cells," *Nature*, vol. 597, pp. 381–386, Sept. 2021.

[116] B. Meier, N. V. Volkova, Y. Hong, P. Schofield, P. J. Campbell, M. Gerstung, and A. Gartner, "and human cancers," *Genome Res*, vol. 28, pp. 666–675, May 2018.

[117] S. Nik-Zainal, J. E. Kucab, S. Morganella, D. Glodzik, L. B. Alexandrov, V. M. Arlt, A. Weninger, M. Hollstein, M. R. Stratton, and D. H. Phillips, "The genome as a record of environmental exposure," *Mutagenesis*, vol. 30, pp. 763–770, Nov 2015.

[118] J. R. BROWN, J. L. THORNTON, and P. POTT, "Percivall Pott (1714-1788) and chimney sweepers' cancer of the scrotum," *Br J Ind Med*, vol. 14, pp. 68–70, Jan 1957.

[119] P. C. Hanawalt and G. Spivak, "Transcription-coupled DNA repair: two decades of progress and surprises," *Nat Rev Mol Cell Biol*, vol. 9, pp. 958–970, Dec 2008.

[120] S. Jinks-Robertson and A. S. Bhagwat, "Transcription-associated mutagenesis," *Annu Rev Genet*, vol. 48, pp. 341–359, 2014.

[121] J. R. Lobry, "Asymmetric substitution patterns in the two DNA strands of bacteria," *Mol Biol Evol*, vol. 13, pp. 660–665, May 1996.

[122] P. Green, B. Ewing, W. Miller, P. J. Thomas, and E. D. Green, "Transcription-associated mutational asymmetry in mammalian evolution," *Nat Genet*, vol. 33, pp. 514–517, Apr 2003.

[123] N. J. Haradhvala, P. Polak, P. Stojanov, K. R. Covington, E. Shinbrot, J. M. Hess, E. Rheinbay, J. Kim, Y. E. Maruvka, L. Z. Braunstein, A. Kamburov, P. C. Hanawalt,

D. A. Wheeler, A. Koren, M. S. Lawrence, and G. Getz, "Mutational Strand Aymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair," *Cell*, vol. 164, pp. 538–549, Jan 2016.

[124] M. Oman, A. Alam, and R. W. Ness, "How Sequence Context-Dependent Mutability Drives Mutation Rate Variation in the Genome," *Genome Biol Evol*, vol. 14, Mar 2022.

[125] C. Greenman, P. Stephens, R. Smith, G. L. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J. Teague, A. Butler, C. Stevens, S. Edkins, S. O'Meara, I. Vastrik, E. E. Schmidt, T. Avis, S. Barthorpe, G. Bhamra, G. Buck, B. Choudhury, J. Clements, J. Cole, E. Dicks, S. Forbes, K. Gray, K. Halliday, R. Harrison, K. Hills, J. Hinton, A. Jenkinson, D. Jones, A. Menzies, T. Mironenko, J. Perry, K. Raine, D. Richardson, R. Shepherd, A. Small, C. Tofts, J. Varian, T. Webb, S. West, S. Widaa, A. Yates, D. P. Cahill, D. N. Louis, P. Goldstraw, A. G. Nicholson, F. Brasseur, L. Looijenga, B. L. Weber, Y. E. Chiew, A. DeFazio, M. F. Greaves, A. R. Green, P. Campbell, E. Birney, D. F. Easton, G. Chenevix-Trench, M. H. Tan, S. K. Khoo, B. T. Teh, S. T. Yuen, S. Y. Leung, R. Wooster, P. A. Futreal, and M. R. Stratton, "Patterns of somatic mutation in human cancer genomes," *Nature*, vol. 446, pp. 153–158, Mar 2007.

[126] M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts, A. Kiezun, P. S. Hammerman, A. McKenna, Y. Drier, L. Zou, A. H. Ramos, T. J. Pugh, N. Stransky, E. Helman, J. Kim, C. Sougnez, L. Ambrogio, E. Nickerson, E. Shefler, M. L. s, D. Auclair, G. Saksena, D. Voet, M. Noble, D. DiCara, P. Lin, L. Lichtenstein, D. I. Heiman, T. Fennell, M. Imielinski, B. Hernandez, E. Hodis, S. Baca, A. M. Dulak, J. Lohr, D. A. Landau, C. J. Wu, J. Melendez-Zajgla, A. Hidalgo-Miranda, A. Koren, S. A. McCarroll, J. Mora, B. Crompton, R. Onofrio, M. Parkin, W. Winckler, K. Ardlie, S. B. Gabriel, C. W. M. Roberts, J. A. Biegel, K. Stegmaier, A. J. Bass, L. A. Garraway, M. Meyerson, T. R. Golub, D. A. Gordenin, S. Sunyaev, E. S. Lander, and G. Getz, "Mutational heterogeneity in cancer and the search for new cancer-associated genes," *Nature*, vol. 499, pp. 214–218, Jul 2013.

[127] P. F. Arndt, T. Hwa, and D. A. Petrov, "Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density, and telomere-specific effects," *J Mol Evol*, vol. 60, pp. 748–763, Jun 2005.

[128] A. R. Poetsch, S. J. Boulton, and N. M. Luscombe, "Genomic landscape of oxidative DNA damage and repair reveals regioselective protection from mutagenesis," *Genome Biol*, vol. 19, p. 215, Dec 2018.

[129] C. K. Kwok, G. Marsico, A. B. Sahakyan, V. S. Chambers, and S. Balasubramanian, "rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome," *Nat Methods*, vol. 13, pp. 841–844, Oct 2016.

[130] R. Linke, M. Limmer, S. A. Juranek, A. Heine, and K. Paeschke, "The Relevance of G-Quadruplexes for DNA Repair," *Int J Mol Sci*, vol. 22, Nov 2021.

[131] C. A. Lewis, J. Crayle, S. Zhou, R. Swanstrom, and R. Wolfenden, "Cytosine deamination and the precipitous decline of spontaneous mutation during Earth's history," *Proc Natl Acad Sci U S A*, vol. 113, pp. 8194–8199, Jul 2016.

[132] T. Misteli, "Beyond the sequence: cellular organization of genome function," *Cell*, vol. 128, pp. 787–800, Feb 2007.

[133] K. D. Makova and R. C. Hardison, "The effects of chromatin organization on variation in mutation rates in the genome," *Nat Rev Genet*, vol. 16, pp. 213–223, Apr 2015.

[134] S. De and F. Michor, "DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes," *Nat Biotechnol*, vol. 29, pp. 1103–1108, Nov 2011.

[135] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold, "Genome-wide mapping of in vivo protein-DNA interactions," *Science*, vol. 316, pp. 1497–1502, Jun 2007.

[136] J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf, "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position," *Nat Methods*, vol. 10, pp. 1213–1218, Dec 2013.

[137] S. Ma and Y. Zhang, "Profiling chromatin regulatory landscape: insights into the development of ChIP-seq and ATAC-seq," *Mol Biomed*, vol. 1, no. 1, p. 9, 2020.

[138] I. Y. Goryshin and W. S. Reznikoff, "Tn5 in vitro transposition," *J Biol Chem*, vol. 273, pp. 7367–7374, Mar 1998.

[139] B. D. Pope, T. Ryba, V. Dileep, F. Yue, W. Wu, O. Denas, D. L. Vera, Y. Wang, R. S. Hansen, T. K. Canfield, R. E. Thurman, Y. Cheng, G. lsoy, J. H. Dennis, M. P. Snyder,

J. A. Stamatoyannopoulos, J. Taylor, R. C. Hardison, T. Kahveci, B. Ren, and D. M. Gilbert, "Topologically associating domains are stable units of replication-timing regulation," *Nature*, vol. 515, pp. 402–405, Nov 2014.

[140] J. H. TAYLOR, "Asynchronous duplication of chromosomes in cultured cells of Chinese hamster," *J Biophys Biochem Cytol*, vol. 7, pp. 455–464, Jun 1960.

[141] K. C. Akdemir, V. T. Le, J. M. Kim, S. Killcoyne, D. A. King, Y. P. Lin, Y. Tian, A. Inoue, S. B. Amin, F. S. Robinson, M. Nimmakayalu, R. E. Herrera, E. J. Lynn, K. Chan, S. Seth, L. J. Klimczak, M. Gerstung, D. A. Gordenin, J. O'Brien, L. Li, Y. L. Deribe, R. G. Verhaak, P. J. Campbell, R. Fitzgerald, A. J. Morrison, J. R. Dixon, and P. Andrew Futreal, "Somatic mutation distributions in cancer genomes vary with three-dimensional chromatin structure," *Nat Genet*, vol. 52, pp. 1178–1188, Nov 2020.

[142] F. F. Diehl, T. P. Miettinen, R. Elbashir, C. S. Nabel, A. M. Darnell, B. T. Do, S. R. Manalis, C. A. Lewis, and M. G. Vander Heiden, "Nucleotide imbalance decouples cell growth from cell proliferation," *Nat Cell Biol*, vol. 24, pp. 1252–1264, Aug 2022.

[143] M. Tomkova, J. Tomek, S. Kriaucionis, and B. ckler, "Mutational signature distribution varies with DNA replication timing and strand asymmetry," *Genome Biol*, vol. 19, p. 129, Sep 2018.

[144] V. B. Seplyarskiy, R. A. Soldatov, E. Koch, R. J. McGinty, J. M. Goldmann, R. D. Hernandez, K. Barnes, A. Correa, E. G. Burchard, P. T. Ellinor, S. T. McGarvey, B. D. Mitchell, R. S. Vasan, S. Redline, E. Silverman, S. T. Weiss, D. K. Arnett, J. Blangero, E. Boerwinkle, J. He, C. Montgomery, D. C. Rao, J. I. Rotter, K. D. Taylor, J. A. Brody, Y. I. Chen, L. de Las Fuentes, C. M. Hwu, S. S. Rich, A. W. Manichaikul, J. C. Mychaleckyj, N. D. Palmer, J. A. Smith, S. L. R. Kardia, P. A. Peyser, L. F. Bielak, T. D. O'Connor, L. S. Emery, C. Gilissen, W. S. W. Wong, P. V. Kharchenko, and S. Sunyaev, "Population sequencing data reveal a compendium of mutational processes in the human germ line," *Science*, vol. 373, pp. 1030–1035, Aug 2021.

[145] A. A. Maklakov and S. Immler, "The Expensive Germline and the Evolution of Ageing," *Curr Biol*, vol. 26, pp. R577–R586, Jul 2016.

[146] M. Nei, "Selectionism and neutralism in molecular evolution," *Mol Biol Evol*, vol. 22, pp. 2318–2342, Dec 2005.

[147] M. Lynch, "Evolution of the mutation rate," *Trends Genet*, vol. 26, pp. 345–352, Aug 2010.

[148] C. Chen, H. Qi, Y. Shen, J. Pickrell, and M. Przeworski, “Contrasting Determinants of Mutation Rates in Germline and Soma,” *Genetics*, vol. 207, pp. 255–267, Sep 2017.

[149] M. Lynch, “Rate, molecular spectrum, and consequences of human mutation,” *Proc Natl Acad Sci U S A*, vol. 107, pp. 961–968, Jan 2010.

[150] M. Lynch, “Evolution of the mutation rate,” *TRENDS in Genetics*, vol. 26, no. 8, pp. 345–352, 2010.

[151] C. Tomasetti, J. Poling, N. J. Roberts, N. R. London Jr, M. E. Pittman, M. C. Haffner, A. Rizzo, A. Baras, B. Karim, A. Kim, *et al.*, “Cell division rates decrease with age, providing a potential explanation for the age-dependent deceleration in cancer incidence,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 41, pp. 20482–20488, 2019.

[152] L. Zhang, X. Dong, M. Lee, A. Y. Maslov, T. Wang, and J. Vijg, “Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in B lymphocytes across the human lifespan,” *Proc Natl Acad Sci U S A*, vol. 116, pp. 9014–9019, Apr 2019.

[153] R. Anandakrishnan, R. T. Varghese, N. A. Kinney, and H. R. Garner, “Estimating the number of genetic mutations (hits) required for carcinogenesis based on the distribution of somatic mutations,” *PLoS Comput Biol*, vol. 15, p. e1006881, Mar 2019.

[154] D. Weghorn and S. Sunyaev, “Bayesian inference of negative and positive selection in human cancers,” *Nat Genet*, vol. 49, pp. 1785–1788, Dec 2017.

[155] K. Voskarides, “Broadening the spectrum of cancer genes under selection in human populations,” *FASEB Bioadv*, vol. 3, pp. 275–277, Apr 2021.

[156] S. Kryazhimskiy and J. B. Plotkin, “The population genetics of dN/dS,” *PLoS Genet*, vol. 4, p. e1000304, Dec 2008.

[157] X. Shen, S. Song, C. Li, and J. Zhang, “Synonymous mutations in representative yeast genes are mostly strongly non-neutral,” *Nature*, vol. 606, no. 7915, pp. 725–731, 2022.

[158] L. Kruglyak, A. Beyer, J. S. Bloom, J. Grossbach, T. D. Lieberman, C. P. Mancuso, M. S. Rich, G. Sherlock, and C. D. Kaplan, “Insufficient evidence for non-neutrality of synonymous mutations,” *Nature*, vol. 616, no. 7957, pp. E8–E9, 2023.

[159] N. Goldman and Z. Yang, "A codon-based model of nucleotide substitution for protein-coding dna sequences.," *Molecular biology and evolution*, vol. 11, no. 5, pp. 725–736, 1994.

[160] I. Martincorena, J. C. Fowler, A. Wabik, A. R. J. Lawson, F. Abascal, M. W. J. Hall, A. Cagan, K. Murai, K. Mahbubani, M. R. Stratton, R. C. Fitzgerald, P. A. Handford, P. J. Campbell, K. Saeb-Parsy, and P. H. Jones, "Somatic mutant clones colonize the human esophagus with age," *Science*, vol. 362, pp. 911–917, Nov 2018.

[161] H. Lee-Six, S. Olafsson, P. Ellis, R. J. Osborne, M. A. Sanders, L. Moore, N. Georgakopoulos, F. Torrente, A. Noorani, M. Goddard, P. Robinson, T. H. H. Coorens, L. O'Neill, C. Alder, J. Wang, R. C. Fitzgerald, M. Zilbauer, N. Coleman, K. Saeb-Parsy, I. Martincorena, P. J. Campbell, and M. R. Stratton, "The landscape of somatic mutation in normal colorectal epithelial cells," *Nature*, vol. 574, pp. 532–537, Oct 2019.

[162] M. Zhu, T. Lu, Y. Jia, X. Luo, P. Gopal, L. Li, M. Odewole, V. Renteria, A. G. Singal, Y. Jang, K. Ge, S. C. Wang, M. Sorouri, J. R. Parekh, M. P. MacConmara, A. C. Yopp, T. Wang, and H. Zhu, "Somatic Mutations Increase Hepatic Clonal Fitness and Regeneration in Chronic Liver Disease," *Cell*, vol. 177, pp. 608–621, Apr 2019.

[163] K. Yoshida, K. H. C. Gowers, H. Lee-Six, D. P. Chandrasekharan, T. Coorens, E. F. Maughan, K. Beal, A. Menzies, F. R. Millar, E. Anderson, S. E. Clarke, A. Pennycuick, R. M. Thakrar, C. R. Butler, N. Kakiuchi, T. Hirano, R. E. Hynds, M. R. Stratton, I. Martincorena, S. M. Janes, and P. J. Campbell, "Tobacco smoking and somatic mutations in human bronchial epithelium," *Nature*, vol. 578, pp. 266–272, Feb 2020.

[164] Z. Wang, S. Zhu, Y. Jia, Y. Wang, N. Kubota, N. Fujiwara, R. Gordillo, C. Lewis, M. Zhu, T. Sharma, L. Li, Q. Zeng, Y. H. Lin, M. H. Hsieh, P. Gopal, T. Wang, M. Hoare, P. Campbell, Y. Hoshida, and H. Zhu, "Positive selection of somatically mutated clones identifies adaptive pathways in metabolic liver disease," *bioRxiv*, Mar 2023.

[165] T. Boveri, "Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris," *J Cell Sci*, vol. 121 Suppl 1, pp. 1–84, Jan 2008.

[166] I. R. Watson, K. Takahashi, P. A. Futreal, and L. Chin, "Emerging patterns of somatic mutations in cancer," *Nat Rev Genet*, vol. 14, pp. 703–718, Oct 2013.

[167] L. A. Aaltonen, F. Abascal, A. Abeshouse, H. Aburatani, D. J. Adams, N. Agrawal, K. S. Ahn, S. M. Ahn, H. Aikata, R. Akbani, K. C. Akdemir, H. Al-Ahmadie, S. T. Al-Sedairy, F. Al-Shahrour, M. Alawi, M. Albert, K. Aldape, L. B. Alexandrov, A. Ally,

K. Alsop, E. G. Alvarez, F. Amary, S. B. Amin, B. Aminou, O. Ammerpohl, M. J. Anderson, Y. Ang, D. Antonello, P. Anur, S. Aparicio, E. L. Appelbaum, Y. Arai, A. Aretz, K. Arihiro, S. I. Ariizumi, J. Armenia, L. Arnould, S. Asa, Y. Assenov, G. Atwal, S. Aukema, J. T. Auman, M. R. R. Aure, P. Awadalla, M. Aymerich, G. D. Bader, A. Baez-Ortega, M. H. Bailey, P. J. Bailey, M. Balasundaram, S. Balu, P. Bandopadhayay, R. E. Banks, S. Barbi, A. P. Barbour, J. Barenboim, J. Barnholtz-Sloan, H. Barr, E. Barrera, J. Bartlett, J. Bartolome, C. Bassi, O. F. Bathe, D. Baumhoer, P. Bavi, S. B. Baylin, W. Bazant, D. Beardsmore, T. A. Beck, S. Behjati, A. Behren, B. Niu, C. Bell, S. Beltran, C. Benz, A. Berchuck, A. K. Bergmann, E. N. Bergstrom, B. P. Berman, D. M. Berney, S. H. Bernhart, R. Beroukhim, M. Berrios, S. Bersani, J. Bertl, M. Betancourt, V. Bhandari, S. G. Bhosle, A. V. Biankin, M. Bieg, D. Bigner, H. Binder, E. Birney, M. Birrer, N. K. Biswas, B. Bjerkehagen, T. Bodenheimer, L. Boice, G. Bonizzato, J. S. De Bono, A. Boot, M. S. Bootwalla, A. Borg, A. Borkhardt, K. A. Boroevich, I. Borozan, C. Borst, M. Bosenberg, M. Bosio, J. Boultwood, G. Bourque, P. C. Boutros, G. S. Bova, D. T. Bowen, R. Bowlby, D. D. L. Bowtell, S. Boyault, R. Boyce, J. Boyd, A. Brazma, P. Brennan, D. S. Brewer, A. B. Brinkman, R. G. Bristow, R. R. Broaddus, J. E. Brock, M. Brock, A. Broeks, A. N. Brooks, D. Brooks, B. Brors, S. Brunak, T. J. C. Bruxner, A. L. Bruzos, A. Buchanan, I. Buchhalter, C. Buchholz, S. Bullman, H. Burke, B. Burkhardt, K. H. Burns, J. Busanovich, C. D. Bustamante, A. P. Butler, A. J. Butte, N. J. Byrne, A. L. rresen Dale, S. J. Caesar-Johnson, A. Cafferkey, D. Cahill, C. Calabrese, C. Caldas, F. Calvo, N. Camacho, P. J. Campbell, E. Campo, C. u, S. Cao, T. E. Carey, J. Carlevaro-Fita, R. Carlsen, I. Cataldo, M. Cazzola, J. Cebon, R. Cerfolio, D. E. Chadwick, D. Chakravarty, D. Chalmers, C. W. Y. Chan, K. Chan, M. Chan-Seng-Yue, V. S. Chandan, D. K. Chang, S. J. Chanock, L. A. Chantrill, A. Chateigner, N. Chatterjee, K. Chayama, H. W. Chen, J. Chen, K. Chen, Y. Chen, Z. Chen, A. D. Cherniack, J. Chien, Y. E. Chiew, S. F. Chin, J. Cho, S. Cho, J. K. Choi, W. Choi, C. Chomienne, Z. Chong, S. P. Choo, A. Chou, A. N. Christ, E. L. Christie, E. Chuah, C. Cibulskis, K. Cibulskis, S. Cingarlini, P. Clapham, A. Claviez, S. Cleary, N. Cloonan, M. Cmero, C. C. Collins, A. A. Connor, S. L. Cooke, C. S. Cooper, L. Cope, V. Corbo, M. G. Cordes, S. M. Cordner, I. s Ciriano, K. Covington, P. A. Cowin, B. Craft, D. Craft, C. J. Creighton, Y. Cun, E. Curley, I. Cutcutache, K. Czajka, B. Czerniak, R. A. Dagg, L. Danilova, M. V. Davi, N. R. Davidson, H. Davies, I. J. Davis, B. N. Davis-Dusenbery, K. J. Dawson, F. M. De La Vega, R. De Paoli-Iseppi, T. Defreitas, A. P. D. Tos, O. Delaneau, J. A. Demchok, J. Demeulemeester, G. M. Demidov, D. lu, N. M. Dennis, R. E. Denroche, S. C. Dentro, N. Desai, V. Deshpande, A. G. Deshwar, C. Desmedt, J. Deu-Pons, N. Dhalla, N. C.

Dhani, P. Dhingra, R. Dhir, A. DiBiase, K. Diamanti, L. Ding, S. Ding, H. Q. Dinh, L. Dirix, H. Doddapaneni, N. Donmez, M. T. Dow, R. Drapkin, O. Drechsel, R. M. Drews, S. Serge, T. Dudderidge, A. Dueso-Barroso, A. J. Dunford, M. Dunn, L. J. Dursi, F. R. Duthie, K. Dutton-Regester, J. Eagles, D. F. Easton, S. Edmonds, P. A. Edwards, S. E. Edwards, R. A. Eeles, A. Ehinger, J. Eils, R. Eils, A. El-Naggar, M. Eldridge, K. Ellrott, S. Erkek, G. Escaramis, S. M. G. Espiritu, X. Estivill, D. Etemadmoghadam, J. E. Eyfjord, B. M. Faltas, D. Fan, Y. Fan, W. C. Faquin, C. Farcas, M. Fassan, A. Fatima, F. Favero, N. Fayzullaev, I. Felau, S. Fereday, M. L. Ferguson, V. Ferretti, L. Feuerbach, M. A. Field, J. L. Fink, G. Finocchiaro, C. Fisher, M. W. Fittall, A. Fitzgerald, R. C. Fitzgerald, A. M. Flanagan, N. E. Fleshner, P. Flicek, J. A. Foekens, K. M. Fong, N. A. Fonseca, C. S. Foster, N. S. Fox, M. Fraser, S. Frazer, M. Frenkel-Morgenstern, W. Friedman, J. Frigola, C. C. Fronick, A. Fujimoto, M. Fujita, M. Fukayama, L. A. Fulton, R. S. Fulton, M. Furuta, P. A. Futreal, A. llgrabe, S. B. Gabriel, S. Gallinger, C. Gambacorti-Passerini, J. Gao, S. Gao, L. Garraway, Ã. Garred, E. Garrison, D. W. Garsed, N. Gehlenborg, J. L. L. Gelpi, J. George, D. S. Gerhard, C. Gerhauser, J. E. Gershenwald, M. Gerstein, M. Gerstung, G. Getz, M. Ghori, R. Ghossein, N. H. Giama, R. A. Gibbs, B. Gibson, A. J. Gill, P. Gill, D. D. Giri, D. Glodzik, V. J. Gnanapragasam, M. E. Goebler, M. J. Goldman, C. Gomez, S. Gonzalez, A. Gonzalez-Perez, D. A. Gordenin, J. Gossage, K. Gotoh, R. Govindan, D. Grabau, J. S. Graham, R. C. Grant, A. R. Green, E. Green, L. Greger, N. Grehan, S. Grimaldi, S. M. Grimmond, R. L. Grossman, A. Grundhoff, G. Gundem, Q. Guo, M. Gupta, S. Gupta, I. G. Gut, M. Gut, J. ke, G. Ha, A. Haake, D. Haan, S. Haas, K. Haase, J. E. Haber, N. Habermann, F. Hach, S. Haider, N. Hama, F. C. Hamdy, A. Hamilton, M. P. Hamilton, L. Han, G. B. Hanna, M. Hansmann, N. J. Haradhvala, O. Harismendy, I. Harliwong, A. O. Harmanci, E. Harrington, T. Hasegawa, D. Haussler, S. Hawkins, S. Hayami, S. Hayashi, D. N. Hayes, S. J. Hayes, N. K. Hayward, S. Hazell, Y. He, A. P. Heath, S. C. Heath, D. Hedley, A. M. Hegde, D. I. Heiman, M. C. Heinold, Z. Heins, L. E. Heisler, E. Hellstrom-Lindberg, M. Helmy, S. G. Heo, A. J. Hepperla, J. M. Heredia-Genestar, C. Herrmann, P. Hersey, J. M. Hess, H. Hilmarsdottir, J. Hinton, S. Hirano, N. Hiraoka, K. A. Hoadley, A. Hobolth, E. Hodzic, J. I. Hoell, S. Hoffmann, O. Hofmann, A. Holbrook, A. Z. Holik, M. A. Hollingsworth, O. Holmes, R. A. Holt, C. Hong, E. P. Hong, J. H. Hong, G. K. Hooijer, H. j, F. Hosoda, Y. Hou, V. Hovestadt, W. Howat, A. P. Hoyle, R. H. Hruban, J. Hu, T. Hu, X. Hua, K. L. Huang, M. Huang, M. N. Huang, V. Huang, Y. Huang, W. Huber, T. J. Hudson, M. Hummel, J. A. Hung, D. Huntsman, T. R. Hupp, J. Huse, M. R. Huska, B. Hutter, C. M. Hutter, D. bschmann, C. A. Iacobuzio-Donahue, C. D. Imbusch, M. Imielinski,

S. Imoto, W. B. Isaacs, K. Isaev, S. Ishikawa, M. Iskar, S. M. A. Islam, M. Ittmann, S. Ivkovic, J. M. G. Izarzugaza, J. Jacquemier, V. Jakrot, N. B. Jamieson, G. H. Jang, S. J. Jang, J. C. Jayaseelan, R. Jayasinghe, S. R. Jefferys, K. Jegalian, J. L. Jennings, S. H. Jeon, L. Jerman, Y. Ji, W. Jiao, P. A. Johansson, A. L. Johns, J. Johns, R. Johnson, T. A. Johnson, C. Jolly, Y. Joly, J. G. Jonasson, C. D. Jones, D. R. Jones, D. T. W. Jones, N. Jones, S. J. M. Jones, J. Jonkers, Y. S. Ju, H. Juhl, J. Jung, M. Juul, R. I. Juul, S. Juul, N. ger, R. Kabbe, A. Kahles, A. Kahraman, V. B. Kaiser, H. Kakavand, S. Kalimuthu, C. von Kalle, K. J. Kang, K. Karaszi, B. Karlan, R. c, D. Karsch, K. Kasaian, K. S. Kassahn, H. Katai, M. Kato, H. Katoh, Y. Kawakami, J. D. Kay, S. H. Kazakoff, M. D. Kazanov, M. Keays, E. Kebebew, R. F. Kefford, M. Kellis, J. G. Kench, C. J. Kennedy, J. N. A. Kerssemakers, D. Khoo, V. Khoo, N. Khuntikeo, E. Khurana, H. Kilpinen, H. K. Kim, H. L. Kim, H. Y. Kim, H. Kim, J. Kim, J. Kim, J. K. Kim, Y. Kim, T. A. King, W. Klapper, K. Kleinheinz, L. J. Klimczak, S. Knappskog, M. Kneba, B. M. Knoppers, Y. Koh, J. Komorowski, D. Komura, M. Komura, G. Kong, M. Kool, J. O. Korbel, V. Korchina, A. Korshunov, M. Koscher, R. Koster, Z. Kote-Jarai, A. Koures, M. Kovacevic, B. Kremeyer, H. Kretzmer, M. Kreuz, S. Krishnamurthy, D. Kube, K. Kumar, P. Kumar, S. Kumar, Y. Kumar, R. Kundra, K. bler, R. ppers, J. Lagergren, P. H. Lai, P. W. Laird, S. R. Lakhani, C. M. Lalansingh, E. Lalonde, F. C. Lamaze, A. Lambert, E. Lander, P. Landgraf, L. Landoni, A. d, A. s, D. Larsimont, E. Larsson, M. Lathrop, L. M. S. Lau, C. Lawerenz, R. T. Lawlor, M. S. Lawrence, A. J. Lazar, A. M. Lazic, X. Le, D. Lee, D. Lee, E. A. Lee, H. J. Lee, J. J. Lee, J. Y. Lee, J. Lee, M. T. M. Lee, H. Lee-Six, K. V. Lehmann, H. Lehrach, D. Lenze, C. R. Leonard, D. A. Leongamornlert, I. Leshchiner, L. Letourneau, I. Letunic, D. A. Levine, L. Lewis, T. Ley, C. Li, C. H. Li, H. I. Li, J. Li, L. Li, S. Li, S. Li, X. Li, X. Li, X. Li, Y. Li, H. Liang, S. B. Liang, P. Lichter, P. Lin, Z. Lin, W. M. Linehan, O. C. rde, D. Liu, E. M. Liu, F. F. Liu, F. Liu, J. Liu, X. Liu, J. Livingstone, D. Livitz, N. Livni, L. Lochovsky, M. Loeffler, G. V. Long, A. Lopez-Guillermo, S. Lou, D. N. Louis, L. B. Lovat, Y. Lu, Y. J. Lu, Y. Lu, C. Luchini, I. Lungu, X. Luo, H. J. Luxton, A. G. Lynch, L. Lype, C. pez, C. n, E. Z. Ma, Y. Ma, G. MacGrogan, S. MacRae, G. Macintyre, T. Madsen, K. Maejima, A. Mafficini, D. T. Maglinte, A. Maitra, P. P. Majumder, L. Malcovati, S. Malikic, G. Malleo, G. J. Mann, L. ffler, K. Marchal, G. Marchegiani, E. R. Mardis, A. A. Margolin, M. G. Marin, F. Markowetz, J. Markowski, J. Marks, T. Marques-Bonet, M. A. Marra, L. Marsden, J. W. M. Martens, S. Martin, J. I. Martin-Subero, I. Martincorena, A. Martinez-Fundichely, Y. E. Maruvka, R. J. Mashl, C. E. Massie, T. J. Matthew, L. Matthews, E. Mayer, S. Mayes, M. Mayo, F. Mbabaali, K. McCune, U. McDermott, P. D. McGillivray, M. D. McLellan, J. D.

McPherson, J. R. McPherson, T. A. McPherson, S. R. Meier, A. Meng, S. Meng, A. Menzies, N. D. Merrett, S. Merson, M. Meyerson, W. Meyerson, P. A. Mieczkowski, G. L. Mihaiescu, S. Mijalkovic, T. Mikkelsen, M. Milella, L. Mileshkin, C. A. Miller, D. K. Miller, J. K. Miller, G. B. Mills, A. Milovanovic, S. Minner, M. Miotto, G. M. Arnau, L. Mirabello, C. Mitchell, T. J. Mitchell, S. Miyano, N. Miyoshi, S. Mizuno, F. bor, M. J. Moore, R. A. Moore, S. Morganella, Q. D. Morris, C. Morrison, L. E. Mose, C. D. Moser, F. os, L. Mularoni, A. J. Mungall, K. Mungall, E. A. Musgrove, V. Mustonen, D. Mutch, F. Muyas, D. M. Muzny, A. oz, J. Myers, O. Myklebost, P. ller, G. Nagae, A. M. Nagrial, H. K. Nahal-Bose, H. Nakagama, H. Nakagawa, H. Nakamura, T. Nakamura, K. Nakano, T. Nandi, J. Nangalia, M. Nastic, A. Navarro, F. C. P. Navarro, D. E. Neal, G. Nettekoven, F. Newell, S. J. Newhouse, Y. Newton, A. W. T. Ng, A. Ng, J. Nicholson, D. Nicol, Y. Nie, G. P. Nielsen, M. M. Nielsen, S. Nik-Zainal, M. S. Noble, K. Nones, P. A. Northcott, F. Notta, B. D. O'Connor, P. O'Donnell, M. O'Donovan, S. O'Meara, B. P. O'Neill, J. R. O'Neill, D. Ocana, A. Ochoa, L. Oesper, C. Ogden, H. Ohdan, K. Ohi, L. Ohno-Machado, K. A. Oien, A. I. Ojesina, H. Ojima, T. Okusaka, L. Omberg, C. K. Ong, S. Ossowski, G. Ott, B. F. F. Ouellette, C. P'ng, M. Paczkowska, S. Paiella, C. Pairojkul, M. Pajic, Q. m, E. Papaemmanuil, I. Papatheodorou, N. Paramasivam, J. W. Park, J. W. Park, K. Park, K. Park, P. J. Park, J. S. Parker, S. L. Parsons, H. Pass, D. Pasternack, A. Pastore, A. M. Patch, I. Ã©, A. Pea, J. V. Pearson, C. S. Pedamallu, J. S. Pedersen, P. Pederzoli, M. Peifer, N. A. Pennell, C. M. Perou, M. D. Perry, G. M. Petersen, M. Peto, N. Petrelli, R. Petryszak, S. M. Pfister, M. Phillips, O. Pich, H. A. Pickett, T. D. Pihl, N. Pillay, S. Pinder, M. Pinese, A. V. Pinho, E. nen, X. Pivot, E. ez, L. Planko, C. Plass, P. Polak, T. Pons, I. Popescu, O. Potapova, A. Prasad, S. R. Preston, M. Prinz, A. L. Pritchard, S. D. Prokopec, E. Provenzano, X. S. Puente, S. Puig, M. s, S. Pulido-Tamayo, G. M. Pupo, C. A. Purdie, M. C. Quinn, R. Rabionet, J. S. Rader, B. Radlwimmer, P. Radovic, B. Raeder, K. M. Raine, M. Ramakrishna, K. Ramakrishnan, S. Ramalingam, B. J. Raphael, W. K. Rathmell, T. Rausch, G. Reifenberger, J. Reimand, J. Reis-Filho, V. Reuter, I. Reyes-Salazar, M. A. Reyna, S. M. Reynolds, E. Rheinbay, Y. Riazalhosseini, A. L. Richardson, J. Richter, M. Ringel, M. r, Y. Rino, K. Rippe, J. Roach, L. R. Roberts, N. D. Roberts, S. A. Roberts, A. G. Robertson, A. J. Robertson, J. B. Rodriguez, B. Rodriguez-Martin, F. G. lez, M. H. A. Roehrl, M. Rohde, H. Rokutan, G. Romieu, I. Rooman, T. Roques, D. Rosebrock, M. Rosenberg, P. C. Rosenstiel, A. Rosenwald, E. W. Rowe, R. Royo, S. G. Rozen, Y. Rubanova, M. A. Rubin, C. Rubio-Perez, V. A. Rudneva, B. C. Rusev, A. Ruzzenente, G. tsch, R. Sabarinathan, V. Y. Sabelnykova, S. Sadeghi, S. C. Sahinalp, N. Saini, M. Saito-Adachi, G. Saksena, A. Salcedo, R. Salgado, L. Salichos, R. Sallari, C. Saller, R. Salvia, M. Sam, J. S. Samra, F. Sanchez-Vega, C. Sander, G. Sanders, R. Sarin, I. Sarrafi, A. Sasaki-Oku, T. Sauer, G. Sauter, R. P. M. Saw, M. Scardoni, C. J. Scarlett, A. Scarpa, G. Scelo, D. Schadendorf, J. E. Schein, M. B. Schilhabel, M. Schlesner, T. Schlomm, H. K. Schmidt, S. J. Schramm, S. Schreiber, N. Schultz, S. E. Schumacher, R. F. Schwarz, R. A. Scolyer, D. Scott, R. Scully, R. Seethala, A. V. Segre, I. Selander, C. A. Semple, Y. Senbabaoglu, S. Sengupta, E. Sereni, S. Serra, D. C. Sgroi, M. Shackleton, N. C. Shah, S. Shahabi, C. A. Shang, P. Shang, O. Shapira, T. Shelton, C. Shen, H. Shen, R. Shepherd, R. Shi, Y. Shi, Y. J. Shiah, T. Shibata, J. Shih, E. Shimizu, K. Shimizu, S. J. Shin, Y. Shiraishi, T. Shmaya, I. Shmulevich, S. I. Shorser, C. Short, R. Shrestha, S. S. Shringarpure, C. Shriver, S. Shuai, N. Sidiropoulos, R. Siebert, A. M. Sieuwerts, L. Sieverling, S. Signoretti, K. O. Sikora, M. Simbolo, R. Simon, J. V. Simons, J. T. Simpson, P. T. Simpson, S. Singer, N. Sinnott-Armstrong, P. Sipahimalani, T. J. Skelly, M. Smid, J. Smith, K. Smith-McCune, N. D. Socci, H. J. Sofia, M. G. Soloway, L. Song, A. K. Sood, S. Sothi, C. Sotiriou, C. M. Soulette, P. N. Span, P. T. Spellman, N. Sperandio, A. J. Spillane, O. Spiro, J. Spring, J. Staaf, P. F. Stadler, P. Staib, S. G. Stark, L. Stebbings, A. nsson, O. Stegle, L. D. Stein, A. Stenhouse, C. Stewart, S. Stilgenbauer, M. D. Stobbe, M. R. Stratton, J. R. Stretch, A. J. Struck, J. M. Stuart, H. G. Stunnenberg, H. Su, X. Su, R. X. Sun, S. Sungalee, H. Susak, A. Suzuki, F. Sweep, M. Szczepanowski, H. ltmann, T. Yugawa, A. Tam, D. Tamborero, B. K. T. Tan, D. Tan, P. Tan, H. Tanaka, H. Taniguchi, T. J. Tanskanen, M. Tarabichi, R. Tarnuzzer, P. Tarpey, M. L. Taschuk, K. Tatsuno, S. Ã©, D. F. Taylor, A. Taylor-Weiner, J. W. Teague, B. T. Teh, V. Tembe, J. Temes, K. Thai, S. P. Thayer, N. Thiessen, G. Thomas, S. Thomas, A. Thompson, A. M. Thompson, J. F. F. Thompson, R. H. Thompson, H. Thorne, L. B. Thorne, A. Thorogood, G. Tiao, N. Tijanic, L. E. Timms, R. Tirabosco, M. Tojo, S. Tommasi, C. W. Toon, U. H. Toprak, D. Torrents, G. Tortora, J. Tost, Y. Totoki, D. Townend, N. Traficante, I. Treilleux, J. R. Trotta, L. H. P. mper, M. Tsao, T. Tsunoda, J. M. C. Tubio, O. Tucker, R. Turkington, D. J. Turner, A. Tutt, M. Ueno, N. T. Ueno, C. Umbricht, H. M. Umer, T. J. Underwood, L. Urban, T. Urushidate, T. Ushiku, L. la Reimand, A. Valencia, D. J. Van Den Berg, S. Van Laere, P. Van Loo, E. G. Van Meir, G. G. Van den Eynden, T. Van der Kwast, N. Vasudev, M. Vazquez, R. Vedururu, U. Veluvolu, S. Vembu, L. P. C. Verbeke, P. Vermeulen, C. Verrill, A. Viari, D. Vicente, C. Vicentini, K. VijayRaghavan, J. Viksna, R. E. Vilain, I. Villasante, A. Vincent-Salomon, T. Visakorpi, D. Voet, P. Vyas, I. a, N. M. Waddell, N. Waddell, C. Wadelius, L. Wadi, R. Wagener, J. A. Wala, J. Wang, J. Wang, L. Wang, Q. Wang, W. Wang, Y. Wang, Z. Wang, P. M. Waring, H. J. Warnatz, J. War-

rell, A. Y. Warren, S. M. Waszak, D. C. Wedge, D. Weichenhan, P. Weinberger, J. N. Weinstein, J. Weischenfeldt, D. J. Weisenberger, I. Welch, M. C. Wendl, J. Werner, J. P. Whalley, D. A. Wheeler, H. C. Whitaker, D. Wigle, M. D. Wilkerson, A. Williams, J. S. Wilmott, G. W. Wilson, J. M. Wilson, R. K. Wilson, B. Winterhoff, J. A. Wintersinger, M. Wiznerowicz, S. Wolf, B. H. Wong, T. Wong, W. Wong, Y. Woo, S. Wood, B. G. Wouters, A. J. Wright, D. W. Wright, M. H. Wright, C. L. Wu, D. Y. Wu, G. Wu, J. Wu, K. Wu, Y. Wu, Z. Wu, L. Xi, T. Xia, Q. Xiang, X. Xiao, R. Xing, H. Xiong, Q. Xu, Y. Xu, H. Xue, S. Yachida, S. Yakneen, R. Yamaguchi, T. N. Yamaguchi, M. Yamamoto, S. Yamamoto, H. Yamaue, F. Yang, H. Yang, J. Y. Yang, L. Yang, L. Yang, S. Yang, T. P. Yang, Y. Yang, X. Yao, M. L. Yaspo, L. Yates, C. Yau, C. Ye, K. Ye, V. D. Yellapantula, C. J. Yoon, S. S. Yoon, F. Yousif, J. Yu, K. Yu, W. Yu, Y. Yu, K. Yuan, Y. Yuan, D. Yuen, C. K. Yung, O. Zaikova, J. Zamora, M. Zapatka, J. C. Zenklusen, T. Zenz, N. Zeps, C. Z. Zhang, F. Zhang, H. Zhang, H. Zhang, H. Zhang, J. Zhang, J. Zhang, J. Zhang, X. Zhang, X. Zhang, Y. Zhang, Z. Zhang, Z. Zhao, L. Zheng, X. Zheng, W. Zhou, Y. Zhou, B. Zhu, H. Zhu, J. Zhu, S. Zhu, L. Zou, X. Zou, A. deFazio, N. van As, C. H. M. van Deurzen, M. J. van de Vijver, L. Van't Veer, and C. von Mering, "Pan-cancer analysis of whole genomes," *Nature*, vol. 578, pp. 82–93, Feb 2020.

[168] F. nez, F. os, I. s, J. Deu-Pons, I. Reyes-Salazar, C. Arnedo-Pac, L. Mularoni, O. Pich, J. Bonet, H. Kranas, A. Gonzalez-Perez, and N. Lopez-Bigas, "A compendium of mutational cancer driver genes," *Nat Rev Cancer*, vol. 20, pp. 555–572, Oct 2020.

[169] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA Cancer J Clin*, vol. 71, pp. 209–249, May 2021.

[170] N. E. Sharpless and D. S. Singer, "Progress and potential: The Cancer Moonshot," *Cancer Cell*, vol. 39, pp. 889–894, Jul 2021.

[171] N. N. C. Institute, "Most recent reported fiscal year budget 2021," https://www.cancer.gov/about-nci/budget/fact-book/data/recent-fiscal-year.

[172] S. Bamford, E. Dawson, S. Forbes, J. Clements, R. Pettett, A. Dogan, A. Flanagan, J. Teague, P. A. Futreal, M. R. Stratton, and R. Wooster, "The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website," *Br J Cancer*, vol. 91, pp. 355–358, Jul 2004.

[173] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton, "A census of human cancer genes," *Nat Rev Cancer*, vol. 4, pp. 177–183, Mar 2004.

[174] D. P. Lane, "Cancer. p53, guardian of the genome," *Nature*, vol. 358, pp. 15–16, Jul 1992.

[175] D. Uprety and A. A. Adjei, "KRAS: From undruggable to a druggable Cancer Target," *Cancer Treat Rev*, vol. 89, p. 102070, Sep 2020.

[176] F. Rascio, F. Spadaccino, M. T. Rocchetti, G. Castellano, G. Stallone, G. S. Netti, and E. Ranieri, "The Pathogenic Role of PI3K/AKT Pathway in Cancer Onset and Drug Resistance: An Updated Review," *Cancers (Basel)*, vol. 13, Aug 2021.

[177] M. miech, P. ski, H. Kono, C. Wardell, and H. Taniguchi, "Mutations in Cancer Progression and Their Possible Effects on Transcriptional Networks," *Genes (Basel)*, vol. 11, Nov 2020.

[178] E. Lakatos, M. J. Williams, R. O. Schenck, W. C. H. Cross, J. Househam, L. Zapata, B. Werner, C. Gatenbee, M. Robertson-Tessi, C. P. Barnes, A. R. A. Anderson, A. Sottoriva, and T. A. Graham, "Evolutionary dynamics of neoantigens in growing tumors," *Nat Genet*, vol. 52, pp. 1057–1066, Oct 2020.

[179] N. Kherreh, S. Cleary, and C. Seoighe, "No evidence that HLA genotype influences the driver mutations that occur in cancer patients," *Cancer Immunol Immunother*, vol. 71, pp. 819–827, Apr 2022.

[180] M. M. Pomerantz and M. L. Freedman, "The genetics of cancer risk," *Cancer J*, vol. 17, no. 6, pp. 416–422, 2011.

[181] P. Anand, A. B. Kunnumakkara, C. Sundaram, K. B. Harikumar, S. T. Tharakan, O. S. Lai, B. Sung, and B. B. Aggarwal, "Cancer is a preventable disease that requires major lifestyle changes," *Pharm Res*, vol. 25, pp. 2097–2116, Sep 2008.

[182] J. M. Hall, M. K. Lee, B. Newman, J. E. Morrow, L. A. Anderson, B. Huey, and M. C. King, "Linkage of early-onset familial breast cancer to chromosome 17q21," *Science*, vol. 250, pp. 1684–1689, Dec 1990.

[183] Y. Miki, J. Swensen, D. Shattuck-Eidens, P. A. Futreal, K. Harshman, S. Tavtigian, Q. Liu, C. Cochran, L. M. Bennett, and W. Ding, "A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1," *Science*, vol. 266, pp. 66–71, Oct 1994.

[184] R. Wooster, G. Bignell, J. Lancaster, S. Swift, S. Seal, J. Mangion, N. Collins, S. Gregory, C. Gumbs, and G. Micklem, "Identification of the breast cancer susceptibility gene BRCA2," *Nature*, vol. 378, no. 6559, pp. 789–792, 1995.

[185] R. Roy, J. Chun, and S. N. Powell, "BRCA1 and BRCA2: different roles in a common pathway of genome protection," *Nat Rev Cancer*, vol. 12, pp. 68–78, Dec 2011.

[186] P. Bhattacharya and T. W. McHugh, "Lynch syndrome," in *StatPearls [Internet]*, StatPearls Publishing, 2022.

[187] S. Haraldsdottir, T. Rafnar, W. L. Frankel, S. Einarsdottir, A. Sigurdsson, H. Hampel, P. Snaebjornsson, G. Masson, D. Weng, R. Arngrimsson, B. Kehr, A. Yilmaz, S. Haraldsson, P. Sulem, T. Stefansson, P. G. Shields, F. Sigurdsson, T. Bekaii-Saab, P. H. Moller, M. Steinarsdottir, K. Alexiusdottir, M. Hitchins, C. C. Pritchard, A. de la Chapelle, J. G. Jonasson, R. M. Goldberg, and K. Stefansson, "Comprehensive population-wide analysis of Lynch syndrome in Iceland reveals founder mutations in MSH6 and PMS2," *Nat Commun*, vol. 8, p. 14755, May 2017.

[188] N. Abu-Ghazaleh, V. Kaushik, A. Gorelik, M. Jenkins, and F. Macrae, "Worldwide prevalence of Lynch syndrome in patients with colorectal cancer: Systematic review and meta-analysis," *Genet Med*, vol. 24, pp. 971–985, May 2022.

[189] E. Barrow, J. Hill, and D. G. Evans, "Cancer risk in Lynch Syndrome," *Fam Cancer*, vol. 12, pp. 229–240, Jun 2013.

[190] M. Tonello, F. Nappo, L. Vassallo, R. Di Gaetano, C. Davoli, E. Pizzolato, O. De Simoni, C. Tassinari, A. Scapinello, P. Pilati, F. Loupakis, S. Lonardi, and A. Sommariva, "Complete pathological response of colorectal peritoneal metastases in Lynch syndrome after immunotherapy case report: is a paradigm shift in cytoreductive surgery needed?," *BMC Gastroenterol*, vol. 22, p. 17, Jan 2022.

[191] V. Laugel, "Cockayne syndrome: the expanding clinical and mutational spectrum," *Mech Ageing Dev*, vol. 134, no. 5-6, pp. 161–170, 2013.

[192] N. L. Batenburg, E. L. Thompson, E. A. Hendrickson, and X. D. Zhu, "Cockayne syndrome group B protein regulates DNA double-strand break repair and checkpoint activation," *EMBO J*, vol. 34, pp. 1399–1416, May 2015.

[193] A. C. Karikkineth, M. Scheibye-Knudsen, E. Fivenson, D. L. Croteau, and V. A. Bohr, "Cockayne syndrome: Clinical features, model systems and pathways," *Ageing Res Rev*, vol. 33, pp. 3–17, Jan 2017.

[194] A. R. Lehmann, D. McGibbon, and M. Stefanini, "Xeroderma pigmentosum," *Orphanet J Rare Dis*, vol. 6, p. 70, Nov 2011.

[195] P. T. Bradford, A. M. Goldstein, D. Tamura, S. G. Khan, T. Ueda, J. Boyle, K. S. Oh, K. Imoto, H. Inui, S. Moriwaki, S. Emmert, K. M. Pike, A. Raziuddin, T. M. Plona, J. J. DiGiovanna, M. A. Tucker, and K. H. Kraemer, "Cancer and neurologic degeneration in xeroderma pigmentosum: long term follow-up characterises the role of DNA repair," *J Med Genet*, vol. 48, pp. 168–176, Mar 2011.

[196] A. A. Yurchenko, I. Padioleau, B. T. Matkarimov, J. Soulier, A. Sarasin, and S. Nikolaev, "XPC deficiency increases risk of hematologic malignancies through mutator phenotype and characteristic mutational signature," *Nat Commun*, vol. 11, p. 5834, Nov 2020.

[197] K. S. Reid-Bayliss, S. T. Arron, L. A. Loeb, V. Bezrookove, and J. E. Cleaver, "Why Cockayne syndrome patients do not get cancer despite their DNA repair deficiency," *Proc Natl Acad Sci U S A*, vol. 113, pp. 10151–10156, Sep 2016.

[198] E. Uribe-Bojanini, S. Hernandez-Quiceno, and A. M. Cock-Rada, "Xeroderma Pigmentosum with Severe Neurological Manifestations/De Sanctis-Cacchione Syndrome and a Novel XPC Mutation," *Case Rep Med*, vol. 2017, p. 7162737, 2017.

[199] A. D. Andrews, S. F. Barrett, and J. H. Robbins, "Xeroderma pigmentosum neurological abnormalities correlate with colony-forming ability after ultraviolet radiation," *Proc Natl Acad Sci U S A*, vol. 75, pp. 1984–1988, Apr 1978.

[200] M. Nishioka, M. Bundo, K. Iwamoto, and T. Kato, "Somatic mutations in the human brain: implications for psychiatric research," *Mol Psychiatry*, vol. 24, pp. 839–856, Jun 2019.

[201] R. Kacher, F. X. Lejeune, S. l, C. Cazeneuve, A. Brice, S. Humbert, and A. Durr, "Propensity for somatic expansion increases over the course of life in Huntington disease," *Elife*, vol. 10, May 2021.

[202] Y. McLennan, J. Polussa, F. Tassone, and R. Hagerman, "Fragile x syndrome," *Curr Genomics*, vol. 12, pp. 216–224, May 2011.

[203] H. Ogawa, K. Horitani, Y. Izumiya, and S. Sano, “Somatic Mosaicism in Biology and Disease,” *Annu Rev Physiol*, vol. 84, pp. 113–133, Feb 2022.

[204] D. Freed, E. L. Stevens, and J. Pevsner, “Somatic mosaicism in the human genome,” *Genes (Basel)*, vol. 5, pp. 1064–1094, Dec 2014.

[205] Z. Wang, S. Zhu, Y. Jia, Y. Wang, N. Kubota, N. Fujiwara, R. Gordillo, C. Lewis, M. Zhu, T. Sharma, L. Li, Q. Zeng, Y. H. Lin, M. H. Hsieh, P. Gopal, T. Wang, M. Hoare, P. Campbell, Y. Hoshida, and H. Zhu, “Positive selection of somatically mutated clones identifies adaptive pathways in metabolic liver disease,” *Cell*, vol. 186, pp. 1968–1984, Apr 2023.

[206] R. Bhattacharya, S. M. Zekavat, J. Haessler, M. Fornage, L. Raffield, M. M. Uddin, A. G. Bick, A. Niroula, B. Yu, C. Gibson, G. Griffin, A. C. Morrison, B. M. Psaty, W. T. Longstreth, J. C. Bis, S. S. Rich, J. I. Rotter, R. P. Tracy, A. Correa, S. Seshadri, A. Johnson, J. M. Collins, K. M. Hayden, T. E. Madsen, C. M. Ballantyne, S. Jaiswal, B. L. Ebert, C. Kooperberg, J. E. Manson, E. A. Whitsel, P. Natarajan, and A. P. Reiner, “Clonal Hematopoiesis Is Associated With Higher Risk of Stroke,” *Stroke*, vol. 53, pp. 788–797, Mar 2022.

[207] S. Jaiswal, P. Natarajan, A. J. Silver, C. J. Gibson, A. G. Bick, E. Shvartz, M. McConkey, N. Gupta, S. Gabriel, D. Ardissino, U. Baber, R. Mehran, V. Fuster, J. Danesh, P. Frossard, D. Saleheen, O. Melander, G. K. Sukhova, D. Neuberg, P. Libby, S. Kathiresan, and B. L. Ebert, “Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease,” *N Engl J Med*, vol. 377, pp. 111–121, Jul 2017.

[208] C. S. Marnell, A. Bick, and P. Natarajan, “Clonal hematopoiesis of indeterminate potential (CHIP): Linking somatic mutations, hematopoiesis, chronic inflammation and cardiovascular disease,” *J Mol Cell Cardiol*, vol. 161, pp. 98–105, Dec 2021.

[209] M. Xie, C. Lu, J. Wang, M. D. McLellan, K. J. Johnson, M. C. Wendl, J. F. McMichael, H. K. Schmidt, V. Yellapantula, C. A. Miller, B. A. Ozenberger, J. S. Welch, D. C. Link, M. J. Walter, E. R. Mardis, J. F. Dipersio, F. Chen, R. K. Wilson, T. J. Ley, and L. Ding, “Age-related mutations associated with clonal hematopoietic expansion and malignancies,” *Nat Med*, vol. 20, pp. 1472–1478, Dec 2014.

[210] G. Genovese, A. K. hler, R. E. Handsaker, J. Lindberg, S. A. Rose, S. F. Bakhoum, K. Chambert, E. Mick, B. M. Neale, M. Fromer, S. M. Purcell, O. Svantesson, M. n, M. glund, S. Lehmann, S. B. Gabriel, J. L. Moran, E. S. Lander, P. F. Sullivan, P. Sklar, H. nberg, C. M. Hultman, and S. A. McCarroll, “Clonal hematopoiesis and blood-cancer

risk inferred from blood DNA sequence," *N Engl J Med*, vol. 371, pp. 2477–2487, Dec 2014.

[211] S. Jaiswal, P. Fontanillas, J. Flannick, A. Manning, P. V. Grauman, B. G. Mar, R. C. Lindsley, C. H. Mermel, N. Burtt, A. Chavez, J. M. Higgins, V. Moltchanov, F. C. Kuo, M. J. Kluk, B. Henderson, L. Kinnunen, H. A. Koistinen, C. Ladenvall, G. Getz, A. Correa, B. F. Banahan, S. Gabriel, S. Kathiresan, H. M. Stringham, M. I. McCarthy, M. Boehnke, J. Tuomilehto, C. Haiman, L. Groop, G. Atzmon, J. G. Wilson, D. Neuberg, D. Altshuler, and B. L. Ebert, "Age-related clonal hematopoiesis associated with adverse outcomes," *N Engl J Med*, vol. 371, pp. 2488–2498, Dec 2014.

[212] J. L. Abkowitz, S. N. Catlin, M. T. McCallie, and P. Guttorp, "Evidence that the number of hematopoietic stem cells per animal is conserved in mammals," *Blood*, vol. 100, pp. 2665–2667, Oct 2002.

[213] H. Holstege, W. Pfeiffer, D. Sie, M. Hulsman, T. J. Nicholas, C. C. Lee, T. Ross, J. Lin, M. A. Miller, B. Ylstra, H. Meijers-Heijboer, M. H. Brugman, F. J. Staal, G. Holstege, M. J. Reinders, T. T. Harkins, S. Levy, and E. A. Sistermans, "Somatic mutations found in the healthy blood compartment of a 115-yr-old woman demonstrate oligoclonal hematopoiesis," *Genome Res*, vol. 24, pp. 733–742, May 2014.

[214] M. A. Fabre, J. G. de Almeida, E. Fiorillo, E. Mitchell, A. Damaskou, J. Rak, V. u, M. Marongiu, M. S. Chapman, M. S. Vijayabaskar, J. Baxter, C. Hardy, F. Abascal, N. Williams, J. Nangalia, I. Martincorena, P. J. Campbell, E. F. McKinney, F. Cucca, M. Gerstung, and G. S. Vassiliou, "The longitudinal dynamics and natural history of clonal haematopoiesis," *Nature*, vol. 606, pp. 335–342, Jun 2022.

[215] N. A. Robertson, E. Latorre-Crespo, M. Terradas-Terradas, J. Lemos-Portela, A. C. Purcell, B. J. Livesey, R. F. Hillary, L. Murphy, A. Fawkes, L. MacGillivray, M. Copland, R. E. Marioni, J. A. Marsh, S. E. Harris, S. R. Cox, I. J. Deary, L. J. Schumacher, K. Kirschner, and T. Chandra, "Longitudinal dynamics of clonal hematopoiesis identifies gene-specific fitness effects," *Nat Med*, vol. 28, pp. 1439–1446, Jul 2022.

[216] K. Wang, H. Liu, Q. Hu, L. Wang, J. Liu, Z. Zheng, W. Zhang, J. Ren, F. Zhu, and G. H. Liu, "Epigenetic regulation of aging: implications for interventions of aging and diseases," *Signal Transduct Target Ther*, vol. 7, p. 374, Nov 2022.

[217] Y. Ishimi, M. Kojima, F. Takeuchi, T. Miyamoto, M. Yamada, and F. Hanaoka, “Changes in chromatin structure during aging of human skin fibroblasts,” *Exp Cell Res*, vol. 169, pp. 458–467, Apr 1987.

[218] E. L. Greer, T. J. Maures, A. G. Hauswirth, E. M. Green, D. S. Leeman, G. S. Maro, S. Han, M. R. Banko, O. Gozani, and A. Brunet, “Members of the H3K4 trimethylation complex regulate lifespan in a germline-dependent manner in C. elegans,” *Nature*, vol. 466, pp. 383–387, Jul 2010.

[219] F. Manders, R. van Boxtel, and S. Middelkamp, “The Dynamics of Somatic Mutagenesis During Life in Humans,” *Front Aging*, vol. 2, p. 802407, 2021.

[220] L. Szilard, “ON THE NATURE OF THE AGING PROCESS,” *Proc Natl Acad Sci U S A*, vol. 45, pp. 30–45, Jan 1959.

[221] L. E. ORGEL, “The maintenance of the accuracy of protein synthesis and its relevance to ageing,” *Proc Natl Acad Sci U S A*, vol. 49, pp. 517–521, Apr 1963.

[222] B. Milholland, Y. Suh, and J. Vijg, “Mutation and catastrophe in the aging genome,” *Exp Gerontol*, vol. 94, pp. 34–40, Aug 2017.

[223] J. W. Gowen, “ON CHROMOSOME BALANCE AS A FACTOR IN DURATION OF LIFE,” *J Gen Physiol*, vol. 14, pp. 447–461, Mar 1931.

[224] A. M. Clark, H. A. Bertrand, and R. E. Smith, “Life span differences between haploid and diploid males of habrobracon serinopae after exposure as adults to x rays,” *The American Naturalist*, vol. 97, no. 895, pp. 203–208, 1963.

[225] P. n, B. Pettersson, and M. n, “Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay,” *Anal Biochem*, vol. 208, pp. 171–175, Jan 1993.

[226] M. Ronaghi, S. Karamohamed, B. Pettersson, M. n, and P. n, “Real-time DNA sequencing using detection of pyrophosphate release,” *Anal Biochem*, vol. 242, pp. 84–89, Nov 1996.

[227] J. M. Heather and B. Chain, “The sequence of sequencers: The history of sequencing DNA,” *Genomics*, vol. 107, pp. 1–8, Jan 2016.

[228] K. V. Voelkerding, S. A. Dames, and J. D. Durtschi, “Next-generation sequencing: from basic research to diagnostics,” *Clin Chem*, vol. 55, pp. 641–658, Apr 2009.

[229] C. V. Van Hout, I. Tachmazidou, J. D. Backman, J. D. Hoffman, D. Liu, A. K. Pandey, C. Gonzaga-Jauregui, S. Khalid, B. Ye, N. Banerjee, A. H. Li, C. O'Dushlaine, A. Marcketta, J. Staples, C. Schurmann, A. Hawes, E. Maxwell, L. Barnard, A. Lopez, J. Penn, L. Habegger, A. L. Blumenfeld, X. Bai, S. O'Keeffe, A. Yadav, K. Praveen, M. Jones, W. J. Salerno, W. K. Chung, I. Surakka, C. J. Willer, K. Hveem, J. B. Leader, D. J. Carey, D. H. Ledbetter, L. Cardon, G. D. Yancopoulos, A. Economides, G. Coppola, A. R. Shuldiner, S. Balasubramanian, M. Cantor, M. R. Nelson, J. Whittaker, J. G. Reid, J. Marchini, J. D. Overton, R. A. Scott, G. R. Abecasis, L. Yerges-Armstrong, and A. Baras, "Exome sequencing and characterization of 49,960 individuals in the UK Biobank," *Nature*, vol. 586, pp. 749–756, Oct 2020.

[230] K. Mitchell, J. J. Brito, I. Mandric, Q. Wu, S. Knyazev, S. Chang, L. S. Martin, A. Karlsberg, E. Gerasimov, R. Littman, B. L. Hill, N. C. Wu, H. T. Yang, K. Hsieh, L. Chen, E. Littman, T. Shabani, G. Enik, D. Yao, R. Sun, J. Schroeder, E. Eskin, A. Zelikovsky, P. Skums, M. Pop, and S. Mangul, "Benchmarking of computational error-correction methods for next-generation sequencing data," *Genome Biol*, vol. 21, p. 71, Mar 2020.

[231] J. Zhang, K. Kobert, T. Flouri, and A. Stamatakis, "PEAR: a fast and accurate Illumina Paired-End reAd mergeR," *Bioinformatics*, vol. 30, pp. 614–620, Mar 2014.

[232] J. J. Salk, M. W. Schmitt, and L. A. Loeb, "Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations," *Nat Rev Genet*, vol. 19, pp. 269–285, May 2018.

[233] F. Pfeiffer, C. ber, M. Blank, K. ndler, M. Beyer, J. L. Schultze, and G. Mayer, "Systematic evaluation of error rates and causes in short samples in next-generation sequencing," *Sci Rep*, vol. 8, p. 10950, Jul 2018.

[234] P. Murat, G. Guilbaud, and J. E. Sale, "DNA polymerase stalling at structured DNA constrains the expansion of short tandem repeats," *Genome Biol*, vol. 21, p. 209, Aug 2020.

[235] H. Do and A. Dobrovic, "Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization," *Clin Chem*, vol. 61, pp. 64–71, Jan 2015.

[236] D. D. Dodani, M. H. Nguyen, R. D. Morin, M. A. Marra, and R. D. Corbett, "Combinatorial and Machine Learning Approaches for Improved Somatic Variant Calling From Formalin-Fixed Paraffin-Embedded Genome Sequence Data," *Front Genet*, vol. 13, p. 834764, 2022.

[237] V. Potapov and J. L. Ong, “Examining Sources of Error in PCR by Single-Molecule Sequencing,” *PLoS One*, vol. 12, no. 1, p. e0169774, 2017.

[238] R. Sun, M. I. Love, T. Zemojtel, A. K. Emde, H. R. Chung, M. Vingron, and S. A. Haas, “Breakpointer: using local mapping artifacts to support sequence breakpoint discovery from single-end reads,” *Bioinformatics*, vol. 28, pp. 1024–1025, Apr 2012.

[239] E. M. Jewett, M. cken, and Y. S. Song, “The Effects of Population Size Histories on Estimates of Selection Coefficients from Time-Series Genetic Data,” *Mol Biol Evol*, vol. 33, pp. 3002–3027, Nov 2016.

[240] A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis, A. Auton, G. R. Abecasis, D. M. Altshuler, R. M. Durbin, G. R. Abecasis, D. R. Bentley, A. Chakravarti, A. G. Clark, P. Donnelly, E. E. Eichler, P. Flicek, S. B. Gabriel, R. A. Gibbs, E. D. Green, M. E. Hurles, B. M. Knoppers, J. O. Korbel, E. S. Lander, C. Lee, H. Lehrach, E. R. Mardis, G. T. Marth, G. A. McVean, D. A. Nickerson, J. P. Schmidt, S. T. Sherry, J. Wang, R. K. Wilson, R. A. Gibbs, E. Boerwinkle, H. Doddapaneni, Y. Han, V. Korchina, C. Kovar, S. Lee, D. Muzny, J. G. Reid, Y. Zhu, J. Wang, Y. Chang, Q. Feng, X. Fang, X. Guo, M. Jian, H. Jiang, X. Jin, T. Lan, G. Li, J. Li, Y. Li, S. Liu, X. Liu, Y. Lu, X. Ma, M. Tang, B. Wang, G. Wang, H. Wu, R. Wu, X. Xu, Y. Yin, D. Zhang, W. Zhang, J. Zhao, M. Zhao, X. Zheng, E. S. Lander, D. M. Altshuler, S. B. Gabriel, N. Gupta, N. Gharani, L. H. Toji, N. P. Gerry, A. M. Resch, P. Flicek, J. Barker, L. Clarke, L. Gil, S. E. Hunt, G. Kelman, E. Kulesha, R. Leinonen, W. M. McLaren, R. Radhakrishnan, A. Roa, D. Smirnov, R. E. Smith, I. Streeter, A. Thormann, I. Toneva, B. Vaughan, X. Zheng-Bradley, D. R. Bentley, R. Grocock, S. Humphray, T. James, Z. Kingsbury, H. Lehrach, R. Sudbrak, M. W. Albrecht, V. S. Amstislavskiy, T. A. Borodina, M. Lienhard, F. Mertes, M. Sultan, B. Timmermann, M. L. Yaspo, E. R. Mardis, R. K. Wilson, L. Fulton, R. Fulton, S. T. Sherry, V. Ananiev, Z. Belaia, D. Beloslyudtsev, N. Bouk, C. Chen, D. Church, R. Cohen, C. Cook, J. Garner, T. Hefferon, M. Kimelman, C. Liu, J. Lopez, P. Meric, C. O’Sullivan, Y. Ostapchuk, L. Phan, S. Ponomarov, V. Schneider, E. Shekhtman, K. Sirotkin, D. Slotta, H. Zhang, G. A. McVean, R. M. Durbin, S. Balasubramaniam, J. Burton, P. Danecek, T. M. Keane, A. Kolb-Kokocinski, S. McCarthy, J. Stalker, M. Quail, R. M. Durbin, S. Balasubramaniam, J. Burton, P. Danecek, T. M. Keane, A. Kolb-Kokocinski, S. McCarthy, J. Stalker, M. Quail, J. P. Schmidt, C. J. Davies, J. Gollub, T. Webster, B. Wong, Y. Zhan, A. Auton, C. L. Campbell, Y. Kong, A. Marcketta, R. A. Gibbs, F. Yu, L. Antunes, M. Bainbridge, D. Muzny, A. Sabo,

Z. Huang, J. Wang, L. J. Coin, L. Fang, X. Guo, X. Jin, G. Li, Q. Li, Y. Li, Z. Li, H. Lin, B. Liu, R. Luo, H. Shao, Y. Xie, C. Ye, C. Yu, F. Zhang, H. Zheng, H. Zhu, C. Alkan, E. Dal, F. Kahveci, G. T. Marth, E. P. Garrison, D. Kural, W. P. Lee, W. F. Leong, M. Stromberg, A. N. Ward, J. Wu, M. Zhang, M. J. Daly, M. A. DePristo, R. E. Handsaker, D. M. Altshuler, E. Banks, G. Bhatia, G. Del Angel, S. B. Gabriel, G. Genovese, N. Gupta, H. Li, S. Kashin, E. S. Lander, S. A. McCarroll, J. C. Nemesh, R. E. Poplin, S. C. Yoon, J. Lihm, V. Makarov, A. G. Clark, S. Gottipati, A. Keinan, J. L. Rodriguez-Flores, J. O. Korbel, T. Rausch, M. H. Fritz, A. M. tz, P. Flicek, K. Beal, L. Clarke, A. Datta, J. Herrero, W. M. McLaren, G. R. Ritchie, R. E. Smith, D. Zerbino, X. Zheng-Bradley, P. C. Sabeti, I. Shlyakhter, S. F. Schaffner, J. Vitti, D. N. Cooper, E. V. Ball, P. D. Stenson, D. R. Bentley, B. Barnes, M. Bauer, R. K. Cheetham, A. Cox, M. Eberle, S. Humphray, S. Kahn, L. Murray, J. Peden, R. Shaw, E. E. Kenny, M. A. Batzer, M. K. Konkel, J. A. Walker, D. G. MacArthur, M. Lek, R. Sudbrak, V. S. Amstislavskiy, R. Herwig, E. R. Mardis, L. Ding, D. C. Koboldt, D. Larson, K. Ye, S. Gravel, A. Swaroop, E. Chew, T. Lappalainen, Y. Erlich, M. Gymrek, T. F. Willems, J. T. Simpson, M. D. Shriver, J. A. Rosenfeld, C. D. Bustamante, S. B. Montgomery, F. M. De La Vega, J. K. Byrnes, A. W. Carroll, M. K. DeGorter, P. Lacroute, B. K. Maples, A. R. Martin, A. Moreno-Estrada, S. S. Shringarpure, F. Zakharia, E. Halperin, Y. Baran, C. Lee, E. Cerveira, J. Hwang, A. Malhotra, D. Plewczynski, K. Radew, M. Romanovitch, C. Zhang, F. C. Hyland, D. W. Craig, A. Christoforides, N. Homer, T. Izatt, A. A. Kurdoglu, S. A. Sinari, K. Squire, S. T. Sherry, C. Xiao, J. Sebat, D. Antaki, M. Gujral, A. Noor, K. Ye, E. G. Burchard, R. D. Hernandez, C. R. Gignoux, D. Haussler, S. J. Katzman, W. J. Kent, B. Howie, A. Ruiz-Linares, E. T. Dermitzakis, S. E. Devine, G. R. Abecasis, H. M. Kang, J. M. Kidd, T. Blackwell, S. Caron, W. Chen, S. Emery, L. Fritsche, C. Fuchsberger, G. Jun, B. Li, R. Lyons, C. Scheller, C. Sidore, S. Song, E. Sliwerska, D. Taliun, A. Tan, R. Welch, M. K. Wing, X. Zhan, P. Awadalla, A. Hodgkinson, Y. Li, X. Shi, A. Quitadamo, G. Lunter, G. A. McVean, J. L. Marchini, S. Myers, C. Churchhouse, O. Delaneau, A. Gupta-Hinch, W. Kretzschmar, Z. Iqbal, I. Mathieson, A. Menelaou, A. Rimmer, D. K. Xifara, T. K. Oleksyk, Y. Fu, X. Liu, M. Xiong, L. Jorde, D. Witherspoon, J. Xing, E. E. Eichler, B. L. Browning, S. R. Browning, F. Hormozdiari, P. H. Sudmant, E. Khurana, R. M. Durbin, M. E. Hurles, C. Tyler-Smith, C. A. Albers, Q. Ayub, S. Balasubramaniam, Y. Chen, V. Colonna, P. Danecek, L. Jostins, T. M. Keane, S. McCarthy, K. Walter, Y. Xue, M. B. Gerstein, A. Abyzov, S. Balasubramanian, J. Chen, D. Clarke, Y. Fu, A. O. Harmanci, M. Jin, D. Lee, J. Liu, X. J. Mu, J. Zhang, Y. Zhang, M. B. Gerstein, A. Abyzov, S. Balasubramanian,

J. Chen, D. Clarke, Y. Fu, A. O. Harmanci, M. Jin, D. Lee, J. Liu, X. J. Mu, J. Zhang, Y. Zhang, Y. Li, R. Luo, H. Zhu, C. Alkan, E. Dal, F. Kahveci, G. T. Marth, E. P. Garrison, D. Kural, W. P. Lee, A. N. Ward, J. Wu, M. Zhang, S. A. McCarroll, R. E. Handsaker, D. M. Altshuler, E. Banks, G. Del Angel, G. Genovese, C. Hartl, H. Li, S. Kashin, J. C. Nemesh, K. Shakir, S. C. Yoon, J. Lihm, V. Makarov, J. Degenhardt, J. O. Korbel, M. H. Fritz, S. Meiers, B. Raeder, T. Rausch, A. M. tz, P. Flicek, F. P. Casale, L. Clarke, R. E. Smith, O. Stegle, X. Zheng-Bradley, D. R. Bentley, B. Barnes, R. K. Cheetham, M. Eberle, S. Humphray, S. Kahn, L. Murray, R. Shaw, E. W. Lameijer, M. A. Batzer, M. K. Konkel, J. A. Walker, L. Ding, I. Hall, K. Ye, P. Lacroute, C. Lee, E. Cerveira, A. Malhotra, J. Hwang, D. Plewczynski, K. Radew, M. Romanovitch, C. Zhang, D. W. Craig, N. Homer, D. Church, C. Xiao, J. Sebat, D. Antaki, V. Bafna, J. Michaelson, K. Ye, S. E. Devine, E. J. Gardner, G. R. Abecasis, J. M. Kidd, R. E. Mills, G. Dayama, S. Emery, G. Jun, X. Shi, A. Quitadamo, G. Lunter, G. A. McVean, K. Chen, X. Fan, Z. Chong, T. Chen, D. Witherspoon, J. Xing, E. E. Eichler, M. J. Chaisson, F. Hormozdiari, J. Huddleston, M. Malig, B. J. Nelson, P. H. Sudmant, N. F. Parrish, E. Khurana, M. E. Hurles, B. Blackburne, S. J. Lindsay, Z. Ning, K. Walter, Y. Zhang, M. B. Gerstein, A. Abyzov, J. Chen, D. Clarke, H. Lam, X. J. Mu, C. Sisu, J. Zhang, Y. Zhang, M. B. Gerstein, A. Abyzov, J. Chen, D. Clarke, H. Lam, X. J. Mu, C. Sisu, J. Zhang, Y. Zhang, R. A. Gibbs, F. Yu, M. Bainbridge, D. Challis, U. S. Evani, C. Kovar, J. Lu, D. Muzny, U. Nagaswamy, J. G. Reid, A. Sabo, J. Yu, X. Guo, W. Li, Y. Li, R. Wu, G. T. Marth, E. P. Garrison, W. F. Leong, A. N. Ward, G. Del Angel, M. A. DePristo, S. B. Gabriel, N. Gupta, C. Hartl, R. E. Poplin, A. G. Clark, J. L. Rodriguez-Flores, P. Flicek, L. Clarke, R. E. Smith, X. Zheng-Bradley, D. G. MacArthur, E. R. Mardis, R. Fulton, D. C. Koboldt, S. Gravel, C. D. Bustamante, D. W. Craig, A. Christoforides, N. Homer, T. Izatt, S. T. Sherry, C. Xiao, E. T. Dermitzakis, G. R. Abecasis, H. Min Kang, G. A. McVean, M. B. Gerstein, S. Balasubramanian, L. Habegger, M. B. Gerstein, S. Balasubramanian, L. Habegger, H. Yu, P. Flicek, L. Clarke, F. Cunningham, I. Dunham, D. Zerbino, X. Zheng-Bradley, K. Lage, J. B. Jespersen, H. Horn, S. B. Montgomery, M. K. DeGorter, E. Khurana, C. Tyler-Smith, Y. Chen, V. Colonna, Y. Xue, M. B. Gerstein, S. Balasubramanian, Y. Fu, D. Kim, M. B. Gerstein, S. Balasubramanian, Y. Fu, D. Kim, A. Auton, A. Marcketta, R. Desalle, A. Narechania, M. A. Sayres, E. P. Garrison, R. E. Handsaker, S. Kashin, S. A. McCarroll, J. L. Rodriguez-Flores, P. Flicek, L. Clarke, X. Zheng-Bradley, Y. Erlich, M. Gymrek, T. F. Willems, C. D. Bustamante, F. L. Mendez, G. D. Poznik, P. A. Underhill, C. Lee, E. Cerveira, A. Malhotra, M. Romanovitch, C. Zhang, G. R. Abecasis, L. Coin, H. Shao, D. Mittelman, C. Tyler-Smith,

Q. Ayub, R. Banerjee, M. Cerezo, Y. Chen, T. W. Fitzgerald, S. Louzada, A. Massaia, S. McCarthy, G. R. Ritchie, Y. Xue, F. Yang, C. Tyler-Smith, Q. Ayub, R. Banerjee, M. Cerezo, Y. Chen, T. W. Fitzgerald, S. Louzada, A. Massaia, S. McCarthy, G. R. Ritchie, Y. Xue, F. Yang, R. A. Gibbs, C. Kovar, D. Kalra, W. Hale, D. Muzny, J. G. Reid, J. Wang, X. Dan, X. Guo, G. Li, Y. Li, C. Ye, X. Zheng, D. M. Altshuler, P. Flicek, L. Clarke, X. Zheng-Bradley, D. R. Bentley, A. Cox, S. Humphray, S. Kahn, R. Sudbrak, M. W. Albrecht, M. Lienhard, D. Larson, D. W. Craig, T. Izatt, A. A. Kurdoglu, S. T. Sherry, C. Xiao, D. Haussler, G. R. Abecasis, G. A. McVean, R. M. Durbin, S. Balasubramaniam, T. M. Keane, S. McCarthy, J. Stalker, R. M. Durbin, S. Balasubramaniam, T. M. Keane, S. McCarthy, J. Stalker, A. Chakravarti, B. M. Knoppers, G. R. Abecasis, K. C. Barnes, C. Beiswanger, E. G. Burchard, C. D. Bustamante, H. Cai, H. Cao, R. M. Durbin, N. P. Gerry, N. Gharani, R. A. Gibbs, C. R. Gignoux, S. Gravel, B. Henn, D. Jones, L. Jorde, J. S. Kaye, A. Keinan, A. Kent, A. Kerasidou, Y. Li, R. Mathias, G. A. McVean, A. Moreno-Estrada, P. N. Ossorio, M. Parker, A. M. Resch, C. N. Rotimi, C. D. Royal, K. Sandoval, Y. Su, R. Sudbrak, Z. Tian, S. Tishkoff, L. H. Toji, C. Tyler-Smith, M. Via, Y. Wang, H. Yang, L. Yang, J. Zhu, W. Bodmer, G. Bedoya, A. Ruiz-Linares, Z. Cai, Y. Gao, J. Chu, L. Peltonen, A. Garcia-Montero, A. Orfao, J. Dutil, J. C. Martinez-Cruzado, T. K. Oleksyk, K. C. Barnes, R. A. Mathias, A. Hennis, H. Watson, C. McKenzie, F. Qadri, R. LaRocque, P. C. Sabeti, J. Zhu, X. Deng, P. C. Sabeti, D. Asogun, O. Folarin, C. Happi, O. Omoniwa, M. Stremlau, R. Tariyal, M. Jallow, F. S. Joof, T. Corrah, K. Rockett, D. Kwiatkowski, J. Kooner, T. T. n, S. J. Dunstan, N. T. Hang, R. Fonnie, R. Garry, L. Kanneh, L. Moses, P. C. Sabeti, J. Schieffelin, D. S. Grant, C. Gallo, G. Poletti, D. Saleheen, A. Rasheed, D. Saleheen, A. Rasheed, L. D. Brooks, A. L. Felsenfeld, J. E. McEwen, Y. Vaydylevich, E. D. Green, A. Duncanson, M. Dunn, J. A. Schloss, J. Wang, H. Yang, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. Min Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. Min Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, and G. R. Abecasis, "A global reference for human genetic variation," *Nature*, vol. 526, pp. 68–74, Oct 2015.

[241] K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. ldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, L. D. Gauthier, H. Brand, M. Solomonson, N. A. Watts, D. Rhodes, M. Singer-Berk, E. M. England, E. G. Seaby, J. A. Kosmicki, R. K. Walters, K. Tashman, Y. Farjoun, E. Banks, T. Poterba, A. Wang, C. Seed, N. Whiffin, J. X. Chong, K. E. Samocha, E. Pierce-Hoffman, Z. Zappala, A. H. O'Donnell-Luria, E. V. Minikel, B. Weisburd, M. Lek, J. S. Ware, C. Vittal,

I. M. Armean, L. Bergelson, K. Cibulskis, K. M. Connolly, M. Covarrubias, S. Donnelly, S. Ferriera, S. Gabriel, J. Gentry, N. Gupta, T. Jeandet, D. Kaplan, C. Llanwarne, R. Munshi, S. Novod, N. Petrillo, D. Roazen, V. Ruano-Rubio, A. Saltzman, M. Schleicher, J. Soto, K. Tibbetts, C. Tolonen, G. Wade, M. E. Talkowski, B. M. Neale, M. J. Daly, D. G. MacArthur, C. A. Aguilar Salinas, T. Ahmad, C. M. Albert, D. Ardissino, G. Atzmon, J. Barnard, L. Beaugerie, E. J. Benjamin, M. Boehnke, L. L. Bonnycastle, E. P. Bottinger, D. W. Bowden, M. J. Bown, J. C. Chambers, J. C. Chan, D. Chasman, J. Cho, M. K. Chung, B. Cohen, A. Correa, D. Dabelea, M. J. Daly, D. Darbar, R. Duggirala, J. Dupuis, P. T. Ellinor, R. Elosua, J. Erdmann, T. Esko, M. Ã¤, J. Florez, A. Franke, G. Getz, B. Glaser, S. J. Glatt, D. Goldstein, C. Gonzalez, L. Groop, C. Haiman, C. Hanis, M. Harms, M. Hiltunen, M. M. Holi, C. M. Hultman, M. Kallela, J. Kaprio, S. Kathiresan, B. J. Kim, Y. J. Kim, G. Kirov, J. Kooner, S. Koskinen, H. M. Krumholz, S. Kugathasan, S. H. Kwak, M. Laakso, T. ki, R. J. F. Loos, S. A. Lubitz, R. C. W. Ma, D. G. MacArthur, J. Marrugat, K. M. Mattila, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, J. B. Meigs, O. Melander, A. Metspalu, B. M. Neale, P. M. Nilsson, M. C. O'Donovan, D. Ongur, L. Orozco, M. J. Owen, C. N. A. Palmer, A. Palotie, K. S. Park, C. Pato, A. E. Pulver, N. Rahman, A. M. Remes, J. D. Rioux, S. Ripatti, D. M. Roden, D. Saleheen, V. Salomaa, N. J. Samani, J. Scharf, H. Schunkert, M. B. Shoemaker, P. Sklar, H. Soininen, H. Sokol, T. Spector, P. F. Sullivan, J. Suvisaari, E. S. Tai, Y. Y. Teo, T. Tiinamaija, M. Tsuang, D. Turner, T. Tusie-Luna, E. Vartiainen, M. P. Vawter, J. S. Ware, H. Watkins, R. K. Weersma, M. Wessman, J. G. Wilson, and R. J. Xavier, "The mutational constraint spectrum quantified from variation in 141,456 humans," *Nature*, vol. 581, pp. 434–443, May 2020.

[242] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, "dbSNP: the NCBI database of genetic variation," *Nucleic Acids Res*, vol. 29, pp. 308–311, Jan 2001.

[243] J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher, "Common SNPs explain a large proportion of the heritability for human height," *Nat Genet*, vol. 42, pp. 565–569, Jul 2010.

[244] H. Liu, L. Han, G. Kang, M. Zhang, and R. Cheng, "Editorial: Statistical Methods, Computing and Resources for Genome-Wide Association Studies," *Front Genet*, vol. 12, p. 714894, 2021.

[245] P. R. Loh, G. Kichaev, S. Gazal, A. P. Schoech, and A. L. Price, "Mixed-model association for biobank-scale datasets," *Nat Genet*, vol. 50, pp. 906–908, Jul 2018.

[246] L. Jiang, Z. Zheng, T. Qi, K. E. Kemper, N. R. Wray, P. M. Visscher, and J. Yang, "A resource-efficient tool for mixed model association analysis of large-scale data," *Nat Genet*, vol. 51, pp. 1749–1755, Dec 2019.

[247] J. Mbatchou, L. Barnard, J. Backman, A. Marcketta, J. A. Kosmicki, A. Ziyatdinov, C. Benner, C. O'Dushlaine, M. Barber, B. Boutkov, L. Habegger, M. Ferreira, A. Baras, J. Reid, G. Abecasis, E. Maxwell, and J. Marchini, "Computationally efficient whole-genome regression for quantitative and binary traits," *Nat Genet*, vol. 53, pp. 1097–1103, Jul 2021.

[248] M. Imamura, A. Takahashi, T. Yamauchi, K. Hara, K. Yasuda, N. Grarup, W. Zhao, X. Wang, A. Huerta-Chagoya, C. Hu, S. Moon, J. Long, S. H. Kwak, A. Rasheed, R. Saxena, R. C. Ma, Y. Okada, M. Iwata, J. Hosoe, N. Shojima, M. Iwasaki, H. Fujita, K. Suzuki, J. Danesh, T. rgensen, M. E. rgensen, D. R. Witte, I. Brandslund, C. Christensen, T. Hansen, J. M. Mercader, J. Flannick, H. as, N. P. Burtt, R. Zhang, Y. J. Kim, W. Zheng, J. R. Singh, C. H. Tam, H. Hirose, H. Maegawa, C. Ito, K. Kaku, H. Watada, Y. Tanaka, K. Tobe, R. Kawamori, M. Kubo, Y. S. Cho, J. C. Chan, D. Sanghera, P. Frossard, K. S. Park, X. O. Shu, B. J. Kim, J. C. Florez, T. Luna, W. Jia, E. S. Tai, O. Pedersen, D. Saleheen, S. Maeda, and T. Kadowaki, "Genome-wide association studies in the Japanese population identify seven novel loci for type 2 diabetes," *Nat Commun*, vol. 7, p. 10531, Jan 2016.

[249] J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher, "GCTA: a tool for genome-wide complex trait analysis," *Am J Hum Genet*, vol. 88, pp. 76–82, Jan 2011.

[250] R. J. Klein, C. Zeiss, E. Y. Chew, J. Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, M. B. Bracken, F. L. Ferris, J. Ott, C. Barnstable, and J. Hoh, "Complement factor H polymorphism in age-related macular degeneration," *Science*, vol. 308, pp. 385–389, Apr 2005.

[251] P. R. Burton, D. G. Clayton, L. R. Cardon, N. Craddock, P. Deloukas, A. Duncanson, D. P. Kwiatkowski, M. I. McCarthy, W. H. Ouwehand, N. J. Samani, J. A. Todd, P. Donnelly, J. C. Barrett, P. R. Burton, D. Davison, P. Donnelly, D. Easton, D. Evans, H. T. Leung, J. L. Marchini, A. P. Morris, C. C. Spencer, M. D. Tobin, L. R. Cardon, D. G. Clayton, A. P. Attwood, J. P. Boorman, B. Cant, U. Everson, J. M. Hussey, J. D.

Jolley, A. S. Knight, K. Koch, E. Meech, S. Nutland, C. V. Prowse, H. E. Stevens, N. C. Taylor, G. R. Walters, N. M. Walker, N. A. Watkins, T. Winzer, J. A. Todd, W. H. Ouwehand, R. W. Jones, W. L. McArdle, S. M. Ring, D. P. Strachan, M. Pembrey, G. Breen, D. St Clair, S. Caesar, K. Gordon-Smith, L. Jones, C. Fraser, E. K. Green, D. Grozeva, M. L. Hamshere, P. A. Holmans, I. R. Jones, G. Kirov, V. Moskvina, I. Nikolov, M. C. O'Donovan, M. J. Owen, N. Craddock, D. A. Collier, A. Elkin, A. Farmer, R. Williamson, P. McGuffin, A. H. Young, I. N. Ferrier, S. G. Ball, A. J. Balmforth, J. H. Barrett, D. T. Bishop, M. M. Iles, A. Maqbool, N. Yuldasheva, A. S. Hall, P. S. Braund, P. R. Burton, R. J. Dixon, M. Mangino, S. Suzanne, M. D. Tobin, J. R. Thompson, N. J. Samani, F. Bredin, M. Tremelling, M. Parkes, H. Drummond, C. W. Lees, E. R. Nimmo, J. Satsangi, S. A. Fisher, A. Forbes, C. M. Lewis, C. M. Onnie, N. J. Prescott, J. Sanderson, C. G. Mathew, J. Barbour, M. K. Mohiuddin, C. E. Todhunter, J. C. Mansfield, T. Ahmad, F. R. Cummings, D. P. Jewell, J. Webster, M. J. Brown, D. G. Clayton, G. M. Lathrop, J. Connell, A. Dominczak, N. J. Samani, C. A. Marcano, B. Burke, R. Dobson, J. Gungadoo, K. L. Lee, P. B. Munroe, S. J. Newhouse, A. Onipinla, C. Wallace, M. Xue, M. Caulfield, M. Farrall, A. Barton, I. N. Bruce, H. Donovan, S. Eyre, P. D. Gilbert, S. L. Hider, A. M. Hinks, S. L. John, C. Potter, A. J. Silman, D. P. Symmmons, W. Thomson, J. Worthington, D. G. Clayton, D. B. Dunger, S. Nutland, H. E. Stevens, N. M. Walker, B. Widmer, J. A. Todd, T. A. Frayling, R. M. Freathy, H. Lango, J. R. Perry, B. M. Shields, M. N. Weedon, A. T. Hattersley, G. A. Hitman, M. Walker, K. S. Elliott, C. J. Groves, C. M. Lindgren, N. W. Rayner, N. J. Timpson, E. Zeggini, M. I. McCarthy, M. Newport, G. Sirugo, E. Lyons, F. Vannberg, A. V. Hill, L. A. Bradbury, C. Farrar, J. J. Pointon, P. Wordsworth, M. A. Brown, J. A. Franklyn, J. M. Heward, M. J. Simmonds, S. C. Gough, S. Seal, M. R. Stratton, N. Rahman, M. Ban, A. Goris, S. J. Sawcer, A. Compston, D. Conway, M. Jallow, M. Newport, G. Sirugo, K. A. Rockett, D. P. Kwiatowski, S. J. Bumpstead, A. Chaney, K. Downes, M. J. Ghori, R. Gwilliam, S. E. Hunt, M. Inouye, A. Keniry, E. King, R. McGinnis, S. Potter, R. Ravindrarajah, P. Whittaker, C. Widden, D. Withers, P. Deloukas, H. T. Leung, S. Nutland, H. E. Stevens, N. M. Walker, J. A. Todd, D. Easton, D. G. Clayton, P. R. Burton, M. D. Tobin, J. C. Barrett, D. Evans, A. P. Morris, L. R. Cardon, N. J. Cardin, D. Davison, T. Ferreira, J. Pereira-Gale, I. B. Hallgrimsdottir, B. N. Howie, J. L. Marchini, C. C. Spencer, Z. Su, Y. Y. Teo, D. Vukcevic, P. Donnelly, D. Bentley, M. A. Brown, L. R. Gordon, M. Caulfield, D. G. Clayton, A. Compston, N. Craddock, P. Deloukas, P. Donnelly, M. Farrall, S. C. Gough, A. S. Hall, A. T. Hattersley, A. V. Hill, D. P. Kwiatkowski, C. Mathew, M. I. McCarthy, W. H. Ouwehand, M. Parkes, M. Pembrey,

N. Rahman, N. J. Samani, M. R. Stratton, J. A. Todd, and J. Worthington, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, vol. 447, pp. 661–678, Jun 2007.

[252] A. R. Wood, J. R. Perry, T. Tanaka, D. G. Hernandez, H. F. Zheng, D. Melzer, J. R. Gibbs, M. A. Nalls, M. N. Weedon, T. D. Spector, J. B. Richards, S. Bandinelli, L. Ferrucci, A. B. Singleton, and T. M. Frayling, "Imputation of variants from the 1000 Genomes Project modestly improves known associations and can identify low-frequency variant-phenotype associations undetected by HapMap based imputation," *PLoS One*, vol. 8, no. 5, p. e64343, 2013.

[253] S. Ripke, B. M. Neale, A. Corvin, J. T. Walters, K. H. Farh, P. A. Holmans, P. Lee, B. Bulik-Sullivan, D. A. Collier, H. Huang, T. H. Pers, I. Agartz, E. Agerbo, M. Albus, M. Alexander, F. Amin, S. A. Bacanu, M. Begemann, R. A. Belliveau, J. Bene, S. E. Bergen, E. Bevilacqua, T. B. Bigdeli, D. W. Black, R. Bruggeman, N. G. Buccola, R. L. Buckner, W. Byerley, W. Cahn, G. Cai, D. Campion, R. M. Cantor, V. J. Carr, N. Carrera, S. V. Catts, K. D. Chambert, R. C. Chan, R. Y. Chen, E. Y. Chen, W. Cheng, E. F. Cheung, S. A. Chong, C. R. Cloninger, D. Cohen, N. Cohen, P. Cormican, N. Craddock, J. J. Crowley, D. Curtis, M. Davidson, K. L. Davis, F. Degenhardt, J. Del Favero, D. Demontis, D. Dikeos, T. Dinan, S. Djurovic, G. Donohoe, E. Drapeau, J. Duan, F. Dudbridge, N. Durmishi, P. Eichhammer, J. Eriksson, V. Escott-Price, L. Essioux, A. H. Fanous, M. S. Farrell, J. Frank, L. Franke, R. Freedman, N. B. Freimer, M. Friedl, J. I. Friedman, M. Fromer, G. Genovese, L. Georgieva, I. Giegling, P. guez, S. Godard, J. I. Goldstein, V. Golimbet, S. Gopal, J. Gratten, L. de Haan, C. Hammer, M. L. Hamshere, M. Hansen, T. Hansen, V. Haroutunian, A. M. Hartmann, F. A. Henskens, S. Herms, J. N. Hirschhorn, P. Hoffmann, A. Hofman, M. V. Hollegaard, D. M. Hougaard, M. Ikeda, I. Joa, A. Ã , R. S. Kahn, L. Kalaydjieva, S. Karachanak-Yankova, J. Karjalainen, D. Kavanagh, M. C. Keller, J. L. Kennedy, A. Khrunin, Y. Kim, J. Klovins, J. A. Knowles, B. Konte, V. Kucinskas, Z. Ausrele Kucinskiene, H. Kuzelova-Ptackova, A. K. hler, C. Laurent, J. L. Keong, S. H. Lee, S. E. Legge, B. Lerer, M. Li, T. Li, K. Y. Liang, J. Lieberman, S. Limborska, C. M. Loughland, J. Lubinski, J. nnqvist, M. Macek, P. K. Magnusson, B. S. Maher, W. Maier, J. Mallet, S. Marsal, M. Mattheisen, M. Mattingsdal, R. W. McCarley, C. McDonald, A. M. McIntosh, S. Meier, C. J. Meijer, B. Melegh, I. Melle, R. I. Mesholam-Gately, A. Metspalu, P. T. Michie, L. Milani, V. Milanova, Y. Mokrab, D. W. Morris, O. Mors, K. C. Murphy, R. M. Murray, I. Myin-Germeys, B. ller Myhsok, M. Nelis, I. Nenadic, D. A. Nertney, G. Nestadt, K. K. Nicodemus, L. Nikitina-Zake, L. Nisen-

baum, A. Nordin, E. O'Callaghan, C. O'Dushlaine, F. A. O'Neill, S. Y. Oh, A. Olincy, L. Olsen, J. Van Os, C. Pantelis, G. N. Papadimitriou, S. Papiol, E. Parkhomenko, M. T. Pato, T. Paunio, M. Pejovic-Milovancevic, D. O. Perkins, O. inen, J. Pimm, A. J. Pocklington, J. Powell, A. Price, A. E. Pulver, S. M. Purcell, D. Quested, H. B. Rasmussen, A. Reichenberg, M. A. Reimers, A. L. Richards, J. L. Roffman, P. Roussos, D. M. Ruderfer, V. Salomaa, A. R. Sanders, U. Schall, C. R. Schubert, T. G. Schulze, S. G. Schwab, E. M. Scolnick, R. J. Scott, L. J. Seidman, J. Shi, E. Sigurdsson, T. Silagadze, J. M. Silverman, K. Sim, P. Slominsky, J. W. Smoller, H. C. So, C. A. Spencer, E. A. Stahl, H. Stefansson, S. Steinberg, E. Stogmann, R. E. Straub, E. Strengman, J. Strohmaier, T. S. Stroup, M. Subramaniam, J. Suvisaari, D. M. Svrakic, J. P. Szatkiewicz, E. derman, S. Thirumalai, D. Toncheva, S. Tosato, J. Veijola, J. Waddington, D. Walsh, D. Wang, Q. Wang, B. T. Webb, M. Weiser, D. B. Wildenauer, N. M. Williams, S. Williams, S. H. Witt, A. R. Wolen, E. H. Wong, B. K. Wormley, H. S. Xi, C. C. Zai, X. Zheng, F. Zimprich, N. R. Wray, K. Stefansson, P. M. Visscher, R. Adolfsson, O. A. Andreassen, D. H. Blackwood, E. Bramon, J. D. Buxbaum, A. D. rglum, S. Cichon, A. Darvasi, E. Domenici, H. Ehrenreich, T. Esko, P. V. Gejman, M. Gill, H. Gurling, C. M. Hultman, N. Iwata, A. V. Jablensky, E. G. nsson, K. S. Kendler, G. Kirov, J. Knight, T. Lencz, D. F. Levinson, Q. S. Li, J. Liu, A. K. Malhotra, S. A. McCarroll, A. McQuillin, J. L. Moran, P. B. Mortensen, B. J. Mowry, M. M. then, R. A. Ophoff, M. J. Owen, A. Palotie, C. N. Pato, T. L. Petryshen, D. Posthuma, M. Rietschel, B. P. Riley, D. Rujescu, P. C. Sham, P. Sklar, D. St Clair, D. R. Weinberger, J. R. Wendland, T. Werge, M. J. Daly, P. F. Sullivan, and M. C. O'Donovan, "Biological insights from 108 schizophrenia-associated genetic loci," *Nature*, vol. 511, pp. 421–427, Jul 2014.

[254] L. Yengo, S. Vedantam, E. Marouli, J. Sidorenko, E. Bartell, S. Sakaue, M. Graff, A. U. Eliasen, Y. Jiang, S. Raghavan, J. Miao, J. D. Arias, S. E. Graham, R. E. Mukamel, C. N. Spracklen, X. Yin, S. H. Chen, T. Ferreira, H. H. Highland, Y. Ji, T. Karaderi, K. Lin, K. ll, D. E. Malden, C. Medina-Gomez, M. Machado, A. Moore, S. eger, X. Sim, S. Vrieze, T. S. Ahluwalia, M. Akiyama, M. A. Allison, M. Alvarez, M. K. Andersen, A. Ani, V. Appadurai, L. Arbeeva, S. Bhaskar, L. F. Bielak, S. Bollepalli, L. L. Bonnycastle, J. Bork-Jensen, J. P. Bradfield, Y. Bradford, P. S. Braund, J. A. Brody, K. S. Burgdorf, B. E. Cade, H. Cai, Q. Cai, A. Campbell, M. adas Garre, E. Catamo, J. F. Chai, X. Chai, L. C. Chang, Y. C. Chang, C. H. Chen, A. Chesi, S. H. Choi, R. H. Chung, M. Cocca, M. P. Concas, C. Couture, G. Cuellar-Partida, R. Danning, E. W. Daw, F. Degenhard, G. E. Delgado, A. Delitala, A. Demirkan, X. Deng, P. Devineni, A. Dietl, M. Dimitriou, L. Dimitrov, R. Dorajoo, A. B. Ekici, J. E. Engmann, Z. Fairhurst-Hunter, A. E. Farmaki,

J. D. Faul, J. C. Fernandez-Lopez, L. Forer, M. Francescatto, S. Freitag-Wolf, C. Fuchsberger, T. E. Galesloot, Y. Gao, Z. Gao, F. Geller, O. Giannakopoulou, F. Giulianini, A. P. Gjesing, A. Goel, S. D. Gordon, M. Gorski, J. Grove, X. Guo, S. Gustafsson, J. Haessler, T. F. Hansen, A. S. Havulinna, S. J. Haworth, J. He, N. Heard-Costa, P. Hebbar, G. Hindy, Y. A. Ho, E. Hofer, E. Holliday, K. Horn, W. E. Hornsby, J. J. Hottenga, H. Huang, J. Huang, A. Huerta-Chagoya, J. E. Huffman, Y. J. Hung, S. Huo, M. Y. Hwang, H. Iha, D. D. Ikeda, M. Isono, A. U. Jackson, S. ger, I. E. Jansen, I. Johansson, J. B. Jonas, A. Jonsson, T. rgensen, I. P. Kalafati, M. Kanai, S. Kanoni, L. L. rhus, A. Kasturiratne, T. Katsuya, T. Kawaguchi, R. L. Kember, K. A. Kentistou, H. N. Kim, Y. J. Kim, M. E. Kleber, M. J. Knol, A. Kurbasic, M. Lauzon, P. Le, R. Lea, J. Y. Lee, H. L. Leonard, S. A. Li, X. Li, X. Li, J. Liang, H. Lin, S. Y. Lin, J. Liu, X. Liu, K. S. Lo, J. Long, L. Lores-Motta, J. Luan, V. Lyssenko, L. P. inen, A. Mahajan, V. Mamakou, M. Mangino, A. Manichaikul, J. Marten, M. Mattheisen, L. Mavarani, A. F. McDaid, K. Meidtner, T. L. Melendez, J. M. Mercader, Y. Milaneschi, J. E. Miller, I. Y. Millwood, P. P. Mishra, R. E. Mitchell, L. T. llehave, A. Morgan, S. Mucha, M. Munz, M. Nakatochi, C. P. Nelson, M. Nethander, C. W. Nho, A. A. Nielsen, I. M. Nolte, S. S. Nongmaithem, R. Noordam, I. Ntalla, T. Nutile, A. Pandit, P. Christofidou, K. rna, M. Pauper, E. R. B. Petersen, L. V. Petersen, N. nen, O. ek, A. Poveda, M. H. Preuss, S. Pyarajan, L. M. Raffield, H. Rakugi, J. Ramirez, A. Rasheed, D. Raven, N. W. Rayner, C. Riveros, R. Rohde, D. Ruggiero, S. E. Ruotsalainen, K. A. Ryan, M. Sabater-Lleal, R. Saxena, M. Scholz, A. Sendamarai, B. Shen, J. Shi, J. H. Shin, C. Sidore, C. M. Sitlani, R. C. Slieker, R. A. J. Smit, A. V. Smith, J. A. Smith, L. J. Smyth, L. Southam, V. Steinthorsdottir, L. Sun, F. Takeuchi, D. S. P. Tallapragada, K. D. Taylor, B. O. Tayo, C. Tcheandjieu, N. Terzikhan, P. Tesolin, A. Teumer, E. Theusch, D. J. Thompson, G. Thorleifsson, P. R. H. J. Timmers, S. Trompet, C. Turman, S. Vaccargiu, S. W. van der Laan, P. J. van der Most, J. B. van Klinken, J. van Setten, S. S. Verma, N. Verweij, Y. Veturi, C. A. Wang, C. Wang, L. Wang, Z. Wang, H. R. Warren, W. Bin Wei, A. R. Wickremasinghe, M. Wielscher, K. L. Wiggins, B. S. Winsvold, A. Wong, Y. Wu, M. Wuttke, R. Xia, T. Xie, K. Yamamoto, J. Yang, J. Yao, H. Young, N. A. Yousri, L. Yu, L. Zeng, W. Zhang, X. Zhang, J. H. Zhao, W. Zhao, W. Zhou, M. E. Zimmermann, M. Zoledziewska, L. S. Adair, H. H. H. Adams, C. A. Aguilar-Salinas, F. Al-Mulla, D. K. Arnett, F. W. Asselbergs, B. O. svold, J. Attia, B. Banas, S. Bandinelli, D. A. Bennett, T. Bergler, D. Bharadwaj, G. Biino, H. Bisgaard, E. Boerwinkle, C. A. ger, K. nnelykke, D. I. Boomsma, A. D. rglum, J. B. Borja, C. Bouchard, D. W. Bowden, I. Brandslund, B. Brumpton, J. E. Buring, M. J. Caulfield, J. C. Chambers, G. R. Chandak, S. J. Chanock, N. Chaturvedi,

Y. I. Chen, Z. Chen, C. Y. Cheng, I. E. Christophersen, M. Ciullo, J. W. Cole, F. S. Collins, R. S. Cooper, M. Cruz, F. Cucca, L. A. Cupples, M. J. Cutler, S. M. Damrauer, T. M. Dantoft, G. J. de Borst, L. C. P. G. M. de Groot, P. L. De Jager, D. P. V. de Kleijn, H. Janaka de Silva, G. V. Dedoussis, A. I. den Hollander, S. Du, D. F. Easton, P. J. M. Elders, A. H. Eliassen, P. T. Ellinor, S. hl, J. Erdmann, M. K. Evans, D. Fatkin, B. Feenstra, M. F. Feitosa, L. Ferrucci, I. Ford, M. Fornage, A. Franke, P. W. Franks, B. I. Freedman, P. Gasparini, C. Gieger, G. Girotto, M. E. Goddard, Y. M. Golightly, C. Gonzalez-Villalpando, P. Gordon-Larsen, H. Grallert, S. F. A. Grant, N. Grarup, L. Griffiths, V. Gudnason, C. Haiman, H. Hakonarson, T. Hansen, C. A. Hartman, A. T. Hattersley, C. Hayward, S. R. Heckbert, C. K. Heng, C. Hengstenberg, A. W. Hewitt, H. Hishigaki, C. B. Hoyng, P. L. Huang, W. Huang, S. C. Hunt, K. Hveem, E. nen, W. G. Iacono, S. Ichihara, M. A. Ikram, C. R. Isasi, R. D. Jackson, M. R. Jarvelin, Z. B. Jin, K. H. ckel, P. K. Joshi, P. Jousilahti, J. W. Jukema, M. nen, Y. Kamatani, K. D. Kang, J. Kaprio, S. L. R. Kardia, F. Karpe, N. Kato, F. Kee, T. Kessler, A. V. Khera, C. C. Khor, L. A. L. M. Kiemeney, B. J. Kim, E. K. Kim, H. L. Kim, P. Kirchhof, M. Kivimaki, W. P. Koh, H. A. Koistinen, G. D. Kolovou, J. S. Kooner, C. Kooperberg, A. ttgen, P. Kovacs, A. Kraaijeveld, P. Kraft, R. M. Krauss, M. Kumari, Z. Kutalik, M. Laakso, L. A. Lange, C. Langenberg, L. J. Launer, L. Le Marchand, H. Lee, N. R. Lee, T. ki, H. Li, L. Li, W. Lieb, X. Lin, L. Lind, A. Linneberg, C. T. Liu, J. Liu, M. Loeffler, B. London, S. A. Lubitz, S. J. Lye, D. A. Mackey, R. gi, P. K. E. Magnusson, G. M. Marcus, P. M. Vidal, N. G. Martin, W. rz, F. Matsuda, R. W. McGarrah, M. McGue, A. J. McKnight, S. E. Medland, D. m, A. Metspalu, B. D. Mitchell, P. Mitchell, D. O. Mook-Kanamori, A. D. Morris, L. A. Mucci, P. B. Munroe, M. A. Nalls, S. Nazarian, A. E. Nelson, M. J. Neville, C. Newton-Cheh, C. S. Nielsen, M. M. then, C. Ohlsson, A. J. Oldehinkel, L. Orozco, K. Pahkala, P. Pajukanta, C. N. A. Palmer, E. J. Parra, C. Pattaro, O. Pedersen, C. E. Pennell, B. W. J. H. Penninx, L. Perusse, A. Peters, P. A. Peyser, D. J. Porteous, D. Posthuma, C. Power, P. P. Pramstaller, M. A. Province, Q. Qi, J. Qu, D. J. Rader, O. T. Raitakari, S. Ralhan, L. S. Rallidis, D. C. Rao, S. Redline, D. F. Reilly, A. P. Reiner, S. Y. Rhee, P. M. Ridker, M. Rienstra, S. Ripatti, M. D. Ritchie, D. M. Roden, F. R. Rosendaal, J. I. Rotter, I. Rudan, F. Rutters, C. Sabanayagam, D. Saleheen, V. Salomaa, N. J. Samani, D. K. Sanghera, N. Sattar, B. Schmidt, H. Schmidt, R. Schmidt, M. B. Schulze, H. Schunkert, L. J. Scott, R. J. Scott, P. Sever, E. J. Shiroma, M. B. Shoemaker, X. O. Shu, E. M. Simonsick, M. Sims, J. R. Singh, A. B. Singleton, M. F. Sinner, J. G. Smith, H. Snieder, T. D. Spector, M. J. Stampfer, K. J. Stark, D. P. Strachan, L. M. 't Hart, Y. Tabara, H. Tang, J. C. Tardif, T. A. Thanaraj, N. J. Timpson, A. njes,

A. Tremblay, T. Tuomi, J. Tuomilehto, M. T. Luna, A. G. Uitterlinden, R. M. van Dam, P. van der Harst, N. Van der Velde, C. M. van Duijn, N. M. van Schoor, V. Vitart, U. lker, P. Vollenweider, H. lzke, N. H. Wacher-Rodarte, M. Walker, Y. X. Wang, N. J. Wareham, R. M. Watanabe, H. Watkins, D. R. Weir, T. M. Werge, E. Widen, L. R. Wilkens, G. Willemsen, W. C. Willett, J. F. Wilson, T. Y. Wong, J. T. Woo, A. F. Wright, J. Y. Wu, H. Xu, C. S. Yajnik, M. Yokota, J. M. Yuan, E. Zeggini, B. S. Zemel, W. Zheng, X. Zhu, J. M. Zmuda, A. B. Zonderman, J. A. Zwart, D. I. Chasman, Y. S. Cho, I. M. Heid, M. I. McCarthy, M. C. Y. Ng, C. J. O'Donnell, F. Rivadeneira, U. Thorsteinsdottir, Y. V. Sun, E. S. Tai, M. Boehnke, P. Deloukas, A. E. Justice, C. M. Lindgren, R. J. F. Loos, K. L. Mohlke, K. E. North, K. Stefansson, R. G. Walters, T. W. Winkler, K. L. Young, P. R. Loh, J. Yang, T. Esko, T. L. Assimes, A. Auton, G. R. Abecasis, C. J. Willer, A. E. Locke, S. I. Berndt, G. Lettre, T. M. Frayling, Y. Okada, A. R. Wood, P. M. Visscher, J. N. Hirschhorn, G. C. Partida, Y. Sun, D. Croteau-Chonka, J. M. Vonk, S. Chanock, and L. Le Marchand, "A saturated map of common genetic variants associated with human height," *Nature*, vol. 610, pp. 704–712, Oct 2022.

[255] A. B. Popejoy and S. M. Fullerton, "Genomics is failing on diversity," *Nature*, vol. 538, pp. 161–164, Oct 2016.

[256] J. H. Barrett, J. C. Taylor, and M. M. Iles, "Statistical perspectives for genome-wide association studies (GWAS)," *Methods Mol Biol*, vol. 1168, pp. 47–61, 2014.

[257] A. Dehghan, "Genome-Wide Association Studies," *Methods Mol Biol*, vol. 1793, pp. 37–49, 2018.

[258] S. Pavan, C. Delvento, L. Ricciardi, C. Lotti, E. Ciani, and N. D'Agostino, "Recommendations for Choosing the Genotyping Method and Best Practices for Quality Control in Crop Genome-Wide Association Studies," *Front Genet*, vol. 11, p. 447, 2020.

[259] A. T. Marees, H. de Kluiver, S. Stringer, F. Vorspan, E. Curis, C. Marie-Claire, and E. M. Derks, "A tutorial on conducting genome-wide association studies: Quality control and statistical analysis," *Int J Methods Psychiatr Res*, vol. 27, p. e1608, Jun 2018.

[260] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham, "PLINK: a tool set for whole-genome association and population-based linkage analyses," *Am J Hum Genet*, vol. 81, pp. 559–575, Sep 2007.

[261] P. R. Loh, G. Tucker, B. K. Bulik-Sullivan, B. J. lmsson, H. K. Finucane, R. M. Salem, D. I. Chasman, P. M. Ridker, B. M. Neale, B. Berger, N. Patterson, and A. L. Price, "Efficient Bayesian mixed-model analysis increases association power in large cohorts," *Nat Genet*, vol. 47, pp. 284–290, Mar 2015.

[262] L. Jiang, Z. Zheng, H. Fang, and J. Yang, "A generalized linear mixed model association tool for biobank-scale data," *Nat Genet*, vol. 53, pp. 1616–1621, Nov 2021.

[263] D. O. Enoma, J. Bishung, T. Abiodun, O. Ogunlana, and V. C. Osamor, "Machine learning approaches to genome-wide association studies," *Journal of King Saud University-Science*, p. 101847, 2022.

[264] A. S. Kaler and L. C. Purcell, "Estimation of a significance threshold for genome-wide association studies," *BMC Genomics*, vol. 20, p. 618, Jul 2019.

[265] K. Watanabe, E. Taskesen, A. van Bochoven, and D. Posthuma, "Functional mapping and annotation of genetic associations with FUMA," *Nat Commun*, vol. 8, p. 1826, Nov 2017.

[266] Y. F. Pei, Q. Tian, L. Zhang, and H. W. Deng, "Exploring the Major Sources and Extent of Heterogeneity in a Genome-Wide Association Meta-Analysis," *Ann Hum Genet*, vol. 80, pp. 113–122, Mar 2016.

[267] C. G. Crossner, J. Carlsson, B. din, A. rnvik, L. Unell, P. Venge, and L. Wranne, "Periodontitis in the primary dentition associated with Actinobacillus actinomycetemcomitans infection and leukocyte dysfunction. A 3 1/2 year follow-up," *J Clin Periodontol*, vol. 17, pp. 264–267, Apr 1990.

[268] G. R. Svishcheva, T. I. Axenovich, N. M. Belonogova, C. M. van Duijn, and Y. S. Aulchenko, "Rapid variance components-based method for whole-genome association analysis," *Nat Genet*, vol. 44, pp. 1166–1170, Oct 2012.

[269] J. Jakobsdottir and M. S. McPeek, "MASTOR: mixed-model association mapping of quantitative traits in samples with related individuals," *Am J Hum Genet*, vol. 92, pp. 652–666, May 2013.

[270] H. M. Kang, N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin, "Efficient control of population structure in model organism association mapping," *Genetics*, vol. 178, no. 3, pp. 1709–1723, 2008.

[271] Z. Zhang, E. Ersoz, C. Q. Lai, R. J. Todhunter, H. K. Tiwari, M. A. Gore, P. J. Bradbury, J. Yu, D. K. Arnett, J. M. Ordovas, and E. S. Buckler, "Mixed linear model approach adapted for genome-wide association studies," *Nat Genet*, vol. 42, pp. 355–360, Apr 2010.

[272] C. Lippert, J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson, and D. Heckerman, "FaST linear mixed models for genome-wide association studies," *Nat Methods*, vol. 8, pp. 833–835, Sep 2011.

[273] X. Zhou and M. Stephens, "Genome-wide efficient mixed-model analysis for association studies," *Nat Genet*, vol. 44, pp. 821–824, Jun 2012.

[274] J. Eu-Ahsunthornwattana, E. N. Miller, M. Fakiola, S. M. Jeronimo, J. M. Blackwell, and H. J. Cordell, "Comparison of methods to account for relatedness in genome-wide association studies with family-based data," *PLoS Genet*, vol. 10, p. e1004445, Jul 2014.

[275] C. Widmer, C. Lippert, O. Weissbrod, N. Fusi, C. Kadie, R. Davidson, J. Listgarten, and D. Heckerman, "Further improvements to linear mixed models for genome-wide association studies," *Sci Rep*, vol. 4, p. 6874, Nov 2014.

[276] A. L. Price, N. A. Zaitlen, D. Reich, and N. Patterson, "New approaches to population stratification in genome-wide association studies," *Nat Rev Genet*, vol. 11, pp. 459–463, Jul 2010.

[277] J. Listgarten, C. Lippert, C. M. Kadie, R. I. Davidson, E. Eskin, and D. Heckerman, "Improved linear mixed models for genome-wide association studies," *Nat Methods*, vol. 9, pp. 525–526, May 2012.

[278] D. Bennett, D. O'Shea, J. Ferguson, D. Morris, and C. Seoighe, "Controlling for background genetic effects using polygenic scores improves the power of genome-wide association studies," *Sci Rep*, vol. 11, p. 19571, Oct 2021.

[279] N. R. Wray, K. E. Kemper, B. J. Hayes, M. E. Goddard, and P. M. Visscher, "Complex Trait Prediction from Genome Data: Contrasting EBV in Livestock to PRS in Humans: Genomic Prediction," *Genetics*, vol. 211, pp. 1131–1141, Apr 2019.

[280] T. H. Meuwissen, B. J. Hayes, and M. E. Goddard, "Prediction of total genetic value using genome-wide dense marker maps," *Genetics*, vol. 157, pp. 1819–1829, Apr 2001.

[281] T. H. Meuwissen, B. J. Hayes, and M. Goddard, "Prediction of total genetic value using genome-wide dense marker maps," *genetics*, vol. 157, no. 4, pp. 1819–1829, 2001.

[282] A. C. Fahed, A. A. Philippakis, and A. V. Khera, “The potential of polygenic scores to improve cost and efficiency of clinical trials,” *Nat Commun*, vol. 13, p. 2922, May 2022.

[283] M. Dehestani, H. Liu, and T. Gasser, “Polygenic Risk Scores Contribute to Personalized Medicine of Parkinson’s Disease,” *J Pers Med*, vol. 11, Oct 2021.

[284] B. Cross, R. Turner, and M. Pirmohamed, “Polygenic risk scores: An overview from bench to bedside for personalised medicine,” *Front Genet*, vol. 13, p. 1000667, 2022.

[285] J. R. Ashenhurst, O. V. Sazonova, O. Svrchek, S. Detweiler, R. Kita, L. Babalola, M. McIntyre, S. Aslibekyan, P. Fontanillas, S. Shringarpure, J. D. Pollard, and B. L. Koelsch, “A Polygenic Score for Type 2 Diabetes Improves Risk Stratification Beyond Current Clinical Screening Factors in an Ancestrally Diverse Sample,” *Front Genet*, vol. 13, p. 871260, 2022.

[286] S. Mistry, J. R. Harrison, D. J. Smith, V. Escott-Price, and S. Zammit, “The use of polygenic risk scores to identify phenotypes associated with genetic risk of schizophrenia: Systematic review,” *Schizophr Res*, vol. 197, pp. 2–8, Jul 2018.

[287] B. J. Coombes, M. Markota, J. J. Mann, C. Colby, E. Stahl, A. Talati, J. Pathak, M. M. Weissman, S. L. McElroy, M. A. Frye, and J. M. Biernacka, “Dissecting clinical heterogeneity of bipolar disorder using multiple polygenic risk scores,” *Transl Psychiatry*, vol. 10, p. 314, Sep 2020.

[288] E. Agerbo, B. B. Trabjerg, A. D. rglum, A. J. Schork, B. J. lmsson, C. B. Pedersen, C. Hakulinen, C. ana, D. M. Hougaard, J. Grove, J. J. McGrath, J. Bybjerg-Grauholm, O. Mors, O. Plana-Ripoll, T. Werge, N. R. Wray, P. B. Mortensen, and K. L. Musliner, “Risk of Early-Onset Depression Associated With Polygenic Liability, Parental Psychiatric History, and Socioeconomic Status,” *JAMA Psychiatry*, vol. 78, pp. 387–397, Apr 2021.

[289] J. Euesden, C. M. Lewis, and P. F. O’Reilly, “PRSice: Polygenic Risk Score software,” *Bioinformatics*, vol. 31, pp. 1466–1468, May 2015.

[290] B. J. lmsson, J. Yang, H. K. Finucane, A. Gusev, S. m, S. Ripke, G. Genovese, P. R. Loh, G. Bhatia, R. Do, T. Hayeck, H. H. Won, S. Kathiresan, M. Pato, C. Pato, R. Tamimi, E. Stahl, N. Zaitlen, B. Pasaniuc, G. Belbin, E. E. Kenny, M. H. Schierup, P. De Jager, N. A. Patsopoulos, S. McCarroll, M. Daly, S. Purcell, D. Chasman, B. Neale, M. Goddard, P. M. Visscher, P. Kraft, N. Patterson, A. L. Price, S. Ripke, B. M. Neale, A. Corvin,

J. T. Walters, K. H. Farh, P. A. Holmans, P. Lee, B. Bulik-Sullivan, D. A. Collier, H. Huang, T. H. Pers, I. Agartz, E. Agerbo, M. Albus, M. Alexander, F. Amin, S. A. Bacanu, M. Begemann, R. A. Belliveau, J. Bene, S. E. Bergen, E. Bevilacqua, T. B. Bigdeli, D. W. Black, R. Bruggeman, N. G. Buccola, R. L. Buckner, W. Byerley, W. Cahn, G. Cai, D. Campion, R. M. Cantor, V. J. Carr, N. Carrera, S. V. Catts, K. D. Chambert, R. C. Chan, R. Y. Chen, E. Y. Chen, W. Cheng, E. F. Cheung, S. A. Chong, C. R. Cloninger, D. Cohen, N. Cohen, P. Cormican, N. Craddock, J. J. Crowley, D. Curtis, M. Davidson, K. L. Davis, F. Degenhardt, J. Del Favero, L. E. DeLisi, D. Demontis, D. Dikeos, T. Dinan, S. Djurovic, G. Donohoe, E. Drapeau, J. Duan, F. Dudbridge, N. Durmishi, P. Eichhammer, J. Eriksson, V. Escott-Price, L. Essioux, A. H. Fanous, M. S. Farrell, J. Frank, L. Franke, R. Freedman, N. B. Freimer, M. Friedl, J. I. Friedman, M. Fromer, G. Genovese, L. Georgieva, E. S. Gershon, I. Giegling, P. Giusti-Rodrguez, S. Godard, J. I. Goldstein, V. Golimbet, S. Gopal, J. Gratten, J. Grove, L. de Haan, C. Hammer, M. L. Hamshere, M. Hansen, T. Hansen, V. Haroutunian, A. M. Hartmann, F. A. Henskens, S. Herms, J. N. Hirschhorn, P. Hoffmann, A. Hofman, M. V. Hollegaard, D. M. Hougaard, M. Ikeda, I. Joa, A. Julia, R. S. Kahn, L. Kalaydjieva, S. Karachanak-Yankova, J. Karjalainen, D. Kavanagh, M. C. Keller, B. J. Kelly, J. L. Kennedy, A. Khrunin, Y. Kim, J. Klovins, J. A. Knowles, B. Konte, V. Kucinskas, Z. A. Kucinskiene, H. Kuzelova-Ptackova, A. K. Kahler, C. Laurent, J. L. Keong, S. H. Lee, S. E. Legge, B. Lerer, M. Li, T. Li, K. Y. Liang, J. Lieberman, S. Limborska, C. M. Loughland, J. Lubinski, J. Lnnqvist, M. Macek, P. K. Magnusson, B. S. Maher, W. Maier, J. Mallet, S. Marsal, M. Mattheisen, M. Mattingsdal, R. W. McCarley, C. McDonald, A. M. McIntosh, S. Meier, C. J. Meijer, B. Melegh, I. Melle, R. I. Mesholam-Gately, A. Metspalu, P. T. Michie, L. Milani, V. Milanova, Y. Mokrab, D. W. Morris, O. Mors, P. B. Mortensen, K. C. Murphy, R. M. Murray, I. Myin-Germeys, B. Mller-Myhsok, M. Nelis, I. Nenadic, D. A. Nertney, G. Nestadt, K. K. Nicodemus, L. Nikitina-Zake, L. Nisenbaum, A. Nordin, E. O'Callaghan, C. O'Dushlaine, F. A. O'Neill, S. Y. Oh, A. Olincy, L. Olsen, J. Van Os, C. Pantelis, G. N. Papadimitriou, S. Papiol, E. Parkhomenko, M. T. Pato, T. Paunio, M. Pejovic-Milovancevic, D. O. Perkins, O. Pietilinen, J. Pimm, A. J. Pocklington, J. Powell, A. Price, A. E. Pulver, S. M. Purcell, D. Quested, H. B. Rasmussen, A. Reichenberg, M. A. Reimers, A. L. Richards, J. L. Roffman, P. Roussos, D. M. Ruderfer, V. Salomaa, A. R. Sanders, U. Schall, C. R. Schubert, T. G. Schulze, S. G. Schwab, E. M. Scolnick, R. J. Scott, L. J. Seidman, J. Shi, E. Sigurdsson, T. Silagadze, J. M. Silverman, K. Sim, P. Slominsky, J. W. Smoller, H. C. So, C. C. Spencer, E. A. Stahl, H. Stefansson, S. Steinberg, E. Stogmann, R. E. Straub,

E. Strengman, J. Strohmaier, T. S. Stroup, M. Subramaniam, J. Suvisaari, D. M. Svrakic, J. P. Szatkiewicz, E. Sderman, S. Thirumalai, D. Toncheva, P. A. Tooney, S. Tosato, J. Veijola, J. Waddington, D. Walsh, D. Wang, Q. Wang, B. T. Webb, M. Weiser, D. B. Wildenauer, N. M. Williams, S. Williams, S. H. Witt, A. R. Wolen, E. H. Wong, B. K. Wormley, J. Q. Wu, H. S. Xi, C. C. Zai, X. Zheng, F. Zimprich, N. R. Wray, K. Stefansson, P. M. Visscher, R. Adolfsson, O. A. Andreassen, D. H. Blackwood, E. Bramon, J. D. Buxbaum, A. D. rglum, S. Cichon, A. Darvasi, E. Domenici, H. Ehrenreich, T. Esko, P. V. Gejman, M. Gill, H. Gurling, C. M. Hultman, N. Iwata, A. V. Jablensky, E. G. Jonsson, K. S. Kendler, G. Kirov, J. Knight, T. Lencz, D. F. Levinson, Q. S. Li, J. Liu, A. K. Malhotra, S. A. McCarroll, A. McQuillin, J. L. Moran, P. B. Mortensen, B. J. Mowry, M. M. Nthen, R. A. Ophoff, M. J. Owen, A. Palotie, C. N. Pato, T. L. Petryshen, D. Posthuma, M. Rietschel, B. P. Riley, D. Rujescu, P. C. Sham, P. Sklar, D. St Clair, D. R. Weinberger, J. R. Wendland, T. Werge, M. J. Daly, P. F. Sullivan, M. C. O'Donovan, P. Kraft, D. J. Hunter, M. Adank, H. Ahsan, K. ki, L. Baglietto, S. Berndt, C. Blomquist, F. Canzian, J. Chang-Claude, S. J. Chanock, L. Crisponi, K. Czene, N. Dahmen, I. d. o. s. S. Silva, D. Easton, A. H. Eliassen, J. Figueroa, O. Fletcher, M. Garcia-Closas, M. M. Gaudet, L. Gibson, C. A. Haiman, P. Hall, A. Hazra, R. Hein, B. E. Henderson, A. Hofman, J. L. Hopper, A. Irwanto, M. Johansson, R. Kaaks, M. G. Kibriya, P. Lichtner, S. m, J. Liu, E. Lund, E. Makalic, A. Meindl, H. Meijers-Heijboer, B. ller Myhsok, T. A. Muranen, H. Nevanlinna, P. H. Peeters, J. Peto, R. L. Prentice, N. Rahman, M. J. nchez, D. F. Schmidt, R. K. Schmutzler, M. C. Southey, R. Tamimi, R. Travis, C. Turnbull, A. G. Uitterlinden, R. B. van der Luijt, Q. Waisfisz, Z. Wang, A. S. Whittemore, R. Yang, and W. Zheng, "Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores," *Am J Hum Genet*, vol. 97, pp. 576–592, Oct 2015.

[291] F. Ã©, J. Arbel, and B. J. lmsson, "LDpred2: better, faster, stronger," *Bioinformatics*, vol. 36, pp. 5424–5431, Apr 2021.

[292] L. R. Lloyd-Jones, J. Zeng, J. Sidorenko, L. Yengo, G. Moser, K. E. Kemper, H. Wang, Z. Zheng, R. Magi, T. Esko, A. Metspalu, N. R. Wray, M. E. Goddard, J. Yang, and P. M. Visscher, "Improved polygenic prediction by Bayesian multiple regression on summary statistics," *Nat Commun*, vol. 10, p. 5086, Nov 2019.

[293] T. Wilkinson, C. Schnier, K. Bush, K. e, D. E. Henshall, C. Lerpiniere, N. E. Allen, R. Flaig, T. C. Russ, D. Bathgate, S. Pal, J. T. O'Brien, and C. L. M. Sudlow, "Identifying dementia outcomes in UK Biobank: a validation study of primary care, hospital admissions and mortality data," *Eur J Epidemiol*, vol. 34, pp. 557–565, Jun 2019.

[294] J. Elliott, B. Bodinier, M. Whitaker, C. Delpierre, R. Vermeulen, I. Tzoulaki, P. Elliott, and M. Chadeau-Hyam, "COVID-19 mortality in the UK Biobank cohort: revisiting and evaluating risk factors," *Eur J Epidemiol*, vol. 36, pp. 299–309, Mar 2021.

[295] A. Mullard, "The UK Biobank at 20," *Nat Rev Drug Discov*, vol. 21, pp. 628–629, Sep 2022.

[296] U. Biobank, "Press release," https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/news/regeneron-announces-major-collaboration-to-exome-sequence-uk-biobank-genetic-data-more-quickly.

[297] S. Inst, "Press release," https://www.sanger.ac.uk/collaboration/uk-biobank-whole-genome-sequencing-project/.

[298] R. Beelen, O. Raaschou-Nielsen, M. Stafoggia, Z. J. Andersen, G. Weinmayr, B. Hoffmann, K. Wolf, E. Samoli, P. Fischer, M. Nieuwenhuijsen, *et al.*, "Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 european cohorts within the multicentre escape project," *The lancet*, vol. 383, no. 9919, pp. 785–795, 2014.

[299] I. P. Gorlov, O. Y. Gorlova, M. L. Frazier, M. R. Spitz, and C. I. Amos, "Evolutionary evidence of the effect of rare variants on disease etiology," *Clin Genet*, vol. 79, pp. 199–206, Mar 2011.

[300] N. authors listed, "STAARpipeline: an all-in-one rare-variant tool for biobank-scale whole-genome sequencing data," *Nat Methods*, vol. 19, pp. 1532–1533, Dec 2022.

[301] Q. Wang, R. S. Dhindsa, K. Carss, A. R. Harper, A. Nag, I. Tachmazidou, D. Vitsios, S. V. V. Deevi, A. Mackay, D. Muthas, M. hn, S. Monkley, H. Olsson, S. Wasilewski, K. R. Smith, R. March, A. Platt, C. Haefliger, S. Petrovski, B. R. Angermann, R. Artzi, C. Barrett, M. Belvisi, M. Bohlooly-Y, O. Burren, L. Buvall, B. Challis, S. Cameron-Christie, S. Cohen, A. Davis, R. F. Danielson, B. Dougherty, B. Georgi, Z. Ghazoui, P. B. L. Hansen, F. Hu, M. Jeznach, X. Jiang, C. Kumar, Z. Lai, G. Lassi, S. H. Lewis, B. Linghu, K. Lythgow, P. Maccallum, C. Martins, A. Matakidou, E. lsson, S. Moosmang, S. O'Dell, Y. Ohne, J. Okae, A. O'Neill, D. S. Paul, A. Reznichenko, M. A. Snowden, A. Walentinsson, J. Zeron, and M. N. Pangalos, "Rare variant contribution to human disease in 281,104 UK Biobank exomes," *Nature*, vol. 597, pp. 527–532, Sep 2021.

[302] P. Dornbos, R. Koesterer, A. Ruttenburg, T. Nguyen, J. B. Cole, A. Leong, J. B. Meigs, J. C. Florez, J. I. Rotter, M. S. Udler, and J. Flannick, "A combined polygenic score of

21,293 rare and 22 common variants improves diabetes diagnosis based on hemoglobin A1C levels," *Nat Genet*, vol. 54, pp. 1609–1614, Nov 2022.

[303] S. J. Jurgens, S. H. Choi, V. N. Morrill, M. Chaffin, J. P. Pirruccello, J. L. Halford, L. C. Weng, V. Nauffal, C. Roselli, A. W. Hall, M. T. Oetjens, B. Lagerman, D. P. vanMaanen, K. G. Aragam, K. L. Lunetta, C. M. Haggerty, S. A. Lubitz, P. T. Ellinor, G. Abecasis, X. Bai, S. Balasubramanian, A. Baras, C. Beechert, B. Boutkov, M. Cantor, G. Coppola, T. De, A. Deubler, A. Economides, G. Eom, M. A. R. Ferreira, C. Forsythe, E. D. Fuller, Z. Gu, L. Habegger, A. Hawes, M. B. Jones, K. Karalis, S. Khalid, O. Krasheninina, R. Lanche, M. Lattari, D. Li, A. Lopez, L. A. Lotta, K. Manoochehri, A. J. Mansfield, E. K. Maxwell, J. Mighty, L. J. Mitnaul, M. Nafde, J. Nielsen, S. O'Keeffe, M. Orelus, J. D. Overton, M. S. Padilla, R. Panea, T. Polanco, M. Pradhan, A. Rasool, J. G. Reid, W. Salerno, T. D. Schleicher, A. Shuldiner, K. Siminovitch, J. C. Staples, R. H. Ulloa, N. Verweij, L. Widom, and S. E. Wolf, "Analysis of rare genetic variation underlying cardiometabolic diseases and traits among 200,000 individuals in the UK Biobank," *Nat Genet*, vol. 54, pp. 240–250, Mar 2022.

[304] X. Wang, E. Lim, C. T. Liu, Y. J. Sung, D. C. Rao, A. C. Morrison, E. Boerwinkle, A. K. Manning, and H. Chen, "Efficient gene-environment interaction tests for large biobank-scale sequencing studies," *Genet Epidemiol*, vol. 44, pp. 908–923, Nov 2020.

[305] S. P. Kar, P. M. Quiros, M. Gu, T. Jiang, J. Mitchell, R. Langdon, V. Iyer, C. Barcena, M. S. Vijayabaskar, M. A. Fabre, P. Carter, S. Petrovski, S. Burgess, and G. S. Vassiliou, "Genome-wide analyses of 200,453 individuals yield new insights into the causes and consequences of clonal hematopoiesis," *Nat Genet*, vol. 54, pp. 1155–1166, Aug 2022.

[306] M. D. Kessler, A. Damask, S. O'Keeffe, M. V. Meter, N. Banerjee, S. Semrau, D. Li, K. Watanabe, J. Horowitz, Y. Houvras, C. Gillies, J. Mbatchou, R. R. White, J. A. Kosmicki, M. G. LeBlanc, M. Jones, R. G. Center, G.-R. D. Collaboration, D. J. Glass, L. A. Lotta, M. N. Cantor, G. S. Atwal, A. E. Locke, M. A. R. Ferreira, R. Deering, C. Paulding, A. R. Shuldiner, G. Thurston, W. Salerno, J. G. Reid, J. D. Overton, J. Marchini, H. M. Kang, A. Baras, G. R. Abecasis, and E. Jorgenson, "Exome sequencing of 628,388 individuals identifies common and rare variant associations with clonal hematopoiesis phenotypes," *medRxiv*, 2022.

[307] L. Chen, P. Liu, T. C. Evans, and L. M. Ettwiller, "DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification," *Science*, vol. 355, pp. 752–756, Feb 2017.

[308] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, “The Sequence Alignment/Map format and SAMtools,” *Bioinformatics*, vol. 25, pp. 2078–2079, Aug 2009.

[309] F. Abascal, L. M. R. Harvey, E. Mitchell, A. R. J. Lawson, S. V. Lensing, P. Ellis, A. J. C. Russell, R. E. Alcantara, A. Baez-Ortega, Y. Wang, E. J. Kwa, H. Lee-Six, A. Cagan, T. H. H. Coorens, M. S. Chapman, S. Olafsson, S. Leonard, D. Jones, H. E. Machado, M. Davies, N. F. Øbro, K. T. Mahubani, K. Allinson, M. Gerstung, K. Saeb-Parsy, D. G. Kent, E. Laurenti, M. R. Stratton, R. Rahbari, P. J. Campbell, R. J. Osborne, and I. Martincorena, “Somatic mutation landscapes at single-molecule resolution,” *Nature*, vol. 593, pp. 405–410, May 2021. Number: 7859 Publisher: Nature Publishing Group.

[310] M. Rodriguez-Galindo, S. Casillas, D. Weghorn, and A. Barbadilla, “Germline de novo mutation rates on exons versus introns in humans,” *Nat Commun*, vol. 11, p. 3304, Jul 2020.

[311] J. Frigola, R. Sabarinathan, L. Mularoni, F. os, A. Gonzalez-Perez, and N. pez Bigas, “Reduced mutation rate in exons due to differential mismatch repair,” *Nat Genet*, vol. 49, pp. 1684–1692, Dec 2017.

[312] A. P. Patel, M. Wang, A. C. Fahed, H. Mason-Suares, D. Brockman, R. Pelletier, S. Amr, K. Machini, M. Hawley, L. Witkowski, C. Koch, A. Philippakis, C. A. Cassa, P. T. Ellinor, S. Kathiresan, K. Ng, M. Lebo, and A. V. Khera, “Association of Rare Pathogenic DNA Variants for Familial Hypercholesterolemia, Hereditary Breast and Ovarian Cancer Syndrome, and Lynch Syndrome With Disease Risk in Adults According to Family History,” *JAMA Netw Open*, vol. 3, p. e203959, Apr 2020.

[313] A. M. D’Gama and C. A. Walsh, “Somatic mosaicism and neurodevelopmental disease,” *Nat Neurosci*, vol. 21, pp. 1504–1514, Nov 2018.

[314] D. J. Burgess, “A body-wide view of somatic mutations,” *Nat Rev Genet*, vol. 22, p. 689, Nov 2021.

[315] J. Hu, S. Adar, C. P. Selby, J. D. Lieb, and A. Sancar, “Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution,” *Genes Dev*, vol. 29, pp. 948–960, May 2015.

[316] T. Lindahl and D. E. Barnes, “Repair of endogenous DNA damage,” *Cold Spring Harb Symp Quant Biol*, vol. 65, pp. 127–133, 2000.

[317] P. Ren, X. Dong, and J. Vijg, “Age-related somatic mutation burden in human tissues,” *Front Aging*, vol. 3, p. 1018119, 2022.

[318] Z. R. Chalmers, C. F. Connelly, D. Fabrizio, L. Gay, S. M. Ali, R. Ennis, A. Schrock, B. Campbell, A. Shlien, J. Chmielecki, F. Huang, Y. He, J. Sun, U. Tabori, M. Kennedy, D. S. Lieber, S. Roels, J. White, G. A. Otto, J. S. Ross, L. Garraway, V. A. Miller, P. J. Stephens, and G. M. Frampton, “Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden,” *Genome Med*, vol. 9, p. 34, Apr 2017.

[319] B. Werner and A. Sottoriva, “Variation of mutational burden in healthy human tissues suggests non-random strand segregation and allows measuring somatic mutation rates,” *PLoS Comput Biol*, vol. 14, p. e1006233, Jun 2018.

[320] J. D. Backman, A. H. Li, A. Marcketta, D. Sun, J. Mbatchou, M. D. Kessler, C. Benner, D. Liu, A. E. Locke, S. Balasubramanian, A. Yadav, N. Banerjee, C. E. Gillies, A. Damask, S. Liu, X. Bai, A. Hawes, E. Maxwell, L. Gurski, K. Watanabe, J. A. Kosmicki, V. Rajagopal, J. Mighty, M. Jones, L. Mitnaul, E. Stahl, G. Coppola, E. Jorgenson, L. Habegger, W. J. Salerno, A. R. Shuldiner, L. A. Lotta, J. D. Overton, M. N. Cantor, J. G. Reid, G. Yancopoulos, H. M. Kang, J. Marchini, A. Baras, G. R. Abecasis, and M. A. R. Ferreira, “Exome sequencing and analysis of 454,787 UK Biobank participants,” *Nature*, vol. 599, pp. 628–634, Nov 2021.

[321] A. Carducci, R. Vannucchi, M. Guidi, D. Reali, and M. A. Ruschi, “Human rotavirus detection in stool specimens using enzyme-linked immunosorbent assays and latex agglutination test,” *Boll Ist Sieroter Milan*, vol. 67, no. 3, pp. 241–244, 1988.

[322] M. L. Hoang, I. Kinde, C. Tomasetti, K. W. McMahon, T. A. Rosenquist, A. P. Grollman, K. W. Kinzler, B. Vogelstein, and N. Papadopoulos, “Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing,” *Proc Natl Acad Sci U S A*, vol. 113, pp. 9846–9851, Aug 2016.

[323] J. R. Huyghe, S. A. Bien, T. A. Harrison, H. M. Kang, S. Chen, S. L. Schmit, D. V. Conti, C. Qu, J. Jeon, C. K. Edlund, P. Greenside, M. Wainberg, F. R. Schumacher, J. D. Smith, D. M. Levine, S. C. Nelson, N. A. Sinnott-Armstrong, D. Albanes, M. H. Alonso, K. Anderson, C. Arnau-Collell, V. Arndt, C. Bamia, B. L. Banbury, J. A. Baron, S. I. Berndt, S. zieau, D. T. Bishop, J. Boehm, H. Boeing, H. Brenner, S. Brezina, S. Buch, D. D. Buchanan, A. Burnett-Hartman, K. Butterbach, B. J. Caan, P. T. Campbell, C. S. Carlson, S. Bel, A. T. Chan, J. Chang-Claude, S. J. Chanock, M. D. Chirlaque, S. H.

Cho, C. M. Connolly, A. J. Cross, K. Cuk, K. R. Curtis, A. de la Chapelle, K. F. Doheny, D. Duggan, D. F. Easton, S. G. Elias, F. Elliott, D. R. English, E. J. M. Feskens, J. C. Figueiredo, R. Fischer, L. M. FitzGerald, D. Forman, M. Gala, S. Gallinger, W. J. Gauderman, G. G. Giles, E. Gillanders, J. Gong, P. J. Goodman, W. M. Grady, J. S. Grove, A. Gsur, M. J. Gunter, R. W. Haile, J. Hampe, H. Hampel, S. Harlid, R. B. Hayes, P. Hofer, M. Hoffmeister, J. L. Hopper, W. L. Hsu, W. Y. Huang, T. J. Hudson, D. J. Hunter, G. ez Sanz, G. E. Idos, R. Ingersoll, R. D. Jackson, E. J. Jacobs, M. A. Jenkins, A. D. Joshi, C. E. Joshu, T. O. Keku, T. J. Key, H. R. Kim, E. Kobayashi, L. N. Kolonel, C. Kooperberg, T. hn, S. ry, S. S. Kweon, S. C. Larsson, C. A. Laurie, L. Le Marchand, S. M. Leal, S. C. Lee, F. Lejbkowicz, M. Lemire, C. I. Li, L. Li, W. Lieb, Y. Lin, A. Lindblom, N. M. Lindor, H. Ling, T. L. Louie, S. Ã¶, S. D. Markowitz, V. n, G. Masala, C. E. McNeil, M. Melas, R. L. Milne, L. Moreno, N. Murphy, R. Myte, A. Naccarati, P. A. Newcomb, K. Offit, S. Ogino, N. C. Onland-Moret, B. Pardini, P. S. Parfrey, R. Pearlman, V. Perduca, P. D. P. Pharoah, M. Pinchev, E. A. Platz, R. L. Prentice, E. Pugh, L. Raskin, G. Rennert, H. S. Rennert, E. Riboli, M. guez Barranco, J. Romm, L. C. Sakoda, C. Schafmayer, R. E. Schoen, D. Seminara, M. Shah, T. Shelford, M. H. Shin, K. Shulman, S. Sieri, M. L. Slattery, M. C. Southey, Z. K. Stadler, C. Stegmaier, Y. R. Su, C. M. Tangen, S. N. Thibodeau, D. C. Thomas, S. S. Thomas, A. E. Toland, A. Trichopoulou, C. M. Ulrich, D. J. Van Den Berg, F. J. B. van Duijnhoven, B. Van Guelpen, H. van Kranen, J. Vijai, K. Visvanathan, P. Vodicka, L. Vodickova, V. Vymetalkova, K. Weigl, S. J. Weinstein, E. White, A. K. Win, C. R. Wolf, A. Wolk, M. O. Woods, A. H. Wu, S. H. Zaidi, B. W. Zanke, Q. Zhang, W. Zheng, P. C. Scacheri, J. D. Potter, M. C. Bassik, A. Kundaje, G. Casey, V. Moreno, G. R. Abecasis, D. A. Nickerson, S. B. Gruber, L. Hsu, and U. Peters, “Discovery of common and rare genetic risk variants for colorectal cancer,” *Nat Genet*, vol. 51, pp. 76–87, Jan 2019.

[324] S. Wang, J. J. Pitt, Y. Zheng, T. F. Yoshimatsu, G. Gao, A. Sanni, O. Oluwasola, M. Ajani, D. Fitzgerald, A. Odetunde, G. Khramtsova, I. Hurley, A. Popoola, A. Falusi, T. Ogundiran, J. Obafunwa, O. Ojengbede, N. Ibrahim, J. Barretina, K. P. White, D. Huo, and O. I. Olopade, “Germline variants and somatic mutation signatures of breast cancer across populations of African and European ancestry in the US and Nigeria,” *Int J Cancer*, vol. 145, pp. 3321–3333, Dec 2019.

[325] M. Vali-Pour, B. Lehner, and F. Supek, “The impact of rare germline variants on human somatic mutation processes,” *Nat Commun*, vol. 13, p. 3724, Jun 2022.

[326] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, and M. R. Stratton, "Deciphering signatures of mutational processes operative in human cancer," *Cell Rep*, vol. 3, pp. 246–259, Jan 2013.

[327] B. C. H. Lee, P. S. Robinson, T. H. H. Coorens, H. H. N. Yan, S. Olafsson, H. Lee-Six, M. A. Sanders, H. C. Siu, J. Hewinson, S. S. K. Yue, W. Y. Tsui, A. S. Y. Chan, A. K. W. Chan, S. L. Ho, P. J. Campbell, I. Martincorena, S. J. A. Buczacki, S. T. Yuen, S. Y. Leung, and M. R. Stratton, "Mutational landscape of normal epithelial cells in Lynch Syndrome patients," *Nat Commun*, vol. 13, p. 2710, May 2022.

[328] M. C. Olave and R. P. Graham, "Mismatch repair deficiency: The what, how and why it is important," *Genes Chromosomes Cancer*, vol. 61, pp. 314–321, Jun 2022.

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

[330] A. A. Bielska, W. K. Chatila, H. Walch, N. Schultz, Z. K. Stadler, J. Shia, D. Reidy-Lagunes, and R. Yaeger, "Tumor Mutational Burden and Mismatch Repair Deficiency Discordance as a Mechanism of Immunotherapy Resistance," *J Natl Compr Canc Netw*, vol. 19, pp. 130–133, Feb 2021.

[331] S.-H. Lin, R. Thakur, and M. J. Machiela, "LDexpress: an online tool for integrating population-specific linkage disequilibrium patterns with tissue-specific expression data," *BMC Bioinformatics*, vol. 22, p. 608, Dec. 2021.

[332] A. Cagan, A. Baez-Ortega, N. Brzozowska, F. Abascal, T. H. H. Coorens, M. A. Sanders, A. R. J. Lawson, L. M. R. Harvey, S. Bhosle, D. Jones, R. E. Alcantara, T. M. Butler, Y. Hooks, K. Roberts, E. Anderson, S. Lunn, E. Flach, S. Spiro, I. Januszczak, E. Wrigglesworth, H. Jenkins, T. Dallas, N. Masters, M. W. Perkins, R. Deaville, M. Druce, R. Bogeska, M. D. Milsom, B. Neumann, F. Gorman, F. Constantino-Casas, L. Peachey, D. Bochynska, E. S. J. Smith, M. Gerstung, P. J. Campbell, E. P. Murchison, M. R. Stratton, and I. Martincorena, "Somatic mutation rates scale with lifespan across mammals," *Nature*, vol. 604, pp. 517–524, Apr 2022.

[333] IDT, "xgen exome research panel v1.0," https://web.archive.org/web/20180403022641/http://eu.idtdna.co generation-sequencing/hybridization-capture/lockdown-panels/xgen-exome-research-panel.

[334] T. Iyama and D. M. Wilson, "Elements That Regulate the DNA Damage Response of Proteins Defective in Cockayne Syndrome," *J Mol Biol*, vol. 428, pp. 62–78, Jan 2016.

[335] V. Laugel, C. Dalloz, M. Durand, F. Sauvanaud, U. Kristensen, M. C. Vincent, L. Pasquier, S. Odent, V. Cormier-Daire, B. Gener, E. S. Tobias, J. L. Tolmie, D. Martin-Coignard, V. Drouin-Garraud, D. Heron, H. Journel, E. Raffo, J. Vigneron, S. Lyonnet, V. Murday, D. Gubser-Mercati, B. Funalot, L. Brueton, J. Sanchez Del Pozo, E. oz, A. R. Gennery, M. Salih, M. Noruzinia, K. Prescott, L. Ramos, Z. Stark, K. Fieggen, B. Chabrol, P. Sarda, P. Edery, A. Bloch-Zupan, H. Fawcett, D. Pham, J. M. Egly, A. R. Lehmann, A. Sarasin, and H. Dollfus, "Mutation update for the CSB/ERCC6 and CSA/ERCC8 genes involved in Cockayne syndrome," *Hum Mutat*, vol. 31, pp. 113–126, Feb 2010.

[336] T. A. Sasani, D. G. Ashbrook, A. C. Beichman, L. Lu, A. A. Palmer, R. W. Williams, J. K. Pritchard, and K. Harris, "A natural mutator allele shapes mutation spectrum variation in mice," *Nature*, vol. 605, pp. 497–502, May 2022.

[337] A. Platt, B. J. lmsson, and M. Nordborg, "Conditions under which genome-wide association studies will be positively misleading," *Genetics*, vol. 186, pp. 1045–1052, Nov 2010.

[338] A. nez Roca, M. Giner-Calabuig, O. Murcia, A. Castillejo, J. L. Soto, A. a Heredia, and R. Jover, "Lynch-like Syndrome: Potential Mechanisms and Management," *Cancers (Basel)*, vol. 14, Feb 2022.

[339] V. Thatikonda, S. M. A. Islam, R. J. Autry, B. C. Jones, S. N. bner, G. Warsow, B. Hutter, D. Huebschmann, S. hling, M. Kool, M. Blattner-Johnson, D. T. W. Jones, L. B. Alexandrov, S. M. Pfister, and N. ger, "Comprehensive analysis of mutational signatures reveals distinct patterns and molecular processes across 27 pediatric cancers," *Nat Cancer*, vol. 4, pp. 276–289, Feb 2023.

[340] A. Mojumdar, N. Mair, N. Adam, and J. A. Cobb, "Changes in DNA double-strand break repair during aging correlate with an increase in genomic mutations," *J Mol Biol*, vol. 434, p. 167798, Oct 2022.

[341] R. R. White and J. Vijg, “Do DNA Double-Strand Breaks Drive Aging?,” *Mol Cell*, vol. 63, pp. 729–738, Sep 2016.

[342] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023.

[343] H. Wickham, “ggplot2: Elegant graphics for data analysis,” 2016.

[344] L. J. Carithers and H. M. Moore, “The Genotype-Tissue Expression (GTEx) Project,” *Biopreserv Biobank*, vol. 13, pp. 307–308, Oct 2015.

[345] C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee, “Second-generation PLINK: rising to the challenge of larger and richer datasets,” *Gigascience*, vol. 4, p. 7, 2015.

[346] Z. Zheng, H. Hu, W. Lei, J. Zhang, M. Zhu, Y. Li, X. Zhang, J. Ma, D. Wan, T. Ma, G. Ren, and D. Ru, “,” *Evol Appl*, vol. 15, pp. 1875–1887, Nov 2022.

[347] D. A. Shagin, I. A. Shagina, A. R. Zaretsky, E. V. Barsova, I. V. Kelmanson, S. Lukyanov, D. M. Chudakov, and M. Shugay, “A high-throughput assay for quantitative measurement of PCR errors,” *Sci Rep*, vol. 7, p. 2718, Jun 2017.

[348] B. Arbeithuber, A. J. Betancourt, T. Ebner, and I. Tiemann-Boege, “Crossovers are associated with mutation and biased gene conversion at recombination hotspots,” *Proc Natl Acad Sci U S A*, vol. 112, pp. 2109–2114, Feb 2015.

[349] A. V. Nesta, D. Tafur, and C. R. Beck, “Hotspots of Human Mutation,” *Trends Genet*, vol. 37, pp. 717–729, Aug 2021.

[350] C. J. Sakofsky, N. Saini, L. J. Klimczak, K. Chan, E. P. Malc, P. A. Mieczkowski, A. B. Burkholder, D. Fargo, and D. A. Gordenin, “Repair of multiple simultaneous double-strand breaks causes bursts of genome-wide clustered hypermutation,” *PLoS Biol*, vol. 17, p. e3000464, Sep 2019.

[351] X. Long and H. Xue, “Genetic-variant hotspots and hotspot clusters in the human genome facilitating adaptation while increasing instability,” *Hum Genomics*, vol. 15, p. 19, Mar 2021.

[352] K. K. Takahashi and H. Innan, “Frequent somatic gene conversion as a mechanism for loss of heterozygosity in tumor suppressor genes,” *Genome Res*, vol. 32, pp. 1017–1025, Jun 2022.

[353] Q. Yi, J. Peng, Z. Xu, Q. Liang, Y. Cai, B. Peng, Q. He, and Y. Yan, "Spectrum of BRAF Aberrations and Its Potential Clinical Implications: Insights From Integrative Pan-Cancer Analysis," *Front Bioeng Biotechnol*, vol. 10, p. 806851, 2022.

[354] C. J. Watson, A. L. Papula, G. Y. P. Poon, W. H. Wong, A. L. Young, T. E. Druley, D. S. Fisher, and J. R. Blundell, "The evolutionary dynamics and fitness landscape of clonal hematopoiesis," *Science*, vol. 367, pp. 1449–1454, Mar 2020.

[355] S. Jaiswal and B. L. Ebert, "Clonal hematopoiesis in human aging and disease," *Science*, vol. 366, Nov 2019.

[356] Y. Watanabe, T. Abe, T. Ikemura, and M. Maekawa, "Relationships between replication timing and GC content of cancer-related genes on human chromosomes 11q and 21q," *Gene*, vol. 433, pp. 26–31, Mar 2009.

[357] M. Heuser, F. Thol, and A. Ganser, "Clonal Hematopoiesis of Indeterminate Potential," *Dtsch Arztebl Int*, vol. 113, pp. 317–322, May 2016.

[358] P. Valent, W. Kern, G. Hoermann, J. D. Milosevic Feenstra, K. Sotlar, M. cker, U. Germing, W. R. Sperr, A. Reiter, D. Wolf, M. Arock, T. Haferlach, and H. P. Horny, "Clonal Hematopoiesis with Oncogenic Potential (CHOP): Separation from CHIP and Roads to AML," *Int J Mol Sci*, vol. 20, Feb 2019.

[359] Z. J. DeBruine, K. Melcher, and J. Timothy J. Triche, "Fast and robust non-negative matrix factorization for single-cell experiments," *bioRxiv*, 2021.

[360] D. Christopoulos, "Introducing unit invariant knee (uik) as an objective choice for elbow point in multivariate data analysis techniques," *Available at SSRN 3043076*, 2016.

[361] T. Yasuzawa, K. Muroi, M. Ichimura, I. Takahashi, T. Ogawa, K. Takahashi, H. Sano, and Y. Saitoh, "Duocarmycins, potent antitumor antibiotics produced by Streptomyces sp. structures and chemistry," *Chem Pharm Bull (Tokyo)*, vol. 43, pp. 378–391, Mar 1995.

[362] J. A. Stamatoyannopoulos, I. Adzhubei, R. E. Thurman, G. V. Kryukov, S. M. Mirkin, and S. R. Sunyaev, "Human mutation rate associated with DNA replication timing," *Nat Genet*, vol. 41, pp. 393–395, Apr 2009.

[363] V. B. Seplyarskiy, R. A. Soldatov, K. Y. Popadin, S. E. Antonarakis, G. A. Bazykin, and S. I. Nikolaev, "APOBEC-induced mutations in human cancers are strongly enriched

on the lagging DNA strand during replication," *Genome Res*, vol. 26, pp. 174–182, Feb 2016.

[364] M. E. Moynahan and M. Jasin, "Mitotic homologous recombination maintains genomic stability and suppresses tumorigenesis," *Nat Rev Mol Cell Biol*, vol. 11, pp. 196–207, Mar 2010.

[365] E. D. Pleasance, R. K. Cheetham, P. J. Stephens, D. J. McBride, S. J. Humphray, C. D. Greenman, I. Varela, M. L. Lin, G. R. ez, G. R. Bignell, K. Ye, J. Alipaz, M. J. Bauer, D. Beare, A. Butler, R. J. Carter, L. Chen, A. J. Cox, S. Edkins, P. I. Kokko-Gonzales, N. A. Gormley, R. J. Grocock, C. D. Haudenschild, M. M. Hims, T. James, M. Jia, Z. Kingsbury, C. Leroy, J. Marshall, A. Menzies, L. J. Mudie, Z. Ning, T. Royce, O. B. Schulz-Trieglaff, A. Spiridou, L. A. Stebbings, L. Szajkowski, J. Teague, D. Williamson, L. Chin, M. T. Ross, P. J. Campbell, D. R. Bentley, P. A. Futreal, and M. R. Stratton, "A comprehensive catalogue of somatic mutations from a human cancer genome," *Nature*, vol. 463, pp. 191–196, Jan 2010.

[366] S. Amatori, S. Tavolaro, S. Gambardella, and M. Fanelli, "The dark side of histones: genomic organization and role of oncohistones in cancer," *Clin Epigenetics*, vol. 13, p. 71, Apr 2021.

[367] T. T. Paull, E. P. Rogakou, V. Yamazaki, C. U. Kirchgessner, M. Gellert, and W. M. Bonner, "A critical role for histone H2AX in recruitment of repair factors to nuclear foci after DNA damage," *Curr Biol*, vol. 10, no. 15, pp. 886–895, 2000.

[368] F. Maura, A. Degasperi, F. Nadeu, D. Leongamornlert, H. Davies, L. Moore, R. Royo, B. Ziccheddu, X. S. Puente, H. Avet-Loiseau, P. J. Campbell, S. Nik-Zainal, E. Campo, N. Munshi, and N. Bolli, "A practical guide for mutational signature analysis in hematological malignancies," *Nat Commun*, vol. 10, p. 2969, Jul 2019.

[369] J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, B. Foster, M. Moser, E. Karasik, B. Gillard, K. Ramsey, S. Sullivan, J. Bridge, H. Magazine, J. Syron, J. Fleming, L. Siminoff, H. Traino, M. Mosavel, L. Barker, S. Jewell, D. Rohrer, D. Maxim, D. Filkins, P. Harbach, E. Cortadillo, B. Berghuis, L. Turner, E. Hudson, K. Feenstra, L. Sobin, J. Robb, P. Branton, G. Korzeniewski, C. Shive, D. Tabor, L. Qi, K. Groch, S. Nampally, S. Buia, A. Zimmerman, A. Smith, R. Burges, K. Robinson, K. Valentino, D. Bradbury, M. Cosentino, N. Diaz-Mayoral, M. Kennedy, T. Engel, P. Williams, K. Erickson, K. Ardlie, W. Winckler, G. Getz, D. DeLuca, D. MacArthur, M. Kellis, A. Thomson, T. Young, E. Gelfand,

M. Donovan, G. Grant, D. Mash, Y. Marcus, M. Basile, J. Liu, J. Zhu, Z. Tu, N. J. Cox, D. L. Nicolae, E. R. Gamazon, H. Kyung, A. Konkashbaev, J. Pritchard, M. Stevens, T. Flutre, X. Wen, T. Dermitzakis, T. Lappalainen, R. Guigo, J. Monlong, M. Sammeth, D. Koller, A. Battle, S. Mostafavi, M. McCarthy, M. Rivas, J. Maller, I. Rusyn, A. Nobel, F. Wright, A. Shabalin, M. Feolo, N. Sharopova, A. Sturcke, J. Paschal, J. M. Anderson, E. L. Wilder, L. K. Derr, E. D. Green, J. P. Struewing, G. Temple, S. Volpi, J. T. Boyer, E. J. Thomson, M. S. Guyer, C. Ng, A. Abdallah, D. Colantuoni, T. R. Insel, S. E. Koester, A. R. Little, P. K. Bender, T. Lehner, Y. Yao, C. C. Compton, J. B. Vaught, S. Sawyer, N. C. Lockhart, J. Demchok, and H. F. Moore, "The Genotype-Tissue Expression (GTEx) project," *Nat Genet*, vol. 45, pp. 580–585, Jun 2013.

[370] A. R. Quinlan and I. M. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, pp. 841–842, Mar 2010.

[371] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu, "Model-based analysis of ChIP-Seq (MACS)," *Genome Biol*, vol. 9, no. 9, p. R137, 2008.

[372] Y. Li, S. Chen, T. Rapakoulia, H. Kuwahara, K. Y. Yip, and X. Gao, "Deep learning identifies and quantifies recombination hotspot determinants," *Bioinformatics*, vol. 38, pp. 2683–2691, May 2022.

[373] E. Peter, "Fbroc: Fast algorithms to bootstrap receiver operating characteristics curves," *R package version 0.4. 0, URL https://CRAN. R-project. org/package= fbroc*, 2016.

[374] F. Manders, A. M. Brandsma, J. de Kanter, M. Verheul, R. Oka, M. J. van Roosmalen, B. van der Roest, A. van Hoeck, E. Cuppen, and R. van Boxtel, "MutationalPatterns: the one stop shop for the analysis of mutational processes," *BMC Genomics*, vol. 23, p. 134, Feb 2022.

[375] J. Yang, N. A. Zaitlen, M. E. Goddard, P. M. Visscher, and A. L. Price, "Advantages and pitfalls in the application of mixed-model association methods," *Nat Genet*, vol. 46, pp. 100–106, Feb 2014.

[376] B. Devlin and K. Roeder, "Genomic control for association studies," *Biometrics*, vol. 55, pp. 997–1004, Dec 1999.

[377] R. A. Fisher, "Design of experiments," *British Medical Journal*, vol. 1, no. 3923, p. 554, 1936.

[378] J. M. Neuhaus, “Estimation efficiency with omitted covariates in generalized linear models,” *Journal of the American Statistical Association*, vol. 93, no. 443, pp. 1124–1129, 1998.

[379] C. Lippert, G. Quon, E. Y. Kang, C. M. Kadie, J. Listgarten, and D. Heckerman, “The benefits of selecting phenotype-specific variants for applications of mixed models in genomics,” *Sci Rep*, vol. 3, p. 1815, 2013.

[380] G. Tucker, A. L. Price, and B. Berger, “Improving the power of GWAS and avoiding confounding from population stratification with PC-Select,” *Genetics*, vol. 197, pp. 1045–1049, Jul 2014.

[381] O. Canela-Xandri, K. Rawlik, and A. Tenesa, “An atlas of genetic associations in UK Biobank,” *Nat Genet*, vol. 50, pp. 1593–1599, Nov 2018.

[382] C. Kadie and D. Heckerman, “Ludicrous speed linear mixed models for genome-wide association studies,” *bioRxiv*, 2019.

[383] V. Tam, N. Patel, M. Turcotte, Y. Ã©, G. Ã©, and D. Meyre, “Benefits and limitations of genome-wide association studies,” *Nat Rev Genet*, vol. 20, pp. 467–484, Aug 2019.

[384] A. Torkamani, N. E. Wineinger, and E. J. Topol, “The personal and clinical utility of polygenic risk scores,” *Nat Rev Genet*, vol. 19, pp. 581–590, Sep 2018.

[385] T. Yanes, M. A. Young, B. Meiser, and P. A. James, “Clinical applications of polygenic breast cancer risk: a critical review and perspectives of an emerging field,” *Breast Cancer Res*, vol. 22, p. 21, Feb 2020.

[386] K. Michailidou, S. m, J. Dennis, J. Beesley, S. Hui, S. Kar, A. on, P. Soucy, D. Glubb, A. Rostamianfar, M. K. Bolla, Q. Wang, J. Tyrer, E. Dicks, A. Lee, Z. Wang, J. Allen, R. Keeman, U. Eilber, J. D. French, X. Qing Chen, L. Fachal, K. McCue, A. E. McCart Reed, M. Ghoussaini, J. S. Carroll, X. Jiang, H. Finucane, M. Adams, M. A. Adank, H. Ahsan, K. ki, H. Anton-Culver, N. N. Antonenkova, V. Arndt, K. J. Aronson, B. Arun, P. L. Auer, F. Bacot, M. Barrdahl, C. Baynes, M. W. Beckmann, S. Behrens, J. Benitez, M. Bermisheva, L. Bernstein, C. Blomqvist, N. V. Bogdanova, S. E. Bojesen, B. Bonanni, A. L. rresen Dale, J. S. Brand, H. Brauch, P. Brennan, H. Brenner, L. Brinton, P. Broberg, I. W. Brock, A. Broeks, A. Brooks-Wilson, S. Y. Brucker, T. ning, B. Burwinkel, K. Butterbach, Q. Cai, H. Cai, T. s, F. Canzian, A. Carracedo, B. D. Carter, J. E. Castelao, T. L. Chan, T. Y. David Cheng, K. Seng Chia, J. Y. Choi,

H. Christiansen, C. L. Clarke, M. e, D. M. Conroy, E. Cordina-Duverger, S. Cornelissen, D. G. Cox, A. Cox, S. S. Cross, J. M. Cunningham, K. Czene, M. B. Daly, P. Devilee, K. F. Doheny, T. rk, I. Dos-Santos-Silva, M. Dumont, L. Durcan, M. Dwek, D. M. Eccles, A. B. Ekici, A. H. Eliassen, C. Ellberg, M. Elvira, C. Engel, M. Eriksson, P. A. Fasching, J. Figueroa, D. Flesch-Janys, O. Fletcher, H. Flyger, L. Fritschi, V. Gaborieau, M. Gabrielson, M. Gago-Dominguez, Y. T. Gao, S. M. Gapstur, J. A. enz, M. M. Gaudet, V. Georgoulias, G. G. Giles, G. Glendon, M. S. Goldberg, D. E. Goldgar, A. lez Neira, G. I. s, M. Grip, J. Gronwald, A. Grundy, P. nel, L. Haeberle, E. Hahnen, C. A. Haiman, N. kansson, U. Hamann, N. Hamel, S. Hankinson, P. Harrington, S. N. Hart, J. M. Hartikainen, M. Hartman, A. Hein, J. Heyworth, B. Hicks, P. Hillemanns, D. N. Ho, A. Hollestelle, M. J. Hooning, R. N. Hoover, J. L. Hopper, M. F. Hou, C. N. Hsiung, G. Huang, K. Humphreys, J. Ishiguro, H. Ito, M. Iwasaki, H. Iwata, A. Jakubowska, W. Janni, E. M. John, N. Johnson, K. Jones, M. Jones, A. Jukkola-Vuorinen, R. Kaaks, M. Kabisch, K. Kaczmarek, D. Kang, Y. Kasuga, M. J. Kerin, S. Khan, E. Khusnutdinova, J. I. Kiiski, S. W. Kim, J. A. Knight, V. M. Kosma, V. N. Kristensen, U. ger, A. Kwong, D. Lambrechts, L. Le Marchand, E. Lee, M. H. Lee, J. W. Lee, C. Neng Lee, F. Lejbkowicz, J. Li, J. Lilyquist, A. Lindblom, J. Lissowska, W. Y. Lo, S. Loibl, J. Long, A. Lophatananon, J. Lubinski, C. Luccarini, M. P. Lux, E. S. K. Ma, R. J. MacInnis, T. Maishman, E. Makalic, K. E. Malone, I. M. Kostovska, A. Mannermaa, S. Manoukian, J. E. Manson, S. Margolin, S. Mariapun, M. E. Martinez, K. Matsuo, D. Mavroudis, J. McKay, C. McLean, H. Meijers-Heijboer, A. Meindl, P. ndez, U. Menon, J. Meyer, H. Miao, N. Miller, N. A. M. Taib, K. Muir, A. M. Mulligan, C. Mulot, S. L. Neuhausen, H. Nevanlinna, P. Neven, S. F. Nielsen, D. Y. Noh, B. G. Nordestgaard, A. Norman, O. I. Olopade, J. E. Olson, H. Olsson, C. Olswold, N. Orr, V. S. Pankratz, S. K. Park, T. W. Park-Simon, R. Lloyd, J. I. A. Perez, P. Peterlongo, J. Peto, K. A. Phillips, M. Pinchev, D. Plaseska-Karanfilska, R. Prentice, N. Presneau, D. Prokofyeva, E. Pugh, K. s, B. Rack, P. Radice, N. Rahman, G. Rennert, H. S. Rennert, V. Rhenius, A. Romero, J. Romm, K. J. Ruddy, T. diger, A. Rudolph, M. Ruebner, E. J. T. Rutgers, E. Saloustros, D. P. Sandler, S. Sangrajrang, E. J. Sawyer, D. F. Schmidt, R. K. Schmutzler, A. Schneeweiss, M. J. Schoemaker, F. Schumacher, P. rmann, R. J. Scott, C. Scott, S. Seal, C. Seynaeve, M. Shah, P. Sharma, C. Y. Shen, G. Sheng, M. E. Sherman, M. J. Shrubsole, X. O. Shu, A. Smeets, C. Sohn, M. C. Southey, J. J. Spinelli, C. Stegmaier, S. Stewart-Brown, J. Stone, D. O. Stram, H. Surowy, A. Swerdlow, R. Tamimi, J. A. Taylor, M. m, S. H. Teo, M. Beth Terry, D. C. Tessier, S. Thanasitthichai, K. ne, R. A. E. M. Tollenaar, I. Tomlinson, L. Tong, D. Torres, T. Truong, C. C. Tseng, S. Tsugane, H. U. Ulmer,

G. Ursin, M. Untch, C. Vachon, C. J. van Asperen, D. Van Den Berg, A. M. W. van den Ouweland, L. van der Kolk, R. B. van der Luijt, D. Vincent, J. Vollenweider, Q. Waisfisz, S. Wang-Gohrke, C. R. Weinberg, C. Wendt, A. S. Whittemore, H. Wildiers, W. Willett, R. Winqvist, A. Wolk, A. H. Wu, L. Xia, T. Yamaji, X. R. Yang, C. Har Yip, K. Y. Yoo, J. C. Yu, W. Zheng, Y. Zheng, B. Zhu, A. Ziogas, E. Ziv, S. R. Lakhani, A. C. Antoniou, A. Droit, I. L. Andrulis, C. I. Amos, F. J. Couch, P. D. P. Pharoah, J. Chang-Claude, P. Hall, D. J. Hunter, R. L. Milne, M. a Closas, M. K. Schmidt, S. J. Chanock, A. M. Dunning, S. L. Edwards, G. D. Bader, G. Chenevix-Trench, J. Simard, P. Kraft, and D. F. Easton, "Association analysis identifies 65 new breast cancer risk loci," *Nature*, vol. 551, pp. 92–94, Nov 2017.

[387] O. Bahcall, "Common variation and heritability estimates for breast, ovarian and prostate cancers," *Nat Genet*, vol. 10, 2013.

[388] S. W. Choi and P. F. O'Reilly, "PRSice-2: Polygenic Risk Score software for biobank-scale data," *Gigascience*, vol. 8, Jul 2019.

[389] F. Dudbridge, "Power and predictive accuracy of polygenic risk scores," *PLoS Genet*, vol. 9, p. e1003348, Mar 2013.

[390] A. R. Martin, M. J. Daly, E. B. Robinson, S. E. Hyman, and B. M. Neale, "Predicting Polygenic Risk of Psychiatric Disorders," *Biol Psychiatry*, vol. 86, pp. 97–109, Jul 2019.

[391] N. Mars, J. T. Koskela, P. Ripatti, T. T. J. Kiiskinen, A. S. Havulinna, J. V. Lindbohm, A. Ahola-Olli, M. Kurki, J. Karjalainen, P. Palta, B. M. Neale, M. Daly, V. Salomaa, A. Palotie, E. n, S. Ripatti, A. Palotie, M. Daly, H. Jacob, A. Matakidou, H. Runz, S. John, R. Plenge, M. McCarthy, J. Hunkapiller, M. Ehm, D. Waterworth, C. Fox, A. Malarstig, K. Klinger, K. Call, T. Ã¤, J. Kaprio, P. Virolainen, K. Pulkki, T. Kilpi, M. Perola, J. Partanen, A. ranta, R. Kaarteenaho, S. Vainio, K. Savinainen, V. M. Kosma, U. Kujala, O. Tuovila, M. Hendolin, R. Pakkanen, J. Waring, B. Riley-Gillis, A. Matakidou, H. Runz, J. Liu, S. Biswas, J. Hunkapiller, D. Waterworth, M. Ehm, D. Diogo, C. Fox, A. Malarstig, C. Marshall, X. Hu, K. Call, K. Klinger, M. Gossel, S. Ripatti, J. Schleutker, M. Perola, M. Arvas, O. Carpen, R. Hinttala, J. Kettunen, R. Laaksonen, A. Mannermaa, J. Paloneva, U. Kujala, O. Tuovila, M. Hendolin, R. Pakkanen, H. Soininen, V. Julkunen, A. Remes, R. inen, M. Hiltunen, J. Peltola, P. Tienari, J. Rinne, A. Ziemann, J. Waring, S. Esmaeeli, N. Smaoui, A. Lehtonen, S. Eaton, H. Runz, S. Ã¤, S. Biswas, J. Michon, G. Kerchner, J. Hunkapiller, N. Bowers, E. Teng, J. Eicher, V. Mehta, P. Gormley, K. Linden, C. Whelan, F. Xu, D. Pulford, M. Ã¤, S. Pikkarainen,

A. Jussila, T. Blomster, M. Kiviniemi, M. Voutilainen, B. Georgantas, G. Heap, J. Waring, N. Smaoui, F. Rahimov, A. Lehtonen, K. Usiskin, J. Maranville, T. Lu, N. Bowers, D. Oh, J. Michon, V. Mehta, K. Kalpala, M. Miller, X. Hu, L. McCarthy, K. Eklund, A. ki, P. ki, L. Ã¤, O. nen, J. Huhtakangas, B. Georgantas, J. Waring, F. Rahimov, A. Lertratanakul, N. Smaoui, A. Lehtonen, D. Close, M. Hochfeld, N. Bowers, J. Michon, D. Diogo, V. Mehta, K. Kalpala, N. Bing, X. Hu, J. Esparza Gordillo, N. Mars, T. Laitinen, M. Pelkonen, P. Kauppi, H. Kankaanranta, T. Harju, N. Smaoui, D. Close, S. Greenberg, H. Chen, N. Bowers, J. Michon, V. Mehta, J. Betts, S. Ghosh, V. Salomaa, T. Niiranen, M. Juonala, K. rinne, M. nen, J. Junttila, M. Laakso, J. ki, J. Sinisalo, M. R. Taskinen, T. Tuomi, J. Laukkanen, B. Challis, A. Peterson, J. Hunkapiller, N. Bowers, J. Michon, D. Diogo, A. Chu, V. Mehta, J. Parkkinen, M. Miller, A. Muslin, D. Waterworth, H. Joensuu, T. Meretoja, O. Carpen, L. Aaltonen, A. Auranen, P. Karihtala, S. Kauppila, P. Auvinen, K. Elenius, R. Popovic, J. Waring, B. Riley-Gillis, A. Lehtonen, A. Matakidou, J. Schutzman, J. Hunkapiller, N. Bowers, J. Michon, V. Mehta, A. Loboda, A. Chhibber, H. Lehtonen, S. McDonough, M. Crohns, D. Kulkarni, K. Kaarniranta, J. Turunen, T. Ollila, S. Seitsonen, H. Uusitalo, V. Aaltonen, H. rvinen, M. Ã¤, N. Hautala, H. Runz, E. Strauss, N. Bowers, H. Chen, J. Michon, A. Podgornaia, V. Mehta, D. Diogo, J. Hoffman, K. Tasanen, L. Huilaja, K. Hannula-Jouppi, T. Salmi, S. Peltonen, L. Koulu, I. Harvima, K. Kalpala, Y. Wu, D. Choy, J. Michon, N. Smaoui, F. Rahimov, A. Lehtonen, D. Waterworth, A. Jalanko, R. Kajanne, U. Lyhs, M. Kaunisto, J. W. Davis, B. Riley-Gillis, D. Quarless, S. Petrovski, J. Liu, C. Y. Chen, P. Bronson, R. Yang, J. Maranville, S. Biswas, D. Chang, J. Hunkapiller, T. Bhangale, N. Bowers, D. Diogo, E. Holzinger, P. Gormley, X. Wang, X. Chen, Ã. Hedman, K. Auro, C. Wang, E. Xu, F. Auge, C. Chatelain, M. Kurki, S. Ripatti, M. Daly, J. Karjalainen, A. Havulinna, A. Jalanko, K. Palin, P. Palta, P. Della Briotta Parolo, W. Zhou, S. Ã¤, M. Rivas, J. Harju, A. Palotie, A. Lehisto, A. Ganna, V. Llorens, A. Karlsson, K. Kristiansson, M. Arvas, K. rinen, J. Ritari, T. Wahlfors, M. Koskinen, O. Carpen, J. Kettunen, K. s, M. Kalaoja, M. Karjalainen, T. Mantere, E. Kangasniemi, S. Heikkinen, A. Mannermaa, E. Laakkonen, J. Kononen, A. Loukola, P. Laiho, T. Sistonen, E. Kaiharju, M. Laukkanen, E. rvensivu, S. ki, L. Ã¶, R. Wong, K. Kristiansson, H. Mattsson, S. Ã¤, T. Hiekkalinna, M. nez, K. Donner, P. Palta, K. rn, J. Nunez-Fontarnau, J. Harju, E. inen, T. P. Ã¤, G. Brein, A. Dada, G. Awaisa, A. Shcherban, T. Ã¤, H. Laivuori, A. Havulinna, S. Ã¤, T. Kiiskinen, T. Laitinen, H. Siirtola, J. Gracia Tabuenca, L. Kallio, S. Soini, J. Partanen, K. nen, S. Vainio, K. Savinainen, V. M. Kosma, and T. Kuopio, “Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers," *Nat Med*, vol. 26, pp. 549–557, Apr 2020.

[392] G. Gibson, "On the utilization of polygenic risk scores for therapeutic targeting," *PLoS Genet*, vol. 15, p. e1008060, Apr 2019.

[393] Q. Zhang, F. Prive, B. J. Vilhjalmsson, and D. Speed, "Improved genetic prediction of complex traits from individual-level data or summary statistics," *Nat com*, 2021.

[394] B. Bulik-Sullivan, "Mixed models for meta-analysis and sequencing," *bioRxiv*, 2015.

[395] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J. C. Sanchez, and M. ller, "pROC: an open-source package for R and S+ to analyze and compare ROC curves," *BMC Bioinformatics*, vol. 12, p. 77, Mar 2011.

[396] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly, and J. Marchini, "The UK Biobank resource with deep phenotyping and genomic data," *Nature*, vol. 562, pp. 203–209, Oct 2018.

[397] A. Manichaikul, J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, and W. M. Chen, "Robust relationship inference in genome-wide association studies," *Bioinformatics*, vol. 26, pp. 2867–2873, Nov 2010.

Appendix A

Table A1: Wilcoxon signed rank test p-values for association between mismatch load and cancer and smoking status.

Context	Cancer	Cancer Norm	Smoking	Smoking Norm
Total load	6.17e-02	6.17e-01	1.23e-01	9.19e-01
CTACGA	4.07e-28	1.41e-01	4.12e-01	1.49e-02
GAGGCG	9.87e-28	4.27e-01	3.93e-01	2.98e-01
TTCTGC	3.65e-25	6.68e-01	3.50e-01	3.54e-03
GAAGCA	4.23e-25	5.11e-01	4.03e-01	5.69e-02
AAAACA	6.38e-25	9.29e-01	6.79e-01	3.94e-01
TTATGA	1.05e-24	8.44e-01	9.22e-01	2.91e-01
CTCCGC	1.49e-24	1.84e-01	5.74e-01	3.21e-01
AAGACG	2.23e-24	4.78e-01	3.19e-01	4.30e-02
TAGTCG	2.71e-24	9.61e-01	8.56e-01	7.77e-01
TAATCA	5.37e-24	9.75e-01	4.90e-01	1.71e-01
CTTCGT	7.57e-24	9.68e-01	2.78e-01	2.90e-02
GTCGGC	2.07e-23	5.99e-01	3.15e-01	7.32e-02
TTTTGT	3.73e-23	1.50e-01	6.41e-01	3.95e-01
CAACCA	1.56e-22	8.98e-01	4.31e-01	1.43e-01
TACTCC	2.12e-22	9.28e-01	9.88e-01	9.65e-01
GACGCC	5.40e-22	2.05e-01	3.68e-01	2.93e-02
CTGCGG	5.60e-22	4.22e-01	3.07e-01	1.55e-02
GTAGGA	1.99e-21	5.12e-01	6.53e-01	2.56e-01
CAGCCG	1.99e-21	2.37e-01	3.44e-01	1.30e-01
AACACC	2.51e-21	9.93e-01	7.46e-01	2.52e-01
GGGGCG	6.77e-21	8.73e-02	1.50e-01	6.61e-01
GTGGGG	7.19e-21	1.69e-01	7.13e-01	1.52e-01
GTTGGT	2.37e-20	2.61e-01	6.23e-01	1.99e-01
TTGTGG	2.77e-20	3.10e-01	3.20e-01	2.23e-01
CACCCC	4.36e-20	9.37e-01	9.47e-01	2.62e-01
CCTCGT	5.76e-20	8.60e-01	1.50e-01	7.97e-01
GGAGCA	2.46e-19	7.86e-01	2.25e-01	7.19e-01
AGGACG	2.47e-19	9.90e-01	2.78e-01	6.14e-01
GGTGTT	5.73e-19	7.76e-02	2.82e-01	6.39e-01
TGGTCG	7.17e-19	1.34e-01	1.38e-01	9.91e-01
TCATGA	3.91e-18	6.62e-01	1.35e-01	3.97e-01
CATCCT	4.13e-18	5.59e-01	3.84e-01	3.76e-01
CCCCGC	4.36e-18	5.24e-01	7.84e-02	1.55e-01
TCCTGC	9.70e-18	5.20e-01	6.19e-02	2.64e-02
TCTTGT	1.78e-17	5.27e-01	6.67e-02	2.52e-03
TGATCA	1.81e-17	7.97e-01	1.69e-01	8.27e-01
CCGCGG	4.30e-17	4.93e-01	4.81e-01	5.48e-01
AGAACA	5.09e-17	2.76e-01	2.22e-01	1.92e-01
CCACGA	7.75e-17	5.89e-01	4.64e-02	5.83e-02
ATGAGG	1.04e-16	3.87e-01	6.07e-01	8.16e-01
ATTAGT	1.23e-16	9.15e-01	5.04e-01	8.59e-01
AATACT	1.56e-16	5.57e-01	7.16e-01	6.27e-01
GCTGGT	3.41e-16	2.87e-01	4.28e-02	1.99e-02
GCGGGG	3.63e-16	2.26e-01	2.13e-02	2.33e-02
ACGAGG	5.28e-16	9.55e-01	1.97e-01	4.12e-01
AGCACC	1.33e-15	5.05e-01	2.93e-01	6.07e-01
AGTATT	1.36e-15	5.21e-01	3.57e-01	3.48e-01
GCCGGC	3.75e-15	3.22e-01	2.35e-01	9.59e-01
CGACCA	1.39e-14	6.49e-01	2.30e-01	4.77e-01
CGGCCG	1.54e-14	3.49e-01	2.73e-01	6.94e-01
CGCCCC	3.28e-14	9.48e-01	4.36e-01	9.07e-01
AGAATA	3.62e-14	8.33e-01	3.18e-01	6.29e-01
GATGCT	4.15e-14	6.60e-01	4.43e-01	8.51e-01
GGAGTA	4.30e-14	1.99e-01	4.40e-01	7.60e-01
TGTTTT	4.82e-14	3.32e-01	6.68e-01	2.84e-01
GGCGCC	6.31e-14	8.88e-01	1.62e-01	5.07e-01
ACTAGT	8.77e-14	9.32e-01	3.82e-01	9.22e-01

ATCAGC	1.31e-13	1.90e-01	3.81e-01	7.32e-01
GCAGGA	1.76e-13	5.83e-01	8.67e-02	9.69e-02
TCGTGG	2.76e-13	4.60e-01	8.08e-01	3.35e-01
CGTCCT	3.56e-13	4.52e-01	3.64e-01	7.27e-01
GGCGTC	9.37e-13	5.03e-01	6.55e-01	1.67e-01
ACCAGC	1.01e-12	5.06e-01	3.05e-01	7.82e-01
ACAATA	1.84e-12	3.04e-01	4.99e-01	1.78e-01
AGCATC	2.22e-12	7.81e-01	3.45e-01	9.20e-01
CGTCTT	6.16e-12	6.08e-01	4.88e-01	7.54e-01
GGTGCT	6.80e-12	6.98e-01	2.43e-01	6.08e-01
TGCTCC	8.15e-12	9.29e-01	1.20e-01	4.00e-01
CGCCTC	1.05e-11	4.75e-01	1.84e-01	6.89e-01
CCACTA	1.26e-11	3.44e-01	4.90e-01	1.65e-01
CGACTA	2.43e-11	9.11e-01	2.75e-01	8.32e-01
AGTACT	3.21e-11	7.93e-02	4.72e-01	4.91e-01
TCATTA	2.09e-10	2.02e-01	6.79e-01	6.50e-01
ATAAGA	3.29e-10	8.59e-01	3.92e-01	9.81e-01
TGTTCT	2.47e-09	9.14e-01	4.17e-01	2.05e-01
CCCCTC	2.52e-09	7.12e-01	4.94e-01	1.36e-01
TATTCT	2.57e-09	4.62e-01	3.79e-01	5.85e-01
AGGATG	6.05e-09	7.04e-01	1.94e-01	8.93e-01
GGGGTG	7.87e-09	3.69e-01	2.04e-01	8.38e-01
CGGCTG	1.44e-08	7.33e-01	3.87e-01	4.14e-01
ACAAGA	1.77e-08	1.44e-01	5.35e-01	3.85e-02
ACCATC	2.26e-08	6.51e-01	5.39e-01	1.70e-01
TGCTTC	2.82e-08	3.96e-01	6.51e-01	1.96e-01
ACTATT	3.41e-08	1.38e-01	1.87e-01	8.64e-01
TGTTAT	5.49e-08	5.94e-01	1.01e-01	3.99e-01
CTACCA	5.89e-08	4.03e-01	8.66e-02	7.43e-01
TCCTTC	6.57e-08	8.71e-01	4.67e-01	6.03e-01
GCAGTA	7.52e-08	6.20e-01	2.08e-01	7.97e-01
ATAACA	1.57e-07	7.01e-01	1.21e-01	8.28e-01
TGGTAG	1.99e-07	6.11e-02	1.74e-01	4.07e-01
GCCGTC	2.04e-07	2.77e-01	1.25e-01	6.45e-01
CTCCCC	2.39e-07	2.06e-01	1.43e-01	7.74e-01
ATCACC	2.62e-07	1.76e-01	1.01e-01	5.56e-01
TACTGC	5.94e-07	8.52e-01	1.26e-01	8.77e-01
TGATTA	7.85e-07	9.53e-01	5.78e-01	4.68e-01
TGATAA	1.18e-06	5.75e-01	2.45e-01	5.71e-01
TGGTTG	1.42e-06	4.98e-01	2.02e-01	8.23e-01
CCGCAG	1.60e-06	6.25e-02	1.82e-01	9.83e-01
TAGTGG	2.01e-06	2.13e-01	1.24e-01	5.62e-01
GTCGCC	2.06e-06	4.56e-01	1.95e-01	3.85e-01
GACGGC	2.51e-06	4.30e-01	7.68e-02	6.16e-01
CAGCGG	2.97e-06	7.03e-01	4.33e-02	6.95e-02
TGCTAC	3.34e-06	4.23e-01	2.38e-01	5.98e-01
TATTGT	3.59e-06	2.26e-01	9.63e-02	8.32e-01
TTCTCC	3.63e-06	9.96e-01	1.14e-01	5.49e-01
GTAGCA	4.38e-06	4.18e-01	6.70e-02	3.63e-01
GGTGAT	5.32e-06	4.74e-02	2.47e-01	2.15e-01
CACCGC	6.07e-06	8.04e-01	8.03e-02	3.97e-01
GATGGT	6.16e-06	6.69e-01	1.23e-01	8.63e-01
GCTGTT	7.48e-06	2.32e-01	1.86e-01	2.97e-01
CTTCCT	9.34e-06	7.06e-01	1.37e-01	7.08e-01
GAGGGG	1.20e-05	3.61e-01	4.48e-02	2.04e-01
GAAGGA	1.63e-05	2.11e-01	8.92e-02	9.14e-01
TTGTCG	2.12e-05	4.01e-01	5.20e-02	2.48e-01
CTGCCG	2.33e-05	7.10e-01	9.45e-02	7.04e-01
TAATGA	2.88e-05	3.87e-02	8.06e-02	4.93e-01
TTATCA	3.07e-05	9.75e-02	7.95e-02	4.60e-01
ATGACG	3.21e-05	3.26e-01	7.12e-02	3.49e-01
CATCGT	4.04e-05	4.30e-01	2.98e-02	4.91e-03
CAACGA	4.39e-05	2.12e-01	7.71e-02	7.70e-01
TCCTAC	4.55e-05	8.19e-01	5.80e-02	6.56e-01
CCTCTT	5.86e-05	5.95e-01	2.77e-01	3.86e-01
TCTTTT	7.84e-05	2.28e-01	6.10e-01	3.16e-01
GTGGCG	8.48e-05	4.70e-01	1.02e-01	3.29e-01
AAGAGG	1.25e-04	4.85e-01	8.73e-02	3.14e-01

GGAGAA	1.78e-04	1.50e-01	2.66e-01	8.02e-01
TCGTTG	2.20e-04	5.21e-01	5.99e-01	3.24e-01
CCTCAT	2.26e-04	8.60e-01	5.95e-02	7.47e-01
ATTACT	3.28e-04	2.34e-01	1.57e-01	7.61e-01
TCTTAT	4.40e-04	5.31e-01	5.47e-02	7.92e-01
GGGGAG	4.47e-04	4.91e-01	6.99e-02	3.80e-01
GTTGCT	4.56e-04	4.33e-01	2.30e-01	4.88e-01
AGCAAC	6.45e-04	8.25e-01	7.58e-02	3.14e-01
GGCGAC	6.50e-04	8.89e-01	1.43e-01	6.63e-01
GCCGAC	6.80e-04	5.01e-01	2.49e-02	9.37e-01
TTTTCT	7.53e-04	6.39e-01	3.63e-01	1.13e-02
CCCCAC	1.02e-03	4.40e-01	2.29e-02	6.83e-01
AGAAAA	1.12e-03	2.89e-01	2.05e-01	3.62e-01
GCTGAT	1.17e-03	9.90e-01	2.62e-02	4.46e-01
TACTTC	1.31e-03	2.53e-01	3.51e-01	4.55e-01
AAAAGA	1.47e-03	4.97e-01	3.09e-01	1.41e-01
AGTAAT	1.55e-03	3.10e-01	5.08e-02	6.56e-01
AAAATA	1.59e-03	8.95e-01	5.03e-01	1.83e-01
AATATT	1.60e-03	7.89e-01	2.73e-01	7.21e-01
AACAGC	1.60e-03	6.86e-01	1.94e-01	3.44e-01
ACCAAC	1.90e-03	2.98e-01	4.38e-02	8.62e-01
TTTTAT	2.33e-03	7.80e-01	3.18e-01	5.34e-01
CTACAA	3.21e-03	5.03e-02	5.22e-01	9.86e-01
ATTAAT	3.40e-03	5.33e-01	1.12e-01	4.44e-01
AATAGT	4.41e-03	6.40e-01	1.48e-01	5.72e-01
CTTCAT	4.54e-03	7.46e-01	1.84e-01	7.90e-01
GTTGAT	5.48e-03	8.80e-01	1.93e-01	3.46e-01
CCACAA	5.49e-03	5.58e-01	5.07e-02	3.58e-01
AGGAAG	7.49e-03	8.89e-01	1.09e-01	3.76e-01
GCGGAG	8.27e-03	9.09e-01	6.02e-01	4.32e-02
AAGATG	9.56e-03	7.93e-01	2.22e-01	9.46e-01
TCATAA	1.13e-02	3.38e-01	7.81e-02	8.40e-01
ATCAAC	1.22e-02	8.55e-01	4.37e-01	7.51e-02
GATGTT	1.23e-02	5.96e-01	2.24e-01	7.75e-01
TCGTAG	1.36e-02	9.45e-01	2.92e-02	5.55e-01
ACTAAT	1.64e-02	1.83e-01	4.99e-02	6.95e-01
GTAGAA	1.67e-02	7.89e-01	5.73e-01	3.52e-01
ACGATG	1.69e-02	5.77e-01	8.62e-01	4.61e-01
TAGTTG	1.87e-02	4.63e-01	6.04e-01	7.34e-01
AACATC	1.99e-02	5.13e-01	3.08e-01	3.02e-01
CGACAA	2.94e-02	9.90e-01	5.09e-02	1.62e-01
CGCCAC	5.80e-02	7.83e-02	1.39e-01	6.99e-01
TTATAA	6.37e-02	4.73e-01	3.52e-01	6.45e-01
GCGGTG	6.43e-02	8.55e-02	9.05e-01	2.97e-01
GCAGAA	6.97e-02	7.42e-01	4.66e-02	6.95e-01
ACGAAG	8.51e-02	6.97e-01	8.32e-01	3.89e-02
TAATTA	8.70e-02	8.10e-01	9.69e-01	4.41e-02
CGGCAG	1.13e-01	1.24e-02	5.47e-01	8.77e-01
TATTTT	1.31e-01	6.15e-01	6.01e-01	1.74e-01
ATAAAA	1.32e-01	9.40e-01	7.13e-01	6.57e-02
CGTCAT	1.54e-01	9.45e-01	2.84e-01	7.86e-01
CACCTC	1.68e-01	9.75e-01	1.96e-01	4.18e-01
GAAGTA	1.90e-01	3.47e-01	5.49e-01	3.82e-01
ACAAAA	2.60e-01	1.15e-01	1.56e-01	7.56e-01
TTCTAC	2.61e-01	2.27e-01	6.34e-01	2.86e-01
ATGAAG	2.93e-01	8.80e-01	2.62e-01	4.16e-01
CCGCTG	3.15e-01	4.14e-01	5.03e-01	8.59e-01
GAGGTG	3.25e-01	7.30e-01	2.38e-01	9.45e-01
CAACTA	3.53e-01	2.04e-01	4.38e-01	3.91e-01
CAGCTG	3.65e-01	8.10e-01	1.73e-01	7.75e-01
GACGTC	3.74e-01	9.07e-01	3.76e-01	4.40e-01
GTCGAC	4.40e-01	8.18e-01	3.05e-01	7.13e-01
TTGTAG	4.89e-01	5.76e-02	2.47e-01	7.65e-01
GTGGAG	5.38e-01	5.64e-01	1.18e-01	8.61e-01
CTGCAG	7.72e-01	6.50e-01	6.47e-02	5.03e-01
CATCTT	8.64e-01	2.21e-01	2.54e-01	8.40e-01
CTCCAC	9.82e-01	1.72e-01	2.38e-01	8.37e-01

Table A2: Linear model fit between mismatch load and age.

Triplet	R^2	P
AGAATA	0.63	9.74e-08
GGAGTA	0.55	1.98e-06
TCATGA	0.49	1.28e-05
CGACTA	0.44	5.04e-05
GGTGAT	0.42	8.34e-05
TCTTAT	0.41	1.02e-04
GCTGAT	0.4	1.40e-04
TGCTTC	0.39	1.60e-04
AGCATC	0.37	2.82e-04
GGCGTC	0.36	3.20e-04
TGATTA	0.36	3.33e-04
TGTTAT	0.31	1.08e-03
TCATAA	0.31	1.17e-03
TAGTGG	0.3	1.56e-03
TGATCA	0.29	1.75e-03
TATTGT	0.29	1.97e-03
ACTATT	0.28	2.41e-03
TACTGC	0.27	2.69e-03
AGGAAG	0.27	3.02e-03
CATCGT	0.26	3.31e-03
GACGCC	0.25	4.04e-03
MismatchLoad	0.25	4.23e-03
CCACTA	0.22	7.90e-03
TTATCA	0.22	8.29e-03
ATAACA	0.21	9.35e-03
ATAAGA	0.21	9.36e-03
TATTCT	0.21	9.97e-03
ACTAGT	0.19	1.41e-02
GCGGGG	0.19	1.44e-02
GAAGGA	0.19	1.55e-02
CCGCAG	0.18	1.58e-02
TTGTCG	0.18	1.69e-02
AAGAGG	0.18	1.76e-02
TCCTGC	0.18	1.77e-02
TGTTTT	0.18	1.88e-02
AGTAAT	0.17	2.08e-02
TCCTAC	0.16	2.75e-02
GATGGT	0.14	3.78e-02
TTATAA	0.14	3.85e-02
CCACGA	0.14	3.90e-02
ACAAAA	0.14	4.11e-02
GGTGTT	0.13	4.29e-02
AGAACA	0.13	4.56e-02
GGAGAA	0.13	4.59e-02
GAGGGG	0.13	4.60e-02
TAATTA	0.12	5.18e-02
TCGTGG	0.12	5.33e-02
ACAAGA	0.12	5.38e-02
ACTAAT	0.12	5.42e-02
AATATT	0.12	5.97e-02
CGACAA	0.12	6.19e-02
ATCAGC	0.11	6.27e-02
GACGGC	0.11	6.67e-02
ATTAGT	0.11	7.08e-02
TAATGA	0.1	7.88e-02
TGGTTG	0.1	7.98e-02
CCTCTT	0.1	7.99e-02
GCCGAC	0.1	8.31e-02
GAGGCG	0.1	8.37e-02
AGAAAA	0.099	8.43e-02
GGCGAC	0.096	8.95e-02
CATCTT	0.095	9.21e-02
TTCTGC	0.092	9.80e-02

GCTGTT	0.091	9.99e-02
GCAGTA	0.09	1.01e-01
CGCCCC	0.09	1.01e-01
TATTTT	0.09	1.02e-01
AAAACA	0.088	1.06e-01
TTTTCT	0.084	1.13e-01
CCGCGG	0.084	1.14e-01
CTTCCT	0.082	1.19e-01
ATGAGG	0.079	1.25e-01
GGAGCA	0.075	1.35e-01
AATACT	0.075	1.35e-01
CAGCCG	0.075	1.36e-01
GCCGTC	0.074	1.39e-01
CGTCCT	0.074	1.39e-01
GGCGCC	0.072	1.44e-01
GCAGAA	0.071	1.46e-01
GTGGGG	0.069	1.53e-01
GCCGGC	0.069	1.54e-01
GTCGGC	0.068	1.56e-01
CACCTC	0.068	1.57e-01
AAAATA	0.066	1.62e-01
TACTTC	0.065	1.66e-01
AGTACT	0.064	1.70e-01
TGCTCC	0.064	1.71e-01
ATGAAG	0.063	1.73e-01
CCTCAT	0.062	1.77e-01
ATTACT	0.062	1.78e-01
CTTCGT	0.06	1.83e-01
GCAGGA	0.059	1.89e-01
GAAGTA	0.058	1.93e-01
CAACCA	0.057	1.95e-01
CAACGA	0.057	1.96e-01
AATAGT	0.057	1.97e-01
AAAAGA	0.056	1.99e-01
TCCTTC	0.054	2.09e-01
GCTGGT	0.053	2.12e-01
CGGCTG	0.052	2.15e-01
CCCCGC	0.052	2.16e-01
TCTTTT	0.052	2.17e-01
GATGCT	0.049	2.33e-01
CTGCCG	0.047	2.44e-01
AACATC	0.046	2.46e-01
GCGGTG	0.045	2.54e-01
ACAATA	0.044	2.56e-01
CACCCC	0.043	2.61e-01
CGCCTC	0.039	2.84e-01
CTCCGC	0.039	2.86e-01
GATGTT	0.036	3.07e-01
ACCATC	0.034	3.20e-01
CACCGC	0.033	3.28e-01
CTGCGG	0.032	3.33e-01
GTTGGT	0.032	3.36e-01
TCGTAG	0.032	3.38e-01
CTACGA	0.032	3.39e-01
ATGACG	0.03	3.50e-01
AGGATG	0.026	3.85e-01
TAGTCG	0.026	3.90e-01
GGGGTG	0.024	4.00e-01
TTGTGG	0.024	4.01e-01
AACACC	0.024	4.10e-01
CAGCGG	0.023	4.14e-01
AGTATT	0.023	4.16e-01
CGCCAC	0.023	4.19e-01
CTTCAT	0.022	4.29e-01
ACGAAG	0.022	4.30e-01
AGCACC	0.021	4.38e-01
GTCGCC	0.02	4.46e-01
CTACCA	0.019	4.54e-01

TGGTAG	0.017	4.79e-01
TAGTTG	0.017	4.80e-01
GAGGTG	0.017	4.81e-01
GTCGAC	0.016	4.97e-01
ATCACC	0.016	5.00e-01
TTATGA	0.016	5.02e-01
CGTCAT	0.016	5.03e-01
ATAAAA	0.016	5.04e-01
TACTCC	0.015	5.12e-01
CTCCCC	0.015	5.15e-01
GTAGCA	0.015	5.17e-01
ATCAAC	0.013	5.34e-01
TTCTAC	0.013	5.37e-01
AAGACG	0.013	5.48e-01
ATTAAT	0.012	5.55e-01
AGCAAC	0.012	5.57e-01
CCCCAC	0.012	5.57e-01
TTTTGT	0.012	5.59e-01
GCGGAG	0.012	5.65e-01
ACGATG	0.011	5.73e-01
GTTGCT	0.011	5.75e-01
TTGTAG	0.011	5.78e-01
TGGTCG	0.01	5.93e-01
TGATAA	0.0097	5.99e-01
TCATTA	0.0095	6.01e-01
AGGACG	0.0094	6.05e-01
CGGCAG	0.0089	6.14e-01
AAGATG	0.0086	6.20e-01
AACAGC	0.0082	6.29e-01
GTGGCG	0.0079	6.34e-01
CTACAA	0.0073	6.47e-01
TCTTGT	0.0043	7.26e-01
CGACCA	0.0042	7.30e-01
ACGAGG	0.0039	7.39e-01
CCGCTG	0.0039	7.40e-01
TGTTCT	0.0036	7.49e-01
CAACTA	0.0033	7.59e-01
CCACAA	0.0033	7.59e-01
CTCCAC	0.0026	7.84e-01
GAAGCA	0.0026	7.85e-01
CTGCAG	0.0026	7.86e-01
TCGTTG	0.0019	8.15e-01
CGTCTT	0.0019	8.17e-01
TTTTAT	0.0017	8.24e-01
CCTCGT	0.0017	8.27e-01
GTTGAT	0.0015	8.36e-01
TGCTAC	0.0013	8.47e-01
GGGGCG	0.0012	8.54e-01
GTAGGA	0.0012	8.56e-01
TTCTCC	0.001	8.65e-01
ACCAAC	0.00092	8.71e-01
CCCCTC	0.0009	8.73e-01
CGGCCG	0.00075	8.84e-01
GACGTC	0.00072	8.86e-01
GTAGAA	0.00071	8.87e-01
GTGGAG	0.00053	9.02e-01
ACCAGC	0.00047	9.08e-01
TAATCA	0.00036	9.19e-01
GGGGAG	0.00011	9.56e-01
GGTGCT	3.4e-05	9.75e-01
CATCCT	1.9e-05	9.82e-01
CAGCTG	4.5e-06	9.91e-01

Table A3: Mismatch asymmetry based on mismatches falling with gene expression groupings.

Context	Expressed	Nonexpressed	Global	Mismatch type	Difference
AAAACA_TTTTGT	0.44	-0.041	0.32	AC	0.48
AAAAGA_TTTTCT	0.44	-0.0076	0.3	AG	0.44
AAAATA_TTTTAT	0.5	0.056	0.36	AT	0.45
AACACC_GTTGGT	0.24	0.2	0.18	AC	0.044
AACAGC_GTTGCT	0.12	0.0088	0.065	AG	0.11
AACATC_GTTGAT	0.23	0.12	0.17	AT	0.11
AAGACG_CTTCGT	0.22	-0.19	0.11	AC	0.41
AAGAGG_CTTCCT	0.21	-0.19	0.097	AG	0.4
AAGATG_CTTCAT	0.28	-0.1	0.17	AT	0.38
AATACT_ATTAGT	-0.084	-0.083	-0.11	AC	-0.00067
AATAGT_ATTACT	-0.19	-0.15	-0.19	AG	-0.032
AATATT_ATTAAT	-0.046	-0.0094	-0.055	AT	-0.037
ACAAAA_TGTTTT	2.7	2.5	2.7	CA	0.25
ACAAGA_TGTTCT	0.2	-0.033	0.12	CG	0.23
ACAATA_TGTTAT	-0.093	-0.35	-0.18	CT	0.26
ACCAAC_GGTGTT	2.4	2.5	2.4	CA	-0.049
ACCAGC_GGTGCT	0.046	0.083	0.047	CG	-0.037
ACCATC_GGTGAT	-0.029	-0.035	-0.041	CT	0.0055
ACGAAG_CGTCTT	0.82	0.38	0.73	CA	0.43
ACGAGG_CGTCCT	0.049	0.38	0.1	CG	-0.33
ACGATG_CGTCAT	0.73	0.8	0.67	CT	-0.07
ACTAAT_AGTATT	2.8	2.9	2.8	CA	-0.081
ACTAGT_AGTACT	-0.0088	-0.018	-0.025	CG	0.0094
ACTATT_AGTAAT	-0.61	-0.58	-0.62	CT	-0.036
CAACCA_TTGTGG	0.46	0.21	0.37	AC	0.25
CAACGA_TTGTCG	0.35	0.17	0.27	AG	0.18
CAACTA_TTGTAG	0.36	0.19	0.28	AT	0.17
CACCCC_GTGGGG	-0.18	-0.0063	-0.16	AC	-0.17
CACCGC_GTGGCG	-0.24	-0.17	-0.24	AG	-0.066
CACCTC_GTGGAG	-0.21	-0.13	-0.2	AT	-0.076
CAGCCG_CTGCGG	-0.099	-0.16	-0.11	AC	0.057
CAGCGG_CTGCCG	-0.14	-0.19	-0.16	AG	0.049
CAGCTG_CTGCAG	-0.074	-0.13	-0.089	AT	0.052
CATCCT_ATGAGG	-0.35	-0.0082	-0.27	AC	-0.34
CATCGT_ATGACG	-0.5	-0.17	-0.43	AG	-0.33
CATCTT_ATGAAG	-0.38	-0.047	-0.31	AT	-0.34
CCACAA_TGGTTG	0.72	0.8	0.73	CA	-0.077
CCACGA_TGGTCG	-0.23	-0.13	-0.21	CG	-0.11
CCACTA_TGGTAG	-0.33	-0.27	-0.33	CT	-0.066
CCCCAC_GGGGTG	0.66	0.78	0.67	CA	-0.12
CCCCGC_GGGGCG	-0.21	-0.11	-0.18	CG	-0.1
CCCCTC_GGGGAG	-0.29	-0.18	-0.28	CT	-0.11
CCGCAG_CGGCTG	-0.38	-0.26	-0.34	CA	-0.12
CCGCGG_CGGCCG	-0.53	-0.053	-0.44	CG	-0.48
CCGCTG_CGGCAG	0.0085	0.43	0.083	CT	-0.42
CCTCAT_AGGATG	1.1	1.4	1.2	CA	-0.34
CCTCGT_AGGACG	-0.33	0.027	-0.26	CG	-0.36
CCTCTT_AGGAAG	-0.74	-0.37	-0.66	CT	-0.36
GAAGCA_TTCTGC	0.24	-0.12	0.12	AC	0.36
GAAGGA_TTCTCC	0.24	-0.16	0.13	AG	0.4
GAAGTA_TTCTAC	0.33	-0.064	0.22	AT	0.4
GACGCC_GTCGGC	0.29	0.04	0.2	AC	0.25
GACGGC_GTCGCC	0.21	0.09	0.18	AG	0.12
GACGTC_GTCGAC	0.29	0.17	0.26	AT	0.12
GAGGCG_CTCCGC	0.33	0.095	0.26	AC	0.23
GAGGGG_CTCCCC	0.3	-0.034	0.23	AG	0.33
GAGGTG_CTCCAC	0.36	0.043	0.3	AT	0.32
GATGCT_ATCAGC	-0.11	-0.18	-0.12	AC	0.071
GATGGT_ATCACC	-0.17	-0.26	-0.18	AG	0.086
GATGTT_ATCAAC	-0.039	-0.12	-0.045	AT	0.077
GCAGAA_TGCTTC	0.74	0.63	0.71	CA	0.11
GCAGGA_TGCTCC	0.011	-0.056	-0.011	CG	0.067
GCAGTA_TGCTAC	0.047	-0.08	0.012	CT	0.13
GCCGAC_GGCGTC	0.54	0.59	0.55	CA	-0.056
GCCGGC_GGCGCC	-0.082	-0.036	-0.063	CG	-0.046
GCCGTC_GGCGAC	0.069	0.11	0.076	CT	-0.039

GCGGAG_CGCCTC	-0.79	-1.2	-0.87	CA	0.43
GCGGGG_CGCCCC	0.1	0.071	0.19	CG	0.034
GCGGTG_CGCCAC	1	1.2	1	CT	-0.17
GCTGAT_AGCATC	0.86	1	0.89	CA	-0.13
GCTGGT_AGCACC	-0.095	0.0014	-0.062	CG	-0.096
GCTGTT_AGCAAC	-0.33	-0.21	-0.3	CT	-0.11
TAATCA_TTATGA	-0.086	-0.26	-0.017	AC	0.18
TAATGA_TTATCA	-0.17	-0.24	-0.18	AG	0.07
TAATTA_TTATAA	-0.099	-0.19	-0.13	AT	0.087
TACTCC_GTAGGA	0.41	0.53	0.41	AC	-0.13
TACTGC_GTAGCA	0.25	0.28	0.24	AG	-0.038
TACTTC_GTAGAA	0.28	0.34	0.27	AT	-0.065
TAGTCG_CTACGA	-0.57	-0.57	-0.51	AC	0.0043
TAGTGG_CTACCA	-0.76	-0.77	-0.73	AG	0.011
TAGTTG_CTACAA	-0.54	-0.54	-0.52	AT	0.00061
TATTCT_ATAAGA	-0.22	-0.0046	-0.17	AC	-0.22
TATTGT_ATAACA	-0.31	-0.042	-0.25	AG	-0.27
TATTTT_ATAAAA	-0.21	0.04	-0.15	AT	-0.25
TCATAA_TGATTA	1.9	2.1	1.9	CA	-0.24
TCATGA_TGATCA	-0.18	0.1	-0.16	CG	-0.28
TCATTA_TGATAA	-0.41	-0.19	-0.37	CT	-0.22
TCCTAC_GGAGTA	1.2	1.6	1.3	CA	-0.37
TCCTGC_GGAGCA	-0.52	-0.13	-0.46	CG	-0.39
TCCTTC_GGAGAA	-0.63	-0.29	-0.56	CT	-0.35
TCGTAG_CGACTA	0.49	0.35	0.52	CA	0.14
TCGTGG_CGACCA	-0.44	0.19	-0.33	CG	-0.64
TCGTTG_CGACAA	-0.14	0.19	-0.11	CT	-0.34
TCTTAT_AGAATA	1.8	2.3	1.9	CA	-0.56
TCTTGT_AGAACA	-0.63	-0.13	-0.49	CG	-0.51
TCTTTT_AGAAAA	-1.1	-0.49	-0.91	CT	-0.56

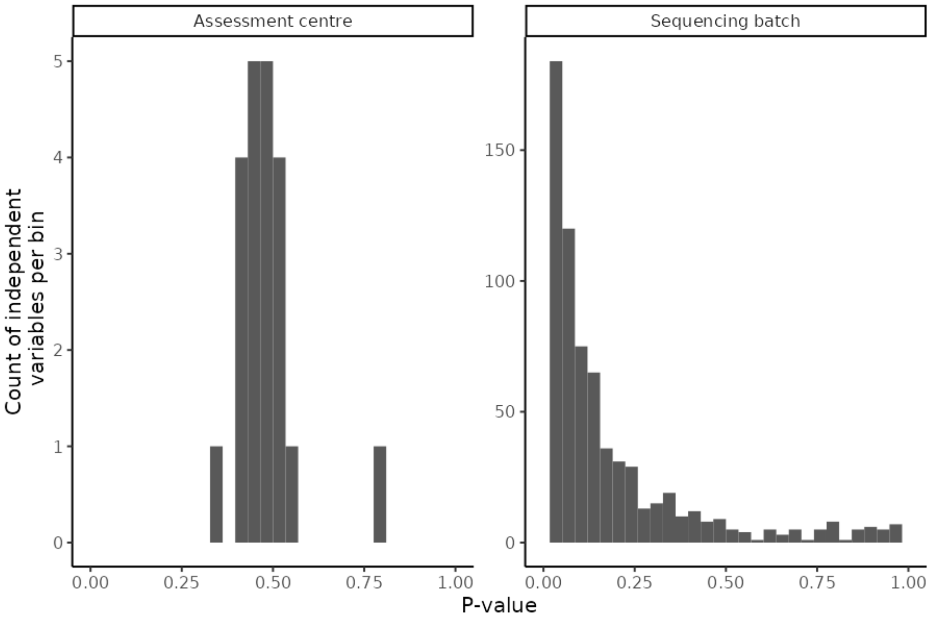

Figure A1: Distribution of p-values derived from the linear regression of median $\log_2$ ratio and assessment centre (n=22) or sequencing batch (n=816). For sequencing batch, there are a majority of batches that have statistical enrichment for association with DNA damage.

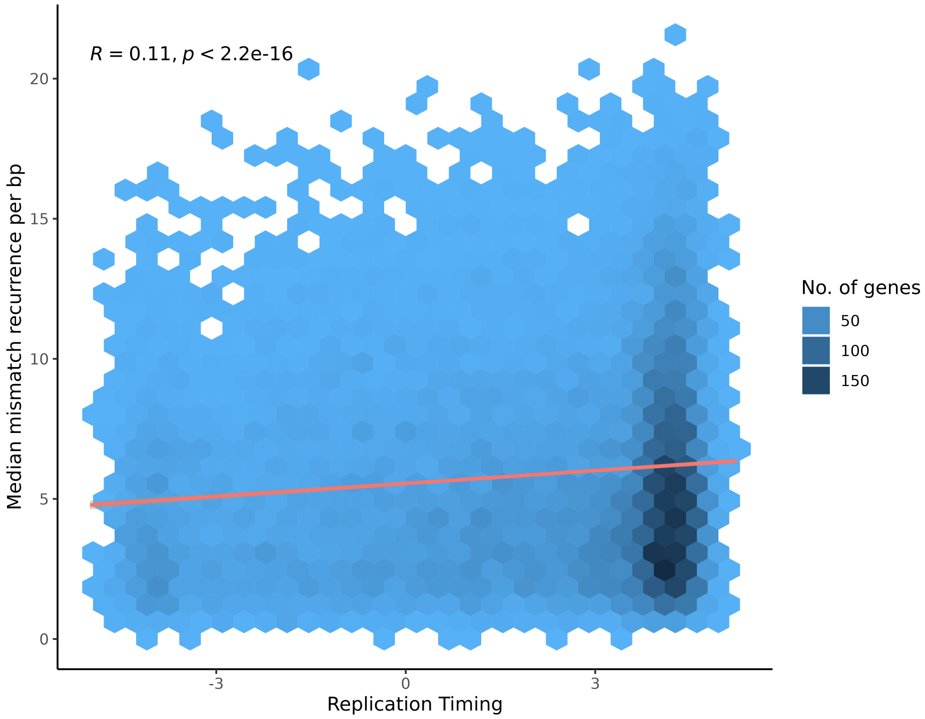

Figure B1: Hexbin plot of the median mutation recurrence as a function of replication timing ranked aggregated score. The red line shows the line fit from a linear model between mismatch recurrence as a function of replication timing.

Appendix B

Table B1: Spearman correlation tests for mismatch load and genomic covariates.

	GC P	GC ρ	Replication (T) P	Replication (T) ρ	Accessibility P	Accessibility ρ
TGATTA	1.2e-61	0.12	0.32	0.0076	0.0065	-0.031
TTGTGG	1.9e-28	0.084	0.46	-0.0056	0.024	-0.026
CGGCCG	7e-32	0.089	0.45	0.0057	0.022	-0.026
TAGTCG	1.7e-66	0.13	0.27	0.0085	6.6e-07	-0.057
GATGGT	4.6e-83	0.15	4.8e-14	0.058	0.016	-0.028
GCAGTA	2.8e-170	0.21	1e-43	0.11	4.6e-05	-0.047
AATATT	0.036	0.016	4.2e-33	-0.092	0.51	-0.0076
TTTTGT	4.4e-256	0.25	5.8e-49	0.11	1.1e-05	-0.05
ACCAAC	2.4e-99	0.16	0.18	-0.01	0.88	-0.0018
GGTGTT	4.9e-196	0.22	5.7e-19	0.068	0.0013	-0.037
TGCTCC	1.2e-172	0.21	3.1e-54	0.12	2e-08	-0.064
TTATCA	1.5e-26	0.081	0.29	-0.0081	0.14	-0.017
GCAGGA	2.3e-226	0.24	3.6e-12	0.053	5e-07	-0.058
TCCTTC	4.7e-79	0.14	0.2	0.0098	7e-05	-0.046
GACGGC	8.1e-165	0.21	1.8e-13	0.056	0.00027	-0.042
GCCGGC	7.2e-208	0.23	4e-58	0.12	1.2e-05	-0.05
GCAGAA	9.6e-158	0.2	2.1e-35	0.095	5.8e-05	-0.046
CCACAA	3.1e-43	0.1	0.0071	0.021	0.0056	-0.032
GTTGGT	1.9e-129	0.18	0.43	-0.0061	9.5e-05	-0.045
AGTACT	9.2e-66	0.13	0.13	-0.012	0.00019	-0.043
CCTCGT	6.1e-295	0.27	1.3e-23	0.077	1.6e-06	-0.055

CGACAA	2.2e-121	0.18	3.4e-34	0.093	2.5e-05	-0.048
GTCGAC	6e-39	0.099	3.5e-09	-0.045	0.071	-0.021
TGTTTT	6.6e-264	0.26	8.8e-16	0.062	2.5e-07	-0.059
TATTGT	1.7e-67	0.13	6.9e-13	0.055	0.17	-0.016
GCTGGT	1.2e-39	0.1	1.5e-12	-0.054	0.038	-0.024
GGAGTA	2.5e-109	0.17	2e-44	0.11	7e-07	-0.057
TCGTAG	1.9e-200	0.23	3.6e-10	0.048	7.2e-05	-0.045
CTCCCC	2.1e-71	0.13	0.97	-0.00027	0.00086	-0.038
GGTGAT	4.8e-151	0.2	7.4e-41	0.1	6.5e-06	-0.052
GTGGAG	1.9e-38	0.098	4.1e-17	0.064	0.0056	-0.032
TCATGA	5.2e-85	0.15	0.36	-0.007	0.035	-0.024
GAGGGG	1.1e-278	0.27	2.5e-49	0.11	8e-11	-0.074
GTAGCA	3.5e-102	0.16	2.9e-23	0.076	0.00091	-0.038
CCGCGG	2.4e-93	0.15	1.1e-09	-0.047	0.04	-0.024
TAATGA	7e-92	0.15	2.4e-24	0.078	8.8e-06	-0.051
ATCACC	5.1e-126	0.18	0.015	0.019	5e-06	-0.052
TCCTGC	5e-106	0.16	6.8e-25	0.079	0.0001	-0.045
ACCAGC	3.9e-255	0.25	4.8e-47	0.11	0.0019	-0.036
TTCTAC	4.1e-27	0.082	0.18	0.01	0.00075	-0.039
CTACCA	8.6e-292	0.27	9e-124	0.18	6.6e-17	-0.096
TTATAA	3.5e-104	0.16	2.1e-38	0.099	1.6e-05	-0.049
GACGCC	7.5e-238	0.25	3.7e-36	0.096	3.6e-05	-0.047
TTATGA	6.7e-56	0.12	0.015	-0.019	0.098	-0.019
GTAGGA	2.9e-102	0.16	0.076	-0.014	0.0038	-0.033
AGAACA	4.2e-111	0.17	0.0016	-0.024	0.26	-0.013
TTCTGC	3.1e-208	0.23	4.7e-56	0.12	0.00031	-0.041
TGATCA	8.7e-68	0.13	0.05	0.015	0.0051	-0.032
GATGCT	6.1e-119	0.17	2.1e-20	0.071	2.2e-07	-0.059
ATAAAA	1.1e-141	0.19	5.2e-42	0.1	0.0055	-0.032
CTACGA	8.8e-148	0.19	1.5e-07	-0.04	0.0038	-0.033
CGACCA	1.6e-154	0.2	0.089	0.013	0.094	-0.019
TCATTA	3.3e-148	0.19	1.3e-38	0.099	3.3e-10	-0.072
ATTAAT	1.7e-248	0.25	7.5e-83	0.15	8.5e-10	-0.07
ATCAGC	6.1e-81	0.14	0.0057	0.021	0.0025	-0.035
CCTCTT	1.5e-56	0.12	8e-06	0.034	0.0016	-0.036
TTTTCT	1.2e-75	0.14	0.005	-0.021	0.062	-0.021
CCGCAG	0	0.3	4.6e-45	0.11	5.5e-07	-0.057
AAAATA	1.3e-179	0.21	1.1e-25	0.08	1.2e-05	-0.05
CTTCCT	0	0.31	1e-118	0.18	2.8e-10	-0.072
GAGGCG	2e-100	0.16	0.0031	0.023	1e-08	-0.066
AGCATC	9.5e-117	0.17	8e-09	0.044	0.00028	-0.042
GCCGTC	8e-75	0.14	0.12	0.012	0.00014	-0.044
GCTGAT	1.9e-37	0.097	0.00021	-0.028	0.18	-0.015
CATCGT	4.3e-84	0.15	7.9e-38	0.098	4.7e-07	-0.058
ACAATA	4.4e-16	0.061	0.00018	-0.029	0.012	-0.029
ACAAGA	2.4e-59	0.12	0.00053	0.027	0.0001	-0.045
ACCATC	3.4e-105	0.16	9.9e-05	0.03	0.0099	-0.03
AAGAGG	3e-103	0.16	8.2e-30	0.087	4e-08	-0.063
TACTTC	4.7e-117	0.17	4.8e-40	0.1	1.2e-08	-0.065
CTGCGG	2.4e-144	0.19	3.6e-06	0.035	0.00061	-0.039
ACAAAA	1.6e-310	0.28	2.1e-80	0.14	1e-09	-0.07
GTTGAT	2.3e-135	0.19	6e-44	0.11	1.9e-05	-0.049
CTACAA	1.2e-141	0.19	2.5e-22	0.074	5.4e-06	-0.052
CCCCAC	1.4e-138	0.19	0.00068	0.026	0.2	-0.015
CGCCAC	4.4e-230	0.24	1.7e-48	0.11	3.2e-07	-0.059
GCGGGG	2.8e-85	0.15	9.7e-06	-0.034	0.0089	-0.03
TCTTTT	1e-149	0.2	1.3e-06	0.037	5.3e-05	-0.046
CCACGA	0.00011	0.029	4.8e-10	-0.048	0.014	-0.028
TTTTAT	1.2e-55	0.12	0.69	0.0031	0.11	-0.018
TGCTAC	9.8e-169	0.21	4.4e-42	0.1	0.00022	-0.042
TGATAA	5.2e-206	0.23	2.4e-56	0.12	1e-09	-0.07
CACCCC	3.3e-317	0.28	1.2e-16	0.063	0.00057	-0.039
ACTAAT	1.1e-160	0.2	9.9e-46	0.11	0.00035	-0.041
GTAGAA	3.6e-135	0.19	2.3e-41	0.1	0.00031	-0.041
TCGTTG	1.7e-263	0.26	1.6e-84	0.15	3.4e-08	-0.063
TGTTAT	4e-221	0.24	1.9e-59	0.12	5.4e-05	-0.046
CTTCAT	1.5e-104	0.16	1.4e-12	0.054	6.2e-07	-0.057
GTCGGC	6.1e-73	0.14	0.00018	0.029	1.1e-06	-0.056

AGAAAA	8.2e-136	0.19	5.2e-18	0.066	0.00024	-0.042
CCACTA	6.2e-51	0.11	0.0055	-0.021	0.38	-0.01
TCCTAC	6.6e-243	0.25	2.1e-42	0.1	5.9e-10	-0.071
ACGAGG	4e-98	0.16	0.52	-0.0049	0.53	-0.0072
AATAGT	2e-53	0.12	0.0052	0.021	1.5e-05	-0.05
TGGTCG	2.4e-271	0.26	8e-47	0.11	1.3e-08	-0.065
CGTCAT	4.1e-170	0.21	1.9e-20	0.071	0.0022	-0.035
CCCCTC	1.4e-152	0.2	1.7e-38	0.099	1.7e-05	-0.049
GTGGCG	4.1e-256	0.25	2.8e-83	0.15	3.6e-10	-0.072
CTGCAG	1.2e-194	0.22	4e-60	0.12	1.5e-06	-0.055
CGTCTT	3.8e-129	0.18	3.4e-31	0.089	0.0019	-0.036
GCGGTG	6.2e-182	0.22	4.2e-16	0.062	1.2e-05	-0.05
ATAACA	6.4e-56	0.12	0.7	-0.0029	0.019	-0.027
AGCACC	6.4e-116	0.17	4.9e-20	0.07	0.005	-0.032
TCATAA	4.9e-77	0.14	0.001	0.025	7.1e-06	-0.051
CGCCCC	1e-148	0.19	6.7e-28	0.084	9.5e-05	-0.045
ACTAGT	2.1e-302	0.28	3.6e-46	0.11	1.3e-07	-0.06
ACGAAG	1.6e-13	0.056	2.5e-21	-0.072	0.63	-0.0055
TACTGC	5.6e-17	0.063	1.9e-05	-0.033	0.023	-0.026
ATTACT	3.2e-183	0.22	1.2e-34	0.094	1.1e-07	-0.061
GTGGGG	1.5e-110	0.17	1.7e-25	0.08	3.3e-06	-0.053
CAGCTG	2.7e-133	0.18	4e-36	0.096	0.00074	-0.039
GCTGTT	1.6e-95	0.16	0.079	-0.013	0.027	-0.025
AGCAAC	4.2e-146	0.19	2e-13	0.056	0.00021	-0.042
TATTTT	2.3e-40	0.1	5.7e-06	0.035	0.061	-0.021
CAGCCG	2.8e-73	0.14	0.011	-0.019	0.22	-0.014
TTGTAG	2.8e-164	0.2	7.4e-76	0.14	1.1e-14	-0.088
ATTAGT	1.1e-42	0.1	0.0011	-0.025	0.029	-0.025
GGCGCC	0	0.32	2.3e-27	0.083	6.1e-06	-0.052
CCCCGC	1.2e-34	0.093	5.2e-22	-0.074	0.17	-0.016
CGACTA	1.3e-130	0.18	0.0088	-0.02	0.024	-0.026
CAACCA	2e-111	0.17	9.1e-24	0.077	0.0045	-0.033
TAGTTG	2e-255	0.25	3.2e-66	0.13	4.5e-06	-0.053
GTCGCC	0	0.33	5.5e-90	0.15	8.7e-05	-0.045
TGGTAG	3.1e-196	0.22	1.3e-37	0.098	0.0088	-0.03
TGTTCT	3.4e-115	0.17	7.2e-25	0.079	0.00016	-0.043
TATTCT	2.4e-184	0.22	2.7e-69	0.13	1.4e-06	-0.055
CGGCAG	5.2e-183	0.22	3.4e-13	0.056	0.00014	-0.044
CATCCT	7e-41	0.1	0.016	-0.019	0.023	-0.026
CTCCAC	0	0.29	6.1e-98	0.16	1.9e-09	-0.069
AGGAAG	4.9e-260	0.26	3.2e-65	0.13	4e-08	-0.063
GAGGTG	0	0.35	3.2e-182	0.22	5.4e-12	-0.079
TAGTGG	7.8e-92	0.15	2e-07	-0.04	0.017	-0.027
AACACC	2.2e-75	0.14	0.0039	-0.022	0.092	-0.019
CCTCAT	4.2e-114	0.17	1e-34	0.094	3.1e-10	-0.072
CACCTC	1.4e-252	0.25	2.6e-47	0.11	1.1e-06	-0.056
CGGCTG	5.3e-28	0.083	5.3e-07	-0.038	0.17	-0.016
TTGTCG	9.1e-117	0.17	0.46	-0.0056	0.00023	-0.042
AACATC	1.3e-161	0.2	4.8e-31	0.089	0.00066	-0.039
GGCGTC	0	0.29	2.7e-90	0.15	2.8e-06	-0.054
CTGCCG	5.4e-61	0.12	0.71	-0.0028	0.0023	-0.035
GAAGCA	1.1e-163	0.2	8.4e-30	0.087	4.9e-06	-0.052
GCCGAC	4.4e-192	0.22	2.8e-23	0.076	1.5e-05	-0.05
CTCCGC	1.9e-202	0.23	6.8e-54	0.12	7.2e-07	-0.057
CGCCTC	3.3e-63	0.13	0.21	-0.0097	0.021	-0.027
GCGGAG	2.3e-201	0.23	3.3e-21	0.072	5.6e-07	-0.057
GGAGCA	1.2e-241	0.25	5.2e-22	0.074	5.8e-08	-0.062
TACTCC	4e-94	0.15	9.6e-25	0.078	0.0022	-0.035
CCGCTG	4e-194	0.22	2.4e-26	0.081	3.3e-06	-0.053
AAAAGA	2e-83	0.15	1.2e-09	0.047	0.013	-0.029
ATAAGA	3.3e-98	0.16	4.5e-23	0.076	0.017	-0.027
ATGAGG	3.6e-114	0.17	2.6e-23	0.076	2.9e-05	-0.048
GGAGAA	8.1e-206	0.23	3.4e-52	0.12	2.6e-05	-0.048
ACTATT	5.8e-311	0.28	2.8e-66	0.13	5.1e-08	-0.062
AACAGC	2.6e-320	0.28	2.1e-21	0.073	4.6e-08	-0.063
AGAATA	7.9e-148	0.19	2.6e-16	0.063	0.00019	-0.043
AAGATG	8.3e-40	0.1	3.7e-12	-0.053	0.00037	-0.041
TGCTTC	4.5e-07	0.038	0.018	-0.018	0.096	-0.019

GAAGGA	0	0.3	4.9e-124	0.18	1.7e-06	-0.055
CTTCGT	5.9e-206	0.23	4.8e-24	0.077	1.3e-07	-0.06
AAAACA	5e-136	0.19	7e-10	0.047	0.0045	-0.033
GAAGTA	9e-143	0.19	3.9e-22	0.074	0.00012	-0.044
GATGTT	6e-294	0.27	8.8e-113	0.17	4.2e-10	-0.071
TCTTGT	0	0.3	4.9e-143	0.19	6.7e-09	-0.066
CAACTA	8.4e-115	0.17	1.4e-05	0.033	0.099	-0.019
ATGACG	3e-185	0.22	4.1e-16	0.062	8.7e-08	-0.061
CAGCGG	1.9e-227	0.24	5.5e-74	0.14	3.7e-09	-0.068
GGCGAC	3.7e-192	0.22	2.7e-24	0.078	5.4e-10	-0.071
GTTGCT	9.2e-317	0.28	4.5e-60	0.12	5.7e-13	-0.082
CACCGC	6.6e-109	0.17	4.6e-05	0.031	0.029	-0.025
AGGACG	2.9e-209	0.23	2.8e-16	0.063	0.034	-0.024
TAATCA	1.1e-58	0.12	0.34	-0.0074	0.099	-0.019
TTCTCC	2.5e-66	0.13	0.42	-0.0062	0.18	-0.015
TCGTGG	9.3e-116	0.17	0.00041	0.027	3.2e-05	-0.048
ACGATG	1e-178	0.21	5.3e-44	0.11	0.00043	-0.04
AGTAAT	1.5e-105	0.16	2.3e-24	0.078	6.1e-06	-0.052
AGGATG	9.8e-150	0.2	0.6	0.0041	0.036	-0.024
AGTATT	1.4e-310	0.28	3.1e-33	0.092	3.1e-05	-0.048
CAACGA	8.1e-122	0.18	6.1e-06	0.035	7.2e-05	-0.045
GGTGCT	3.5e-200	0.23	6e-12	0.053	0.064	-0.021
CGTCCT	0.28	-0.0081	6e-16	-0.062	0.41	-0.0095
GGGGCG	3.3e-95	0.16	3.4e-06	0.036	0.044	-0.023
ATGAAG	2e-33	0.091	0.65	-0.0035	0.00081	-0.038
GACGTC	1.2e-155	0.2	2.1e-36	0.096	0.00024	-0.042
AATACT	6.7e-60	0.12	2.9e-13	0.056	6.5e-06	-0.052
GGGGTG	3.3e-205	0.23	1.7e-21	0.073	3.2e-05	-0.048
TAATTA	6.1e-96	0.16	0.011	0.019	4.9e-05	-0.046
TCTTAT	1.4e-209	0.23	4.3e-51	0.11	4.5e-10	-0.071
AAGACG	6.9e-159	0.2	4.8e-13	0.055	3.4e-06	-0.053
ATCAAC	1.6e-204	0.23	7e-66	0.13	1.6e-09	-0.069
TGGTTG	1.5e-73	0.14	0.0017	0.024	0.0018	-0.036
CATCTT	2.9e-164	0.2	9.1e-21	0.071	2.5e-09	-0.068
GGGGAG	2.6e-285	0.27	4.3e-56	0.12	1.3e-08	-0.065
rank_mean	0	0.52	4.1e-102	0.16	7.5e-27	-0.12

Table B2: Nucleotide content proportions in recombination hotspots, the exome and genome wide.

	GC	As	Cs	Gs	Ts	Total Bases
Recombination hotspots	43.23	28.37	21.6	21.63	28.4	22778529
Exome	51.75	24.11	25.9	25.84	24.14	37495327
Genome	39.19	28.31	19.56	19.63	28	2824183054

Table B3: Linear model of mismatch load and recombination adjusted for chromatin structure and GC content.

Context	P	R^2	Effect size
GCTGAT	4.6e-83	0.072	2.7e+03
GCAGAA	3.8e-69	0.089	2.9e+03
ACCAAC	7.2e-54	0.062	1.9e+03
TCTTAT	4.1e-51	0.06	1.3e+03
ACAAAA	6.5e-50	0.081	2e+03
CAGCTG	9.8e-42	0.073	1.2e+03
GCCGAC	2.8e-40	0.053	2.3e+03
CACCTC	2.4e-37	0.074	1.2e+03
CCCCAC	1.8e-31	0.066	1.4e+03
GTGGAG	2.5e-27	0.033	1.4e+03
CCACTA	2.7e-27	0.042	9.7e+02
CCACAA	4e-27	0.051	1.7e+03
GCCGTC	7.4e-25	0.045	1.1e+03
ACTAAT	3.8e-24	0.071	1.3e+03
TGCTAC	2.1e-23	0.086	1.3e+03
TACTTC	5.9e-19	0.089	9.7e+02
TCATAA	1.5e-18	0.057	1.4e+03
GGCGAC	5.2e-17	0.08	1.3e+03
GCAGTA	2.3e-14	0.05	1.1e+03
GCGGTG	6.5e-14	0.062	9.7e+02
TCCTAC	2.7e-13	0.053	1e+03
CTCCGC	3.9e-13	0.083	-1.1e+03
TGATCA	4.5e-13	0.05	-8.9e+02
TCCTTC	1.9e-11	0.053	4.8e+02
GAGGCG	4.3e-11	0.07	-7e+02
TAATCA	4.6e-11	0.038	-7.2e+02
GTCGAC	1.6e-10	0.035	7.2e+02
GACGTC	2e-10	0.069	6.8e+02
GTAGAA	2.5e-10	0.053	7.5e+02
AGGAAG	4.8e-10	0.08	5.8e+02
CTGCAG	9.1e-10	0.079	8.2e+02
AAGAGG	9.5e-10	0.058	5.1e+02
TCTTGT	1.3e-09	0.092	-5.4e+02
CCTCAT	1.4e-09	0.061	9.1e+02
TAGTCG	1.6e-09	0.046	-6.9e+02
GAAGCA	4.6e-09	0.035	-7.1e+02
CTCCCC	5.2e-09	0.05	5e+02
GCGGGG	7.3e-09	0.027	-8.6e+02
GCGGAG	1.4e-08	0.063	9.6e+02
GAAGGA	3.1e-08	0.099	-7.7e+02
AGCAAC	4.3e-08	0.074	6.3e+02
AAGACG	5.5e-08	0.059	-7.9e+02
CAACCA	9.3e-08	0.046	-8.6e+02
TGGTAG	1e-07	0.04	8.7e+02
CGGCCG	3.2e-07	0.026	-6.7e+02
GGGGCG	4.2e-07	0.045	-8.8e+02

TCCTGC	5.1e-07	0.047	-8.3e+02
CGACAA	5.8e-07	0.054	6e+02
TAGTTG	5.9e-07	0.077	3.7e+02
TTATGA	9.6e-07	0.026	-6.4e+02
CTTCGT	1e-06	0.046	-5e+02
GGAGCA	1.5e-06	0.054	-3.2e+02
CCGCGG	1.8e-06	0.054	-4.6e+02
CGCCCC	3.1e-06	0.039	-5.3e+02
TTCTGC	3.7e-06	0.044	-4.6e+02
CCCCTC	4.2e-06	0.061	5.5e+02
ATAAGA	6.3e-06	0.051	-5.7e+02
TGTTTT	7.5e-06	0.077	-6e+02
TTTTGT	9.6e-06	0.07	-5.2e+02
TGCTTC	1.1e-05	0.008	3.9e+02
ACGAGG	1.2e-05	0.039	-6.7e+02
GGTGCT	1.2e-05	0.064	-5.7e+02
ACTATT	1.2e-05	0.1	-5.3e+02
TATTCT	2.3e-05	0.06	-5.8e+02
GATGCT	2.5e-05	0.05	-3.9e+02
AGCACC	2.6e-05	0.05	-6.7e+02
TACTCC	2.6e-05	0.041	-5.8e+02
GCTGGT	4e-05	0.025	-5.5e+02
CAGCCG	4.1e-05	0.04	-5.8e+02
AGTACT	4.2e-05	0.036	-4.1e+02
CTCCAC	5e-05	0.093	4.2e+02
TCATGA	5.7e-05	0.034	-4.6e+02
AGAACA	5.9e-05	0.035	-5.6e+02
CGGCAG	6.1e-05	0.059	6.3e+02
AGGACG	7.3e-05	0.029	-5.2e+02
GACGCC	7.8e-05	0.062	-4.2e+02
GCAGGA	8.5e-05	0.066	-3.4e+02
AATACT	0.00011	0.026	-2.8e+02
TTGTGG	0.00012	0.029	-3.8e+02
GGGGAG	0.00016	0.09	5e+02
AAAACA	0.00016	0.049	-4.5e+02
ACCAGC	0.00021	0.085	-4e+02
GGTGTT	0.00021	0.051	-5.7e+02
TGCTCC	0.00022	0.047	-4.1e+02
TAATGA	0.00025	0.041	-4.5e+02
GTTGGT	0.00025	0.056	-4.2e+02
TGATAA	0.00025	0.075	4.2e+02

Table B4: Linear model of mismatch load and gene expression adjusted for replication timing and GC content.

Context	P	Effect size	R^2
TGATTA	0.00028	-0.034	0.0019
TTGTGG	0.044	-0.019	0.0021
CGGCCG	0.008	-0.032	0.0027
TAGTCG	0.35	-0.011	0.006
GATGGT	0.00017	-0.035	0.012
GCAGTA	0.22	-0.016	0.03
AATATT	9.4e-07	-0.047	0.013
TTTTGT	0.88	0.0018	0.027
ACCAAC	0.024	-0.026	0.015
GGTGTT	0.76	0.0037	0.019
TGCTCC	0.33	0.012	0.012
TTATCA	0.4	0.011	0.00089
GCAGGA	0.0092	-0.025	0.012
TCCTTC	0.00024	-0.041	0.0014
GACGGC	0.0054	-0.033	0.014
GCCGGC	0.7	-0.0041	0.02
GCAGAA	0.72	0.0047	0.018
CCACAA	0.025	-0.03	0.0017

GTTGGT	0.32	0.013	0.015
AGTACT	1.3e-05	-0.047	0.006
CCTCGT	0.37	-0.0092	0.028
CGACAA	0.4	0.012	0.021
GTCGAC	0.02	-0.033	0.0097
TGTTTT	0.85	0.0024	0.028
TATTGT	0.028	-0.022	0.0011
GCTGGT	0.098	-0.021	0.0081
GGAGTA	0.12	-0.02	0.0045
TCGTAG	0.89	0.0019	0.013
CTCCCC	0.0062	-0.029	0.0058
GGTGAT	0.013	-0.031	0.0065
GTGGAG	5.9e-09	-0.055	0.0042
TCATGA	0.0079	-0.031	0.005
GAGGGG	0.45	0.0098	0.032
GTAGCA	0.029	-0.026	0.012
CCGCGG	0.55	-0.0082	0.021
TAATGA	0.51	0.0078	0.0027
ATCACC	0.45	-0.0091	0.0082
TCCTGC	0.083	-0.022	0.0077
ACCAGC	0.41	0.0088	0.013
TTCTAC	0.3	-0.013	0.00053
CTACCA	0.029	0.027	0.031
TTATAA	0.43	-0.0098	0.0056
GACGCC	0.22	0.014	0.02
TTATGA	0.34	-0.011	0.0018
GTAGGA	0.49	-0.0084	0.011
AGAACA	0.33	-0.012	0.018
TTCTGC	0.00019	-0.033	0.036
TGATCA	0.2	-0.017	0.0046
GATGCT	0.34	-0.013	0.013
ATAAAA	0.065	-0.022	0.011
CTACGA	0.011	-0.027	0.027
CGACCA	0.037	0.028	0.028
TCATTA	3.9e-09	-0.056	0.008
ATTAAT	0.15	-0.016	0.017
ATCAGC	0.69	-0.0049	0.006
CCTCTT	0.013	-0.027	0.0014
TTTTCT	0.17	0.018	0.0031
CCGCAG	0.66	0.0067	0.031
AAAATA	0.046	-0.025	0.0096
CTTCCT	0.34	0.013	0.046
GAGGCG	0.5	0.0085	0.0077
AGCATC	0.91	0.0017	0.017
GCCGTC	0.0062	-0.035	0.0083
GCTGAT	1.3e-10	-0.066	0.0042
CATCGT	0.0007	-0.038	0.004
ACAATA	0.00028	-0.044	0.0039
ACAAGA	0.0045	-0.032	0.0016
ACCATC	0.041	-0.023	0.012
AAGAGG	0.027	-0.027	0.00082
TACTTC	0.67	-0.0052	6.6e-05
CTGCGG	3e-07	-0.051	0.018
ACAAAA	0.023	0.027	0.04
GTTGAT	0.31	0.012	0.0029
CTACAA	0.14	-0.019	0.013
CCCCAC	0.0002	-0.041	0.018
CGCCAC	0.23	0.02	0.035
GCGGGG	0.013	0.033	0.02
TCTTTT	0.72	-0.0046	0.016
CCACGA	0.0023	-0.035	0.012
TTTTAT	0.084	-0.023	0.0061
TGCTAC	0.015	-0.032	0.0088
TGATAA	0.93	-0.0011	0.01
CACCCC	0.55	-0.0064	0.05
ACTAAT	0.035	-0.022	0.003
GTAGAA	0.031	-0.027	0.0096
TCGTTG	0.23	0.016	0.025

TGTTAT	0.08	-0.018	0.023
CTTCAT	0.041	-0.024	0.0051
GTCGGC	0.95	0.00067	0.00083
AGAAAA	0.13	-0.019	0.014
CCACTA	2e-06	-0.052	0.0059
TCCTAC	0.0096	-0.028	0.031
ACGAGG	0.38	-0.012	0.0085
AATAGT	0.54	0.0086	0.0025
TGGTCG	0.54	-0.0071	0.032
CGTCAT	0.0019	0.051	0.025
CCCCTC	0.046	-0.026	0.022
GTGGCG	0.36	-0.011	0.029
CTGCAG	0.013	0.034	0.03
CGTCTT	0.98	-0.00028	0.014
GCGGTG	0.0013	0.045	0.014
ATAACA	0.51	0.0078	0.0057
AGCACC	0.092	-0.022	0.015
TCATAA	0.098	-0.022	0.012
CGCCCC	0.54	0.0079	0.011
ACTAGT	0.73	0.0039	0.042
ACGAAG	0.22	-0.015	0.0077
TACTGC	0.0006	-0.041	0.0024
ATTACT	0.54	-0.0067	0.012
GTGGGG	0.63	-0.0056	0.0043
CAGCTG	0.0012	-0.037	0.013
GCTGTT	0.57	-0.0074	0.013
AGCAAC	2.4e-09	-0.059	0.0032
TATTTT	0.061	-0.018	0.00036
CAGCCG	0.39	-0.011	0.01
TTGTAG	0.097	-0.021	0.02
ATTAGT	0.42	0.011	0.0053
GGCGCC	0.39	0.011	0.049
CCCCGC	5.1e-07	-0.046	0.0075
CGACTA	0.63	-0.0055	0.0085
CAACCA	0.6	-0.0064	0.0068
TAGTTG	0.82	-0.0025	0.013
GTCGCC	0.018	-0.026	0.033
TGGTAG	0.31	-0.014	0.033
TGTTCT	0.02	-0.029	0.016
TATTCT	0.0092	-0.029	0.01
CGGCAG	0.66	-0.0056	0.022
CATCCT	0.47	-0.0079	0.0046
CTCCAC	2.5e-05	-0.045	0.025
AGGAAG	7.5e-07	-0.055	0.014
GAGGTG	0.58	0.0072	0.078
TAGTGG	0.049	-0.021	0.0068
AACACC	0.95	-0.00083	0.012
CCTCAT	8.9e-06	-0.057	0.0027
CACCTC	0.29	-0.012	0.024
CGGCTG	0.96	-0.00063	0.0072
TTGTCG	0.038	-0.026	0.013
AACATC	0.022	-0.028	0.0096
GGCGTC	0.41	0.0096	0.039
CTGCCG	0.4	-0.012	0.0057
GAAGCA	0.16	-0.017	0.023
GCCGAC	0.0017	-0.03	0.037
CTCCGC	0.24	0.016	0.023
CGCCTC	0.18	-0.016	0.007
GCGGAG	0.003	0.041	0.021
GGAGCA	0.014	-0.022	0.0089
TACTCC	0.1	-0.022	0.014
CCGCTG	0.27	0.017	0.026
AAAAGA	0.23	-0.013	0.0068
ATAAGA	0.018	-0.029	0.0069
ATGAGG	0.8	0.0035	0.018
GGAGAA	0.00044	-0.041	0.023
ACTATT	0.011	-0.03	0.016
AACAGC	0.13	-0.016	0.031

AGAATA	0.24	-0.014	0.012
AAGATG	0.0018	-0.031	0.0024
TGCTTC	0.018	-0.029	0.0019
GAAGGA	0.97	0.00035	0.041
CTTCGT	0.046	-0.021	0.028
AAAACA	0.0068	-0.032	0.012
GAAGTA	0.047	-0.021	0.0025
GATGTT	0.54	0.0073	0.027
TCTTGT	0.34	-0.01	0.044
CAACTA	0.53	-0.0076	0.003
ATGACG	0.64	0.0064	0.033
CAGCGG	0.29	-0.013	0.028
GGCGAC	0.61	0.0066	0.025
GTTGCT	0.016	0.032	0.03
CACCGC	0.0068	-0.031	0.0065
AGGACG	0.72	0.0043	0.023
TAATCA	0.76	-0.0039	0.0069
TTCTCC	0.14	-0.017	0.007
TCGTGG	0.7	-0.0052	0.01
ACGATG	0.00087	0.048	0.015
AGTAAT	0.42	-0.0089	0.01
AGGATG	0.76	-0.004	0.017
AGTATT	0.73	0.0035	0.036
CAACGA	0.86	0.0024	0.02
GGTGCT	0.14	-0.016	0.044
CGTCCT	0.0057	-0.04	0.0046
GGGGCG	0.047	-0.024	0.0077
ATGAAG	0.47	-0.0095	0.00041
GACGTC	8.6e-08	-0.051	0.01
AATACT	0.021	-0.024	0.00049
GGGGTG	0.98	-0.00027	0.015
TAATTA	0.78	0.0034	0.0017
TCTTAT	0.0002	-0.043	0.015
AAGACG	0.022	-0.027	0.0062
ATCAAC	0.076	-0.019	0.003
TGGTTG	0.0012	-0.042	0.0042
CATCTT	0.23	-0.013	0.013
GGGGAG	0.41	-0.012	0.027
rank_mean2	3.1e-23	-0.013	0.16

Table B5: dNdS results for the top 20 ranked haematological genes.

Gene name	N-syn	N-mis	ω-mis	P-mis	CDS length	P-adj
SRSF2	30146	189634	2.322876	-11244.9547	666	-16.884316
U2AF2	83303	505291	1.940647	-18337.7656	1428	-12.841573
NPM1	10133	47335	1.490590	-726.8417	885	-0.821290
DDX41	122384	545952	1.470771	-8056.4894	1923	-4.189542
STAT3	208089	909437	1.409478	-10700.6964	2313	-4.626328
DHX15	255217	960539	1.393180	-11797.8082	2388	-4.940456
ABL1	229318	948077	1.367368	-9621.0455	3450	-2.788709
IDH2	83657	380479	1.351408	-3315.8351	1359	-2.439908
WT1	67222	287218	1.349230	-2600.4189	1554	-1.673371
DNMT3A	118109	509999	1.340271	-4369.6349	2739	-1.595339

SF3B1	418724	1469848	1.316580	-12954.7957	3915	-3.309015
KRAS	38502	151979	1.311895	-1197.1147	570	-2.100201
NOTCH1	367914	1639304	1.307178	-11399.9642	7668	-1.486693
MYD88	69599	253709	1.305917	-2047.8061	954	-2.146547
XPO1	361907	1414859	1.295563	-10161.1907	3216	-3.159574
ETV6	104629	411400	1.292756	-2894.4140	1359	-2.129812
CALR	65982	321637	1.279577	-1759.3411	1254	-1.402983
MAP2K1	75993	294054	1.277377	-1900.8293	1182	-1.608147
RPS14	43185	152801	1.276941	-1055.7144	456	-2.315163
BRAF	157443	549267	1.261917	-3457.1107	2424	-1.426201

Appendix C

Theory

The following supplement gives more technical arguments that conditioning on a polygenic gene score, that is constructed from SNPs on off-target chromosomes, selected for signficance of association with the outcome, improves statistical power while conserving type I error in a standard linear mixed model. To simplify arguments, we will compare the two approaches fastGWA and fastGWA-PGS. The argument will be in 3 stages. First, we derive an expression for the variance of the association estimate when the PGS is not adjusted for. Second, we derive an expression for the variance of the association estimate in the PGS adjusted model, under the assumption that the genetic and environomental residuals remain independent of the target SNP genotype conditional on the PGS - importantly, this independence condition also implies that the association parameter being estimated is the same in the models with and without adjustment for PGS. This PGS-adjusted association will be seen to have smaller variance than the corresponding estimator from the unadjusted model. Finally, we argue this independence condition (of the residuals in the PGS-adjusted model and SNP genotype) is approximately true assuming that the PGS is statistically independent of the selected SNP. In practice, the PGS should be approximately independent of the selected SNP under the null hypothesis of no causal association at the target SNP, since the off-target SNPs that constitute the polygenic score are selected independently and are on differing chromosomes (that is they are not in LD with the target SNP and there is no-collider bias between the target SNPs and off-target SNPs since the null hypothesis is true). This proves the conservation of type I error. Under the alternative hypothesis that the target SNP has a causal association with the outcome, collider bias might result in some correlation between the PGS and target SNP genotype; however, the extent of this correlation is likely extremely weak when there are a large number of variants that are associated with the trait in question, and unlikely to invalidate the following argument.

We first list the assumptions and notation we will use for the remainder of the argument.

Assumptions

- Let X correspond to the standardized SNP genotype at a particular location
- Without loss of generality, assume that $Var(X) = 1$ and $E(X) = 0$ (that is if X^* is the original genotype data, $X = (X^* - E(X^*))/SD(X^*)$
- Similarly, the outcome Y is standardized, so that $E(Y) = 0$ and $Var(Y) = 1$
- Data collected on outcome, Y, target SNP X, and offtarget genetic SNPs, $G_1, ..., G_K$ for samples $i = 1 \dots N$
- The estimated LOCO polygenic score, $\hat{P} = \sum_{k \in \hat{S}} \hat{\beta}_k G_k$, constructed over SNPs in the selection set $\hat{S}$. Again SNP variables G_k for $k \in S$ are standardized to have mean 0, variance 1. By construction, $\hat{P}$ has expected value 0. We assume that $\hat{\beta}_k$ are scaled so that the empirical variance of $\hat{P}$ over samples $i \leq N$ is 1.
- Finally, we consider the LOCO polygenic score P that corresponds to SNPs in S but weighted according to their "true" associations β_k, $P = \sum_{k \in \hat{S}} \beta_k G_k$
- Subscript notation. i and j refer to individuals $i, j \leq N$; $k \leq K$ refers to genetic location

Variance of $\hat{\beta}$ in fastGWA model

The fastGWA model takes the form:

$$Y = \beta X + g^{(0)} + \varepsilon^{(0)} \tag{1}$$

where $Var(\varepsilon_1^{(0)}, ..., \varepsilon_N^{(0)})$= $\sigma_0^2 I$ and $Var(\mathbf{g^{(0)}})$ = $Var(g_1^{(0)}, ..., g_N^{(0)})$ = $\Pi\tau_0^2$, where the family matrix Π is assumed known (or can be estimated using the original genotypes). The overall variance matrix of $Var(\mathbf{Y}) = (Y_1, ..., Y_N)$ in (1) accounting for both the environmental variance and genetic random effect is $V = \sigma_0^2 I + \Pi\tau_0^2$ Assuming consistent REML estimates,$\hat{\tau}_0$ and $\hat{\sigma}_0$, of τ_0 and σ_0, estimated by fastGWA, fastGWA estimates β by genalized least squares:

$$\hat{\beta} = \mathbf{X}^t \hat{V}^{-1} \mathbf{Y}$$

Since, $\hat{\beta}$ is computed using generalized least squares, it is easily shown that:

$$Var(\hat{\beta}) = (\mathbf{X}^t \hat{V}^{-1}\mathbf{X})^{-1}$$

with $\mathbf{X}$ being the vector of the target SNP over $i = 1, ..., N$

Henceforth, we will assume that estimation error in the estimated variance components: $\hat{\sigma}_0$ and $\hat{\tau}_0$ is negligible, so can effectively leave out the hat-notation when referring to variance components.

To examine the effect of the extent of family correlation structure on $Var(\hat{\beta})$ in a simplistic setting, we will assume that Π has a compound symmetry structure (implying that all individuals are equally related. That is

$$\Pi = \rho\mathbf{J} + (1-\rho)\mathbf{I}$$

where J is the $N \times N$ matrix of 1's. That is Π has elements $-1 \leq \rho \leq 1$ on its off diagonals and 1 on its diagonals. It follows that the matrix V has also a compound symmetry form:

$$V = \rho\tau_0^2\mathbf{J} + ((1-\rho)\tau_0^2 + \sigma_0^2)\mathbf{I}$$

The inverse of V (if it exists) can be calculated analytically and is equal to:

$$V^{-1} = \mathbf{I}/((1-\rho)\tau_0^2 + \sigma_0^2) - \mathbf{J}\frac{\rho\tau_0^2}{((1-\rho)\tau_0^2 + \sigma_0^2)((1-\rho)\tau_0^2 + \sigma_0^2 + N\rho\tau_0^2)}$$

It follows that:

$$Var(\hat{\beta}) = (\mathbf{X}^t\hat{V}^{-1}\mathbf{X})^{-1} = [\frac{\sum_{i \leq N} X_i^2}{(1-\rho)\tau_0^2 + \sigma_0^2} - \frac{\sum_{i,j \leq N} X_i X_j \rho\tau_0^2}{((1-\rho)\tau_0^2 + \sigma_0^2)((1-\rho)\tau_0^2 + \sigma_0^2 + N\rho\tau_0^2)}]^{-1}$$

Now, noting that $E(X_i^2)$ =1 and assuming that $E(X_iX_j) = \rho$, the genetic correlation, for large N one can show that the above is approximately equal to

$$Var(\hat{\beta}) = \frac{\sigma_0^2 + (1-\rho)\tau_0^2}{N(1-\rho)} \qquad (2)$$

indicating that $Var(\hat{\beta})$ is smallest when fastGWA is run on unrelated individuals, that is where $\rho = 0$. From this, we see that the inclusion of a genetic-random effect (with a particular correlation matrix) in fastGWA does little to increase power (although the association estimate will be slightly more efficient than the corresponding estimate from a regression not taking into account family structure when $\rho \neq 0$. The goal in Fast-GWA is instead to properly incorporate family structure in the estimation of $Var(\hat{\beta})$. In particular, related-ness in the GWAS reduces the power of finding associated SNPs (which is indicated in that $Var(\hat{\beta})$ is a increasing function of ρ).

Variance of $\hat{\beta}$ in fastGWA-PGS model

The fastGWA-PGS model takes the form:

$$Y = \beta X + g^{(1)} + \gamma \hat{P} + \varepsilon^{(1)} \tag{3}$$

where $\hat{P} = P + \varepsilon_P$ is the estimated polygenic risk score, assumed to be independent of X, and estimated in a LOCO fashion. We will later justify that the modified residual terms $\varepsilon^{(1)}$ and $g^{(1)}$, are zero mean random variables that are independent of X conditional on $\hat{P}$ provided $\hat{P}$ is independent of X. Comparing with equation (1) we have that:

$$Var(\varepsilon^{(0)}) + Var(g^{(0)}) = Var(\varepsilon^{(1)}) + Var(g^{(1)}) + \gamma^2 \tag{4}$$

Importantly, these independence conditions imply that conditional on $\hat{P}$, $Cov(X,Y|\hat{P}) = \beta Var(X|\hat{P}) = \beta Var(X)$. Noting then that $Cov(X,Y|\hat{P})$ is constant, it must equal $Cov(X,Y)$, which implys that $\beta = Cov(X,Y)/Var(X)$. This indicates that the coefficient β multiplying the SNP genotype is the same in (3) and (1). Note that the variances of both residual terms may be reduced due to addition of the polygenic risk score, that is $Var(\varepsilon^{(1)}) = \sigma_1^2 < Var(\varepsilon^{(0)}) = \sigma_0^2$ and $Var(g^{(1)}) = \tau_1^2 < Var(g^{(0)}) = \tau_0^2$. As vector equations we again assume that $Var(\varepsilon_1^{(0)},...,\varepsilon_N^{(1)})= \sigma_1^2 I$ and $Var(\mathbf{g^{(1)}}) = Var(g_1^{(1)},...,g_N^{(1)}) = \Pi\tau_1^2$. Comparing equations (1) and (3), it follows that adjustment for the polygenic score will reduce the variance of the environmental noise and genetic components in (1), by the quantities: $Corr(\hat{P},\varepsilon^{(0)})$ and $Corr(\hat{P},g^{(0)})$. Note if we instead adjusted for the ”true” polygenic score, P, in the regression, we might reduce more of the noise in the genetic random effect but would not reduce noise in the environmental random effect.

The model can be approximately fit in 2 stages. First, we orthogonalize the outcome, Y with respect to $\hat{P}$. That is we set $Y^{(1)} = Y - Y_{\hat{P}} = Y - \hat{\gamma}\hat{P}$, where $Y_{\hat{P}}$ is the predicted outcome from a regression using $\hat{P}$. Second, we orthogonalize X with respect to $\hat{P}$, that is calculate $X^{(1)} = X - X_{\hat{P}}$. Assuming X is truly independent of $\hat{P}$ one would expect that $X^{(1)} \sim X$. Finally, β is estimated by a generalized least squares fit, regressing $Y^{(1)}$ on $X^{(1)}$, in the following model

$$Y^{(1)} = \beta X^{(1)} + g^{(1)} + \varepsilon^{(1)} \tag{5}$$

where the variance matrix

$$V^{(1)} = \sigma_1^2 + \Pi\tau_1^2. \tag{6}$$

Similarly to before, $\hat{\beta} = \mathbf{X^{(1)}}^t \mathbf{V}^{(1)-1}\mathbf{Y^{(1)}}$ and the variance of $\hat{\beta}$ is

$$Var(\hat{\beta}) = [\mathbf{X}^{(1)t}\mathbf{V}^{(1)-1}\mathbf{X}^{(1)}]^{-1} \tag{7}$$

and under the circumstance that the off-diagonal elements of Π are all equal to ρ, and $X^{(1)} \sim X$, this is approximately

$$Var(\hat{\beta}) = \frac{\sigma_1^2 + (1-\rho)\tau_1^2}{N(1-\rho)} \tag{8}$$

noting that $\sigma_1^2 < \sigma_0^2$ and $\tau_1^2 < \tau_0^2$ and comparing to (2) indicates the variance of $\hat{\beta}$ is reduced by adding the informative (and independent) estimated PGS to the regression. Because of near-orthogonality of X and $\hat{P}$, one would not expect the absolute-size of $\hat{\beta}$ to be altered (indeed we argued previously that the β coefficient in the two regression formulae (1) and (5) should be equal), indicating that a test based on $\hat{\beta}^2/Var(\hat{\beta})$ should have improved power.

Justification of independence of modified residuals and SNP genotype X under approximate independence of X and $\hat{P}$

As previously noted, if residuals, $\varepsilon^{(1)}$ and $g^{(1)}$ and genotype, X, in equation (3) are truly independent of each other, and $\varepsilon^{(1)}$ and $g^{(1)}$ are zero mean and finite variance, standard calculations as demonstrated later show that the variance calculated as (7) is asymptotically correct. In addition, the β parameters will 'match' in equations (1) and (3), and hence the PGS adjusted model will have improved power under the alternative whilst conserving type I error under the null. The following is an argument to justfify this condition. By assumption, in equation (1), the residual terms $\varepsilon^{(0)}$ and $g^{(0)}$ are independent of the genotype vector X. We also have assumed that the selected polygenic score, $\hat{P}$ is statistically independent of X. This implies that once standardized to have mean 0, X and $\hat{P}$ should be approximately orthogonal. Now, conditional on the vector of polygenic scores, $\hat{\mathbf{P}}$ Let $Y_{\hat{P}} = \hat{\gamma}\hat{\mathbf{P}}$ be the projection of the response vector Y onto the vector $\hat{\mathbf{P}}$. By examining the right hand side of equation (1), and the approximate orthogonality of X and $\hat{\mathbf{P}}$, this projection is also equal to the sum of the projections of the vectors $\varepsilon^{(0)}$ and $g^{(0)}$ onto $\hat{\mathbf{P}}$, which we denote $\varepsilon^{(0)}_{\hat{P}} + \gamma^{(0)}_{\hat{P}}$. Now denoting $\varepsilon^{(1)} = \varepsilon^{(0)} - \varepsilon^{(0)}_{\hat{P}}$ and $g^{(1)} = g^{(0)} - g^{(0)}_{\hat{P}}$, we have the equation:

$$Y_i - \hat{\gamma}\hat{P}_i \approx \beta X_i + \varepsilon^{(1)} + g^{(1)} \tag{9}$$

where β is the same coefficient as in equation (1). Noting that conditional on $\hat{\mathbf{P}}$, the vectors $\varepsilon^{(1)}$ and $g^{(1)}$ are functions of the vectors $\varepsilon^{(0)}$ and $g^{(0)}$, which are all independent of X, $\varepsilon^{(1)}$ and $g^{(1)}$ are also independent of X. In addition, $\varepsilon^{(0)}$, $g^{(0)}$ and $\hat{P}$ are 0-mean random variables by assumption. Since, as vectors $\varepsilon^{(1)}$ and $g^{(1)}$ can be viewed as the difference of a zero mean vector and a projection onto a zero mean vector they can also be viewed as zero mean vectors, which completes the argument.

Conservation of Type I error, after adjustment for $\hat{P}$, assuming independence of X and modified residuals

Under the scenario that we have sucessfully reduced residual noise by incorporating a polygenic risk score as above, the association test checks the orthogonality of the genotype vector for the SNP, X with the noise reduced outcome vector (after subtracting off the predicted outcome based on the polygenic score). Since the polygenic risk score is approximately othogonal to the SNP in question, and was constructed with no reference to the SNP, the Type I error of this test should not be affected. This follows in a straightforward way from the observations that the modified genetic and environmental residuals are independent of $\mathbf{X}$ and have 0 mean and the variance matrix listed above as we have justified above.

In more detail, suppose that $\beta = 0$. If $E(\hat{\beta}) = 0$ and the variance of $Var(\hat{\beta})$ is really given by (7), it follows that the test statistic: $\hat{\beta}^2/Var(\hat{\beta})$ should be approximately chi-squared with 1 degree of freedom, and p-values will be uniform as required for a valid statistical test.

First $E(\hat{\beta}) = E(\mathbf{X}^{(\mathbf{1})^t}\mathbf{V}^{(1)-1}\mathbf{Y}^{(\mathbf{1})}) = \mathbf{X}^{(\mathbf{1})^t}\mathbf{V}^{(1)-1}E(\mathbf{Y}^{(\mathbf{1})})$. Now since β=0, $E(\mathbf{Y}^{(\mathbf{1})}) = E(g^{(1)} + \varepsilon^{(1)}) = 0$ from the model.

Second, $Var(\hat{\beta}) = Var(\mathbf{X}^{(\mathbf{1})^t}\mathbf{V}^{(1)-1}\mathbf{Y}^{(\mathbf{1})}) = \mathbf{X}^{(\mathbf{1})^t}\mathbf{V}^{(1)-1}Var(\mathbf{Y}^{(\mathbf{1})})\mathbf{V}^{(1)-1}\mathbf{X}^{(\mathbf{1})}$. Now $Var(\mathbf{Y}^{(\mathbf{1})}) = Var(\varepsilon^{(\mathbf{1})}) + Var(\mathbf{g}^{(\mathbf{1})})$, which by definition is given by (6), implying that $Var(\hat{\beta})$ is indeed given by (7)

Supplementary figures & Tables

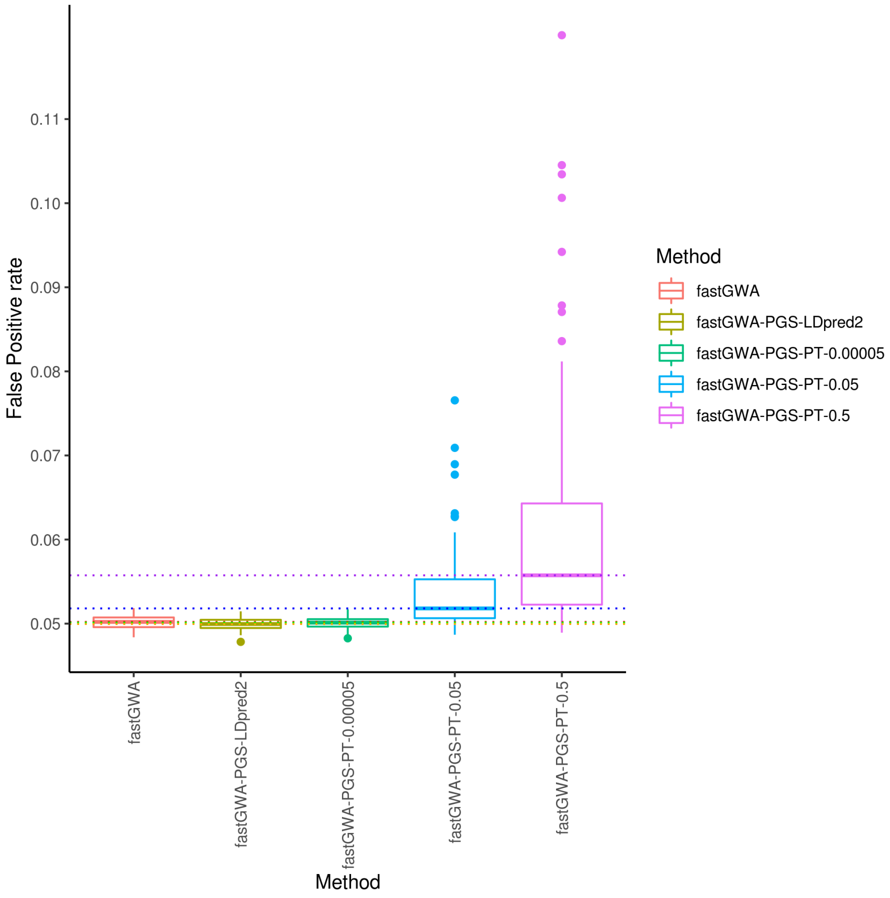

Figure C1: Assessment of the false positive rate in 100 simulations, causal variants were simulated on the even chromosomes leaving the odd chromosomes to carry information on the false positive rate. The results of fastGWA-PGS are shown for three P&T p-value thresholds (LOCO PGS is calculated using 5 x 10^{-5}, 0.05 & 0.5 p-value cut off points) and LDpred2.

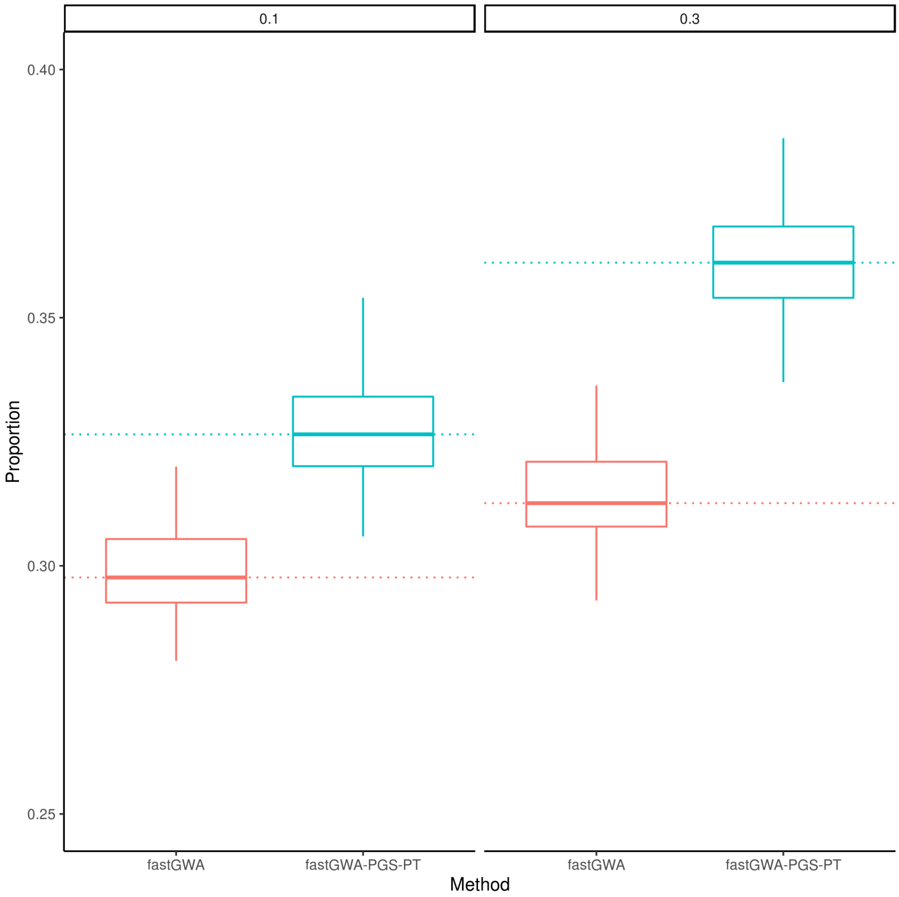

Figure C2: Proportion of causal variants recovered in 100 case-control simulations of a disease with prevalence 0.1 (left) or 0.3 (right), heritability of 0.5 and 1,000 causal variants.

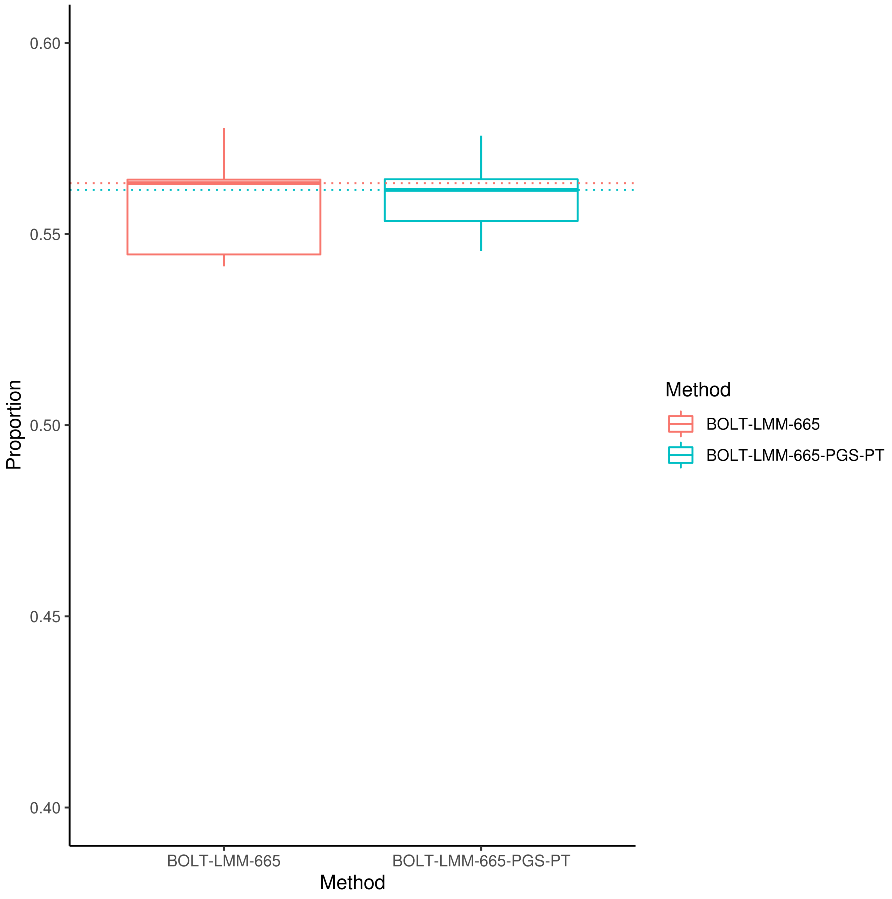

Figure C3: The effect on power of adding the LOCO PGS fixed effect to BOLT-LMM with a GRM that included all variants in the simulation. The LOCO PGS is calculated using the P&T method. The plot shows the proportion of causal variants recovered over 10 simulations.

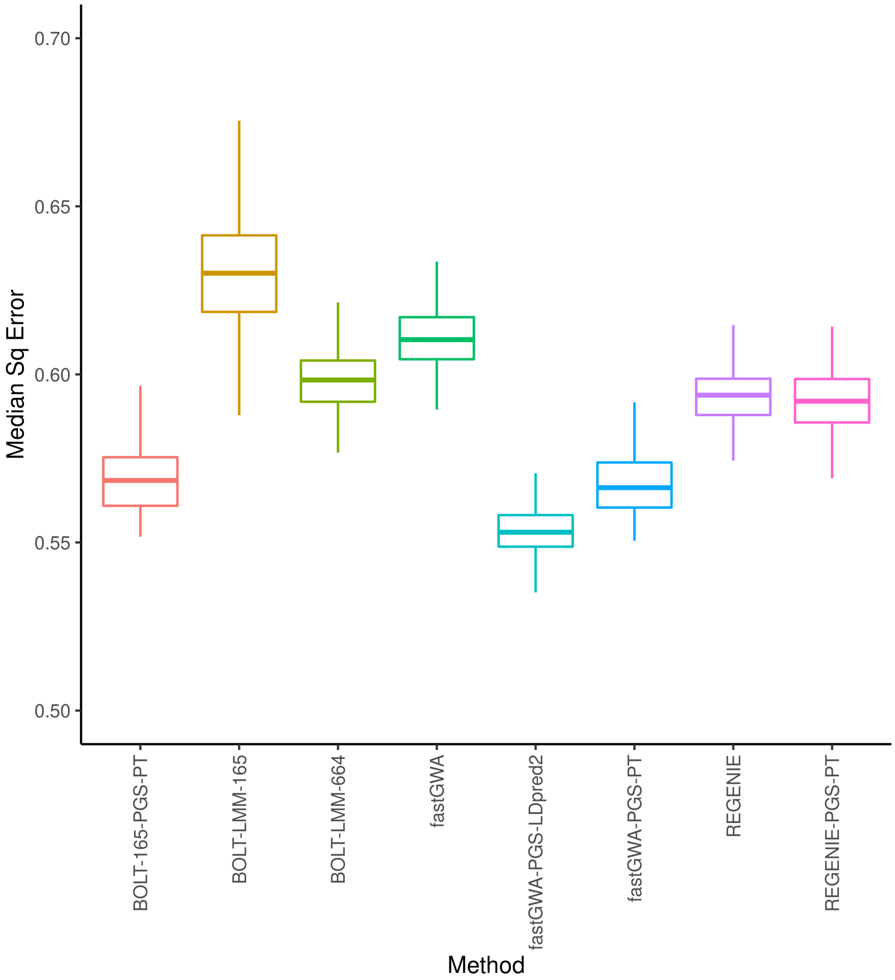

Figure C4: Median squared error of effect size estimates over 100 simulations of a quantitative trait with heritability of 0.5 and 1,000 causal variants in 100,000 individuals .

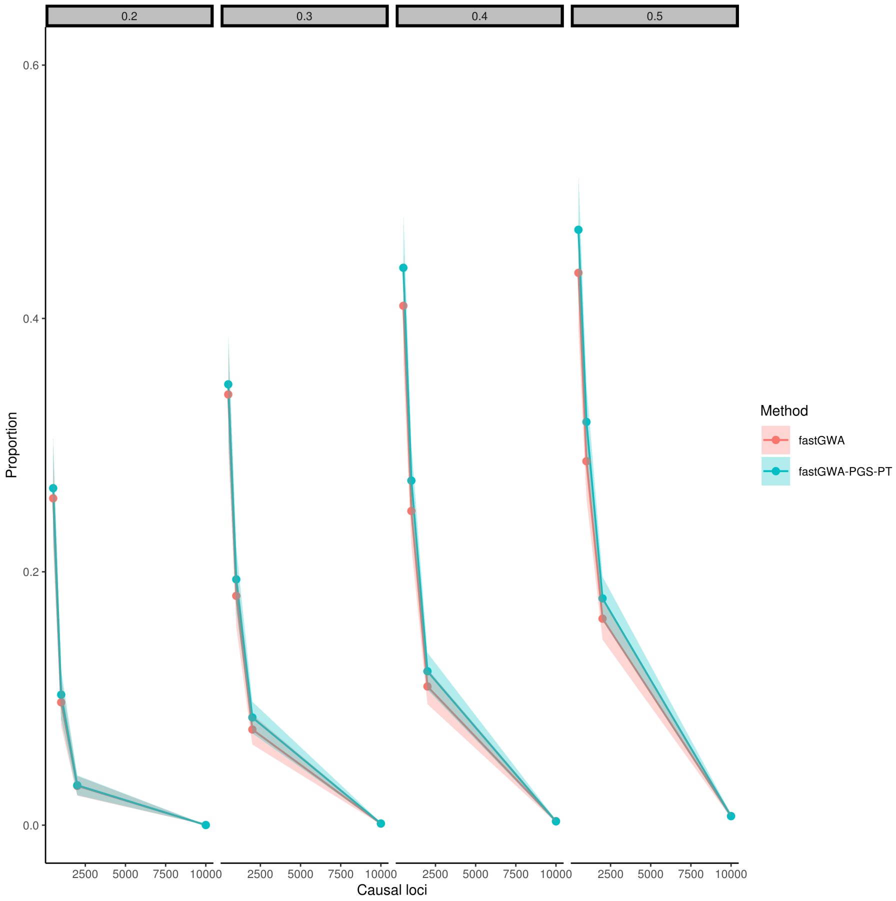

Figure C5: The proportion of causal variants recovered as a function of the number of causal variants in case-control simulations of a disease with prevalence 0.1. The plots show the results for h2 ranging from 0.2 to 0.5.

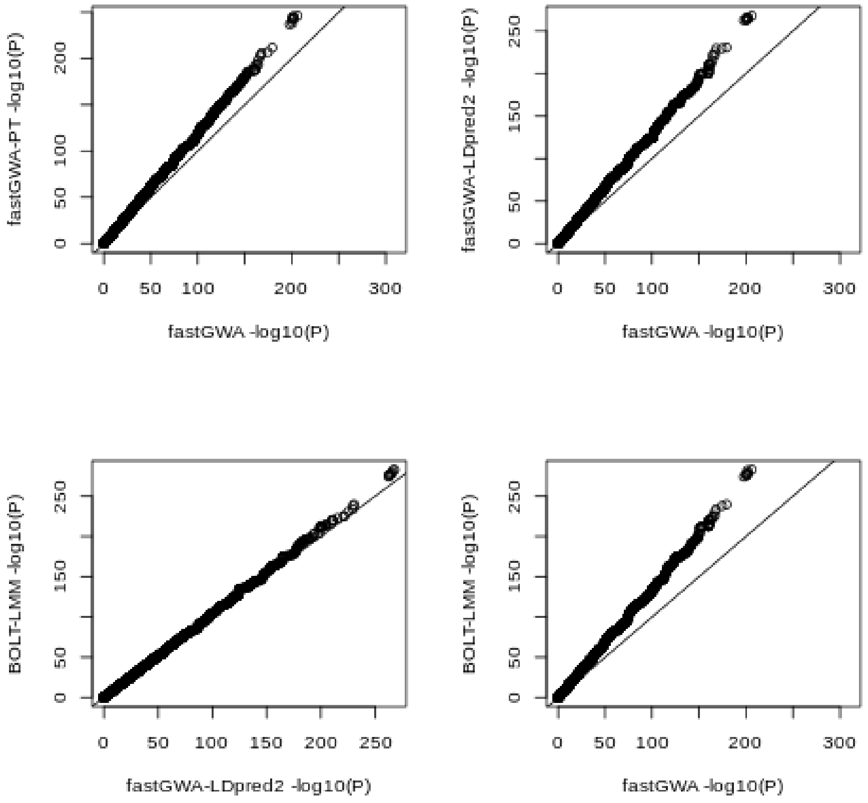

Figure C6: QQ plots comparing the distributions of the negative logarithm of the p-values obtained when different methods were applied to the height phenotype from the UK Biobank.

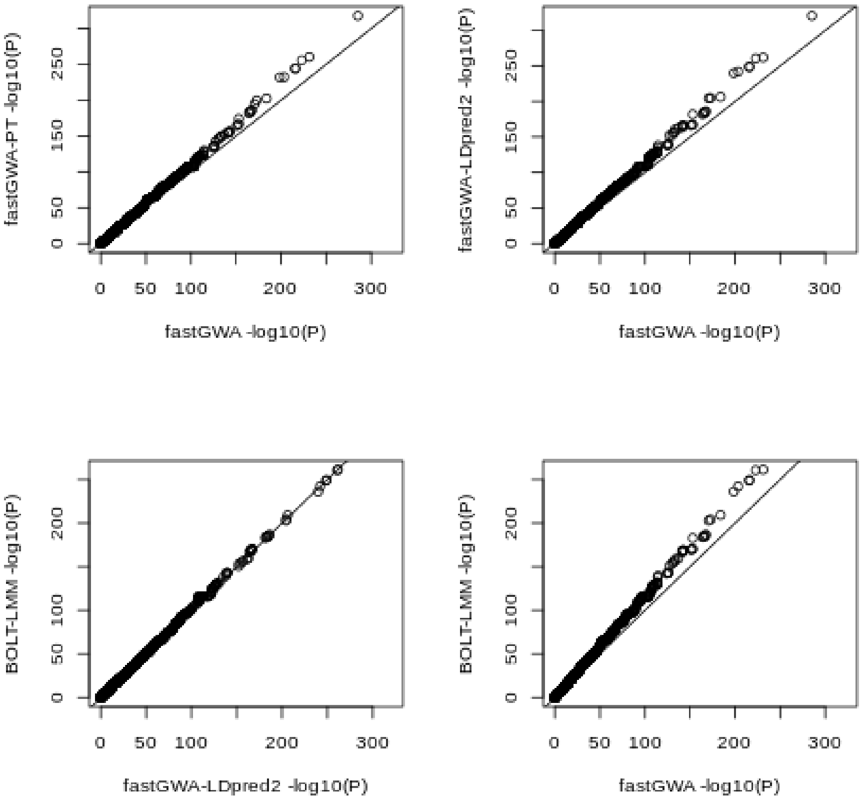

Figure C7: QQ plots comparing the distributions of the negative logarithm of the p-values obtained when different methods were applied to the heel bone mineral density (HBMD) phenotype from the UK Biobank

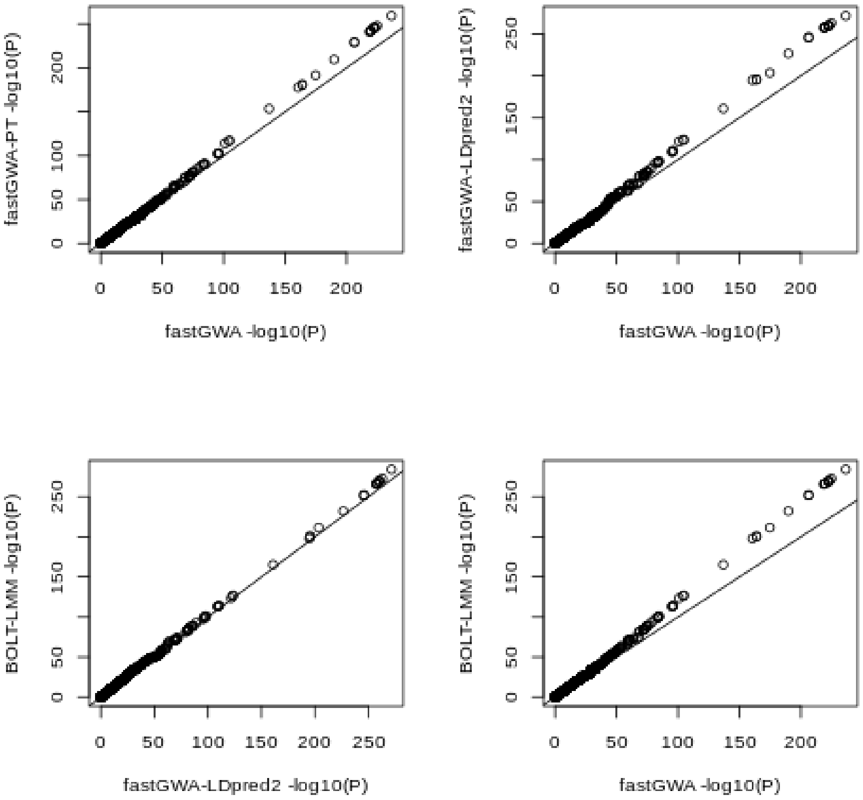

Figure C8: QQ plots comparing the distributions of the negative logarithm of the p-values obtained when different methods were applied to the body mass index (BMI) phenotype from the UK Biobank

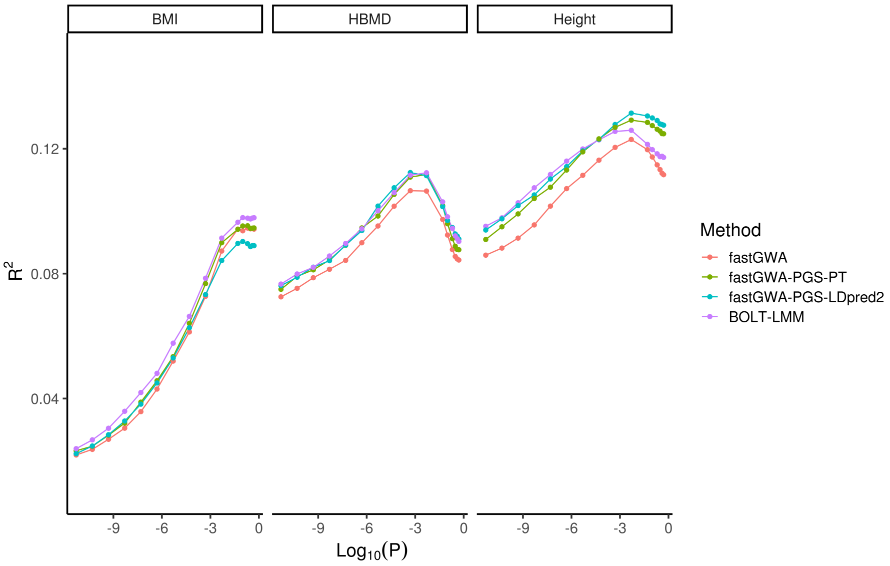

Figure C9: Proportion of phenotypic variance explained by differing GWAS methods Proportion of phenotypic variance (in height, BMI, & HBMD) explained by polygenic scores, calculated using the P&T method, as a function of the p-value thresholds applied in the P&T method. The polygenic scores were calculated from summary statistics obtained using the methods shown.

Table C1: Mean proportion of causal variants recovered in 100 simulations of a quantitative trait (h^2=0.5, N=100,000 & 1,000 causal loci).

Method	Mean	Change (%) relative to fastGWA
fastGWA	0.445	0.00
fastGWA-PGS-PT	0.527	18.4
fastGWA-PGS-LDpred2	0.561	25.9
BOLT-LMM-165	0.491	10.3
BOLT-LMM-165-PGS-PT	0.545	22.4
BOTL-LMM-664	0.558	25.3
REGENIE	0.481	8.1
REGENIE-PGS-PT	0.485	8.9

Table C2: Paired t-tests for fastGWA vs all other methods (based on 100 simulations of a quantitative trait with h^2=0.5, N=100,000 & 1,000 causal loci).

Method	Mean difference	Conf-95	Conf+95	P
fastGWA-PGS-PT	82	78	86	3e-32
fastGWA-PGS-LDpred2	115	110	120	2.3e-36
BOTL-LMM-664	112	108	116	2.3e-40
BOLT-LMM-165	45	42	47	3.1e-31
BOLT-LMM-165-PGS-PT	100	96	103	1.4e-39
REGENIE	36	34	39	3e-28
REGENIE-PGS-PT	39	34	43	3.9e-20

Table C3: Paired t-tests for BOLT-LMM-664 vs all other methods (based on 100 simulations of a quantitative trait with h^2=0.5, N=100,000, & 1,000 causal loci).

Method	Mean difference	Conf-95	Conf+95	P
fastGWA	112	108	116	2.3e-40
fastGWA-PGS-PT	30	26	34	3.1e-18
fastGWA-PGS-LDpred2	-2.7	-5.9	0.47	0.092
BOLT-LMM-165	68	65	70	2.9e-37
BOLT-LMM-165-PGS-PT	12	9.9	15	1.9e-12
REGENIE	76	73	79	1.7e-37
REGENIE-PGS-PT	73	69	77	1.4e-31

Table C4: Median proportion of recovered variants in 100 case control simulations with disease prevalence of 0.1 & 0.3 (h^2=0.5, N=100,000, & 1,000 causal loci).

Method	Prevalence	Median
fastGWA	0.10	0.30
fastGWA	0.30	0.31
fastGWA-PGS-PT	0.10	0.33
fastGWA-PGS-PT	0.30	0.36

Table C5: Paired t-tests between fastGWA and fastGWA-PGS-PT for 100 case control simulations with a disease prevalence of 0.1 & 0.3 (h^2=0.5, N=100,000 & 1,000 causal loci).

Method	Prevalence	Mean difference	Conf-95	Conf+95	P
fastGWA-PGS-PT	0.1	29.3	28.1	30.6	2.17e-65
fastGWA-PGS-PT	0.3	38.2	32.9	43.5	3.66e-25

Table C6: Maximum difference in sensitivity between methods, and the corresponding specificity at which this maximum occurs (from 100 simulations with h^2=0.5, N=100,000 & 1,000 causal loci).

Method comparison	Relative increase	Max ΔSensitivity	Corresponding specificity
fastGWA-PGS-LDpred2 vs fastGWA	0.1135	0.0728	0.9988
fastGWA-PGS-LDpred2 vs BOTL-LMM-664	0.0016	0.0015	0.2000
REGENIE vs fastGWA	0.0315	0.0217	1.0000
REGENIE-PGS-PT vs fastGWA	0.0347	0.0239	1.0000
BOLT-LMM-PGS-PT vs BOLT-LMM-165	0.0419	0.0278	0.9991
BOTL-LMM-664 vs fastGWA	0.1185	0.0766	0.9986
fastGWA-PGS-PT vs fastGWA	0.0847	0.0531	0.9992
fastGWA-PGS-LDpred2 vs fastGWA-PGS-PT	0.0287	0.0208	0.9966

Table C7: Average of the median squared error (MEDSE) of effect size estimates for causal variants across 100 simulations (h^2=0.5, N=100,000 & 1,000 causal loci) and relative change to fastGWA.

Method	Mean	Improvement relative to fastGWA
fastGWA	0.6196	0.0%
fastGWA-PGS-PT	0.5756	7.1%
fastGWA-PGS-LDpred	0.5612	9.4%
BOLT-LMM-165	0.6510	-5.0%
BOLT-LMM-165-PT	0.5764	7.0%
BOTL-LMM-664	0.6070	2.0%
REGENIE	0.6032	2.6%
REGENIE-PT	0.6022	2.8%

Table C8: Paired t-tests applied to the median squared error (MEDSE) of effect size estimates for causal variants across 100 simulations, relative to fastGWA (h^2=0.5, N=100,000 & 1,000 causal loci).

Method	Mean difference	Conf-95	Conf+95	P
fastGWA-PGS-PT	-0.044	-0.046	-0.042	3e-76
fastGWA-PGS-LDpred2	-0.058	-0.06	-0.057	5.9e-91
BOTL-LMM-664	-0.013	-0.014	-0.011	3.7e-30
BOLT-LMM-165	0.031	0.015	0.048	0.00022
BOLT-165-PGS-PT	-0.043	-0.045	-0.041	2.6e-71
REGENIE-PGS-PT	-0.017	-0.02	-0.015	2.2e-26
REGENIE	-0.016	-0.018	-0.015	1.9e-38

Table C9: Paired t-test of MEDSE beta estimates of 100 quantitative trait simulations relative to BOLT-LMM-165.

Method	Mean difference	Conf-95	Conf+95	P
fastGWA	-0.031	-0.048	-0.015	0.00022
fastGWA-PGS-PT	-0.075	-0.092	-0.059	5.2e-15
fastGWA-PGS-LDpred2	-0.09	-0.11	-0.073	3e-18
BOTL-LMM-664	-0.044	-0.061	-0.027	1.1e-06
BOLT-165-PGS-PT	-0.075	-0.09	-0.059	2e-15
REGENIE-PGS-PT	-0.049	-0.063	-0.034	1.8e-09
REGENIE	-0.048	-0.063	-0.033	1.2e-08

Table C10: Proportion of causal variants recovered for simulations of a quantitative trait over a range of parameter values (N=100,000; Nc = number of causal variants).

Heritability	Method	Nc	Proportion
0.1	fastGWA	500	0.25
0.1	fastGWA	1,000	0.10
0.1	fastGWA	2,000	0.03
0.1	fastGWA	5,000	0.00
0.1	fastGWA	10,000	0.00
0.1	fastGWA-PGS-PT	500	0.26
0.1	fastGWA-PGS-PT	1,000	0.11
0.1	fastGWA-PGS-PT	2,000	0.03
0.1	fastGWA-PGS-PT	5,000	0.00
0.1	fastGWA-PGS-PT	10,000	0.00
0.2	fastGWA	500	0.35
0.2	fastGWA	1,000	0.23
0.2	fastGWA	2,000	0.10
0.2	fastGWA	5,000	0.02
0.2	fastGWA	10,000	0.00
0.2	fastGWA-PGS-PT	500	0.39
0.2	fastGWA-PGS-PT	1,000	0.26
0.2	fastGWA-PGS-PT	2,000	0.12
0.2	fastGWA-PGS-PT	5,000	0.02
0.2	fastGWA-PGS-PT	10,000	0.00
0.3	fastGWA	500	0.47
0.3	fastGWA	1,000	0.33
0.3	fastGWA	2,000	0.18
0.3	fastGWA	5,000	0.04
0.3	fastGWA	10,000	0.01
0.3	fastGWA-PGS-PT	500	0.50
0.3	fastGWA-PGS-PT	1,000	0.38
0.3	fastGWA-PGS-PT	2,000	0.20
0.3	fastGWA-PGS-PT	5,000	0.05
0.3	fastGWA-PGS-PT	10,000	0.01
0.4	fastGWA	500	0.53

0.4	fastGWA	1,000	0.39
0.4	fastGWA	2,000	0.23
0.4	fastGWA	5,000	0.07
0.4	fastGWA	10,000	0.02
0.4	fastGWA-PGS-PT	500	0.58
0.4	fastGWA-PGS-PT	1,000	0.45
0.4	fastGWA-PGS-PT	2,000	0.29
0.4	fastGWA-PGS-PT	5,000	0.09
0.4	fastGWA-PGS-PT	10,000	0.02
0.5	fastGWA	500	0.56
0.5	fastGWA	1,000	0.43
0.5	fastGWA	2,000	0.28
0.5	fastGWA	5,000	0.11
0.5	fastGWA	10,000	0.03
0.5	fastGWA-PGS-PT	500	0.62
0.5	fastGWA-PGS-PT	1,000	0.52
0.5	fastGWA-PGS-PT	2,000	0.36
0.5	fastGWA-PGS-PT	5,000	0.16
0.5	fastGWA-PGS-PT	10,000	0.04

Table C11: Proportion of causal variants recovered for simulations of a quantitative trait over a range of parameter values (N=430,000; Nc = number of causal variants).

Heritability	Method	Nc	Proportion
0.1	fastGWA	500	0.54
0.1	fastGWA	1,000	0.40
0.1	fastGWA	2,000	0.25
0.1	fastGWA	5,000	0.08
0.1	fastGWA	10,000	0.02
0.1	fastGWA-PGS-PT	500	0.55
0.1	fastGWA-PGS-PT	1,000	0.41
0.1	fastGWA-PGS-PT	2,000	0.27
0.1	fastGWA-PGS-PT	5,000	0.08
0.1	fastGWA-PGS-PT	10,000	0.02
0.2	fastGWA	500	0.68
0.2	fastGWA	1,000	0.56
0.2	fastGWA	2,000	0.40
0.2	fastGWA	5,000	0.21
0.2	fastGWA	10,000	0.08
0.2	fastGWA-PGS-PT	500	0.69
0.2	fastGWA-PGS-PT	1,000	0.59
0.2	fastGWA-PGS-PT	2,000	0.44
0.2	fastGWA-PGS-PT	5,000	0.23
0.2	fastGWA-PGS-PT	10,000	0.09
0.3	fastGWA	500	0.73
0.3	fastGWA	1,000	0.64
0.3	fastGWA	2,000	0.50
0.3	fastGWA	5,000	0.30
0.3	fastGWA	10,000	0.15
0.3	fastGWA-PGS-PT	500	0.76
0.3	fastGWA-PGS-PT	1,000	0.68
0.3	fastGWA-PGS-PT	2,000	0.54
0.3	fastGWA-PGS-PT	5,000	0.34
0.3	fastGWA-PGS-PT	10,000	0.18
0.4	fastGWA	500	0.77

0.4	fastGWA	1,000	0.68
0.4	fastGWA	2,000	0.56
0.4	fastGWA	5,000	0.37
0.4	fastGWA	10,000	0.21
0.4	fastGWA-PGS-PT	500	0.81
0.4	fastGWA-PGS-PT	1,000	0.72
0.4	fastGWA-PGS-PT	2,000	0.62
0.4	fastGWA-PGS-PT	5,000	0.40
0.4	fastGWA-PGS-PT	10,000	0.26
0.5	fastGWA	500	0.76
0.5	fastGWA	1,000	0.70
0.5	fastGWA	2,000	0.58
0.5	fastGWA	5,000	0.41
0.5	fastGWA	10,000	0.26
0.5	fastGWA-PGS-PT	500	0.80
0.5	fastGWA-PGS-PT	1,000	0.74
0.5	fastGWA-PGS-PT	2,000	0.66
0.5	fastGWA-PGS-PT	5,000	0.49
0.5	fastGWA-PGS-PT	10,000	0.33

Table C12: Proportion of causal variants recovered for simulations of a binary trait over a range of parameter values (N=100,000; disease prevalence = 0.1; Nc = number of causal variants).

Heritability	Nc	Method	Proportion
0.2	10,000	fastGWA	0.0001
0.2	10,000	fastGWA-PGS-PT	0.0001
0.2	1,000	fastGWA	0.0970
0.2	1,000	fastGWA-PGS-PT	0.1030
0.2	2,000	fastGWA	0.0310
0.2	2,000	fastGWA-PGS-PT	0.0315
0.2	500	fastGWA	0.2580
0.2	500	fastGWA-PGS-PT	0.2660
0.3	10,000	fastGWA	0.0012
0.3	10,000	fastGWA-PGS-PT	0.0013
0.3	1,000	fastGWA	0.1810
0.3	1,000	fastGWA-PGS-PT	0.1940
0.3	2,000	fastGWA	0.0755
0.3	2,000	fastGWA-PGS-PT	0.0850
0.3	500	fastGWA	0.3400
0.3	500	fastGWA-PGS-PT	0.3480
0.4	10,000	fastGWA	0.0032
0.4	10,000	fastGWA-PGS-PT	0.0030
0.4	1,000	fastGWA	0.2480
0.4	1,000	fastGWA-PGS-PT	0.2720
0.4	2,000	fastGWA	0.1095
0.4	2,000	fastGWA-PGS-PT	0.1215
0.4	500	fastGWA	0.4100
0.4	500	fastGWA-PGS-PT	0.4400
0.5	10,000	fastGWA	0.0072
0.5	10,000	fastGWA-PGS-PT	0.0071
0.5	1,000	fastGWA	0.2873
0.5	1,000	fastGWA-PGS-PT	0.3183
0.5	2,000	fastGWA	0.1630
0.5	2,000	fastGWA-PGS-PT	0.1790
0.5	500	fastGWA	0.4360

0.5	500	fastGWA-PGS-PT	0.4700

Table C13: Two-sample tests of equality of proportions applied to the proportions of significant loci identified using the method shown, compared to fastGWA. The results shown are for the three UK Biobank quantitative traits analyzed. Prop 1 and Prop 2 show the proportions of significant loci for the method on the row and for fastGWA, respectively. Conf-95 and Conf+95 show the low and upper 95% confidence interval for the difference in these proportions.

Method	Phenotype	P	X-sq	Prop 1	Prop 2	Conf-95	Conf+95
BOLT-LMM	BMI	3.133e-05	17.3357	0.0159	0.0123	0.0019	0.0054
fastGWA-PGS-LDpred2		0.1083	2.5795	0.0136	0.0123	-0.0003	0.0030
fastGWA-PGS-PT		0.1563	2.0093	0.0135	0.0123	-0.0005	0.0029
BOLT-LMM	HBMD	0.009114	6.8004	0.0106	0.0087	0.0005	0.0033
fastGWA-PGS-LDpred2		0.02029	5.3864	0.0104	0.0087	0.0003	0.0031
fastGWA-PGS-PT		0.1066	2.6034	0.0098	0.0087	-0.0002	0.0026
BOLT-LMM	Height	5.458e-19	79.2553	0.0493	0.0360	0.0104	0.0163
fastGWA-PGS-LDpred2		1.953e-13	54.0515	0.0468	0.0360	0.0079	0.0138
fastGWA-PGS-PT		1.166e-06	23.6328	0.0430	0.0360	0.0042	0.0099