Title	Analysis of clonal mutations in cancer as a means of studying variation in somatic mutation processes
Author(s)	Cleary, Siobhán
Publication Date	2023-09-13
Publisher	NUI Galway
Item record	http://hdl.handle.net/10379/17895

OLLSCOIL NA GAILLIMHE
UNIVERSITY OF GALWAY

Analysis of clonal mutations in cancer as a means of studying variation in somatic mutation processes

Siobhán Cleary

A thesis presented in fulfilment of the requirements for the degree of Doctor of Philosophy

Supervisor: Professor Cathal Seoighe

School of Mathematical and Statistical Sciences
University of Galway
Galway City, Ireland

July, 2023

Table of Contents

Abstract

Somatic mutations are mutations that arise throughout a person's lifetime. They contribute to ageing, cancer and other age-related disorders. Recent technological advances led to many studies investigating somatic mutations in normal tissues. However, somatic mutations are hard to identify in normal tissues due to their low frequency and the difficulty distinguishing between real mutations and errors incorporated during the experimental processes. Studies of somatic mutations in normal tissues suggest that there is still much unknown about how somatic mutations contribute to cancer. Somatic mutations can be studied by analysing cancer samples. Generally, somatic mutations in cancer samples are studied to understand cancer progression and response to treatment. This thesis aimed to investigate somatic mutations present in all cancer cells of a sample (clonal mutations) as a means to understand what is happening in normal tissue.

Chapter 2 describes a method to predict the total clonal mutation load of a cancer sample and the use of this approach to investigate the relationship between variation in clonal somatic mutation load and differences between tissues in the risk of developing cancer. Before predicting the total clonal load, we first needed to distinguish between clonal mutations and mutations present in only a subset of cells (subclonal). We adjusted variant frequency for tumour purity and local copy number variation to classify variants as clonal or subclonal. We used the linear relationship between clonal variants and age to predict the total clonal burden for each tissue type. Under the assumption that subclonal mutation accumulation does not correlate with age, we determined what proportion of true clonal variants were classified as clonal. By adjusting various thresholds for classifying variants as clonal variants, we could classify, at best, 45% of the true clonal variants. We then used the relationship between clonal mutation burden and age to estimate the true clonal load for our samples. To investigate whether the estimated clonal mutation burden could be used as a proxy for the number of somatic mutations in healthy cells, we compared our results to somatic mutation burdens that have been measured directly in normal tissues (matched for age and tissue type with the cancer samples). We also found that the predicted clonal load was correlated with lifetime cancer risk. Our findings suggest that we can use predicted clonal load from cancer samples to investigate somatic mutations in the normal tissue and has the advantage of being able to use the large volume of cancer genomics data that has already been generated to extend our understanding of the accumulation of somatic mutations in normal tissues.

The major histocompatibility complex (MHC) can present neoantigens resulting from somatic mutations on the cell surface, potentially directing an immune response against it. In Chapter 3, we investigated whether gene expression explains the lack of signal of immunoediting observed among clonal passenger mutations. This hypothesis stemmed from two publications that reported that driver mutations arise in gaps in the capacity of the immune system to recognize them. We investigated whether passenger mutations ca-

pable of eliciting an immune response occur preferentially on lowly expressed genes or if the mutant allele has a lower expression than the reference allele through a process termed allele-specific expression (ASE). The neoantigen must be expressed to be presented by the MHC on the cell surface, so a reduction in expression could be a means by which the immunogenic mutations are tolerated. After accounting for gene length and sequence context, we found no difference in the expression of genes harbouring immunogenic mutations compared to nonimmunogenic or synonymous mutations. Additionally, there was no evidence that the mutant allele exhibited ASE more often for immunogenic mutations than nonimmunogenic mutations. Using simulations, we also estimated an upper bound for the impact of immunoediting on the mutational landscape in cancer, showing that at most 5% of missense mutations could be removed by this process. To our knowledge, this was the first attempt to quantify the proportion of missense mutations removed through immunoediting.

Finally, in Chapter 4, we extended our analysis on the relationship between gene expression and somatic mutation accumulation by investigating the relationship between germline ASE and cancer risk. Here, we investigated the hypothesis that a single score representing germline ASE in all TSGs for an individual would be associated with an increased cancer risk because only mutations on the expressed copy would be required to disrupt the function of the gene. To assess this, we first tested the ability of two methods to predict ASE using genotype data. We modified a tool called PrediXcan which predicts overall gene expression to predict the expression of each haplotype and generated a ratio with the predicted values. We also applied logistic regression models using heterozygous SNP status as predictors and ASE status as the outcome. Although the performance of ASE predictions was poor for many genes using both methods, our results indicate that it may be possible to generate more accurate predictions using genotype data as input as more data becomes available. As a pilot study, we generated a single TSG ASE score using the genes for which the predictions worked well and assessed the relationship with breast cancer risk. We found no statistically significant relationship between TSG ASE and cancer risk, which is likely due to our inability to predict ASE in the TSGs that contribute to cancer risk in this tissue type, as assessed using cancer data.

In conclusion, this thesis presented a novel approach to predict the true clonal load of cancer samples and demonstrated its similarity to the observed somatic mutation load in normal tissue. We also provided further insight into the role of the immune system in shaping the mutational landscape of cancer samples and, using a novel method, generated an estimate for the proportion of missense mutations removed through immunoediting. Finally, we also presented a novel approach to predict germline ASE using genotype data showing it is feasible for some genes and performance is likely to be improved as more data becomes available.

Acknowledgements

First and foremost, I would like to express my sincerest thanks to my supervisor Prof. Cathal Seoighe for all your support, patience, and guidance over the past four years. I learned a lot under your supervision.

I would like to thank my GRC members, Dr Andrew Simpkin, Dr Emma Holian and Dr Pilib Ó Broin for your feedback and advice each year.

To my fellow Bioinformatic PhD students and colleagues, particularly those who worked with me in ADB1018, it has been an absolute pleasure. My sincerest thanks to Declan- you've been a great friend and a massive support from the very beginning, always willing to help when I was stuck, particularly when it came to lugh!! Noor, the other half of team Noor and Siobhán, it has been a great joy to work with you but also to get to know you and to experience this journey with you. You and Sumaira are two of the kindest souls I've met and I've really appreciated all of the support you've both given and all of the laughs we've shared. Brian I've really enjoyed all of our discussions, both related and unrelated to somatic mutations. They've been very thought provoking. Thank you! Adib I've greatly appreciated the help you've given when it came to understanding complex mathematical and statistical methods and Im in awe at the way you can break it down into simple terms that are easy to understand. Laura, Barbara, Andrew and Barry you were all so welcoming when I first started in Galway and I really enjoyed all of the fun times we had. Lydia it has been a pleasure to get to know you over the last couple of years. I've really enjoyed our coffee dates which gave me a much needed break from thesis writing. Thank you to Aaron for your words of wisdom and encouragement.

To all of my friends, particularly Karen, Ellen, Marta, Carol, Sinead, Simone, Joey, Caroline, Lorraine, Amy-Louise and Amy. Thank you for always believing in me, for always encouraging me and for reminding me to have fun and enjoy life. You have all helped me in some way or another during the PhD and I'm very lucky to have such a great group of people in my life.

To Dr. Jaine Blayney and Dr. Samuel Clokie who both played a massive role in shaping me into the Bioinformatician I am today. I learnt a lot from you during the beginning of my career and you gave me the confidence to pursue my PhD degree.

Last but definitely not least, to my family, who have provided me with so much love and support throughout my life. I would not have been able undertake this PhD if it had not been for your help. Fiona - not everyone was lucky enough to have been born with a best friend. Thank you for moving to Galway and for feeding me and forcing me to step away from the laptop to join you and Conor Mc for walks. Thank you for always being there for me and for encouraging me to keep going whenever things have felt difficult. To Conor and Elena, thank you both for always being willing to offer solutions

when I had questions about linear models and always being so generous with your time. And finally thank you to my parents. I genuinely cant thank you enough for supporting every decision I've made, no matter how out of the blue they seemed, for believing that I could achieve everything I set out to achieve and for making the path a lot easier. You don't know how much your belief in what I can achieve has meant to me. You have always been there for me and for that I am truly grateful.

Declaration

I hereby declare that this thesis which I now submit for assessment as partial fulfillment of the requirements for the award of Doctor of Philosophy is all my own work and I have acknowledged any assistance or contributions and cited the published work of others where applicable. I have not obtained a degree from the University of Galway or elsewhere on the basis of any of this work.

Signed: ______________________

Date: 14 July 2023

List of Figures

List of Tables

1 Chapter 1: Introduction

1.1 Somatic mutations in normal tissues

Mutations accumulate in humans throughout their lifetime. These are somatic mutations, that arise in a single stem cell and are only present in cells derived from that cell [1]. They do not occur in the germline and are not passed on to offspring [2]. These mutations can be changes in the DNA at a single position (single nucleotide variants; SNVs), small insertions and deletions of sequence (indels), somatic copy number alterations (SCNA) or structural variants (SVs) [3]. However, our research was focused on investigating SNVs, the most numerous type of somatic mutations [4], and as such, they will be the focus of this thesis. Due to the redundancy of the genetic code, which means different codons can code for the same amino acid, SNVs in coding regions can have different functional effects on the coding sequence [5]. These changes can be synonymous (do not change the amino acid sequence) or nonsynonymous (changes the amino acid sequence), with nonsynonymous variants classified as missense (the amino acid is changed to a different amino acid) or nonsense (the protein sequence is terminated prematurely) [6]. Although recently, a third category has been proposed, “unsense”, to account for those changes that appear synonymous but have a functional consequence, such as changing the gene expression or impacting protein production [7].

Somatic mutations play a role in the ageing process [8, 9] as well as in age-related disorders such as neurodegenerative diseases [10, 11] and cancer [12–14]. As a result, they are a key focus of research in these areas. Until recently, it has been challenging to assess somatic mutation in normal tissue. This is due to the difficulty of detecting somatic mutations at low frequencies using bulk sequencing techniques. When performing bulk sequencing, a large number of cells is required from a sample. However, due to the polyclonal nature of most tissues, somatic mutations are present in only a small number of cells, meaning they will be only present in a small proportion of the cells in the sample taken. Owing to inaccuracies caused by the sequencing and bioinformatic processes, it is difficult to distinguish the true somatic mutations from artefactual variants [15].

1.1.1 Methods to detect somatic mutations in normal tissues

Due to recent advances in technology, the following strategies are being used to study somatic mutations in normal tissues:

1. Single-cell clonal expansions: Tissue cells are cultured *in vitro*, and single cells are sorted by flow cytometry before being clonally expanded. Following this, a second round of clonal expansion of single cells picked from these cultures generates enough cells to perform whole genome sequencing [16, 17]. A limitation of this method is the introduction of artefactual variants through the culturing process itself. However, most of these mutations will only be present in a subset of the cells, while

true somatic variants should be present in all cells of the culture [18].

2. Laser capture microdissection (LCM): this approach involves using a laser under a microscope to remove unwanted cells leaving only the area of interest [19]. This allows the retention of a clonal population of cells, which can then be sequenced by whole exome or whole genome sequencing. A disadvantage to this method is that it requires well-defined clonal structures, so it is limited to large clones.

3. Deep Targeted Sequencing: For epithelial structures that do not have well-defined structures, such as skin or oesophagus, punch biopsies are taken, and small sections are sequenced to a very high depth [20–23]. However, this method focuses on a panel of informative genes rather than all genes within a sample.

4. Consensus sequencing with molecular barcodes: The duplex sequencing method was designed specifically to identify variants present at extremely low sequencing depth in a sample by labelling each strand of double-stranded DNA molecule with a sequence tag and performing PCR amplification and sequencing of each strand [24]. The sequences from both strands are only kept if they match each other exactly, resulting in high accuracy for variant calls. A limitation of this method is that it requires a larger sequencing volume than standard sequencing to achieve appropriate sequencing depth for analysis. Another method called bottleneck sequencing (BotSeq) was developed, which built upon this approach using limiting dilutions prior to PCR amplification, creating a bottleneck that results in random sampling of the double-stranded molecules resulting in a smaller library for sequencing [25]. A more recent version of this sequencing approach, nanorate sequencing (NanoSeq), has a reduced error rate of 5 errors per 1 billion base pairs [26]. However, as it randomly samples the genome not all genes are covered, but it does give an indication of mutational burden and patterns [27]. Another method called enzymatically cleaved and optimal sequencing (EcoSeq) performs similar analyses but reduces the number of genomic regions required to analyse a sample to reduce sequencing costs[28]. However, it also only analyses a portion of the genome. A similar approach creates independent copies of each strand that have been labelled with a unique molecular identifier, using a rolling circle amplification step before PCR amplification [29]. It has the advantage that it only requires one strand of the DNA, so it is more cost-effective than duplex sequencing techniques, and the results are comparable to those obtained by single cell-based approaches.

5. Single-cell DNA sequencing techniques: Whole genome application methods such as multiple displacement amplification [30], multiple annealing and looping-based amplification cycles [31], and degenerate oligonucleotide-primed PCR [32] have been used to detect somatic mutations in single cells. However, each method is prone to error with

differences in coverage of alleles and complete allelic dropout common due to the amplification process, making it difficult to accurately call somatic mutations using these methods [33, 34].

6. Single-cell RNA sequencing: A tool called SCmut has been developed to successfully call somatic mutations using single-cell RNA sequencing (scRNA-seq) data [35]. However, this requires that mutations are first identified using bulk DNA-sequencing techniques or that the user supplies a list of somatic mutations as input [35]. Additionally, methods incorporating scRNA-seq with data from single-nucleus assay for transposase-accessible chromatin sequencing (snATAC-seq) have also been developed to identify somatic mutations in normal tissue [36, 37].

7. Bulk RNA-Sequencing: A recent study used bulk RNA-sequencing data of normal samples from the Genotype-Tissue Expression (GTEx) project to identify somatic mutations [38]. Additionally, a pipeline called RNA-MuTect has been developed to identify somatic mutations in RNA-seq data [39]. However this requires a matched DNA sequencing sample which is used to identify germline mutations. Calling variants in RNA-sequencing data, in general, is extremely difficult due to splicing and RNA-editing events, as well as technical factors such as sequencing errors, and mapping errors [38]. Additionally, the specificity of calling variants decreases as coverage increases likely due to sequencing errors passing quality control filters when more reads are present [40]. Therefore, this method requires strict quality control screens and filtering. It also has the disadvantage that only expressed genes can be used to detect somatic mutations [40].

1.1.2 Somatic mutation rates in normal tissue

Somatic mutations have been shown to accumulate with age [41]. However, the rate at which they accumulate varies depending on tissue and cell type (Table 5.1 in Appendix A). Multiple studies of the same tissue have estimated consistent mutation burdens, with bile ducts showing the lowest rate of 9 SNVs per year and appendix, large and small intestines showing the highest rates of 56, 49-51 and 49 SNVs per year, respectively, over the entire genome. A study of different cell types in kidneys showed that mutation burden differs between cell types within a tissue due to different mutagen exposures [42]. A subset of cells from the proximal tubule showed the highest yearly increase of 56.6 SNVs per year, while subcutaneous, visceral adipose tissue and visceral adipose tissue had increases of 17.5 and 27.2 SNVs per year [42]. The majority of studies have focused on mutation accumulation in stem cell tissues due to the technical difficulties of studying differentiated cells. However, there have been a growing number of studies investigating the differences in mutations in non-dividing cells compared to stem cells of the same tissues. Mutations may accumulate at a higher rate in differentiated cells compared to stem cells because the consequences of a mutation in a stem cell are far-reaching and, as a result, a stem cell would be under more stringent error control than differentiated cells [43]. However, differentiated cells would not accumulate mutations caused by DNA replication [44] which means fewer opportunities exist to acquire mutations. Also, differentiated cells are short-lived, meaning there is a limited amount of time for the cells to acquire additional mutations post-mitosis [44]. A study investigating somatic mutations in differentiated liver hepatocytes compared to liver stem cells (LSCs) found that there is a higher mutation frequency in the hepatocytes (21 SNVs per cell per mitosis compared to 11 SNVs in LSCs) [45]. The higher mutation load in differentiated cells could be due to mutations accumulating during the differentiation process itself [44].

1.1.3 Mutational processes that contribute to somatic mutation

As well as assessing the mutation rate of somatic mutations in normal tissues, mutational signatures within samples have also been investigated. Different mutational processes leave a characteristic mark on the type and frequency of mutations found in a cell. These “signatures” have been characterised in cancer samples with 94 single base substitution signatures recorded in the catalogue of somatic mutations in cancer (COSMIC) [46]. These were identified by considering each of the six possible mutation types (C >A, C >G, C >T, T >A, T >C, and T >G) and the nucleotide on its 5’ and 3’ sides which gives a total of 96 trinucleotide contexts. The frequencies of these 96 mutation types are assessed, and the signatures that contribute the most to the observed mutation pattern are identified. While the aetiology of many mutational signatures remains unknown, mutational signature analysis can, in some cases, identify exogenous and endogenous mutational processes that have contributed to the observed mutations. Two signatures, SBS1 and SBS5, which are associated with ageing, have been consistently found in all tissue types (Table 5.1 in Appendix A). SBS1 is caused by spontaneous deamination of 5-methylcytosine [47]. The rates of mutations from this process correlate with estimated rates of stem cell divisions and this signature is, therefore, thought to be a cell division or mitotic clock. SBS5 is another clock-like signature [48]. Although the aetiology is unknown, the number of mutations attributed to this signature correlates with an individual’s age [48]. SBS18 is common across tissue types and has been shown to correlate with alcohol consumption [49]. SBS7 is thought to be associated with UV exposure [50] and has been found in skin, skeletal muscle, lymphocytes and kidney. SBS18 is common in colorectal tissues and small intestine. Several signatures have been identified in normal tissues that have not been found in cancer samples. These may correspond to mutational processes that have been masked in cancer samples but are associated with clonal expansion rather than cancer progression [18].

1.1.4 Positive selection of driver mutations in normal tissue

Studies of somatic mutations in normal tissues have provided evidence that clonal expansions are not exclusive to cancer (Table 5.1 in Appendix A), and there is positive selection of mutations in common cancer-associated genes in morphologically non-cancerous tissue. In fact, some of the driver genes were found to be more frequently mutated in normal tissue than in the corresponding cancer type [51]. Studies of inflammatory tissue have also shown that positive selection of cancer-associated genes is frequent but rarely develops into cancer. Interestingly the number of drivers undergoing positive selection varies between tissue types with colon tissue having a low number of positively selected drivers. These findings in normal tissues indicate that mutations in these driver genes may not be sufficient to drive tumourigenesis and that more needs to be learned about the role of clonal expansion in cancer progression [52]. It is possible that driver genes found in cancer do not play a role in cancer progression but are present because they were in the normal cell [53]. It is worth noting that the screens used in the studies to identify positive selection in normal tissues generally focused on cancer associated genes and it is possible that non-driver genes could have greater positive selection in normal tissues compared to cancer associated genes [51]. It is clear from studies of normal tissue that there is still a lot to be discovered about the transition from normal cells into cancer cells. It is likely that clonal expansion is still important for cancer initiation but that more driver events may be required than initially thought [51]. It is also likely that increased genomic instability as well as clonal expansion is required because genomic instability is common in cancer but not common in normal cells [51]. Additionally, driver mutations may need to occur in the correct order and in the correct combination for cancer to develop [51].

1.2 Contribution of somatic mutations to cancer transformation

As we have seen in the previous section, somatic mutations are common in normal cells, but they also contribute to carcinogenesis. Cancer is characterised by the unregulated growth of cells which form a mass of cells which can be benign or malignant [54]. This is caused by disrupting the processes that control cell division and cell death so that the cells can continue to grow indefinitely and evade normal inhibitory signals [54]. Somatic mutations occur randomly in the genome, and the majority of somatic mutations that accumulate throughout a lifetime are neutral, generally causing changes in the DNA sequence that do not impact fitness [55]. These are termed “passenger” mutations. It is only when a cell acquires mutations that impact genes that are vital for growth and survival that cancer can arise. These mutations occur in tumour suppressor genes and oncogenes and are called “driver” mutations [56]. Tumour suppressor genes (TSGs) normally function to suppress cell growth and proliferation [54]. As a result, these genes can stop cancer from forming and they are often inactivated in tumour cells. Both copies of the gene need to be inactivated in order to inhibit its normal function [54]. Oncogenes are genes whose normal function is to promote cell growth and proliferation and are important for cancer transformation [57–59]. Loss of function of TSGs and gain of function of oncogenes are important for cancer growth and proliferation [60]. The more somatic mutations that arise, the greater the chance of a mutation occurring in one of these cancer-associated genes.

1.2.1 Factors that affect the accumulation of somatic mutations

There are a number of risk factors that increase the likelihood of cancer occurring, both due to endogenous processes as well as exposure to certain environmental factors such as ultraviolet light and smoke that increase somatic mutation accumulation. Age is one of the most significant risk factors for developing cancer because as we get older, there are more opportunities for somatic mutations to accumulate [61]. Sex plays a role in cancer risk, with males having a higher incidence of cancer compared to women, which is thought to be due to differences in environmental exposure and hormones [62]. Race and ethnicity are also risk factors, but this may be due to socioeconomic factors [62]. Environmental exposures, including smoking, alcohol, exposure to UV, and exposure to aristolochic acid, increase the risk of developing cancer. Individuals that smoke have an increased mutational burden compared to non-smokers [63, 64], and respiratory cancer risk is higher for smokers than non-smokers [65]. Point mutation prevalence has been shown to be higher in smoke and aristolochic acid-exposed individuals than non-exposed (27 and 36 fold respectively) [25]. Alcohol consumption is associated with increased cancer risk for upper esophageal, pharyngeal and liver cancers [66]. Exposure to UV light is the main risk factor for developing skin cancer [67]. Aristolochic acids are natural compounds that are found in the Aristolochiaceae family of plants. It is a carcinogen which leaves a characteristic mutational signature in the genome and is associated with an increased risk of developing urological cancers [68].

Inherited mutations can predispose an individual to developing cancer. Patients with an inherited mutation in TSGs or oncogenes tend to develop cancer at an earlier age than patients who do not. Examples include germline mutations in BRCA1 and BRCA2 which increase the risk of ovarian and breast cancers in women, prostate cancer in men and pancreatic cancer in both men and women [69]. Lynch syndrome is an example of an inherited disorder that increases the risk of multiple cancers, particularly colorectal cancer [70]. It is due to germline mutations in MLH1, MSH2, MSH6 [71] or PMS2 [72], which are genes important in DNA mismatch repair. Mutations in these genes increase the mutational burden with a 130-fold increase in the number of nuclear mutations observed in patients who had inactivating mutations in PMS2 compared to PMS2 wild-type patients [25]. Cancer generally originates in stem cells, and the number of stem cell divisions of tissues have been shown to correlate with an increased risk of developing cancer in that tissue [73]. This was attributed to the accumulation of somatic mutations incorporated through errors in DNA replication during each stem cell division.

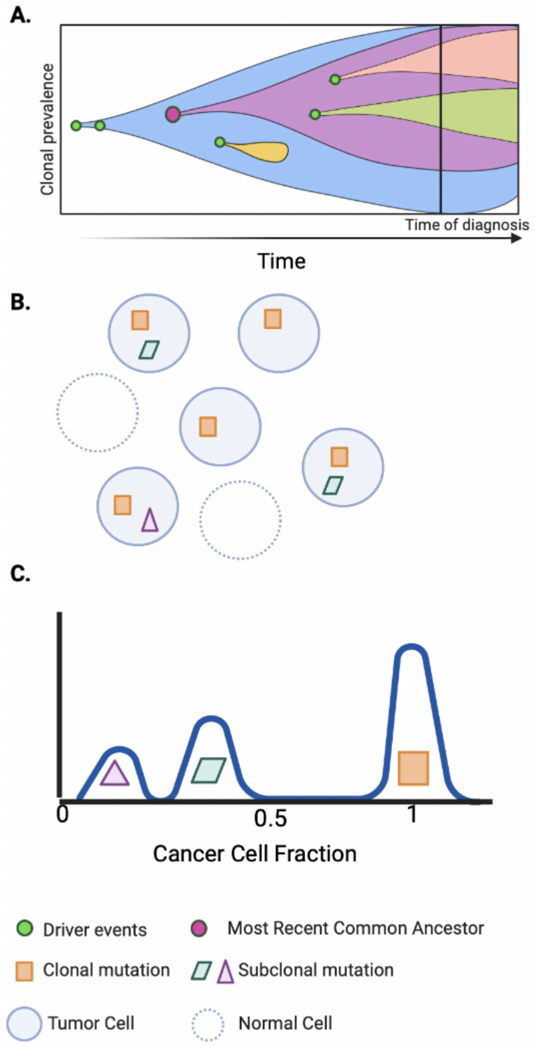


Figure 1.1: **Principles of clonal mutations in tumor samples.** (**A**) Driver mutations (plus symbol) can occur in a cell resulting in clonal expansion so that all cells in the tumor sample have the same mutations that were present in the most recent common ancestor (MRCA) (grey plus sign). (**B**) As the tumor continues to grow, additional mutations can arise, creating sub-populations of cells with distinct sets of mutations. The tumour sample contains a mixture of tumor cells (solid circles) and normal cells (dashed circles). (**C**) Mutations present in the MRCA will be in all tumor cells (clonal) and will have a cancer cell fraction (CCF) of 1 (square) while other mutations which occurred at a later stage will only be present in a subset of cells (subclonal) and will have CCF less than 0.5. Adapted, with permission, from Figure 1 from Dentro *et al.* [74] (copyright to Cold Spring Harbor Laboratory Press).

1.3 Clonal Theory of Cancer

Studies investigating the evolution of tumour growth determined that most tumours are monoclonal, such that a single cell transforms into a cancerous cell and expands to generate a mass of cells with a single common ancestor [75]. In 1976 Peter Nowell proposed the clonal theory of cancer evolution whereby normal cells transform into malignant cancerous cells through a multi-step process that results in the accumulation of a series of mutations [75]. The majority of somatic mutations that accumulate are neutral and do not give a growth advantage to the cell, but they create a population of heterogeneous cells that compete with each other for resources [76]. Occasionally a mutation occurs in a gene that either promotes cell proliferation or decreases cell death, giving the cell a growth advantage so that it can out-compete the other cells [77]. If this cell continues to divide and grow unchecked, it will result in an expansion of cells, all harbouring the same mutations present in the initial founder clone (Figure 1.1). After a cell has transformed into a cancer cell, it can acquire additional somatic mutations, called subclonal mutations, that are only present in a subset of cells [74]. The majority of subclonal mutations are selectively neutral [76, 78]. However, occasionally subclonal mutations can result in late clonal expansions that create distinct cellular populations within the tumour that can lead to intra-tumour heterogeneity (ITH) [79]. This has important implications in terms of resistance to cancer therapy [80].

1.4 Bioinformatic analysis of cancer data

1.4.1 Data generated by Cancer Consortiums

Several consortiums have been created with the aim of bringing groups of scientists together from different research institutions to develop and validate methods, pool resources and expertise, and generate large datasets to expand our knowledge and understanding of cancer. Some of the main ones which have focused on multiple cancer types are detailed below. However, there are also cancer-specific consortiums such as Tracking the evolution of non-small-cell lung cancer (TRACER) [81], Multiple Myeloma Research Foundation (MMRF)(mrf.org) and Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO) [82].

Data from pan-cancer studies:
The Cancer Genome Project was launched by the Welcome Trust Sanger Institute in 2000 to identify genetic changes and patterns across cancer genomes using high-throughput sequencing. The data from this project is available in the Catalogue of Somatic Mutations in Cancer (COSMIC) database (cancer.sanger.ac.uk), which includes thousands of somatic mutations as well as a collection of mutational signatures found in human cancers [83]. The most commonly used dataset in cancer research was generated by The Cancer Genome Atlas (TGCA) [84], a consortium launched jointly by the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI) in December 2005. Initially started as a pilot program with

three cancer types; glioblastoma, serous cystadenocarcinoma of the ovary, and lung squamous carcinoma, it is now comprised of data from primary cancer and matched normal samples of 33 cancer types. There are 20 collaborating institutions located across Canada and the US participating in the program. TCGA has generated genomic, transcriptomic, epigenomic and proteomic data from over 11,000 individuals, and the data is publicly available to researchers under open and controlled access types, hosted on the Genomic Data Commons (GDC) Data Portal (https://portal.gdc.cancer.gov). The Pan-Cancer Analysis of Whole Genomes (PCAWG) project was established to aggregate whole genome analyses from projects such as TCGA and ICGC [85]. This built upon previous studies to identify coding and non-coding variations in cancer genomes of 2,834 individuals across 38 tumour types. The data was derived from primary tumour with matched normal samples and is publicly available through the ICGC database (https://dcc.icgc.org/pcawg). The International Cancer Genome Consortium (ICGC) was launched in 2008 to coordinate cancer genome projects worldwide [86]. Cancer projects such as TCGA and PCAWG are included within the ICGC. Foundation Medicine released genomic data for 18,004 adult cancers that were profiled using the FoundationOne assay [87]. The data comes from 162 tumour subtypes with the majority being thoracic, gastrointestinal, breast, gynaecologic and hepato-pancreato-biliary cancers. Genomics Evidence Neoplasia Information Exchange (GENIE) was launched by the American Association for Cancer Research (AACR) to encourage data sharing from 19 different institutions with the aim of generating enough data to aid in clinical decision making. It is comprised of genomic and clinical data from 44,756 patients from more than 50 cancer types [88]. Clinical Proteomic Tumour Analysis Consortium (CPTAC) was launched in the US in 2011 as a national effort to accelerate understanding of cancer using genomic and proteomic data for 1527 samples from 9 cancer types [89]. International Cancer Proteogenomic Consortium (ICPC) is a global partnership of scientists sharing genomic and proteomic data from cancer samples from 12 tissue types with the aim of using proteogenomic data to predict cancer treatment outcome (cpc.cancer.gov).

Data from studies of rare cancers:
The Cancer Genome Characterization Initiative (CGCI) was launched to characterise rare cancers, including Burkitt's Lymphoma (BLGSP), HIV+ Diffuse Large B-Cell Lymphoma (HTMCP-DLBCL), HIV+ Cervical Cancers (HTMCP-CC), HIV+ Lung Cancers (HTMCP-LC), Diffuse Large B-Cell Lymphoma (NHL-DLBCL) and Follicular Lymphoma (NHL-FL). The majority of projects are ongoing, with NHL-DLBCL and NHL-Fl both complete. Genomic, exomic and transcriptomic data is publicly available through the GDC (https://ocg.cancer.gov/programs/cgci).

Data from studies of paediatric cancers:
The Therapeutically Applicable Research to Generate Effective Treatments (TARGET) program (https://ocg.cancer.gov/programs/target) was launched in 2006 to develop targets and biomarkers for treating paediatric cancers and to further understand the molecular landscape of childhood can-

cers. There are five cancer types included in this study; Acute Lymphoblastic Leukemia (ALL), Acute Myeloid Leukemia (AML), Neuroblastoma (NBL), Osteosarcoma (OS) and Wilms' Tumour (WT). The research has been carried out by a collaborative team comprised mostly of Children's Oncology Group (COG) members and researchers from the NCI, who worked together to generate, analyze, integrate, and interpret high-quality genomic, transcriptomic, epigenomic and kinomic data which is publicly available through the GDC. There is also the Children's Brain Tumour Tissue Consortium (CBTTC) which is a collaborative project from 32 institutions in the US which share clinical and molecular data from 4842 individuals with brain and spinal cord tumours [90]. St Jude Children's Research Hospital launched the Pediatric Cancer Genome Project study in 2011 with the goal of sequencing genomes of paediatric cancer patients in order to gain better understanding of more than 20 different cancer types [91]. They have whole genome sequencing data from 800 patients and whole exome and whole transcriptome data from 1200 patients.

1.4.2 NGS workflow for detection of somatic variants in cancer samples

Somatic mutations are generally identified using next-generation sequencing (NGS) technologies, typically in tumour samples. Targeted sequencing panels or whole exome sequencing (WXS) panels tend to be used in clinical practice to identify mutations in genes of interest. Targeted sequencing identifies mutations in a subset of genes of particular importance for the disease in question. Limiting the sequenced regions of the genome to these genes allows each position to be sequenced to a high depth (ranging from 100-1000s of reads at each position) [92]. WXS targets the roughly 20,000 protein-coding genes in the human genome and typically sequences to a depth of 100X. [93] WGS is an unbiased sequence technique which covers the entire genome [94]. However, due to the high costs associated with WGS, depths between 30-50X are typically achieved for cancer samples, making it challenging to identify somatic mutations present at a low frequency or in samples with low tumour purity [95]. A description is provided below of a typical WXS workflow. WGS follows a similar workflow with some alterations to the processing steps and WGS-specific tools used for specific steps, such as alignment. The majority of the data used in this thesis came from WXS, which is why we chose to use this as our example workflow here.

Sample Types:
Samples for bulk sequencing are generally preserved by formalin-fixed paraffin embedded (FFPE) or fresh frozen (FF) methods. The FFPE method is used to store tissue for a long period of time [96]. Samples are first fixed with a formaldehyde solution that stops cell metabolism, and then paraffin is used to seal the tissue and reduce the rate of oxidation [97]. The FF technique requires that the sample is frozen in liquid nitrogen 30-60 minutes after surgery [98]. The FFPE method preserves morphology, while FF does not. FFPE is the most common method due to the lower cost, ability to store at room

temperature and the larger time frame to process the sample after surgery [99]. FF needs to be kept frozen because once it starts to thaw the DNA or RNA starts to degrade[100]. FF has the advantage that DNA/RNA is preserved better than FFPE, with FFPE related artefacts observed as C >T mutations in the sample [101]. To identify somatic mutations in a sample matched paired tumour normal samples are generally taken from a patient. The normal sample is usually taken from blood but sometimes normal tissue adjacent to the tumour is used [102]. This allows germline variants to be identified and removed. Occasionally, a tumour sample is taken without a matched normal sample (referred to as tumour only). This makes it more difficult to identify germline mutations and relies instead on databases of common variants or information from a panel of normal samples to remove variants that are likely to be germline [103].

Alignment and preprocessing:
The first step of any bioinformatics pipeline is to assess the quality of the sequence reads output from the sequencer. These reads are typically 75, 100 or 150 base pairs in length and can either be single or paired-end reads. Paired-end reads are sequenced from both ends of the DNA fragment which gives higher quality and better sequence alignment. The format of the files are generally binary base call (bcl) files which are converted to FASTQ files that contain a readout of the nucleotides called for each read along with a quality score for each base. These FASTQ files are assessed using a tool called FASTQC [104] to check that the reads are of sufficient quality for downstream processing. Typically the ends of reads are trimmed using a tool such as Trimmomatic [105] because base quality drops off at the 3' end in the sequencing by synthesis process of Illumina sequencing. To get positional information for the reads, alignment to a reference genome is required. Several tools are available to align reads to the reference genome, with BWA-MEM [106] the most commonly used. After alignment, preprocessing steps are required before variant calling. If amplification by PCR is performed during library preparation, duplicate marking is required to identify reads generated from the same molecule. This could introduce bias in the variant calling step, with some reads over-represented in the results. Depending on downstream processing, an optional local realignment step using GATK [107] or ABRA [108] can be performed to limit errors caused by indels and SNPs. The original alignment step aligned each read separately, but these tools use information from all reads at that location to determine the best alignment of the reads. Base quality recalibration is another optional but highly recommended step. It is performed using the GATK suite of tools [107] and accounts for inaccuracy in base quality scores assigned by the sequencer. Base quality scores are essential for variant calling, and this step improves the accuracy of variant calls. However, it is computationally expensive and time-consuming. Due to improvements in sequencing technologies that have increased the accuracy of base quality scores, some researchers may not include this step in order to reduce analysis turnaround time.

Somatic Variant Calling, Annotation and Filtering:

Typically somatic variant calling is performed with a matched tumour and normal sample from an individual. Variants in the tumour are detected based on comparison to a reference genome and the normal sample to remove germline variants. Variant callers can be position-based callers or haplotype-based callers (examples are given in Table 1.1). Position-based callers directly compare the aligned sequence to the reference base while haplotype-based callers perform a local realignment step that identifies regions of variation and uses these haplotype blocks to identify variants. Mutect2 is the most commonly used variant caller [109] but there have been many studies to compare the sensitivity and specificity of the different variant callers available [110–112]. Best practice guidelines recommend that a consensus-based approach [109] using multiple variant calling tools is used to identify high confidence variants such as the approach used by the mutation calling workgroup who analysed TCGA variant calls [113]. However, this approach increases the processing time of the calls so it depends on whether specificity or sensitivity is important for downstream analysis. After variants have been identified further annotation using tools such as Annovar [114] or VEP [115] are performed to identify the functional impact of variants. Filtering based on the variant frequency and variant consequence is typically performed to reduce the list of candidate variants for analysis by clinicians and researchers [103].

Variant Caller	**Position-based**	**Haplotype-based**	**Cite**
MuTect2		Y	[109]
VarScan2	Y		[116]
Strelka2		Y	[117]
MuSe	Y		[118]
Pisces	Y		[119]
deepSNV	Y		[120]
SomaticSniper	Y		[121]
VarDict		Y*	[122]

Table 1.1: **Available somatic variant callers with details of whether they are position or haplotype based callers.** Y indicates which class the variant caller belongs to. Y* indicates the tool is not haplotype based but has its own inbuilt realignment step.

1.4.3 Methods to identify clonal variants in tumour samples

An important aspect of assessing somatic variants in tumours is identifying those variants that are clonal and subclonal. In terms of therapy for cancer treatment it may be important to identify which variants are clonal and are therefore present in all cells of the tumour in order to decide on genes to target. However, in order to identify variants that could aid in relapse identification of subclonal mutations may be important. While it is possible to use variant frequency to classify a variant as clonal or subclonal, under the assumption that a variant with a frequency of 0.5 is clonal, this is not always accurate due to confounding factors such as tumour purity, copy number alterations (CNA) and intra-tumour heterogeneity (ITH).

A number of tools are available to identify clonal and subclonal variants in bulk sequencing tumour samples. These are outlined in Table 1.2. Tools can be classified as multi-region bulk sequencing or single sample bulk sequencing. Multi-region tools have the advantage that they can account for ITH. By sampling from multiple sites within a tumour it is easier to identify true clonal mutations and not suffer the "illusion of clonality" problem of single sample bulk sequencing. Illusion of clonality occurs due to sampling bias which means a subclonal mutation can appear clonal because it is present in all cells from the sample but is not present in all cells of the tumour [74]. However, it is rare that a researcher would have multiple samples of the same tumour with single sample bulk sequencing more common. Copy number alterations can impact estimation of the cancer cell fraction (CCF) required to call a clonal variant [74](estimation of CCF is discussed in more detail in Chapter2). The majority of tools take copy number status of the variant loci into account when calculating CCF (references in Table 1.2). The ploidy is either estimated by the tool itself, as in the case of PureCN, or copy number information must be supplied as input (Table 1.2). An additional factor to consider when calculating CCF for a variant from bulk sequencing data is tumour purity. Normal cells are present in samples taken from a tumour for bulk sequencing which can dilute the variant frequency of true clonal variants [74]. Therefore, the tumour purity of a sample should be accounted for when classifying a variant is clonal or not. Some tools infer tumour purity or take tumour purity estimates as an input variable while others do not account for this at all (Table 1.2).

Tool	**Input Sample**	**Comments**	**Cite**
PureCN	Single	Infers CNA and purity	[123]
PyClone	Single or Multi	Uses CN information from other tools, accounts for normal contamination	[124]
DPClust	Single or Multi	CNA and purity as input	[47]
FastClone	Single	CN data and purity as input	[125]
CliP	Sinle	CNA and purity as input	[126]
CloneFinder	Multi	Does not account for CNA	[127]
TrAp	Single	Does not account for CNA, accounts for normal contamination	[128]
PhylogicNDT	Single or Multi	CNA and purity as input	[129]
SciClone	Single or Multi	CNA and purity as input	[130]
CITUP	Multiple	Does not account for CNA, estimates purity	[131]
SCHISM	Multiple	CNA and purity as input	[132]
Chimaera	Multiple	CNA and purity as input	[133]
LiquidCNA	Longitudinal liquid biopsies	CNA and purity estimation	[134]
QuantumClone	Muli	CNA and purity as input	[135]
SuperFreq	Multiple	CNA as input	[136]
MOBSTER	Single, multi and longitudinal	CNA and purity as input	[137]

DEVOLUTION	Multiregion	CNA as input. Uses genotype data	[138]
THEMIS	Multiple	CNA as input	[139]
Canopy	Multi	CNA and purity as input	[140]
AncesTree	Multi	Does not account for CNA, estimates purity	[141]
BayClone	Single	CNA as input	[142]
Clomial	Multi	Does not account for CNA or purity	[143]
cloneHD	Multi or longitudinal	CNA as input	[144]
LICHeE	Multi	Does not account for CNA or purity	[145]

Table 1.2: **Tools that determine clonal status of SNVs using targeted or whole exome sequencing data.** CNA= copy number alteration, CN= Copy Number, purity=tumor purity.

1.5 The role of antigen presentation in immunoediting and the subsequent impact on immunotherapy efficacy

The immune system has evolved primarily to protect us from microbial threats, preserve our integrity, and avoid death [146]. Humans have developed sophisticated innate and adaptive mechanisms for this purpose. The innate system can recognise a threat that has not been encountered before, while the adaptive system only recognises specific threats and also has a memory so it can recall prior exposure [147]. The role of the immune system in cancer has been debated since the 1800s, but it is now well established that the immune system actively removes cancerous cells before they have a chance to take hold [148]. The first evidence that the immune system played a role in preventing cancer came from treating cancer with a toxin known as Coley's Toxin, which caused the tumor to disappear [149]. The response to the toxin was assumed to be caused by immune cells attacking and eliminating the tumor cells. However, not all patients responded to this treatment, and the modality was unknown, causing scientists to question the importance of the immune system's role in preventing cancer, until more recently when additional research has provided further evidence to support Coley's principles (as discussed below) [150].

1.5.1 Cancer cell antigens and the major histocompatibility complex

The immune system can recognise aberrant cells and remove them from our body. The major histocompatibility complex (MHC) which is expressed in all nucleated cells plays a major role in this process [147] because it can bind non-self peptides and present them on the cell surface for possible recognition and removal by immune cells [147]. Alterations to the DNA sequences can generate peptides that are different from peptides present in the normal cell (self-antigens). There are three ways in which non-self antigens can arise in a cancer cell: 1) a mutation in the DNA sequence that changes the protein in the cancer cell compared to the normal cell (neoantigen), 2) overexpression of a protein or expression of the protein in a cell type that does not usually express it, 3) post-transcriptional modification such as alternative splicing, glycosylation or phosphorylation [151]. Here, we focus on neoantigens derived from somatic mutations. The first evidence that neoantigens play an important role in T cell response to tumours came with the identification of tumour-associated RNA transcripts that allowed tumour cells to be recognised by specific T-cells [152, 153]. Further studies showed it is possible to identify potential neoantigens that may elicit a T-cell reaction using cancer genomic data [154, 155], and that both CD4 and CD8 T-cells respond to neoantigens in many cancers[156–158]. Additionally, neoantigens are proving to be good targets for T-cell mediated therapies such as immune checkpoint inhibitor (ICI) therapies and neoantigen-specific T-cell reactivity therapies [151].

The MHC is a cluster of genes on chromosome 6 that encode the proteins involved in antigen presentation to T-cells [159]. Human Leukocyte Antigen (HLA) is the term for the human MHC [160]. There are two MHC classes, MHC-I and MHC-II. The human MHC is polygenic, meaning there are multiple genes within each class, and polymorphic, meaning there are multiple variants of each gene. There is a high degree of sequence variability within genes between individuals. There are 3 MHC-I genes (HLA-A, HLA-B and HLA-C) and 3 MHC-II genes (HLA-DR,HLA-DP and HLA-DQ) [160]. MHC-I genes are expressed on nucleated cells, while MHC-II genes are expressed on antigen-presenting cells (APCs) such as macrophages, dendritic cells, and B-lymphocytes [159]. Dendritic cells can activate both CD4 and CD8 T-cells, while other specialised APCs such as macrophages and B-lymphocytes can only activate CD4 T-cells [147]. The role of the MHC is to display antigens present inside a cell on the cell surface, potentially targeting them for destruction upon recognition by T-cells. The MHC classes detect peptides originating from different compartments [159]. MHC-I detect peptides generated in the cytosol of the cell while MHC-II detects peptides that have been generated from phagocytosis of an extracellular protein [147].

Cytosolic proteins are processed by the MHC-I pathway (Figure 1.2 left hand panel) [147]. Upon synthesis by the ribosome, proteins are released into the cytosol, where the proteasome degrades defective proteins [161]. MHC-I molecules are formed in the lumen of the endoplasmic reticulum (ER) [162]. Peptides produced by the proteasome are delivered to the ER through the transporter associated with antigen processing (TAP1 and TAP2) where they can bind to the MHC-I [163]. MHC-1 molecules test the binding of peptides until they find one that is the right size and can stably bind to the MHC-1 molecules [162]. Peptides that bind to the MHC-1 are usually 8-10 amino acids in length and they bind to both ends of the MHC-1 molecule [164]. The MHC can bind many, but not all, antigens and can bind them with varying degrees of affinity [165]. Once a peptide is bound then the peptide:MHC-1 complex is transported to the cell surface for detection by the T-cells [162].

Similarly, extracellular proteins are processed by the MHC-II pathway (Figure 1.2 right hand panel) [147]. These proteins have been engulfed by the APC into an intracellular vesicle called an endosome. This endosome fuses with a lysosome in the cell which is acidic and causes the protein to break down into peptides [147]. MHC-II molecules are present in the ER. Once they are synthesized they interact with the Ii protein (CD74) which prevents other ER proteins from binding to the MHC-II [166]. The Ii protein chaperones the MHC-II molecule out of the ER through the Golgi apparatus and forms a new endosome [166]. The acidic environment of this endosome breaks down the Ii protein so that only a fragment called CLIP remains bound to the MHC-II molecule groove [167]. HLA-DM then interacts with the MHC-II molecule, releasing CLIP [167]. The structure containing the MHC-II molecule fuses with the endo-lysosome allowing the MHC-II molecule to interact with the fragmented peptides. Like MHC-I, MHC-II molecules will only bind to specific peptides. Once bound the peptide:MHC-II complex is transported to the cell surface for detection by CD4 T cells.

T-cells have receptors (TCRs) that are capable of binding the peptide:MHC complex [168]. There are two classes of T-cells, CD4 and CD8 T-cells. T-cell development takes place in the thymus and only those T-cells that have receptors that do not bind strongly to self-antigens can differentiate into mature T-cells [168]. These naïve T-cells then enter lymphoid organs where they can come into contact with APCs that have antigens presented on their cell surface [169]. TCRs are heterogeneous which means they can bind many different peptides [170]. The specificity of T-cell recognition involves both the peptide and the MHC molecule. CD4 T-cells, also known as helper T-cells (Th), bind to MHC-II molecules while CD8 T-cells, cytotoxic T-cells (CTLs), bind to MHC-I molecules. Once a T-cell binds to the peptide:MHC complex, it becomes activated and clonally expands and differentiates [169]. Activated CTLs release cytotoxins leading to programmed cell death of the target cells, while activated Th cells recruit other immune cells to attack the target cell [147].

1.5.2 Immune-Surveillance and Immunoediting

In 1906 Ehrlich postulated that the immune system has the ability to recognise and eliminate cancer cells [171]. However, it was not until half a century later that work from Burnet and Thomas gave the first evidence that the immune system is involved in preventing cancer[172, 173]. The large number of mutations present in some cancer types which had the potential to activate the immune system had confused researchers because if the immune systems was playing a role in preventing cancer then these cells should have been eliminated by the immune system. Thomas and Burnet postulated that tumour cells might continue to develop and provoke an immune response that would clear the tumour without hint of its existence, a process termed “immune surveillance” [172, 173]. This theory suggests that the immune system is constantly monitoring cells and once detected destroys cancerous and precancerous cells before they have a chance to grow. It is only when the cancer cells evolve and obtain mechanisms to avoid detection that they are then able to proliferate. In 2002 Dunn *et al.* expanded this theory and introduced the concept of immunoediting whereby the immune system shapes the mutational landscape of the tumour such that the tumour is composed of mutations that are successfully able to evade the immune system [174]. There are three stages in the process of immunoediting (Figure 1.3); 1) Elimination phase; 2) Equilibrium phase, 3) Escape phase.

1. Elimination phase \Immunosurveillance: This phase encompasses the early ideas of immunosurveillance in which tumour cells are suppressed due to the constant detection and removal by innate and adaptive immune cells. As a tumour cell forms, it is recognised and removed before it has a chance to grow and proliferate, leaving no evidence of its existence behind. This made it difficult to prove the existence of the theory of immune-surveillance because there was no way to observe it in the tumour. Early mouse studies showed that chemically induced tumours could be immunogenic in genetically similar mice [175, 176]. If a mouse immunised with irradiated tumour cells is subsequently injected with viable cells of the same tumour, the tumour rejection antigens were eliminated while they were not removed when tumour cells of a different cell type were injected into the mouse [175, 176].

   A lot of additional research has been carried out using animal models since then to provide evidence for the role of the immune system in preventing cancer. Several studies in immunocompromised mice (RAG gene knockouts) demonstrated the process of immunoediting in practice [177, 178]. The purpose of the RAG gene is to introduce double-strand breaks in lymphocytes in order to initiate V(D)J recombination [179]. This is an important step for creating diverse antigen receptors in mature lymphocytes. When this gene was knocked out, the mouse lacked mature T cells and was immunodeficient [177]. It was also shown that RAG knockout mice were more prone to cancerous tumours when exposed to carcinogens compared to wild-type mice. Additionally, they were more likely to get spontaneous tumours [178]. The protective effects seen when a mouse was injected with irradiated cells before injection with viable cells were not seen when the experiments were performed on immune-deficient mouse models [178].

   Studies using immune-suppressed mice showed they were at a higher risk of developing tumours due to chemically induced and spontaneously arising mutations than their non-immune compromised counterparts [178]. Additionally, transplanting tumour cells from immunodeficient mice into mice with a fully functioning immune system showed that the tumour cells were effectively removed but tumour cells where not removed in these mice when they were derived from syngenic immunocompetent mice [178]. Shankaran *et al.* showed that rejection of the tumour cells was due to the activation of T-cell immune responses [178]. Further evidence came from the observation that humans with primary

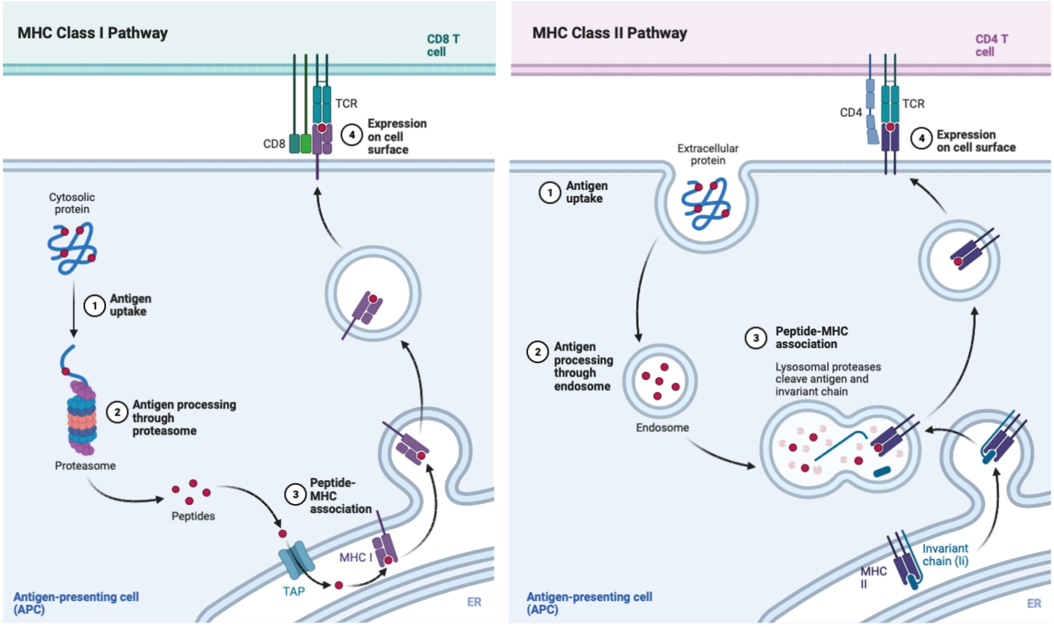

MHC Class I Pathway
CD8 T cell
TCR
CD8
4 Expression on cell surface
Cytosolic protein
1 Antigen uptake
2 Antigen processing through proteasome
Proteasome
Peptides
3 Peptide-MHC association
MHC I
TAP
Antigen-presenting cell (APC)
ER
MHC Class II Pathway
CD4 T cell
CD4
TCR
4 Expression on cell surface
Extracellular protein
1 Antigen uptake
2 Antigen processing through endosome
Endosome
3 Peptide-MHC association
Lysosomal proteases cleave antigen and invariant chain
Invariant chain (Ii)
MHC II
Antigen-presenting cell (APC)
ER

Figure 1.2: **Schematic of the major histocompatibility I (MHC-I) and major histocompatibility II (MHC-II) pathways..** MHC-I pathway (left hand panel) and the MHC-II pathway (right hand panel), illustrating the process of antigen uptake (**1**), processing (**2**), binding to MHC (**3**) and presentation on the cell surface (**4**). TAP=transporter associated with antigen processing, TCR= T-cell Receptor. Reprinted from "MHC Class I and II Pathways", by BioRender.com (2023). Retrieved from https://app.biorender.com/biorender-templates.

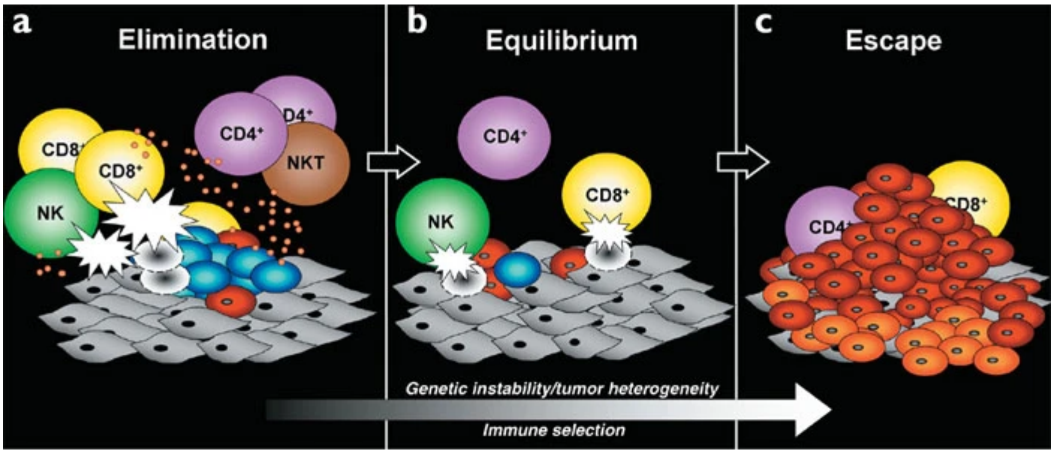


Figure 1.3: **The three phases of tumor immunoediting.** (**A**) Elimination phase: Developing tumor cells (blue) and tumor cell variants (red) are recognised and removed by immune cells (white flashes), (**B**) Equilibrium phase: the immune cells do not eliminate all tumor cells but keep their levels at bay, (**C**) Escape phase: the immune system can no longer recognise the tumor cells which are now free to grow uncontrollably. Additional tumor variants have formed (orange). Different lymphocyte populations include CD4 (purple) and CD8 (yellow) T cells, NK= Natural Kill cells (green), NKT= Natural Killer T-cells (brown). Grey cells=underlying stroma and non-transformed cell, Small orange circles= cytokines. Reprinted with permission from Dunn *et al.* 2002.

immune deficiencies [180–182], HIV-infected patients [183, 184] and immunosuppressed transplantation patients [185–190] were at increased risk of developing tumours. These patients often develop viral-related cancers, with non-Hodgkins lymphoma caused by the Epstein Barr Virus the most common [191–193]. However, non-viral cancers also occur at an increased rate [184, 194–196]. Thus, highlighting the role of the immune system in protecting against such cancers. A population-based study of patients who received transplants in Finland spanning 30 years showed that cancer risk decreased in the later period, which was associated with changes in immunosuppression regimes [197].

2. Equilibrium phase \Immune Selection: The equilibrium phase is the process by which immune cells do not completely eradicate the tumour but keep the numbers at bay so it cannot grow and progress. This is termed tumour dormancy and may last many years [198] with the equilibrium phase believed to be the longest of all the phases. It is during this phase that the immune system shapes the mutational landscape of the tumour cell. Cells that harbour neoantigens that fail to be recognised by the immune system are retained while cells with immunogenic neoantigens are selectively removed [199]. During this phase there is a balance between stimulation and inhibition of the immune system depending on the presence of pro or anti immune molecules present at each stage of the immune life cycle [200]. The first evidence for this immune selection of cancer cells came from injecting mice with carcinogens [198]. Those mice that had stable masses of cells which did not transform into a tumour initially, subsequently had components of their immune system disabled [198]. This resulted in the formation of tumours, suggesting the immune system played a role in preventing the tumour [198].

   Immune selection is difficult to assess in practice because the levels of tumour cells are below the level of detection until they have evaded the immune system and are able to grow into a large mass of cells [198]. However, there have been studies that have used predicted MHC binding affinity of neoantigens to identify evidence of immune selection. Support for immune-mediated negative selection in tumours came from studies that used the ratio of nonsynonymous mutations to synonymous mutations (dN/dS) to detect regions of negative selection in the tumour genome [201, 202]. Human epitopes (the part of the antigen that binds to the T cell) showed stronger evidence of negative selection compared to non-exposed regions of the same protein, with epitopes bound to common HLA alleles showing stronger negative selection than epitopes bound to rare HLA alleles [201]. Additionally, dN/dS scores were negatively correlated with cytolytic activity of immune infiltrates which the authors proposed was evidence of immune selection [201]. A more recent study using dN/dS to measure immune selection in cancer patients, classified patients with negative selection as "immune edited" and patients who had a missense or a truncating mutation in a gene involved in antigen presentation as “immune-escaped” [202]. Those patients that were immune-escaped were no longer under the same selection pressures as patients that were not immune-escaped and showed evidence of neutral neoantigens [202]. Immune-edited patients were also shown to have low tumour antigenicity and have a poor response to ICI treatment [202]. Other studies investigated the relationship between predicted neoantigens and cytolytic activity in the tumour [203, 204]. Predicted neoantigen load was found to be correlated with cytolytic activity, particularly in cancers with a viral contribution [203]. Additionally, these tumors showed evidence of escaping the immune system with enrichment of mutations within the antigen presentation machinery and over-expression of immunosuppressive genes that protect against CTL mediated destruction observed in these tumors [203]. Two studies using predicted MHC-I and MHC-II binding affinity of observed neoantigens in TCGA samples found evidence of selection based on patient HLA type for driver mutations. They proposed that common driver mutations are common because common HLA alleles cannot bind them and present them for removal by the immune system[205, 206]. These results have been questioned by us, as discussed in Chapter 3, [207] and others [208], due to the author’s misinterpretation of results and failure to account for the trinucleotide context, which gave a false indication of negative selection. While debates on this are ongoing [209–212], the consensus seems to be that prediction of MHC binding affinity is not sufficient to detect negative selection. Incorporating the probability of being recognised by TCRs with predicted MHC binding affinity was shown to be a better predictor of survival after immunotherapy treatment than neoantigen load alone [204]. Using this scoring method meant that a neoantigen with a high score would have a high predicted binding affinity and low similarity to self-antigens [204].

3. Immune Escape: Cancer that can grow has developed mechanisms to escape the immune system. They can do this by engaging the signals that prevent activation of the immune system . These signals are referred to as immune checkpoints and when they are engaged, they will prevent the T cell from killing [213]. The cancer cell can engage these checkpoints but they can also be activated by dendritic cells and macrophages [214]. This is because the immune system has learned to constantly modulate its ability to respond and regulation of this response is as important as activation in order to prevent immune cells from attacking healthy normal cells [147]. While dendritic cells send signals to activate the T cell to kill, they also modulate T-cell ability to react against a threat [215]. There is a fine balance between activation and inhibition, and in some cases, the tip to inhibition has gone too far so that the T cells are suppressed and cannot act on the cancer cell to remove it [216].

   Avoiding immune destruction is one of the hallmarks of cancer [217]. Cancer cells counteract the ability of the immune system to eliminate them, allowing them to grow and progress. In order for the immune system to impact tumour growth and shape the tumour genome, there needs to be a fully functioning antigen presentation machinery (APM) and also the presence of immune cells capable of recognising and destroying the cancerous cells within the tumour microenvironment. To avoid detection cancer cells have acquired ways to disrupt the APM. Recurrent mutations in key players, such as MHC-1 and β2-microglobulin (B2M), are common in cancers, with downregulation of the genes more frequently observed [218, 219]. B2M is a component of the MHC that is required for its formation and stabilization on the cell surface. The MHC cannot form without B2M and loss of this gene is a form of immune escape in cancers [220]. However, mutations in this gene are rare with downregulation of the gene more common [219]. HLA-G, a non-classical HLA molecule, is an immunosuppressive protein that inhibits natural killer cells and CTLs and plays a role in pregnancy tolerance by protecting the fetus from attack by maternal NK cells [221]. It is usually expressed in immune-privileged tissues but is also frequently over-expressed in tumours [222]. Additionally, TAP1 is downregulated in colorectal cancer as a means to escape immune recognition [223]. Downregulation of key players in the APM inhibits the ability of the immune system to identify antigens on the cancer cell surface. Cancer cells also overexpress proteins such as PDL1 and NF-kb, which are checkpoint inhibitors that block the immune system [224–226].

T-cell suppression is also common in cancer. Regulatory T-cells (Tregs), whose normal function is to monitor and regulate effector T cell activity, are present in high levels in several cancer types [227]. Additionally, Tregs in tumours have been shown to have greater suppressive functionality compared to Tregs in normal samples [228, 229]. Cytokines such as TGF-ß produced by the tumour cells themselves or by other cells in the tumour microenvironment can also suppress T-cell function [230].

It has also been suggested that low mutational burden is another means by which the tumour escapes the immune system[231]. With fewer mutations, there is a decreased likelihood of neoantigens that will elicit an immune response. The theory behind this is that those tumour cells that had neoantigens capable of eliciting a strong immune response would have been removed by the immune system, leaving behind cancer cells that can fly under the radar [199, 232]. In this sense, decrease in tumour immunogenicity is a method to escape immune detection. However, recent work has questioned this, showing there is limited evidence for depletion of neoantigens in cancer samples [233].

HLA loss of heterozygosity (LOH) is another proposed mechanism of immune evasion, with loss of HLA-C08:02 which is required for KRAS G12D neoantigen detection observed in a lesion from a tumour that was resistant to treatment with CD8+ cells targeting mutant KRAS [234]. Subsequently HLA LOH was shown to be prevalent in lung cancer, occurring in 40% of early-stage NSCLCs [235].

1.5.3 Response to Immunotherapy

The premise of the development of immunotherapies is to reactivate the immune system so that it can fight the cancerous cell as it had been able to do prior to cancer progression. It was discovered that inhibition of the checkpoint blockade, enables T cells once again to target the cancer cell and remove it [236]. Therefore, instead of targeting the cancer cell itself, it is possible to target the immune system, which allows us to predict how the immune system will respond because it is so heavily regulated. Current treatments perform a combination of targeting the cancer cell as well as targeting the tumour microenvironment. Currently, only three checkpoint blockade molecules are targeted by therapies, CTL4, PD-1 and PDL1 [237], but there are many other molecules that could also be targeted and are the focus of ongoing research [238]. ICI treatments are only effective in about 20-30% of patients and their efficacy varies among cancer types [239]. Due to the high costs associated with immunotherapy treatment current research aims to identify biomarkers that will predict which patients would respond well to immunotherapy. Currently, there are only three FDA-approved biomarkers of response to immunotherapy; PDL1, microsatellite instability (MSI) and tumour mutational burden (TMB)[237].

Programmed Cell Death Protein 1 (PD-1) is found on the surface of T-cells and, when it is bound to its ligand, PDL1 or PDL2, it prevents T-cells from attacking other cells [240]. Anti PD-1 therapy blocks the interaction of PD-1 with PDL-1 allowing the T-cell to attack cancer cells. PDL1 is highly expressed in some tumours and its expression is correlated with better response in tumours treated with anti PD-1 therapy [241–243]. PDL1 was the first approved biomarker of response to immunotherapy [244]and is the most frequently used in the clinic [237]. However, it has a low predictive potential with a high proportion (estimates of 20-30%) of patients with negative PDL1 responding well to anti PD-1 treatment [245, 246]. This is likely due to biological factors such as the temporal and spatial regulation of PDL1 gene expression [247] and the impact of prior treatment on PDL1 expression [248], as well as technical factors such as differences in antibodies and platforms used to detect PDL1 and differences in thresholds and scoring methods to call PDL1 high [249].

Tumours with defects in the DNA mismatch repair (dMMR) machinery accumulate thousands of mutations and are considered hypermutated [250]. Microsatellite regions are 1-6 nucleotides of repetitive sequences found in the human genome [251]. These regions are particularly prone to mutation in tumours that have defective dMMR [250]. Microsatellite instability (MSI)/deficient dMMR was the first biomarker approved as a companion diagnostic for all solid tumours rather than for a specific tumour type [252]. MSI tumours have increased expression of PD-1 and PDL1 [253] with MSI high (MSI-H) tumours shown to respond well to anti PD-1 treatment [254].

TMB is the number of nonsynonymous mutations per million bases

(Mb) in a sample [255, 256]. However, there is some inconsistency in the definition of TMB in the literature with some publications using nonsynonymous or missense mutations only in the calculation and others, generally studies using targeted gene panels, including indels and synonymous mutations too [257]. TMB is correlated with neoantigen load and as a result was proposed as a potential biomarker for ICI response [258]. The idea behind this is that the higher the mutation load in a sample, the higher the number of mutations with the potential to elicit an immune response. High TMB is associated with a good response to immunotherapy [259] and was the third biomarker for response to pembrolizumab approved by the FDA [237]. Like MSI, this biomarker is not cancer type specific but approved for the treatment of adult and paediatric patients with unresectable or metastatic solid tumours. Multiple studies have assessed the predictive potential of TMB as a biomarker for response to ICI across a wide range of cancer types with mixed results[255, 260] indicating a need for more tumour type specific analysis.

There are many other factors, other than overall mutation numbers, that should be considered when predicting response to immunotherapy [261]. Clonal load ie. the total number of mutations that are present in all cells of the cancer [262, 263] has important implications in terms of response to therapy and likelihood of relapse. The expression level of genes containing the neoantigens is important as only the antigens that are expressed can be processed and presented to the immune system [264]. Another important factor is the binding affinity of those neoantigens to the HLA alleles because this affects which neoantigens will be presented on the cell surface [264, 265]. Neoantigens that are highly different to self or known antigens would also be more likely to be recognised by the immune system [261]. TMB estimates focus on SNV numbers only, but indels are also common and can create neoantigens that are highly distinct [266]. Tumour purity impacts TMB estimates with stromal cells shown to be a confounding factor when assessing TMB [264]. Intra-tumour heterogeneity confounds immune checkpoint blockade response, with mixed results reported for a subset of lung cancer patients [264]. Finally, a recent study [267] argues that in addition to overall TMB, copy number states and the sequence alteration load should be considered when making clinical decisions. Therefore, there are many additional factors that can impact response to immunotherapy and should be considered when developing biomarkers.

1.6 Allele-Specific Expression

Most of this section has been published in: Cleary S and Seoighe C. Perspectives on Allele Specific Expression. Annu. Rev. Biomed. Data Sci. 2021.4:1, doi:10.1146/annurev-biodatasci-021621-122219

Allelic imbalance arises when there is a difference in the states or activities of the alleles of a locus in a diploid (or higher ploidy) organism. Much of the research on allelic imbalance has focused on differences in messenger RNA (mRNA) abundance, which we will refer to as allelic expression imbalance. Imbalance in mRNA abundance between alleles has been referred to as allele-specific expression (ASE) [268]. This term is often used to refer to gene expression imbalance, without regard to whether the difference in expression is due to genetic variants or epigenetic effects, such as imprinting or random monoallelic expression [269, 270]. However, as it is suggestive of an effect that arises from the allele itself, we propose that the term allele-specific expression should be reserved for imbalance with a genetic origin and adopt that usage here. We use the term allelic expression imbalance when the cause of the differences in expression between alleles is not specified. Similarly, we use allelic imbalance to refer to any differences between alleles in chromatin state, expression level or relative isoform abundance and allele-specific imbalance when these differences are genetic in origin.

1.6.1 Mechanisms of Allelic Imbalance and ASE

Genetic variants can have an impact on chromatin structure [271, 272], on gene transcription [271, 273, 274], and post-transcriptional processes (Figure 1.4). In many cases, these variants can affect the expression level of the linked allele, leading to ASE, as well as leading to other measurable forms of allele-specific imbalance. Some of the main mechanisms leading to allele-specific imbalance, highlighting the potential of some of these to give rise to ASE, are:

- Transcription factor binding: The alleles of a heterozygous SNP can have different affinities for a transcription factor resulting in allelic imbalance in transcription factor occupancy [274] and distinct rates of transcription for each allele [275]. Analysis of allele-specific transcription factor binding has played an important role in understanding how non-coding DNA can affect gene expression and contribute to disease phenotypes. In order to dissect fully the implications of altered binding, the causal gene regulatory variant, the transcription factor that binds to it and the target gene should all be identified [276]. Differences in chromatin accessibility can also result from allele-specific transcription factor binding and may make a substantial contribution to complex diseases [272].

- Nonsense-mediate decay: NMD is a key cellular quality control mechanism that results in the elimination of mRNAs carrying premature termination codons (PTCs) that might result in malformed proteins [277]. This process takes place in the cytoplasm and is associated with the termination of translation and mRNA degradation [278]. NMD also plays a role in controlling mRNA expression level, contributing to the regulation of a large number of human genes [279]. A heterozygous SNP at which one of the alleles results in a PTC can result in degradation of mRNA derived from that allele, resulting in ASE [280].

- Alternative splicing: Genetic variants can affect mRNA splicing by altering splicing signals in the transcript. Such mutations can occur within or close to splice donor or acceptor sites, around the branch point or in exonic or intronic enhancer or suppressor sites [281]. Common effects on splicing include exon skipping, intron retention, alternate $3'$ or $5'$ exon ends and mutually exclusive exons [282]. Because they act in cis, transcribed splicing mutations typically result in allele-specific splicing [281, 283]. When a mutation that alters mRNA splicing introduces an in-frame stop codon (e.g. by skipping an exon within the coding region that is not a multiple of three nucleotides in length) it can trigger nonsense-mediated decay [282] targeted towards the affected allele. This results in a lower abundance of the mature mRNA for the allele causing mis-splicing than from the wild-type allele and consequently results in ASE. Even when NMD is not triggered, differences between the protein isoforms resulting from genetic variants that affect splicing can have important functional consequences [284].

- Variants affecting mRNA binding: RNA-binding proteins (RBPs) play a role in post-transcriptional gene regulation by binding to RNA in a sequence specific manner, modulating the fate of the bound RNA. Genetic variants on the mRNA can disrupt the interaction of RBPs with the mRNA, resulting in allelic imbalance in RNA binding and, potentially, ASE or allelic variation in mRNA localization or translation [285]. Application of a method developed to detect allelic imbalance in RNA binding to enhanced crosslinking and immunoprecipitation sequencing (eCLIP-Seq) data from ENCODE revealed genomic variants that alter mRNA splicing as well as gene expression level [286, 287], illustrating the potential of allele-specific RNA binding to cause ASE. miRNAs and long non-coding RNAs (lncRNA) contribute to post-transcriptional regulation of gene expression. These non-coding RNAs can themselves display allele-specific imbalance in their expression, as well as inducing ASE in the genes they regulate [288, 289]. Compared to protein-coding mRNAs, lncRNAs show greater levels of allelic imbalance in their expression [290]. The interaction of miRNAs with their target mRNAs can be affected by SNPs within sites in the mRNA that are complementary to the miRNA [291] and again this is likely to result in allele-specific expression.

1.6.2 Computational pipelines for measuring ASE

Analysis of ASE from high throughput sequencing data typically involves generation of counts of sequence reads mapped to each allele. Generating this data involves multiple steps, as detailed below, each of which is associated with potential biases and confounding factors. Several efficient and scalable pipelines are available for these tasks, such as AlleleWorkbench [293], WASP [294], CloudASM [295] and ALEA [296].

1. Sequencing In order to have sufficient power to discriminate between the expression levels of alternative alleles, analysis of allelic imbalance requires higher coverage than is generated in a typical RNA-Seq ex-

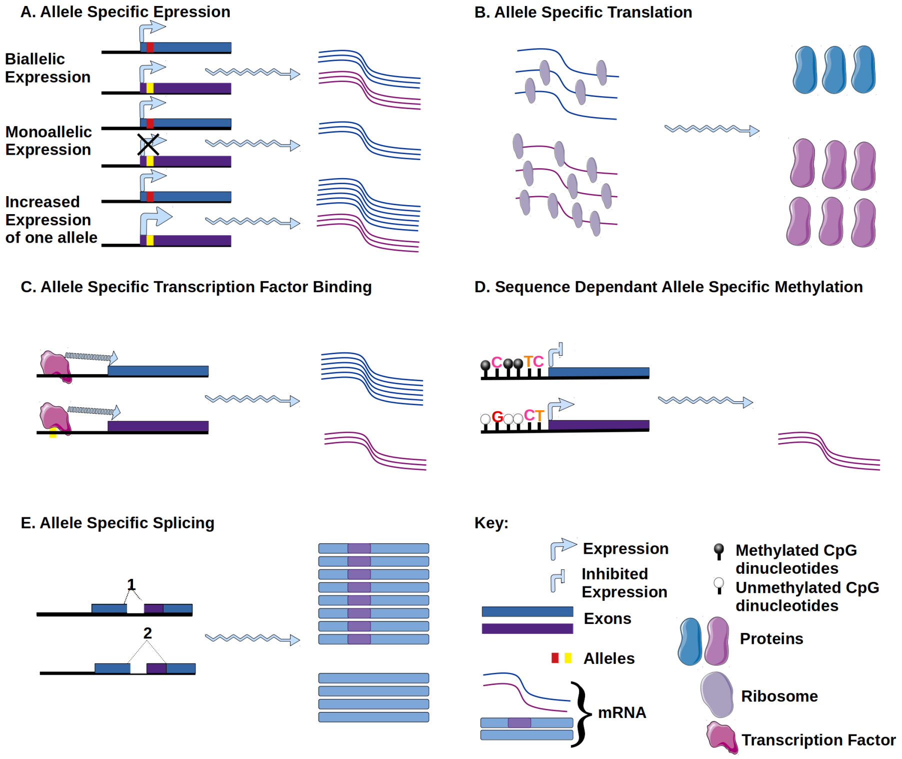

A. Allele Specific Epression
Biallelic Expression
Monoallelic Expression
Increased Expression of one allele
B. Allele Specific Translation
C. Allele Specific Transcription Factor Binding
D. Sequence Dependant Allele Specific Methylation
C
TC
G
CT
E. Allele Specific Splicing
1
2
Key:
Expression
Inhibited Expression
Exons
Alleles
mRNA
Methylated CpG dinucleotides
Unmethylated CpG dinucleotides
Proteins
Ribosome
Transcription Factor

Figure 1.4: **Types of allelic imbalance.** (**A**) Allelic expression imbalance. Three cases are shown: equal expression of both alleles (top); exclusive expression of one allele (middle); higher expression of one allele (bottom). (**B**) Allelic imbalance in translation. Genetic variants can alter the rate of mRNA translation, resulting in different levels of ribosome occupancy between alleles. (**C**) Imbalance in transcription factor binding. In the example shown a sequence variant reduces transcription factor binding affinity resulting in allele-specific expression. (**D**) DNA methylation imbalance: methylation inhibiting the expression of one allele. If the difference in methylation results from cis-acting genetic variants it can lead to allele-specific expression (**E**) Allele-Specific Splicing: A variant that alters splicing results in different isoforms from the two alleles. Reprinted with permission from Cleary *et al.*[292].

periment focusing on total expression analysis [297]. A threshold of 30 reads spanning the location of interest is often applied to infer allelic imbalance at individual heterozygous sites [298, 299]. This can limit the number of genes with sufficient coverage to detect allelic imbalance.

2. Alignment and removal of PCR duplicates One of the first steps in software pipelines for the analysis of allelic imbalance is to align the sequence reads to a reference genome or transcriptome. Errors in the alignment, or mapping, can have a substantial impact on the results obtained [300]. Mapping errors (mapping a read to the wrong location or failure to map a read) can occur with greater frequency for reads containing the alternative than the reference allele at heterozygous SNPs [301], leading to false-positive signals of allelic imbalance. A number of strategies have been proposed to mitigate sequence alignment biases. These include the use of a masked reference [302], personalised diploid genomes [303] or transcriptomes [304], haplotype genomes for alignment [305], the use of SNP-tolerant mappers such as GSNAP [306],STAR-WASP [307], ASE-lux [308] and SNP-omatic [309] and methods that use remapping strategies such as WASP [294]. Methods that align sequence reads to a diploid transcriptome that includes genetic variants have been reported to result in improved estimation of ASE [304].

   The polymerase chain reaction (PCR) amplification step in the preparation of sequencing libraries can result in the same cDNA fragment being sequenced more than once. This results in sequence reads with identical mapping coordinates. Although it is straightforward to identify these duplicate reads and remove them, this is generally not recommended for RNA-Seq data due to loss of information for highly expressed genes. However, statistical tests of allelic imbalance are often not robust to the presence of duplicate reads and therefore potential PCR duplicates should be removed prior to analysis of allelic imbalance [300]. Many tools for removing duplicates retain the read with the best mapping score, but for analysis of allelic imbalance it is essential to use tools,such as WASP [294], that select the retained reads at random, to avoid mapping bias in favour of the reference allele.

3. Genotyping and haplotype phasing Generation of allele-specific read counts requires at least one heterozygous SNP within the targeted feature (gene, transcription factor binding site etc). Heterozygous SNPs can be identified separately using genotyping arrays or genomic DNA sequencing. Alternatively, the heterozygous SNPs can be inferred from the reads that map to the feature of interest. In the case of allele-specific expression, for example, genotype can be inferred from the RNA-Seq reads. However, this carries the risk that features that show extreme imbalance can be mistakenly called as homozygous, leading to false negatives in the inference of allelic imbalance. Conversely, sequencing errors, transcription errors or even rare somatic mutations that result in a site that is homozygous in the germline incorrectly being called het-

erozygous can lead to false positive inference of allelic imbalance. Errors may also occur when genotyping is performed on genomic DNA. In this case, homozygous sites incorrectly called heterozygous can lead to false positive inference of allelic imbalance [294]. More recent methods for the analysis of allelic imbalance take account of uncertainty in genotyping [294, 310, 311].

Accurate SNP phase data supports the inference of allelic imbalance, by allowing reads to be mapped to haplotypes spanning multiple heterozygous SNPs. The information contained in the sequence reads can be used for this purpose, with the higher accuracy obtained when long read data are available [312]. Haplotypes inferred from population phasing can be combined with the information contained in RNA-Seq reads spanning heterozygous SNPs to improve accuracy [313]. However, this tends to be accurate for common variants but uncertain for rare variants.

Allele-specific read counts are the required input for many ASE tools [310, 314–317]. These can be determined for heterozygous SNPs using tools such as ASEReadCounter [300]. However, mapping reads to haplotypes rather than individual heterozygous SNPs provides greater power for ASE analysis [313]. Haplotype-specific expression levels can be estimated from RNA-Seq data using phASER [313] and haplotypes obtained from the RNA-Seq reads can be integrated with population-level phasing using phASER-pop [313] to extend haplotypes to putative regulatory variants in untranscribed regions (Figure 1.5 C). Some tools such as IDP-ASE [312] and BYASE [318] perform haplotyping as part of ASE estimation. For tools, such as EAGLE [319], that take read counts as input it is possible to supply gene level haplotypic counts instead of heterozygous SNP counts [320] as phaser generates one count per gene [313].

1.6.3 Considerations for analysis of ASE in cancer

Somatic copy number alteration (SCNA) can be a confounding factor in analysing allelic imbalance in cancer studies, leading to false positives for ASE [321]. A recent pan-cancer study revealed that SCNAs accounted for 84.3% of the observed allelic imbalance [322]. Some studies address this by filtering positions that overlap with copy number variation [303, 323]. Methods have been developed to take account of copy number variation and tumour purity when assessing allelic imbalance of somatic mutations [324]. Due to the presence of high frequency somatic mutations and copy number alterations in cancer, genotyping is usually based on the normal sample. Comparison of the cancer and normal sample can then reveal the allele that is retained in cancer in the case of loss of heterozygosity, which can be informative about the process leading to cancer development [325]. Alternatively, ASE can be estimated for tumour and normal samples separately and the proportions of SNPs showing ASE can be compared between the two groups

[326]. Other studies have compared the variant allele frequency of heterozygous SNPs in whole exome sequences and transcriptome sequences [327–329] or used the allelic ratios in genomic DNA to correct for the effects of copy number variants [330].

1.6.4 Statistical Methods

A wide range of statistical models have been developed for the analysis of ASE. Broadly, they can be characterized by whether the goal is to detect allelic imbalance within individual samples or to combine data across multiple samples, either to characterize ASE or to use it to help estimate the effects of putative regulatory variants (Figure 1.5). For the former goal the simplest method is to treat the number of reads mapping to the reference (or alternative) allele as a binomial random variable. Several Bayesian methods [312, 331, 332] have also been proposed to analyze ASE within individuals. Methods focused on estimating ASE can be differentiated based on whether they are applied on a gene by gene basis in individual samples, as is the case with the binomial test and also some more specialist methods [312, 333], or whether they attempt to learn model parameters by considering multiple genes simultaneously (Exemplified by Skelly *et al.* [331], McCoy *et al* [332]). Of particular note has been the development of models designed to learn about the effects of regulatory variants by combining ASE with variation in gene expression levels across individuals [294, 311, 334]. Building on these, recent work has leveraged ASE to estimate the expected variance in gene expression for human genes, with important implications for understanding genetic disease mechanisms [335].

1. Binomial model and its limitations Applied to individual heterozygous SNPs, it is straightforward to evaluate a null hypothesis that a randomly sampled sequence read has the same probability of being generated from the reference or alternative allele. This null hypothesis can be modified to account for mapping bias in favour of the reference allele [301] by setting a slightly higher probability of a read being generated from the reference allele, under the null hypothesis of no imbalance [298]. Further improvements in power can be obtained by mapping reads to phased haplotypes rather than to individual heterozygous sites and information within RNA-Seq reads, including allelic imbalance, can be leveraged to obtain phased information even for rare variants[312, 313, 336]. Statistical models have also been developed for joint inference of heterozygous SNPs and detection of ASE from RNA-Seq reads [310]. In common with many other methods to infer ASE (Exemplified by Liu *et al.* [316]), the latter method uses a likelihood ratio test to evaluate a null hypothesis corresponding to equal representation of alternative alleles, while accounting for uncertainty in the inferred genotypes.

   Inference of allelic imbalance using the binomial test, and its variants, has several major caveats. Allele-specific count data tends to be overdispersed, relative to the binomial distribution, meaning that the variance in the count of reads mapping to an allele is higher than expected for

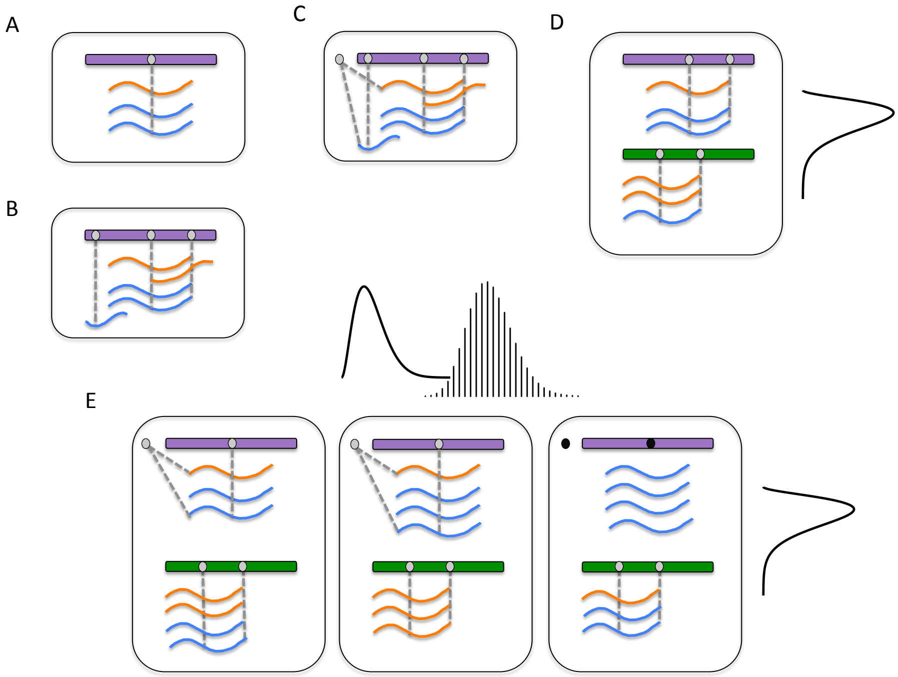


Figure 1.5: **Illustration of the types of statistical models used in the analysis of allelic expression imbalance.** Boxes represent individuals. Filled grey circles represent heterozygous SNPs and black circles represent homozygous SNPs. Sequence reads mapped to alleles of (**A**) a single heterozygous SNP or (**B**) haplotypes spanning multiple expressed heterozygous SNPs can be tested for unequal representation of the two alleles. (**C**) Haplotypes can be extended to putative regulatory SNPs when population-based phasing is taken into account. If data from multiple individuals are available this allows the extent and direction of expression imbalance to be correlated with the allele at the putative regulatory SNP. (**D**) Statistical models can learn parameters of distributions describing the variation of ASE across genes within a single sample. (**E**) Models can combine evidence from ASE in heterozygous individuals with gene expression level in all individuals. These models include distributions describing allelic expression ratios across SNPs in the same gene and across different genes as well as distributions for total expression level in different individuals. Reprinted with permission from Cleary *et al.* [292]

a binomial random variable [294, 300]. This overdispersion is likely to have both biological causes, reflecting a high prevalence of true allelic imbalance, as well as technical causes. It is possible to treat the number of reads derived from one of the alleles or haplotypes as a beta-binomial (or a binomial-logit-normal [335]) instead of a binomial random variable [294, 300, 333]. The beta-binomial is a two-parameter distribution that arises when the parameter of a binomial random variable is itself a beta-distributed random variable. It can be parameterized with a mean and an overdispersion parameter [333], with the latter controlling the extent of the increase in variance relative to the binomial parameter. However, if the overdispersion is primarily biological in origin, reflecting a high frequency of allelic imbalance, including an overdispersion parameter estimated from the data in the null hypothesis may result in a reduction in power to detect ASE.

One of the technical sources of overdispersion is the presence of duplicate reads, but this can be addressed by removal of duplicates as discussed previously, or through the use of molecular barcodes [337]. A lack of reproducibility of allelic imbalance results between technical replicates has recently been reported and interpreted to suggest that other steps in library preparation may be more important sources of bias than PCR amplification for allelic expression analysis [338]. This lack of reproducibility is in contrast to earlier results, obtained from technical replicates in the Geuvadis study, which suggested that the variance across technical replicates was similar to its expectation under the binomial distribution following implementation of quality control steps [300]. A key shortcoming of hypothesis testing for allelic imbalance is that it places the emphasis on evaluating a null hypothesis, which may be unrealistic and sensitive to sequencing depth, rather than on estimating the extent of the imbalance between alleles. Lastly, methods to detect allelic imbalance in single individuals cannot easily distinguish between genetic and epigenetic causes and therefore cannot be used to infer ASE (which as used here implies a genetic origin). Despite the above potential limitations the binomial test remains in use for detecting allelic expression imbalance [339], perhaps due to the ease of interpretation and use.

2. Bayesian models for allelic imbalance Several Bayesian methods have been developed for the analysis of allelic imbalance. Considering data from just a single gene and a single individual, but multiple SNPs, IDP-ASE [312] simultaneously performs haplotype reconstruction and inference of allelic expression imbalance from RNA-Seq data. Taking a flat prior it samples from the joint posterior probability of the reconstructed haplotypes and the probability that a random read is derived from one or other of the haplotypes in an individual. Skelly *et al.* [331] developed a hierarchical Bayesian model for allelic imbalance that considers data from multiple genes simultaneously (Figure 1.5 D). This was first used with RNA-Seq data derived from crosses of *Saccharomyces cerevisiae*

strains and data from a single human cell line [331]. The study also included genomic data, which allowed technical artifacts, such as mapping bias, to be taken into account. The model for the RNA-Seq data consisted of a mixture prior with a component corresponding to allelic imbalance genes and another for non-allelic imbalance genes, for which the allele-specific read counts have the same distribution as in the genomic data. For the imbalance component, allele-specific read counts in a given gene were modelled using a beta-binomial, parameterized with the expected value and overdispersion. Across all genes, both the expected value and overdispersion were themselves beta-distributed, with independent parameters, allowing for genes with variable or relatively constant allelic imbalance across heterozygous SNPs. Markov Chain Monte Carlo (MCMC) was used to obtain samples from the joint posterior distribution of the proportion of genes with imbalanced expression, expected value and overdispersion of the imbalance for each gene as well as parameters describing how these vary across the genes with allelic imbalance. An advantage of this Bayesian approach is the capacity to make inferences about the overall proportion of genes affected by allelic imbalance and the effect size distribution across these genes. A Bayesian implementation of a mixed effects binomial regression model was used by the same group to combine information across individuals and across tissues to estimate ASE associated with Neanderthal introgression [332]. The parameter of the binomial distribution describing the number of non-reference reads was modeled as a sum of a fixed intercept term (corresponding to the ASE effect) and random effects for tissue and individual. Recently, Dong *et al.* developed a Bayesian model, together with a Python library [318] to estimate gene and isoform level expression imbalance for any ploidy >1. The authors claim that their method compares favourably to existing methods and gives consistent results across technical replicates. To the best of our knowledge, however, no independent benchmarking has been carried out to evaluate the performance of these methods.

1.6.5 Prevalence of Allele Specific Expression

Several studies have reported the frequency with which allelic expression imbalance is observed [273, 298, 299, 331, 332, 339–341]. As discussed above, there are multiple genetic and epigenetic mechanisms that can lead to allelic expression imbalance; however, most allelic expression imbalance is reported to arise from genetic variation [298]. Therefore, estimates of the overall prevalence of allelic expression imbalance provide an indication of ASE prevalence. There are at least two different quantities that can be considered. The first is the frequency with which the alleles are imbalanced within an individual. This has been estimated by testing heterozygous SNPs for evidence of imbalance [298]. However, rejection of the simple null hypothesis of equal expression of two alleles does not guarantee that the imbalance is biologically meaningful. Any sequence heterogeneity between the alleles may have some effect on gene regulation and rejection of the null hypothesis may then become a question of the precision of the measurement, which tends to be greater for more highly expressed genes. Methods that consider all genes simultaneously and estimate the proportion of imbalanced genes and the effect size distribution are, therefore, preferable [331]. A second measure of prevalence of allelic imbalance that has been reported is the proportion of genes that show imbalance in at least some subset of individuals, when data from a cohort of individuals is analyzed. Given a large enough sample of individuals, high sequencing depth and samples from sufficient tissues, this proportion is likely to approach one, and it therefore requires thresholds on the strength of imbalance and the proportion of individuals displaying imbalance in a particular tissue type [339] to be meaningful.

1. Divergent reports of ASE frequency In 2002, Yan *et al.* [342] developed an experimental method to assess differences in expression between alleles of heterozygous SNPs and applied the method to data from 13 genes in 96 individuals from the Centre d'Étude du Polymorphisme Humain (CEPH) pedigrees. For six of these genes, there was evidence of allelic imbalance, and this imbalance followed a pattern consistent with Mendelian inheritance. This was followed in 2003 by an estimate of the prevalence of ASE in human using microarrays [273]. Of 602 genes that could be tested, 54% showed evidence of allelic expression imbalance. Using reciprocal crosses of two mouse subspecies and a method based on consistent rejection of the null hypothesis of balanced allelic expression (p-value < 0.05) across replicates, Pinter *et al.* [340] estimated that 20% of mouse genes show evidence of allelic expression imbalance in any given tissue. The majority of the imbalance resulted from genetic effects rather than imprinting or random monoallelic expression. By crossing inbred mice from three subspecies and applying a slightly different method that also focused on rejection of the null hypothesis of balanced expression, Crowley *et al.* [341] reported that over 80% of genes showed evidence of allelic imbalance. Using Bayesian modeling Skelly also estimated a high frequency (80%) of ASE in a hybrid of two diverse Saccharomyces cerevisiae strains [331]. Applying the same method to a single human cell line, they estimated a frequency of approximately 20% of allelic imbalance [331]. Studies that have investigated allelic imbalance in humans rely on standing genetic variation, rather than crosses of divergent strains and the prevalence of ASE may therefore depend on the heterozygosity of the individual. Data from human lymphoblastoid cell lines, generated by the Geuvadis consortium [298], suggested that 6.5% of human genes show evidence of ASE, again using a binomial test (with a significance level of 0.005). A similar frequency of ASE (390 out of 6385 sites interrogated, or 6.1%) was reported by the pilot study of GTEx [299], using the same p-value threshold. This was reduced to 2.3%, when reads were downsampled to achieve a common sequence depth of 30 reads. This decrease by nearly a factor of three illustrates that the reported frequency of ASE based on statistical hypothesis tests

is not a reliable indicator of the underlying prevalence. Estimation of the prevalence requires parameterized models, such as those described earlier, that can provide estimates of the proportion of genes affected within or across individuals and the distribution of the effect size. If the estimate of 20% for the weight of the allelic imbalance component in the model of Skelly *et al.* [331] referred to above is reasonable, this suggests that locus-specific tests may fail to detect a substantial proportion of ASE. This may be due to limitations in sequencing depth and insufficient power to detect weaker ASE effects.

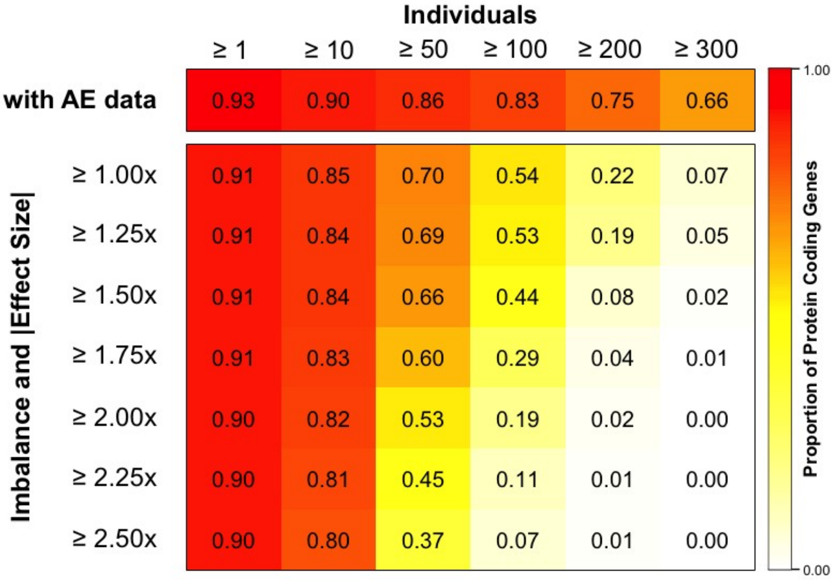
Approximately 25% of heterozygous SNPs that tag an introgressed haplotype from Neanderthals showed evidence of ASE [332]. In some sense, this resembles a natural experiment analogous to the reciprocal crosses that were used to estimate ASE prevalence in mouse [340, 341], except that the crossed populations are outbred and the data are collected many generations after the hybridizations, so that the introgressed segments may have been affected by evolutionary selection. Interestingly, there was no significant difference in the prevalence of ASE between heterozygous SNPs that tagged a Neanderthal allele compared to other heterozygous SNPs matched for minor allele frequency. This is surprising, given that the Neanderthal alleles should be associated with more divergent regulatory regions, creating more opportunities for allelic imbalance. The lack of a difference was interpreted as evidence of post-introgression purifying selection acting on variants that affect gene regulation [332]. However, it is worth noting that the comparison involves Neanderthal haplotypes that are at low frequency in modern humans, potentially due to the relatively small contribution of the Neanderthal introgression and modern human haplotypes at comparable frequencies, some of which will have been suppressed by purifying selection in modern humans. Although no differences are reported in ASE prevalence between introgressed and non-introgressed haplotypes, a cross-tissue analysis suggested lower relative expression of Neanderthal haplotypes in brain and testis, compared to other tissues [332].

2. Survey of ASE across tissues and over time Generation of RNA-Seq data from over 838 individuals across 49 human tissues by the GTEx consortium [343] has provided a real opportunity to gain insights into the prevalence and patterns of ASE. Analysis of the most recent release of GTex suggested that a very high proportion of genes show evidence of ASE in at least some of the samples [339]. Among protein-coding genes, 53% showed evidence of strong ASE (at least two fold difference in expression between the alleles) in at least 50 individuals in at least one of the 49 tissues, (Figure 1.6). Given the mean number of samples per tissue (311) this corresponds to strong imbalance in a substantial fraction of the samples. Note that these results show that most genes can be affected by ASE, but does not translate easily into an estimate of the probability that a given gene will show expression imbalance in a given sample.

Analysis of the prevalence of ASE across samples suggested some differences across GTEx tissues, with testis having the largest number of genes with detected imbalance, though this appeared to have been driven largely by the number of expressed genes [339]. An earlier analysis of whole-blood RNA-Seq data from 65 individuals at age 70 and at age 80 from the Prospective Investigation of the Vasculature in Uppsala Seniors (PIVUS) cohort [344] suggested a small (2.7%) but statistically significant increase in the prevalence of ASE with age [345], though there were examples of genes for which ASE tended to decrease as well as increase with age. Many of the genes that showed changes in ASE over time were associated with the immune response and suggested to be involved in the aging process [345]. Changes in ASE with age suggest that it may be valuable to evaluate the frequency and effect size distribution of ASE across genes at a sample level. This is likely to reflect sequence heterozygosity, but given the relationship with age, may also have associations with phenotype or disease risk.

ASE does not occur at the same level across the human genome. A recent study investigating ASE in the GTEx and CARTaGENE cohorts discovered that ASE is less frequently observed in low recombinant regions and that recombination is used as a mechanism to mask deleterious mutations ensuring they are expressed at a lower frequency than the wild type allele [346].

3. Caveats Recent results suggest that ASE is affected by technical artifacts arising most likely during the preparation of sequencing libraries [338]. Mendelevich *et al.* simulated replicate RNA-Seq datasets and found that the differences in allelic imbalance between technical replicates were greater than expected from the simulations. They used this difference to calculate an overdispersion factor, which was found to be relatively stable for a given sample. This was then used as a correction factor for the inference of imbalance, resulting in substantial reductions in false positive rates. The absence of technical replicates makes it difficult to assess the potential of false positives to contribute to the high rates of allelic imbalance reported by the GTEx consortium. However, it is difficult to envision how such technical artifacts could result in strong and consistent signals of ASE across such high proportions of individuals. Encouragingly, the allelic fold changes reported by GTEx from ASE were also highly consistent (Spearman rho = 0.83) with fold changes estimated orthogonally on the basis of eQTL analysis. Interestingly, excluding individuals who were heterozygous for a known eQTL led to a relatively small drop (median of 7.5%) in the number of genes with evidence of ASE in at least one sample of a given tissue [339]. It is therefore possible that some of the ASE that is not supported by eQTLs is artifactual; however, it seems more likely that this result points to a large number of low-frequency eQTLs that have not been discovered.



	≥ 1	≥ 10	≥ 50	≥ 100	≥ 200	≥ 300
with AE data	0.93	0.90	0.86	0.83	0.75	0.66
≥ 1.00x	0.91	0.85	0.70	0.54	0.22	0.07
≥ 1.25x	0.91	0.84	0.69	0.53	0.19	0.05
≥ 1.50x	0.91	0.84	0.66	0.44	0.08	0.02
≥ 1.75x	0.91	0.83	0.60	0.29	0.04	0.01
≥ 2.00x	0.90	0.82	0.53	0.19	0.02	0.00
≥ 2.25x	0.90	0.81	0.45	0.11	0.01	0.00
≥ 2.50x	0.90	0.80	0.37	0.07	0.01	0.00

Figure 1.6: **Proportion of protein-coding genes with allelic imbalance in normal tissue.** The top row gives the proportion of protein-coding genes with allelic imbalance data in at least the number of individuals shown in the column for at least one GTEx tissue. The remaining rows show the proportion of protein-coding genes with statistically significant allelic imbalance (binomial test FDR <0.05) in at least the number of individuals shown in the column in at least one tissue, as a function of the minimum effect size (expression ratio between the alleles) given in the rows. Reprinted with permission from Castel /emphet al. [339] under the Creative Commons licence.

1.6.6 ASE in cancer

A recent study by the PCAWG consortium found that about 10% of allelic expression imbalance found in cancer is attributed to germline regulatory variants while the remaining 90% is due to somatic events [322]. Inherited regulatory variants can increase cancer risk in certain tissue types while acquired somatic mutations on one copy of the gene or chromosome can disrupt gene expression or can change the dosage level due to copy number changes. Additionally, studies of precancerous and normal tissue bordering cancerous cells have implicated the role of AI in the development of cancer. AI is increased in normal tissue adjacent to the tumour compared to distant normal tissue [347]. Although the level of AI is similar in tumour and adjacent tissue, the pattern differs, indicating a large degree of genomic instability in surrounding cells which may have implications for disease progression and response to therapy [347]. Indeed, ASE of PIK3CA has been recently shown to be prognostic in breast cancer [348]. Analysis of precancerous legions with matched lung adenocarcinoma revealed shared chromosomal AI events and highlighted its role in tumourigenesis [349].

1. Germline ASE:

   Germline variants associated with ASE of specific genes were found to increase the risk of certain cancers. The first evidence came from studies of colorectal cancer where it was discovered that the rs6983267 SNP was associated with ASE in the MYC gene [350, 351]. This SNP was first associated with colorectal cancer by Tomlinson *et al.* [352] and a further study in the Finnish population confirmed AI as a mechanism for the contribution of this SNP to the risk of cancer development [353]. They showed that, compared to first-degree relatives, patients with colorectal cancer favored the G allele at this location [353]. Additional studies identified genes with ASE that conferred an increased risk of other cancer types. These included ASE of BRCA1, BRCA2, FGFR2, DMBT1, STXBP4 and COX11, PALB2 in breast cancer [354–360], APC and TGFBR2 in colorectal cancer [361, 362], PARP1 in melanoma [363], BRCA1 in ovarian cancer [360], DAPK1 in chronic lymphocytic leukemia [364] and SCARB1 in renal cancer [365].

   Familial studies of *Li–Fraumeni* syndrome demonstrated the role of ASE in the modified penetrance of germline TP53 mutations through an analysis of unaffected carriers of the mutation with affected offspring [366, 367]. Variable penetrance of missense variants in LRRC34 contributing to papillary thyroid carcinoma has also been attributed to ASE caused by an upstream regulatory variant that reduces the expression of the mutant allele [368].

   As previously discussed, ASE is common in normal tissue, but the magnitude of differences between alleles is often small for non-imprinted genes. This differs in cancer samples with tumours showing significantly different patterns of ASE of germline SNPs compared to the matched normal sample [365, 369]. A study of kindreds with familial pancreatic cancer showed that changes in ASE can include loss of ASE, gain of ASE and extreme, almost mono-allelic, gain of ASE and that extreme ASE is common in the germline of patients relative to normal control samples [369]. They also showed that those patients that developed pancreatic cancer were at the higher end of the ASE spectrum, indicating that increased ASE is a risk factor for developing cancer [369]. Similar investigations comparing tumour and matched normal samples in prostate cancer showed significant changes in the regulatory effect of germline variants between normal and tumour [370].

A recent study by Luft *et al.* explored the presence of germline mutations in the human genome that were selected for/against during cancer development, with loss of heterozygosity, an extreme form of AI, observed predominantly in genes involved in the repair of double-stranded breaks and homologous repair [325]. These mutations increased the likelihood of a second hit that removes the wild-type allele and retains the mutation damaging to the tumour suppressor genes [325].

2. Somatic ASE

Although germline variants can contribute to AI in cancer by altering the expression of one allele, it is not the main mechanism for allelic imbalance in cancer samples. Changes in DNA allelic ratios caused by somatic copy number events is the main contributor, with different reports estimating it accounts for between 35%-85% of observed AI in genes [322, 326] Somatic mutations leading to nonsense-mediated decay are common in cancer genomes, leading to downregulation of the allele containing the protein truncating mutation [371]. This is a common cause of disruption of tumour suppressor gene function. Rhee *et al.* reported a high degree of allelic imbalance in known cancer genes, consistent with frequent dysregulation of cancer genes [328]. Although loss of function (LOF) mutations in tumour suppressor genes tend to be recessive, allelic imbalance can contribute to cancer development when the functional gene copy is downregulated [328]. Expression of wild type alleles can be affected by LOH or changes in CNV [328]. Clayton *et al.*, discovered that LOF mutations are frequent in the normal genome. However, changes in the expression of the allele containing the mutations distinguishes cancer from normal cells [326]. This implies that healthy individuals are at an increased risk of developing cancer and functionally important changes in ASE are associated with cancer onset and progression [326]. In cancer samples LOF mutations are not limited to cancer-associated genes and are common in other genes suggesting that there is a general loss of regulatory control and almost half of genes exhibiting ASE in cancer showed evidence of exon-skipping indicating that alternative splicing plays an important role in changing ASE patterns in cancer [326].

Gain of function mutations can also be associated with AI in cancer genomes [324]. These mutations are typically present in oncogenes that drive proliferation and progression of cancer. In contrast to tumour suppressor genes, heterozygous mutations in oncogenes are generally sufficient for tumourigenesis [372]. However, they are usually present in the genome along with copy number variation leading to changes in dosages for mutant versus wild type alleles [324]. Bielski *et al.* discovered that allelic imbalance in oncogenes is common in untreated cancers [324]. Previously, it had been suggested that mutant AI was a consequence of cancer therapy but these results indicate that AI likely provides a fitness advantage to the evolving clone [324]. They also discovered that copy number changes leading to AI of mutant alleles arose independently of the mutant SNV [324].

Examples of genes exhibiting AI in cancer include KRAS and telomerase reverse transcriptase (TERT). KRAS is one of the most frequently mutated gene in human cancers and activating mutations of the gene have been found to be the driving force behind a number of cancers including colorectal, pancreatic and lung adenocarcinomas [373]. KRAS is a protein within the RAS/MAPK pathway and is responsible for growth and proliferation of the cell, making it an attractive target for cancer [374]. AI of KRAS has been shown to be extremely frequent in cancers with loss of heterozygosity and copy number variation being the two main mechanisms [373]. Loss of the wild type allele is common in lung adenocarcinomas and copy number gain of the mutant allele is common in pancreatic cancer [373]. Normal cells exhibit cellular senescence meaning that they have a limited number of cell divisions before they die [375]. This is due to telomere shortening with each cell division [375]. TERT is responsible for telomere lengthening, allowing continued cellular replication and its expression is regularly up-regulated in cancer [376]. TERT expression is affected by epigenetic regulation with mutant TERT showing the histone activating modification H3K4me2/3 on its promoter along with RNA polymerase II binding [377]. In contrast, the wild-type promoter has the histone silencing modification H3K27me3 [378]. TERT has also been shown to be regulated through alternate splicing with highest expressed isoform in cancer shown to be the full-length isoform, resulting in active telomerase [377].

1.7 Thesis Overview and Research Questions

It is clear from the studies of somatic mutations in normal tissues that there is much that remains to be discovered about the transformation of normal cells into cancerous cells and the role of somatic mutations in this process. Therefore, this thesis focused on identifying mutations that occurred prior to cancer transformation in cancer samples, understanding the role of the immune system in shaping this mutational landscape and investigating the role of germline ASE in cancer predisposition.

In Chapter 2 we identified clonal mutations in WXS data from TCGA by adjusting variant frequency for tumor purity and local copy number variation because these factors can affect our ability to identify mutations present in all cells of the tumor. In this chapter we make the assumption that the majority of clonal variants are likely to have been present in the cell prior to cancer initiation and this would give us a good indication of the somatic mutations present in the normal cell. We then investigated which cut-offs for calling clonal and subclonal variant calls give the best sensitivity and specificity scores by comparing results to calls generated for TCGA samples in the PCAWG cohort. Next, we used the relationship between age and somatic mutation accumulation to estimate the true clonal load for each sample and compared predicted values to observed somatic mutations from the same tissue type and age. Finally, we investigated the relationship between median predicted clonal load and cancer risk and also perform a genome-wide association study analysis with the predicted clonal load as the phenotype to identify any germline variants that may be associated with the accumulation of somatic mutations.

In Chapter 3, we aimed to understand the role of the immune system in shaping which somatic mutations are clonal in these samples, with the assumption that the majority of clonal mutations occurred prior to immune escape. Two previous studies [205, 206] showed that driver mutations in cancer are common because of an inability of common HLA alleles to present them to the immune system. However, the same was not true for passenger mutations. Therefore, we focused our analysis on passenger clonal mutations and investigated the relationship between gene expression and immune escape. We hypothesized that the lack of a relationship between clonal passenger mutations and HLA genotype may reflect alternative mechanisms through which these mutations may be hidden from the immune system. Specifically, clonal passenger mutations may occur preferentially on lowly expressed genes or the genes on which they occur may be selectively down-regulated. The latter could result in allele-specific expression if the effects that cause the mutant allele to be down-regulated to avoid detection by the immune system act in cis. We also used simulations to identify an upper bound for the detection and removal of missense mutations by the immune system.

In Chapter 4, we extended our analysis of the relationship between gene expression and somatic mutation accumulation by investigating the relationship between germline ASE and cancer risk. We hypothesized that ASE in tumor suppressor genes could be associated with cancer risk because if one copy of the gene is down-regulated compared to the other then the probability of getting cancer would be greater as the cell would only need to acquire loss of function somatic mutations on the expressed copy for cancer to arise. First we aimed to predict germline ASE using genotype data from heterozygous SNPs and tested this approach using data from the GTEx cohort. We used two methods to predict ASE; 1) a modified version of a gene expression prediction tool called PrediXcan [379] 2) logistic regression models using gene-level ASE generated by Castel *et al.* [339] as the response variable and heterozygous status of SNPs as the predictor variables. We then performed a pilot study to predict ASE in UK Biobank data and tested its association with breast cancer risk.

Our research questions can be summarised as follows:

1. Is it possible to predict the true clonal somatic mutation load in a cancer sample by utilizing the known relationship between somatic mutation accumulation and age? (Chapter 2)

2. Is predicted clonal load associated with cancer risk? (Chapter 2)

3. Is it possible to identify germline variants that are associated with increased clonal load? (Chapter 2)

4. Does gene expression level explain the lack of a signal of immunoediting among clonal passenger mutations? (Chapter 3)

5. Is it possible to predict gene level germline ASE using genotype data? (Chapter 4)

6. Is there a relationship between predicted germline ASE of tumour suppressor genes and cancer risk? (Chapter 4)

2 Chapter 2: Investigating somatic mutation load in normal tissues using clonal cancer mutations

2.1 Abstract

Somatic mutations are difficult to measure in normal tissues due to the low frequency of the mutations and our inability to distinguish these variations from noise introduced by the methodological processes. We use clonal mutations in cancer samples in place of normal tissues to understand what has occurred in the cell before cancer transformation. We adjust the variant frequency of somatic mutations identified in samples from the Cancer Genome Atlas (TCGA) to account for tumour purity and ploidy and then determined the clonal status of those mutations. We use the relationship of age with somatic mutation accumulation to estimate the true clonal mutation load for these samples. Using our model, we can predict the total clonal burden of a sample for a particular cancer at a particular age. We find a positive correlation between the clonal mutations estimated using our method and the somatic mutation load determined in normal tissue by other studies. We also find a correlation between the clonal mutation load and lifetime cancer risk of developing cancer. Our findings suggest that this method can be used to estimate somatic mutation load in normal tissues from cancer samples and has the advantage of being able to use the multitude of already published cancer data sets to accentuate our understanding of the somatic mutations that accumulate throughout a person's lifetime.

2.2 Introduction

Somatic mutations arise in cells throughout a person's lifetime. They contribute to ageing [8, 9], neuro-degenerative diseases [10, 11], cancer and other age-related disorders [12–14]. A somatic mutation arises in a stem cell. It is, therefore, only present in the initial stem cell and any cell derived from it, meaning it is generally only present in a small subset of cells within the tissue. This makes somatic mutations in normal tissue samples challenging to detect using standard bulk sequencing approaches.

In recent years, due to advances in technology such as single-cell clonal expansions [16], laser capture microdissection [380] and single-cell RNA Sequencing [381], there has been a growing number of publications studying the number and types of somatic mutations present in normal tissue cells. These have primarily focused on individual tissue types, such as skin [20, 382–384], oesophagus [385, 386], colon [387], endometrial [388], liver [389, 390], brain [391], prostate [392], urethral tissue [393, 394], or on a variety of tissue types within individuals but from a small number of donors [16, 388, 395–398]. The most surprising finding from these studies was the identification of cancer driver mutations under positive selection in normal tissues that showed no evidence of cancer growth. This finding led scientists to question the previous ideas about cancer transformation. These driver mutations were considered sufficient to initiate cancer transformation; however, although these mutations drive clonal expansion, additional mutations are required for tumorigenesis. Additionally, it could be the timing and order of these mutations, the combination of mutations or tissue specificity rather than the presence of a driver mutation that is important [51].

There have been many attempts to estimate the normal somatic mutation rate in humans. The first estimates used inactivating mutations in "sentinel genes" such as PIG-A[399] and HPRT [400] to calculate the somatic mutation rate. However, more recent technological advances in high throughput sequencing, such as duplex sequencing[24], BotSeqS [25], NanoSeq[26], SMM-Seq [29] and EcoSeq [28] , and methods that incorporate culturing single cells followed by sequencing [401] have allowed direct estimation using cells from normal tissues. Despite the advances in our ability to analyse somatic mutations in normal tissues, current technologies still have a lot of problems, such as high error rates [381], the introduction of laboratory-induced mutations [402] or an inability to determine monoclonal structures in the tissue types [403]. These issues mean a tissue's true somatic mutation load can be underestimated.

Here, we used clonal mutations derived from cancer samples to study somatic mutations in normal tissue. Due to the nature of cancer evolution, i.e. the clonal expansion of an initial cancerous cell, every cell in the cancer sample contains a record of all of the mutations that accumulated before the last common ancestor of the cancer cells (clonal mutations), as well as those that occurred subsequently (subclonal mutations). Therefore, subclonal mutations will only be present in a fraction of the cells. An advantage to this method is that we can use the already thousands of available cancer sets to study somatic mutations and their role in cancer tumorigenesis without the need to identify mutations in normal tissues. In this study, we identified clonal mutations in single sample bulk sequencing data from The Cancer Genome Atlas (TCGA) and estimated the true clonal mutation load using the known correlation of somatic mutation accumulation with age [404]. To our knowledge, this is the first attempt to estimate the true somatic clonal mutation load in cancer samples in this way. We also explored factors that affect clonal mutation load and investigated the relationship between somatic mutation load and cancer risk.

2.3 Results

2.3.1 Clonal classification of variants

We followed the method outlined by Dentro *et al.* [74] to infer the clonal status of 1,096,100 single nucleotide variants (SNVs) in 6,807 TCGA samples. This approach uses variant frequency (VF) to determine clonality, with a threshold that is adjusted for tumour purity and copy number variation (CNV) status (explained in detail in Methods). The adjusted variant frequency allows the cancer cell fraction (CCF), i.e. the proportion of cancer cells that harbour the mutation, to be calculated, and we use this to determine clonality. As explained in Methods, several thresholds can be applied to classify variants as clonal or subclonal, with ambiguous calls falling into an "undetermined" class. We applied binomial tests to calculate the CCF and used the confidence intervals from these tests to classify variants. We set upper and lower confidence intervals for clonal variants and an upper limit for subclonal. Determining which value to use for these limits is explained in the following section.

2.3.2 Adjusting thresholds to improve clonal variant calling

A recent study by the Pan-Cancer Analysis of Whole Genomes Consortium [405] investigated the clonal and subclonal architecture of 2,658 cancers using whole genome sequencing. They performed a comprehensive analysis using a consensus approach that incorporated the output of 11 different subclonal reconstruction callers to assign clonal status to each mutation. This dataset includes 527 TCGA samples from our analysis, which we used to compare our results to assess the accuracy of the calls. Limiting our comparisons to positions analysed by both, we had 64,393 variant calls to compare. An advantage of using this dataset to gain confidence in our calls is that the samples for WGS are taken from a different region of the tumour and, therefore, can act as a second sample for the tumour.

To investigate the impact of adjusting the confidence interval thresholds for classifying variants as clonal, subclonal and undetermined (Figure 2.1), we compared our calls to the clonal and subclonal calls from PCAWG. This ensured that we had high confidence that the clonal calls were truly clonal and subclonal calls were truly subclonal. Adjusting the upper confidence interval for classifying a subclonal variant increases the number of PCAWG clonal variants classified as subclonal in our dataset (Figure 2.1A.) As the CCF increases, it is more challenging to distinguish clonal from subclonal calls. It can be that specific variants arose early on in cancer progression and rose to prominence, appearing clonal in the region where the PCAWG sample was taken, but it is, in fact, subclonal in the tumour as a whole. If a variant appears subclonal in one region of the tumour, it is deemed subclonal in the whole tumour. For this reason, we can keep the upper CI threshold for subclonal as high as 0.7 and still be confident that we have not misclassified subclonal mutations as clonal.

Next, we investigated the impact of changing the thresholds to call a variant clonal (Figure 2.1B and Figure 2.1C). Changing the lower CCF confidence interval (Figure 1B), has a greater effect on our clonal calls than changing the upper confidence interval (Figure 2.1C). Changing the lower confidence interval for classifying a variant as clonal impacts the number of subclonal mutations misclassified as clonal. Therefore increasing the lower limit decreases the proportion of subclonal variants misclassified as clonal.

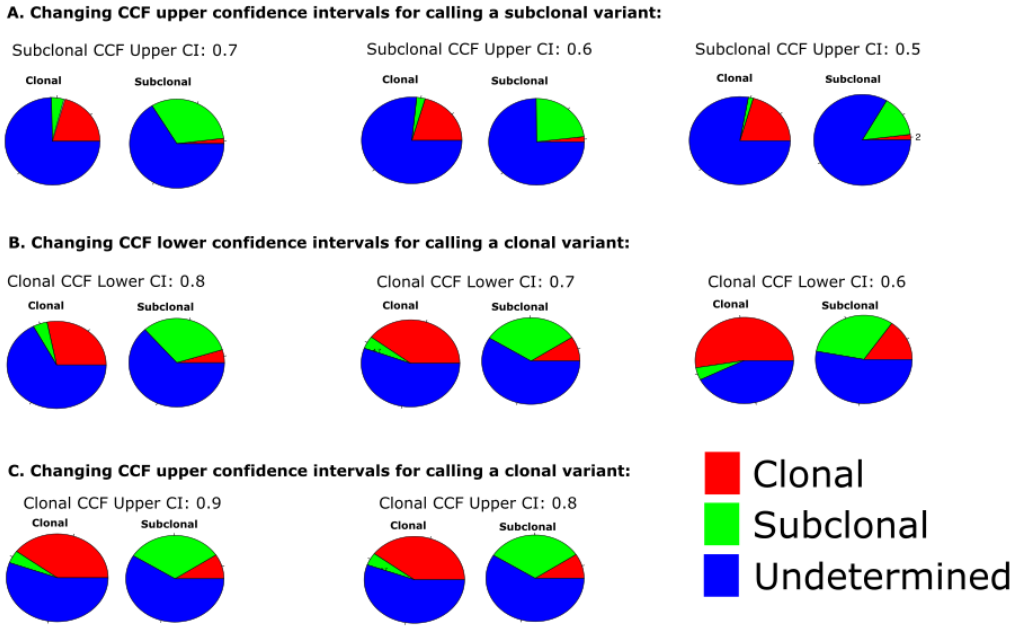


Figure 2.1: **Comparison of clonal and subclonal calls when cancer cell fraction (CCF) thresholds are changed.** Comparison of calls from the TCGA dataset to clonal and subclonal calls from PCAWG when changing CCF confidence interval thresholds. The circles represent clonal and subclonal calls from PCAWG, and the colours represent calls from TCGA. 50X minimum depth was required for variants used in this analysis. (**A**) Clonal CCF confidence intervals were kept at 1 for the upper limit and 0.8 for the lower limit. Subclonal upper CCF confidence intervals varied. (**B**) Subclonal CCF confidence interval was kept at 0.7, and upper CCF for calling clonal variant was kept at 1. Clonal lower CCF confidence interval varied. (**C**) Subclonal upper CCF confidence interval was kept at 0.7, and the lower CCF for calling clonal variant was kept at 0.8. Clonal upper CCF confidence interval varied.

2.3.3 Impact of read depth on classifying variants

As total read depth increases, our ability to classify variants also increases (Figure 2.2). Therefore, we assessed the impact of changing the depth threshold on our ability to call clonal variants. Below depths of 100X we could not classify over half of the variants, while the total proportion of unclassified variants drops to about 25% when we achieve minimum depths of 400X. The proportion of variants called clonal by PCAWG but are called subclonal in our analysis also increases as depth increases. As there are more reads covering these variants, we can be confident that these variants really are subclonal but appear clonal in the region from which the PCAWG sample was taken. The proportion of variants called subclonal by PCAWG but called clonal in our analysis remains high until we achieve depths of 1000X. However, the number of variants present at 1000X is extremely low so it is difficult to draw strong conclusions. The number of variants dramatically decreases as read depth increases. The mean read depth for a variant in our dataset is 95X. Therefore we use a depth of 100X going forward in our analysis to use as much data as

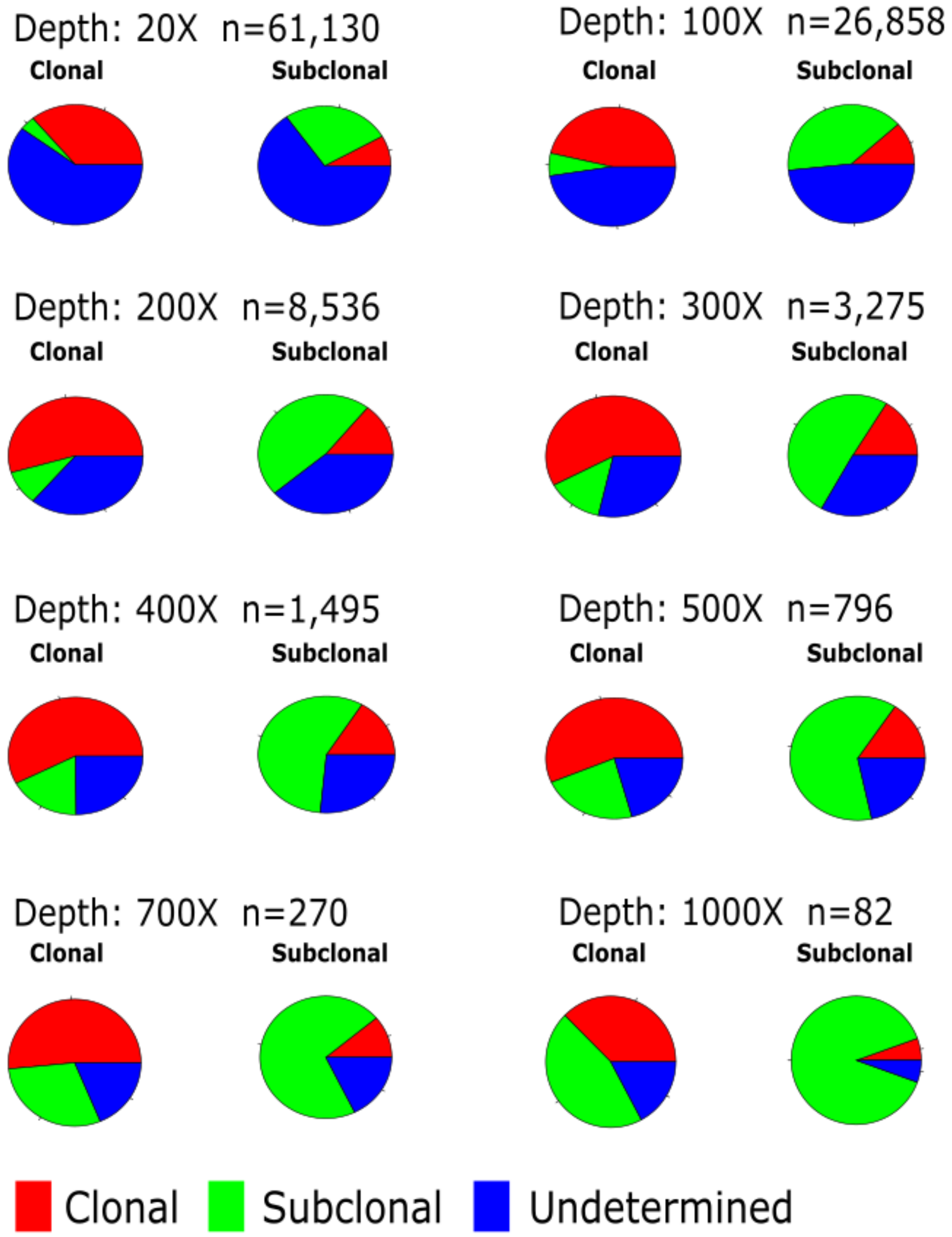


Figure 2.2: **Comparison of clonal and subclonal calls when cancer cell fraction (CCF) thresholds are changed.** Comparison of calls from the TCGA dataset to clonal and subclonal calls from PCAWG over different sequencing depths. The circles represent clonal and subclonal calls from PCAWG, and the colours represent calls from TCGA. Percentages of total PCAWG clonal or subclonal are given for each pie piece. Subclonal CCF upper CI is kept at 0. Clonal CCF confidence intervals were kept at 1 for the upper limit and 0.8 for the lower limit. Each pair of pie charts represent the overlaps for varying minimum read depth values. The number of variants (n) present after filtering for depth is also given.

possible while also confidently classifying at least half the variants. At 100X there is also a small proportion of variants called subclonal in our analysis but clonal by PCAWG (11.7%) and called clonal in our analysis but subclonal by PCAWG (6.2%).

2.3.4 Relationship of mutation load with purity

The tumour purity of a sample has a major impact on our ability to call variants. This is especially true in the case of subclonal variants because it is much more difficult to identify subclonal mutations when the proportion of cancer cells is low in the sample. As expected, the inferred number of subclonal mutations increases as tumour purity increases (Figure 2.3C).

Variant allele frequencies are lower for clonal variants in samples with low tumour purity compared to samples with higher purity. This is because the reference allele will be present at a higher fraction than the mutant allele due to the presence of normal cells in the sample. The purity distribution varies among cancer types (Figure 2.3A), which is reflected in the corresponding variant frequency distributions (Figure 2.3B). Cancers with lower purity, such as LUAD or LUSC, have a clonal peak around 0.25. However, cancer types for which samples tended to have higher tumour purity, such as UVM, peak at 0.5. This highlights the impact of purity on the inferred mutant allele frequency and why it needs to be accounted for when classifying variants as clonal or subclonal instead of using variant frequency alone.

There is a decline in the total number of clonal variants as purity increases (Figure 2.3C). This is because most samples with high tumour purity come from ACC and UVM cancers which have a low overall mutational burden and therefore their clonal burden will be low in comparison to samples from other cancer types.

2.3.5 Comparison of raw mutation calls versus curated call set

For our analysis, we used the publicly available mutation calls from the Multi-Center Mutation Calling in Multiple Cancers (MC3) working group [113]. This is a highly curated set of mutation calls for which MC3 working group applied stringent filtering criteria and performed a comprehensive assessment of variants in order to remove germline variants and artefacts before making the dataset available for public release. We repeated our analysis using the original raw mutation calls from all seven callers, using the same consensus approach that a variant must be called by at least two variant callers in order to assess the impact of noise on determining clonal and subclonal calls. The curated calls contains 22.6% of the variants present in the raw calls (Table 2.1). Interestingly, applying a depth filter of 100X to both sets reduced the raw calls by 3% but reduced the curated calls by 70%. This is likely due to differences in combining results from various mutation callers. When combining variants from the raw calls, we used the depth results from the variant calls of one variant caller, while the MC3 group averaged counts from

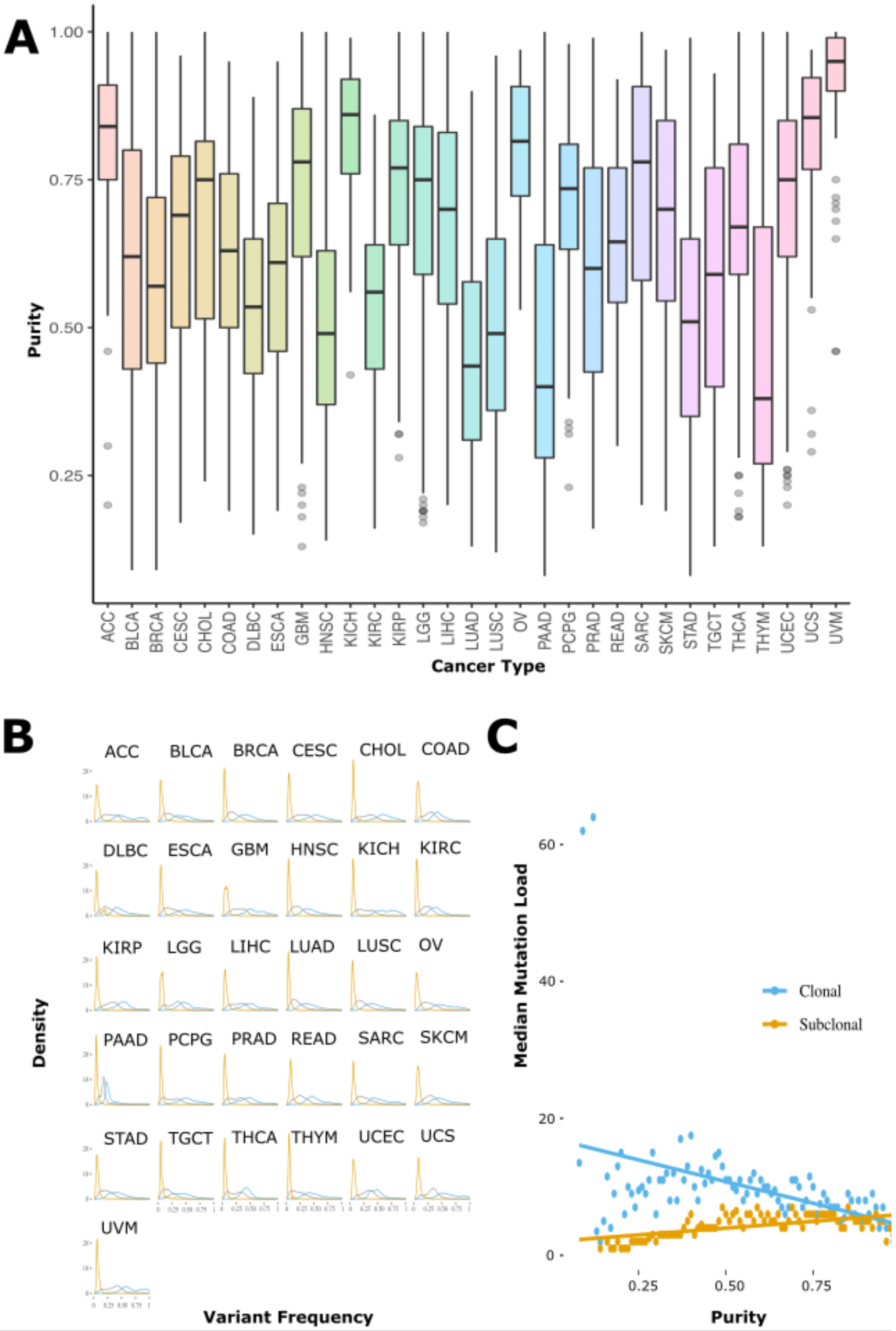


Figure 2.3: **Overview of purity estimates and variant frequency for variants within each class of clonality.** (**A**) Purity distributions per sample split by cancer type. (**B**) Variant frequency spectrum for clonal (blue), subclonal (yellow) and undetermined (grey) variants split by cancer type. (**C**) Relationship between sample purity and median clonal (blue), subclonal (yellow) and undetermined (grey) mutational loads for each purity value.

	Raw variants	Curated variants
Total SNV	8,326,463	1,884,295
Total after > 100X depth filter	8,082,966	550,601
Number of hypermutated samples	129	29
SNVs classified as clonal	2,383,316	192,403
SNVs classified as subclonal	512,914	82,854
SNVs not classified	5,186,736	275,344

Table 2.1: **Comparison of results when using variants present in the raw mutation calls to those present in the curated calls.**

all variant callers. Identifying hypermutated samples as those which have >1000 missense mutations removes 100 more samples from the analysis when using the raw calls than when using the curated calls. This is likely due to artefacts and germline variants contributing to the total missense count, inflating the mutation load for these samples. The proportion of variants classified as clonal, subclonal and undetermined is similar when using both datasets, indicating that there is no difference in our ability to classify variants in both datasets. If the majority of variants in the raw calls are, in fact, germline variants or artefacts, as identified by the MC3 group, then using the raw calls for our analysis would result in a high overestimation of the mutational loads.

2.3.6 Impact of total reads for calling variants

The number of reads sequenced for each individual sample has an impact on the total number of variants called for that sample. Each variant caller requires a certain depth for which a variant can be called. Therefore, if the number of reads present in a sample is low, the total number of variants that reach the minimum depth threshold will also be low. Additionally, we apply a read depth filter to our dataset to increase our confidence in clonal classification. The number of reads sequenced for each sample ranges from 25,964,716 to 850,543,992. There was a positive correlation between total mutational burden and total sequenced reads (Spearman Rho=0.25, p-value= 7.2×10^{-196}). Therefore, it is important to consider this when analysing the total mutational burden.

2.3.7 Relationship of mutational load with age

We are confident that the thresholds we used to classify variants resulted in accurate clonal and subclonal calls. However, there was still a large proportion of variants we could not classify (undetermined). In order to further assess the impact of the confidence interval thresholds on our calls, we exploited the known relationship of age with mutation accumulation. Because somatic mutations accumulate with age in healthy cells, the number of clonal mutations in a cancer sample should also increase with patient age since the clonal mutation burden includes all of the mutations that arose prior to cancer transformation. This is not the case for subclonal mutations, as the subclonal mutation burden depends on the rate at which mutations accumulate in the cancer cells and the time since the development of the cancer (rather than on patient age). Mutation load increases with age in most cancer types except for lung and endometrial cancers, which show a negative trend [406–408] . We also saw this in our data when we assessed the relationship between age and mutation load for each cancer type (Table 2.2). For this reason, we removed the UCEC, LUAD and LUSC cancer types from our analysis. Although LUSC showed a negative correlation which did not reach statistical significance, we still removed it from the analysis due to the negative correlation reported previously.

The positive correlation between the remaining cancer types and age (Spearman's rho=0.32, p-value= 9.4×10^{-114}) provides a means to estimate the total number of clonal mutations per sample. We assume the relationship with age exists only for clonal variants. In that case, we can say that the total clonal load, T, of a sample can be modelled as follows, with age, A, as the predictor variable:

$$T = \beta_{0\mathrm{T}} + \beta_{1\mathrm{T}} A + \varepsilon$$

The number of clonal variants we have classified as clonal is a proportion of the total true clonal load with the remaining clonal variants present in the undetermined group. The slope of the model ($\beta_{1\mathrm{T}}$) tells us how much the total clonal mutational load changes with each year increase in age. The slope of the model using the clonal variants we have accurately classified as clonal will be a proportion, p, of the total slope. The remaining slope will be present in the model that relates the unclassified variants to age in the model. As a result, we can estimate the proportion of all clonal variants that were called clonal using the slopes from two models 1. relating the clonal mutation load to age (see methods equation 7), 2. relating the undetermined mutation load to age (see methods equation 8),.

The intercept of the regression model relating mutation load to age ($\beta_{0\mathrm{T}}$) can be used to estimate the specificity of calling clonal mutations. This is because the clonal mutation load, but not the subclonal load, correlates with age. If all clonal calls are truly clonal and no subclonal calls have been misclassified as clonal, we would expect this intercept to be 0 (i.e. the extrapolated clonal mutation burden at age zero is zero). However, if the clonal calls contain subclonal mutations, the intercept will no longer pass through or close to 0. Because the variants classified as clonal are only a proportion of the true clonal load, we first estimated the true clonal mutation load and fit a model using these values to obtain the intercept value of the estimated total clonal load value at age zero.

In order to build a linear regression model predicting clonal mutation load using age as the predictor variable, we first needed to assess whether our data met the assumptions of linearity, normality and homoscedasticity for linear models (Figure 2.5). Using the clonal mutational load values as our response variable, we did not meet the normality assumptions for the residuals (first column of Figure 2.5B) which means our estimates of clonal load would not be reliable using this

Cancer Type	**Rho**	**P-value**
ACC	↑ 0.27	0.01
BLCA	↑ 0.12	0.02
BRCA	↑ 0.04	0.30
CESC	↑ 0.18	0.04
CHOL	↑ 0.31	0.08
COAD	↑ 0.08	0.24
DLBC	↑ 0.24	0.41
ESCA	↑ 0.18	0.02
GBM	↑ 0.28	0.00
HNSC	↑ 0.16	0.00
KICH	↑ 0.47	0.00
KIRC	↑ 0.4	0.00
KIRP	↑ 0.41	0.00
LGG	↑ 0.39	0.00
LIHC	↑ 0.11	0.17
∗ LUAD	↓ -0.11	0.02
∗ LUSC	↓ -0.02	0.64
OV	↑ 0.19	0.19
PAAD	↑ 0.16	0.04
PCPG	↑ 0.15	0.12
PRAD	↑ 0.13	0.01
READ	↑ 0.19	0.10
SARC	↑ 0.32	0.00
SKCM	↑ 0.22	0.03
STAD	↑ 0.25	0.00
TGCT	↑ 0.08	0.36
THCA	↑ 0.29	0.00
THYM	↑ 0.16	0.27
∗ UCEC	↓ -0.17	0.00
UCS	↑ 0.29	0.04
UVM	↓ -0.05	0.69

Table 2.2: **Spearman correlation coefficient for age and total mutation load per cancer type.** Arrows represent the direction of the correlation, with the up arrows representing a positive relationship and the down arrows representing a negative relationship. P-values that reach statistical significance (<0.05) are coloured in green. Stars indicate the cancer types that were removed due to known negative correlations.

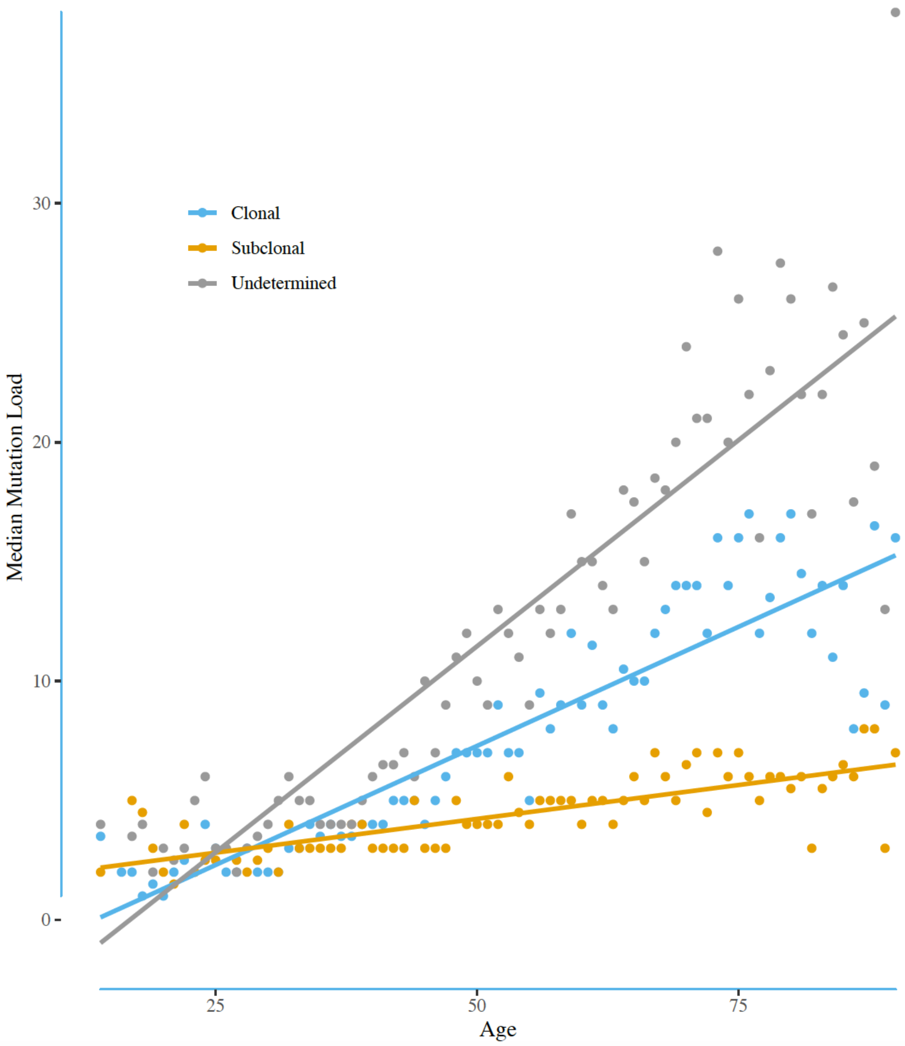


Figure 2.4: **Relationship between age and median clonal, subclonal and undetermined mutational loads for each age value.** Clonal= blue, Subclonal= yellow, Undetermined= grey. Medians were calculated over all samples with a given integer-valued age.

data. We also have a problem with heteroscedasticity in this data set (Figure 2.5C), (Breusch-Pagan test p-value=5.63×10^{-15}). Applying log transformation to the response variable (second column of Figure 2.5) fixed the normality violation. However, the response and predictor variables no longer have a linear relationship. Scaling the clonal mutational load by total sequencing read depth (third column Figure 2.5) gave similar results to using the original data. Instead of using all data to fit our model, we used the median clonal load at each age, taking cancer type into account (column four of Figure 2.5). Although this improved the heteroscedasticity of the model (Breusch-Pagan test p-value=0.51), it did not improve the normality. We also tried a weighted least squares approach (column 5 of Figure 2.5) which again improved heteroscedasticity but failed to improve normality. Therefore, we instead used a generalised linear model method which does not assume a normal distribution for residuals.

The thresholds used to identify clonal mutations can be further optimised using the proportion of true clonal mutations that have been called clonal as well as the intercept of the model using the estimated total clonal load. The slopes from both models were calculated each time the thresholds were altered and used to assess the impact of changing the thresholds (Figure 2.6). Again, changing the lower confidence interval for calling a clonal variant had the biggest impact. We achieved the lowest intercept value using a lower CCF confidence interval of 0.8. The upper confidence interval had no impact at this level, so we kept it at 0.8 in order to capture as many variants as possible. Although the proportion of variants is still small (0.45), we are confident that the variants classified as clonal really are clonal, and so our predictions of total clonal load will be more reliable. Our final thresholds for calling a clonal variant are 0.8 for both the lower and upper CCF confidence intervals, and the final lower CCF confidence interval threshold for calling a subclonal variant is 0.7.

2.3.8 Effect of Cancer Type on clonal load

The model with age as the only predictor assumes that the relationship of clonal mutations with age is consistent across all cancer types. However, we know this is not the case and that mutation rates differ between tissue types. Using the estimated total clonal load, we can perform a linear regression which accounts for cancer type and age to understand the impact of cancer type on the clonal mutational load. When we include the cancer type in our model, the goodness of fit of this model (McFadden R squared=0.29) is better than a model with age alone (McFadden R squared=0.07).

2.3.9 Relationship of sample features with clonal mutational load

There are a number of biological features that can have an impact on the total clonal mutational burden of a sample. These include race, gender, stage, and grade. Gender [409–411] and race [412] have been shown to have an impact on the mutational burden for certain cancer types and therefore, could impact our ability to predict clonal load accurately. Additionally, the stage and grade of a tumour could impact our ability to call a clonal variant accurately because if the tumour is a late stage or high grade, there has been a long time since cancer initiation, meaning it could be a very heterogeneous sample. The heterogeneity of the sample could mean that a subclonal mutation that occurred after the initial cancer expansion

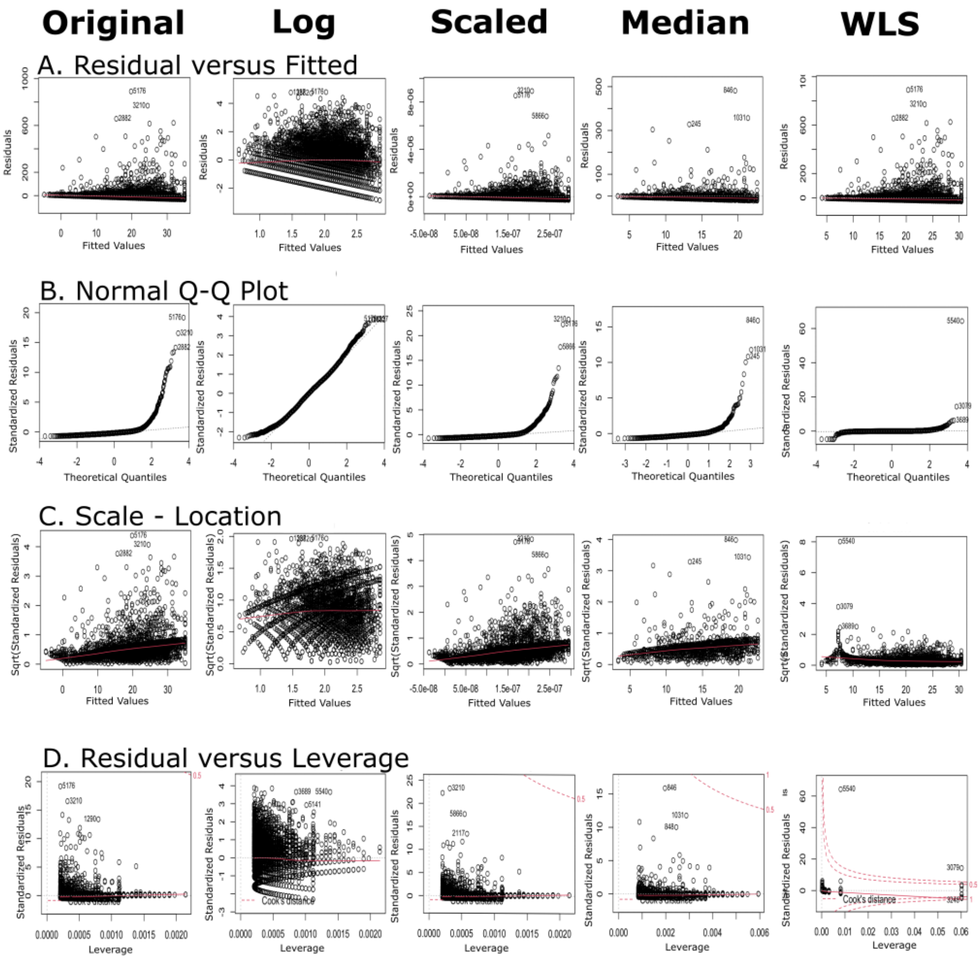


Figure 2.5: **Diagnostic plots for regression models.** (Diagnostic plots for) models using the original data as the response variable, (2) log-transformed response variable (3) response value scaled using total read depth, 3) response variable as the median clonal load at a particular age, depending on cancer type (4) original data as the response variable using a weighted least squares (WLS) approach. (**A**) Residual versus fitted values plot to assess linear pattern between response and predictor (**B**) Q-Q plot to assess if residuals follow a normal distribution (**C**) Scale-Location plot to test the assumption of equal variance (homoscedasticity) (**D**) Residual versus leverage plot to identify influential cases.

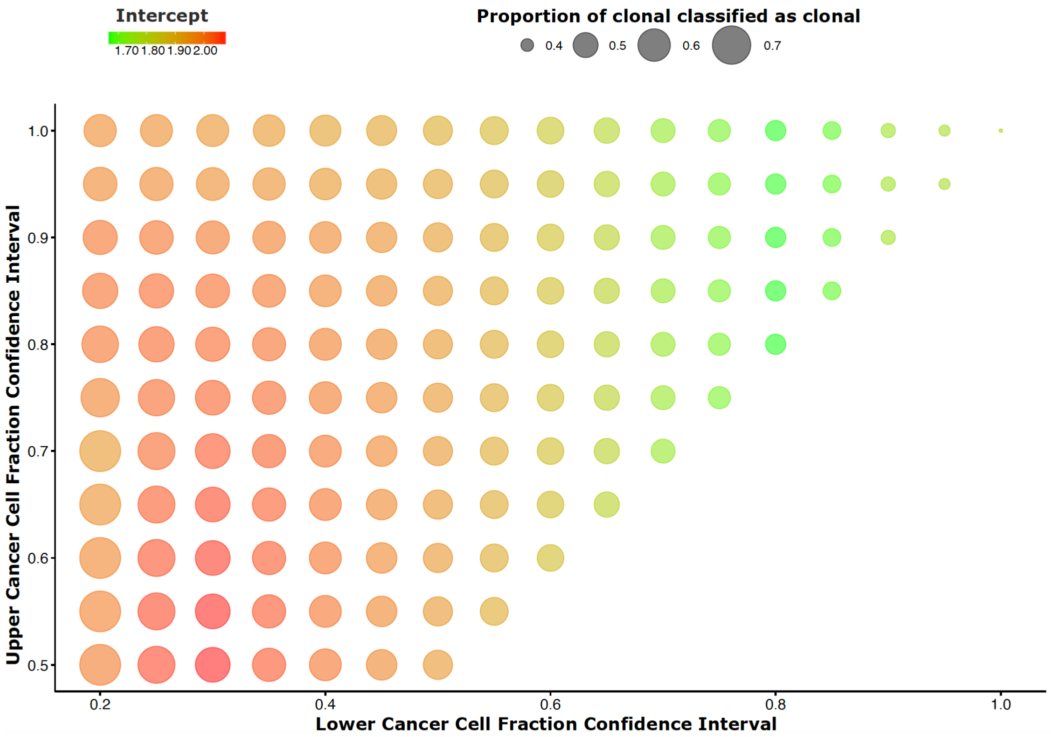


Figure 2.6: **Sensitivity and specificity of clonal mutation calls as a function of confidence interval thresholds based on the linear models relating the clonal mutation load to age and the undetermined mutation load to age.** The size of the circles represents the proportion of true clonal variants called clonal. The colour of the circle depicts the intercept value of a model with the estimated clonal load as the response and age as the predictor. Red indicates higher values, and green indicates low intercept values. The Y-axis shows the different clonal upper confidence interval thresholds applied, and the X-axis shows the different clonal lower confidence interval thresholds.

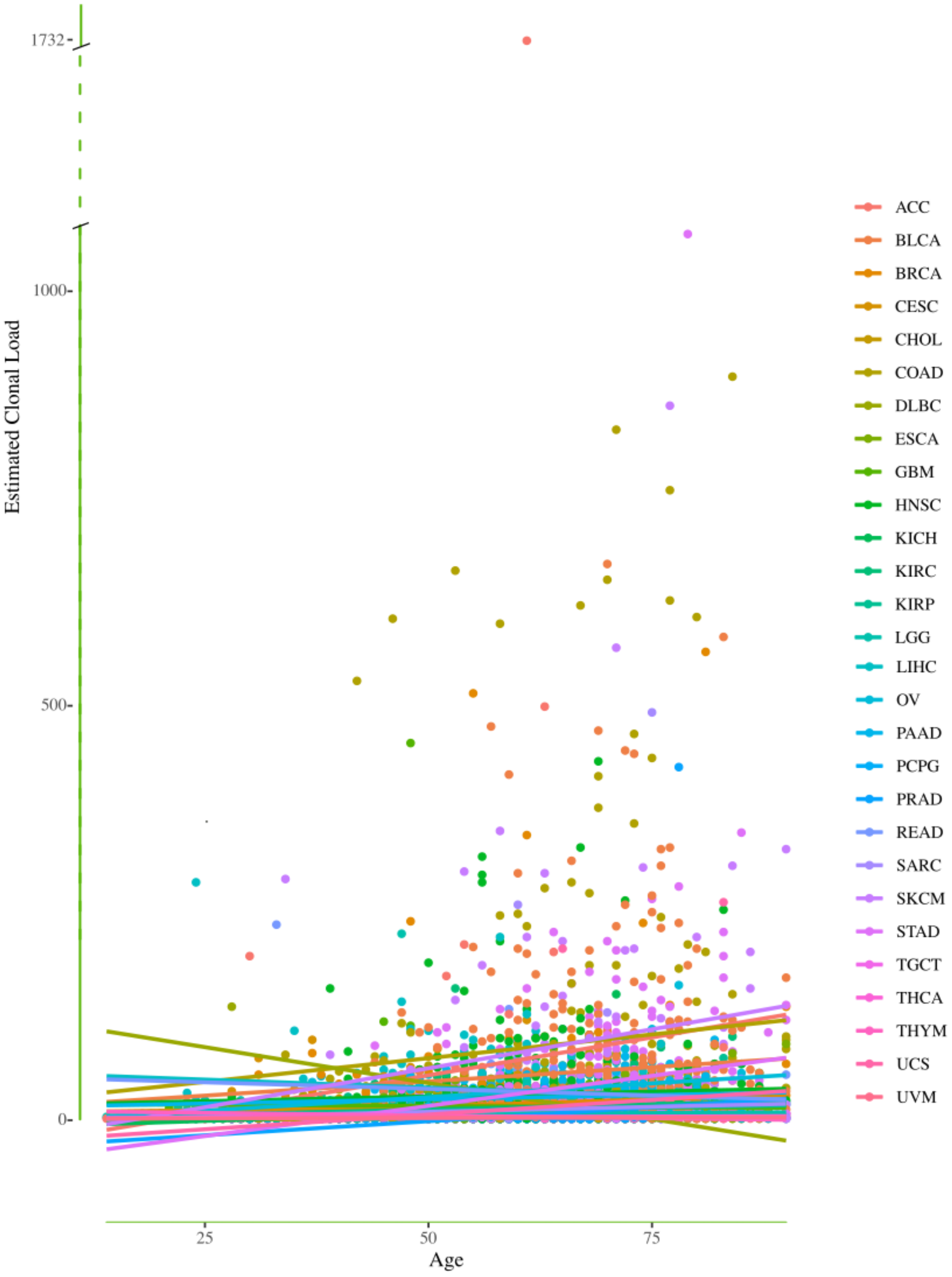


Figure 2.7: **Relationship between median predicted clonal load and age for each cancer type.** Colours correspond to each cancer type. Cancer codes are explained in Table 2.4.

Predictors	**Goodness of Fit (McFadden R^2)**
Age	0.07
Age + Cancer Type	0.29
Age + Gender	0.07
Age + Race	0.07
Age + Grade	0.07
Age +Stage	0.07

Table 2.3: **Assessment of adding predictors to the model.** The two models with * were fit using fewer samples because grade and stage information was not available for all samples.

has taken over now appears clonal. This would result in a higher clonal mutation load for these samples compared to samples from an earlier stage or lower grade tumour.

We assessed the impact of adding each factor separately as predictors in the model that includes age as a predictor (Table 2.3). Adding each factor to the model did not improve the goodness of fit, indicating that cancer type with age is enough to capture the variation in clonal mutational load.

2.3.10 Comparison with somatic mutation load in normal tissues

Next, we investigated whether the clonal counts estimated for each tissue type correlated with the expected somatic mutation load for the normal tissues from which the cancer originated. This gives us an indication as to how well the clonal counts represent the normal somatic mutation load. A recent study by Moore *et al.*[413] generated somatic mutation burdens using whole genome sequencing for 29 cell types. Using our model, we predicted the total clonal mutation burden for each cell type and age present for seven cancer types that correspond to tissue types in the data; see Methods for details of the normal data.

Our predictions were derived from whole exome sequencing (WXS) data, so we scaled the whole genome sequencing (WGS) values from the normal tissue, assuming that the exome corresponds to approximately 1.5% of the whole genome, for the comparisons. There was a high correlation (Pearson r=0.63, p-value=0.03) between the estimated values and those observed in normal cells (Figure 2.9). The regression line (blue) deviates from the x=y line (black), indicating that our estimates do not perfectly match the normal somatic mutation load. However, it does fall within the 95% confidence interval of the regression. This indicates that although the estimated clonal mutation burden in the cancer samples was higher than the mutation burdens observed in the corresponding normal cells, the difference was not significant.

Some predictions fall outside the 95% confidence interval range (grey area), which indicates that we may be over and underestimating the total clonal load for these cancers. Interestingly, we had normal data from two individuals for each of these tissue types (skin, prostate and oesophagus), but only one sample for

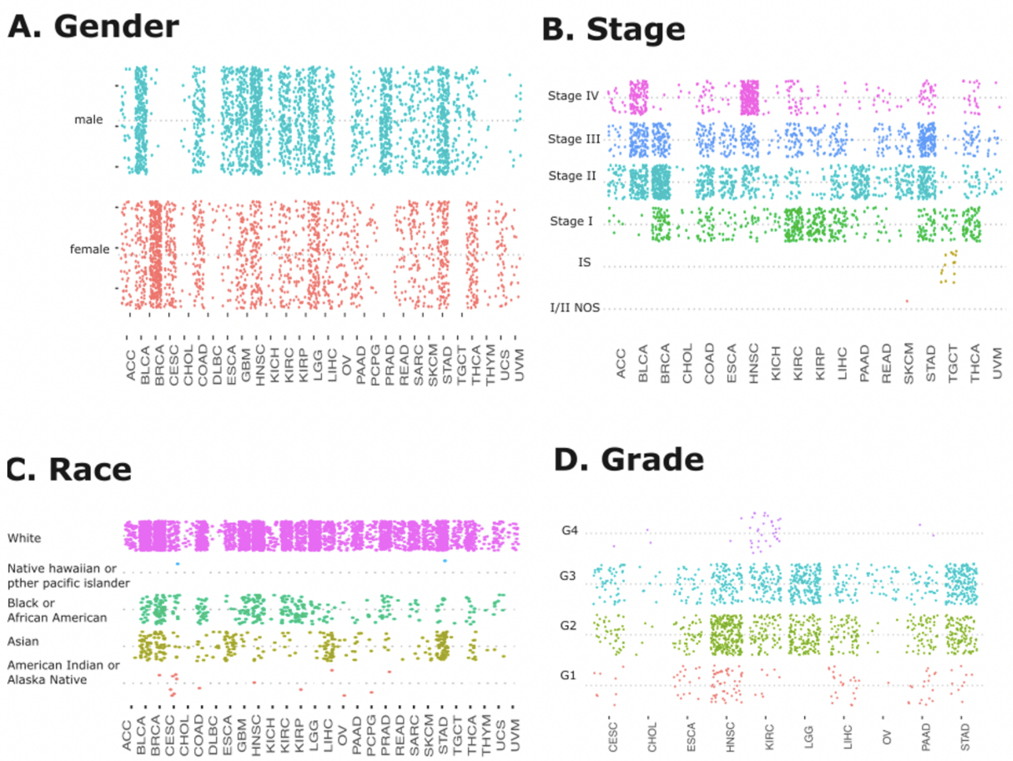


Figure 2.8: **Total number of samples for a variety of sample features split by cancer type.** The number of samples (points) for each (**A**) gender, (**B**) stage, (**C**) race and (**D**) grade for each cancer type for which we have data available. Cancer codes are explained in Table 2.4.

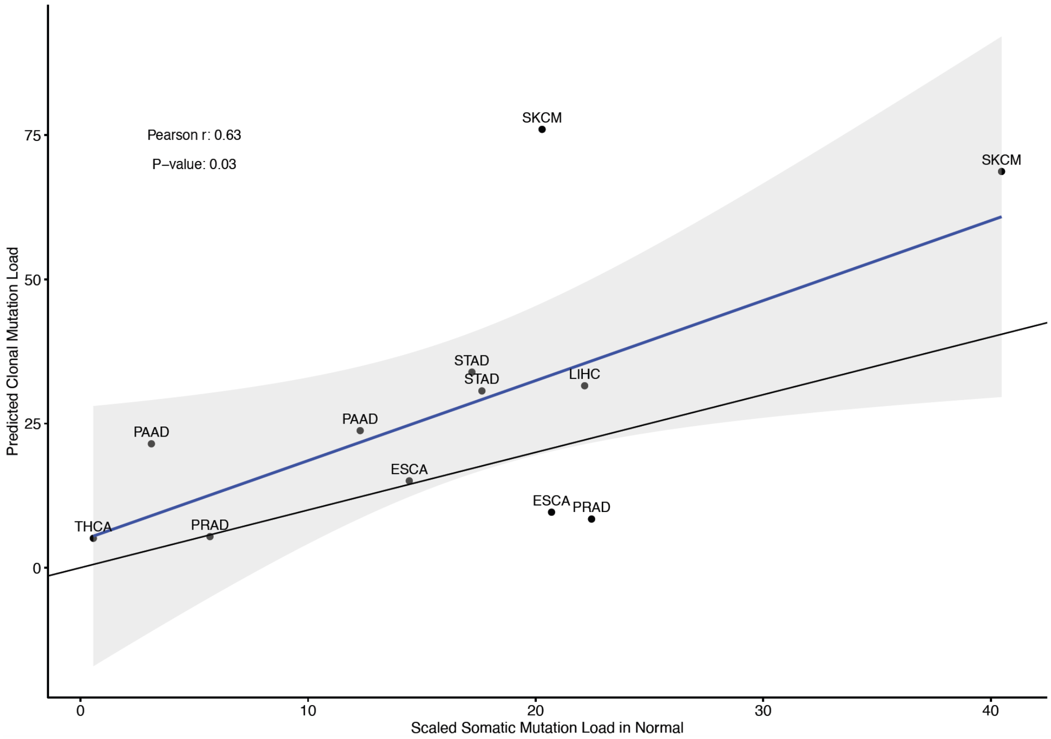


Figure 2.9: **Scatterplot of predicated clonal mutation load versus scaled somatic mutation load from normal tissues.** The fitted linear regression line is shown in blue, with the light grey area indicating the 95% confidence interval on the regression line. The x=y line is shown in black. Samples are labelled by cancer type. Cancer codes are explained in Table 2.4.

each fell outside the range. Further investigation of the mutational load from the normal samples showed that the normal skin tissue for the individual at age 54 was lower than the mutational load from the normal skin sample at age 47. While the prediction for skin at age 47 was close to the observed mutational load for the skin sample at age 47, the prediction for age 54 was almost three times higher than the observed mutational load for the skin sample at age 54. This could be due to skin exposure, or lack thereof, to UV light. It may be that the individual at age 47 had more exposure to UV compared to the individual at age 54, which would result in a higher mutational burden. This is likely to be true, too, for the samples used to build the model used for predictions. It is likely that individuals with skin cancer have had high exposure to UV light, explaining why the prediction at age 47 is closer to the observed mutational load in the normal skin sample at age 47. Investigating the normal oesophageal samples showed that the individual at age 47 had 1.5 times more mutations compared to the individual at age 78 and more than twice the number of mutations as predicted for age 47. It is possible that the individual at age 47 had some unreported exposure to environmental factors that increased mutational load or had a defect in DNA mismatch repair, causing hypermutation. It is interesting to note that the individual at age 78 died of metastatic oesophageal cancer. The prostate sample at age 78 had more than four times the somatic mutational load of the individual at age 47 and 2.75 times more than the predicted value at age 78. These results demonstrate that our predictions are close to the observed values in normal tissue, but we require more data from normal tissues to properly assess the accuracy of the predictions.

2.3.11 Association with lifetime cancer risk

Age is the most significant risk factor for cancer [61] , and this has been attributed to the accumulation of somatic mutations throughout a person's lifetime. We investigated the relationship between the log predicted clonal mutation counts at age 80 per cancer type and log lifetime cancer risk (LCR) (Figure 2.10). There was a positive correlation between the values, but it did not reach statistical significance. There is also a positive correlation between lifetime stem cell division (LSCD) and predicted clonal counts for the same cancer types (Figure 2.10B). Even with the small sample size, there is a strong positive relationship between LCR and LSCD (Figure 2.10C), as previously reported by Tomasetti and Volgestein [73]. It should be possible to better understand the relationships as LSCD data becomes available for more cancer types, as we were limited to the 11 cancer types for which we had LSCD data.

2.3.12 Genome-Wide Association Study

In order to examine germline genetic contribution to the predicted clonal mutational load, we performed a genome-wide association analysis (GWAS) using predicted clonal mutational load as the phenotype. No SNPs reached genome-wide significance ($p < 5 \times 10^{-8}$), and there was only one SNP that passed the threshold for suggestive loci ($p < 1 \times 10^{-6}$) (Figure 2.11A). There was one gene, CKM, that passed the significance threshold in the gene-based analysis (Figure 2.11B). There is little evidence that there is a germline effect on the clonal mutational load of the cancer samples. However, it may be that the effect is confounded by the large number of cancer types in this study. A GWAS analysis for individual cancer types with large sample sizes would help to elucidate this.

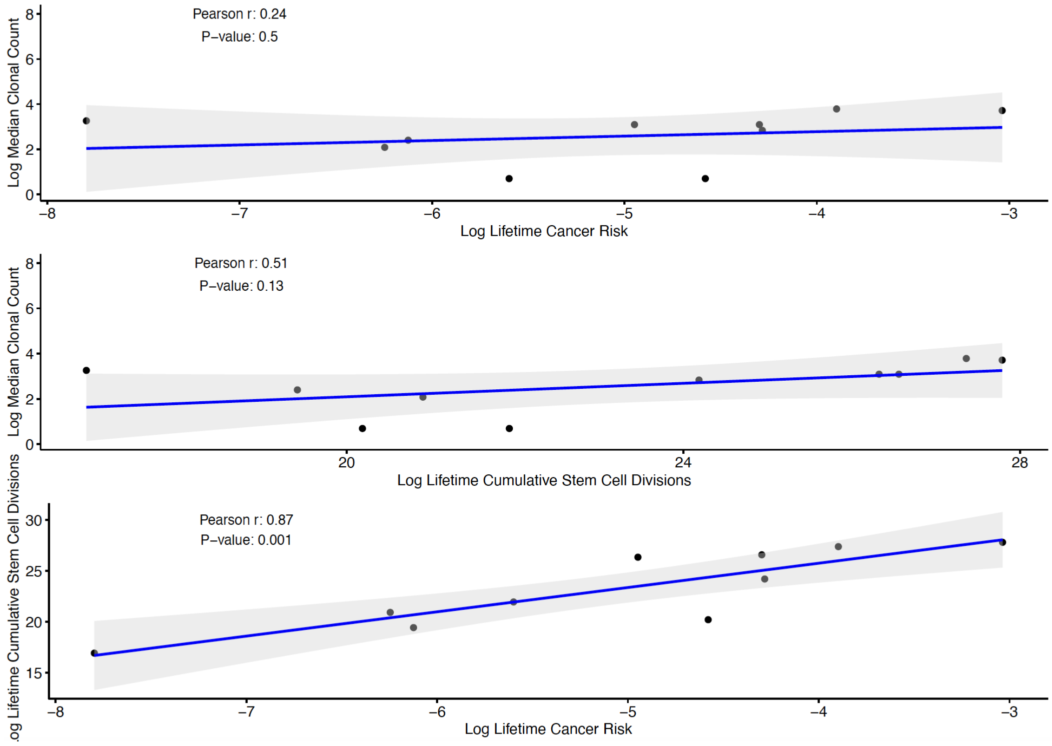


Figure 2.10: **Relationship between predicted clonal load, cancer risk and lifetime stem cell divisions.** Scatterplots of (**A**) log median estimated clonal load versus log lifetime cancer risk, (**B**) log median estimated clonal load versus log lifetime stem cell division (LSCD), (**C**) log lifetime stem cell division versus log lifetime cancer risk. Blue lines are the fitted linear regression lines with the light grey area indicating 95% confidence intervals.

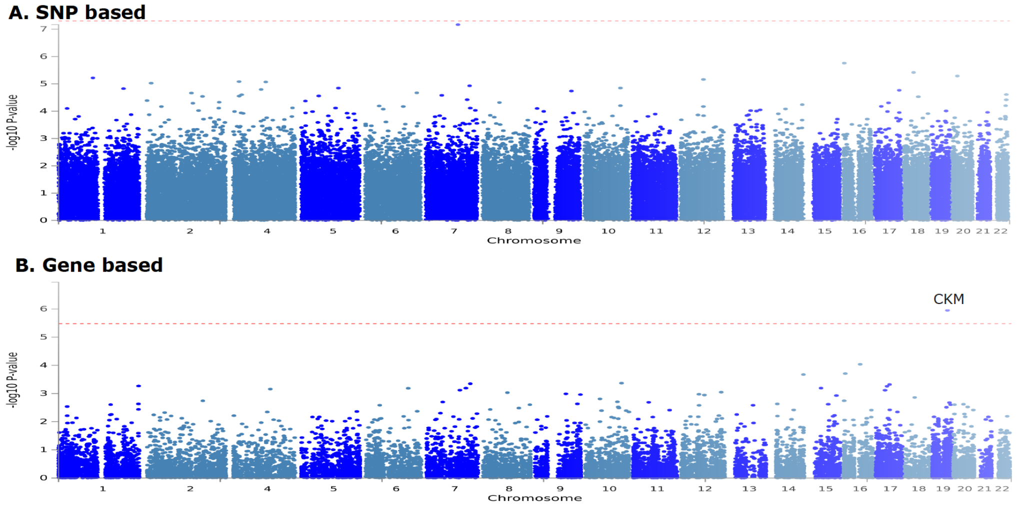


Figure 2.11: **Manhattan plots using predicted clonal load as phenotype in a genome wide association study.** (**A**) SNP-based Manhattan plot with genetic coordinates on the x-axis and negative logarithm of the association p-value for each single nucleotide polymorphism (SNP) displayed on the y-axis. Each dot represents a SNP. (**B**) Gene-based Manhattan plot with genetic coordinates on the x-axis and negative logarithm of the association p-value for each gene displayed on the y-axis. Each dot represents a gene. The dotted red lines indicate the significance level. Any SNP/gene that achieved significance is labelled.

2.4 Discussion

Recent studies determining somatic mutations in normal tissue have highlighted that there is still a lot unknown about somatic mutation accumulation in the normal cell [20, 39, 382–398, 414–418]. These studies showed that driver mutations are present at high frequencies and that clonal expansions are common in normal tissues. Thus, highlighting the need for further studies into somatic mutation accumulation in normal tissues. However, there is still a long way to go to improve technology, reduce costs and eliminate noise from these datasets. Therefore, it is essential to glean as much information about the normal tissue from the myriad of cancer datasets that already exist. Historically, studies investigating clonal mutations in cancer primarily focused on identifying driver mutations, understanding somatic mutation rates in the normal tissue and predicting response to immunotherapy. Passenger mutations were less studied until more recently when it was discovered that they aid in reducing tumour fitness [419, 420], can be used as a molecular clock to calculate the age of the tumour [421], can be used to more accurately classify tumour types for tumour biopsies [422], and some may even accumulate to slow down tumour progression [423, 424]. Cancer samples have not been fully utilised to understand what is happening in the normal cell, which is essential to understand the path to cancer fully. Here, we predicted the total clonal mutation burden in the exome of cancer samples to investigate if it could be used to understand mutations in normal tissue further. We also investigated the relationship between clonal mutation load with factors such as tissue of origin, grade, and stage, as well as the association with LCR, LSCD and somatic mutation load in the normal tissue.

Variant frequency can be used to infer the clonality of a variant, with low variant frequency called subclonal and high variant frequency called clonal. However, using a fixed variant frequency threshold can result in some variants being misclassified as either clonal or subclonal because a cancer sample is never purely made up of cancer cells and is a heterogeneous mix of normal and cancer cell types. Therefore it is necessary to adjust variant frequency to account for this. While many methods, such as Canopy [140] and cloneHD [425], take copy number into account, few methods also incorporate tumour purity. Our analysis used adjusted variant frequencies to account for tumour purity and local copy number to calculate the CCF, which is used to determine clonality. Even with adjusted variant frequencies, our results show it is still challenging to classify variants accurately as clonal or subclonal using set thresholds.

Several factors affect our ability to identify all clonal mutations in a sample and cause us to underestimate the total clonal burden. These include the sequencing depth at each position and the total reads sequenced in the sample. We applied a depth filter of 100X, which means that any clonal variant present at a location that did not meet this depth requirement would be lost from our analysis. This is impacted by the total number of reads sequenced per sample because samples with lower total reads will have a reduced ability to meet the depth requirement. A limitation of our analysis is that we were not able to obtain the callable size of the genome for each sample. This is a measurement of the total number of positions which reached the required depth for calling a variant. A better estimate of the total clonal mutational burden would have included this value in the calculation.

We assumed that at age zero, the somatic mutation load for an individual was zero. However, somatic mutations can occur during early human embryogenesis at an estimated 3 mutations per cell per cell doubling [426]. Most variant calling pipelines in cancer studies use a matched normal from the same individual to remove germline variants which will also remove somatic variants that occurred during embryogenesis. These would not be identified by our method. Therefore, we would expect the number of somatic mutations at birth to be zero or close to zero and were able to use this as another way to determine the specificity of the model. We were unable to achieve an intercept of zero, even with the most stringent CCF thresholds applied, which indicates that there was some level of subclonal contamination in our clonal calls. It is estimated that roughly 26% of clonal calls may actually be subclonal calls when using a single sample bulk sequencing approach [79, 427]. This means that using a single sample would result in an overestimation of the total clonal load.

We also assumed that there is no relationship between subclonal mutations and age. However, the number of accumulated subclonal mutations might also depend on the time since cancer initiation and the amount of time that has passed until cancer diagnosis. This is because for some cancer types a large amount of time may have passed between cancer initiation and detection, during which time subclones are able to accumulate mutation. Although the rate of mutation may be different to the rate prior to cancer initiation, this rate could also be correlated with time. There is a deficiency in analyses investigating the relationship between age and subclonal mutations. Therefore we rely on the studies investigating mutational signatures in subclonal mutations to understand the impact. A recent study investigating intra-tumour heterogeneity of cancer samples has shown a change in signature activity in subclonal mutations compared with clonal mutations [79].

SBS1 clock-like signature is no longer active in subclones, which indicates that the C >T dominant signature, characteristic of the relationship between somatic mutation and age, is not present in subclones and can act as evidence that there is no or low relationship between age and subclonal mutation accumulation.

The somatic mutation rate can differ between human tissues and cell types, which is reflected in differences in the mutational load between cancer types. Including the cancer type in our model increased the goodness of fit for the model predicting the total clonal mutation load. Clonal load differs depending on the tissue of origin for the cancer, with endometrial (UCEC), lung (LUAD and LUSC), skin (SKCM), colorectal (COAD) and bladder (BLCA) cancers all showing high clonal load and thyroid (THCA), PCPG, thymus (THYM) and eye (UVM) cancers exhibiting low clonal load. This is all in line with results previously reported in the literature showing the relationship between the mutational burden of cancer types and response to immunotherapy. Cancer types such as melanoma, lung, urothelial and endometrial all have a better response to immunotherapy, possibly as a result of their high mutational burden [259, 428–432]. The high mutational burden of skin and lung cancers is attributed to mutagenic exposure such as UV light and tobacco smoke, respectively [48] while the APOBEC signature, which causes an increased mutation rate, has been found in 70% of bladder cancers[48, 433]. Increased mutational burden observed in colorectal cancers is caused by microsatellite instability, mutations in the POL genes and defects in the WNT signalling pathway, particularly in the APC gene, which are common in the non-hypermutated samples [434]. Uveal melanoma has a much lower mutation rate compared to other cutaneous melanomas and, as a result, has a low response to immunotherapy [435]. Comparison of metastatic uveal melanomas with paired primary tumours has shown stability at the nucleotide level in this cancer type[436]. Thymomas have the lowest average mutational burdens among human cancers and have low somatic copy number alterations, which is thought to explain why they do not respond well to molecular targeted therapies [437, 438].

A limitation of our analysis is that only a single sample was available per patient, which can cause an illusion of clonality when a mutation is shared by all cancer cells in the sample but not all cells in the wider tumour from which the sample was derived. By using only a single sample from these patients, we may be misclassifying subclonal mutations as clonal mutations and, as a result, overestimating the true clonal count in these samples. This is especially true for late-stage and high-grade tumours, which have had more time to evolve and are more differentiated. In addition, it is difficult to determine from the data if a subclone has expanded to replace the original clone in the more advanced cancers because the CCF of the variants are all 1. It would only be possible to determine this by using multiple samples from the same tumour.

When comparing the predicted clonal counts estimated in cancer samples to the corresponding tissue type, we see a positive correlation between the two values. Although our estimates are higher than those of the somatic mutation load in normal tissue, they follow a similar trend. One factor that may affect these comparisons is that we only have one sample for an individual at a particular age for each normal tissue type. Therefore, the somatic mutation load in the normal tissue does not account for variability between individuals. While on the other hand, we do not have multiple samples from different regions in a tumour from the cancer samples, so the predictions do not account for variability within an individual. There may also be variability in the mutation load in clones from normal tissue; therefore, the samples used may not reflect the mutational burden of a normal cell before cancer initiation. Suppose specific clones within a normal tissue have different mutational burdens, we would expect that the cancer cell is more likely to arise from a cell that has a mutational burden in the higher end compared to other clones in the tissue. Additional analysis using single-cell sequencing data is needed to clarify this.

A high-profile study by Tomasetti and Volgestien in 2015 [73] investigated the relationship between the lifetime number of stem cell divisions (LSCD) of a tissue and lifetime cancer risk (LCR). They found a high correlation between these two values and claimed that this suggested that intrinsic factors, i.e. stem cell divisions, play a far greater role in cancer risk than previously thought. They attributed this correlation to the accumulation of somatic mutations that are incorporated through errors in DNA replication during each stem cell division, i.e. the higher the number of stem cell divisions, the greater the chance of somatic mutations occurring. If this is the case, and somatic mutation burden does explain the relationship observed, we would expect to see a stronger correlation between somatic mutation load and LCR. However, this was not evident from our analysis and similar results were reported by Milholland et. al. [439]. Although we saw a positive correlation between predicted clonal mutation burden and LCR, there is a stronger correlation between LCR and LSCD, which suggests that the relationship between LCR and LSCD is not mediated (or at least not entirely mediated) by increased mutation accumulation.

2.5 Conclusion

In conclusion, the results of this chapter highlight the difficulty of determining clonal mutations within a tumour sample using single-sample bulk sequencing data. Adjusting for tumour purity and copy number variance, the two factors most likely to bias results, still does not solve the problem, with only 45% of clonal variants classified as truly clonal after applying the thresholds that achieve the best sensitivity and specificity. Our results do indicate that it is possible to estimate the true clonal load of a tumour sample by leveraging the known associations of somatic mutation accumulation in normal tissues with age. In this way we can determine how many variants we have accurately classified as clonal for a sample at a specific age and of a specific tissue type.

Comparing the predicted clonal load to the measured somatic mutation load of samples taken from normal tissue indicated that the predicted clonal loads are in line with the expected number of somatic mutations for that tissue type of a particular age. However, there were some deviations which were indicative of additional factors other than age that may be contributing. It will be necessary to develop models for individual tissues types using larger sample numbers in order to further elucidate this.

2.6 Materials and Methods

2.6.1 Data Acquisition

Access to controlled TCGA data was granted through the Genomic Data Commons (GDC). Details of the samples used as provided in Table 2.4. Somatic variant calls from whole exome sequencing (WXS) were obtained from harmonised data produced by the MC3 working group, which aimed to reduce artefact contamination and produce high-confidence somatic calls for the TCGA data [113]. We used publicly available MAF and raw VCF files from each variant caller for our analysis. Focal copy number variant (CNV) files and HTSeq count expression files were downloaded from the GDC. Tumour purity calls were obtained from Thorsson *et al.* [440]. Clinical information and metadata were obtained for the TCGA samples using the TCGAbiolinks (v2.14.1)[441] and TCGAutils (v1.6.2) packages in R (v3.6.2). TCGA genotype data were obtained from Carrot-Zhang *et al.* [442] Lifetime cumulative stem cell division rates and lifetime cancer risk values were obtained from Tomasetti and Vogelstein [73].

Study Code	**Study Name**	**Number of Samples**
LAML	Acute Myeloid Leukemia	98
ACC	Adrenocortical carcinoma	88
BLCA	Bladder Urothelial Carcinoma	380
LGG	Brain Lower Grade Glioma	470
BRCA	Breast invasive carcinoma	571
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	137
CHOL	Cholangiocarcinoma	35
COAD	Colon adenocarcinoma	220
ESCA	Esophageal carcinoma	159
GBM	Glioblastoma multiforme	280
HNSC	Head and Neck squamous cell carcinoma	481
KICH	Kidney Chromophobe	61
KIRC	Kidney renal clear cell carcinoma	246
KIRP	Kidney renal papillary cell carcinoma	155
LIHC	Liver hepatocellular carcinoma	161
LUAD	Lung adenocarcinoma	402
LUSC	Lung squamous cell carcinoma	386
DLBC	Lymphoid Neoplasm Diffuse Large B cell Lymphoma	14
OV	Ovarian serous cystadenocarcinoma	50
PAAD	Pancreatic adenocarcinoma	157
PCPG	Pheochromocytoma and Paraganglioma	126
PRAD	Prostate adenocarcinoma	439
READ	Rectum adenocarcinoma	78
SARC	Sarcoma	218
SKCM	Skin Cutaneous Melanoma	103
STAD	Stomach adenocarcinoma	405
TGCT	Testicular Germ Cell Tumours	127
THYM	Thymoma	92
THCA	Thyroid carcinoma	384

UCS	Uterine Carcinosarcoma	52
UCEC	Uterine Corpus Endometrial Carcinoma	378
UVM	Uveal Melanoma	73

Table 2.4: **Description of the TCGA cancer types used in this study, including the total number of samples available for each.**

2.6.2 Variant Processing

Only single nucleotide variants that passed the filtering criteria for each of the seven variant callers used by the MC3 working group [113] and were called by two or more variant callers were included in the analysis. Only primary tumour samples were used for this analysis. Hyper-mutated samples with >1000 missense mutations were removed from the analysis.

2.6.3 Identification of clonal variants

Clonal status for samples that had CNV and purity estimates were calculated for all variants. This was achieved following the principles outlined in Dentro *et al.* [74] and briefly summarised below.

Variant frequency, f_i, is calculated as:

$$f_i = \frac{r_{\text{mut}i}}{r_{\text{mut}i} + r_{\text{ref}i}} \tag{1}$$

where

r_{muti} = Number of reads supporting the mutation

r_{refi} = Number of reads supporting the reference

Copy number changes and tumour purity can affect the allele frequency of the mutation. Therefore, the adjusted variant frequency, u_i, can be written as follows:

$$u_i = f_i * \frac{1}{p}[(p * n_{\text{tot,t,i}}) + ((1-p) * n_{\text{normal,t,i}})] \tag{2}$$

where

f_i = observed variant frequency calculated in (1)

p = tumour purity

$n_{\text{tot,t,i}}$ = copy number of tumour cells

$n_{\text{normal,t,i}}$ = copy number of normal cells. Assumed to be 2 for autosomes, 2 for the X chromosome in females and 1 for X and Y chromosome in males.

u_i can also be written as a function of the cancer cell fraction (CCF_i) and multiplicity (m_i)), the number of chromosomes that carry the mutation.:

$$u_\text{i} = CCF_\text{i} * m_\text{i} \tag{3}$$

Equation (3) can be rewritten as:

$$CCF_{\mathrm{i}} = \frac{u_{\mathrm{i}}}{m_{\mathrm{i}}} \tag{4}$$

A clonal mutation will have $u_i \geq 1$ while a sub-clonal variant will have CCF less than 1 and will only be carried by a single chromosome copy (assuming it has not been affected by a sub-clonal CNV) so will have $u_{\mathrm{i}} < 1$.

Using these observations we can determine m_{i}:

$$m_{\mathrm{i}} = \begin{cases} |u_{\mathrm{i}}| & \text{if } \mathrm{u_i} \geq 1 \\ 1 & \text{if } \mathrm{u_i} \text{ ¡ } 1 \end{cases} \tag{5}$$

First, $\mathrm{CCF_i}$ was determined for each mutation using (4) by calculating (3) and (5). Then, a Clopper and Pearson 99% confidence interval (CI) [443] was calculated for $\mathrm{u_i}$ based on the number of reads harboring the mutation and the total number of reads covering the mutation assuming a binomial process. Finally, intergenic mutations or mutations in genes that did not have CNV information were removed.

2.6.4 Adjusting thresholds to call a variant

The first set of analyses examined the impact of changing various thresholds for calling a variant clonal or subclonal. We aimed to choose a threshold that would classify as many variants as possible as clonal or subclonal, reducing the number in the undetermined category while ensuring we did not misclassify any variant. To do this, we needed to take the following factors into account:

1. Depth: Choose a depth as low as possible while still being able to classify variants accurately.
2. Confidence Intervals from binomial tests to calculate the cancer cell fraction (CCF):
   (a) Clonal variants: Choose upper and lower confidence intervals that capture as many true clonal variants as possible without misclassifying subclonal variants as clonal. For clonal variants, we expect the CCF to be close to 1, i.e. all cancer cells carry the mutation. However, due to fluctuations in the read counts for a variant, true clonal mutations may be present at a CCF lower than 1. Therefore, we used cut-offs for the true CCF's upper and lower confidence intervals to call a variant clonal.
   (b) Subclonal variant: In this case, we only needed to consider the upper bound for the true CCF, as any variant with a low CCF was assumed to be subclonal (or a technical artefact). We choose an upper bound high enough to capture as many subclonal variants as possible.
3. The total number of variants classified as clonal or subclonal: We aimed to limit the number of variants that cannot be accurately classified and are, therefore, in the ”Undetermined” category. However, this was the least important factor because our models estimate the true clonal load based on

the clonal and undetermined variants. Therefore, it was more important for these models that the clonal mutations are truly clonal and do not incorporate subclonal mutations that have been misclassified.

2.6.5 Calculating Sensitivity and Specificity

We obtained the mutation timing files from Gerstung *et al.* [405] to identify subclonal and clonal variants in the PCAWG data to calculate our calls' sensitivity and specificity. As a result, sensitivity and specificity were calculated as follows:

		Truth Set (PCAWG)	
		Clonal	Not Clonal
Test Results	Clonal	True Positive	False Positive
	Not Clonal	False Negative	True Negative

Table 2.5: **Confusion matrix for comparing TCGA and PCAWG clonal calls.**

$$Sensitivity = \frac{TruePositive}{TruePositive + FalseNegative}$$

$$Specificity = \frac{TrueNegative}{TrueNegative + FalsePositive}$$

2.6.6 Estimation of true clonal load

We utilised the relationship between clonal mutation counts and age to estimate the true clonal load for a sample.

$$T = \beta_0 + \beta_1 A + \varepsilon \tag{6}$$

where

T = True clonal counts

A = Age

β_0 = intercept of equation

β_1 = slope of equation

ε= error

When we used conservative criteria to distinguish clonal, and subclonal mutations, such that only mutations that could be confidently called clonal or subclonal were classified, a large proportion of variants remained unclassified. The following describes an approach to estimate the faction of these unclassified mutations that were clonal. This fraction was then used to estimate the total number of clonal mutations (those confidently classified as clonal plus the estimated proportion of the unclassified variants).

Let the total number of clonal mutations be T, and suppose that a proportion, p, of these have been classified as clonal. We used the following equations to determine T:

$$y_{\mathrm{c}} = \beta_{0\mathrm{c}} + \beta_{1\mathrm{c}}A + \varepsilon \tag{7}$$

$$y_{\mathrm{u}} = \beta_{0\mathrm{u}} + \beta_{1\mathrm{u}}A + \varepsilon \tag{8}$$

$$p = \frac{\beta_{1\mathrm{c}}}{\beta_{1\mathrm{c}} + \beta_{1\mathrm{u}}} \tag{9}$$

$$T = \frac{y_{\mathrm{c}}}{p} \tag{10}$$

where

y_{c} = Observed counts in the clonal category

y_{u} = Observed counts in the undetermined category

We were also able to calculate the proportion of the undetermined category that was truly clonal (q) as:

$$q = \frac{(1-p)T}{y_{\mathrm{u}}} \tag{11}$$

2.6.7 Estimates of somatic mutation load in the normal tissue

We obtained data for seven tissue types from Moore *et al.* [388], who performed single-cell sequencing analysis of biopsies from multiple tissues from 3 individuals. They took multiple samples from each site for most tissues. We used the median clonal mutation burden per genome values from monoclonal (proportion clonal=1) samples for each tissue for each individual (Table 2.6).

Sample ID	**Tissue Type**	**Age**	**Cause of Death**
PD28690	Oesophagus	78	Metastatic oesophageal adenocarcinoma
PD28690	Prostate	78	Metastatic oesophageal adenocarcinoma
PD28690	Thyroid	78	Metastatic oesophageal adenocarcinoma
PD42565/PD43851	Oesophagus	47	Acute coronary syndrome
PD42565/PD43851	Pancreas	47	Acute coronary syndrome
PD42565/PD43851	Prostate	47	Acute coronary syndrome
PD42565/PD43851	Skin	47	Acute coronary syndrome
PD42565/PD43851	Stomach	47	Acute coronary syndrome
PD43850	Liver	54	Traumatic injuries
PD43850	Pancreas	54	Traumatic injuries
PD43850	Skin	54	Traumatic injuries

Table 2.6: **Details of tissue types and age for samples taken from Moore *et al.***

2.6.8 Genome-Wide Association Study

We obtained quality-controlled, stranded and imputed genotyping files from Sayaman *et al.* [444] for the TCGA samples. Chr3 was excluded from this analysis as the imputed file was corrupt, and we could not retrieve it from the authors. We also obtained principle component analysis results with the first 3-4 principle components (PCs) capturing population structure information and PCs 5 and 6 capturing outliers. We used their PAM ancestry calls to limit our analysis to samples of European ancestry and reduce the impact of population structure on our results. The European group was chosen as it is the largest group in the dataset (n=3,383). We used PLINK v2.00a3LM [445] to filter SNPs based on the following criteria: maf >0.005, hwe with midp adjustment p $<1.0 \times 10^{-6}$, variants with missing call rates >0.02, variant pruning: window size =100-, shift=500, r^2 threshold =0.2, excluded all instances of variants with duplicate ID and excluded variants with more than two alleles. There were 581,875 SNPs remaining for our analysis. We applied an inverse normal rank transformation to the estimated clonal load and used this as the phenotype for GWAS analysis. Sex, cancer type, age and PC1-7 were used as covariates. Sex and cancer type were used as categorical variables, and all others were continuous. We used GCTA 1.93.2beta MLMA for the association testing. FUMA online tool was used to annotate, visualise and interpret GWAS results [446].

3 Chapter 3: Investigating the relationship between gene expression and immune evasion in the context of clonal passenger mutations

The results presented in this chapter motivated Kherreh N, Cleary S, Seoighe C. No evidence that HLA genotype influences the driver mutations that occur in cancer patients. Cancer Immunol Immunother. 2022 Apr;71(4):819-827. doi: 10.1007/s00262-021-03028-w. Epub 2021 Aug 21. PMID: 34417841; PMCID: PMC8921139. Part of the introduction and paragraphs entitled "Susceptibility to cancer based on HLA alleles" and "No evidence that HLA genotype influences the driver mutations that occur in cancer patients" in the results section of this chapter were published in this article.

I performed all data analysis except for susceptibility to cancer based on HLA alleles which was carried out by Cathal Seoighe, and re-analysis of the Marty et al. *papers, which Noor Kherreh carried out.*

3.1 Abstract

The major histocompatibility (MHC) complex can present neoantigens resulting from somatic mutations on cell surfaces, potentially directing immune responses against cancer. It has been reported that driver mutations observed in a cancer occur in the gaps in a patient's ability to present that particular mutation to the immune system, which is controlled by the patient's MHC genotype. Although this finding was reported for driver mutations, the same was not observed for passenger mutations. Therefore, we hypothesised that immunogenic passenger mutations escaped immune recognition by different mechanisms related to gene expression. Here, we investigated whether immunogenic passenger mutations were tolerated due to their presence in lowly expressed genes. We also investigated whether genes harbouring these mutations were downregulated compared to the normal expression of that gene, and we specifically assessed whether the mutant allele was downregulated compared to the normal allele in a process known as allele-specific expression. When controlling for gene length and sequence context, we found no evidence of immune evasion by these mechanisms. In addition, we estimated an upper bound for the impact of immunoediting on the mutation landscape in cancer. This upper bound indicates that at most 5% of missense mutations are removed by immunoediting; however, the data are consistent with no mutations being lost through immunoediting.

3.2 Introduction

The immune system has evolved to recognise aberrant and non-self-molecules resulting from pathogen infection, somatic mutations and malformed proteins. The major histocompatibility complex (MHC) plays a crucial role in this process. There are two classes of MHC molecules, class I (MHC-I) and class II (MHC-II), encoded, in humans, by a cluster of genes on chromosome 6. The human MHC genes and proteins, often termed Human Leukocyte Antigens (HLA), are diverse, with over 15,000 alleles identified [447]. Somatic mutations in genes encoding self-proteins can result in an altered amino acid sequence, thereby generating so-called neoantigens that have the potential to elicit an immune response upon presentation by the MHC to T-cells [448]. By killing cells carrying immunogenic neoantigens, the immune system has been proposed to play a vital role in shaping the cancer genome in a process referred to as immunoediting [449, 450].

Dunn *et al.* first proposed the term immunoediting to describe the dual ability of the immune system to defend the host by suppressing tumour growth and shaping the immunogenicity of tumours [174]. It is characterised by three phases – elimination, equilibrium and escape, collectively termed the three Es of cancer immunoediting [174, 199]. The elimination phase involves the recognition and destruction of tumour cells by the immune system before they are clinically detectable. Some cells are thought to escape elimination and enter into the equilibrium phase, during which the immune system keeps tumour growth in check but cannot entirely eliminate it. The tumour may continue to develop mutations that enable it to evade immune recognition, resulting in a population of cells resistant to the immune response [199, 450]. The final stage occurs when the cancer escapes immune control, leading to uncontrolled proliferation, due potentially to reduced immunogenicity of cancer cells or to mutations that create an immunosuppressive environment [199, 449].

One of the mechanisms through which cancer evades the immune response is to acquire mutations that alter antigen presentation [451]. The most selective step of antigen presentation to the immune cells is the binding of antigenic peptides to the MHC, which has been inferred by a variety of studies of the implications of mutating the HLA genes or the B2M gene, whose product, β_2m, forms an integral part of MHC Class I molecules [264, 452–455]. Loss or mutation of HLA or B2M genes is associated with increased tumour mutation burden [455]. A lack of neoantigens capable of eliciting an immune response could also allow cancers to avoid immune responses, and several studies have reported selection against immunogenic somatic mutations in cancer [201, 203, 453]. However, the evidence for depletion of mutations that give rise to neoantigens has recently been questioned [233].

Two high-profile studies [205, 206] reported that the driver mutations found in cancer patients could be predicted from the capacity of the patient's HLA alleles to bind the resulting neoantigens. The patient harmonic mean best rank (PHBR) score was proposed in Marty *et al.* [205] and Marty Pyke *et al.* [206] as a measure of whether MHC molecules can bind a neoantigen resulting from a somatic mutation, given the HLA genotype of a patient. The score is derived from predicted binding affinities of the patient's MHC molecules for the peptides spanning the mutation. The conclusions of both studies are based on an analysis of 1,018 cancer driver mutations in patients from the cancer genome atlas (TCGA). The focus of the 2017 study is on MHC class I alleles, and the primary focus of the 2018 study is on the presentation of cancer neoantigens by MHC class II molecules. Although they found that the patient's HLA alleles could predict the driver mutation landscape, the same was not true for passenger mutations. This finding is surprising as passenger mutations should be under the same selection pressures as driver mutations, with immunogenic passenger mutations also removed by the immune system. We hypothesised that immunogenic passenger mutations occur preferentially in lowly expressed genes, which are less likely to be presented to the immune system, and those passenger mutations that occurred in highly expressed genes were removed by immune cells and are not present in the cancer cell. Here we investigated the differences between the expression of genes harbouring missense passenger mutations to those harbouring synonymous mutations that are not subjected to selection. We investigated if missense mutations occur in lowly expressed genes or if genes harbouring missense mutations are downregulated. We also specifically investigated differences in the expression of the mutant and reference alleles to test for evidence of downregulation of the mutant allele as a potential mechanism of immune escape.

3.3 Results

To assess whether passenger mutations are preferentially occurring in lowly expressed genes, we performed several comparisons that used synonymous mutations as a proxy for neutrality. Synonymous mutations do not change the amino acid composition of the gene and therefore tend to have less selection pressure acting upon them. The 68,936 missense mutations included in this analysis were found in 11,926 unique genes, while the 24,679 genes with synonymous mutations were found in 9,221 individual genes. Of those, 8,586 genes overlapped between the two groups, i.e., the gene harboured missense mutations in one or more samples but synonymous mutations in others. Several genes had multiple missense mutations for a single sample. We ensured that we only included such genes once in our analysis.

Some genes harboured multiple missense mutations within an individual and were also recurrently mutated across samples (Table 3.1). Most of these genes were longer than the average gene length (6,507 base pairs), meaning there were more sites within the gene at which a mutation could occur. The highest number of missense mutations in a gene for an individual sample was 22, while the highest number of synonymous mutations in a gene was three. It could be that our method to identify genes with missense or synonymous mutations biases it so that genes with a lower mutation rate were found predominantly in the synonymous group because genes with a missense mutation could also have a synonymous one. However, this was unlikely due to the considerable overlap of genes seen in the two groups. The average gene length in the missense group was 6,529 base pairs, while the average gene length in the synonymous group was 6,932 base pairs.

3.3.1 Tolerance of missense mutations in lowly expressed genes

We compared the expression of genes harbouring missense mutations to the expression of all genes within a sample. We hypothesised that if missense mutations occur preferentially in lowly expressed genes, then there should be a high proportion of genes with a missense mutation that have expression values lower than the median expression of all genes within that sample (Figure 3.1). We compared the proportion of genes with missense mutations with expression lower than the expression of all genes within the sample to the proportion of genes with synonymous mutations that have expression lower than all genes (Figure 3.1) and found that there was a slightly higher proportion of missense genes with expression lower than the average (P=0.004). This would indicate that missense mutations are occurring preferentially in genes that have lower expression than half the genes within that sample.

Gene Symbol	Number of samples with a missense mutation	Highest number of missense mutations in a sample	Gene Length	GC Content
TTN	417	22	118976	0.37
MUC16	205	12	43830	0.45
AHNAK2	162	5	18788	0.58
SYNE1	109	7	47523	0.39
PCLO	108	5	22874	0.35
NEB	90	9	33502	0.39
MUC5B	72	5	18598	0.65
RYR1	52	5	16282	0.60
DNAH3	47	7	15023	0.46
DNAH9	47	6	16941	0.51
DST	45	6	48142	0.39
F5	45	5	9373	0.38
FAT2	36	5	14712	0.48
NCKAP5	36	5	9558	0.39
MYH3	33	6	6684	0.48
FBN1	33	5	18525	0.40

Table 3.1: **The most recurrently mutated genes across samples and the highest number of missense mutations within the gene for an individual sample.**

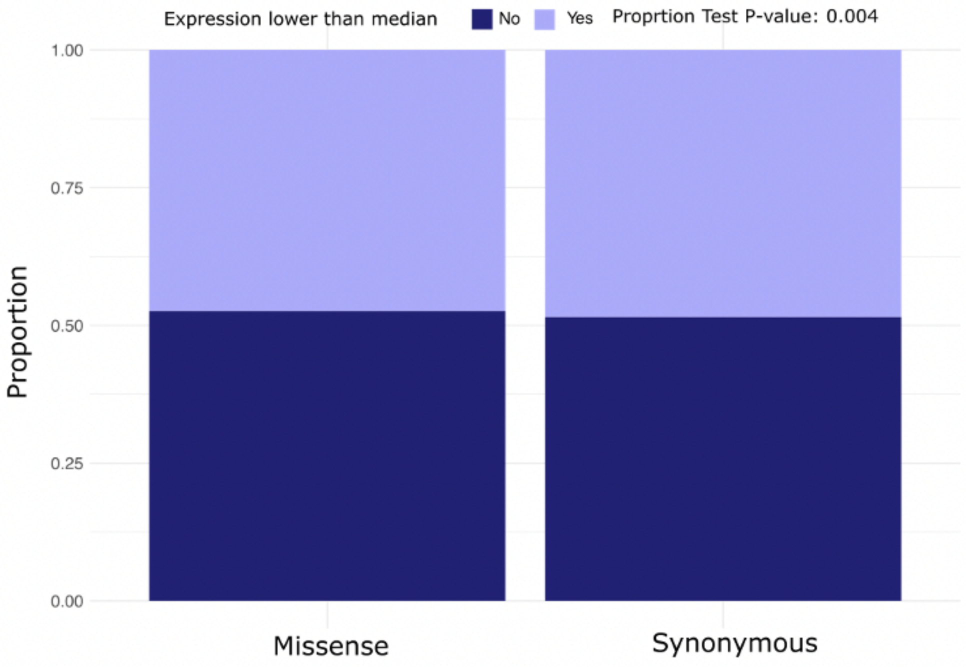


Figure 3.1: **Comparison of genes whose expression is lower than the median expression of genes for missense versus synonymous mutations.** Stacked bar plots comparing the conditional proportion of genes whose gene expression is lower than the median gene expression for individual samples for the two mutation types.

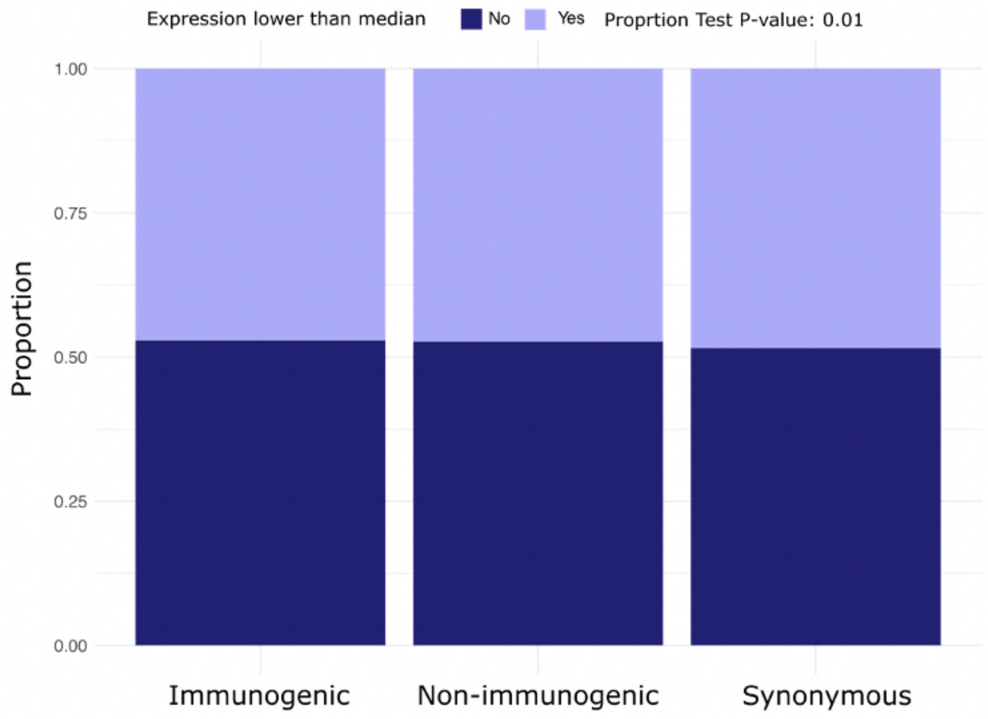


Figure 3.2: **Comparison of genes whose expression is lower than the median expression of genes with missense mutations split into those that are immunogeneic and non-immunogenic.** Stacked bar plots comparing the conditional proportion of genes whose gene expression is lower than the median gene expression for individual samples for missense mutations split into immunogenic and nonimmunogenic based on their PHBR score, compared to synonymous mutations.

3.3.2 Comparison of immunogenic, nonimmunogenic and synonymous gene expression

If neoantigens are being downregulated to escape immune recognition, we would expect that genes harbouring mutations deemed immunogenic based on their PHBR score would have lower expression than genes harbouring nonimmunogenic mutations or mutations that do not change the amino acid composition of the protein. To test this, we split missense mutations into immunogenic (PHBR <2) and nonimmunogenic (PHBR ≥ 2) genes and compared the TPM expression score for genes with immunogenic, nonimmunogenic and synonymous mutations. We compared the proportion of genes that have expression lower than the median expression within a sample for all three groups (Figure 3.2). We do see a significant difference ($P = 0.001$) between the groups, with a trend as expected for immunogenic mutations to have a higher proportion of genes with expression lower than the median compared to nonimmunogenic or synonymous mutations. These results indicate that immunogenic mutations are preferentially occurring on lowly expressed genes. However, the difference is marginal (proportions lower than the median of 0.53, 0.53, 0.52 for genes with immunogenic, nonimmunogenic and synonymous mutations, respectively).

3.3.3 Simulated mutations in the same sequence context as observed mutations

To assess whether the differences in gene expression were due to immunoediting and not due to differences in the mutagenic processes between genes, we assigned each mutation to a random position in the same gene with the same sequence context as the observed mutation (Figure 3.3). We then computed the consequence of these changes and calculated the proportion of missense mutations for each gene with the randomly assigned mutations. By doing this, we controlled for gene length and sequence context in our analysis. If the difference in expression was due to selection by the immune system, we would expect to see a difference between the proportion of missense mutations in the observed and random datasets. When we compared the proportions of missense mutations in the random set with those observed for each gene we found no significant difference (P=0.9) (Figure 3.4). If the immune system resulted in the removal of missense mutations in the real dataset, we would expect the proportion of missense mutations to be lower in the observed data than in the random data. However, this is not the case. Although we see a slight difference between the densities of the observed and random datasets, coloured in blue and yellow, respectively, the random density deviates from the observed at both high and low proportions.

To assess the level of immunoediting required to observe a difference in the missense rate, we randomly removed 1,2,3,4 5,6,7,8,9, 10, 15 and 20% of missense mutations from the random dataset (Figure 3.5). We then randomly assigned mutations of the same mutation context within the same gene for the subsets of data, using the same process as before (Figure 3.3), and compared the proportions. When 5% or fewer of mutations were selectively removed, we saw no difference between the medians of the two groups (Figure 3.6). Although a statistical test comparing the medians of both groups shows a significant difference (p-value <0.05) at 3%, we do not observe a stable result until we remove 5% of missense variants. At 5% the result is significant (p=0.003), but the medians are the same indicating that the significance is caused by a slight difference in the shape of the distribution of the two sets of data. When 6% or more of missense mutations were removed, we saw a significant difference (P=0.0004), including a difference between the two medians. The data for which we simulated immunoediting had a lower proportion of missense mutations compared to randomised data (median missense proportion of 0.74 compared to 0.75), which is what is expected for negative selection. This indicates that we should observe selection by immunoediting when the immune system removes at least 5% of missense mutations. [456].

We also assessed the relationship between the proportion of missense mutations and gene expression, expecting that genes with high expression would have a lower proportion of missense mutations (Figure 3.7). We see a slight difference in the pattern (as indicated by fitting generalised additive models) for the observed mutations compared to the random dataset. However, the confidence bands (grey area) overlap between the random and observed fitted lines, indicating no significant difference. To confirm this, we assessed the difference in the missense proportion for the random and observed groups for genes whose log TPM expression was greater than 4 (cutoff determined from Figure 3.7) using a paired Wilcox test and found no statistically significant difference (P=-0.59). Therefore, as suggested by the overlapping confidence bands, the slight decrease in the proportion of observed

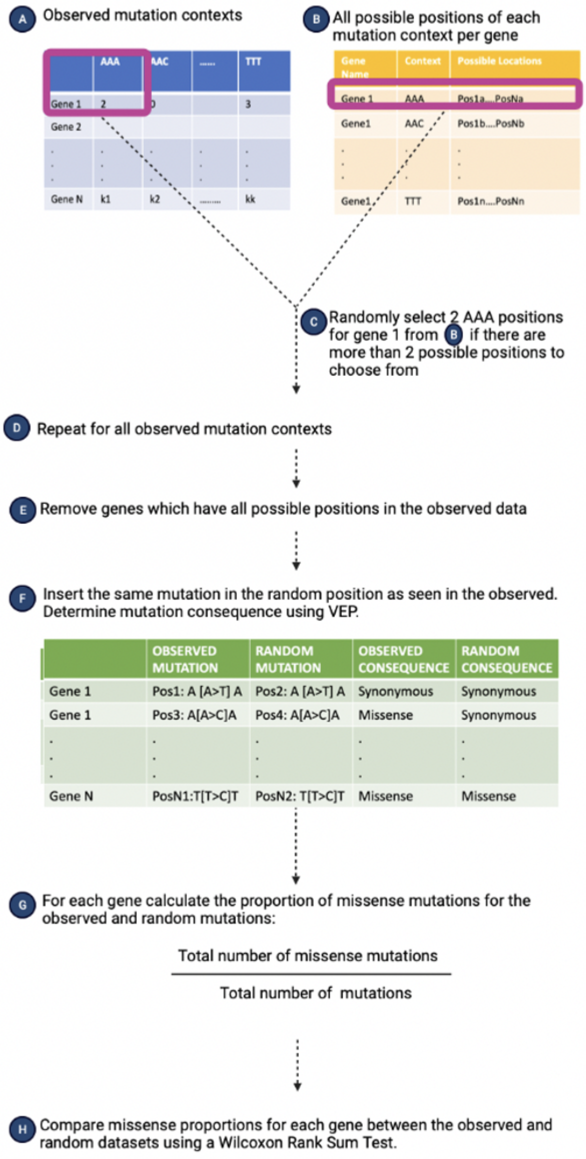

A
Observed mutation contexts
AAA
AAC
....
TTT
Gene 1
2
0
3
Gene 2
Gene N
k1
k2
....
kk
B
All possible positions of each mutation context per gene
Gene Name
Context
Possible Locations
Gene 1
AAA
Pos1a....PosNa
Gene1
AAC
Pos1b....PosNb
Gene1
TTT
Pos1n....PosNn
C
Randomly select 2 AAA positions for gene 1 from B if there are more than 2 possible positions to choose from
D
Repeat for all observed mutation contexts
E
Remove genes which have all possible positions in the observed data
F
Insert the same mutation in the random position as seen in the observed. Determine mutation consequence using VEP.
OBSERVED MUTATION
RANDOM MUTATION
OBSERVED CONSEQUENCE
RANDOM CONSEQUENCE
Gene 1
Pos1: A [A>T] A
Pos2: A [A>T] A
Synonymous
Synonymous
Gene 1
Pos3: A[A>C]A
Pos4: A[A>C]A
Missense
Synonymous
Gene N
PosN1:T[T>C]T
PosN2: T[T>C]T
Missense
Missense
G
For each gene calculate the proportion of missense mutations for the observed and random mutations:
Total number of missense mutations
Total number of mutations
H
Compare missense proportions for each gene between the observed and random datasets using a Wilcoxon Rank Sum Test.

Figure 3.3: **Workflow describing the creation of a random dataset of mutations with the same mutational context as observed mutations.** (**A**) Mutation contexts are assigned to each observed mutation, and the total number of observed mutations of each context type is counted for each gene. (**B**) A list of all possible positions that could be mutated for each context for each gene. (**C** & **D**) The exact number of positions as observed in the real dataset were randomly sampled from the list of all possible positions for that context in that gene. (**E**) The gene was removed from the analysis if all possible positions were present in the observed data. (**F**) The random position was mutated to the same allele as in the observed data, and the variant consequence was annotated using VEP online tool. (**G**) The proportion of missense mutations for each gene was calculated for the observed and random datasets. (**H**) A paired Wilcoxon rank sum test was performed to assess whether the two datasets differed.

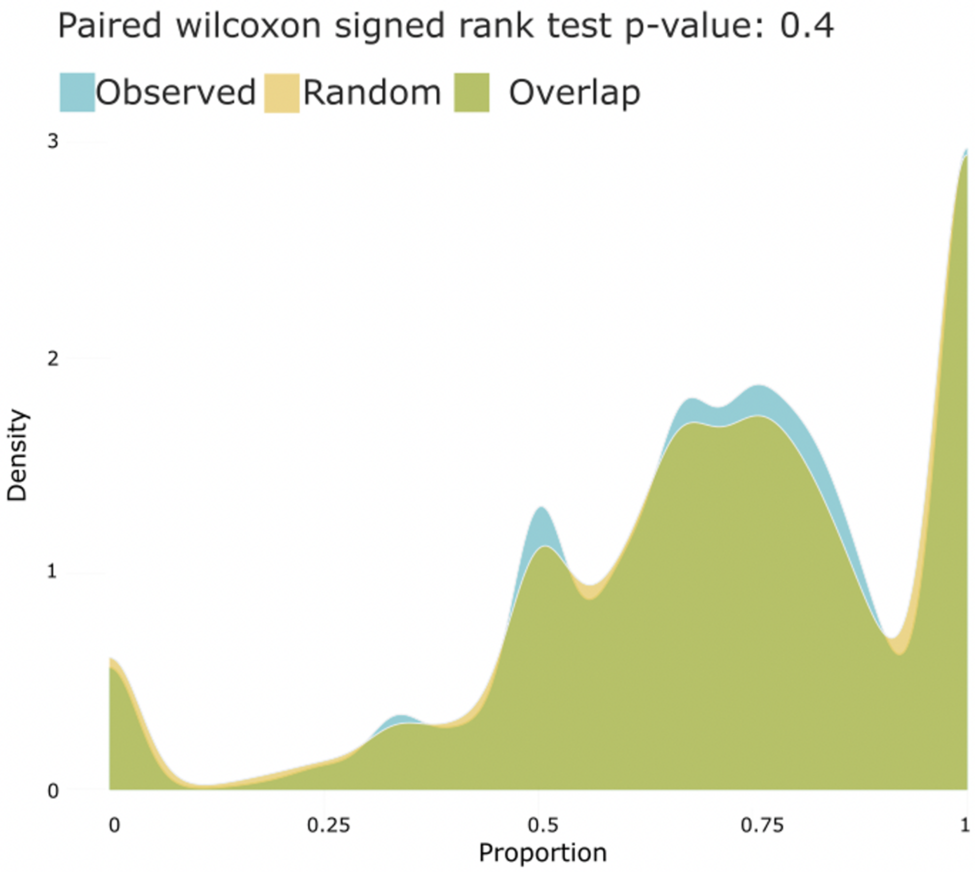


Figure 3.4: **Comparison of the proportion of missense mutations per gene for the observed versus random dataset.** Overlapping density plots showing the proportion of mutations classified as missense for each gene in the observed data (blue) and in the randomly assigned mutations for the same mutational context (yellow).

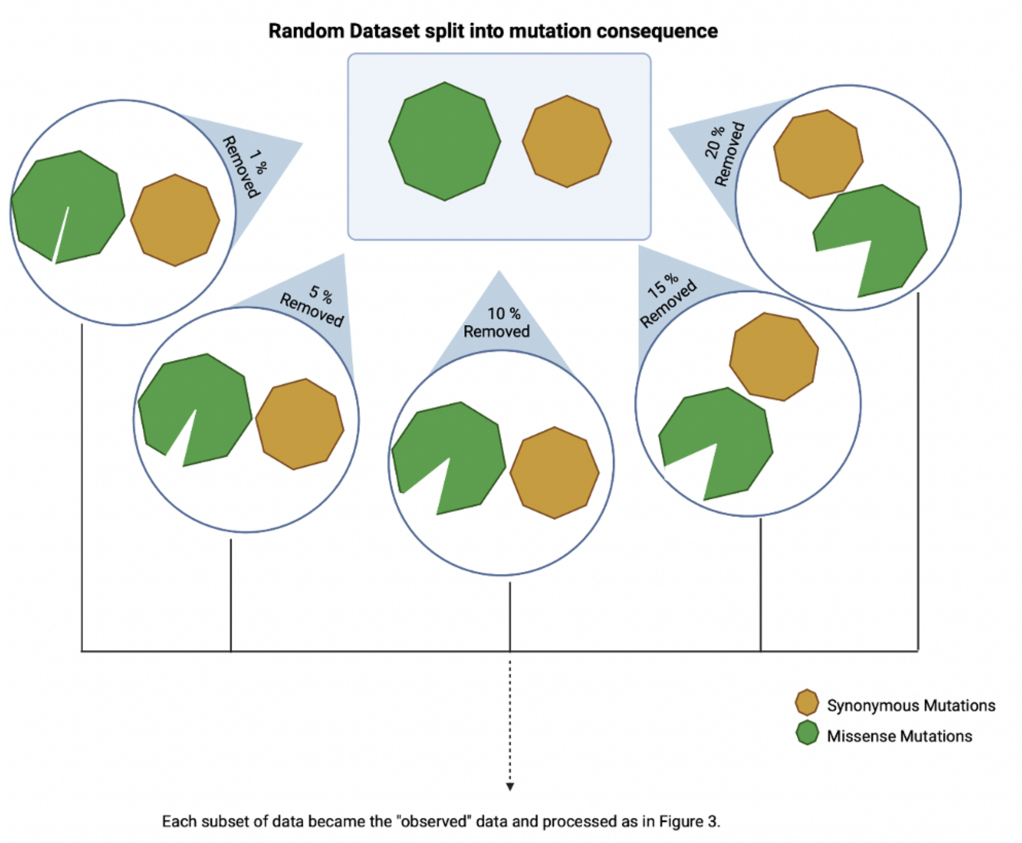

Random Dataset split into mutation consequence
1 % Removed
5 % Removed
10 % Removed
15 % Removed
20 % Removed
Synonymous Mutations
Missense Mutations
Each subset of data became the "observed" data and processed as in Figure 3.

Figure 3.5: **Schematic to illustrate the process of randomly removing different proportions of missense mutations from the data.** Randomly removed 1, 5, 10, 15 and 20% of the missense mutations from the random dataset created in Figure 3.3.

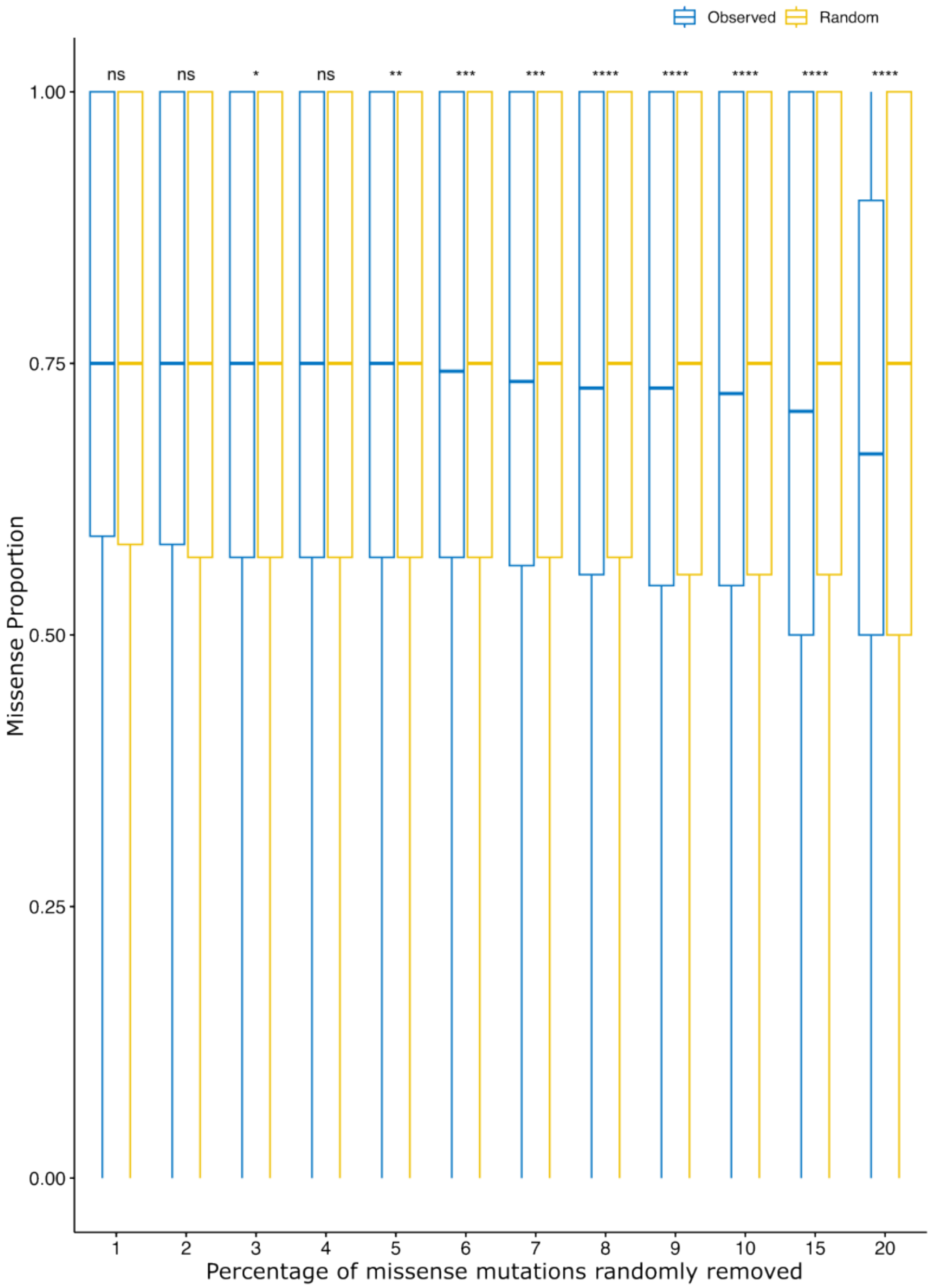


Figure 3.6: **Comparison of the proportion of missense mutations per gene for the simulated datasets with a proportion of missense mutations removed versus the corresponding random dataset.** Boxplot P-values are from paired Wilcoxon rank sum tests.

missense mutations for highly expressed genes is not dissimilar to the proportion of missense mutations in the randomly selected mutations. This means that the decline is not a result of immunoediting but is likely due to other factors, such as repair mechanisms that act on highly expressed genes in general

3.3.4 Downregulation of genes harbouring missense mutations as a means of immune escape

We previously compared expression of a gene with a missense mutation to the expression of all genes within a sample. We next compared the expression of a gene with a missense gene to the expression of the same gene between samples of the same tissue. If the expression of the gene with the missense mutation was lower than the median expression across all samples of that particular cancer type, it would suggest that the gene was being downregulated compared to the normal expression of that gene. A gene can be mutated in multiple samples, so we used the number of genes with lower expression compared to the median as evidence of downregulation for that gene. We performed binomial tests for each gene to test the probability of a gene with a missense mutation having lower expression than the median expression of that gene. After applying multiple test corrections, we found that none of the genes had a statistically significant difference in expression. Thus, there is a lack of evidence that genes harbouring a missense mutation are downregulated as a means to escape the immune system.

3.3.5 Downregulation of mutated allele compared to the normal allele

A potential mechanism through which cancer could evade an immune response directed against a neoantigen is through downregulation of the allele carrying the mutation targeted by the immune system. This could result in allele-specific expression (ASE). We investigated whether the mutant allele's expression was lower than the reference allele for the clonal nonsynonymous mutations in our dataset. For variants exhibiting ASE at the SNV level, we compared the number of variants with lower expression of the mutant allele (relative to the reference allele) for both missense and synonymous variants (Figure 3.8). Although the proportion of variants with lower mutant allele expression compared to the normal is high, as we would expect, the proportion of variants in the synonymous group is not significantly different from the proportion of variants in the missense group (P= 0.46). We also investigated ASE of other nonsynonymous mutation types, stop lost, stop gained and start lost (Figure 3.9) and showed that these mutation types were behaving as expected, with the mutated allele showing lower expression than the normal for stop gained and start lost mutations.

Next, we split missense mutations into those that should elicit an immune response (immunogenic) and those that would not (nonimmunogenic) based on their PHBR score for samples with scores available (n= 4,796) and found no statistically significant difference (P= 0.66), thus providing no evidence that the mutant allele is downregulated compared to the normal allele as a means to evade the immune system.

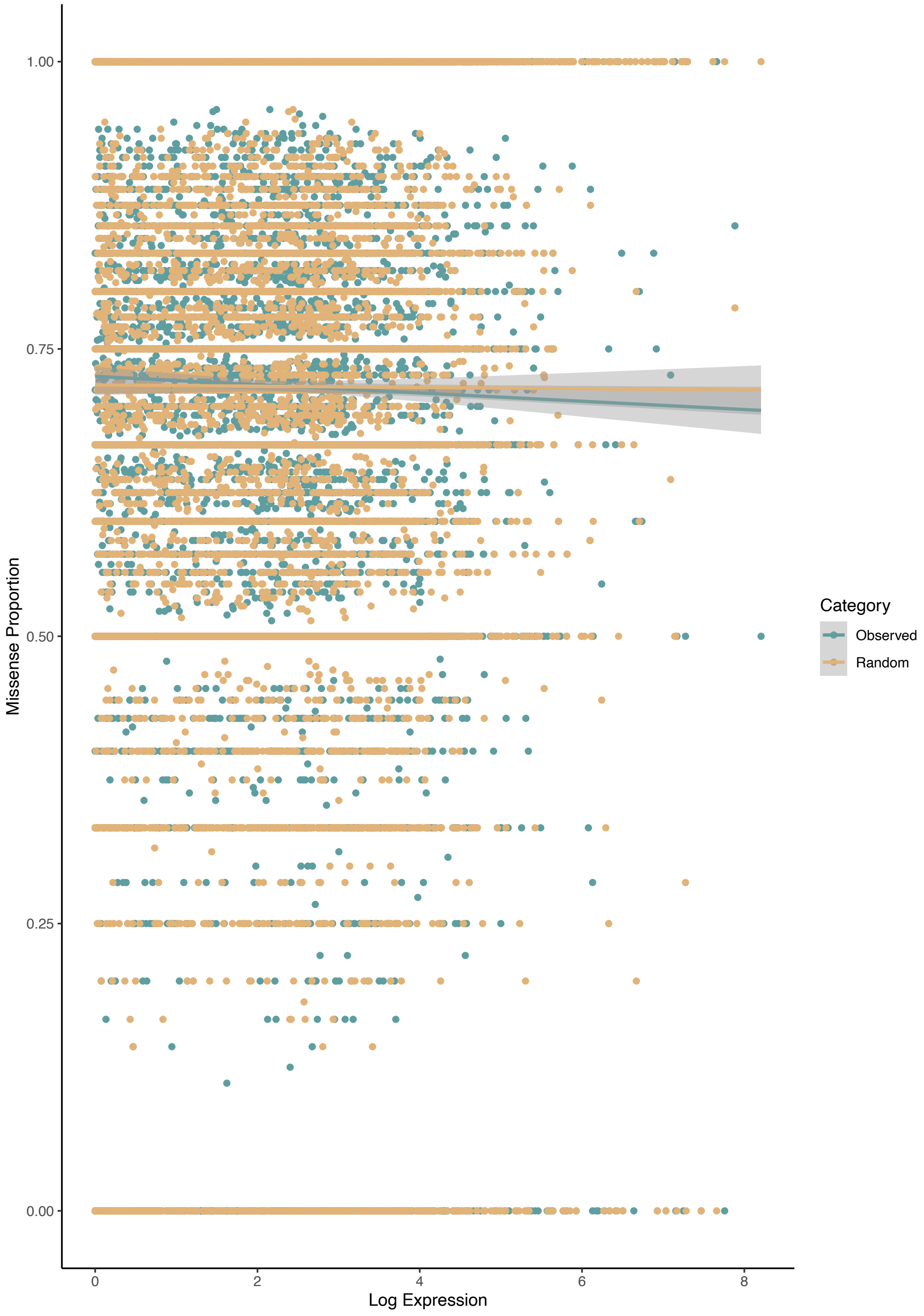


Figure 3.7: **The impact of gene expression on the proportion of missense mutations present in each gene.** The proportion of mutations classified as non-synonymous for each gene in the observed data (**A**) and in the randomly assigned mutations for the same mutational context (**B**) as a function of gene expression.

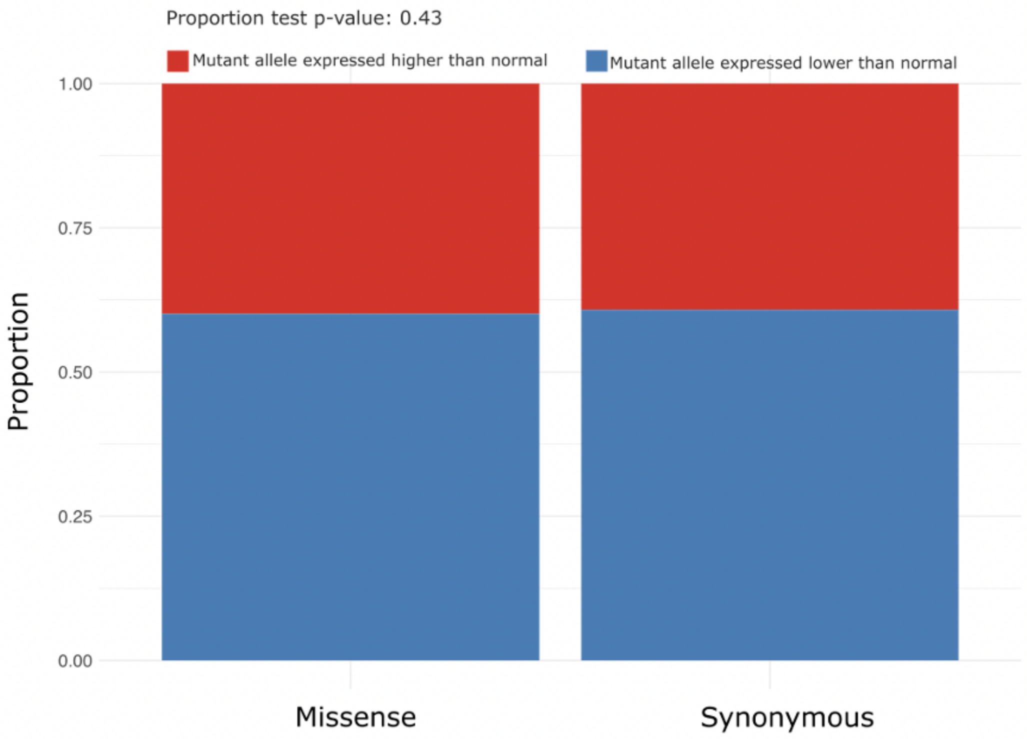


Figure 3.8: **Comparison of allele expression for mutations that are missense versus those that are synonymous.** Stacked Bar plots comparing the conditional proportion of missense and synonymous mutants whose allele expression is greater than (red) or lower than (blue) the normal allele

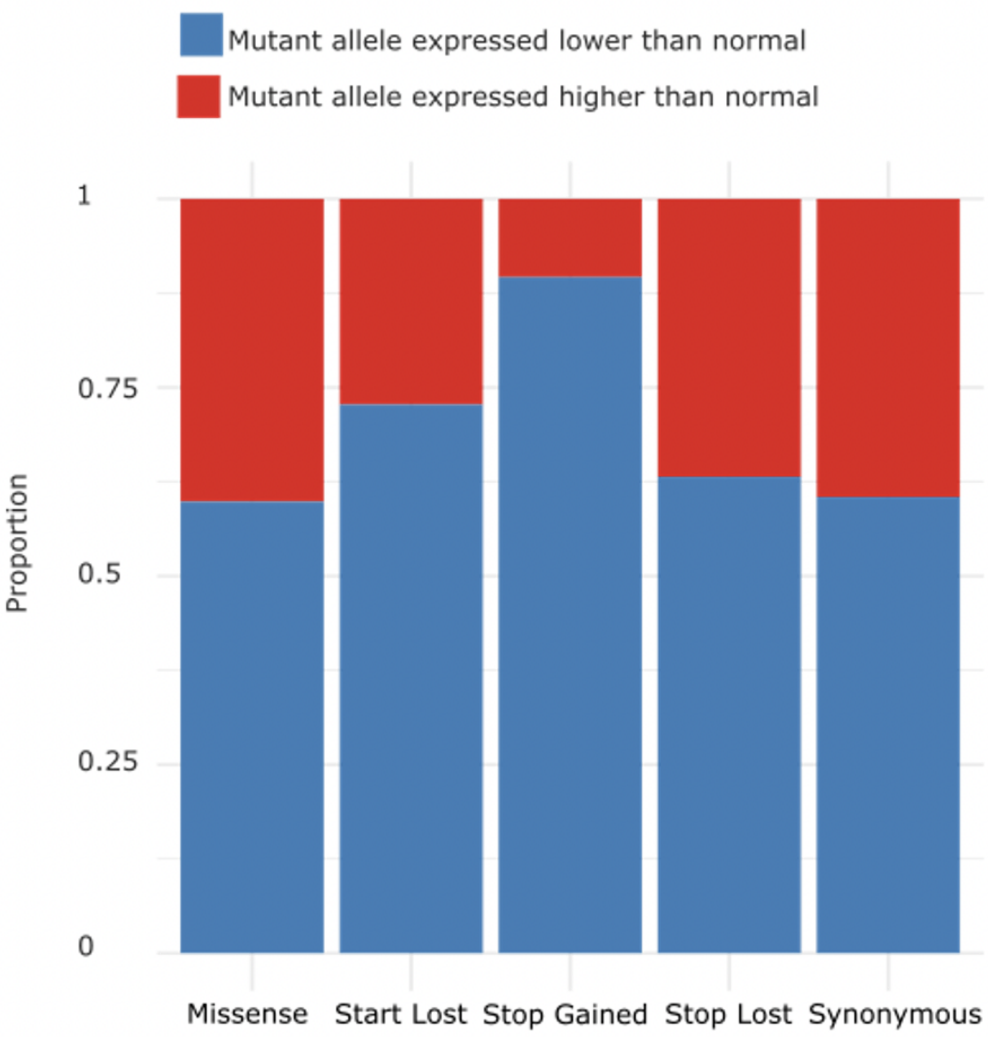


Figure 3.9: **Comparison of allele expression for synonymous and all nonsyonymous mutation types.** Stacked Bar plots comparing the conditional proportion of synonymous and all nonsynonymous mutation types whose allele expression is greater than (red) or lower than (blue) the normal allele.

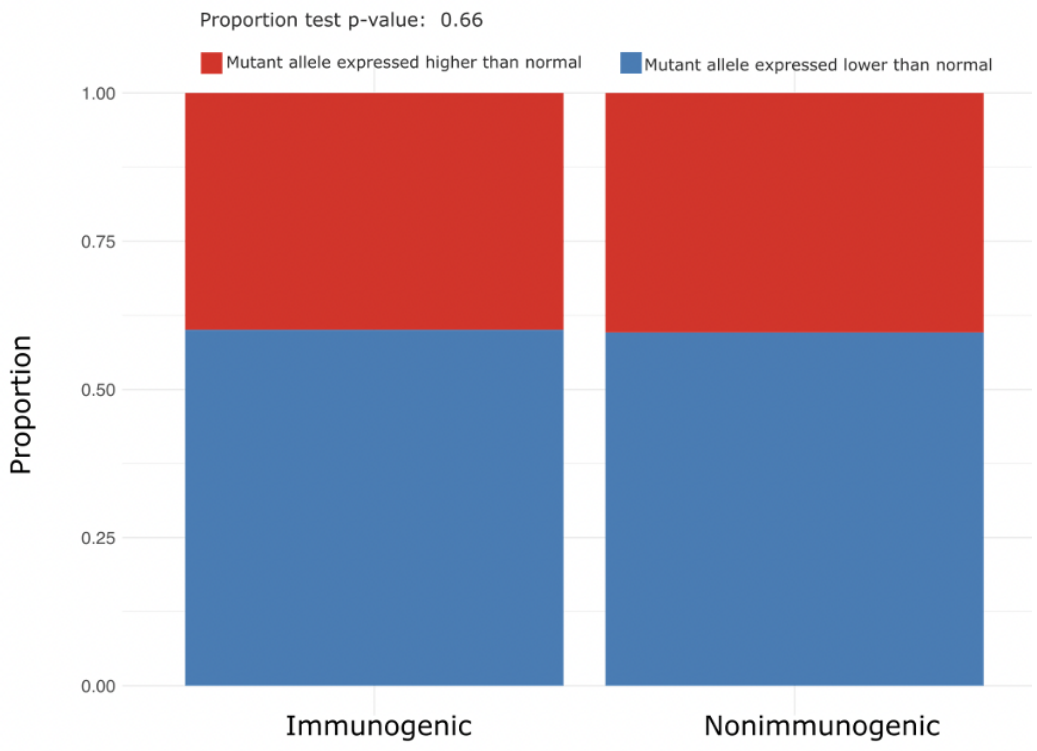


Figure 3.10: **Comparison of allele expression for missense mutations that are immunogenic versus those that are nonimmunogenic.** Stacked Bar plots comparing the conditional proportion of missense mutations split into immunogenic and nonimmunogenic based on their PHBR score whose allele expression is greater than (red) or lower than (blue) the normal allele.

3.3.6 Immune escaped versus non-immune escaped

The selection pressures differ between samples that have acquired mechanisms to escape immune evasion versus samples that have not. Therefore, we investigated whether including these samples affected our results. Samples that have evaded the immune system would no longer be under the same constraint and, therefore, should be free to accumulate mutations in highly expressed genes. However, this would depend on the time at which the sample escaped immune evasion, as mutations that occurred before immune evasion would be under the same constraint as samples that have not evaded the immune system at all.

342 samples contained a clonal nonsynonymous mutation in one of the antigen presentation or immune evasion genes. We removed these samples and repeated the analyses (Figures 5.1-5.4 in Appendix B). We saw no difference in the expression of genes harbouring missense mutations compared to synonymous mutations with regards to the proportion of genes that were expressed lower than the median gene expression within a sample (P=0.3). Furthermore, we saw no difference when we split missense mutations into immunogenic and nonimmunogenic based on their PHBR score (P=0.8). The slight difference we saw between the groups using the complete data set was absent when we excluded these samples. This was surprising because if the immunogenic passenger mutations were preferentially occurring in lowly expressed genes as an alternative means to escape detection by the immune system, we would expect to see a greater effect when immune escaped samples were removed. However, it is important to note again that this would depend on the time at which the sample escaped immune evasion.

We saw no statistically significant difference when we compared the proportion of mutant alleles expressed lower than the normal allele for missense and synonymous mutations (P=0.9) and immunogenic versus nonimmunogenic (P=0.64). We also performed binomial tests to test if genes were downregulated compared to their normal expression across samples and found no statistically significant difference after multiple test correction.

3.3.7 Impact of changing threshold for calling clonal variants

The thresholds we used to call a clonal variant could also influence our results. We used strict thresholds to call a clonal variant, as determined in Chapter 2. However, we might have excluded some real clonal mutations, which would result in some genes being classed as having a synonymous mutation only when they do have a clonal missense mutation. Therefore, we applied much lower thresholds for calling a clonal variant, with cancer cell fraction upper limit >0.6 and lower limit >0.4. We saw similar results as reported for the more stringent thresholds (Figures 5.1-5.4 in Appendix B), which indicated that this was not affecting results.

3.3.8 Susceptibility to cancer based on HLA alleles

Marty *et al.*'s findings have significant implications in terms of the potential to predict susceptibility to cancer occurrence based on a patient's MHC alleles. We investigated this by fitting a logistic regression model to the log odds of cancer status to PHBR coverage, using age and sex as covariates, in the UK Biobank data and showed that HLA homozygosity was not a potential cancer risk factor (P=0.15). Coverage is the number of common driver mutations that can be presented to the immune system by patients in the UK Biobank. The lack of association between PHBR coverage and cancer risk did not support the finding that cancer driver mutations occur in gaps in the ability of a patient's MHC alleles to bind the mutations.

3.3.9 No evidence that HLA genotype influences the driver mutations that occur in cancer patients

Based on our inability to prove both of our hypotheses that stemmed from the Marty *et al.* papers, our group re-analysed their results. In doing so, we discovered that specific mutations in the dataset could be seen in many patients and, as a result, contributed many data points in the comparisons of PHBR scores between the ”no mutation” and ”mutation” groups. If a mutation that occurred many times happened to have high PHBR scores (indicating low immunogenicity), then it would skew the results, especially since there was a high correlation between scores across patients (Figure 3.11A). This would mean that there was a lack of independence between the scores, in violation of the assumption of the statistical tests performed. When we shuffled the MHC genotypes between patients, we found no difference compared to the real data (Figure 3.11B), with the difference between the mutation group and no mutation just as large in the shuffled dataset. This indicated that the difference between the no mutation and mutation groups was not driven by the patient genotype.

Marty *et al.* used mixed effects models to account for the non-independence of observations in their study. Two separate models were used to account for differences in the frequencies of different driver mutations (referred to as the ’within-mutation model’) and differences in the number of driver mutations between patients (referred to as the ’within-patient model’). However, they did not find a significant result when they used the within-mutation model, suggesting that their results were driven by differences in driver mutation frequencies, which were not accounted for in the within-patient model. The failure to detect a result using the within-mutation model was explained as being because the mutation frequencies were high in the cases where they cannot be presented to the immune system, and this was due to the failure of common MHC alleles to bind the peptides. We explored this explanation by shuffling the MHC genotypes and comparing the PHBR scores between the real and shuffled data. We saw no difference between the two groups (P=0.69), suggesting that it was not the HLA alleles that were responsible for this result. We also found that common mutations in the dataset, such as BRAF V600E, were found to be immunogenic in a large number of patients. Therefore, the reason it was common could not be explained by a failure of the most common HLA alleles to present it to the immune system.

When we compared scores split by the recurrence number of the mutations instead of combining all results, we saw no difference between the mutation and no mutation groups (Figure 3.11C). This further shows that the difference Marty *et al.* observed between the mutation and no mutation groups was caused by a small number of highly recurrent mutations that happened to have high PHBR scores. We also showed that highly recurrent mutations tend to occur in the same genes and that there was a high correlation between PHBR scores for mutations within the same gene. This indicated that mutations within the same gene, which could be of the same amino acid class which can affect binding scores [233], could be responsible for what appears to be a link between the recurrence of the mutation and

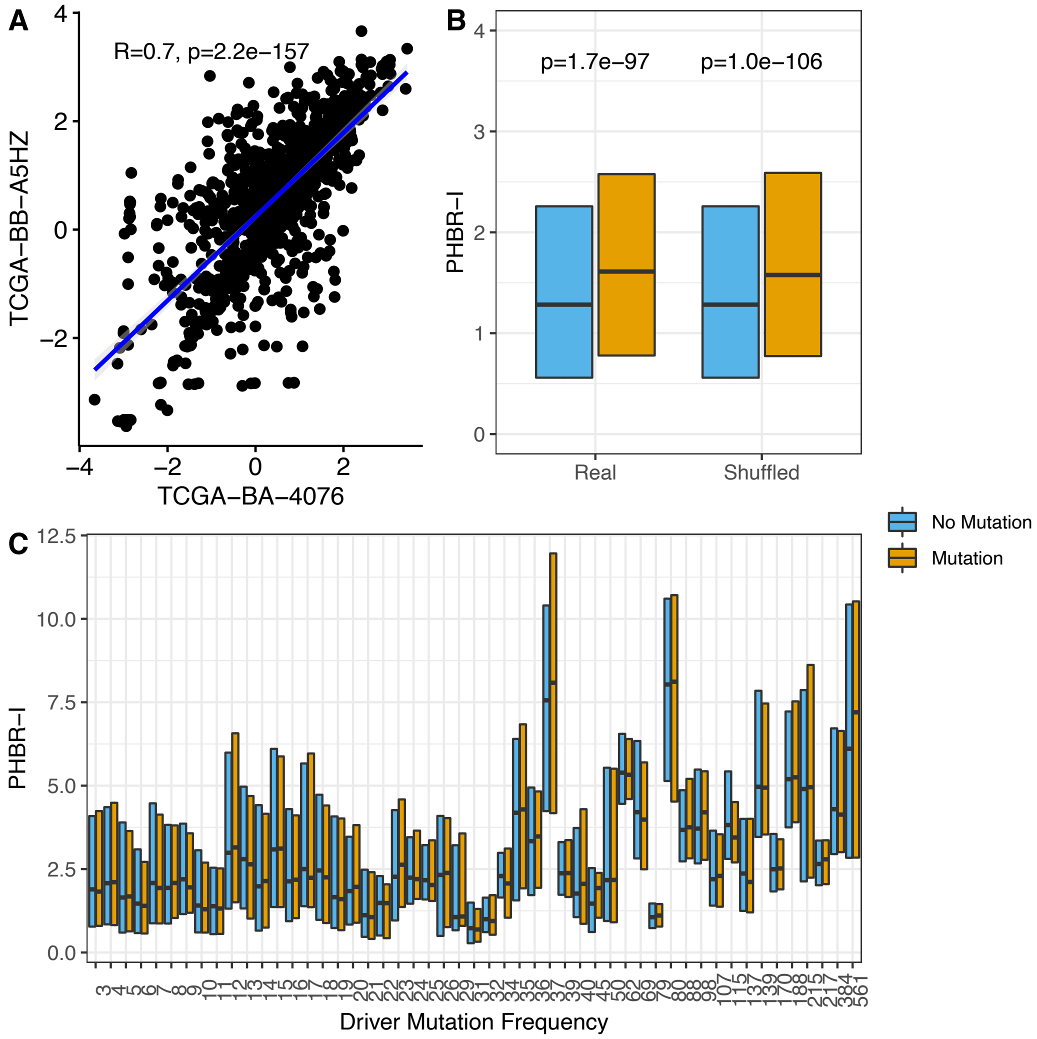


Figure 3.11: **Comparison of patient harmonic best rank (PHBR) scores for driver genes within TCGA data.** (**A**) Scatterplot of log PHBR-I scores of all driver mutations, calculated using the HLA genotypes of two randomly selected patients from TCGA. (**B**) Median and interquartile range of PHBR-I score in the No Mutation (blue) and Mutation (orange) groups for the real data and for data in which the MHC genotypes have been randomised between patients. (**C**) Median and interquartile range of PHBR-I scores in the No Mutation (blue) and Mutation (orange) groups in bins of mutation recurrence.

the inability of the patient's antigen to present the mutation to the immune system. When we restricted the analysis to include only the most frequent mutation for a driver gene, we saw no difference between the mutation and no mutation groups (P=0.28).

3.4 Discussion

Marty *et al.* hypothesised that the driver mutations observed in a cancer occurred in the gaps in a patient's ability to present that particular mutation to the immune system, which is controlled by the patient's MHC genotype [205, 206]. However, they reported that this phenomenon is only seen for driver mutations and not for passenger mutations. If this is the case, then it would appear that the landscape of passenger mutations in the tumour was not shaped by the immune system. However, passenger mutations should be more likely to be removed by the immune system as they offer no benefit to the cancer, and there is no advantage to keeping them. Therefore, we expected the selection effect to be greater for passenger mutations than for drivers. To understand Marty *et al.*'s findings, we investigated whether there were alternative mechanisms for removing these passenger mutations. We hypothesised that observed passenger mutations are in genes that were either lowly expressed or were being downregulated compared to the normal gene. The idea behind this was that the cancer downregulated that gene to avoid detection by the immune system as an alternative method of immune escape. This mechanism would not be advantageous for the cancer in terms of driver mutations as their expression is required for cancer survival and progression.

Our initial analyses assessed whether passenger mutations were tolerated due to their occurrence in lowly expressed genes. Genes harbouring immunogenic mutations appeared to have lower expression than the median expression across samples, supporting our theory that they were preferentially occurring in lowly expressed genes. However, it is difficult to distinguish between differences caused by mutagenic processes themselves and selection by the immune system. Many factors can affect the presence of a mutation. Mutation rates differ at specific sequence contexts, with higher mutation rates in GC-rich regions [48, 457–461]. The length of the gene can also affect the mutation rate. Longer genes can accumulate more mutations because there are more positions at which a mutation can occur. However, even accounting for gene length, genes that encode larger proteins, such as TTN, are enriched with mutations [462]. This is due to differences in mutation rates related to the transcription of genes as well as differences in replication timing. Larger genes tend to have lower expression and to be late replicating, both of which were linked to increased mutation [462]. To account for both of these factors, we randomised mutations within a gene with the same mutational context as observed mutations and saw no statistically significant difference in the proportion of missense mutations within genes. Thus indicating that the difference between the expression of genes with missense mutations and synonymous mutations was due to factors other than immunoediting.

We were also able to determine an upper bound on the percentage of missense mutations that were required to be removed by the immune system for a difference to be observed. We found that if at least 5% of missense mutations were removed due to the immune response, then we would expect to see evidence of purifying selection within the dataset. Our results are consistent with no or low contribution of the immune system on removing passenger mutations. However, it may be that the effect of immunoediting is subtle and that its effect is not large enough to be detectable in this data. It could be that only 5% or less of missense mutations are capable of being recognised and eliminated by the immune system.

The idea that genes harbouring immunogenic passenger mutations were downregulated or that the mutant allele is downregulated compared to the normal allele to evade the immune system was not supported by our results. Rosenthal *et al.*[264] reported that only 33% of clonal neoantigens were present in expressed genes when analysing non-small-cell lung cancer. However, we found no evidence of a difference in the expression of missense, and more specifically, immunogenic missense mutations being present in lowly expressed genes as a consequence of immunoediting. As shown in our analyses, any difference in expression of genes with immunogenic mutations were likely due to differences in sequence context for the genes, which was not accounted for by Rosenthal *et al.*

The lack of evidence for immunoediting of passenger mutations and the finding that individual-specific driver mutation coverage inferred from PHBR scores shows no association with cancer risk led us to re-evaluate the results from Marty *et al.* [205] and Marty-Pyke *et al.*[206] . We discovered that the relationship between HLA genotypes and the oncogenic landscape of driver mutations reported in these papers is due to misinterpretation and unjustified statistical assumptions. The results of this paper were also questioned by Claeys *et al.* [208], who noted that the immunogenic selection signals were due to oncogenic mechanisms that lower binding affinities of 13 common driver mutations in 6 different genes and not due to HLA binding affinity.

Our results show that the relationship between MHC genotype and missense mutations present in an individual is independent of PHBR scores, as reported by Marty *et al.* However, it is possible that the MHC genotype does play a role in shaping the mutational landscape in an individual but that the PHBR score does not capture the effect. This seems unlikely due to the experimental evidence for the capacity of PHBR scores to predict the binding affinity of neoantigens [205, 206]. However, it is likely that binding affinity alone is not sufficient to determine the immunogenicity of a neoantigen. As well as presenting the neoantigen to the immune system by the MHC activation of cytotoxic T cells is required to destroy the tumour cell. PHBR score does not consider the cytolytic activity of the tumour or the similarity of neoantigens to self antigens. Using HLA-binding affinity to classify mutations as immunogenic is routinely used. However, a study by a global consortium which assessed predicted epitope immunogenicity found that only 6% of predictions were actually immunogenic [463], consistent with a previous report [464]. This indicates that our results may have been impacted by noise introduced due to the method we used for predicting immunogenicity and that the level of immunogenic mutations capable of eliciting an immune response is lower than observed in our dataset. As our expression analyses depended on the correct classification of immunogenic mutations, this could have confounded our results. Using tools, such as DeepNeo [465, 466], which consider T -cell reactivity as well as MHC binding affinity predict immunogenicity may be necessary to observe a signal.

A limitation of our analysis when identifying immune-escaped samples is that we used a simple classification method to identify them. By solely using the presence of a clonal nonsynonymous mutation in one of 88 genes involved in antigen presentation and immune escape, we may be misclassifying some samples as immune escaped that were not. Samples that have a mutation in one of these genes tend to have many mutations. This may be because the mutations can no longer be recognised and eliminated by the immune system. However, it is also likely that these samples have a mutation in one of the 88 genes because it is a highly mutated sample and therefore has a higher chance of having a mutation in these genes. We could also have excluded some samples that have evaded the immune system due to the strict thresholds we applied for calling clonal variants. A more accurate method to classify these samples is required to elucidate this.

3.5 Conclusion

In conclusion, we assessed the theory that passenger immunogenic mutations are preferentially occurring on lowly expressed genes or are exhibiting ASE with down-regulation of the mutant allele as a means of immune escape. Our results show that the difference in the expression of genes harbouring immunogenic mutations compared to non-immunogenic and synonymous mutations is likely due to differences in sequencing context rather than the effect of immunoediting. We also estimated an upper bound of 5% for the level of immunoediting that could be acting on the tumour, at a level that is undetectable in our data. This would suggest that immunoediting does not play a role in shaping the passenger mutation landscape of tumors or that if it does have a role it is too low to detect.

3.6 Materials and Methods

3.6.1 Clonal Variants

Variants classified as clonal from the analysis in Chapter 2 were used in this analysis. Specifically, these mutations reached 100X total depth and had an upper confidence interval for the cancer cell fraction (CCF) >0.8 and a lower confidence interval of >0.7. Driver mutations were removed from the analysis. Oncogene [467] and tumour suppressor gene [468] information was downloaded on 13 January 2020 from their respective websites. A list of 56 weak driver genes was obtained from [469]. We also included the pan-cancer TCGA driver genes from Oncovar [470]. This gave a list of 2,020 driver genes.

3.6.2 Identification of immunogenic mutations

The immunogenicity of mutations was determined using PHBR scores provided by Noor Kherreh. The scores were calculated by considering all peptides of a specific length or range of lengths that contain the mutation. First, a rank-based presentation score was obtained for each peptide using NetMHCpan3.0 [471], and for each of the patient's HLA alleles, the best rank value was retained. The PHBR score is then the harmonic mean (across the patient's HLA alleles) of these best-rank scores (see Marty *et al.* [205] and Marty-Pyke *et al.* [206] for details). This score was calculated for class I MHC alleles in [205], based on peptides with lengths ranging from 8 to 11 amino acids and for class II alleles in[206], where it was based on peptides of length 15 amino acids.

3.6.3 Expression Analysis

HTSeq Count files for each primary tumour sample were merged into one matrix for each cancer type using a custom R script. Genes were filtered using the filterByExpr function in edgeR (v3.28.1) [472] with the default settings so that at least 70% of samples had a minimum count of 10. Genes were then normalised to transcripts per million (TPM) for comparing gene expression within samples in a data set or to the weighted trimmed mean of M-values (TMM) when comparing the expression of a gene across multiple samples. We used log2 TPM and TMM values in our analysis. Boxplots were produced using ggplot2 (v3.3.0) [473], and pairwise comparison tests were performed using the Wilcoxon test. Fisher's exact test was performed to test the significance of contingency tables. For all analyses that split mutations into immunogenic, nonimmunogenic, and synonymous mutations, only samples for which HLA data were available were included. For analyses looking at nonsynonymous and synonymous mutations, all samples were included. Oncogenic genes were removed from the study.

3.6.4 Randomisation of mutations in the same mutational context as observed mutations

Mutation contexts for all trinucleotides within the coding region of the GRCh38 genome were provided by Noor Kherreh. These regions were annotated with gene information by finding the overlapping regions with gene coordinates using the GenomicRanges (v1.46.1) [474] R package. Trinucleotide contexts of all clonal passenger mutations were annotated using the mutSignatures (v2.1.4) [475] R package. Random mutations within genes containing trinucleotide contexts of the same type as the observed mutations within that gene were assigned for each mutation. If the possible mutations of that context for a gene were the same as the observed mutations, it was removed. Mutations were annotated with Variant Effect Predictor online tool [476] against GRCh38 using the option to show one selected consequence per variant.

3.6.5 Allele-Specific Expression Analysis

Pileup files from RNA-Seq genomic realigned bam files downloaded from GDC were generated using samtools v1.9 [477] mpileup with the -C50 and –B options. We limited the output to the list of clonal mutations for each sample. This was then used as input to cisASE v 1.0.2 [330], run in RNA-only mode. cisASE was run using SNV and gene level detection. The annotation file used for gene level detection was generated with Ensembl Biomart [478] for human GRCh38. The minimum required depth was set to 10. The log-likelihood ratio (LLR) thresholds for significance at 0.05 level for each sample were obtained from cisASE output. These thresholds were computed by simulating the null distribution 2,000 times using the input data. Genes or SNVs with an LLR value greater than the cutoff value for that specific sample were deemed to show ASE. We only included mutations within the autosomes for ASE analysis.

3.6.6 Immune Escaped Samples

To assign samples as immune escaped or immune non-escaped, we followed the method of Gourmet *et al.* [479] and characterised samples as immune escaped and non-immune escaped based on the presence of nonsynonymous clonal mutations in at least one of 88 genes known to be involved in antigen presentation and immune escape; B2M, CALR, CANX, CD4, CD74, CD8A, CD8B, CIITA, CREB1, CTSB, CTSL, CTSS, ERAP1, ERAP2, FAS, HLA-A, HLA-B, HLA-C, HLA-DMB, HLA-DMA, HLA-DOA, HLA-DOB, HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DRA, HLA-DRB1, HLA-DRB3, HLA-DRB4, HLA-DRB5, HLA-E, HLA-F, HLA-G, HSPA1A, HSP90AA1, HSP90AB1, HSPA1B, HSPA1L, HSPA2, HSPA4, HSPA5, HSPA6, HSPA8, HSPBP1, IFI30, IFNG, IRF1, KIR2DL1, KIR2DL2, KIR2DL3, KIR2DL4, KIR2DS1, KIR2DS2, KIR2DS4, KIR2DS5, KIR3DL1, KIR3DL2, KIR3DL3, KLRC1, KLRC2, KLRC3, KLRC4, KLRD1, LGMN, MEX3B, NFYA, NFYB, NFYC, PDIA3, PSMA7, PSMB10, PSMB11, PSMB6, PSMB8, PSMB9, PSME1, PSME2, PSME3, PSMF1, RFX5, RFXANK, RFXAP, TAP1, TAP2, TAPBP, TNF

4 Chapter 4: Predicting allele specific expression in tumor suppressor genes and analysing its relationship with breast cancer risk

4.1 Abstract

Allele-specific expression (ASE) is the process of expressing one allele at a different level to a second allele at a specific location in a diploid organism. ASE of tumor suppressor genes (TSGs) occurs in both normal and cancer cells and ASE of specific TSGs has previously been shown to be associated with cancer risk. Here, we assess the feasibility of predicting ASE in TSGs using data from The Genotype-Tissue Expression project, to generate a single ASE score for TSGs and test its association with cancer risk. We assessed the ability of two methods, an adapted version of the gene expression prediction tool PrediXcan and application of logistic regression models, using genotype data to predict ASE. While both methods showed it is possible to predict ASE to some extent using this data, the overall performance was poor, with each method having its own limitations. As a pilot study, we applied the prediction methods to UK Biobank data and generated a single TSG ASE score for each sample and assessed its association with breast cancer risk. Although we found no statistically significant association between TSG ASE and breast cancer risk, the lack of association may be due to our inability to predict ASE in TSGs that contribute to cancer risk in this tissue type. We could not predict ASE for the TSGs commonly containing driver mutations in breast cancer, as assessed using data from The Cancer Genome Project. It is likely that as more data becomes available, it will become possible to improve the prediction of ASE and generate a more accurate score to be used to test for an association with cancer risk.

4.2 Introduction

Allele-specific expression (ASE) is the process of expressing one allele at a specific locus in a diploid organism at a different level than the second allele. ASE can be caused by genetic variation that alters the expression of alleles. Mutations can cause a difference in the ability of a transcription factor to bind to a gene to initiate transcription [480], can result in nonsense-mediated decay of one version of the gene [481] or can disrupt splicing sites resulting in aberrant gene expression [482, 483]. ASE has a high prevalence in normal tissues, with a recent report using data from 838 individuals across 49 tissue types by the Genotype-Tissue Expression (GTEx) project reporting that at least 53% of protein coding genes show evidence of ASE in at least 50 individuals [339]. However, this imbalance varied across tissues within individuals and between individuals.

Tumour suppressor genes (TSGs) play an essential role in preventing cancer. downregulation or disruption of these genes is frequently required in order for cancer to progress. Inherited mutations in these genes can increase the risk of an individual acquiring cancer [484]. This is because it only takes one driver mutation that disrupts the normal functioning version of the gene for cancer to develop. ASE in TSGs has been shown to occur in normal tissues as well as in cancer samples [326]. However, the proportion of loci showing ASE is much higher in some cancer tissues, such as breast, head and neck, lung and thyroid, compared to their corresponding normal tissues. The majority of genes showing ASE in tumor tissues were due to somatic events such as copy number alterations or exon-skipping within the tumour tissues [322, 326]. ASE caused by germline mutations has been indicated as a risk factor for cancer in several tissue types such as melanoma [485], prostate [370, 486–488], colorectal [353, 361, 362, 489–491], pancreas [492], lung [493], breast [356, 359, 494, 495], kidney [365]and blood[496, 497]. However, these studies focused on particular ASE SNPs in one or more TSGs rather than investigating the level of ASE across the whole class of TSGs. Here, we propose to generate a single TSG ASE score to assess the level of ASE in tumour suppressor genes within an individual and to investigate its relationship to an individual's risk of developing cancer.

ASE is normally measured by comparing the number of reads for each allele or haplotype using RNA sequencing data to the ratio of the reads using DNA sequencing data from an individual [498]. This allows us to account for any biases that may be introduced due to the sequencing or mapping process. However, RNA and DNA sequencing data are not always available for individual samples due to the large costs associated, limited amounts of specimen taken from biopsies as well as storage capabilities for the large amounts of data generated. Another issue is that ASE analysis requires more extensive coverage than is generated in normal RNA-sequencing experiments [499]. While models such as cisASE [330] have been developed, which allow us to measure ASE data using RNA-sequencing data only, the same is not true when only DNA sequencing data is available. A previous study [500] assessed the feasibility of predicting ASE at the SNP level, using genomic features such as GerpN, a score of neutral evolution, to predict if a specific SNP shows an imbalance. The goal of their study was to use predicted SNP imbalance to prioritise candidate variants in clinical diagnostics when RNA-Sequencing data is not available. However, this requires large amounts of gene annotation and downstream processing to obtain the relevant data to develop the predictions. The results of the predictions rely on the correct annotation of the variants. Additionally, this method generates predictions at the SNP level only, which may not correspond to the gene level imbalance in which we are interested.

Here we assess the ability of phased genotype data to predict ASE. We then use these predictions to generate an ASE score which can be used to measure the level of ASE of a target gene set within a single sample. Generating scores for specific classes of genes can be used to assess potential risk for diseases such as cancer. We are specifically interested in predicting ASE in TSGs as we hypothesise that ASE in this class of genes may be associated with increased cancer risk. As a pilot study, we test the association between predicted TSG ASE and breast cancer risk using data from the UK Biobank. We also investigate the presence of predicted TSG ASE in breast cancer samples from the pan-cancer analysis of whole genome (PCAWG) to further assess the prevalence of ASE in TSGs in the normal tissues of individuals with cancer.

4.3 Results

4.3.1 Prevalence of ASE in tumour suppressor genes for GTEx samples

The most comprehensive database of allelic expression to date comes from Castel *et al.* [339] who generated SNP and haplotype level ASE results from 54 human tissues of the GTEx project . We refer to this data as the phASER results going forward, based on the method used to generate the data. We assessed the prevalence of ASE in TSGs within the GTEx cohort (Figure 4.1), and found varying numbers of genes exhibiting ASE at the gene level within different tissue types. All samples showed ASE in at least one TSG gene. Of the possible 1018 protein-coding TSGs, 837 showed ASE in at least one sample in this dataset (822 genes when imprinted genes were removed). The three artery tissue types had the highest number of TSGs with ASE, with a median value of 7 across samples within that tissue type, while all brain tissue types clustered at the lower end of the spectrum, with the majority having ASE in fewer than five genes. It is reassuring to see that tissue types of the same organ tended to cluster together and that the number of TSGs per sample type remains relatively consistent within tissue type, suggesting that the result is not artefactual and is robust.

There were a number of samples that had a higher number of TSGs showing ASE than the majority (Table 4.1). It is interesting to note that the majority of these samples are males. This observation is not due to imprinting, as we removed imprinted genes prior to analysis because it is impossible to predict ASE in these genes using genotype data. However, the median per sex per tissue type was the same ± 1 (Figure 5.7 in Appendix C), indicating that, in general, there are no sex biases in terms of the number of TSGs per tissue type. There was no obvious pattern in terms of tissue type, age, or cause of death, with a variety of each category observed. One sample was present twice (GTEX-QMR6), while the rest of the samples were unique to one observation. Curiously, the cause of death for this sample was an illness such as heart disease or cancer (no specific details of illness was provided by GTEx).

4.3.2 ASE Prediction in GTEx

We first assessed the feasibility of using a tool called PrediXcan [379] to predict allele-specific expression in the GTEx cohort. PrediXcan was developed to predict gene expression using phased genotype data as input. By assessing the genotype of specific model SNPs for an individual, it predicts a measure for each gene for a particular tissue by applying multivariate adaptive shrinkage in R

Subject ID	**Total TSGs with ASE**	**Tissue**	**Age**	**Sex**	**Cause of Death**
GTEX-QMR6	109	Lung	50-59	Male	4
GTEX-Q2AH	96	Skin Sun Exposed Lower leg	40-49	Male	0
GTEX-QMR6	58	Brain Hippocampus	50-59	Male	4
GTEX-1497J	49	Skin Sun Exposed Lower leg	60-69	Male	0

GTEX-POYW	45	Lung	60-69	Male	0
GTEX-VJYA	42	Muscle Skeletal	60-69	Male	0
GTEX-WWYW	41	Pituitary	50-59	Female	3
GTEX-QDVN	41	Skin Sun Exposed Lower leg	50-59	Male	0
GTEX-V955	41	Breast Mammary Tissue	60-69	Male	0
GTEX-1RB15	37	Adipose Visceral Omentum	60-69	Male	0
GTEX-VJYA	36	Skin Not Sun Exposed Suprapubic	60-69	Male	0
GTEX-OIZI	33	Whole Blood	40-49	Male	0
GTEX-1JMQI	33	Adipose Subcutaneous	50-59	Male	0
GTEX-VUSG	32	Whole Blood	50-59	Male	0
GTEX-QDVN	31	Adipose Subcutaneous	50-59	Male	0

Table 4.1: **amples that have the highest number of genes showing alleles specific expression for a particular tissue type.**

(MASHR) based expression models. We generated predictions for all genes for each tissue type in the GTEx cohort for which models were available (n=49) using PrediXcan for each haplotype, as explained in Methods. We then compared the z-score ratio of haplotype A and haplotype B for each gene to the allelic fold change (aFC) score generated with the phASER results for each individual sample (Figure 4.2). While we do see a positive correlation between the aFC and z-score ratio, the correlation is not strong, with the highest Pearson r of 0.31 observed for any sample in any tissue type. This equates to an r^2 of 0.09, meaning that only 9% of the variation of the observed aFC is explained by the predicted score.

However, when we calculate gene level correlation (Figure 4.3) instead of sample correlation, we see that the PrediXcan method performs well at predicting ASE for some genes. For some genes, there is a strong positive correlation; for others, there is a strong negative correlation; and for others, there is no correlation. Therefore, it may be genes for which there is no correlation that is causing the poor performance in sample-level correlations. To check whether the difference in performance for individual genes was due to the expression level of genes, with genes that had no expression or low expression in that tissue type having poor correlation, we removed genes that had a TPM expression value less than 10 for that tissue (Figure 5.8 in Appendix C). We also removed genes that had no observed ASE in the phASER results because there would be no variability in the aFC values (Figure 5.9 in Appendix C). Although the distributions are flatter, we see no difference in the range of scores, indicating that this is not the reason for the variability in performance.

We next looked to see if converting the ASE score to a binary classifier of ASE and no ASE would improve results. We applied different combinations of

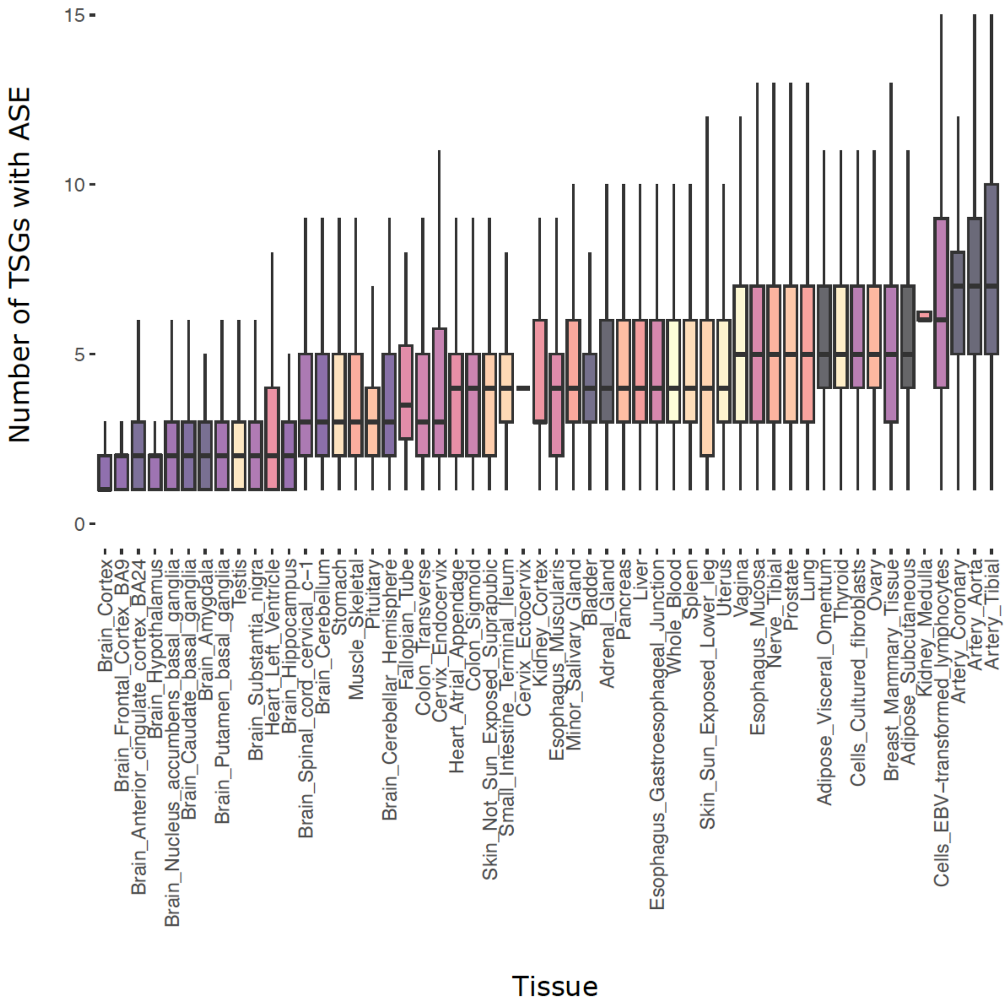


Figure 4.1: **Distribution of the number of samples showing allele specific expression in tumor suppressor genes in normal tissues.** Boxplot of number of TSGs showing ASE, classified using a binomial threshold, per sample split into Tissue Type. Outliers (Number of TSGs >15) were removed from plot.

thresholds to classify ASE status and calculated the area under the curve (AUC) to assess the performance within each tissue type. The highest AUC we achieve for any tissue type is 0.813 (Figure 4.4A), with the highest results observed in testis. The lowest AUC score achieved is 0.491 in brain amygdala (Figure 4.4B), performing no better than by chance. We observe the poorest performance when a high threshold (6) is applied to phASER results, and either low (0.5) or high (>9) is applied to the predicted values (Table 5.2 in Appendix C). The top AUC scores within individual tissue types (Figures 5.10-5.13 in Appendix C) tended to occur with phASER thresholds of 3 or 4 and predicted thresholds ranging between 1 and 3. The AUC results suggest that as we increase the predicted threshold value, performance worsens, but as the phASER threshold increases, performance improves. This might indicate that the predictions work better for extreme ASE. Investigating the results in more detail, we see that specificity tends to be poor for the predictions. This means that although the predictions called a large number of true ASE genes, it also misclassified a large number. Also, when we looked

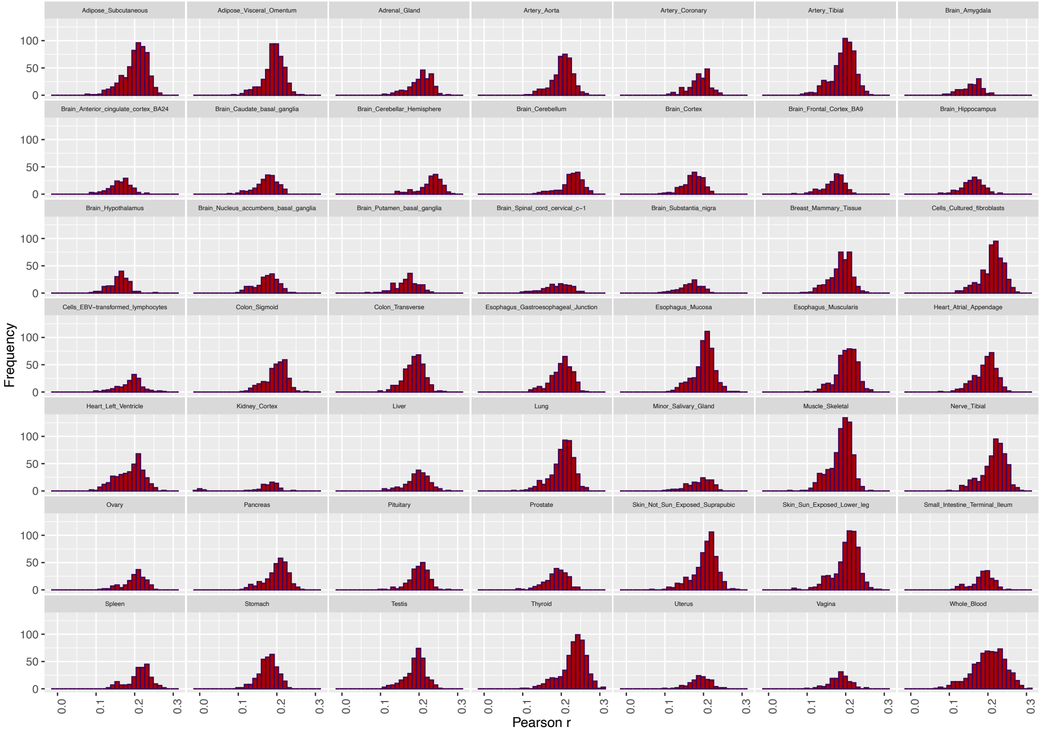


Figure 4.2: **Comparison of predicted allele specific expression using Predixcan versus allelic fold change from phASER results for all genes within a sample**. Histogram of Pearson r values for comparing PrediXcan z-score ratios to aFC for all genes within a sample. Plots are split by tissue type.

at the number of ASE genes in the phASER results compared to the PrediXcan results we saw that the number of samples that had ASE for a particular gene in the phASER results was small while the number in the PrediXcan results was large. This would indicate that the good performance values observed are likely because the PrediXcan results called a large number of samples as having ASE for particular genes, which happens to overlap with the majority of phASER ASE calls. We also applied a significance level threshold to phASER aFC, with genes showing ASE also required to have an adjusted p-value of <0.05 based on a binomial test assessing the difference in counts between the two haplotypes. If they did not reach this significance level and meet the aFC threshold, the gene was classified as not showing ASE. This did not improve results, with the AUC values ranging from 0.49 to 0.81 with the best performance still achieved in testis. Due to the poor performance of binary classification of ASE with PrediXcan results, we use the z-score ratio going forward in our analysis.

4.3.3 Logistic regression models models applied to individual genes in GTEx breast samples

Due to the poor overall performance of the PrediXcan models to predict ASE in the GTEx data, we next sought to apply logistic regression models using ASE status for each gene as the response variable and SNP genotypes within 100 kilobases of the gene region classified as heterozygous or homozygous as the predictor variables. We focused our analysis on one tissue type, breast tissue, so that we would not have single individuals with multiple tissue types contributing many data points to

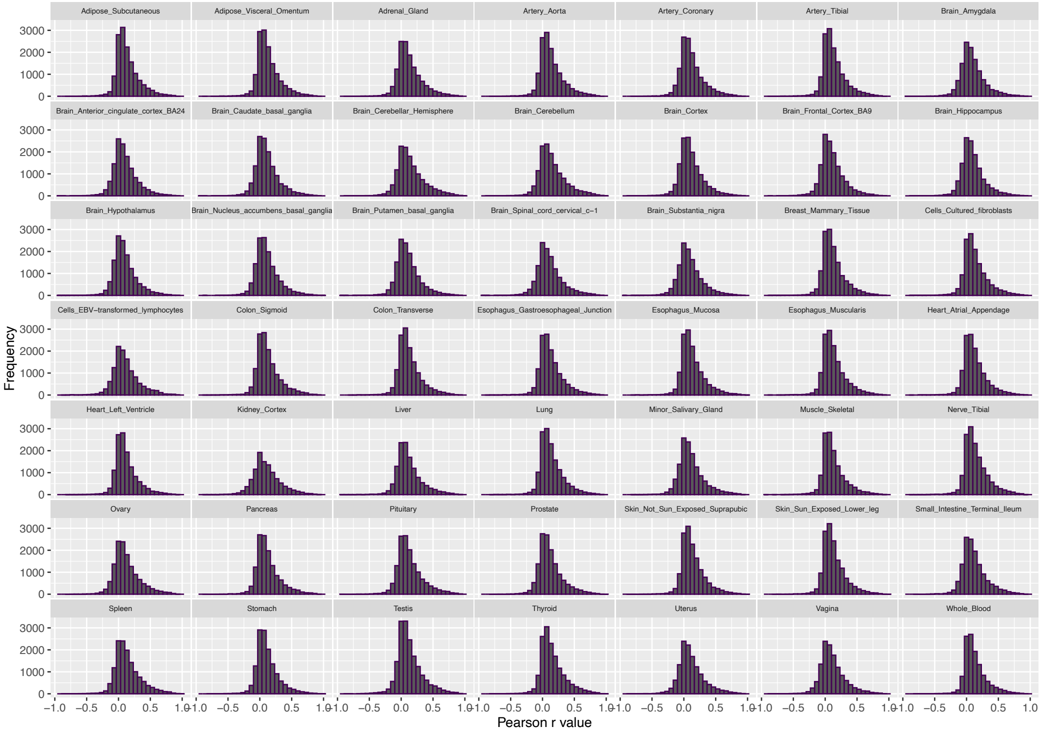


Figure 4.3: **Comparison of predicted allele specific expression using Predixcan versus allelic fold change from phASER results for each gene across all samples.** Histogram of Pearson r values for comparing PrediXcan z-score ratios to aFC for all samples within a tissue type for a particular gene. Plots are split by tissue type.

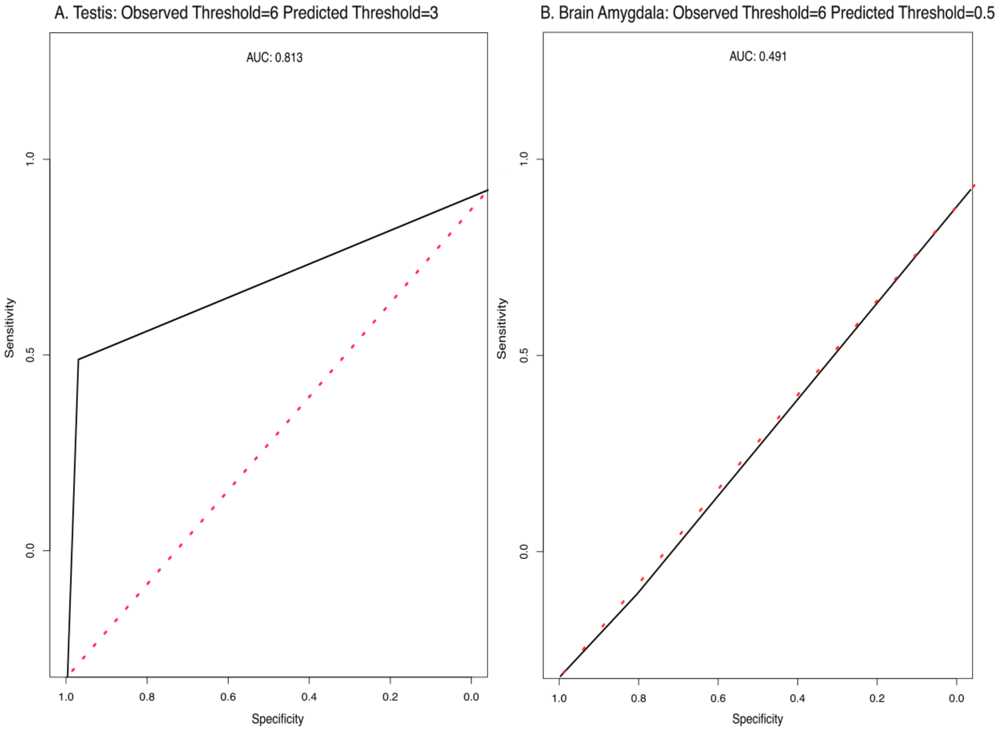


Figure 4.4: **Example Receiver Operator Curves when predicting a binary outcome of allele specific expression using PrediXcan.** (**A**) Best and (**B**) Worst binary classification comparisons. The red dashed line corresponds to the performance of a random classifier while the black line corresponds to the model performance. Area under the curve (AUC) are given for both.

the models. This could inflate the relationship of a specific SNP to the response variable due to the same individual's genotype contributing many times. Using SNPs that were present in both GTEx and UKB datasets and were present within 100 kilobases of the gene of interest, we generated models following the workflow shown in Figure 4.9.

Using a binary response variable of ASE or no ASE with an aFC of at least ±1 and binomial p-value <0.05 required to deem a value as ASE, we were able to analyse 2580 genes that exhibited ASE. We were only able to generate models for 472 of these genes. Of the genes that did not have a model, 1625 failed because there were insufficient samples with ASE numbers to balance the classes in the training set, 157 failed due to the information value filter step, 314 failed because there were no SNPs available in the dataset for the genes and 12 failed due to too few heterozygote SNPs or cross-validation error. Of the 472 genes which we could model, the prediction performance, as measured by AUC using the test dataset, varied between 0.2532 and 1. The median AUC achieved was 0.6538, meaning that at least half of the models had poor performance (AUC <0.7) when we used a probability score of 0.5 as the cutoff to classify predictions as ASE and no ASE. Most models (434) use only 1 SNP for predictions and there was a slight negative correlation between AUC and the number of heterozygous SNPs available for the gene (Pearson's r= -0.103, p-value = 0.026; Figure 4.5). It appears that genes with a small number of samples exhibiting ASE is accounting for the low and high AUC values, with genes that had a large number of samples with ASE tending to have AUC between 0.5 and 0.9. This is unsurprising since AUC values are more stochastic in these cases. Of the 472 genes, only 26 were TSGs, and their AUC values ranged from 0.3462 to 0.9177 with a median of 0.6175. There were 12 TSGs which had good performance (AUC >0.7).

Applying a binomial threshold to classify genes as ASE or not ASE may be too stringent. Therefore, we applied logistic regression models to the data using only aFC cut-off to classify variants. By doing this we were able to build models for 7850 genes. Of those that failed 3561 were due to IV filter step, 2456 were because there were no SNPs available in the dataset for the genes, 254 failed because there were insufficient samples with ASE numbers to balance the classes in the training set, 56 failed due to too few heterozygote SNPs and 34 due to cross-validation error. Of the 7850 genes which we could model, the prediction performance, as measured by AUC using the test dataset, varied between 0.2532 and 1. Again the majority of genes (7407) only had 1 SNP in the model. In this case, there was no significant correlation between AUC and the number of SNPs used in the model (Pearson's r= 0.004,p-value=0.7465), which was unsurprising due to the high number of genes that only had 1 SNP in the model. There was a significant correlation between the performance of the models and the number of samples showing ASE (Pearson r= -0.1, p-value= 2.7×10^{-17}). This slight negative correlation was surprising as we would expect that predictive ability would be worse when fewer samples are exhibiting ASE but it is likely due to the small number of genes that have a large number of samples exhibiting ASE. The median AUC achieved was 0.5586, meaning that at least half of the models could not classify genes as ASE when we used a probability of 0.5 as the cut-off to classify predictions as ASE and no ASE. It is likely that by removing the binomial threshold filter, we are classifying samples as ASE for a particular gene that may not be true ASE, resulting in models that cannot accurately predict ASE. Of the 7850 genes, there were 483 TSGs, and their

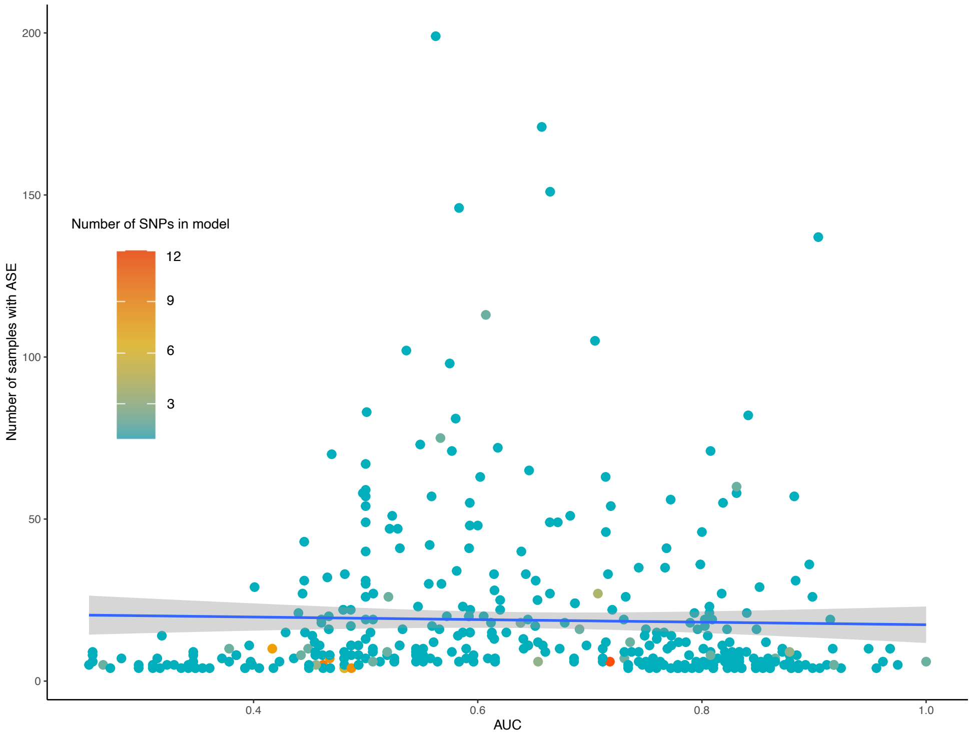


Figure 4.5: **Relationship between area under the curve (AUC) performance value for each gene model when a binomial threshold was applied compared to the number of samples that have ASE for that gene in the phASER results.** Data points are coloured by the number of SNPs used in the model.

AUC ranged from 0.2785 to 0.9304 with a median of 0.5560. There were 76 TSGs which had good performance (AUC>0.7).

4.3.4 Comparison of PrediXcan results with binary logistic regression models

We compared the results from both methods for predicting ASE in order to assess which performed better and should be used for downstream analysis. We compared the performance of predictions for the TSGs which achieved AUC >0.7 when we generated logistic regression models to the z-score ratio for the same genes generated from PrediXcan breast tissue results. We generated a single ASE score per sample using each prediction method and compared the results to a score generated using phASER aFC. Instead of applying a probability cut-off to classify genes as predicted ASE or not ASE using the logistic regression model results, we used the probability scores themselves to generate the score. We added the probabilities for all TSGs into one score per sample. By doing this, a gene which had a high probability of being ASE contributed more to the score than a gene that had a low probability, so a sample that had a large number of genes showing ASE would have a larger score than a sample with few or no genes showing ASE. This would also negate the need to apply gene-level thresholds, which might improve model performance, to classify predictions. For the PrediXcan results, we generated an ASE score by combining the absolute z-score ratio of predictions for the same genes present in the logistic regression results. We also generated a single score for these genes in the phASER results by adding the absolute aFC score. By adding the scores, genes that had large aFC or z-score ratio, meaning they had more extreme ASE, contributed more than genes that had a low aFC or z-score ratio. Therefore samples that had a low number of TSGs with ASE but had high ASE had a similar score to samples that had a lot of genes with small amounts of ASE but scored higher than samples with low number of TSGs with low ASE. Comparing the predicted scores to the score generated using the observed phASER aFC values, we saw a positive correlation for both methods. Comparing ASE score for the 12 TSGs and 76 TSGs that achieved good predictive performance in Section 4.3.3 (Figure 4.6) with the phASER scores generated using the same genes, we saw positive correlations (Pearson's r = 0.29, p-value= 6.4×10^{-9} and r = 0.21, p-value = 3.1×10^{-5} , respectively). When we assessed the PrediXcan score with the phASER score for the same genes we also got a positive correlation (Pearson's r = 0.2 p-value = 6.9×10^{-5} and r = 0.08, p-value = 0.11, respectively) but it was not as strong as from the logistic regression models (Figure 4.6).

4.3.5 Pilot Study: Predicting Breast Tissue TSG ASE in UK Biobank

We next generated PrediXcan results for breast tissue in 21,036 UKB white British females, half of whom had self-reported breast cancer, with the remaining samples as matched controls, as explained in Methods. We limited our analysis to white British to limit the impact of population structure on our results. The input data we used for this came from phased haplotype data instead of phased whole genome data, and as a result, only a subset of the SNPs used by the MASHR models were available to generate the predictions. There were 11% of the overall SNPs available for the breast tissue MASHR models. Of the 14,654 genes present in this dataset, only 848 had the full model SNPs available for generating the predictions. We

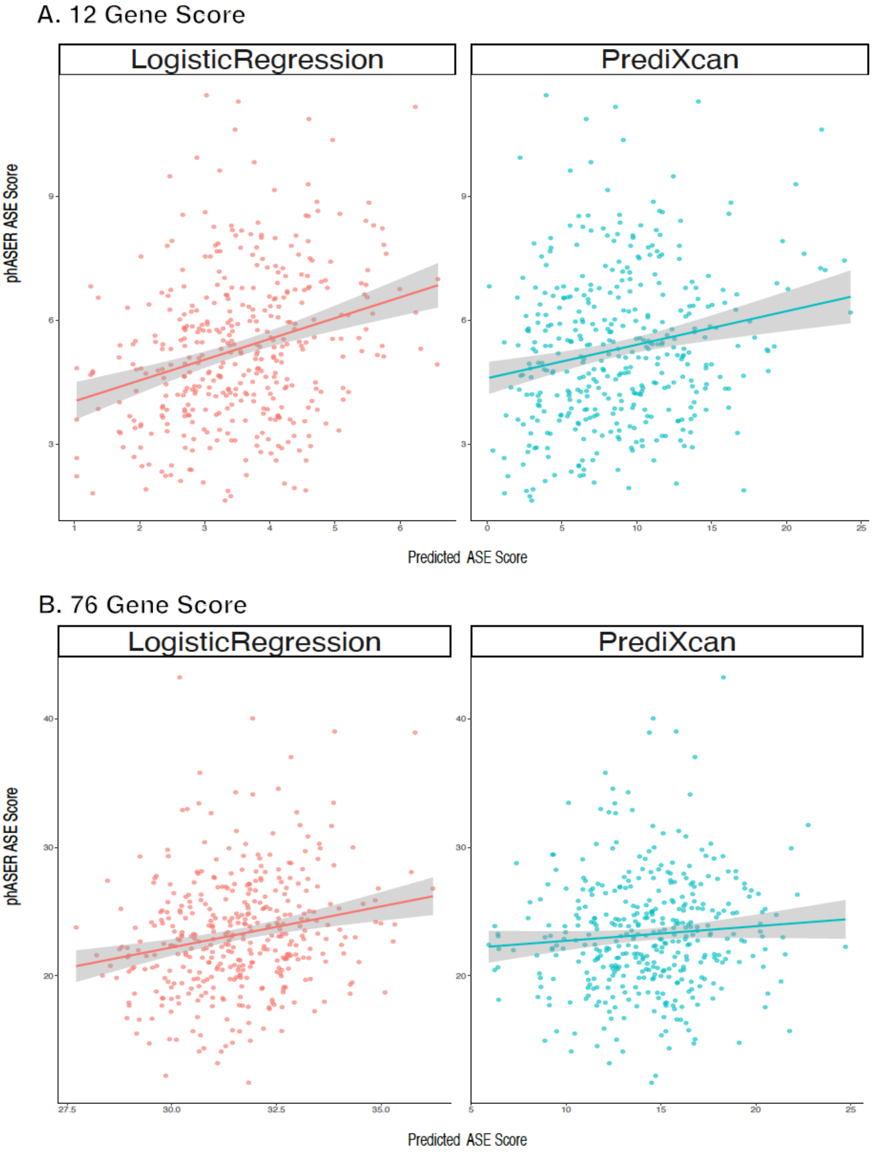


Figure 4.6: **Tumor suppressor gene (TSG) allele specific expression (ASE) scores generated from phASER allelic fold change (aFC) compared to scores generated from predictions.** Scatterplot of TSG ASE scores generated using the observed phASER aFC results against scores generated using logistic regression predictions (blue) or PrediXcan (red). ASE scores for all three types of data were generated using the (**A**) 12 or (**B**) 76 TSGs for which we could generate logistic regression models (**A** included a binomial threshold filter for classifying ASE and **B** did not). Each data point is a score for an individual breast tissue sample.

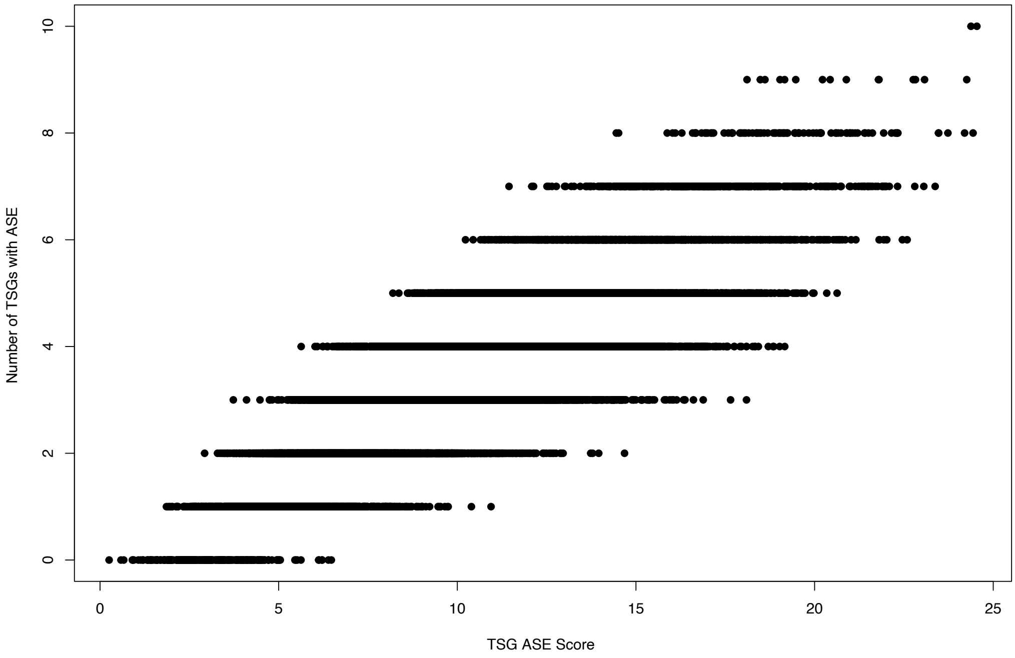


Figure 4.7: **Relationship between allele specific expression (ASE) score generated using phASER aFC against total number of tumor supressor genes (TSGs) showing ASE in this dataset.**

could only assess ASE in the genes for which the model could accurately generate predictions. Of these 848 genes, there were only 47 TSGs available.

We generated an ASE score by combining the absolute z-score ratio of predictions for the 47 TSGs, which had all model SNPs available for PrediXcan to provide gene expression predictions. The TSG score allowed us to generate a score that considers the magnitude of ASE. There was a positive correlation between the ASE score and the number of TSGs showing predicted ASE (z-score >1) (Figure 4.7). However, ASE scores tend to vary compared to the number of TSGs showing ASE, with similar ASE scores observed with different numbers of TSGs with ASE. This was because some genes which had a higher z-score ratio contributed more to the ASE score than a gene with a low ASE score and a sample that had a small number of TSGs showing large magnitude of ASE had a similar score to a sample with a large number of TSGs showing a small level of ASE. Therefore, the ASE score was likely a better measure of ASE than just counting the number of genes that exhibited some level of ASE because it captures more information.

We fitted a binary generalised logistic regression model to determine if there was an association between ASE score and cancer risk. Using cancer status (0 = no cancer, 1 = cancer) as our response variable, we found no significant association between PrediXcan ASE Score and cancer risk (Figure 4.8A). We also fitted models using ASE scores generated using genes that achieved good performance after fitting logistic regression models with (Figure 4.8B) and without (Figure 4.8C) binomial thresholds applied as the predictor variables and again found no significant association with cancer risk.

4.3.6 TSGs with driver mutations in TCGA breast cancer samples

We identified TSGs that commonly have driver mutations in breast cancer data in order to prioritise genes that have been shown to play a role in breast cancer development for our analysis. There were 19 TSGs that had driver mutations in TCGA breast cancer data. However, only one of these genes (CASP8) has ASE in the phASER results, and it displayed ASE in just one sample. This was an important observation as it means that TSGs commonly mutated in cancer are absent from ASE results derived from normal tissue. Therefore, there was no data available to generate models for these genes using the GTEx data.

PrediXcan can be used to identify ASE in genes that do not have ASE in the phASER results. However, we cannot assess the accuracy of predictions for these genes. Unfortunately, none of the 19 genes were present in the 47 TSGs for which all SNPs were available for determining predicted ASE in the UKB data. Therefore, the lack of association between predicted ASE score and cancer risk in the UKB dataset might be due to our inability to predict ASE in TSGs important for breast cancer development.

We were able to predict ASE in 80 TCGA breast cancer samples using PCAWG genotype data. The advantage of using this data was that the majority (91%) of SNPs are available for predicting expression using PrediXcan. There were 739 TSGs in the MASHR breast tissue PrediXcan results, with 659 genes having all SNPs available to make predictions. Of the 19 genes that had driver mutations in TCGA breast cancer samples, there were 13 for which we could predict ASE with PrediXcan, and none of them had predicted ASE when we calculated the z-score ratio.

4.4 Discussion

Allele-specific expression is common in tumour suppressor genes within normal tissues. Based on Knudson's two-hit hypothesis model of cancer [501], it seems likely that downregulation of one copy of a TSG could be a risk factor for cancer as it would only take a mutation that disrupts the function of the normal version of the gene to knock out its function completely. Imprinted genes offer us a natural example of this theory. Genomic imprinting is the process by which only one copy of the gene is expressed, dependent on the epigenetic activation of a specific parental chromosome [502]. Aberrant expression and gene copy loss of imprinted TSGs such as MEG3 is common in tumourigenesis [503]. Additionally, specific SNPs that cause allelic imbalance in a tumour suppressor gene have been shown to increase cancer risk in various cancer types[353, 356, 359, 361, 362, 365, 370, 485–497].

The ability to predict gene level ASE has applications for diseases other than cancer. ASE was shown to play a role in the variability in penetrance of certain disease-causing variants [504]. A mystery that had confused scientists, why individuals with the same disease-causing variant had different severity of disease, was explained, in part, by the variants that affect gene expression. The presence of the disease-causing variant combined with differences in gene expression regulation contributed to the modified penetrance of the disease. Therefore, the ability to predict ASE may prove useful when predicting the severity of disease-causing variants for diseases other than cancer such as autism spectrum disorder which has

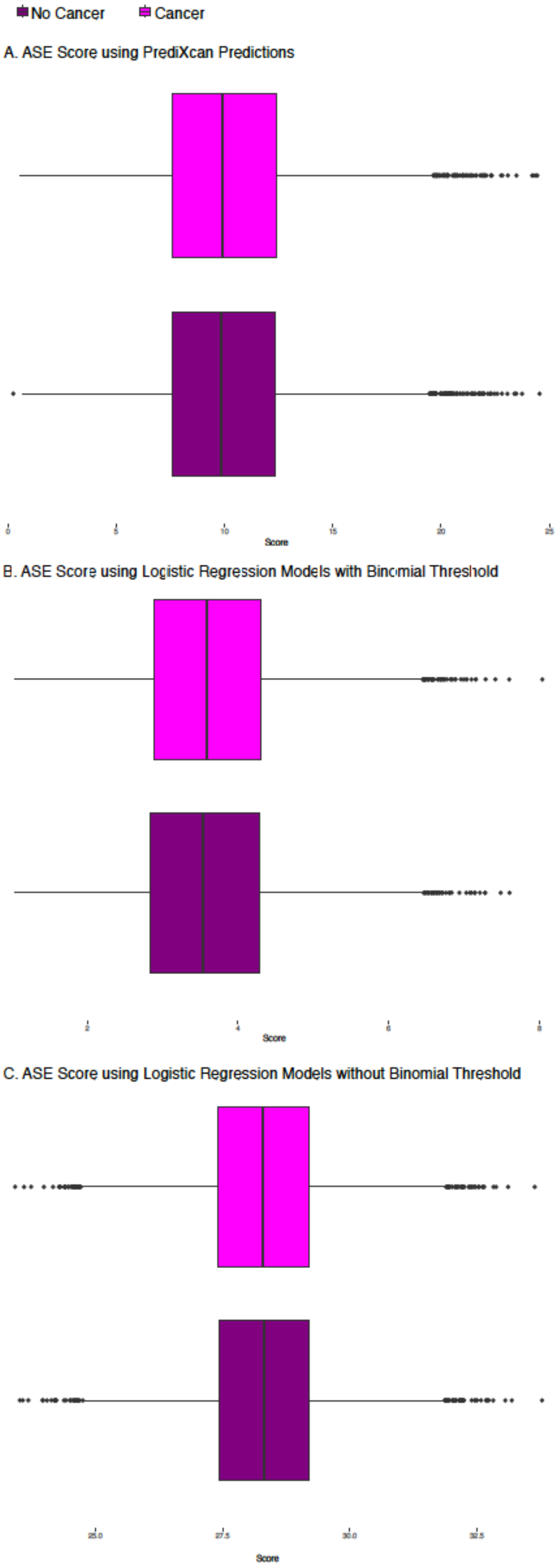


Figure 4.8: **Comparison of predicted tumor supressor gene allele specific expression scores for breast cancer versus non-cancer samples within UK Biobank data.** Boxplots of TSG ASE Scores for UKB samples with and without cancer generated using predictions from PrediXcan (**A**) and from logistic regression models for genes that had AUC >0.7 when the binomial threshold was (**B**) and was not (**C**) applied.

variable penetrance that could be explained by ASE [504].

The initial aim of our analyses was to determine if it is feasible to predict gene level ASE. Although the predictive performance is low using both methods to predict ASE, our results do indicate that it is possible to predict ASE to a certain extent. It is important to note that using the results from Castel *et al.* may not be fully appropriate for what we are trying to do. If we look at the occurrence of TSGs that are genomically imprinted (Table 5.4 in Appendix C), we observe that they are not as consistent as might be expected in the phASER results. While the detection of ASE for these genes may differ between tissue types due to gene expression differences between tissues, we would expect that an imprinted gene would be consistently called ASE within a tissue type. Therefore, there may be additional factors in the data, such as sequencing depth or sample quality, that affect the ability to detect ASE in the truth set. This may account for some of the discrepancies between observed and predicted ASE.

PrediXcan was designed to predict overall gene expression for particular tissue types. The variability in gene level correlation with the phASER results is likely due to the poor performance of PrediXcan for predicting overall gene expression for some genes [505]. If it does not work well for predicting overall gene expression, which it was developed for, in some genes we cannot expect it to perform well at predicting allele specific expression for all genes, as seen in our results. It is also important to note that the PrediXcan models were developed using the GTEx data, and therefore we would expect it to perform best when predicting ASE in this dataset. Migrating the method to other datasets, such as UKB and PCAWG, would result in poorer predictive potential. Without ASE data in these datasets to further assess the predictive performance of PrediXcan, it is difficult to determine the degree of model degradation in these datasets and the accuracy of the ASE scores. It is possible that the lack of relationship between predicted ASE and cancer risk is due to poor predictive performance rather than a lack of association. As PrediXcan models improve so should our ability to predict ASE using PrediXcan.

The ability to predict ASE in the UKB data was restricted due to the availability of appropriate input files to use with PrediXcan. The method requires the use of phased data, and as such, we could only use the haplotype data. There was a small overlap between the SNPs in the haplotype data compared to the GTEx data, with only 10% of the SNPs used in PrediXcan models available. In contrast, the majority of SNPs were available for predicting ASE in the PCAWG, for which we had whole genome phased data available. This meant that our ASE score, which was used to assess cancer risk, was generated using only a small number of TSGs and is likely not representative of the true extent of ASE in these samples, limiting our ability to assess cancer risk.

Applying logistic regression models using the phASER data worked well for a small subset of TSGs. A disadvantage to this method is that we could only generate models for genes that had samples exhibiting some level of ASE in the phASER results for breast tissue data. This limited our analysis to 2580 genes, when we applied a significance threshold (binomial p-value <0.05 to call ASE event) and 14221 genes otherwise. Another disadvantage to the method is that we were limited to SNPs that were present in the UKB to apply our models in downstream analysis. If we had used all SNPs available in the GTEx dataset, we might have been able to generate models for a larger number of genes, and we may have had SNPs that improved the predictive performance, but we would not have been able to apply the models to the UKB data.

Additionally, we are likely unable to predict ASE in the TSGs that are most associated with cancer risk. From our investigation of ASE in TSGs with common driver genes in breast cancer, we found that there is no overlap with our predicted results. Either ASE is not common in these genes, or our method could not predict ASE accurately in the relevant genes. When we applied the PrediXcan method to 80 breast cancer samples from the PCAWG we could not predict ASE in any of the TSGs. The PCAWG dataset had the advantage that we could use whole genome genotype data and were able to generate results for the majority of TSGs. The lack of predicted ASE in these genes in this dataset could indicate that the PrediXcan method is not appropriate for this analysis. However, it could also be the low sample size that resulted in no predicted TSGs.

As previously mentioned, it is possible that the phASER dataset is not appropriate for our analysis. The allelic expression results in this data was derived in normal samples who died of causes other than cancer. Therefore, this dataset may be biased towards individuals who do not have ASE in their TSGs and therefore do not have ASE in the genes required for our analysis. Using normal tissue samples from patients with cancer to derive ASE data which can be used to develop models, may be more appropriate. However, most cancer datasets have gene expression from tumour samples only, or if a matched sample is available, it is generally from adjacent tissue, which is not appropriate for assessing germline ASE. Tumour samples have increased rates of acquired ASE compared to the normal sample and adjacent normal samples have been shown to have rates similar to tumour tissue but distinct from normal tissue at more distant sites [347]. Alternatively, a dataset of allelic expression in normal tissue with a much larger sample size than available in GTEx could yield enough information to predict ASE in TSGs more commonly mutated in particular cancer types.

4.5 Conclusion

In conclusion, we have shown that it is possible to predict ASE using genotype data for some genes. As a pilot study, we assessed the association between predicted TSG ASE and cancer risk in breast tissue. Although we found no association, this may be due to the poor performance of the methods used to predict ASE rather than a lack of relationship between the two. Additionally, it could be that we need to restrict our analysis to TSGs that have driver mutations in a particular cancer type. Our results showed that we were unable to predict ASE in TSGs that have driver mutations in breast cancer, and so we could not assess this. Our ability to detect ASE is limited by the availability of appropriate datasets to develop methods and the poor performance of existing methods designed to predict gene expression that we have modified for ASE prediction. Our results do show that when more ASE datasets become available, it would be possible to build up a resource of gene-level models that can be used to predict ASE. This will allow us to better assess the relationship between TSG ASE and cancer risk and may also prove helpful for researchers focused on specific Mendelian disorders associated with these genes. Predicting ASE in these genes would aid in determining the expected penetrance of disease and could help in identifying therapeutic targets.

4.6 Materials and Methods

4.6.1 Data Acquisition

We retrieved haplotype-level ASE generated by Castel *et al.* for 838 GTEx samples for 49 tissue types that were generated using WASP filters [339]. We used the haplotype-level data in order to determine which genes show ASE to use as the truth set for comparing the results of our predictions. Phased whole genome genotype data were obtained from GTEx [506]. We obtained haplotype level data for normal tissues from UKB [507] and phased whole genome genotype data for cancer tissues from PCAWG [508]. A list of 1217 TSGs was downloaded from the TSGene database on 10 January 2023. A list of 127 imprinted genes was downloaded from geneimprint [509] on 23 January 2023 (Supplementary Table2).

4.6.2 Determining aFC for each gene

Using the haplotype expression matrix file containing counts for haplotype A and haplotype B based on WASP-corrected RNA-seq alignments for each GTEx sample generated by Castel et. al. [339] we calculated allelic fold change (aFC) for each gene as:

$$aFC = log_2 \frac{haplotypeAcounts+1}{haplotypeBcounts+1}$$

In order to calculate significant ASE, we performed a binomial test for each gene and performed multiple test corrections using the Holm method [510]. Any gene which had an adjusted p-value <0.05 was considered to have significant ASE. We applied the p-value threshold when assessing the number of TSGs that show ASE in the phASER results. However, for the comparisons of phASER aFC with the predicted results, we assessed the results with and without the threshold applied in case the requirement of p-value <0.05 was too stringent.

4.6.3 Predicting gene level ASE using PrediXcan

Using the phased genotype files from GTEX, we created two new files for each allele of the haplotype to supply as input to PrediXcan. For heterozygote SNPs, we changed 0|1 phased genotype to 0|0 for haplotype A and 1|1 for haplotype B and changed 1|0 to 1|1 for haplotype A, and 0|0 for haplotype B. Homozygote SNPs remained unchanged as they are the same for both alleles. We predicted expression for each haplotype file using MASHR models for each of the 49 tissue types available using the following options with Predict.py: vcf mode= genotype, on the fly mapping = METADATA {}_ {}_ {}_{}_ b38; –vcf_genotypes and model_db_SNP_key=varID.

The results from PrediXcan gave us predicted z-score transformed values for each gene available for each tissue type. We then computed a z-score ratio, as described in [511], for the values from haplotype A and haplotype B to determine ASE for a particular gene. To obtain the z-score ratio by dividing the difference between the z-scores from haplotype A and haplotye B by the standard deviation of differences for all genes for that sample. We divide by the standard deviation in order to determine if the difference between the two scores is statistically significant [511].

4.6.4 Assessing the performance of PrediXcan

In order to assess how well PrediXcan performed when predicting ASE, we calculated sample level Pearson correlation for each tissue type. We also performed gene-level Pearson correlation analysis.

In order to assess whether a binary classification worked better than using a continuous value for ASE, we applied different combinations of thresholds to the real and predicted results and calculated the area under the curve (AUC) for each pair. We applied the following cutoff to the real data; increments of 1 ranging from aFC=1 up to aFC=6, using absolute values for aFC. We applied thresholds in 0.5 increments for the z-score rations ranging from z-score=1 to z-score=10, using absolute values for the z-score. We used the pROC package [512] in R to compute AUC with the default "DeLong" method for confidence intervals. Delong calculates the variance for the AUCs using the method present in Delong et al. [513] and calculates the confidence intervals using qnorm.

4.7 Logistic Regression Model Development

We applied logistic regression models for each gene as described in the workflow below (Figure 4.9). Firstly we created our two input datasets as follows:

1. Using the haplotype data phASER results for breast tissue samples we generated binary response variable of ASE (1) and no ASE (0) for each gene. We classified samples based on their aFC for each gene. If absolute aFC >1 the gene for a sample was classified as having ASE and if absolute aFC ≤ 1 it was classified as having no ASE. We also generated models incorporating a binomial p-value threshold to classify a sample as having ASE for that gene.
2. Using phased genotype data from GTEx, we generated all possible predictor variables for a gene by identifying SNPs that were present within 100KB of the gene and also present in the UKB dataset. We classified genes as heterozygous (1) and homozygous (0) based on their genotypes.

We then developed the models as follows:

1. We ensured that heterozygous SNPs were available to use as potential predictor variables. If there were none available we could not proceed with model development.
2. We split the samples so that 80% were in the training set and 20% were in the test set, ensuring that ASE status was balanced between the two sets using the groupdata2 v2.0.2 R package [514].
3. We developed the logistic regression models using the training set as follows:
   (a) To account for imbalanced response variables, we used the ROSE R package v0.0-4 [515] to create synthetic balanced samples. ROSE requires at least two majority and two minority class examples.
   (b) We assessed the predictive performance of SNPs using the Information Value (IV) v1.2.3 R package.

(c) We removed any SNPs that had an IV lower than 0.1 as these SNPs would have weak predictive performance. We also removed SNPs that had an IV higher than 0.5 as this are deemed suspicious and may cause overfitting [516].

(d) Before performing best subset selection of predictors, we prioritised SNPs based on their IV value. If a gene had more than 15 SNPs available to use for predictions we picked the 15 with the highest IV. This is because the function to perform best subset selection requires 15 or less values due to the high computational requirements of the method.

(e) If there were more than 1 SNP we performed 5 fold cross validation model selection using bestglm v 0.37.3 R package [517] to find a small subset of predictors with the best prediction accuracy, following the method outlined in http://www.science.smith.edu/ jcrouser/SDS293/labs/lab9-r.html.

(f) If best subset selection was performed we chose the model that had the lowest cross validation error.

4. By applying the models to the test dataset we generated prediction probability scores for each sample. Samples which had a predicted probability score of >0.5 was deemed to have ASE in the gene and no ASE otherwise. Finally, we assessed the model performance using pROC v1.18.0 R package [512] to generate area under the curve (AUC) values for each gene.

4.7.1 Predicting ASE in UKB

We limited our analysis in the UKB dataset to samples of white British heritage in order to limit the influence of population structure on our results. As breast cancer is the predominant cancer type in the UKB dataset, we focused on this cancer type for our analysis. In order to create the relevant input for use with PrediXcan, we created two-phased genotype VCFs, one for each haplotype as previously described. We analysed data from the 10,518 females with self-reported breast cancer and 10,518 white British females who had no reported cancer diagnosis of any kind as control samples. To ensure we were not introducing any biases into our dataset by reducing the data to these numbers, we randomly selected samples that matched the age and smoking status of the samples with breast cancer. In total, we analysed 21,036 samples.

We extracted variants for these 21,036 samples from the phased haplotype data using plink2 (v2.00a3LM AVX2 Intel (15 Jun 2020)) with the default options and exported the results to VCF format. We then created two haplotype files, as previously described for the GTEx data, to use as input to PrediXcan. UKB data were aligned to the hg19 genome build, so we used the liftover command within Predict.py to convert to hg38 using the hg19 to hg38 chain file. All other options were the same as those used for the GTEx data.

The 47 TSGs for which all SNPs were available for predicting gene expression in breast tissue were DLEC1, EED, SIRT6, SMARCA2, CNOT3, PNN, RASSF2, STUB1, FZR1, TGFB1, TBL2, RNF8, E2F3, CYB561D2, RAP1A, ARID1A, MXI1, TNFRSF10B, TNFSF9, MAX, ZFP36, MYBBP1A, XAF1, EDNRB, BRCA2, CDH13, SCYL1, EPHA2, SELENBP1, RBMS3, CTDSPL, LRIG1,

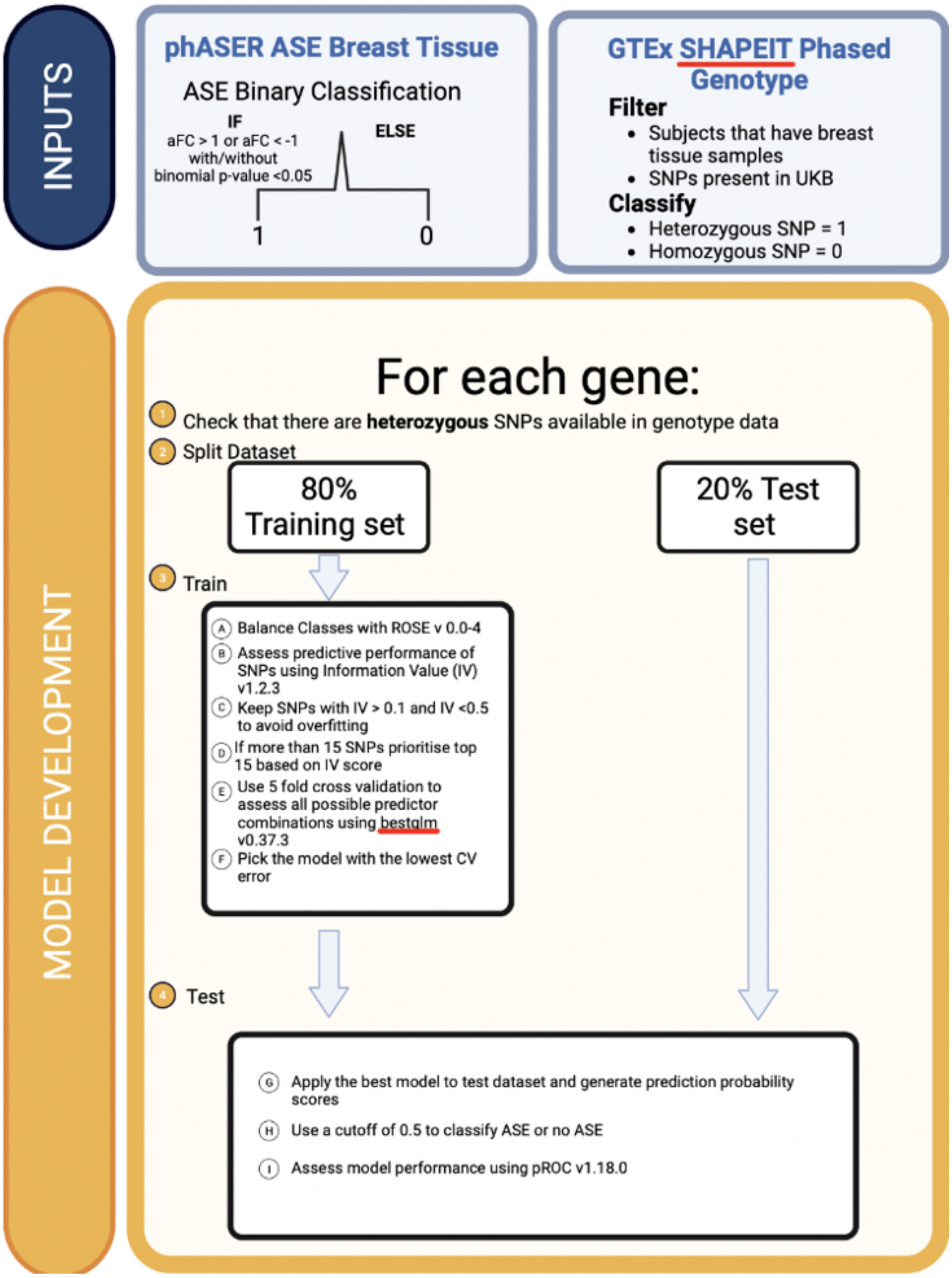

INPUTS
phASER ASE Breast Tissue
ASE Binary Classification
IF
aFC > 1 or aFC < -1
with/without
binomial p-value <0.05
ELSE
1
0
GTEx SHAPEIT Phased Genotype
Filter
• Subjects that have breast tissue samples
• SNPs present in UKB
Classify
• Heterozygous SNP = 1
• Homozygous SNP = 0
MODEL DEVELOPMENT
For each gene:
1 Check that there are **heterozygous** SNPs available in genotype data
2 Split Dataset
80% Training set
20% Test set
3 Train
A Balance Classes with ROSE v 0.0-4
B Assess predictive performance of SNPs using Information Value (IV) v1.2.3
C Keep SNPs with IV > 0.1 and IV <0.5 to avoid overfitting
D If more than 15 SNPs prioritise top 15 based on IV score
E Use 5 fold cross validation to assess all possible predictor combinations using bestglm v0.37.3
F Pick the model with the lowest CV error
4 Test
G Apply the best model to test dataset and generate prediction probability scores
H Use a cutoff of 0.5 to classify ASE or no ASE
I Assess model performance using pROC v1.18.0

Figure 4.9: **Workflow for generating gene level models using phASER allelic fold change as predictor and GTEx genotype data as response variable.**

NR1I2, TSLP, CDKN2A, CCAR2, HPGD, HEPACAM, STAT3, KISS1, IRX1, SAA1, PAWR, ALOX15B, ZFP82, SPRY4, TRIM31.

4.7.2 ASE Score Generation

In order to generate a single score to act as a measure of the level of TSG within a sample, we summed the absolute score of all predicted values. We used the absolute score because the direction of ASE is not important, but the magnitude of ASE is. The important thing is that one allele of the gene is showing ASE relative to the other. We did not apply a threshold but instead summed the values of all genes available. We also applied a threshold in order to produce a binary classification. We then generated a score summing the total number of TSGs showing ASE. We then assessed the association of the continuous score and binary score with the risk of breast cancer in UKB using by fitting a generalised linear regression model with cancer status as the response variable and TSG ASE score as the predictor.

4.7.3 ASE Prediction in PCAWG

We generated predictions for the 819 TCGA samples present within the PCAWG cohort. We used the phased whole genome genotype files available from [508] and generated two files for each allele as previously described. We ran PrediXcan using the same commands as for the UKB data, again using the function to liftOver from hg19 to hg38 genome. We used the Breast Mammary Tissue MASHR model from PrediXcan to predict ASE in breast cancer samples of those with infiltrating duct carcinoma, as this is the cancer type that corresponds to the GTEx tissue type.

5 Chapter5: Conclusions and Future Work

5.1 Summary and Main Findings

The aim of this thesis was to investigate somatic mutations present in cancer cells prior to the initiation of cancer as a means to understand somatic mutations in normal tissues. Recent studies of somatic mutation load in normal tissues have shown that there is still a lot unknown about how somatic mutations contribute to cancer initiation. While technologies to detect somatic mutations in normal tissue are improving, these analyses still have many limitations. Somatic mutations can be studied by analyzing cancer samples. However, cancer samples contain mutations that have occurred post-cancer initiation. Consequently, investigating all mutations in the sample does not give us a full understanding of what has occurred pre-cancer initiation. Therefore, it is essential to identify those mutations that accumulated throughout the history of the cell and its progenitors prior to the development of cancer. Once these mutations were identified, we were able to assess some of the factors influencing their abundance in the sample. We investigated the role of the immune system in shaping the mutational landscape of the cancer cell before it has escaped immune recognition. We also investigated hereditary variants that result in allele-specific expression (ASE). Germline ASE could reduce the number of somatic mutations that are required for a normal cell to transition to a cancer cell. For example, a single somatic mutation may be sufficient to inactivate a tumour suppressor gene in individuals who have a high level of allele-specific expression of the gene.

To understand what is happening in the cell before cancer transformation, it was first necessary to accurately determine which somatic mutations were present in the initial cancer cell. Chapter 2 focused on identifying clonal somatic mutations using bulk sequencing data from TCGA samples. Bulk sequencing has the disadvantage that it is generally a mix of tumor and normal cells so it is difficult to distinguish true clonal mutations when using variant frequency alone. In the absence of copy number changes, the expected frequency of autosomal clonal variants in samples that consist only of tumor cells, with no contamination from normal cells, is close to 0.5 (with some deviation from 0.5 due to bias introduced from aligning reads with a mutated allele compared to the normal allele). It is essential to account for tumor purity when determining which somatic mutations were present prior to the occurrence of the most recent common ancestor of the tumor. However, even when accounting for tumor purity and local copy number status of a variant it can still be difficult to distinguish clonal from subclonal mutations using read counts from bulk sequencing data. We estimated that using thresholds based on binomial tests comparing alternative and reference allele read accounts, which allows us to account for alignment bias, could only accurately classify approximately 45% of true clonal calls. This means that the clonal status of the majority of variants could not be established using read counts from variants with bulk-sequencing data.

To avoid the need to determine the total number of clonal mutations in a sample using sequencing data we developed a linear model to predict the true clonal load of a sample using the known relationship between somatic mutation accumulation and age. In Chapter 2 we generated a generalized linear regression model using age and cancer type as covariates to predict the total clonal mutational burden for a sample. To our knowledge, this was the first time this approach has been taken to estimate the total number of clonal mutations within a sample. We showed that predictions from this model closely matched somatic mutation numbers observed for the same age and tissue type as supplied to the model, indicting this approach worked well for estimating the somatic mutation load of a cell pre-cancer initiation. However, there were some deviations from predictions which were likely due to environmental factors that were not accounted for in our model. Although these predictions do not tell us which mutations are clonal it does give an indication of the estimated total load and therefore can be informative about the proportion of clonal mutations have accurately been identified within a sample. However, more accurate predictions that incorporate environmental factors in the model would be beneficial to get more precise estimates of clonal load.

In Chapter 2 we also assessed the relationship between somatic mutation numbers and lifetime cancer risk using predicted clonal load at age 80. A previous study found an association between the lifetime stem cell division numbers for a tissue type and risk of developing cancer [73]. The positive relationship between these two quantities was explained as arising from an increased somatic mutation burden for tissues with many stem cell divisions. However, we found a lower association between predicted clonal mutation burden and cancer risk than between lifetime stem cell divisions and cancer risk. This was surprising because if the relationship between lifetime stem cell division and cancer risk is due to increased somatic mutation burden with every stem cell division, we would expect a stronger association. This suggests that there may be factors in addition to mutation burden responsible for the relationship between lifetime stem cell division and tissue-specific cancer risk reported by Tomasetti and Vogelstein [73]. This is an important finding, given the controversy generated by the results of Tomasetti and Vogelstein, who were accused in the literature of downplaying the role of environmental factors such as UV exposure and smoking on the risk of developing cancer.

Once we identified the clonal mutations in TCGA samples, we next sought to understand the impact of the immune system on shaping the mutational landscape. Chapter 3 focused on investigating the relationship between the expression of genes harboring passenger mutations and immune evasion. We hypothesized that passenger mutations successfully evaded detection and removal by the immune system due to lowered expression compared to other genes within a sample or alternatively the mutated allele was downregulated compared to the normal allele in a process termed allele-specific expression. We found no evidence that passenger mutations occur preferentially on lowly expressed genes or that the mutated allele was downregulated compared to the normal allele, after we accounted for sequence context and gene length.

Additionally in Chapter 3, using simulations, we estimated an upper bound for the proportion of missense mutations that could be eliminated by the immune system without detection in our data. We estimated that if at least 5% of missense mutations were removed by immunoediting we would be able to detect it in our analysis. To our knowledge this was the first attempt to estimate an upper bound for the impact of immunoediting on the mutation landscape in cancer. Our results are consistent with no mutations being lost through immunoediting or if immunoediting is playing a role on the mutation landscape of cancer the effect is too subtle to observe. It is possible that the number of missense mutations that can be recognized and eliminated by the immune system is lower than currently thought. This may be due to the methods generally used to predict the immunogenicity of a mutation. Most methods rely on predicting MHC binding affinity. However, although this is an important step in the process, it is not enough for the peptide to be presented on the cell surface. There must also be a TCR capable of recognizing the peptide for the immune system to remove it. A recent study by an international consortium estimated that only 6% of predicted neoantigens can bind to a TCR for effective removal by the immune system [463]. Therefore, only a small proportion of missense mutations can generate an immune response.

Furthermore, in Chapter 3, we showed that reanalysis of two previous studies reporting that certain driver mutations are commonly seen in cancer because of a gap in a patient's MHC genotype to recognize them was due to unjustified statistical assumptions. The observed relationship between MHC genotype and driver mutation occurrence was due to the presence of multiple high frequency variants that had highly correlated MHC binding affinity scores. When accounting for this we found that the signal reported by Marty *et al.* and Marty-Pyke *et al.* [205, 206] disappeared and that there was no evidence that HLA genotype influences the driver mutations that occur in cancer patients.

In Chapter 4, we assessed the impact of germline variation of tumor suppressor genes (TSGs) on cancer risk. Inherited germline mutations of certain tumor suppressor genes such as BRCA1 and BRCA2 confer an increased risk of developing certain cancers [69] and certain SNPs resulting in ASE of these genes have also been shown to increase cancer risk [354–360]. Here, we investigated the effect of inherited germline variation that could result in ASE in TSGs in general instead of studying specific SNPs. ASE could disrupt the expression of one version of a tumor suppressor gene so that mutations are only required in one copy, the expressed copy, to initiate cancer. We hypothesized that inherited ASE of TSGs would result in fewer somatic mutations being required to disrupt the overall expression of TSGs, thereby increasing cancer risk. To assess this in the UK Biobank data we first needed to predict gene level ASE within a sample. ASE is usually studied by comparing the ratio of reads covering the mutant allele or mutant haplotype to the normal allele/haplotype using RNA-Sequencing data. This is then compared to the ratio of mutant and normal reads in matched DNA-Sequencing data to eliminate the effect of alignment bias. However, RNA-Seq and DNA-Seq data are not always available for a sample. Moreover, RNA-Seq data was not available for UK Biobank data for which we used to assess cancer risk. Therefore, we first aimed to predict gene level ASE using genotype data only.

Although methods have been developed to predict gene expression level from genotype data, to our knowledge, this was the first attempt to predict gene-level ASE using DNA-Seq data only. Our results highlighted the difficulty of predicting gene-level ASE with the available data. The largest resource of allelic imbalance in humans available to date was generated in the GTEx data set. While this provided ASE results for a large variety of tissue types (n=54) there was only data available from 838 individuals and although ASE was observed in a large number of TSGs there were small numbers of samples with ASE of a particular TSG. This made it difficult to generate models to predict ASE for all TSGs. As a result, our initial approach was to utilize a tool, PrediXcan, designed to predict overall gene expression using genotype data by providing information for each haplotype independently and determining a ratio of ASE using the predicted score for each. While this appeared to work well for some genes it did not work well for all. This was likely caused by the poor performance of PrediXcan to predict overall gene expression for some genes. If it does not work well for the predictions it was designed for it is unsurprising that it does not work well for predicting ASE either. As PrediXcan models improve for predicating overall expression it is possible that it would also improve our ability to predict ASE using them. We next sought to fit generalized linear models for each gene with a binary outcome of ASE or no ASE using the genotype (classified as heterozygous or not) of SNPs within 100 kilobase of the gene. This only performed well for a subset of genes, likely due to an insufficient number of individuals showing ASE for the majority of TSGs. Additionally, the UK Biobank data did not have whole genome phased genotype data which meant there were some genes that did not have SNPs available for building the models.

Finally in Chapter 4, as a pilot study, we predicted ASE in TSGs for which the models performed well and tested an association with breast cancer risk. Although, we found no evidence of a relationship between ASE in TSGs and cancer risk it is possible this was due to poor performance of ASE predictions. When we analysed the TSGs common in breast cancer samples from TCGA we discovered that the TSGs for which we could make predictions were not common drivers of breast cancer. Therefore, it is likely that genes for which ASE could confer an increased cancer risk in breast tissue were not included in the analysis.

5.2 Future Work

Although our method for predicting the total clonal burden of a sample generated estimates comparable to those observed in the corresponding normal tissue for the same age some improvements could be made. Our model was generated using pan-cancer data as input with cancer type used as a covariate. However, there was varying numbers of samples available for each cancer type which may impact the results. It would be advantageous to generate models using larger cancer data sets to improve estimates. It may be better to generate models on a per-cancer basis using cancer sub-type as a covariate to get more accurate predictions for individual cancer types. Additionally, it would be useful to incorporate environmental factors as covariates as exposure to these would likely affect the mutational burden within a sample.

Finally, our results indicate that it is feasible to predict ASE for some genes using genotype data but that is difficult with the limited amount of data currently available. A larger resource of gene expression from normal samples is required to predict ASE for all genes. Additionally, in the case of ASE in TSGs, it may be more beneficial to use samples from normal tissues of patients with cancer. Germline ASE in TSGs might be too rare in the general population to generate sufficient data for developing models. Therefore, it could be useful to generate the data from cancer samples which could have increased numbers of samples with germline ASE. However, somatic ASE is common in cancer tissue and is increased in normal tissue adjacent to the tumor [347]. Therefore, it is essential that the normal sample be taken from a site in the tissue that is distal from the tumor. Generating models to predict ASE is relevant to diseases other than cancer and therefore a resource of models to predict ASE in all genes, not just TSGs, using this data could prove useful for diseases with variable penetrance that could be explained by ASE [504].

Bibliography

[1] Rossi, D. J.; Jamieson, C. H.; Weissman, I. L. Stems cells and the pathways to aging and cancer. *Cell* **2008**, *132*, 681–696.

[2] Miles, B.; Tadi, P. Genetics, somatic mutation. **2020**,

[3] De, S. Somatic mosaicism in healthy human tissues. *Trends in Genetics* **2011**, *27*, 217–223.

[4] Consortium, . G. P., *et al.* A map of human genome variation from population scale sequencing. *Nature* **2010**, *467*, 1061.

[5] Spencer, P. S.; Barral, J. M. Genetic code redundancy and its influence on the encoded polypeptides. *1*, e201204006.

[6] Cote, R. G.; Jones, P.; Apweiler, R.; Hermjakob, H. The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *7*, 97.

[7] Vihinen, M. When a Synonymous Variant Is Nonsynonymous. *13*, 1485.

[8] McNeely, T.; Leone, M.; Yanai, H.; Beerman, I. DNA damage in aging, the stem cell perspective. *139*, 309–331.

[9] Vijg, J.; Dong, X. Pathogenic Mechanisms of Somatic Mutation and Genome Mosaicism in Aging. *182*, 12–23.

[10] Proukakis, C. Somatic mutations in neurodegeneration: An update. *144*, 105021.

[11] Jourdon, A.; Fasching, L.; Scuderi, S.; Abyzov, A.; Vaccarino, F. M. The role of somatic mosaicism in brain disease. *65*, 84–90.

[12] Olafsson, S. *et al.* Somatic Evolution in Non-neoplastic IBD-Affected Colon. *182*, 672–684.e11.

[13] Nanki, K. *et al.* Somatic inflammatory gene mutations in human ulcerative colitis epithelium. *577*, 254–259, Number: 7789 Publisher: Nature Publishing Group.

[14] Mustjoki, S.; Young, N. S. Somatic Mutations in Benign Disease. *384*, 2039–2052, Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMra2101920.

[15] Dou, Y.; Gold, H. D.; Luquette, L. J.; Park, P. J. Detecting Somatic Mutations in Normal Cells. *34*, 545–557, Publisher: Elsevier.

[16] Blokzijl, F. *et al.* Tissue-specific mutation accumulation in human adult stem cells during life. *538*, 260–264, Number: 7624 Publisher: Nature Publishing Group.

[17] Lodato, M. A. *et al.* Aging and neurodegeneration are associated with increased mutations in single human neurons. *359*, 555–559, Publisher: American Association for the Advancement of Science.

[18] Kakiuchi, N.; Ogawa, S. Clonal expansion in non-cancer tissues. *21*, 239–256, Number: 4 Publisher: Nature Publishing Group.

[19] Espina, V.; Wulfkuhle, J. D.; Calvert, V. S.; VanMeter, A.; Zhou, W.; Coukos, G.; Geho, D. H.; Petricoin, E. F.; Liotta, L. A. Laser-capture microdissection. *1*, 586–603, Number: 2 Publisher: Nature Publishing Group.

[20] Martincorena, I.; Campbell, P. J. Somatic mutation in cancer and normal cells. *349*, 1483–1489, Publisher: American Association for the Advancement of Science.

[21] Simons, B. D. Deep sequencing as a probe of normal stem cell fate and preneoplasia in human epidermis. *113*, 128–133, Publisher: Proceedings of the National Academy of Sciences.

[22] Lynch, M. D.; Lynch, C. N. S.; Craythorne, E.; Liakath-Ali, K.; Mallipeddi, R.; Barker, J. N.; Watt, F. M. Spatial constraints govern competition of mutant clones in human epidermis. *8*, 1119, Number: 1 Publisher: Nature Publishing Group.

[23] Wei, L. *et al.* Ultradeep sequencing differentiates patterns of skin clonal mutations associated with sun-exposure status and skin cancer burden. *7*, eabd7703.

[24] Kennedy, S. R.; Schmitt, M. W.; Fox, E. J.; Kohrn, B. F.; Salk, J. J.; Ahn, E. H.; Prindle, M. J.; Kuong, K. J.; Shen, J.-C.; Risques, R.-A.; Loeb, L. A. Detecting ultralow-frequency mutations by Duplex Sequencing. *9*, 2586–2606, Number: 11 Publisher: Nature Publishing Group.

[25] Hoang, M. L.; Kinde, I.; Tomasetti, C.; McMahon, K. W.; Rosenquist, T. A.; Grollman, A. P.; Kinzler, K. W.; Vogelstein, B.; Papadopoulos, N. Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *113*, 9846–9851, Publisher: Proceedings of the National Academy of Sciences.

[26] Abascal, F. *et al.* Somatic mutation landscapes at single-molecule resolution. *593*, 405–410.

[27] Fowler, J. C.; Jones, P. H. Somatic Mutation: What Shapes the Mutational Landscape of Normal Epithelia? *12*, 1642–1655.

[28] Ueda, S. *et al.* A quantification method of somatic mutations in normal tissues and their accumulation in pediatric patients with chemotherapy. *119*, e2123241119.

[29] Maslov, A. Y.; Makhortov, S.; Sun, S.; Heid, J.; Dong, X.; Lee, M.; Vijg, J. Single-molecule, quantitative detection of low-abundance somatic mutations by high-throughput sequencing. *8*, eabm3259, Publisher: American Association for the Advancement of Science.

[30] Spits, C.; Le Caignec, C.; De Rycke, M.; Van Haute, L.; Van Steirteghem, A.; Liebaers, I.; Sermon, K. Whole-genome multiple displacement amplification from single cells. *1*, 1965–1970, Number: 4 Publisher: Nature Publishing Group.

[31] Zong, C.; Lu, S.; Chapman, A. R.; Xie, X. S. Genome-Wide Detection of Single Nucleotide and Copy Number Variations of a Single Human Cell. *338*, 1622, Publisher: NIH Public Access.

[32] Telenius, H.; Carter, N. P.; Bebb, C. E.; Nordenskjld, M.; Ponder, B. A.; Tunnacliffe, A. Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *13*, 718–725.

[33] Kashima, Y.; Sakamoto, Y.; Kaneko, K.; Seki, M.; Suzuki, Y.; Suzuki, A. Single-cell sequencing techniques from individual to multiomics analyses. *52*, 1419–1427, Number: 9 Publisher: Nature Publishing Group.

[34] Lahnemann, D.; Kster, J.; Fischer, U.; Borkhardt, A.; McHardy, A. C.; Schnhuth, A. Accurate and scalable variant calling from single cell DNA sequencing data with ProSolo. *12*, 6744.

[35] Vu, T. N.; Nguyen, H.-N.; Calza, S.; Kalari, K. R.; Wang, L.; Pawitan, Y. Cell-level somatic mutation detection from single-cell RNA sequencing. *35*, 4679–4687.

[36] Bizzotto, S.; Dou, Y.; Ganz, J.; Doan, R. N.; Kwon, M.; Bohrson, C. L.; Kim, S. N.; Bae, T.; Abyzov, A.; Network†, N. B. S. M., *et al.* Landmarks of human embryonic development inscribed in somatic mutations. *Science* **2021**, *371*, 1249–1253.

[37] Muyas, F.; Li, R.; Rahbari, R.; Mitchell, T. J.; Hormoz, S.; Cortes-Ciriano, I. Accurate de novo detection of somatic mutations in high-throughput single-cell profiling data sets. *bioRxiv* **2022**, 2022–11.

[38] Garcia-Nieto, P. E.; Morrison, A. J.; Fraser, H. B. The somatic mutation landscape of the human body. *20*, 298.

[39] Yizhak, K. *et al.* RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *364*, eaaw0726.

[40] Cirulli, E. T.; Singh, A.; Shianna, K. V.; Ge, D.; Smith, J. P.; Maia, J. M.; Heinzen, E. L.; Goedert, J. J.; Goldstein, D. B.; the Center for HIV/AIDS Vaccine Immunology (CHAVI), Screening the human exome: a comparison of whole genome and whole transcriptome sequencing. *11*, R57.

[41] Grist, S.; McCarron, M.; Kutlaca, A.; Turner, D.; Morley, A. In vivo human somatic mutation: frequency and spectrum with age. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **1992**, *266*, 189–196.

[42] Franco, I.; Johansson, A.; Olsson, K.; Vrtačnik, P.; Lundin, P.; Helgadottir, H. T.; Larsson, M.; Revêchon, G.; Bosia, C.; Pagnani, A., *et al.* Somatic mutagenesis in satellite cells associates with human skeletal muscle aging. *Nature communications* **2018**, *9*, 800.

[43] Wyles, S. P.; Brandt, E. B.; Nelson, T. J. Stem Cells: The Pursuit of Genomic Stability. *15*, 20948–20967, Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

[44] Manders, F.; van Boxtel, R.; Middelkamp, S. The Dynamics of Somatic Mutagenesis During Life in Humans. *2*.

[45] Brazhnik, K.; Sun, S.; Alani, O.; Kinkhabwala, M.; Wolkoff, A. W.; Maslov, A. Y.; Dong, X.; Vijg, J. Single-cell analysis reveals different age-related somatic mutation profiles between stem and differentiated cells in human liver. *6*, eaax2659, Publisher: American Association for the Advancement of Science.

[46] Alexandrov, L. B.; Kim, J.; Haradhvala, N. J.; Huang, M. N.; Tian Ng, A. W.; Wu, Y.; Boot, A.; Covington, K. R.; Gordenin, D. A.; Bergstrom, E. N., *et al.* The repertoire of mutational signatures in human cancer. *Nature* **2020**, *578*, 94–101.

[47] Nik-Zainal, S.; Van Loo, P.; Wedge, D. C.; Alexandrov, L. B.; Greenman, C. D.; Lau, K. W.; Raine, K.; Jones, D.; Marshall, J.; Ramakrishna, M., *et al.* The life history of 21 breast cancers. *Cell* **2012**, *149*, 994–1007.

[48] Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *500*, 415–421, Number: 7463 Publisher: Nature Publishing Group.

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

[50] Hayward, N. K.; Wilmott, J. S.; Waddell, N.; Johansson, P. A.; Field, M. A.; Nones, K.; Patch, A.-M.; Kakavand, H.; Alexandrov, L. B.; Burke, H., *et al.* Whole-genome landscapes of major melanoma subtypes. *Nature* **2017**, *545*, 175–180.

[51] Wijewardhane, N.; Dressler, L.; Ciccarelli, F. D. Normal Somatic Mutations in Cancer Transformation. *39*, 125–129.

[52] Kennedy, S. R.; Zhang, Y.; Risques, R. A. Cancer-Associated Mutations but No Cancer: Insights into the Early Steps of Carcinogenesis and Implications for Early Cancer Detection. *5*, 531–540.

[53] Fiala, C.; Diamandis, E. P. Mutations in normal tissues-some diagnostic and clinical implications. *18*, 283.

[54] Cooper, G.; Adams, K. *The cell: a molecular approach*; Oxford University Press, 2022.

[55] Bozic, I.; Antal, T.; Ohtsuki, H.; Carter, H.; Kim, D.; Chen, S.; Karchin, R.; Kinzler, K. W.; Vogelstein, B.; Nowak, M. A. Accumulation of driver and passenger mutations during tumor progression. *Proceedings of the National Academy of Sciences* **2010**, *107*, 18545–18550.

[56] Haber, D. A.; Settleman, J. Drivers and passengers. *Nature* **2007**, *446*, 145–146.

[57] Harvey, J. An unidentified virus which causes the rapid production of tumours in mice. *Nature* **1964**, *204*, 1104–1105.

[58] Kirsten, W.; Mayer, L. A. Morphologic responses to a murine erythroblastosis virus. *Journal of the National Cancer Institute* **1967**, *39*, 311–335.

[59] Weinberg, R. A. Cellular oncogenes. *9*, 131–133, Publisher: Elsevier.

[60] Hahn, W. C.; Weinberg, R. A. Modelling the molecular circuitry of cancer. *Nature Reviews Cancer* **2002**, *2*, 331–341.

[61] White, M. C.; Holman, D. M.; Boehm, J. E.; Peipins, L. A.; Grossman, M.; Henley, S. J. Age and Cancer Risk. *46*, S7–15.

[62] Siegel, R. L.; Miller, K. D.; Jemal, A. Cancer statistics, 2019. *69*, 7–34, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21551.

[63] Alexandrov, L. B.; Ju, Y. S.; Haase, K.; Van Loo, P.; Martincorena, I.; Nik-Zainal, S.; Totoki, Y.; Fujimoto, A.; Nakagawa, H.; Shibata, T.; Campbell, P. J.; Vineis, P.; Phillips, D. H.; Stratton, M. R. Mutational signatures associated with tobacco smoking in human cancer. *354*, 618–622, Publisher: American Association for the Advancement of Science.

[64] Wang, R.; Li, S.; Wen, W.; Zhang, J. Multi-Omics Analysis of the Effects of Smoking on Human Tumors. *8*.

[65] Gandini, S.; Botteri, E.; Iodice, S.; Boniol, M.; Lowenfels, A. B.; Maisonneuve, P.; Boyle, P. Tobacco smoking and cancer: A meta-analysis. *122*, 155–164, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ijc.23033.

[66] Blot, W. J. Alcohol and cancer. *Cancer research* **1992**, *52*, 2119s–2123s.

[67] Narayanan, D. L.; Saladi, R. N.; Fox, J. L. Ultraviolet radiation and skin cancer. *International journal of dermatology* **2010**, *49*, 978–986.

[68] Das, S.; Thakur, S.; Korenjak, M.; Sidorenko, V. S.; Chung, F. F.-L.; Zavadil, J. Aristolochic acid-associated cancers: a public health risk in need of global action. *22*, 576–591, Number: 10 Publisher: Nature Publishing Group.

[69] Liede, A.; Karlan, B. Y.; Narod, S. A. Cancer Risks for Male Carriers of Germline Mutations in BRCA1 or BRCA2: A Review of the Literature. *22*, 735–742, Publisher: Wolters Kluwer.

[70] Barrow, E.; Hill, J.; Evans, D. G. Cancer risk in Lynch Syndrome. *12*, 229–240.

[71] Lynch, H. T.; de la Chapelle, A. Hereditary Colorectal Cancer. *348*, 919–932, Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMra012242.

[72] Senter, L.; Clendenning, M.; Sotamaa, K.; Hampel, H.; Green, J.; Potter, J. D.; Lindblom, A.; Lagerstedt, K.; Thibodeau, S. N.; Lindor, N. M., *et al.* The clinical phenotype of Lynch syndrome due to germ-line PMS2 mutations. *Gastroenterology* **2008**, *135*, 419–428.

[73] Tomasetti, C.; Vogelstein, B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *347*, 78–81.

[74] Dentro, S. C.; Wedge, D. C.; Van Loo, P. Principles of Reconstructing the Subclonal Architecture of Cancers. *7*, a026625.

[75] Nowell, P. C. The Clonal Evolution of Tumor Cell Populations. *194*, 23–28, Publisher: American Association for the Advancement of Science.

[76] Williams, M. J.; Werner, B.; Barnes, C. P.; Graham, T. A.; Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nature genetics* **2016**, *48*, 238–244.

[77] Vogelstein, B.; Papadopoulos, N.; Velculescu, V. E.; Zhou, S.; Diaz, L. A.; Kinzler, K. W. Cancer Genome Landscapes. *339*, 1546–1558, Publisher: American Association for the Advancement of Science.

[78] Williams, M. J.; Werner, B.; Heide, T.; Curtis, C.; Barnes, C. P.; Sottoriva, A.; Graham, T. A. Quantification of subclonal selection in cancer from bulk sequencing data. *50*, 895–903, Number: 6 Publisher: Nature Publishing Group.

[79] Dentro, S. C. *et al.* Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *184*, 2239–2254.e39, Publisher: Elsevier.

[80] Schmitt, M. W.; Loeb, L. A.; Salk, J. J. The influence of subclonal resistance mutations on targeted cancer therapy. *13*, 335–347, Number: 6 Publisher: Nature Publishing Group.

[81] Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non-Small-Cell Lung Cancer. *376*, 2109–2121.

[82] Huyghe, J. R.; Bien, S. A.; Harrison, T. A.; Kang, H. M.; Chen, S.; Schmit, S. L.; Conti, D. V.; Qu, C.; Jeon, J.; Edlund, C. K., *et al.* Discovery of common and rare genetic risk variants for colorectal cancer. *Nature genetics* **2019**, *51*, 76–87.

[83] Tate, J. G.; Bamford, S.; Jubb, H. C.; Sondka, Z.; Beare, D. M.; Bindal, N.; Boutselakis, H.; Cole, C. G.; Creatore, C.; Dawson, E., *et al.* COSMIC: the catalogue of somatic mutations in cancer. *Nucleic acids research* **2019**, *47*, D941–D947.

[84] Collins, A.; Project Team, T. C. G. A. The cancer genome atlas (TCGA) pilot project. *Cancer Research* **2007**, *67*, LB–247.

[85] Pan-cancer analysis of whole genomes. *Nature* **2020**, *578*, 82–93.

[86] Jennings, J. L.; Hudson, T. J. International Cancer Genome Consortium (ICGC). *Cancer Research* **2010**, *70*, 2226–2226.

[87] Hartmaier, R. J.; Albacker, L. A.; Chmielecki, J.; Bailey, M.; He, J.; Goldberg, M. E.; Ramkissoon, S.; Suh, J.; Elvin, J. A.; Chiacchia, S., *et al.* High-Throughput Genomic Profiling of Adult Solid Tumors Reveals Novel Insights into Cancer PathogenesisPublic Release of Adult Cancer Genomic Data. *Cancer research* **2017**, *77*, 2464–2475.

[88] Micheel, C. M.; Sweeney, S. M.; LeNoue-Newton, M. L.; André, F.; Bedard, P. L.; Guinney, J.; Meijer, G. A.; Rollins, B. J.; Sawyers, C. L.; Schultz, N., *et al.* American Association for Cancer Research Project Genomics Evidence Neoplasia Information Exchange: from inception to first data release and beyond—lessons learned and member institutions' perspectives. *JCO clinical cancer informatics* **2018**, *2*, 1–14.

[89] Ellis, M. J.; Gillette, M.; Carr, S. A.; Paulovich, A. G.; Smith, R. D.; Rodland, K. K.; Townsend, R. R.; Kinsinger, C.; Mesri, M.; Rodriguez, H., *et al.* Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer discovery* **2013**, *3*, 1108–1112.

[90] Lilly, J. V.; Rokita, J. L.; Mason, J. L.; Patton, T.; Stefankiewiz, S.; Higgins, D.; Trooskin, G.; Larouci, C. A.; Arya, K.; Appert, E., *et al.* The children's brain tumor network (CBTN)-Accelerating research in pediatric central nervous system tumors through collaboration and open science. *Neoplasia* **2023**, *35*, 100846.

[91] McLeod, C.; Gout, A. M.; Zhou, X.; Thrasher, A.; Rahbarinia, D.; Brady, S. W.; Macias, M.; Birch, K.; Finkelstein, D.; Sunny, J., *et al.* St. Jude Cloud: a pediatric cancer genomic data-sharing ecosystem. *Cancer discovery* **2021**, *11*, 1082–1099.

[92] Qin, D. Next-generation sequencing and its clinical application. *Cancer biology & medicine* **2019**, *16*, 4.

[93] Goldfeder, R. L.; Wall, D. P.; Khoury, M. J.; Ioannidis, J. P.; Ashley, E. A. Human genome sequencing at the population scale: a primer on high-throughput DNA sequencing and analysis. *American journal of epidemiology* **2017**, *186*, 1000–1009.

[94] Haworth, A.; Savage, H.; Lench, N. Diagnostic Genomics and Clinical Bioinformatics. *Medical and Health Genomics* **2016**, 37–50.

[95] Griffith, M.; Miller, C. A.; Griffith, O. L.; Krysiak, K.; Skidmore, Z. L.; Ramu, A.; Walker, J. R.; Dang, H. X.; Trani, L.; Larson, D. E., *et al.* Optimizing cancer genome sequencing and analysis. *Cell systems* **2015**, *1*, 210–223.

[96] Lou, J. J.; Mirsadraei, L.; Sanchez, D. E.; Wilson, R. W.; Shabihkhani, M.; Lucey, G. M.; Wei, B.; Singer, E. J.; Mareninov, S.; Yong, W. H. A review of room temperature storage of biospecimen tissue and nucleic acids for anatomic pathology laboratories and biorepositories. *47*, 267–273.

[97] O'Rourke, M. B.; Padula, M. P. Analysis of formalin-fixed, paraffin-embedded (FFPE) tissue via proteomic techniques and misconceptions of antigen retrieval. *60*, 229–238, Publisher: Future Science.

[98] Mager, S. R.; Oomen, M. H. A.; Morente, M. M.; Ratcliffe, C.; Knox, K.; Kerr, D. J.; Pezzella, F.; Riegman, P. H. J. Standard operating procedure for the collection of fresh frozen tissue samples. *43*, 828–834.

[99] Gao, X. H.; Li, J.; Gong, H. F.; Yu, G. Y.; Liu, P.; Hao, L. Q.; Liu, L. J.; Bai, C. G.; Zhang, W. Comparison of fresh frozen tissue with formalin-fixed paraffin-embedded tissue for mutation analysis using a multi-gene panel in patients with colorectal cancer. *Frontiers in oncology* **2020**, *10*, 310.

[100] Auer, H. *et al.* The effects of frozen tissue storage conditions on the integrity of RNA and protein. *89*, 518–528.

[101] Robbe, P. *et al.* Clinical whole-genome sequencing from routine formalin-fixed, paraffin-embedded specimens: pilot study for the 100,000 Genomes Project. *20*, 1196–1205, Number: 10 Publisher: Nature Publishing Group.

[102] Gross, A. M.; Kreisberg, J. F.; Ideker, T. Analysis of Matched Tumor and Normal Profiles Reveals Common Transcriptional and Epigenetic Signals Shared across Cancer Types. *10*, e0142618, Publisher: Public Library of Science.

[103] Koboldt, D. C. Best practices for variant calling in clinical sequencing. *12*, 91.

[104] FastQC. `https://qubeshub.org/resources/fastqc`.

[105] Bolger, A. M.; Lohse, M.; Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *30*, 2114–2120.

[106] Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* **2013**,

[107] Van der Auwera, G. A.; O'Connor, B. D. *Genomics in the cloud: using Docker, GATK, and WDL in Terra*; O'Reilly Media, 2020.

[108] Mose, L. E.; Wilkerson, M. D.; Hayes, D. N.; Perou, C. M.; Parker, J. S. ABRA: improved coding indel detection via assembly-based realignment. *30*, 2813, Publisher: Oxford University Press.

[109] Benjamin, D.; Sato, T.; Cibulskis, K.; Getz, G.; Stewart, C.; Lichtenstein, L. Calling Somatic SNVs and Indels with Mutect2. `https://www.biorxiv.org/content/10.1101/861054v1`, Pages: 861054 Section: New Results.

[110] Kroigrd, A. B.; Thomassen, M.; Lnkholm, A.-V.; Kruse, T. A.; Larsen, M. J. Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data. *11*, e0151664, Publisher: Public Library of Science.

[111] Bian, X.; Zhu, B.; Wang, M.; Hu, Y.; Chen, Q.; Nguyen, C.; Hicks, B.; Meerzaman, D. Comparing the performance of selected variant callers using synthetic data and genome segmentation. *19*, 429.

[112] Wang, M.; Luo, W.; Jones, K.; Bian, X.; Williams, R.; Higson, H.; Wu, D.; Hicks, B.; Yeager, M.; Zhu, B. SomaticCombiner: improving the performance of somatic variant calling based on evaluation tests and a consensus approach. *10*, Publisher: Nature Publishing Group.

[113] Ellrott, K.; Bailey, M. H.; Saksena, G.; Covington, K. R.; Kandoth, C.; Stewart, C.; Hess, J.; Ma, S.; McLellan, M.; Sofia, H. J.; Hutter, C.; Getz, G.; Wheeler, D.; Ding, L. Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *6*, 271–281.e7.

[114] Wang, K.; Li, M.; Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **2010**, *38*, e164–e164.

[115] McLaren, W.; Gil, L.; Hunt, S. E.; Riat, H. S.; Ritchie, G. R.; Thormann, A.; Flicek, P.; Cunningham, F. The ensembl variant effect predictor. *Genome biology* **2016**, *17*, 1–14.

[116] Koboldt, D. C.; Zhang, Q.; Larson, D. E.; Shen, D.; McLellan, M. D.; Lin, L.; Miller, C. A.; Mardis, E. R.; Ding, L.; Wilson, R. K. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *22*, 568–576, Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[117] Kim, S.; Scheffler, K.; Halpern, A. L.; Bekritsky, M. A.; Noh, E.; Kallberg, M.; Chen, X.; Kim, Y.; Beyter, D.; Krusche, P.; Saunders, C. T. Strelka2: fast and accurate calling of germline and somatic variants. *15*, 591–594, Number: 8 Publisher: Nature Publishing Group.

[118] Fan, Y.; Xi, L.; Hughes, D. S. T.; Zhang, J.; Zhang, J.; Futreal, P. A.; Wheeler, D. A.; Wang, W. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *17*, 178.

[119] Dunn, T.; Berry, G.; Emig-Agius, D.; Jiang, Y.; Lei, S.; Iyer, A.; Udar, N.; Chuang, H.-Y.; Hegarty, J.; Dickover, M.; Klotzle, B.; Robbins, J.; Bibikova, M.; Peeters, M.; Strmberg, M. Pisces: an accurate and versatile variant caller for somatic and germline next-generation sequencing data. *35*, 1579–1581.

[120] Gerstung, M.; Beerenwinkel, N. Calling subclonal mutations with deepSNV. 2015.

[121] Larson, D. E.; Harris, C. C.; Chen, K.; Koboldt, D. C.; Abbott, T. E.; Dooling, D. J.; Ley, T. J.; Mardis, E. R.; Wilson, R. K.; Ding, L. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *28*, 311–317.

[122] Lai, Z.; Markovets, A.; Ahdesmaki, M.; Chapman, B.; Hofmann, O.; McEwen, R.; Johnson, J.; Dougherty, B.; Barrett, J. C.; Dry, J. R. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *44*, e108.

[123] Riester, M.; Singh, A. P.; Brannon, A. R.; Yu, K.; Campbell, C. D.; Chiang, D. Y.; Morrissey, M. P. PureCN: copy number calling and SNV classification using targeted short read sequencing. *11*, 13.

[124] Roth, A.; Khattra, J.; Yap, D.; Wan, A.; Laks, E.; Biele, J.; Ha, G.; Aparicio, S.; Bouchard-Cote, A.; Shah, S. P. PyClone: statistical inference of clonal population structure in cancer. *11*, 396–398, Number: 4 Publisher: Nature Publishing Group.

[125] Xiao, Y.; Wang, X.; Zhang, H.; Ulintz, P. J.; Li, H.; Guan, Y. FastClone is a probabilistic tool for deconvoluting tumor heterogeneity in bulk-sequencing samples. *11*, 4469, Number: 1 Publisher: Nature Publishing Group.

[126] Jiang, Y.; Yu, K.; Ji, S.; Shin, S. J.; Cao, S.; Montierth, M. D.; Huang, L.; Kopetz, S.; Msaouel, P.; Wang, J. R., *et al.* CliP: subclonal architecture reconstruction of cancer cells in DNA sequencing data using a penalized likelihood model. *bioRxiv* **2021**, 2021–03.

[127] Miura, S.; Gomez, K.; Murillo, O.; Huuki, L. A.; Vu, T.; Buturla, T.; Kumar, S. Predicting clone genotypes from tumor bulk sequencing of multiple samples. *34*, 4017–4026.

[128] Strino, F.; Parisi, F.; Micsinai, M.; Kluger, Y. TrAp: a tree approach for fingerprinting subclonal tumor composition. *41*, e165.

[129] Leshchiner, I.; Livitz, D.; Gainor, J. F.; Rosebrock, D.; Spiro, O.; Martinez, A.; Mroz, E.; Lin, J. J.; Stewart, C.; Kim, J., *et al.* Comprehensive analysis of tumour initiation, spatial and temporal progression under multiple lines of treatment. *biorxiv* **2018**, 508127.

[130] Miller, C. A. *et al.* SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution. *10*, e1003665, Publisher: Public Library of Science.

[131] Malikic, S.; McPherson, A. W.; Donmez, N.; Sahinalp, C. S. Clonality inference in multiple tumor samples using phylogeny. *31*, 1349–1356.

[132] Niknafs, N.; Beleva-Guthrie, V.; Naiman, D. Q.; Karchin, R. SubClonal Hierarchy Inference from Somatic Mutations: Automatic Reconstruction of Cancer Evolutionary Trees from Multi-region Next Generation Sequencing. *11*, e1004416.

[133] Manica, M. *et al.* Inferring clonal composition from multiple tumor biopsies. *6*, 1–13, Number: 1 Publisher: Nature Publishing Group.

[134] Lakatos, E.; Hockings, H.; Mossner, M.; Huang, W.; Lockley, M.; Graham, T. A. LiquidCNA: Tracking subclonal evolution from longitudinal liquid biopsies using somatic copy number alterations. *24*, Publisher: Elsevier.

[135] Deveau, P. *et al.* QuantumClone: clonal assessment of functional mutations in cancer based on a genotype-aware method for clonal reconstruction. *34*, 1808–1816.

[136] Flensburg, C.; Sargeant, T.; Oshlack, A.; Majewski, I. J. SuperFreq: Integrated mutation detection and clonal tracking in cancer. *16*, e1007603, Publisher: Public Library of Science.

[137] Caravagna, G.; Heide, T.; Williams, M. J.; Zapata, L.; Nichol, D.; Chkhaidze, K.; Cross, W.; Cresswell, G. D.; Werner, B.; Acar, A.; Chesler, L.; Barnes, C. P.; Sanguinetti, G.; Graham, T. A.; Sottoriva, A. Subclonal reconstruction of tumors by using machine learning and population genetics. *52*, 898–907, Number: 9 Publisher: Nature Publishing Group.

[138] Andersson, N.; Chattopadhyay, S.; Valind, A.; Karlsson, J.; Gisselsson, D. DEVOLUTION-A method for phylogenetic reconstruction of aneuploid cancers based on multiregional genotyping data. *4*, 1103.

[139] Liu, J.; Halloran, J. T.; Bilmes, J. A.; Daza, R. M.; Lee, C.; Mahen, E. M.; Prunkard, D.; Song, C.; Blau, S.; Dorschner, M. O.; Gadi, V. K.; Shendure, J.; Blau, C. A.; Noble, W. S. Comprehensive statistical inference of the clonal structure of cancer from multiple biopsies. *7*, 16943.

[140] Jiang, Y.; Qiu, Y.; Minn, A. J.; Zhang, N. R. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *113*.

[141] El-Kebir, M.; Oesper, L.; Acheson-Field, H.; Raphael, B. J. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics* **2015**, *31*, i62–i70.

[142] Sengupta, S.; Wang, J.; Lee, J.; Müller, P.; Gulukota, K.; Banerjee, A.; Ji, Y. Bayclone: Bayesian nonparametric inference of tumor subclones using NGS data. Pacific Symposium on Biocomputing Co-Chairs. 2014; pp 467–478.

[143] Zare, H.; Wang, J.; Hu, A.; Weber, K.; Smith, J.; Nickerson, D.; Song, C.; Witten, D.; Blau, C. A.; Noble, W. S. Inferring clonal composition from multiple sections of a breast cancer. *PLoS computational biology* **2014**, *10*, e1003703.

[144] Fischer, A.; Vázquez-García, I.; Illingworth, C. J.; Mustonen, V. High-definition reconstruction of clonal composition in cancer. *Cell reports* **2014**, *7*, 1740–1752.

[145] Popic, V.; Salari, R.; Hajirasouliha, I.; Kashef-Haghighi, D.; West, R. B.; Batzoglou, S. Fast and scalable inference of multi-sample cancer lineages. *Genome biology* **2015**, *16*, 1–17.

[146] Kono, H.; Rock, K. L. How dying cells alert the immune system to danger. *8*, 279–289.

[147] Weaver, C.; Murphy, K. Janeway's immunobiology. *Garland Sci.* **2016**,

[148] Adams, J. L.; Smothers, J.; Srinivasan, R.; Hoos, A. Big opportunities for small molecules in immuno-oncology. *Nature reviews Drug discovery* **2015**, *14*, 603–622.

[149] Coley, W. B. The Therapeutic Value of the Mixed Toxins of the Streptococcus of Erysipelas and Bacillus Prodigiosus in the Treatment of Inoperable Malignant Tumors, with a Report of One Hundred and Sixty Cases.1: Bibliography. *112*, 251, Num Pages: 31 Place: Philadelphia, United States Publisher: American Periodicals Series II.

[150] McCarthy, E. F. The Toxins of William B. Coley and the Treatment of Bone and Soft-Tissue Sarcomas. *26*, 154–158.

[151] Schumacher, T. N.; Scheper, W.; Kvistborg, P. Cancer Neoantigens. *37*, 173–200, _eprint: https://doi.org/10.1146/annurev-immunol-042617-053402.

[152] Wolfel, T.; Hauer, M.; Schneider, J.; Serrano, M., *et al.* A p16INK4a-insensitive CDK4 mutant targeted by cytolytic T lymphocytes in a human melanoma. *Science* **1995**, *269*, 1281.

[153] Coulie, P. G.; Lehmann, F.; Lethe, B.; Herman, J.; Lurquin, C.; Andrawiss, M.; Boon, T. A mutated intron sequence codes for an antigenic peptide recognized by cytolytic T lymphocytes on a human melanoma. *Proceedings of the National Academy of Sciences* **1995**, *92*, 7976–7980.

[154] Matsushita, H.; Vesely, M. D.; Koboldt, D. C.; Rickert, C. G.; Uppaluri, R.; Magrini, V. J.; Arthur, C. D.; White, J. M.; Chen, Y.-S.; Shea, L. K., *et al.* Cancer exome analysis reveals a T-cell-dependent mechanism of cancer immunoediting. *Nature* **2012**, *482*, 400–404.

[155] Castle, J. C.; Kreiter, S.; Diekmann, J.; Löwer, M.; Van de Roemer, N.; de Graaf, J.; Selmi, A.; Diken, M.; Boegel, S.; Paret, C., *et al.* Exploiting the Mutanome for Tumor VaccinationB16 Melanoma T-cell–Druggable Mutanome. *Cancer research* **2012**, *72*, 1081–1091.

[156] Robbins, P. F.; Lu, Y.-C.; El-Gamil, M.; Li, Y. F.; Gross, C.; Gartner, J.; Lin, J. C.; Teer, J. K.; Cliften, P.; Tycksen, E.; Samuels, Y.; Rosenberg, S. A. Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. *19*, 747–752.

[157] van Rooij, N.; van Buuren, M. M.; Philips, D.; Velds, A.; Toebes, M.; Heemskerk, B.; van Dijk, L. J.; Behjati, S.; Hilkmann, H.; El Atmioui, D., *et al.* Tumor exome analysis reveals neoantigen-specific T-cell reactivity in an ipilimumab-responsive melanoma. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology* **2013**, *31*.

[158] Tran, E.; Robbins, P. F.; Rosenberg, S. A. 'Final common pathway'of human cancer immunotherapy: targeting random somatic mutations. *Nature immunology* **2017**, *18*, 255–262.

[159] Choo, S. Y. The HLA System: Genetics, Immunology, Clinical Testing, and Clinical Implications. *48*, 11–23.

[160] Terasaki, P. I.; Dausset, J. History of HLA: Ten recollections. *(No Title)* **1990**,

[161] Rock, K. L.; York, I. A.; Saric, T.; Goldberg, A. L. Protein degradation and the generation of MHC class I-presented peptides. **2002**,

[162] Hewitt, E. W. The MHC class I antigen presentation pathway: strategies for viral immune evasion. *110*, 163–169.

[163] Antoniou, A. N.; Ford, S.; Pilley, E. S.; Blake, N.; Powis, S. J. Interactions formed by individually expressed TAP1 and TAP2 polypeptide subunits. *Immunology* **2002**, *106*, 182–189.

[164] Matsumura, M.; Fremont, D. H.; Peterson, P. A.; Wilson, l. A. Emerging principles for the recognition of peptide antigens by MHC class I molecules. *Science* **1992**, *257*, 927–934.

[165] Wieczorek, M.; Abualrous, E. T.; Sticht, J.; Álvaro-Benito, M.; Stolzenberg, S.; Noé, F.; Freund, C. Major histocompatibility complex (MHC) class I and MHC class II proteins: conformational plasticity in antigen presentation. *Frontiers in immunology* **2017**, *8*, 292.

[166] Stumptner-Cuvelette, P.; Benaroch, P. Multiple roles of the invariant chain in MHC class II function. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* **2002**, *1542*, 1–13.

[167] Thibodeau, J.; Bourgeois-Daigneault, M.-C.; Lapointe, R. Targeting the MHC Class II antigen presentation pathway in cancer immunotherapy. *1*, 908–916.

[168] Zniga-Pflucker, J. C. T-cell development made simple. *4*, 67–72, Number: 1 Publisher: Nature Publishing Group.

[169] Cantrell, D. A. T-cell antigen receptor signal transduction. *105*, 369–374.

[170] Mason, D. A very high level of crossreactivity is an essential feature of the T-cell receptor. *Immunology today* **1998**, *19*, 395–404.

[171] Ehrlich, P. Ueber den jetzigen stand der karzinomforschung. vortrag gehalten vor den studenten der amsterdamer universitaet, vereinigung fuer wissenschaftliche arbeit 1 june 1908. printed in: P. ehrlich. *Beitraege zur Experimentellen Pathologie und Chemotherapie, Akademische Verlagsgesellschaft, Leipzig* **1909**, 118–164.

[172] Burnet, M. Cancer-A Biological Approach. *1*, 841–847.

[173] Thomas, L.; Lawrence, H. Cellular and humoral aspects of the hypersensitive states. *New York: Hoeber-Harper* **1959**, 529–32.

[174] Dunn, G. P.; Bruce, A. T.; Ikeda, H.; Old, L. J.; Schreiber, R. D. Cancer immunoediting: from immunosurveillance to tumor escape. *3*, 991–998.

[175] Gross, L. Intradermal Immunization of C3H Mice against a Sarcoma That Originated in an Animal of the Same Line. *3*, 326–333.

[176] Prehn, R. T.; Main, J. M. Immunity to methylcholanthrene-induced sarcomas. *18*, 769–778.

[177] Shinkai, Y.; Rathbun, G.; Lam, K.-P.; Oltz, E. M.; Stewart, V.; Mendelsohn, M.; Charron, J.; Datta, M.; Young, F.; Stall, A. M.; Alt, F. W. RAG-2-deficient mice lack mature lymphocytes owing to inability to initiate V(D)J rearrangement. *68*, 855–867.

[178] Shankaran, V.; Ikeda, H.; Bruce, A. T.; White, J. M.; Swanson, P. E.; Old, L. J.; Schreiber, R. D. IFNγ and lymphocytes prevent primary tumour development and shape tumour immunogenicity. *Nature* **2001**, *410*, 1107–1111.

[179] Christie, S. M.; Fijen, C.; Rothenberg, E. V(D)J Recombination: Recent Insights in Formation of the Recombinase Complex and Recruitment of DNA Repair Machinery. *10*.

[180] Gatti, R. A.; Good, R. A. Occurrence of malignancy in immunodeficiency diseases. A literature review. *28*, 89–98.

[181] Salavoura, K.; Kolialexi, A.; Tsangaris, G.; Mavrou, A. Development of Cancer in Patients with Primary Immunodeficiencies. *28*, 1263–1269, Publisher: International Institute of Anticancer Research Section: Clinical Studies.

[182] Kebudi, R.; Kiykim, A.; Sahin, M. K. Primary immunodeficiency and cancer in children; a review of the literature. *Current pediatric reviews* **2019**, *15*, 245–250.

[183] Boshoff, C.; Weiss, R. Aids-related malignancies. *2*, 373–382, Number: 5 Publisher: Nature Publishing Group.

[184] Clifford, G. M.; Polesel, J.; Rickenbach, M.; Dal Maso, L.; Keiser, O.; Kofler, A.; Rapiti, E.; Levi, F.; Jundt, G.; Fisch, T.; Bordoni, A.; De Weck, D.; Franceschi, S.; on behalf of the Swiss HIV Cohort Study, Cancer Risk in the Swiss HIV Cohort Study: Associations With Immunodeficiency, Smoking, and Highly Active Antiretroviral Therapy. *97*, 425–432.

[185] Birkeland, S. A. *et al.* Cancer risk after renal transplantation in the nordic countries, 1964-1986. *60*, 183–189, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ijc.2910600209.

[186] Huo, Z.; Li, C.; Xu, X.; Ge, F.; Wang, R.; Wen, Y.; Peng, H.; Wu, X.; Liang, H.; Peng, G., *et al.* Cancer risks in solid organ transplant recipients: results from a comprehensive analysis of 72 cohort studies. *Oncoimmunology* **2020**, *9*, 1848068.

[187] Jung, S. W.; Lee, H.; Cha, J. M. Risk of malignancy in kidney transplant recipients: a nationwide population-based cohort study. *23*, 160.

[188] Sherston, S. N.; Carroll, R. P.; Harden, P. N.; Wood, K. J. Predictors of Cancer Risk in the Long-Term Solid-Organ Transplant Recipient. *97*, 605.

[189] Adami, J.; Gabel, H.; Lindelf, B.; Ekstrm, K.; Rydh, B.; Glimelius, B.; Ekbom, A.; Adami, H.-O.; Granath, F. Cancer risk following organ transplantation: a nationwide cohort study in Sweden. *89*, 1221–1227, Number: 7 Publisher: Nature Publishing Group.

[190] Haagsma, E. B.; Hagens, V. E.; Schaapveld, M.; van den Berg, A. P.; de Vries, E. G. E.; Klompmaker, I. J.; Slooff, M. J. H.; Jansen, P. L. M. Increased cancer risk after liver transplantation: a population-based study. *34*, 84–91.

[191] Bavinck, J. N. B.; Hardie, D. R.; Green, A.; Cutmore, S.; MacNaught, A.; O'Sullivan, B.; Siskind, V.; Van Der Woude, F. J.; Hardie, I. R. THE RISK OF SKIN CANCER IN RENAL TRANSPLANT RECIPIENTS IN QUEENSLAND, AUSTRALIA: A Follow-up Study: 1. *Transplantation* **1996**, *61*, 715–721.

[192] Krynitz, B.; Edgren, G.; Lindelöf, B.; Baecklund, E.; Brattström, C.; Wilczek, H.; Smedby, K. E. Risk of skin cancer and other malignancies in kidney, liver, heart and lung transplant recipients 1970 to 2008—a Swedish population-based study. *International journal of cancer* **2013**, *132*, 1429–1438.

[193] Corthay, A. Does the Immune System Naturally Protect Against Cancer? *5*.

[194] Engels, E. A. *et al.* Spectrum of Cancer Risk Among US Solid Organ Transplant Recipients. *306*, 1891–1901.

[195] Shiels, M. S.; Cole, S. R.; Kirk, G. D.; Poole, C. A meta-analysis of the incidence of non-AIDS cancers in HIV-infected individuals. *Journal of acquired immune deficiency syndromes (1999)* **2009**, *52*, 611.

[196] Sigel, K.; Wisnivesky, J.; Gordon, K.; Dubrow, R.; Justice, A.; Brown, S. T.; Goulet, J.; Butt, A. A.; Crystal, S.; Rimland, D., *et al.* HIV as an independent risk factor for incident lung cancer. *AIDS (London, England)* **2012**, *26*, 1017.

[197] Friman, T. K.; Jaamaa-Holmberg, S.; berg, F.; Helantera, I.; Halme, M.; Pentikainen, M. O.; Nordin, A.; Lemstrm, K. B.; Jahnukainen, T.; Raty, R.; Salmela, B. Cancer risk and mortality after solid organ transplantation: A population-based 30-year cohort study in Finland. *150*, 1779–1791, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ijc.33934.

[198] Koebel, C. M.; Vermi, W.; Swann, J. B.; Zerafa, N.; Rodig, S. J.; Old, L. J.; Smyth, M. J.; Schreiber, R. D. Adaptive immunity maintains occult cancer in an equilibrium state. *450*, 903–907, Number: 7171 Publisher: Nature Publishing Group.

[199] Dunn, G. P.; Old, L. J.; Schreiber, R. D. The Three Es of Cancer Immunoediting. *22*, 329–360, _eprint: https://doi.org/10.1146/annurev.immunol.22.012703.104803.

[200] Cicchese, J. M.; Evans, S.; Hult, C.; Joslyn, L. R.; Wessler, T.; Millar, J. A.; Marino, S.; Cilfone, N. A.; Mattila, J. T.; Linderman, J. J., *et al.* Dynamic balance of pro-and anti-inflammatory signals controls disease and limits pathology. *Immunological reviews* **2018**, *285*, 147–167.

[201] Zapata, L.; Pich, O.; Serrano, L.; Kondrashov, F. A.; Ossowski, S.; Schaefer, M. H. Negative selection in tumor genome evolution acts on essential cellular functions and the immunopeptidome. *19*, 67.

[202] Zapata, L.; Caravagna, G.; Williams, M. J.; Lakatos, E.; AbdulJabbar, K.; Werner, B.; Chowell, D.; James, C.; Gourmet, L.; Milite, S.; Acar, A.; Riaz, N.; Chan, T. A.; Graham, T. A.; Sottoriva, A. Immune selection determines tumor antigenicity and influences response to checkpoint inhibitors. *55*, 451–460, Number: 3 Publisher: Nature Publishing Group.

[203] Rooney, M. S.; Shukla, S. A.; Wu, C. J.; Getz, G.; Hacohen, N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *160*, 48–61.

[204] Luksza, M.; Riaz, N.; Makarov, V.; Balachandran, V. P.; Hellmann, M. D.; Solovyov, A.; Rizvi, N. A.; Merghoub, T.; Levine, A. J.; Chan, T. A.; Wolchok, J. D.; Greenbaum, B. D. A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *551*, 517–520, Number: 7681 Publisher: Nature Publishing Group.

[205] Marty, R.; Kaabinejadian, S.; Rossell, D.; Slifker, M. J.; van de Haar, J.; Engin, H. B.; de Prisco, N.; Ideker, T.; Hildebrand, W. H.; Font-Burgada, J.; Carter, H. MHC-I Genotype Restricts the Oncogenic Mutational Landscape. *171*, 1272–1283.e15.

[206] Marty Pyke, R.; Thompson, W. K.; Salem, R. M.; Font-Burgada, J.; Zanetti, M.; Carter, H. Evolutionary Pressure against MHC Class II Binding Cancer Mutations. *175*, 416–428.e13.

[207] Kherreh, N.; Cleary, S.; Seoighe, C. No evidence that HLA genotype influences the driver mutations that occur in cancer patients. *Cancer Immunology, Immunotherapy* **2022**, *71*, 819–827.

[208] Claeys, A.; Luijts, T.; Marchal, K.; Van den Eynden, J. Low immunogenicity of common cancer hot spot mutations resulting in false immunogenic selection signals. *17*, e1009368.

[209] Wang, S.; Wang, X.; Wu, T.; He, Z.; Li, H.; Sun, X.; Liu, X.-S. Revisiting neoantigen depletion signal in the untreated cancer genome.

[210] Wu, T.; Wang, G.; Wang, X.; Wang, S.; Zhao, X.; Wu, C.; Ning, W.; Tao, Z.; Chen, F.; Liu, X.-S. Quantification of Neoantigen-Mediated Immunoediting in Cancer Evolution. *82*, 2226–2238.

[211] Claeys, A.; Van den Eynden, J. Quantification of Neoantigen-Mediated Immunoediting in Cancer Evolution-Letter. *83*, 971–972.

[212] Wu, T.; Diao, K.; Liu, X.-S. Quantification of Neoantigen-Mediated Immunoediting in Cancer Evolution-Reply. *83*, 973.

[213] Sharpe, A. H. Introduction to checkpoint inhibitors and cancer immunotherapy. *276*, 5–8.

[214] He, X.; Xu, C. Immune checkpoint signaling and cancer immunotherapy. *30*, 660–669, Number: 8 Publisher: Nature Publishing Group.

[215] Maldonado, R. A.; von Andrian, U. H. How tolerogenic dendritic cells induce regulatory T cells. *Advances in immunology* **2010**, *108*, 111–165.

[216] Seliger, B.; Massa, C. The dark side of dendritic cells: development and exploitation of tolerogenic activity that favor tumor outgrowth and immune escape. *Frontiers in immunology* **2013**, *4*, 419.

[217] Hanahan, D.; Weinberg, R. A. Hallmarks of cancer: the next generation. *144*, 646–674.

[218] Lee, J. H.; Shklovskaya, E.; Lim, S. Y.; Carlino, M. S.; Menzies, A. M.; Stewart, A.; Pedersen, B.; Irvine, M.; Alavi, S.; Yang, J. Y., *et al.* Transcriptional downregulation of MHC class I and melanoma de-differentiation in resistance to PD-1 inhibition. *Nature communications* **2020**, *11*, 1897.

[219] Zhao, Y.; Cao, Y.; Chen, Y.; Wu, L.; Hang, H.; Jiang, C.; Zhou, X. B2M gene expression shapes the immune landscape of lung adenocarcinoma and determines the response to immunotherapy. *164*, 507–523, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/imm.13384.

[220] del Campo, A. B.; Kyte, J. A.; Carretero, J.; Zinchencko, S.; Mendez, R.; Gonzalez-Aseguinolaza, G.; Ruiz-Cabello, F.; Aamdal, S.; Gaudernack, G.; Garrido, F.; Aptsiauri, N. Immune escape of cancer cells with beta2-microglobulin loss over the course of metastatic melanoma. *134*, 102–113, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ijc.28338.

[221] Rouas-Freiss, N.; Gonçalves, R. M.-B.; Menier, C.; Dausset, J.; Carosella, E. D. Direct evidence to support the role of HLA-G in protecting the fetus from maternal uterine natural killer cytolysis. *Proceedings of the National Academy of Sciences* **1997**, *94*, 11520–11525.

[222] Loustau, M.; Anna, F.; Drean, R.; Lecomte, M.; Langlade-Demoyen, P.; Caumartin, J. HLA-G Neo-Expression on Tumors. *11*, 1685.

[223] Ling, A.; Lfgren-Burstrm, A.; Larsson, P.; Li, X.; Wikberg, M. L.; Oberg, A.; Stenling, R.; Edin, S.; Palmqvist, R. TAP1 down-regulation elicits immune escape and poor prognosis in colorectal cancer. *6*, e1356143, Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/2162402X.2017.1356143.

[224] Coelho, M. A.; de Carné Trécesson, S.; Rana, S.; Zecchin, D.; Moore, C.; Molina-Arcas, M.; East, P.; Spencer-Dene, B.; Nye, E.; Barnouin, K., *et al.* Oncogenic RAS signaling promotes tumor immunoresistance by stabilizing PD-L1 mRNA. *Immunity* **2017**, *47*, 1083–1099.

[225] Martinez-Gonzalez, L. J.; Pascual Geler, M.; Robles Fernandez, I.; Cozar, J. M.; Lorente, J. A.; Alvarez Cubero, M. J. Improving the genetic signature of prostate cancer, the somatic mutations. *36*, 312.e17–312.e23.

[226] Maeda, T.; Hiraki, M.; Jin, C.; Rajabi, H.; Tagde, A.; Alam, M.; Bouillez, A.; Hu, X.; Suzuki, Y.; Miyo, M., *et al.* MUC1-C induces PD-L1 and immune evasion in triple-negative breast cancer. *Cancer research* **2018**, *78*, 205–215.

[227] Ohue, Y.; Nishikawa, H. Regulatory T (Treg) cells in cancer: Can Treg cells be a new therapeutic target? *110*, 2080–2089.

[228] Yokokawa, J.; Cereda, V.; Remondo, C.; Gulley, J. L.; Arlen, P. M.; Schlom, J.; Tsang, K. Y. Enhanced functionality of CD4+CD25(high)FoxP3+ regulatory T cells in the peripheral blood of patients with prostate cancer. *14*, 1032–1040.

[229] Gasparoto, T. H.; de Souza Malaspina, T. S.; Benevides, L.; de Melo, E. J. F.; Costa, M. R. S. N.; Damante, J. H.; Ikoma, M. R. V.; Garlet, G. P.;

Cavassani, K. A.; da Silva, J. S.; Campanelli, A. P. Patients with oral squamous cell carcinoma are characterized by increased frequency of suppressive regulatory T cells in the blood and tumor microenvironment. *59*, 819–828.

[230] Xia, A.; Zhang, Y.; Xu, J.; Yin, T.; Lu, X.-J. T Cell Dysfunction in Cancer Immunity and Immunotherapy. *10*.

[231] Beatty, G. L.; Gladney, W. L. Immune escape mechanisms as a guide for cancer ImmunotherapyTailoring cancer immunotherapy. *Clinical cancer research* **2015**, *21*, 687–692.

[232] Coulie, P. G.; Van den Eynde, B. J.; van der Bruggen, P.; Boon, T. Tumour antigens recognized by T lymphocytes: at the core of cancer immunotherapy. *14*, 135–146, Number: 2 Publisher: Nature Publishing Group.

[233] Van den Eynden, J.; Jimenez-Sanchez, A.; Miller, M. L.; Larsson, E. Lack of detectable neoantigen depletion signals in the untreated cancer genome. *51*, 1741–1748, Number: 12 Publisher: Nature Publishing Group.

[234] Tran, E.; Robbins, P. F.; Lu, Y.-C.; Prickett, T. D.; Gartner, J. J.; Jia, L.; Pasetto, A.; Zheng, Z.; Ray, S.; Groh, E. M., *et al.* T-cell transfer therapy targeting mutant KRAS in cancer. *New England Journal of Medicine* **2016**, *375*, 2255–2262.

[235] McGranahan, N.; Rosenthal, R.; Hiley, C. T.; Rowan, A. J.; Watkins, T. B. K.; Wilson, G. A.; Birkbak, N. J.; Veeriah, S.; Loo, P. V.; Herrero, J.; Swanton, C.; Consortium, t. T. Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *171*, 1259, Publisher: Elsevier.

[236] Ribas, A.; Wolchok, J. D. Cancer Immunotherapy Using Checkpoint Blockade. *359*, 1350–1355.

[237] Wang, Y.; Tong, Z.; Zhang, W.; Zhang, W.; Buzdin, A.; Mu, X.; Yan, Q.; Zhao, X.; Chang, H.-H.; Duhon, M.; Zhou, X.; Zhao, G.; Chen, H.; Li, X. FDA-Approved and Emerging Next Generation Predictive Biomarkers for Immune Checkpoint Inhibitors in Cancer Patients. *11*, 683419.

[238] Kerr, W. G.; Chisholm, J. D. The next generation of immunotherapy for cancer: Small molecules could make big waves. *202*, 11–19.

[239] Gascon, M.; Isla, D.; Cruellas, M.; Galvez, E. M.; Lastra, R.; Ocariz, M.; Pano, J. R.; Ramirez, A.; Sesma, A.; Torres-Ramon, I.; Yubero, A.; Pardo, J.; Martinez-Lostao, L. Intratumoral versus Circulating Lymphoid Cells as Predictive Biomarkers in Lung Cancer Patients Treated with Immune Checkpoint Inhibitors: Is the Easiest Path the Best One? *9*, 1525, Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

[240] Han, Y.; Liu, D.; Li, L. PD-1/PD-L1 pathway: current researches in cancer. *10*, 727–742.

[241] Topalian, S. L.; Hodi, F. S.; Brahmer, J. R.; Gettinger, S. N.; Smith, D. C.; McDermott, D. F.; Powderly, J. D.; Carvajal, R. D.; Sosman, J. A.; Atkins, M. B., *et al.* Safety, activity, and immune correlates of anti–PD-1 antibody in cancer. *New England Journal of Medicine* **2012**, *366*, 2443–2454.

[242] Garon, E. B. *et al.* Pembrolizumab for the treatment of non-small-cell lung cancer. *372*, 2018–2028.

[243] Ribas, A.; Tumeh, P. C. The Future of Cancer Therapy: Selecting Patients Likely to Respond to PD1/L1 BlockadeSelecting Patients for PD1/L1 Blockade Therapy. *Clinical Cancer Research* **2014**, *20*, 4982–4984.

[244] Jorgensen, J. T. Companion diagnostic assays for PD-1/PD-L1 checkpoint inhibitors in NSCLC. *16*, 131–133, Publisher: Taylor & Francis _eprint: https://doi.org/10.1586/14737159.2016.1117389.

[245] Keenan, T. E.; Burke, K. P.; Van Allen, E. M. Genomic correlates of response to immune checkpoint blockade. *25*, 389–402, Number: 3 Publisher: Nature Publishing Group.

[246] Davis, A. A.; Patel, V. G. The role of PD-L1 expression as a predictive biomarker: an analysis of all US Food and Drug Administration (FDA) approvals of immune checkpoint inhibitors. *7*, 278.

[247] Mansfield, A. S.; Aubry, M. C.; Moser, J. C.; Harrington, S. M.; Dronca, R. S.; Park, S. S.; Dong, H. Temporal and spatial discordance of programmed cell death-ligand 1 expression and lymphocyte tumor infiltration between paired primary lesions and brain metastases in lung cancer. *27*, 1953–1958, Publisher: Elsevier.

[248] Zhang, J.; Dang, F.; Ren, J.; Wei, W. Biochemical aspects of PD-L1 regulation in cancer immunotherapy. *Trends in biochemical sciences* **2018**, *43*, 1014–1032.

[249] Yang, F.; Wang, J. F.; Wang, Y.; Liu, B.; Molina, J. R. Comparative Analysis of Predictive Biomarkers for PD-1/PD-L1 Inhibitors in Cancers: Developments and Challenges. *14*, 109.

[250] Roberts, S. A.; Gordenin, D. A. Hypermutation in human cancer genomes: footprints and mechanisms. *14*, 786–800.

[251] Subramanian, S.; Mishra, R. K.; Singh, L. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *4*, R13.

[252] Marcus, L.; Lemery, S. J.; Keegan, P.; Pazdur, R. FDA Approval Summary: Pembrolizumab for the Treatment of Microsatellite Instability-High Solid Tumors. *25*, 3753–3758.

[253] Llosa, N. J. *et al.* The vigorous immune microenvironment of microsatellite instable colon cancer is balanced by multiple counter-inhibitory checkpoints. *5*, 43–51.

[254] Yi, M.; Jiao, D.; Xu, H.; Liu, Q.; Zhao, W.; Han, X.; Wu, K. Biomarkers for predicting efficacy of PD-1/PD-L1 inhibitors. *17*, 129.

[255] Klempner, S. J.; Fabrizio, D.; Bane, S.; Reinhart, M.; Peoples, T.; Ali, S. M.; Sokol, E. S.; Frampton, G.; Schrock, A. B.; Anhorn, R.; Reddy, P. Tumor Mutational Burden as a Predictive Biomarker for Response to Immune Checkpoint Inhibitors: A Review of Current Evidence. *25*, e147–e159.

[256] Melendez, B.; Campenhout, C. V.; Rorive, S.; Remmelink, M.; Salmon, I.; D'Haene, N. Methods of measurement for tumor mutational burden in tumor tissue. *7*, Publisher: AME Publishing Company.

[257] Allgauer, M. *et al.* Implementing tumor mutational burden (TMB) analysis in routine diagnostics-a primer for molecular pathologists and clinicians. *7*, Publisher: AME Publishing Company.

[258] Chabanon, R. M.; Pedrero, M.; Lefebvre, C.; Marabelle, A.; Soria, J.-C.; Postel-Vinay, S. Mutational Landscape and Sensitivity to Immune Checkpoint Blockers. *22*, 4309–4321.

[259] Hellmann, M. D. *et al.* Genomic Features of Response to Combination Immunotherapy in Patients with Advanced Non-Small-Cell Lung Cancer. *33*, 843–852.e4.

[260] McGrail, D. J. *et al.* High tumor mutation burden fails to predict immune checkpoint blockade response across all cancer types. *32*, 661–672.

[261] McGranahan, N.; Swanton, C. Neoantigen quality, not quantity. *11*, eaax7918.

[262] Gejman, R. S.; Chang, A. Y.; Jones, H. F.; DiKun, K.; Hakimi, A. A.; Schietinger, A.; Scheinberg, D. A. Rejection of immunogenic tumor clones is limited by clonal fraction. *7*, e41090, Publisher: eLife Sciences Publications, Ltd.

[263] McGranahan, N. *et al.* Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *351*, 1463–1469, Publisher: American Association for the Advancement of Science.

[264] Rosenthal, R. *et al.* Neoantigen-directed immune escape in lung cancer evolution. *567*, 479–485, Number: 7749 Publisher: Nature Publishing Group.

[265] Chowell, D. *et al.* Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. *359*, 582–587, Publisher: American Association for the Advancement of Science.

[266] Turajlic, S. *et al.* Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. *18*, 1009–1021.

[267] Niknafs, N. *et al.* Persistent mutation burden drives sustained anti-tumor immune responses. *29*, 440–449, Number: 2 Publisher: Nature Publishing Group.

[268] Knight, J. C. Allele-specific gene expression uncovered. *Trends in Genetics* **2004**, *20*, 113–116.

[269] Buckland, P. R. Allele-specific gene expression differences in humans. *Human molecular genetics* **2004**, *13*, R255–R260.

[270] Gregg, C. Known unknowns for allele-specific expression and genomic imprinting effects. *F1000Prime Reports* **2014**, *6*.

[271] Lee, M. P. Allele-specific gene expression and epigenetic modifications and their application to understanding inheritance and cancer. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* **2012**, *1819*, 739–742.

[272] Zhang, S.; Zhang, H.; Zhou, Y.; Qiao, M.; Zhao, S.; Kozlova, A.; Shi, J.; Sanders, A. R.; Wang, G.; Luo, K., *et al.* Allele-specific open chromatin in human iPSC neurons elucidates functional disease variants. *Science* **2020**, *369*, 561–565.

[273] Lo, H. S.; Wang, Z.; Hu, Y.; Yang, H. H.; Gere, S.; Buetow, K. H.; Lee, M. P. Allelic variation in gene expression is common in the human genome. *Genome research* **2003**, *13*, 1855–1862.

[274] Reddy, T. E.; Gertz, J.; Pauli, F.; Kucera, K. S.; Varley, K. E.; Newberry, K. M.; Marinov, G. K.; Mortazavi, A.; Williams, B. A.; Song, L., *et al.* Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome research* **2012**, *22*, 860–869.

[275] Cavalli, M.; Pan, G.; Nord, H.; Arzt, E. W.; Wallerman, O.; Wadelius, C. Allele-specific transcription factor binding in liver and cervix cells unveils many likely drivers of GWAS signals. *Genomics* **2016**, *107*, 248–254.

[276] Cavalli, M.; Pan, G.; Nord, H.; Arzt, E. W.; Wallerman, O.; Wadelius, C. Allele specific chromatin signals, 3D interactions, and motif predictions for immune and B cell related diseases. *Scientific reports* **2019**, *9*, 1–14.

[277] Hug, N.; Longman, D.; Cáceres, J. F. Mechanism and regulation of the nonsense-mediated decay pathway. *Nucleic acids research* **2016**, *44*, 1483–1495.

[278] Kervestin, S.; Jacobson, A. NMD: a multifaceted response to premature translational termination. *Nature reviews Molecular cell biology* **2012**, *13*, 700–712.

[279] Alonso, C. R. Nonsense-mediated RNA decay: A molecular system micromanaging individual gene activities and suppressing genomic noise. *Bioessays* **2005**, *27*, 463–466.

[280] Montgomery, S. B.; Lappalainen, T.; Gutierrez-Arcelus, M.; Dermitzakis, E. T. Rare and common regulatory variation in population-scale sequenced human genomes. *PLoS Genet* **2011**, *7*, e1002144.

[281] Nembaware, V.; Lupindo, B.; Schouest, K.; Spillane, C.; Scheffler, K.; Seoighe, C. Genome-wide survey of allele-specific splicing in humans. *BMC genomics* **2008**, *9*, 265.

[282] Park, E.; Pan, Z.; Zhang, Z.; Lin, L.; Xing, Y. The expanding landscape of alternative splicing variation in human populations. *The American Journal of Human Genetics* **2018**, *102*, 11–26.

[283] Nembaware, V.; Wolfe, K. H.; Bettoni, F.; Kelso, J.; Seoighe, C. Allele-specific transcript isoforms in human. *FEBS letters* **2004**, *577*, 233–238.

[284] Li, Y. I.; Van De Geijn, B.; Raj, A.; Knowles, D. A.; Petti, A. A.; Golan, D.; Gilad, Y.; Pritchard, J. K. RNA splicing is a primary link between genetic variation and disease. *Science* **2016**, *352*, 600–604.

[285] Sheinberger, J.; Hochberg, H.; Lavi, E.; Kanter, I.; Avivi, S.; Reinitz, G.; Schwed, A.; Aizler, Y.; Varon, E.; Kinor, N., *et al.* CD-tagging-MS2: detecting allelic expression of endogenous mRNAs and their protein products in single cells. *Biology Methods and Protocols* **2017**, *2*, bpx004.

[286] Yang, E.-W.; Bahn, J. H.; Hsiao, E. Y.-H.; Tan, B. X.; Sun, Y.; Fu, T.; Zhou, B.; Van Nostrand, E. L.; Pratt, G. A.; Freese, P., *et al.* Allele-specific binding of RNA-binding proteins reveals functional genetic variants in the RNA. *Nature communications* **2019**, *10*, 1–15.

[287] Bahrami-Samani, E.; Xing, Y. Discovery of allele-specific protein-RNA interactions in human transcriptomes. *The American Journal of Human Genetics* **2019**, *104*, 492–502.

[288] Messemaker, T. C.; van Leeuwen, S. M.; van den Berg, P. R.; EJ‘t Jong, A.; Palstra, R.-J.; Hoeben, R. C.; Semrau, S.; Mikkers, H. M. Allele-specific repression of Sox2 through the long non-coding RNA Sox2ot. *Scientific reports* **2018**, *8*, 1–13.

[289] Võsa, U.; Esko, T.; Kasela, S.; Annilo, T. Altered gene expression associated with microRNA binding site polymorphisms. *PloS one* **2015**, *10*, e0141351.

[290] Johnsson, P. A.; Hartmanis, L.; Ziegenhain, C.; Hendriks, G.-J.; Hagemann-Jensen, M.; Reinius, B.; Sandberg, R. Deducing transcriptional kinetics and molecular functions of long non-coding RNAs using allele-sensitive single-cell RNA-sequencing. *bioRxiv* **2020**,

[291] Kim, J.; Bartel, D. P. Allelic imbalance sequencing reveals that single-nucleotide polymorphisms frequently alter microRNA-directed repression. *Nature biotechnology* **2009**, *27*, 472–477.

[292] Cleary, S.; Seoighe, C. Perspectives on allele-specific expression. *Annual Review of Biomedical Data Science* **2021**, *4*, 101–122.

[293] Soderlund, C. A.; Nelson, W. M.; Goff, S. A. Allele Workbench: transcriptome pipeline and interactive graphics for allele-specific expression. *PloS one* **2014**, *9*, e115740.

[294] Van De Geijn, B.; McVicker, G.; Gilad, Y.; Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nature methods* **2015**, *12*, 1061–1063.

[295] Dumont, E. L.; Tycko, B.; Do, C. CloudASM: an ultra-efficient cloud-based pipeline for mapping allele-specific DNA methylation. *Bioinformatics* **2020**, *36*, 3558–3560.

[296] Younesy, H.; Möller, T.; Heravi-Moussavi, A.; Cheng, J. B.; Costello, J. F.; Lorincz, M. C.; Karimi, M. M.; Jones, S. J. ALEA: a toolbox for allele-specific epigenomics analysis. *Bioinformatics* **2014**, *30*, 1172–1174.

[297] Pastinen, T. Genome-wide allele-specific analysis: insights into regulatory variation. *Nature Reviews Genetics* **2010**, *11*, 533–538.

[298] Lappalainen, T.; Sammeth, M.; Friedländer, M. R.; Ac't Hoen, P.; Monlong, J.; Rivas, M. A.; Gonzalez-Porta, M.; Kurbatova, N.; Griebel, T.; Ferreira, P. G., *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **2013**, *501*, 506–511.

[299] Consortium, G., *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **2015**, *348*, 648–660.

[300] Castel, S. E.; Levy-Moonshine, A.; Mohammadi, P.; Banks, E.; Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome biology* **2015**, *16*, 195.

[301] Degner, J. F.; Marioni, J. C.; Pai, A. A.; Pickrell, J. K.; Nkadori, E.; Gilad, Y.; Pritchard, J. K. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **2009**, *25*, 3207–3212.

[302] Lee, C.; Kang, E. Y.; Gandal, M. J.; Eskin, E.; Geschwind, D. H. Profiling allele-specific gene expression in brains from individuals with autism spectrum disorder reveals preferential minor allele usage. *Nature neuroscience* **2019**, *22*, 1521–1532.

[303] Rozowsky, J.; Abyzov, A.; Wang, J.; Alves, P.; Raha, D.; Harmanci, A.; Leng, J.; Bjornson, R.; Kong, Y.; Kitabayashi, N., *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Molecular systems biology* **2011**, *7*, 522.

[304] Raghupathy, N.; Choi, K.; Vincent, M. J.; Beane, G. L.; Sheppard, K. S.; Munger, S. C.; Korstanje, R.; Pardo-Manual de Villena, F.; Churchill, G. A. Hierarchical analysis of RNA-seq reads improves the accuracy of allele-specific expression. *Bioinformatics* **2018**, *34*, 2177–2184.

[305] Wood, D. L.; Nones, K.; Steptoe, A.; Christ, A.; Harliwong, I.; Newell, F.; Bruxner, T. J.; Miller, D.; Cloonan, N.; Grimmond, S. M. Recommendations for accurate resolution of gene and isoform allele-specific expression in RNA-Seq data. *PloS one* **2015**, *10*, e0126911.

[306] Pandey, R. V.; Franssen, S. U.; Futschik, A.; Schlötterer, C. Allelic imbalance metre (A llim), a new tool for measuring allele-specific gene expression with RNA-seq data. *Molecular ecology resources* **2013**, *13*, 740–745.

[307] Dobin, A.; Davis, C. A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T. R. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **2013**, *29*, 15–21.

[308] Miao, Z.; Alvarez, M.; Pajukanta, P.; Ko, A. ASElux: an ultra-fast and accurate allelic reads counter. *Bioinformatics* **2018**, *34*, 1313–1320.

[309] Manske, H. M.; Kwiatkowski, D. P. SNP-o-matic. *Bioinformatics* **2009**, *25*, 2434–2435.

[310] Harvey, C. T.; Moyerbrailean, G. A.; Davis, G. O.; Wen, X.; Luca, F.; Pique-Regi, R. QuASAR: quantitative allele-specific analysis of reads. *Bioinformatics* **2015**, *31*, 1235–1242.

[311] Kumasaka, N.; Knights, A. J.; Gaffney, D. J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nature genetics* **2016**, *48*, 206–213.

[312] Deonovic, B.; Wang, Y.; Weirather, J.; Wang, X.-J.; Au, K. F. IDP-ASE: haplotyping and quantifying allele-specific expression at the gene and gene isoform level by hybrid sequencing. *Nucleic acids research* **2017**, *45*, e32–e32.

[313] Castel, S. E.; Mohammadi, P.; Chung, W. K.; Shen, Y.; Lappalainen, T. Rare variant phasing and haplotypic expression from RNA sequencing with phASER. *Nature communications* **2016**, *7*, 1–6.

[314] Fan, J.; Hu, J.; Xue, C.; Zhang, H.; Susztak, K.; Reilly, M. P.; Xiao, R.; Li, M. ASEP: Gene-based detection of allele-specific expression across individuals in a population by RNA sequencing. *PLoS Genetics* **2020**, *16*, e1008786.

[315] Xie, J.; Ji, T.; Ferreira, M. A.; Li, Y.; Patel, B. N.; Rivera, R. M. Modeling allele-specific expression at the gene and SNP levels simultaneously by a Bayesian logistic mixed regression model. *BMC bioinformatics* **2019**, *20*, 530.

[316] Liu, Z.; Gui, T.; Wang, Z.; Li, H.; Fu, Y.; Dong, X.; Li, Y. cisASE: a likelihood-based method for detecting putative cis-regulated allele-specific expression in RNA sequencing data. *Bioinformatics* **2016**, *32*, 3291–3297.

[317] Edsgärd, D.; Iglesias, M. J.; Reilly, S.-J.; Hamsten, A.; Tornvall, P.; Odeberg, J.; Emanuelsson, O. GeneiASE: Detection of condition-dependent and static allele-specific expression from RNA-seq data without haplotype information. *Scientific reports* **2016**, *6*, 21134.

[318] Dong, L.; Wang, J.; Wang, G. BYASE: A Python library for estimating gene and isoform level allele-specific expression. *Bioinformatics* **2020**,

[319] Knowles, D. A.; Davis, J. R.; Edgington, H.; Raj, A.; Favé, M.-J.; Zhu, X.; Potash, J. B.; Weissman, M. M.; Shi, J.; Levinson, D. F., *et al.* Allele-specific expression reveals interactions between genetic variation and environment. *Nature Methods* **2017**, *14*, 699–702.

[320] Oliva, M.; Muñoz-Aguirre, M.; Kim-Hellmuth, S.; Wucher, V.; Gewirtz, A. D.; Cotter, D. J.; Parsana, P.; Kasela, S.; Balliu, B.; Viñuela, A., *et al.* The impact of sex on gene expression across human tissues. *Science* **2020**, *369*.

[321] de Santiago, I.; Liu, W.; Yuan, K.; O'Reilly, M.; Chilamakuri, C. S. R.; Ponder, B. A.; Meyer, K. B.; Markowetz, F. BaalChIP: Bayesian analysis of allele-specific transcription factor binding in cancer genomes. *Genome biology* **2017**, *18*, 39.

[322] Calabrese, C. *et al.* Genomic basis for RNA alterations in cancer. *578*, 129–136, Number: 7793 Publisher: Nature Publishing Group.

[323] Li, G.; Bahn, J. H.; Lee, J.-H.; Peng, G.; Chen, Z.; Nelson, S. F.; Xiao, X. Identification of allele-specific alternative mRNA processing via transcriptome sequencing. *Nucleic acids research* **2012**, *40*, e104–e104.

[324] Bielski, C. M. *et al.* Widespread Selection for Oncogenic Mutant Allele Imbalance in Cancer. *34*, 852–862.e4, Publisher: Elsevier.

[325] Luft, J.; Young, R. S.; Meynert, A. M.; Taylor, M. S. Detecting oncogenic selection through biased allele retention in The Cancer Genome Atlas. `https://www.biorxiv.org/content/10.1101/2020.07.03.186593v1`, Pages: 2020.07.03.186593 Section: New Results.

[326] Clayton, E. A.; Khalid, S.; Ban, D.; Wang, L.; Jordan, I. K.; McDonald, J. F. Tumor suppressor genes and allele-specific expression: mechanisms and significance. *11*, 462–479.

[327] Batcha, A. M. N. *et al.* Allelic Imbalance of Recurrently Mutated Genes in Acute Myeloid Leukaemia. *9*, 11796.

[328] Rhee, J.-K.; Lee, S.; Park, W.-Y.; Kim, Y.-H.; Kim, T.-M. Allelic imbalance of somatic mutations in cancer genomes and transcriptomes. *7*, 1653.

[329] Halabi, N. M.; Martinez, A.; Al-Farsi, H.; Mery, E.; Puydenus, L.; Pujol, P.; Khalak, H. G.; McLurcan, C.; Ferron, G.; Querleu, D., *et al.* Preferential allele expression analysis identifies shared germline and somatic driver genes in advanced ovarian cancer. *PLoS genetics* **2016**, *12*, e1005755.

[330] Liu, Z.; Gui, T.; Wang, Z.; Li, H.; Fu, Y.; Dong, X.; Li, Y. cisASE: a likelihood-based method for detecting putative *cis* -regulated allele-specific expression in RNA sequencing data. *32*, 3291–3297.

[331] Skelly, D. A.; Johansson, M.; Madeoy, J.; Wakefield, J.; Akey, J. M. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome research* **2011**, *21*, 1728–1737.

[332] McCoy, R. C.; Wakefield, J.; Akey, J. M. Impacts of Neanderthal-introgressed sequences on the landscape of human gene expression. *Cell* **2017**, *168*, 916–927.

[333] Mayba, O.; Gilbert, H. N.; Liu, J.; Haverty, P. M.; Jhunjhunwala, S.; Jiang, Z.; Watanabe, C.; Zhang, Z. MBASED: allele-specific expression detection in cancer tissues and cell lines. *Genome biology* **2014**, *15*, 405.

[334] Sun, W. A statistical framework for eQTL mapping using RNA-seq data. *Biometrics* **2012**, *68*, 1–11.

[335] Mohammadi, P.; Castel, S. E.; Cummings, B. B.; Einson, J.; Sousa, C.; Hoffman, P.; Donkervoort, S.; Jiang, Z.; Mohassel, P.; Foley, A. R., *et al.* Genetic regulatory variation in populations informs transcriptome analysis in rare disease. *Science* **2019**, *366*, 351–356.

[336] Berger, E.; Yorukoglu, D.; Zhang, L.; Nyquist, S. K.; Shalek, A. K.; Kellis, M.; Numanagić, I.; Berger, B. Improved haplotype inference by exploiting long-range linking and allelic imbalance in RNA-seq datasets. *Nature Communications* **2020**, *11*, 1–9.

[337] Marx, V. How to deduplicate PCR. *Nature methods* **2017**, *14*, 473–476.

[338] Mendelevich, A.; Vinogradova, S.; Gupta, S.; Mironov, A. A.; Sunyaev, S.; Gimelbrant, A. A. Unexpected variability of allelic imbalance estimates from RNA sequencing. *bioRxiv* **2020**,

[339] Castel, S. E. *et al.* A vast resource of allelic expression data spanning human tissues. *21*, 234.

[340] Pinter, S. F.; Colognori, D.; Beliveau, B. J.; Sadreyev, R. I.; Payer, B.; Yildirim, E.; Wu, C.-t.; Lee, J. T. Allelic imbalance is a prevalent and tissue-specific feature of the mouse transcriptome. *Genetics* **2015**, *200*, 537–549.

[341] Crowley, J. J.; Zhabotynsky, V.; Sun, W.; Huang, S.; Pakatci, I. K.; Kim, Y.; Wang, J. R.; Morgan, A. P.; Calaway, J. D.; Aylor, D. L., *et al.* Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nature genetics* **2015**, *47*, 353–360.

[342] Yan, H.; Yuan, W.; Velculescu, V. E.; Vogelstein, B.; Kinzler, K. W., *et al.* Allelic variation in human gene expression. *Science* **2002**, *297*, 1143–1143.

[343] Consortium, G., *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **2020**, *369*, 1318–1330.

[344] Lind, L.; Fors, N.; Hall, J.; Marttala, K.; Stenborg, A. A comparison of three different methods to evaluate endothelium-dependent vasodilation in the elderly: the Prospective Investigation of the Vasculature in Uppsala Seniors (PIVUS) study. *Arteriosclerosis, thrombosis, and vascular biology* **2005**, *25*, 2368–2375.

[345] Balliu, B.; Durrant, M.; de Goede, O.; Abell, N.; Li, X.; Liu, B.; Gloudemans, M. J.; Cook, N. L.; Smith, K. S.; Knowles, D. A., *et al.* Genetic regulation of gene expression and splicing during a 10-year period of human aging. *Genome biology* **2019**, *20*, 230.

[346] Harwood, M. P.; Alves, I.; Edgington, H.; Agbessi, M.; Bruat, V.; Soave, D.; Lamaze, F. C.; Fave, M.-J.; Awadalla, P. Recombination affects allele-specific expression of deleterious variants in human populations. *8*, eabl3819, Publisher: American Association for the Advancement of Science.

[347] Ellsworth, D. L.; Ellsworth, R. E.; Love, B.; Deyarmin, B.; Lubert, S. M.; Mittal, V.; Shriver, C. D. Genomic patterns of allelic imbalance in disease free tissue adjacent to primary breast carcinomas. *88*, 131–139.

[348] Correia, L.; Magno, R.; Xavier, J. M.; de Almeida, B. P.; Duarte, I.; Esteves, F.; Ghezzo, M.; Eldridge, M.; Sun, C.; Bosma, A., *et al.* Allelic expression imbalance of PIK3CA mutations is frequent in breast cancer and prognostically significant. *NPJ Breast Cancer* **2022**, *8*, 71.

[349] Sivakumar, S.; San Lucas, F. A.; Jakubek, Y. A.; McDowell, T. L.; Lang, W.; Kallsen, N.; Peyton, S.; Davies, G. E.; Fukuoka, J.; Yatabe, Y., *et al.* Genomic landscape of allelic imbalance in premalignant atypical adenomatous hyperplasias of the lung. *EBioMedicine* **2019**, *42*, 296–303.

[350] Pomerantz, M. M.; Ahmadiyeh, N.; Jia, L.; Herman, P.; Verzi, M. P.; Doddapaneni, H.; Beckwith, C. A.; Chan, J. A.; Hills, A.; Davis, M., *et al.* The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nature genetics* **2009**, *41*, 882–884.

[351] Wright, J. B.; Brown, S. J.; Cole, M. D. Upregulation of c-MYC in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. *Molecular and cellular biology* **2010**, *30*, 1411–1420.

[352] Tomlinson, I.; Webb, E.; Carvajal-Carmona, L.; Broderick, P.; Kemp, Z.; Spain, S.; Penegar, S.; Chandler, I.; Gorman, M.; Wood, W., *et al.* A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24. 21. *Nature genetics* **2007**, *39*, 984–988.

[353] Tuupanen, S.; Niittymaki, I.; Nousiainen, K.; Vanharanta, S.; Mecklin, J.-P.; Nuorva, K.; Jarvinen, H.; Hautaniemi, S.; Karhu, A.; Aaltonen, L. A. Allelic Imbalance at rs6983267 Suggests Selection of the Risk Allele in Somatic Colorectal Tumor Evolution. *68*, 14–17.

[354] Cox, D. G.; Simard, J.; Sinnett, D.; Hamdi, Y.; Soucy, P.; Ouimet, M.; Barjhoux, L.; Verny-Pierre, C.; McGuffog, L.; Healey, S., *et al.* Common variants of the BRCA1 wild-type allele modify the risk of breast cancer in BRCA1 mutation carriers. *Human molecular genetics* **2011**, *20*, 4732–4747.

[355] Duran-Lozano, L.; Montalban, G.; Bonache, S.; Moles-Fernández, A.; Tenés, A.; Castroviejo-Bermejo, M.; Carrasco, E.; López-Fernández, A.; Torres-Esquius, S.; Gadea, N., *et al.* Alternative transcript imbalance underlying breast cancer susceptibility in a family carrying PALB2 c. 3201+ 5G¿ T. *Breast Cancer Research and Treatment* **2019**, *174*, 543–550.

[356] Maia, A.-T. *et al.* Effects of BRCA2 cis-regulation in normal breast and cancer risk amongst BRCA2 mutation carriers. *14*, R63.

[357] Meyer, K. B.; Maia, A.-T.; O'Reilly, M.; Teschendorff, A. E.; Chin, S.-F.; Caldas, C.; Ponder, B. A. J. Allele-specific up-regulation of FGFR2 increases susceptibility to breast cancer. *PLoS biology* **2008**, *6*, e108.

[358] Gao, C.; Devarajan, K.; Zhou, Y.; Slater, C. M.; Daly, M. B.; Chen, X. Identifying breast cancer risk loci by global differential allele-specific expression (DASE) analysis in mammary epithelial transcriptome. *13*, 570.

[359] Esteves, F.; Xavier, J. M.; Ford, A. M.; Rocha, C.; Pharoah, P. D. P.; Caldas, C.; Chin, S.-F.; Maia, A.-T. Germline allelic expression of genes at 17q22 locus associates with risk of breast cancer. *172*, 146–157.

[360] Montalban, G.; Bonache, S.; Moles-Fernández, A.; Gisbert-Beamud, A.; Tenés, A.; Bach, V.; Carrasco, E.; López-Fernández, A.; Stjepanovic, N.; Balmaña, J., *et al.* Screening of BRCA1/2 deep intronic regions by targeted gene sequencing identifies the first germline BRCA1 variant causing pseudoexon activation in a patient with breast/ovarian cancer. *Journal of medical genetics* **2019**, *56*, 63–74.

[361] Curia, M. C.; De Iure, S.; De Lellis, L.; Veschi, S.; Mammarella, S.; White, M. J.; Bartlett, J.; Di Iorio, A.; Amatetti, C.; Lombardo, M.; Di Gregorio, P.; Battista, P.; Mariani-Costantini, R.; Williams, S. M.; Cama, A. Increased Variance in Germline Allele-Specific Expression of APC Associates With Colorectal Cancer. *142*, 71–77.e1.

[362] Valle, L.; Serena-Acedo, T.; Liyanarachchi, S.; Hampel, H.; Comeras, I.; Li, Z.; Zeng, Q.; Zhang, H.-T.; Pennison, M. J.; Sadim, M.; Pasche, B.; Tanner, S. M.; de la Chapelle, A. Germline Allele-Specific Expression of TGFBR1 Confers an Increased Risk of Colorectal Cancer. *321*, 1361–1365, Publisher: American Association for the Advancement of Science.

[363] Choi, K.; Raghupathy, N.; Churchill, G. A. A Bayesian mixture model for the analysis of allelic expression in single cells. *Nature communications* **2019**, *10*, 1–11.

[364] Wei, Q.-X.; Claus, R.; Hielscher, T.; Mertens, D.; Raval, A.; Oakes, C. C.; Tanner, S. M.; de la Chapelle, A.; Byrd, J. C.; Stilgenbauer, S.; Plass, C. Germline allele-specific expression of DAPK1 in chronic lymphocytic leukemia. *8*, e55261.

[365] Gusev, A.; Spisak, S.; Fay, A. P.; Carol, H.; Vavra, K. C.; Signoretti, S.; Tisza, V.; Pomerantz, M.; Abbasi, F.; Seo, J.-H.; Choueiri, T. K.; Lawrenson, K.; Freedman, M. L. Allelic imbalance reveals widespread germline-somatic regulatory differences and prioritizes risk loci in Renal Cell Carcinoma. `https://www.biorxiv.org/content/10.1101/631150v1`, Pages: 631150 Section: New Results.

[366] Buzby, J. S.; Williams, S. A.; Schaffer, L.; Head, S. R.; Nugent, D. J. Allele-specific wild-type TP53 expression in the unaffected carrier parent of children with Li-Fraumeni syndrome. *211*, 9–17.

[367] Buzby, J. S.; Williams, S. A.; Nugent, D. J. Unaffected Li-Fraumeni Syndrome Carrier Parent Demonstrates Allele-Specific mRNA Stabilization of Wild-Type TP53 Compared to Affected Offspring. *13*, 2302, Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

[368] Comiskey Jr, D. F.; He, H.; Liyanarachchi, S.; Sheikh, M. S.; Genutis, L. K.; Hendrickson, I. V.; Yu, L.; Brock, P. L.; de la Chapelle, A. Variants in LRRC34 reveal distinct mechanisms for predisposition to papillary thyroid carcinoma. *Journal of Medical Genetics* **2020**, *57*, 519–527.

[369] Tan, A. C. *et al.* Allele-specific expression in the germline of patients with familial pancreatic cancer: An unbiased approach to cancer gene discovery. *7*, 135–144, Publisher: Taylor & Francis _eprint: https://doi.org/10.4161/cbt.7.1.5199.

[370] Shetty, A.; Seo, J.-H.; Bell, C. A.; O'Connor, E. P.; Pomerantz, M. M.; Freedman, M. L.; Gusev, A. Allele-specific epigenetic activity in prostate cancer and normal prostate tissue implicates prostate cancer risk mechanisms. *108*, 2071–2085.

[371] Supek, F.; Lehner, B.; Lindeboom, R. G. H. To NMD or Not To NMD: Nonsense-Mediated mRNA Decay in Cancer and Other Genetic Diseases. *37*, 657–668.

[372] Varmus, H. E. The molecular genetics of cellular oncogenes. *Annual review of genetics* **1984**, *18*, 553–612.

[373] Timar, J.; Kashofer, K. Molecular epidemiology and diagnostics of KRAS mutations in human cancer. *Cancer and Metastasis Reviews* **2020**, 1–10.

[374] Jančík, S.; Drábek, J.; Radzioch, D.; Hajdúch, M. Clinical relevance of KRAS in human cancers. *Journal of Biomedicine and Biotechnology* **2010**, *2010*.

[375] Bernadotte, A.; Mikhelson, V. M.; Spivak, I. M. Markers of cellular senescence. Telomere shortening as a marker of cellular senescence. *Aging (Albany NY)* **2016**, *8*, 3.

[376] Dratwa, M.; Wysoczańska, B.; Łacina, P.; Kubik, T.; Bogunia-Kubik, K. TERT—regulation and roles in cancer formation. *Frontiers in Immunology* **2020**, *11*, 589929.

[377] McKelvey, B. A.; Umbricht, C. B.; Zeiger, M. A. Telomerase Reverse Transcriptase (TERT) Regulation in Thyroid Cancer: A Review. *Frontiers in Endocrinology* **2020**, *11*, 485.

[378] Stern, J. L.; Paucek, R. D.; Huang, F. W.; Ghandi, M.; Nwumeh, R.; Costello, J. C.; Cech, T. R. Allele-specific DNA methylation and its interplay with repressive histone marks at promoter-mutant TERT genes. *Cell reports* **2017**, *21*, 3700–3707.

[379] Gamazon, E. R.; Wheeler, H. E.; Shah, K.; Mozaffari, S. V.; Aquino-Michaels, K.; Carroll, R. J.; Eyler, A. E.; Denny, J. C.; Nicolae, D. L.; Cox, N. J.; Im, H. K.; GTEx Consortium, PrediXcan: Trait Mapping Using Human Transcriptome Regulation.

[380] Ellis, P.; Moore, L.; Sanders, M. A.; Butler, T. M.; Brunner, S. F.; Lee-Six, H.; Osborne, R.; Farr, B.; Coorens, T. H. H.; Lawson, A. R. J.; Cagan, A.; Stratton, M. R.; Martincorena, I.; Campbell, P. J. Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. *16*, 841–871.

[381] Gawad, C.; Koh, W.; Quake, S. R. Single-cell genome sequencing: current state of the science. *17*, 175–188, Number: 3 Publisher: Nature Publishing Group.

[382] Tang, S.; Ning, Q.; Yang, L.; Mo, Z.; Tang, S. Mechanisms of immune escape in the cancer immune cycle. *86*, 106700.

[383] Jonason, A.; Kunala, S.; Price, G.; Restifo, R.; Spinelli, H.; Persing, J.; Leffell, D.; Tarone, R.; Brash, D. Frequent clones of p53-mutated keratinocytes in normalhumanskin. *93*, 14025–14029, Publisher: Proceedings of the National Academy of Sciences.

[384] Hernando, B.; Dietzen, M.; Parra, G.; Gil-Barrachina, M.; Pitarch, G.; Mahiques, L.; Valcuende-Cavero, F.; McGranahan, N.; Martinez-Cadenas, C. The effect of age on the acquisition and selection of cancer driver mutations in sun-exposed normal skin. *32*, 412–421.

[385] Martincorena, I.; Fowler, J. C.; Wabik, A.; Lawson, A. R. J.; Abascal, F.; Hall, M. W. J.; Cagan, A.; Murai, K.; Mahbubani, K.; Stratton, M. R.; Fitzgerald, R. C.; Handford, P. A.; Campbell, P. J.; Saeb-Parsy, K.; Jones, P. H. Somatic mutant clones colonize the human esophagus with age. *362*, 911–917, Publisher: American Association for the Advancement of Science.

[386] Yokoyama, A. *et al.* Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *565*, 312–317, Number: 7739 Publisher: Nature Publishing Group.

[387] Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *574*, 532–537, Number: 7779 Publisher: Nature Publishing Group.

[388] Moore, L. *et al.* The mutational landscape of normal human endometrial epithelium. *580*, 640–646, Number: 7805 Publisher: Nature Publishing Group.

[389] Brunner, S. F.; Roberts, N. D.; Wylie, L. A.; Moore, L.; Aitken, S. J.; Davies, S. E.; Sanders, M. A.; Ellis, P.; Alder, C.; Hooks, Y.; Abascal, F.; Stratton, M. R.; Martincorena, I.; Hoare, M.; Campbell, P. J. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *574*, 538–542, Number: 7779 Publisher: Nature Publishing Group.

[390] Ng, S. W. K. *et al.* Convergent somatic mutations in metabolism genes in chronic liver disease. *598*, 473–478, Number: 7881 Publisher: Nature Publishing Group.

[391] Rodin, R. E. *et al.* The landscape of somatic mutation in cerebral cortex of autistic and neurotypical individuals revealed by ultra-deep whole-genome sequencing. *24*, 176–185.

[392] Grossmann, S.; Hooks, Y.; Wilson, L.; Moore, L.; O'Neill, L.; Martincorena, I.; Voet, T.; Stratton, M. R.; Heer, R.; Campbell, P. J. Development, maturation, and maintenance of human prostate inferred from somatic mutations. *28*, 1262–1274.e5.

[393] Lawson, A. R. J. *et al.* Extensive heterogeneity in somatic mutation and selection in the human bladder. *370*, 75–82, Publisher: American Association for the Advancement of Science.

[394] Li, R.; Du, Y.; Chen, Z.; Xu, D.; Lin, T.; Jin, S.; Wang, G.; Liu, Z.; Lu, M.; Chen, X.; Xu, T.; Bai, F. Macroscopic somatic clonal expansion in morphologically normal human urothelium. *370*, 82–89, Publisher: American Association for the Advancement of Science.

[395] Li, R. *et al.* A body map of somatic mutagenesis in morphologically normal human tissues. *597*, 398–403.

[396] Poon, G. Y. P.; Watson, C. J.; Fisher, D. S.; Blundell, J. R. Synonymous mutations reveal genome-wide levels of positive selection in healthy tissues. *53*, 1597–1605.

[397] Muyas, F.; Zapata, L.; Guigo, R.; Ossowski, S. The rate and spectrum of mosaic mutations during embryogenesis revealed by RNA sequencing of 49 tissues. *12*, 49.

[398] Son, H.; Kim, J. H.; Kim, I. B.; Kim, M.-H.; Sim, N. S.; Kim, D.-S.; Lee, J.; Lee, J. H.; Kim, S. Multi-organ analysis of low-level somatic mosaicism reveals stage- and tissue-specific mutational features in human development. `https://www.biorxiv.org/content/10.1101/2021.08.23.457440v1`, Pages: 2021.08.23.457440 Section: New Results.

[399] Araten, D. J.; Golde, D. W.; Zhang, R. H.; Thaler, H. T.; Gargiulo, L.; Notaro, R.; Luzzatto, L. A Quantitative Measurement of the Human Somatic Mutation Rate. *65*, 8111–8117.

[400] Peruzzi, B.; Araten, D. J.; Notaro, R.; Luzzatto, L. The use of PIG-A as a sentinel gene for the study of the somatic mutation rate and of mutagenic agents in vivo. *705*, 3–10.

[401] Saini, N. *et al.* The Impact of Environmental and Endogenous Damage on Somatic Mutation Load in Human Skin Fibroblasts. *12*, e1006385, Publisher: Public Library of Science.

[402] Kuijk, E.; Jager, M.; van der Roest, B.; Locati, M. D.; Van Hoeck, A.; Korzelius, J.; Janssen, R.; Besselink, N.; Boymans, S.; van Boxtel, R.; Cuppen, E. The mutational impact of culturing human pluripotent and adult stem cells. *11*, 2493, Number: 1 Publisher: Nature Publishing Group.

[403] Olafsson, S.; Anderson, C. A. Somatic mutations provide important and unique insights into the biology of complex diseases. *37*, 872–881.

[404] Failla, G. The Aging Process and Cancerogenesis. *71*, 1124–1140, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1749-6632.1958.tb54674.x.

[405] Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *578*, 122–128.

[406] Chatsirisupachai, K.; Lagger, C.; Magalhes, J. P. d. Age-associated differences in the cancer molecular landscape. *0*, Publisher: Elsevier.

[407] Li, C. H.; Haider, S.; Boutros, P. C. Age influences on the molecular presentation of tumours. *13*, 208, Number: 1 Publisher: Nature Publishing Group.

[408] Campbell, B. B. *et al.* Comprehensive Analysis of Hypermutation in Human Cancer. *171*, 1042–1056.e10.

[409] Xiao, D.; Pan, H.; Li, F.; Wu, K.; Zhang, X.; He, J. Analysis of ultra-deep targeted sequencing reveals mutation burden is associated with gender and clinical outcome in lung adenocarcinoma. *7*, 22857–22864, Publisher: Impact Journals.

[410] Gupta, S.; Artomov, M.; Goggins, W.; Daly, M.; Tsao, H. Gender Disparity and Mutation Burden in Metastatic Melanoma. *107*, djv221.

[411] Li, C. H.; Haider, S.; Shiah, Y.-J.; Thai, K.; Boutros, P. C. Sex Differences in Cancer Driver Genes and Biomarkers. *78*, 5527–5537.

[412] Zhang, W.; Edwards, A.; Flemington, E. K.; Zhang, K. Racial disparities in patient survival and tumor mutation burden, and the association between tumor mutation burden and cancer incidence rate. *7*, 13639.

[413] Moore, L. *et al.* The mutational landscape of human somatic and germline cells. *597*, 381–386, Number: 7876 Publisher: Nature Publishing Group.

[414] Lac, V.; Nazeran, T. M.; Tessier-Cloutier, B.; Aguirre-Hernandez, R.; Albert, A.; Lum, A.; Khattra, J.; Praetorius, T.; Mason, M.; Chiu, D.; Kbel, M.; Yong, P. J.; Gilks, B. C.; Anglesio, M. S.; Huntsman, D. G. Oncogenic mutations in histologically normal endometrium: the new normal? *249*, 173–181, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/path.5314.

[415] Lac, V. *et al.* Iatrogenic endometriosis harbors somatic cancer-driver mutations. *34*, 69–78.

[416] Anglesio, M. *et al.* Cancer-Associated Mutations in Endometriosis without Cancer. *376*, 1835–1848.

[417] Suda, K.; Nakaoka, H.; Yoshihara, K.; Ishiguro, T.; Tamura, R.; Mori, Y.; Yamawaki, K.; Adachi, S.; Takahashi, T.; Kase, H.; Tanaka, K.; Yamamoto, T.; Motoyama, T.; Inoue, I.; Enomoto, T. Clonal Expansion and Diversification of Cancer-Associated Mutations in Endometriosis and Normal Endometrium. *24*, 1777–1789.

[418] Oh, J.-H.; Sung, C. O. Comprehensive characteristics of somatic mutations in the normal tissues of patients with cancer and existence of somatic mutant clones linked to cancer development. *58*, 433–441, Publisher: BMJ Publishing Group Ltd Section: Cancer genetics.

[419] Wodarz, D.; Newell, A. C.; Komarova, N. L. Passenger mutations can accelerate tumour suppressor gene inactivation in cancer evolution. *15*, 20170967.

[420] Pavel, A. B.; Korolev, K. S. Genetic load makes cancer cells more sensitive to common drugs: evidence from Cancer Cell Line Encyclopedia. *7*, 1938, Number: 1 Publisher: Nature Publishing Group.

[421] Youn, A.; Simon, R. Using passenger mutations to estimate the timing of driver mutations and identify mutator alterations. *14*, 363.

[422] Salvadores, M.; Mas-Ponte, D.; Supek, F. Passenger mutations accurately classify human tumors. *15*, e1006953, Publisher: Public Library of Science.

[423] McFarland, C. D.; Korolev, K. S.; Kryukov, G. V.; Sunyaev, S. R.; Mirny, L. A. Impact of deleterious passenger mutations on cancer progression. *110*, 2910–2915.

[424] McFarland, C. D.; Yaglom, J. A.; Wojtkowiak, J. W.; Scott, J. G.; Morse, D. L.; Sherman, M. Y.; Mirny, L. A. The damaging effect of passenger mutations on cancer progression. *77*, 4763–4772.

[425] Fischer, A.; Vazquez-Garcia, I.; Illingworth, C. J.; Mustonen, V. High-Definition Reconstruction of Clonal Composition in Cancer. *7*, 1740–1752.

[426] Ju, Y. S. *et al.* Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *543*, 714–718, Number: 7647 Publisher: Nature Publishing Group.

[427] Werner, B.; Traulsen, A.; Sottoriva, A.; Dingli, D. Detecting truly clonal alterations from multi-region profiling of tumours. *7*, 44991, Number: 1 Publisher: Nature Publishing Group.

[428] Rizvi, N. A. *et al.* Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *348*, 124–128.

[429] Kowanetz, M. *et al.* Differential regulation of PD-L1 expression by immune and tumor cells in NSCLC and the response to treatment with atezolizumab (anti-PD-L1). *115*, E10119–E10126, Publisher: Proceedings of the National Academy of Sciences.

[430] Ready, N. E. *et al.* Nivolumab Monotherapy and Nivolumab Plus Ipilimumab in Recurrent Small Cell Lung Cancer: Results From the CheckMate 032 Randomized Cohort. *15*, 426–435.

[431] Balar, A. V.; Castellano, D.; O'Donnell, P. H.; Grivas, P.; Vuky, J.; Powles, T.; Plimack, E. R.; Hahn, N. M.; de Wit, R.; Pang, L.; Savage, M. J.; Perini, R. F.; Keefe, S. M.; Bajorin, D.; Bellmunt, J. First-line pembrolizumab in cisplatin-ineligible patients with locally advanced and unresectable or metastatic urothelial cancer (KEYNOTE-052): a multicentre, single-arm, phase 2 study. *18*, 1483–1492.

[432] Snyder, A. *et al.* Genetic basis for clinical response to CTLA-4 blockade in melanoma. *371*, 2189–2199.

[433] Burns, M. B.; Temiz, N. A.; Harris, R. S. Evidence for APOBEC3B mutagenesis in multiple human cancers. *45*, 977–983, Number: 9 Publisher: Nature Publishing Group.

[434] Muzny, D. M. *et al.* Comprehensive molecular characterization of human colon and rectal cancer. *487*, 330–337, Number: 7407 Publisher: Nature Publishing Group.

[435] Smit, K. N.; Jager, M. J.; de Klein, A.; Kili, E. Uveal melanoma: Towards a molecular understanding. *75*, 100800.

[436] Rodrigues, M. *et al.* Evolutionary Routes in Metastatic Uveal Melanomas Depend on MBD4 Alterations. *25*, 5513–5524.

[437] Chen, Y.; Gharwan, H.; Thomas, A. Novel Biologic Therapies for Thymic Epithelial Tumors. *4*.

[438] Radovich, M. *et al.* The integrated genomic landscape of thymic epithelial tumors. *33*, 244–258.e10.

[439] Milholland, B.; Auton, A.; Suh, Y.; Vijg, J. Age-related somatic mutations in the cancer genome. *6*, 24627–24635, Publisher: Impact Journals.

[440] Thorsson, V. *et al.* The Immune Landscape of Cancer. *48*, 812–830.e14.

[441] Colaprico, A.; Silva, T. C.; Olsen, C.; Garofano, L.; Cava, C.; Garolini, D.; Sabedot, T. S.; Malta, T. M.; Pagnotta, S. M.; Castiglioni, I.; Ceccarelli, M.; Bontempi, G.; Noushmehr, H. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *44*, e71.

[442] Carrot-Zhang, J. *et al.* Comprehensive Analysis of Genetic Ancestry and Its Molecular Correlates in Cancer. *37*, 639–654.e6.

[443] Clopper, C. J.; Pearson, E. S. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *26*, 404–413, Publisher: [Oxford University Press, Biometrika Trust].

[444] Sayaman, R. W. *et al.* Germline genetic contribution to the immune landscape of cancer. *54*, 367–386.e8.

[445] Chang, C. C.; Chow, C. C.; Tellier, L. C.; Vattikuti, S.; Purcell, S. M.; Lee, J. J. Second-generation PLINK: rising to the challenge of larger and richer datasets. *4*, 7.

[446] Watanabe, K.; Taskesen, E.; van Bochoven, A.; Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *8*, 1826.

[447] Robinson, J.; Halliwell, J. A.; Hayhurst, J. D.; Flicek, P.; Parham, P.; Marsh, S. G. E. The IPD and IMGT/HLA database: allele variant databases. *43*, D423–431.

[448] Rock, K. L.; Shen, L. Cross-presentation: underlying mechanisms and role in immune surveillance. *207*, 166–183.

[449] Swann, J. B.; Smyth, M. J. Immune surveillance of tumors. *117*, 1137–1146.

[450] Vesely, M. D.; Schreiber, R. D. Cancer immunoediting: antigens, mechanisms, and implications to cancer immunotherapy. *1284*, 1–5.

[451] Campoli, M.; Ferrone, S. HLA antigen changes in malignant cells: epigenetic mechanisms and biologic significance. *27*, 5869–5885.

[452] Shukla, S. A. *et al.* Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *33*, 1152–1158.

[453] Davoli, T.; Uno, H.; Wooten, E. C.; Elledge, S. J. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *355*, eaaf8399.

[454] Brown, S. D.; Warren, R. L.; Gibb, E. A.; Martin, S. D.; Spinelli, J. J.; Nelson, B. H.; Holt, R. A. Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *24*, 743–750, Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[455] Castro, A.; Ozturk, K.; Pyke, R. M.; Xian, S.; Zanetti, M.; Carter, H. Elevated neoantigen levels in tumors with somatic mutations in the HLA-A, HLA-B, HLA-C and B2M genes. *12*, 107.

[456] Chen, C.; Qi, H.; Shen, Y.; Pickrell, J.; Przeworski, M. Contrasting Determinants of Mutation Rates in Germline and Soma. *207*, 255–267.

[457] Chapman, M. A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *471*, 467–472.

[458] Jones, S. *et al.* Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *321*, 1801–1806.

[459] Lee, W. *et al.* The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *465*, 473–477.

[460] Pleasance, E. D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *463*, 184–190.

[461] Hodgkinson, A.; Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *12*, 756–766.

[462] Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *499*, 214–218, Number: 7457 Publisher: Nature Publishing Group.

[463] Wells, D. K. *et al.* Key Parameters of Tumor Epitope Immunogenicity Revealed Through a Consortium Approach Improve Neoantigen Prediction. *183*, 818–834.e13.

[464] Yadav, M.; Jhunjhunwala, S.; Phung, Q. T.; Lupardus, P.; Tanguay, J.; Bumbaca, S.; Franci, C.; Cheung, T. K.; Fritsche, J.; Weinschenk, T.; Modrusan, Z.; Mellman, I.; Lill, J. R.; Delamarre, L. Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *515*, 572–576, Number: 7528 Publisher: Nature Publishing Group.

[465] Kim, K.; Kim, H. S.; Kim, J. Y.; Jung, H.; Sun, J.-M.; Ahn, J. S.; Ahn, M.-J.; Park, K.; Lee, S.-H.; Choi, J. K. Predicting clinical benefit of immunotherapy by antigenic or functional mutations affecting tumour immunogenicity. *11*, 951.

[466] Kim, J. Y.; Cha, H.; Kim, K.; Sung, C.; An, J.; Bang, H.; Kim, H.; Yang, J. O.; Chang, S.; Shin, I.; Noh, S.-J.; Shin, I.; Cho, D.-Y.; Lee, S.-H.; Choi, J. K. MHC II immunogenicity shapes the neoepitope landscape in human tumors. *55*, 221–231, Number: 2 Publisher: Nature Publishing Group.

[467] Oncogene Database. https://ongene.bioinfo-minzhao.org, Accessed: 2020-01-19.

[468] Tumor Suppressor Gene Database. https://bioinfo.uth.edu/TSGene/, Accessed: 2020-01-19.

[469] Kumar, S. *et al.* Passenger Mutations in More Than 2,500 Cancer Genomes: Overall Molecular Functional Impact and Consequences. *180*, 915–927.e16.

[470] Wang, T.; Ruan, S.; Zhao, X.; Shi, X.; Teng, H.; Zhong, J.; You, M.; Xia, K.; Sun, Z.; Mao, F. OncoVar: an integrated database and analysis platform for oncogenic driver variants in cancers. *49*, D1289–D1301.

[471] Nielsen, M.; Andreatta, M. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *8*, 33.

[472] Robinson, M. D.; McCarthy, D. J.; Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *26*, 139–140.

[473] Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer.

[474] Lawrence, M.; Huber, W.; Pages, H.; Aboyoun, P.; Carlson, M.; Gentleman, R.; Morgan, M. T.; Carey, V. J. Software for Computing and Annotating Genomic Ranges. *9*, e1003118, Publisher: Public Library of Science.

[475] Fantini, D.; Vidimar, V.; Yu, Y.; Condello, S.; Meeks, J. J. MutSignatures: an R package for extraction and analysis of cancer mutational signatures. *10*, 18217, Number: 1 Publisher: Nature Publishing Group.

[476] McLaren, W.; Pritchard, B.; Rios, D.; Chen, Y.; Flicek, P.; Cunningham, F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *26*, 2069–2070.

[477] Danecek, P.; Bonfield, J. K.; Liddle, J.; Marshall, J.; Ohan, V.; Pollard, M. O.; Whitwham, A.; Keane, T.; McCarthy, S. A.; Davies, R. M.; Li, H. Twelve years of SAMtools and BCFtools. *10*, giab008.

[478] Cunningham, F. *et al.* Ensembl 2022. *50*, D988–D995.

[479] Gourmet, L.; Sottoriva, A.; Secrier, M.; Zapata, L. Immune evasion impacts the selective landscape of driver genes during tumorigenesis. https://www.biorxiv.org/content/10.1101/2022.06.20.496910v1, Pages: 2022.06.20.496910 Section: New Results.

[480] Reddy, T. E.; Gertz, J.; Pauli, F.; Kucera, K. S.; Varley, K. E.; Newberry, K. M.; Marinov, G. K.; Mortazavi, A.; Williams, B. A.; Song, L.; Crawford, G. E.; Wold, B.; Willard, H. F.; Myers, R. M. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *22*, 860–869.

[481] Montgomery, S. B.; Lappalainen, T.; Gutierrez-Arcelus, M.; Dermitzakis, E. T. Rare and Common Regulatory Variation in Population-Scale Sequenced Human Genomes. *7*, e1002144, Publisher: Public Library of Science.

[482] Nembaware, V.; Wolfe, K. H.; Bettoni, F.; Kelso, J.; Seoighe, C. Allele-specific transcript isoforms in human. *577*, 233–238.

[483] Nembaware, V.; Lupindo, B.; Schouest, K.; Spillane, C.; Scheffler, K.; Seoighe, C. Genome-wide survey of allele-specific splicing in humans. *9*, 265.

[484] Payne, S. R.; Kemp, C. J. Tumor suppressor genetics. *26*, 2031–2045.

[485] Helias-Rodzewicz, Z.; Funck-Brentano, E.; Zimmermann, U.; Terrones, N.; Saiag, P.; Emile, J.-F. Frequent allelic imbalance in NRAS mutant melanomas. *34*, 9578–9578, Publisher: Wolters Kluwer.

[486] Ren, N.; Liu, Q.; Yan, L.; Huang, Q. Parallel Reporter Assays Identify Altered Regulatory Role of rs684232 in Leading to Prostate Cancer Predisposition. *22*, 8792, Number: 16 Publisher: Multidisciplinary Digital Publishing Institute.

[487] Larson, N. B.; McDonnell, S.; French, A. J.; Fogarty, Z.; Cheville, J.; Middha, S.; Riska, S.; Baheti, S.; Nair, A. A.; Wang, L.; Schaid, D. J.; Thibodeau, S. N. Comprehensively Evaluating cis-Regulatory Variation in the Human Prostate Transcriptome by Using Gene-Level Allele-Specific Expression. *96*, 869–882.

[488] Baca, S. C. *et al.* Genetic determinants of chromatin reveal prostate cancer risk mediated by context-dependent gene regulation. *54*, 1364–1375, Number: 9 Publisher: Nature Publishing Group.

[489] Liu, Z.; Dong, X.; Li, Y. A Genome-Wide Study of Allele-Specific Expression in Colorectal Cancer. *9*.

[490] Rosic, J.; Miladinov, M.; Dragicevic, S.; Eric, K.; Bogdanovic, A.; Krivokapic, Z.; Nikolic, A. Genetic analysis and allele-specific expression of SMAD7 3UTR variants in human colorectal cancer reveal a novel somatic variant exhibiting allelic imbalance. *859*, 147217.

[491] Palin, K. *et al.* Contribution of allelic imbalance to colorectal cancer. *9*, 3664, Number: 1 Publisher: Nature Publishing Group.

[492] Zhang, M. *et al.* Characterising *cis* -regulatory variation in the transcriptome of histologically normal and tumour-derived pancreatic tissues. *67*, 521–533.

[493] Sivakumar, S. *et al.* Genomic landscape of allelic imbalance in premalignant atypical adenomatous hyperplasias of the lung. *42*, 296–303, Publisher: Elsevier.

[494] Przytycki, P. F.; Singh, M. Differential Allele-Specific Expression Uncovers Breast Cancer Genes Dysregulated by Cis Noncoding Mutations. *10*, 193–203.e4.

[495] Zaccaria, S.; Raphael, B. J. Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL. *39*, 207–214, Number: 2 Publisher: Nature Publishing Group.

[496] Song, H.; Liu, Y.; Tan, Y.; Zhang, Y.; Jin, W.; Chen, L.; Wu, S.; Yan, J.; Li, J.; Chen, Z.; Chen, S.; Wang, K. Recurrent noncoding somatic and germline WT1 variants converge to disrupt MYB binding in acute promyelocytic leukemia. *140*, 1132–1144.

[497] Mulet-Lazaro, R.; van Herk, S.; Erpelinck, C.; Bindels, E.; Sanders, M. A.; Vermeulen, C.; Renkens, I.; Valk, P.; Melnick, A. M.; de Ridder, J.; Rehli, M.; Gebhard, C.; Delwel, R.; Wouters, B. J. Allele-specific expression of GATA2 due to epigenetic dysregulation in CEBPA double-mutant AML. *138*, 160–177.

[498] Castel, S. E.; Levy-Moonshine, A.; Mohammadi, P.; Banks, E.; Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *16*, 195.

[499] Pastinen, T. Genome-wide allele-specific analysis: insights into regulatory variation. *11*, 533–538, Number: 8 Publisher: Nature Publishing Group.

[500] Zhang, Z.; van Dijk, F.; de Klein, N.; van Gijn, M. E.; Franke, L. H.; Sinke, R. J.; Swertz, M. A.; van der Velde, K. J. Feasibility of predicting allele specific expression from DNA sequencing using machine learning. *11*, 10606, Number: 1 Publisher: Nature Publishing Group.

[501] Knudson, A. G. Mutation and Cancer: Statistical Study of Retinoblastoma. *68*, 820–823.

[502] Reik, W.; Walter, J. Genomic imprinting: parental influence on the genome. *2*, 21–32.

[503] Benetatos, L.; Vartholomatos, G.; Hatzimichael, E. MEG3 imprinted gene contribution in tumorigenesis. *129*, 773–779.

[504] Castel, S. E.; Cervera, A.; Mohammadi, P.; Aguet, F.; Reverter, F.; Wolman, A.; Guigo, R.; Iossifov, I.; Vasileva, A.; Lappalainen, T. Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *50*, 1327–1334, Number: 9 Publisher: Nature Publishing Group.

[505] Wheeler, H. E.; Shah, K. P.; Brenner, J.; Garcia, T.; Aquino-Michaels, K.; Consortium, G.; Cox, N. J.; Nicolae, D. L.; Im, H. K. Survey of the Heritability and Sparse Architecture of Gene Expression Traits across Human Tissues. *12*, e1006423, Publisher: Public Library of Science.

[506] THE GTEX CONSORTIUM, The GTEx Consortium atlas of genetic regulatory effects across human tissues. *369*, 1318–1330, Publisher: American Association for the Advancement of Science.

[507] O'Connell, J.; Sharp, K.; Shrine, N.; Wain, L.; Hall, I.; Tobin, M.; Zagury, J.-F.; Delaneau, O.; Marchini, J. Haplotype estimation for biobank scale datasets. *48*, 817–820.

[508] Aaltonen, L. A. *et al.* Pan-cancer analysis of whole genomes. *578*, 82–93, Number: 7793 Publisher: Nature Publishing Group.

[509] Oncogene Database. https://www.geneimprint.com, Accessed: 2022-01-23.

[510] Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* **1979**, 65–70.

[511] Cheadle, C.; Vawter, M. P.; Freed, W. J.; Becker, K. G. Analysis of Microarray Data Using Z Score Transformation. *5*, 73–81.

[512] Robin, X.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J.-C.; Müller, M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics* **2011**, *12*, 1–8.

[513] DeLong, E. R.; DeLong, D. M.; Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **1988**, 837–845.

[514] Olsen, L. groupdata2: creating groups from data. R package ver. 1.0. 0. 2017.

[515] Lunardon, N.; Menardi, G.; Torelli, N. ROSE: a package for binary imbalanced learning. *R journal* **2014**, *6*.

[516] Siddiqi, N. *Intelligent credit scoring: Building and implementing better credit risk scorecards*; John Wiley & Sons, 2017.

[517] McLeod, A.; Xu, C. bestglm: Best subset GLM. *URL http://CRAN. R-project. org/package= bestglm* **2010**,

[518] Machado, H. E. *et al.* Diverse mutational landscapes in human lymphocytes. *608*, 724–732, Number: 7924 Publisher: Nature Publishing Group.

[519] Mitchell, E. *et al.* Clonal dynamics of haematopoiesis across the human lifespan. *606*, 343–350, Number: 7913 Publisher: Nature Publishing Group.

[520] Osorio, F. G.; Huber, A. R.; Oka, R.; Verheul, M.; Patel, S. H.; Hasaart, K.; de la Fonteijne, L.; Varela, I.; Camargo, F. D.; van Boxtel, R. Somatic mutations reveal lineage relationships and age-related mutagenesis in human hematopoiesis. *Cell reports* **2018**, *25*, 2308–2316.

[521] Bae, T. *et al.* Analysis of somatic mutations in 131 human brains reveals aging-associated hypermutability. *377*, 511–517, Publisher: American Association for the Advancement of Science.

[522] Luquette, L. J. *et al.* Single-cell genome sequencing of human neurons identifies somatic point mutation and indel enrichment in regulatory elements. *54*, 1564–1571, Number: 10 Publisher: Nature Publishing Group.

[523] Abascal, F.; Harvey, L. M.; Mitchell, E.; Lawson, A. R.; Lensing, S. V.; Ellis, P.; Russell, A. J.; Alcantara, R. E.; Baez-Ortega, A.; Wang, Y., *et al.* Somatic mutation landscapes at single-molecule resolution. *Nature* **2021**, *593*, 405–410.

[524] Suda, K.; Nakaoka, H.; Yoshihara, K.; Ishiguro, T.; Tamura, R.; Mori, Y.; Yamawaki, K.; Adachi, S.; Takahashi, T.; Kase, H., *et al.* Clonal expansion and diversification of cancer-associated mutations in endometriosis and normal endometrium. *Cell reports* **2018**, *24*, 1777–1789.

[525] Zhang, L.; Cecco, M. D.; Lee, M.; Hao, X.; Maslov, A. Y.; Montagna, C.; Campisi, J.; Dong, X.; Sedivy, J. M.; Vijg, J. Analysis of somatic mutations in senescent cells using single-cell whole-genome sequencing. `https://www.biorxiv.org/content/10.1101/2022.09.16.508266v1`, Pages: 2022.09.16.508266 Section: New Results.

[526] Brazhnik, K.; Sun, S.; Alani, O.; Kinkhabwala, M.; Wolkoff, A. W.; Maslov, A. Y.; Dong, X.; Vijg, J. Single-cell analysis reveals different age-related somatic mutation profiles between stem and differentiated cells in human liver. *Science Advances* **2020**, *6*, eaax2659.

[527] Brunner, S. F.; Roberts, N. D.; Wylie, L. A.; Moore, L.; Aitken, S. J.; Davies, S. E.; Sanders, M. A.; Ellis, P.; Alder, C.; Hooks, Y., *et al.* Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature* **2019**, *574*, 538–542.

[528] Huang, Z.; Sun, S.; Lee, M.; Maslov, A. Y.; Shi, M.; Waldman, S.; Marsh, A.; Siddiqui, T.; Dong, X.; Peter, Y.; Sadoughi, A.; Shah, C.; Ye, K.; Spivack, S. D.; Vijg, J. Single-cell analysis of somatic mutations in human bronchial epithelial cells in relation to aging and smoking. *54*, 492–498, Number: 4 Publisher: Nature Publishing Group.

[529] Grossmann, S.; Hooks, Y.; Wilson, L.; Moore, L.; O'Neill, L.; Martincorena, I.; Voet, T.; Stratton, M. R.; Heer, R.; Campbell, P. J. Development, maturation, and maintenance of human prostate inferred from somatic mutations. *Cell Stem Cell* **2021**, *28*, 1262–1274.

[530] Hernando, B.; Dietzen, M.; Parra, G.; Gil-Barrachina, M.; Pitarch, G.; Mahiques, L.; Valcuende-Cavero, F.; McGranahan, N.; Martinez-Cadenas, C. The effect of age on the acquisition and selection of cancer driver mutations in sun-exposed normal skin. *Annals of Oncology* **2021**, *32*, 412–421.

[531] Wang, Y. *et al.* APOBEC mutagenesis is a common process in normal human small intestine. *55*, 246–254, Number: 2 Publisher: Nature Publishing Group.

Tissue	Other Information	Exome Rate	Genome Rate	Mutational Signatures	Positive Selection of Genes	Method	Cite
Appendix			56 per year	SBS1,SBS5, SBS18, SBS32, SBS35,SBS88		LCM	[388]
Bile Duct			9 per year	SBS1,SBS5, SBS18		LCM	[388]
Blood	Mature granulocytes		19.8 per year			NanoSeq	[26]
Blood	HSC/MPPs		19.9 per year			NanoSeq	[26]
Blood	HSPCs		16 per year	SBS1, SBSblood		Clonal expansion	[518]
Blood	B-cells		Naïve: 15 per cell per year Memory: 17 per cell per year	SBS1, SBSblood, SBS7a, SBS8, SBS9, SBS17b	ACTG1	Clonal expansion	[518]
Blood	HSC/MPPs		17 per year		DNMT3A, ZNF318, HIST2H3D	Clonal expansion	[519]
Blood	HSPCs		14.2 per year	SBS1, SBS5, SBS32		Clonal expansion	[520]
Blood	T-cells		Naïve: 22 per cell per year Memory: 25 per cell per year	SBS1, SBSblood, Memory: SBS7a,SBS8, SBS9, SBS17b	ACTG1	Clonal expansion	[518]
Brain	Neurons		17.1 per year	SBS5, SBS16,SBS1		NanoSeq	[26]

Brain	Bulk Seq		20-60			FACS sorted WGS/bulk sequencing	[521]
Brain	Neuron: HDG		~40 sSNVs per year	SBS5		single cell WGS	[17]
Brain	Neuron: PFC		~23 sSNVs per year	SBS5		single cell WGS	[17]
Brain	Neuron: PFC		16 SNV per year	SBS1		single cell WGs	[522]
Cardia		39 per exome		SBS5, SBS1, SBS4	MUC6, ARID1A	LCM and mini-bulk exome sequencing	[395]
Colon	Epithelium		similar to estimates of colonic stem cells	SBS1, SBS5 and colibactin signature		NanoSeq	[523]
Colon	ASC		36 per year	SBS1, SBS5		Clonal expansion	[16]
Colon			1,508 to 15,329	SBS1, SBS5, SBS18	AXIN2, STAG2, PIK3CA, ERBB2, ERBB3, FBXW7	LCM	[387]
Colon		25 per exome		SBS5, SBS1, SBS45	NOTCH3,BCOR, SETBP1	LCM and mini-bulk exome sequencing	[395]
Colon		0.019 per Mb per year		age-related	PIGR (UC tissue: NFKBIZ, ZC3H12A)	WES	[13]

Endometrium			29 SBS per gland per year	SBS5, SBS40, SBS18, SBS23, SBS1	PIK3CA, PIK3R1, ARHGAP35, FBXW7, ZFHX3, FOXA2, ERBB2, CHD4, KRAS, SPOP, PPP2R1A and ERBB3	LCM and WGS	[388]
Endometrium		128 per exome				LCM and WXS	[524]
Esophagus		31.5 per exome		SBS5, SBS1,SBS22, SBS4	NOTCH1, FAT1, TP53	LCM and mini-bulk exome sequencing	[395]
Esophagus			100s-1000s over lifetime	SBS1, SBS5, SBS16	NOTCH1, TP53	Ultra-deep targeted sequencing	[385]
Esophagus			41.5 per year	SBS2, SBS13, SBS1, SBS16	NOTCH1, PPMID	WES	[386]
Fibroblasts	Senescent		2618 per cell			single cell WGS	[525]
Fibroblasts	Early Passage		1103 per cell			single cell WGS	[525]
Gastric Gland			25 per year			LCM	[388]
Kidney	Kidney Tubules (KT)		KT1: 11.7 SNsV per year KT2: 56.6 SNVs per year	SBS1, SBS5, SBS3/8, SBS40, SBS18, SBS7a, tumor specific SBS2+SBS17		Clonal expansion and WGS	[42]

Kidney	SAT		17.5 SNVs per year	SBS1, SBS5, SBS3/8, SBS40, SBS18, SBS7a, tumor specific SBS2+SBS17		Clonal expansion and WGS	[42]
Kidney	VAT		27.2 SNVs per year	SBS1, SBS5, SBS3/8, SBS40, SBS18, SBS7a, tumor specific SBS2+SBS17		Clonal expansion and WGS	[42]
Large Intestine			49 per year			LCM	[388]
Liver	Hepatocytes		21 per cell per mitosis	SBS5, SBS18 and SBS36		Clonal expansion	[526]
Liver	ASC		36 per year	SBS5		Clonal expansion	[16]
Liver		73 per exome		SBS4, SBS22, SBS4, SBS1	PTCH1, MUC6, APOB, KMT2D, AMER1, NOTCH2	LCM and mini-bulk exome sequencing	[395]
Liver			33 per year	SBS5, SigA		LCM	[527]
Lung	Bronchial epithelium	26 per exome		SBS5, SBS1, SBS22		LCM and mini-bulk exome sequencing	[395]
Lung	Bronchial epithelium		28 SNV per cell per year	SBS5, SBS36, SBS30, SBS39, SBS20, SBS2	FAM135B, CREBBP, BCR, APC, FAT1, PBRM1, FAT3, BRAF, LRP1B, PRKCB, SALL4, KEAP1	single cell WGS	[528]
Pancreas		11 per exome		SBS5, SBS1, SBS45		LCM and mini-bulk exome sequencing	[395]

Pancreas	Pancreatic acini		15 per year	SBS1,SBS5, SBS18		LCM	[388]
Prostate			19 per year	SBS1,SBS5, SBS18		LCM	[388]
Prostate			16 per year	SBS1, SBS5, and SBS40	FOXA1	LCM	[529]
Rectum		47 per exome		SBS5, SBS1	ERBB3, NOTCH1	LCM and mini-bulk exome sequencing	[395]
Skeletal muscle				SBS1,SBS5, SBS3/8, SBS40, SBS18, SBS7a, tumor-specific SBS2+SBS17		Clonal expansion and WGS	[42]
Skin		2-6 per Mb		UV signature	NOTCH1, NOTCH2, NOTCH3, FAT1, TP53, RBM10	Ultra-deep targeted sequencing	[20]
Skin		42.39 per exome		SBS7, SBS17a,	P53,NOTCH1, NOTCH2 and FAT1	Deep Sequencing	[530]
Small Intestine	ASC		36 per year	SBS1, SBS5		Clonal expansion	[16]
Small Intestine		41 per exome		SBS5, SBS1, SBS22		LCM and mini-bulk exome sequencing	[395]
Small Intestine		32 per exome		SBS3		LCM and mini-bulk exome sequencing	[395]
Small Intestine			49 per year	SBS1,SBS5, SBS18, SBS2		LCM	[388]

Small Intestine			51 per year	SBS1,SBS5,SBS18. some: SBS2, SBS13, SBS88, SBS35, SBS40, SBS17b		LCM and WGS	[531]
Smooth muscle			20.7 per year			NanoSeq	[523]
Stomach		32 per exome		SBS5, SBS1	MUC6, ARID1A	LCM and mini-bulk exome sequencing	[395]
Urothelium		40 per exome	1879 per genome or 500-2,000 per genome by middle age	SBS2+SBS13, SBS5	:KMT2D, KDM6A,ARID1A, RBM10, EP300, STAG2, NOTCH2, CDKN1A, CREBBP, FOXQ1, RHOA and ERCC2	LCM and WGS	[393]
Urothelium		2.2 per Mb (AA) and 0.22 per Mb (nonAA)		SBS22,SBS1 SBS5, SBS2 SBS13	KMT2D,KDM6A,ATM, CREBBP,FAT1, KMT2C	LCM and WXS	[394]

Table 5.1: **Summary of studies of somatic mutations in normal tissues.** HSC= Hematopoietic Stem Cells, MPP=Multi-Potent Progenitor cells, HDG= Hippocampal Dentate Gyrus, SAT= Subcutaneous Adipose Tissue, VAT= Visceral Adipose Tissue, PFC= PreFrontal Cortex, SBS= Single Base Substitutions, AA= Aristolochic Acid, LCM=Laser Capture Microdisection, WXS= Whole Exome Sequencing, WGS= Whole Genome Sequencing

Appendix B

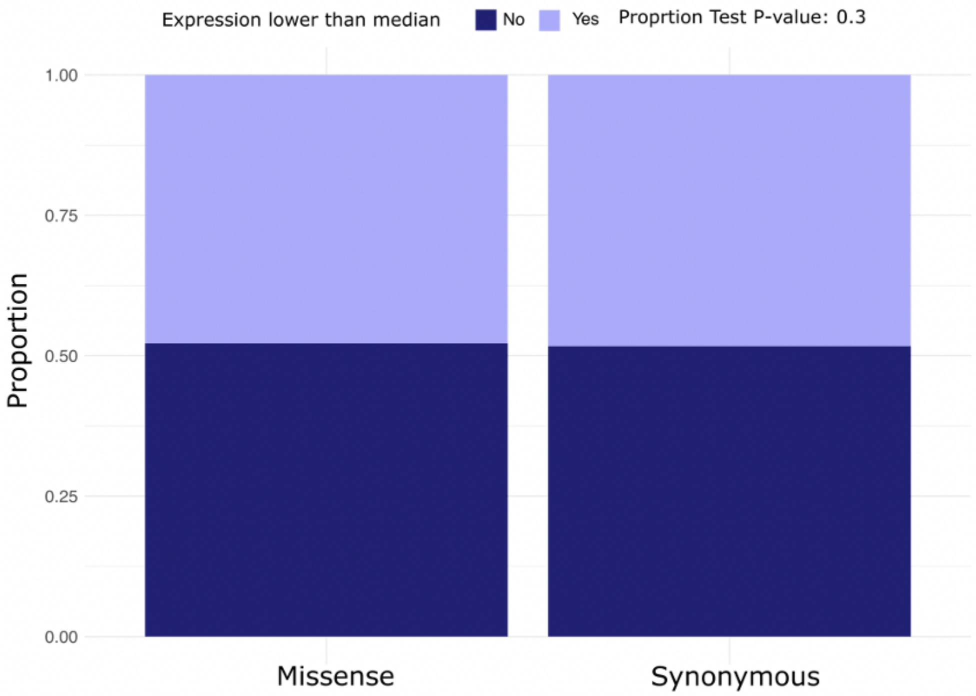


Figure 5.1: **Comparison of genes whose expression is lower than the median expression of genes for missense versus synonymous mutations when immune escaped samples are removed.** Stacked Bar plots comparing the conditional proportion of genes whose gene expression is lower than the median gene expression for individual samples for the two mutation types with immune escaped samples removed.

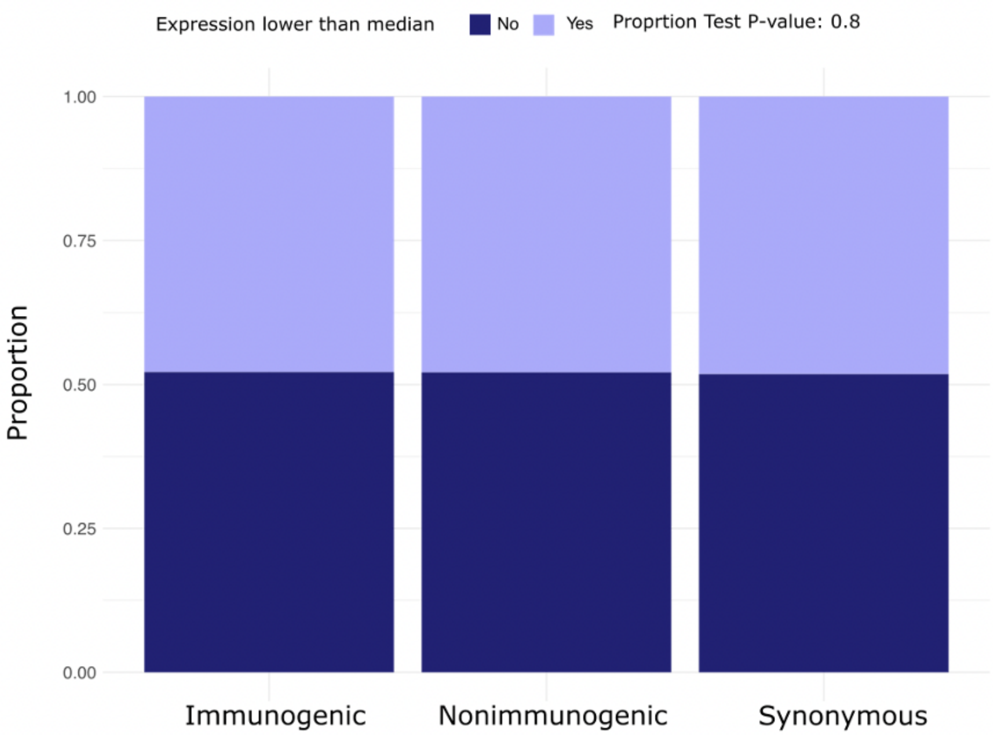


Figure 5.2: **Comparison of genes whose expression is lower than the median expression of genes with missense mutations split into those that are immunogeneic and non-immunogenic when immune escaped samples are removed.** Stacked Bar plots comparing the conditional proportion of genes whose gene expression is lower than the median gene expression for individual samples for missense mutations split into immunogenic and nonimmunogenic based on their PHBR score, compared to synonymous mutations types with immune escaped samples removed.

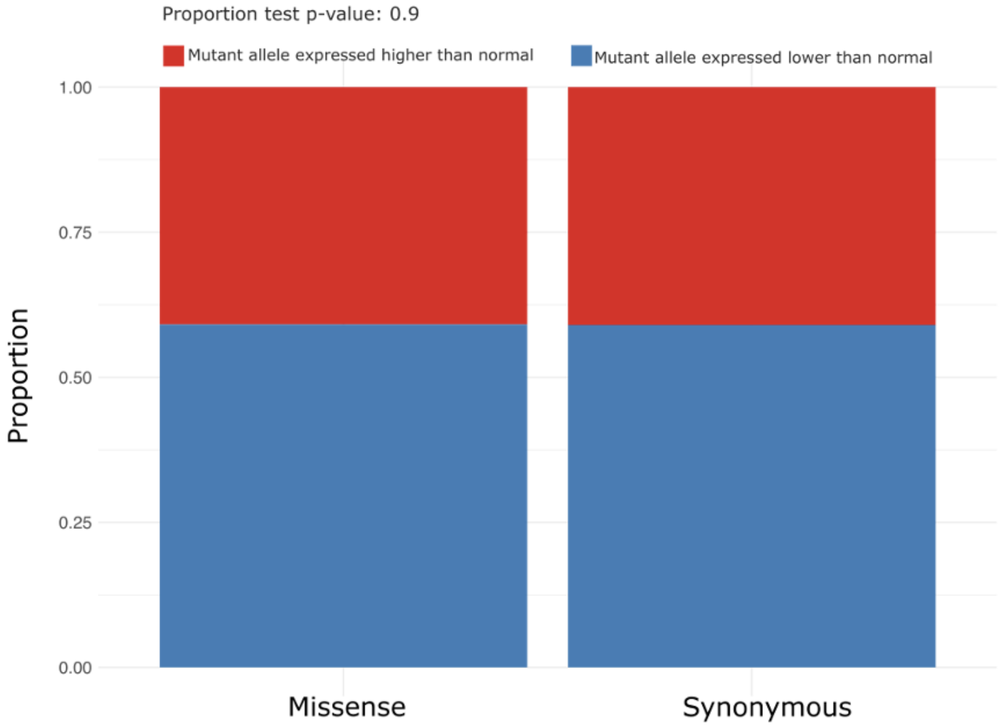


Figure 5.3: **Comparison of allele expression for mutations that are missense versus those that are synonymous when immune escaped samples are removed.** Stacked Bar plots comparing the conditional proportion of missense and synonymous mutants whose allele expression is greater than (red) or lower than (blue) the normal allele with immune escaped samples removed.

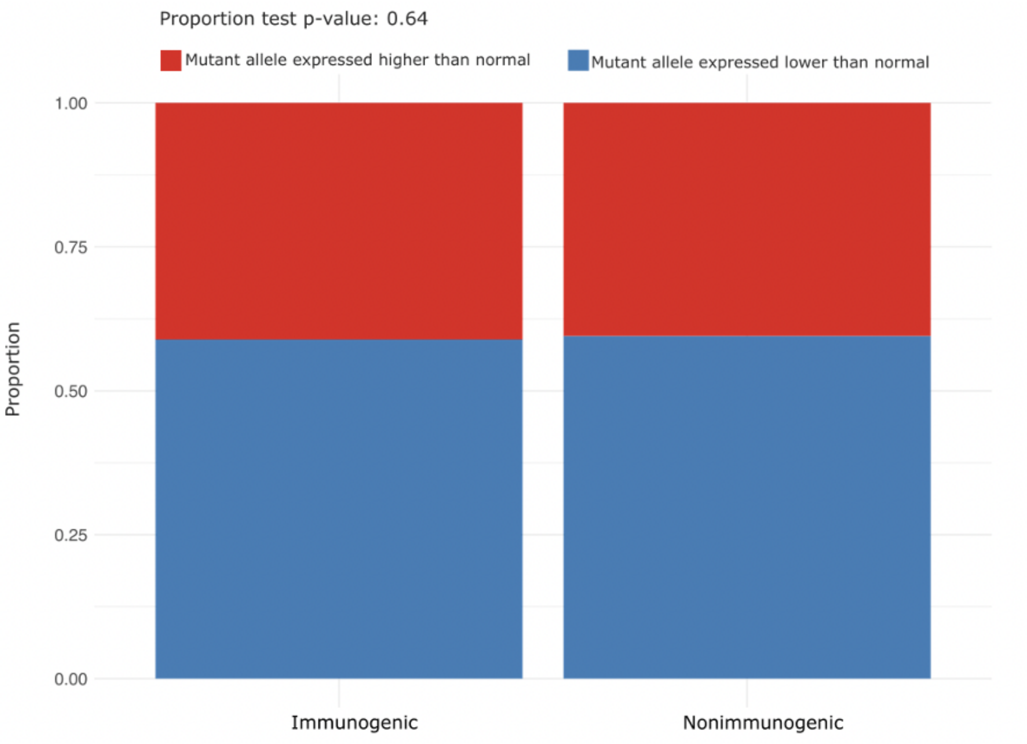


Figure 5.4: **Comparison of allele expression for missense mutations that are split into immunogenic and nonimmunogenic mutations when immune escaped samples are removed.** Stacked Bar plots comparing the conditional proportion of missense mutations split into immunogenic and nonimmunogenic based on their PHBR score whose allele expression is greater than (red) or lower than (blue) the normal allele.

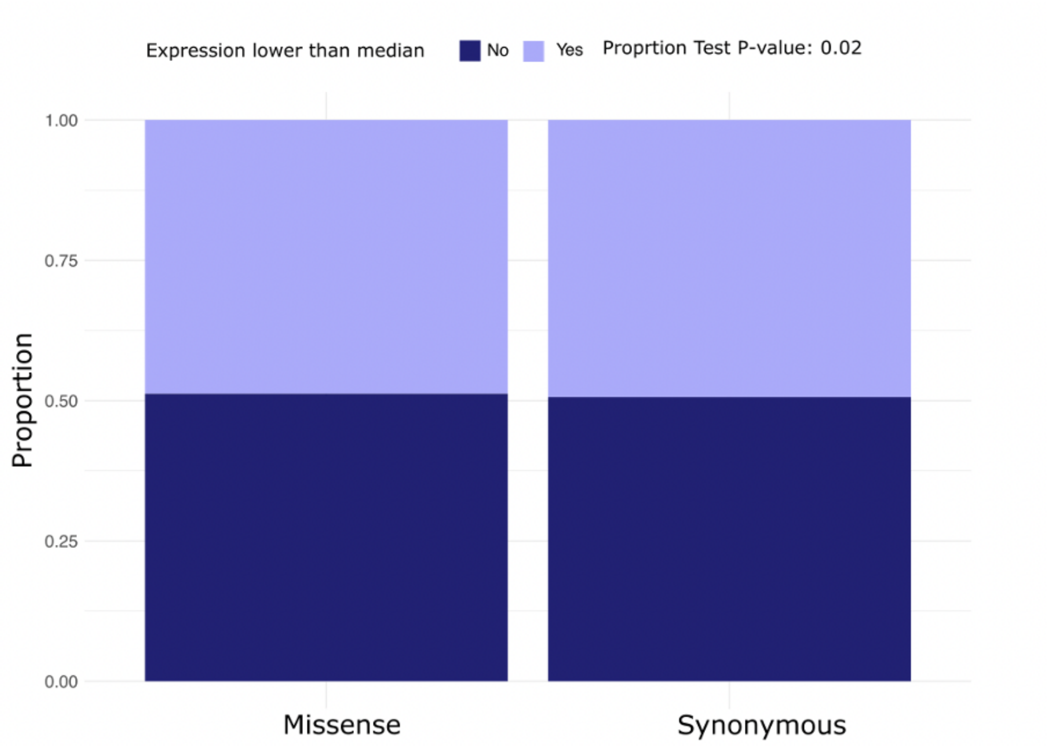


Figure 5.5: **Comparison of genes whose expression is lower than the median expression of genes for missense versus synonymous mutations when less stringent thresholds are applied for classifying clonal mutations.** Stacked Bar plots comparing the conditional proportion of genes whose gene expression is lower than the median gene expression for individual samples for the two mutation types with adjusted thresholds for calling a clonal variant.

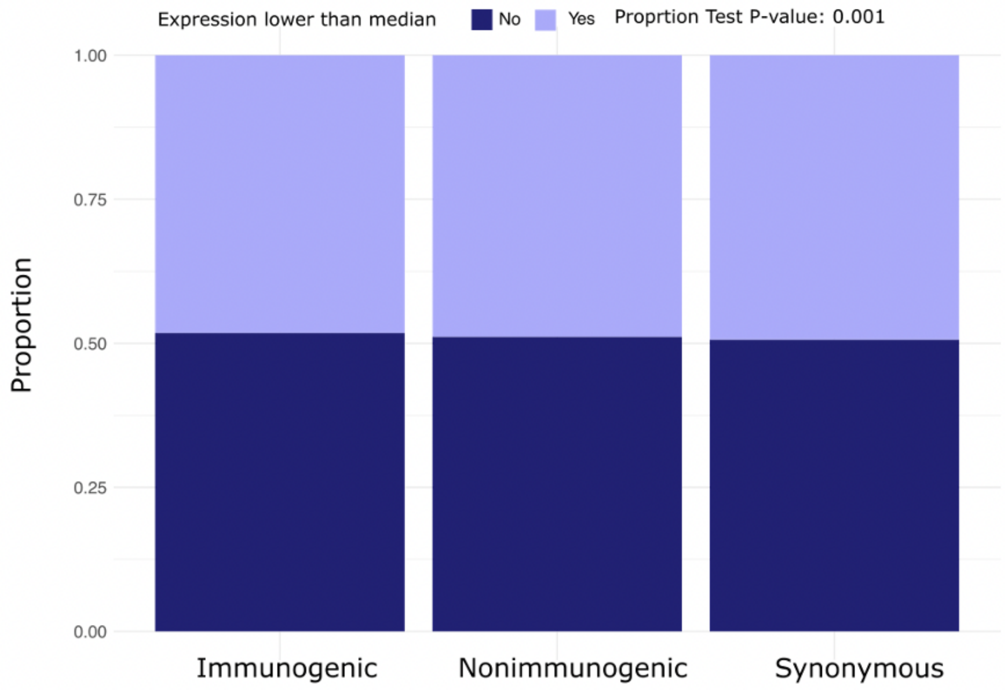


Figure 5.6: **Comparison of genes whose expression is lower than the median expression of genes with missense mutations split into those that are immunogeneic and non-immunogenic when less stringent thresholds are applied for classifying clonal mutations.** Stacked Bar plots comparing the conditional proportion of genes whose gene expression is lower than the median gene expression for individual samples for missense mutations split into immunogenic and nonimmunogenic based on their PHBR score, compared to synonymous mutations with adjusted thresholds for calling a clonal variant.

Appendix C

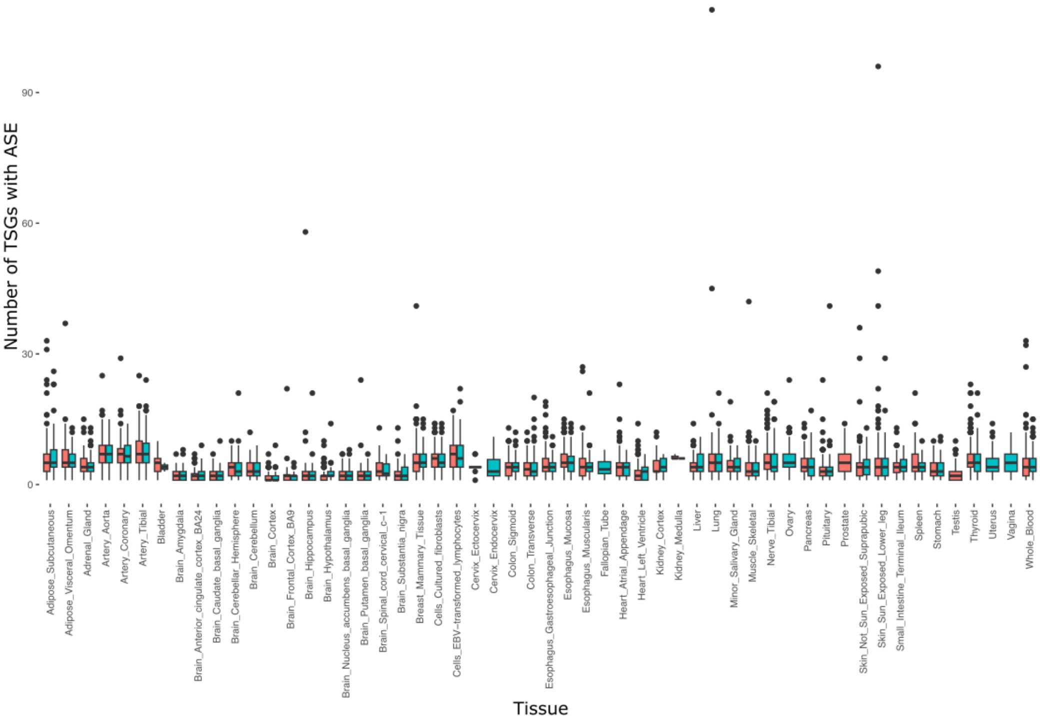


Figure 5.7: **Boxplot of the number of samples showing allele specific expression in tumor suppressor genes in normal tissues split by sex.** Red=male, Blue=female.

Tissue	**Threshold**		**Upper CI**	**AUC**	**Lower CI**	**Sens.**	**Spec.**
	Real	**Predicted**					
Adipose Subcutaneous	4	1	0.68	0.69	0.7	0.87	0.51
Adipose Subcutaneous	4	1.5	0.68	0.69	0.7	0.92	0.45
Adrenal Gland	4	0.5	0.69	0.7	0.71	0.78	0.62
Adrenal Gland	4	1	0.68	0.69	0.7	0.87	0.51
Adrenal Gland	4	1.5	0.69	0.7	0.71	0.92	0.48
Adrenal Gland	4	2	0.69	0.7	0.71	0.95	0.44
Adrenal Gland	4	2.5	0.68	0.69	0.7	0.96	0.42
Artery Coronary	4	1.5	0.68	0.69	0.7	0.92	0.46
Artery Coronary	4	2	0.68	0.69	0.7	0.95	0.44
Brain Cerebellar Hemisphere	3	1	0.7	0.7	0.71	0.86	0.55
Brain Cerebellar Hemisphere	3	1.5	0.7	0.71	0.72	0.91	0.51
Brain Cerebellar Hemisphere	3	2	0.7	0.71	0.71	0.94	0.47

Brain Cerebellar Hemisphere	3	2.5	0.69	0.7	0.7	0.96	0.43
Brain Cerebellar Hemisphere	4	0.5	0.68	0.69	0.7	0.77	0.61
Brain Cerebellar Hemisphere	4	1	0.71	0.73	0.74	0.86	0.59
Brain Cerebellar Hemisphere	4	1.5	0.73	0.74	0.76	0.91	0.58
Brain Cerebellar Hemisphere	4	2	0.74	0.75	0.76	0.94	0.55
Brain Cerebellar Hemisphere	4	2.5	0.73	0.74	0.76	0.96	0.53
Brain Cerebellar Hemisphere	4	3	0.72	0.73	0.74	0.97	0.49
Brain Cerebellar Hemisphere	4	3.5	0.7	0.71	0.72	0.98	0.43
Brain Cerebellar Hemisphere	4	4	0.68	0.69	0.7	0.99	0.39
Brain Cerebellar Hemisphere	5	1.5	0.67	0.69	0.71	0.91	0.47
Brain Cerebellar Hemisphere	5	2	0.68	0.7	0.72	0.94	0.45
Brain Cerebellar Hemisphere	5	2.5	0.68	0.7	0.72	0.96	0.44
Brain Cerebellar Hemisphere	5	3	0.68	0.7	0.72	0.97	0.42
Brain Cerebellum	3	1	0.68	0.69	0.7	0.86	0.52
Brain Cerebellum	3	1.5	0.69	0.69	0.7	0.91	0.48
Brain Cerebellum	4	1	0.68	0.69	0.7	0.86	0.53
Brain Cerebellum	4	1.5	0.7	0.71	0.72	0.91	0.51
Brain Cerebellum	4	2	0.7	0.71	0.72	0.94	0.48
Brain Cerebellum	4	2.5	0.69	0.7	0.72	0.96	0.45
Cells Cultured fibroblasts	3	0.5	0.7	0.7	0.71	0.76	0.64
Cells Cultured fibroblasts	3	1	0.72	0.72	0.72	0.87	0.57
Cells Cultured fibroblasts	3	1.5	0.72	0.72	0.73	0.92	0.52
Cells Cultured fibroblasts	3	2	0.71	0.72	0.72	0.95	0.49
Cells Cultured fibroblasts	3	2.5	0.7	0.7	0.7	0.97	0.43

Cells Cultured fibroblasts	4	0.5	0.7	0.71	0.72	0.76	0.65
Cells Cultured fibroblasts	4	1	0.72	0.73	0.74	0.87	0.59
Cells Cultured fibroblasts	4	1.5	0.74	0.75	0.76	0.92	0.57
Cells Cultured fibroblasts	4	2	0.74	0.75	0.76	0.95	0.55
Cells Cultured fibroblasts	4	2.5	0.73	0.74	0.75	0.97	0.51
Cells Cultured fibroblasts	4	3	0.72	0.72	0.73	0.98	0.47
Cells Cultured fibroblasts	4	3.5	0.7	0.71	0.72	0.98	0.43
Cells Cultured fibroblasts	5	1	0.68	0.69	0.7	0.87	0.51
Cells Cultured fibroblasts	5	1.5	0.7	0.71	0.72	0.92	0.5
Cells Cultured fibroblasts	5	2	0.71	0.72	0.73	0.95	0.49
Cells Cultured fibroblasts	5	2.5	0.7	0.71	0.72	0.97	0.46
Cells Cultured fibroblasts	5	3	0.68	0.7	0.71	0.98	0.42
Cells Cultured fibroblasts	6	1.5	0.67	0.69	0.71	0.92	0.46
Cells Cultured fibroblasts	6	2	0.68	0.7	0.72	0.95	0.45
Cells Cultured fibroblasts	6	2.5	0.68	0.7	0.72	0.97	0.44
Colon Sigmoid	4	1.5	0.68	0.69	0.7	0.92	0.46
Colon Sigmoid	4	2	0.68	0.69	0.7	0.95	0.44
Colon Transverse	3	1	0.69	0.69	0.7	0.88	0.5
Colon Transverse	4	0.5	0.69	0.7	0.7	0.79	0.6
Colon Transverse	4	1	0.71	0.72	0.73	0.88	0.56
Colon Transverse	4	1.5	0.71	0.72	0.73	0.93	0.51
Colon Transverse	4	2	0.71	0.72	0.72	0.95	0.48
Colon Transverse	4	2.5	0.68	0.69	0.7	0.97	0.42
Colon Transverse	5	1	0.68	0.69	0.71	0.88	0.5
Colon Transverse	5	1.5	0.68	0.69	0.71	0.93	0.46

Colon Transverse	5	2	0.68	0.69	0.71	0.95	0.44
Esophagus Gastroesophageal Junction	4	1	0.69	0.7	0.71	0.87	0.53
Esophagus Gastroesophageal Junction	4	1.5	0.7	0.71	0.72	0.92	0.49
Esophagus Gastroesophageal Junction	4	2	0.69	0.7	0.71	0.95	0.46
Esophagus Gastroesophageal Junction	4	2.5	0.69	0.7	0.7	0.97	0.42
Esophagus Mucosa	3	1	0.69	0.69	0.7	0.87	0.52
Esophagus Mucosa	3	1.5	0.68	0.69	0.69	0.92	0.46
Esophagus Mucosa	4	0.5	0.7	0.71	0.71	0.76	0.66
Esophagus Mucosa	4	1	0.73	0.74	0.74	0.87	0.61
Esophagus Mucosa	4	1.5	0.73	0.74	0.75	0.92	0.56
Esophagus Mucosa	4	2	0.72	0.72	0.73	0.95	0.5
Esophagus Mucosa	4	2.5	0.7	0.7	0.71	0.97	0.44
Esophagus Mucosa	5	1	0.69	0.71	0.72	0.87	0.55
Esophagus Mucosa	5	1.5	0.7	0.71	0.73	0.92	0.51
Esophagus Mucosa	5	2	0.69	0.71	0.72	0.95	0.46
Esophagus Mucosa	5	2.5	0.68	0.69	0.7	0.97	0.42
Esophagus Muscularis	3	1	0.69	0.69	0.7	0.87	0.51
Esophagus Muscularis	4	1	0.7	0.71	0.72	0.87	0.55
Esophagus Muscularis	4	1.5	0.71	0.72	0.73	0.92	0.52
Esophagus Muscularis	4	2	0.7	0.7	0.71	0.95	0.46
Esophagus Muscularis	4	2.5	0.68	0.69	0.7	0.97	0.42
Heart Atrial Appendage	5	1.5	0.68	0.69	0.7	0.92	0.46

Heart Atrial Appendage	5	2	0.68	0.7	0.71	0.95	0.45
Heart Atrial Appendage	5	2.5	0.68	0.7	0.71	0.97	0.43
Heart Atrial Appendage	5	3	0.68	0.69	0.71	0.98	0.41
Heart Atrial Appendage	5	3.5	0.68	0.69	0.7	0.98	0.4
Heart Atrial Appendage	6	1	0.68	0.71	0.73	0.87	0.54
Heart Atrial Appendage	6	1.5	0.7	0.72	0.75	0.92	0.52
Heart Atrial Appendage	6	2	0.71	0.73	0.75	0.95	0.51
Heart Atrial Appendage	6	2.5	0.71	0.73	0.75	0.97	0.5
Heart Atrial Appendage	6	3	0.71	0.73	0.75	0.98	0.49
Heart Atrial Appendage	6	3.5	0.71	0.73	0.75	0.98	0.48
Heart Atrial Appendage	6	4	0.71	0.73	0.76	0.99	0.48
Heart Atrial Appendage	6	4.5	0.69	0.72	0.74	0.99	0.44
Heart Left Ventricle	4	1	0.7	0.71	0.72	0.88	0.55
Heart Left Ventricle	4	1.5	0.71	0.72	0.73	0.92	0.52
Heart Left Ventricle	4	2	0.71	0.72	0.73	0.95	0.48
Heart Left Ventricle	4	2.5	0.7	0.71	0.72	0.97	0.46
Heart Left Ventricle	4	3	0.69	0.7	0.71	0.98	0.43
Heart Left Ventricle	4	3.5	0.68	0.7	0.7	0.98	0.41
Heart Left Ventricle	5	1	0.68	0.7	0.71	0.88	0.52
Heart Left Ventricle	5	1.5	0.7	0.71	0.73	0.92	0.5
Heart Left Ventricle	5	2	0.7	0.72	0.73	0.95	0.48
Heart Left Ventricle	5	2.5	0.7	0.72	0.73	0.97	0.47
Heart Left Ventricle	5	3	0.7	0.71	0.73	0.98	0.45
Heart Left Ventricle	5	3.5	0.7	0.71	0.73	0.98	0.44

Heart Left Ventricle	5	4	0.68	0.7	0.71	0.99	0.41
Heart Left Ventricle	6	1	0.68	0.7	0.73	0.88	0.53
Heart Left Ventricle	6	1.5	0.7	0.72	0.74	0.92	0.52
Heart Left Ventricle	6	2	0.7	0.73	0.75	0.95	0.5
Heart Left Ventricle	6	2.5	0.7	0.73	0.75	0.97	0.49
Heart Left Ventricle	6	3	0.71	0.73	0.76	0.98	0.48
Heart Left Ventricle	6	3.5	0.7	0.73	0.75	0.98	0.47
Heart Left Ventricle	6	4	0.69	0.72	0.74	0.99	0.45
Heart Left Ventricle	6	4.5	0.69	0.71	0.74	0.99	0.43
Heart Left Ventricle	6	5	0.67	0.69	0.72	0.99	0.39
Heart Left Ventricle	6	5.5	0.67	0.69	0.72	0.99	0.39
Heart Left Ventricle	6	6	0.66	0.69	0.71	1	0.38
Kidney Cortex	4	2	0.67	0.69	0.72	0.95	0.43
Kidney Cortex	4	2.5	0.67	0.69	0.72	0.97	0.42
Liver	3	1	0.68	0.69	0.7	0.88	0.5
Liver	3	1.5	0.68	0.69	0.7	0.92	0.46
Liver	4	0.5	0.7	0.72	0.73	0.8	0.64
Liver	4	1	0.73	0.74	0.75	0.88	0.6
Liver	4	1.5	0.74	0.75	0.76	0.92	0.58
Liver	4	2	0.74	0.75	0.76	0.95	0.55
Liver	4	2.5	0.73	0.74	0.75	0.96	0.51
Liver	4	3	0.7	0.72	0.73	0.98	0.46
Liver	4	3.5	0.69	0.7	0.72	0.98	0.42
Liver	5	0.5	0.67	0.7	0.72	0.8	0.6
Liver	5	1	0.69	0.72	0.74	0.88	0.56
Liver	5	1.5	0.71	0.73	0.75	0.92	0.54
Liver	5	2	0.72	0.74	0.76	0.95	0.53
Liver	5	2.5	0.71	0.74	0.76	0.96	0.5
Liver	5	3	0.69	0.71	0.73	0.98	0.45
Liver	5	3.5	0.68	0.7	0.72	0.98	0.41
Liver	6	2	0.66	0.7	0.74	0.95	0.45
Liver	6	2.5	0.66	0.7	0.74	0.96	0.44
Minor Salivary Gland	4	0.5	0.7	0.72	0.73	0.81	0.62
Minor Salivary Gland	4	1	0.68	0.7	0.71	0.88	0.52

Minor Salivary Gland	4	1.5	0.69	0.71	0.72	0.92	0.49
Minor Salivary Gland	4	2	0.69	0.7	0.72	0.95	0.46
Minor Salivary Gland	4	2.5	0.68	0.7	0.71	0.96	0.43
Minor Salivary Gland	5	0.5	0.72	0.74	0.77	0.81	0.67
Minor Salivary Gland	5	1.5	0.67	0.7	0.72	0.92	0.47
Minor Salivary Gland	5	2	0.67	0.69	0.72	0.95	0.44
Minor Salivary Gland	5	2.5	0.67	0.69	0.72	0.96	0.42
Minor Salivary Gland	6	0.5	0.67	0.71	0.76	0.81	0.62
Nerve Tibial	4	1	0.7	0.71	0.72	0.86	0.56
Nerve Tibial	4	1.5	0.71	0.72	0.72	0.92	0.52
Nerve Tibial	4	2	0.71	0.72	0.72	0.95	0.49
Nerve Tibial	4	2.5	0.71	0.71	0.72	0.97	0.46
Nerve Tibial	4	3	0.69	0.7	0.71	0.98	0.42
Nerve Tibial	5	1	0.69	0.7	0.71	0.86	0.53
Nerve Tibial	5	1.5	0.7	0.72	0.73	0.92	0.51
Nerve Tibial	5	2	0.71	0.72	0.73	0.95	0.5
Nerve Tibial	5	2.5	0.71	0.72	0.73	0.96	0.48
Nerve Tibial	5	3	0.7	0.71	0.73	0.98	0.45
Nerve Tibial	6	1.5	0.68	0.7	0.72	0.92	0.48
Nerve Tibial	6	2	0.68	0.71	0.73	0.95	0.46
Nerve Tibial	6	2.5	0.68	0.7	0.73	0.96	0.44
Nerve Tibial	6	3	0.68	0.7	0.72	0.98	0.43
Ovary	4	1	0.68	0.69	0.71	0.88	0.51
Ovary	4	1.5	0.69	0.7	0.72	0.92	0.49
Ovary	4	2	0.69	0.7	0.72	0.95	0.46
Pancreas	4	0.5	0.69	0.7	0.71	0.77	0.63
Pancreas	4	1	0.72	0.73	0.74	0.87	0.59
Pancreas	4	1.5	0.73	0.74	0.75	0.92	0.56
Pancreas	4	2	0.72	0.73	0.74	0.95	0.52
Pancreas	4	2.5	0.71	0.72	0.72	0.96	0.47
Pancreas	5	0.5	0.69	0.7	0.72	0.77	0.63
Pancreas	5	1	0.71	0.73	0.74	0.87	0.59
Pancreas	5	1.5	0.73	0.74	0.76	0.92	0.57
Pancreas	5	2	0.73	0.74	0.76	0.95	0.54
Pancreas	5	2.5	0.7	0.72	0.74	0.96	0.48
Pancreas	5	3	0.68	0.69	0.71	0.98	0.41
Pancreas	6	0.5	0.69	0.71	0.74	0.77	0.66
Pancreas	6	1	0.7	0.73	0.75	0.87	0.59
Pancreas	6	1.5	0.72	0.74	0.77	0.92	0.57
Pancreas	6	2	0.72	0.75	0.78	0.95	0.55

Pancreas	6	2.5	0.7	0.73	0.75	0.96	0.49
Pancreas	6	3	0.68	0.71	0.74	0.98	0.45
Pancreas	6	3.5	0.68	0.7	0.73	0.98	0.42
Pituitary	6	0.5	0.68	0.7	0.72	0.78	0.62
Skin Not Sun Exposed Suprapubic	4	1	0.69	0.7	0.71	0.87	0.53
Skin Not Sun Exposed Suprapubic	4	1.5	0.7	0.71	0.72	0.92	0.5
Skin Not Sun Exposed Suprapubic	4	2	0.7	0.71	0.72	0.95	0.47
Skin Not Sun Exposed Suprapubic	4	2.5	0.7	0.71	0.71	0.97	0.45
Skin Not Sun Exposed Suprapubic	4	3	0.69	0.69	0.7	0.98	0.41
Skin Not Sun Exposed Suprapubic	5	1	0.69	0.7	0.71	0.87	0.53
Skin Not Sun Exposed Suprapubic	5	1.5	0.71	0.72	0.73	0.92	0.52
Skin Not Sun Exposed Suprapubic	5	2	0.71	0.72	0.73	0.95	0.49
Skin Not Sun Exposed Suprapubic	5	2.5	0.71	0.72	0.73	0.97	0.48
Skin Not Sun Exposed Suprapubic	5	3	0.7	0.71	0.72	0.98	0.45
Skin Not Sun Exposed Suprapubic	5	3.5	0.69	0.7	0.72	0.98	0.42
Skin Not Sun Exposed Suprapubic	5	4	0.68	0.7	0.71	0.99	0.4
Skin Not Sun Exposed Suprapubic	6	1	0.67	0.69	0.72	0.87	0.51
Skin Not Sun Exposed Suprapubic	6	1.5	0.69	0.71	0.73	0.92	0.5
Skin Not Sun Exposed Suprapubic	6	2	0.69	0.71	0.74	0.95	0.48

Skin Not Sun Exposed Suprapubic	6	2.5	0.7	0.72	0.74	0.97	0.47
Skin Not Sun Exposed Suprapubic	6	3	0.69	0.71	0.73	0.98	0.44
Skin Not Sun Exposed Suprapubic	6	3.5	0.68	0.7	0.72	0.98	0.42
Skin Not Sun Exposed Suprapubic	6	4	0.67	0.69	0.71	0.99	0.4
Skin Sun Exposed Lower leg	5	1.5	0.68	0.69	0.7	0.92	0.46
Skin Sun Exposed Lower leg	5	2	0.68	0.7	0.7	0.95	0.44
Skin Sun Exposed Lower leg	5	2.5	0.68	0.69	0.7	0.97	0.42
Skin Sun Exposed Lower leg	6	1.5	0.67	0.69	0.71	0.92	0.46
Skin Sun Exposed Lower leg	6	2	0.68	0.7	0.71	0.95	0.44
Skin Sun Exposed Lower leg	6	2.5	0.67	0.69	0.71	0.97	0.42
Skin Sun Exposed Lower leg	6	3	0.67	0.69	0.71	0.98	0.4
Small Intestine Terminal Ileum	4	1	0.68	0.69	0.7	0.88	0.5
Small Intestine Terminal Ileum	4	1.5	0.68	0.69	0.71	0.93	0.46
Spleen	4	1	0.7	0.71	0.72	0.86	0.56
Spleen	4	1.5	0.7	0.72	0.73	0.92	0.51
Spleen	4	2	0.7	0.71	0.72	0.95	0.47
Spleen	4	2.5	0.69	0.7	0.71	0.96	0.44
Spleen	4	3	0.68	0.69	0.7	0.97	0.41
Spleen	5	1	0.67	0.69	0.71	0.86	0.52
Spleen	5	1.5	0.68	0.7	0.72	0.92	0.48
Spleen	5	2	0.67	0.69	0.71	0.95	0.43
Spleen	5	2.5	0.67	0.69	0.71	0.96	0.41
Stomach	4	1	0.68	0.69	0.7	0.88	0.5
Stomach	4	1.5	0.68	0.7	0.7	0.93	0.46

Stomach	4	2	0.69	0.7	0.71	0.95	0.44
Testis	3	0.5	0.68	0.69	0.69	0.77	0.61
Testis	3	1	0.71	0.71	0.72	0.86	0.56
Testis	3	1.5	0.71	0.71	0.72	0.92	0.51
Testis	3	2	0.71	0.71	0.72	0.94	0.48
Testis	3	2.5	0.7	0.7	0.71	0.96	0.44
Testis	4	0.5	0.72	0.73	0.74	0.77	0.7
Testis	4	1	0.76	0.77	0.78	0.86	0.68
Testis	4	1.5	0.78	0.78	0.8	0.91	0.66
Testis	4	2	0.78	0.79	0.8	0.94	0.64
Testis	4	2.5	0.78	0.79	0.8	0.96	0.62
Testis	4	3	0.77	0.78	0.79	0.97	0.59
Testis	4	3.5	0.75	0.76	0.77	0.98	0.53
Testis	4	4	0.72	0.73	0.74	0.99	0.47
Testis	5	0.5	0.71	0.72	0.74	0.77	0.68
Testis	5	1	0.75	0.77	0.78	0.86	0.68
Testis	5	1.5	0.77	0.79	0.8	0.91	0.66
Testis	5	2	0.78	0.8	0.82	0.94	0.65
Testis	5	2.5	0.79	0.8	0.82	0.96	0.65
Testis	5	3	0.79	0.8	0.82	0.97	0.63
Testis	5	3.5	0.77	0.79	0.81	0.98	0.6
Testis	5	4	0.75	0.77	0.78	0.99	0.55
Testis	5	4.5	0.7	0.72	0.73	0.99	0.44
Testis	5	5	0.67	0.69	0.71	0.99	0.39
Testis	6	0.5	0.7	0.72	0.75	0.77	0.68
Testis	6	1	0.74	0.77	0.8	0.86	0.68
Testis	6	1.5	0.76	0.79	0.82	0.91	0.67
Testis	6	2	0.78	0.8	0.83	0.94	0.66
Testis	6	2.5	0.78	0.81	0.84	0.96	0.66
Testis	6	3	0.79	0.81	0.84	0.97	0.65
Testis	6	3.5	0.78	0.81	0.84	0.98	0.64
Testis	6	4	0.78	0.81	0.83	0.99	0.62
Testis	6	4.5	0.75	0.78	0.81	0.99	0.57
Testis	6	5	0.74	0.76	0.79	0.99	0.54
Testis	6	5.5	0.72	0.75	0.78	1	0.5
Testis	6	6	0.72	0.75	0.78	1	0.5
Testis	6	6.5	0.71	0.74	0.77	1	0.48
Testis	6	7	0.7	0.73	0.76	1	0.46
Testis	6	7.5	0.7	0.73	0.76	1	0.46
Testis	6	8	0.7	0.73	0.76	1	0.46
Testis	6	8.5	0.67	0.7	0.73	1	0.4
Thyroid	3	1	0.69	0.69	0.69	0.86	0.52
Thyroid	4	0.5	0.69	0.7	0.71	0.75	0.65
Thyroid	4	1	0.71	0.72	0.72	0.86	0.57
Thyroid	4	1.5	0.72	0.72	0.73	0.92	0.53
Thyroid	4	2	0.71	0.72	0.73	0.95	0.49
Thyroid	4	2.5	0.7	0.71	0.72	0.96	0.45
Thyroid	4	3	0.69	0.7	0.7	0.98	0.42

Thyroid	5	0.5	0.68	0.69	0.7	0.75	0.63
Thyroid	5	1.5	0.68	0.7	0.7	0.92	0.47
Thyroid	5	2	0.69	0.7	0.71	0.95	0.45
Thyroid	5	2.5	0.69	0.7	0.7	0.96	0.42
Thyroid	5	3	0.68	0.69	0.7	0.98	0.4
Thyroid	6	0.5	0.68	0.69	0.7	0.75	0.63
Uterus	4	1.5	0.67	0.69	0.7	0.92	0.45
Uterus	4	2	0.68	0.69	0.71	0.95	0.44
Vagina	4	1	0.68	0.69	0.71	0.88	0.5
Vagina	4	1.5	0.68	0.7	0.71	0.92	0.47
Vagina	4	2	0.68	0.69	0.71	0.95	0.44
Whole Blood	4	1	0.69	0.69	0.7	0.88	0.51
Whole Blood	4	1.5	0.69	0.69	0.7	0.92	0.46
Whole Blood	5	0.5	0.68	0.69	0.7	0.77	0.62
Whole Blood	5	1	0.71	0.72	0.73	0.88	0.57
Whole Blood	5	1.5	0.72	0.73	0.74	0.92	0.53
Whole Blood	5	2	0.71	0.72	0.73	0.95	0.48
Whole Blood	5	2.5	0.69	0.7	0.71	0.97	0.44
Whole Blood	6	1	0.7	0.72	0.73	0.88	0.56
Whole Blood	6	1.5	0.71	0.72	0.74	0.92	0.52
Whole Blood	6	2	0.7	0.72	0.73	0.95	0.48
Whole Blood	6	2.5	0.68	0.69	0.71	0.97	0.42

Table 5.2: **Top 5% best AUC predictions.**CI= Confidence Interval, Sens=Sensitivity, Spec=Specificity

Tissue	**Threshold**		**Upper CI**	**AUC**	**Lower CI**	**Sens.**	**Spec.**
	Real	**Predicted**					
Adipose Subcutaneous	1	8	0.5	0.5	0.5	1	0.01
Adipose Subcutaneous	1	8.5	0.5	0.5	0.5	1	0.01
Adipose Subcutaneous	1	9	0.5	0.5	0.5	1	0.01
Adipose Subcutaneous	1	9.5	0.5	0.5	0.5	1	0
Adipose Subcutaneous	1	10	0.5	0.5	0.5	1	0
Adipose Subcutaneous	2	10	0.5	0.5	0.5	1	0.01
Adipose Visceral Omentum	1	8	0.5	0.5	0.5	1	0.01
Adipose Visceral Omentum	1	8.5	0.5	0.5	0.5	1	0.01

Adipose Visceral Omentum	1	9	0.5	0.5	0.5	1	0.01
Adipose Visceral Omentum	1	9.5	0.5	0.5	0.5	1	0
Adipose Visceral Omentum	1	10	0.5	0.5	0.5	1	0
Adrenal Gland	1	8	0.5	0.5	0.5	1	0.01
Adrenal Gland	1	8.5	0.5	0.5	0.5	1	0.01
Adrenal Gland	1	9	0.5	0.5	0.5	1	0.01
Adrenal Gland	1	9.5	0.5	0.5	0.5	1	0
Adrenal Gland	1	10	0.5	0.5	0.5	1	0
Artery Aorta	1	8	0.5	0.5	0.5	1	0.01
Artery Aorta	1	8.5	0.5	0.5	0.5	1	0.01
Artery Aorta	1	9	0.5	0.5	0.5	1	0
Artery Aorta	1	9.5	0.5	0.5	0.5	1	0
Artery Aorta	1	10	0.5	0.5	0.5	1	0
Artery Coronary	1	8	0.5	0.5	0.5	1	0.01
Artery Coronary	1	8.5	0.5	0.5	0.5	1	0.01
Artery Coronary	1	9	0.5	0.5	0.5	1	0.01
Artery Coronary	1	9.5	0.5	0.5	0.5	1	0.01
Artery Coronary	1	10	0.5	0.5	0.5	1	0
Artery Tibial	1	8	0.5	0.5	0.5	1	0.01
Artery Tibial	1	8.5	0.5	0.5	0.5	1	0.01
Artery Tibial	1	9	0.5	0.5	0.5	1	0.01
Artery Tibial	1	9.5	0.5	0.5	0.5	1	0
Artery Tibial	1	10	0.5	0.5	0.5	1	0
Brain Amygdala	1	7.5	0.5	0.5	0.5	1	0.01
Brain Amygdala	1	8	0.5	0.5	0.5	1	0.01
Brain Amygdala	1	8.5	0.5	0.5	0.5	1	0.01
Brain Amygdala	1	9	0.5	0.5	0.5	1	0
Brain Amygdala	1	9.5	0.5	0.5	0.5	1	0
Brain Amygdala	1	10	0.5	0.5	0.5	1	0
Brain Amygdala	6	0.5	0.46	0.49	0.52	0.81	0.17
Brain Amygdala	6	9	0.5	0.5	0.51	1	0.01
Brain Amygdala	6	9.5	0.5	0.5	0.5	1	0
Brain Amygdala	6	10	0.5	0.5	0.5	1	0
Brain Anterior cingulate cortex A24	1	7.5	0.5	0.5	0.5	1	0.01

Brain Anterior cingulate cortex BA24	1	8	0.5	0.5	0.5	1	0.01
Brain Anterior cingulate cortex BA24	1	8.5	0.5	0.5	0.5	1	0.01
Brain Anterior cingulate cortex A24	1	9	0.5	0.5	0.5	1	0
Brain Anterior cingulate cortex BA24	1	9.5	0.5	0.5	0.5	1	0
Brain Anterior cingulate cortex BA24	1	10	0.5	0.5	0.5	1	0
Brain Anterior cingulate cortex BA24	2	9.5	0.5	0.5	0.5	1	0.01
Brain Anterior cingulate cortex BA24	2	10	0.5	0.5	0.5	1	0.01
Brain Anterior cingulate cortex BA24	6	7.5	0.5	0.5	0.51	1	0.01
Brain Anterior cingulate cortex BA24	6	8	0.5	0.5	0.51	1	0.01
Brain Anterior cingulate cortex BA24	6	8.5	0.5	0.5	0.51	1	0.01
Brain Anterior cingulate cortex BA24	6	9	0.5	0.5	0.5	1	0
Brain Anterior cingulate cortex BA24	6	9.5	0.5	0.5	0.5	1	0
Brain Anterior cingulate cortex BA24	6	10	0.5	0.5	0.5	1	0
Brain Caudate basal ganglia	1	8	0.5	0.5	0.5	1	0.01
Brain Caudate basal ganglia	1	8.5	0.5	0.5	0.5	1	0.01
Brain Caudate basal ganglia	1	9	0.5	0.5	0.5	1	0
Brain Caudate basal ganglia	1	9.5	0.5	0.5	0.5	1	0
Brain Caudate basal ganglia	1	10	0.5	0.5	0.5	1	0

Brain Cerebellar Hemisphere	1	7.5	0.5	0.5	0.5	1	0.01
Brain Cerebellar Hemisphere	1	8	0.5	0.5	0.5	1	0.01
Brain Cerebellar Hemisphere	1	8.5	0.5	0.5	0.5	1	0
Brain Cerebellar Hemisphere	1	9	0.5	0.5	0.5	1	0
Brain Cerebellar Hemisphere	1	9.5	0.5	0.5	0.5	1	0
Brain Cerebellar Hemisphere	1	10	0.5	0.5	0.5	1	0
Brain Cerebellar Hemisphere	2	10	0.5	0.5	0.5	1	0.01
Brain Cerebellum	1	7.5	0.5	0.5	0.5	1	0.01
Brain Cerebellum	1	8	0.5	0.5	0.5	1	0.01
Brain Cerebellum	1	8.5	0.5	0.5	0.5	1	0
Brain Cerebellum	1	9	0.5	0.5	0.5	1	0
Brain Cerebellum	1	9.5	0.5	0.5	0.5	1	0
Brain Cerebellum	1	10	0.5	0.5	0.5	1	0
Brain Cerebellum	2	9.5	0.5	0.5	0.5	1	0.01
Brain Cerebellum	2	10	0.5	0.5	0.5	1	0.01
Brain Cortex	1	7.5	0.5	0.5	0.5	1	0.01
Brain Cortex	1	8	0.5	0.5	0.5	1	0.01
Brain Cortex	1	8.5	0.5	0.5	0.5	1	0.01
Brain Cortex	1	9	0.5	0.5	0.5	1	0.01
Brain Cortex	1	9.5	0.5	0.5	0.5	1	0
Brain Cortex	1	10	0.5	0.5	0.5	1	0
Brain Frontal Cortex BA9	1	7.5	0.5	0.5	0.5	1	0.01
Brain Frontal Cortex BA9	1	8	0.5	0.5	0.5	1	0.01
Brain Frontal Cortex BA9	1	8.5	0.5	0.5	0.5	1	0.01
Brain Frontal Cortex BA9	1	9	0.5	0.5	0.5	1	0
Brain Frontal Cortex BA9	1	9.5	0.5	0.5	0.5	1	0
Brain Frontal Cortex BA9	1	10	0.5	0.5	0.5	1	0

Brain Hippocampus	1	7.5	0.5	0.5	0.5	1	0.01
Brain Hippocampus	1	8	0.5	0.5	0.5	1	0.01
Brain Hippocampus	1	8.5	0.5	0.5	0.5	1	0.01
Brain Hippocampus	1	9	0.5	0.5	0.5	1	0.01
Brain Hippocampus	1	9.5	0.5	0.5	0.5	1	0
Brain Hippocampus	1	10	0.5	0.5	0.5	1	0
Brain Hypothalamus	1	7.5	0.5	0.5	0.5	1	0.01
Brain Hypothalamus	1	8	0.5	0.5	0.5	1	0.01
Brain Hypothalamus	1	8.5	0.5	0.5	0.5	1	0.01
Brain Hypothalamus	1	9	0.5	0.5	0.5	1	0.01
Brain Hypothalamus	1	9.5	0.5	0.5	0.5	1	0
Brain Hypothalamus	1	10	0.5	0.5	0.5	1	0
Brain Nucleus accumbens basal ganglia	1	7.5	0.5	0.5	0.5	1	0.01
Brain Nucleus accumbens basal ganglia	1	8	0.5	0.5	0.5	1	0.01
Brain Nucleus accumbens basal ganglia	1	8.5	0.5	0.5	0.5	1	0.01
Brain Nucleus accumbens basal ganglia	1	9	0.5	0.5	0.5	1	0.01
Brain Nucleus accumbens basal ganglia	1	9.5	0.5	0.5	0.5	1	0
Brain Nucleus accumbens basal ganglia	1	10	0.5	0.5	0.5	1	0
Brain Nucleus accumbens basal ganglia	6	0.5	0.48	0.5	0.52	0.8	0.2
Brain Nucleus accumbens basal ganglia	6	7.5	0.5	0.5	0.51	1	0.01

Brain Nucleus accumbens basal ganglia	6	8	0.5	0.5	0.51	1	0.01
Brain Nucleus accumbens basal ganglia	6	8.5	0.5	0.5	0.51	1	0.01
Brain Nucleus ccumbens basal ganglia	6	9	0.5	0.5	0.51	1	0.01
Brain Nucleus accumbens basal ganglia	6	9.5	0.5	0.5	0.51	1	0.01
Brain Nucleus accumbens basal ganglia	6	10	0.5	0.5	0.51	1	0.01
Brain Putamen basal ganglia	1	8	0.5	0.5	0.5	1	0.01
Brain Putamen basal ganglia	1	8.5	0.5	0.5	0.5	1	0.01
Brain Putamen basal ganglia	1	9	0.5	0.5	0.5	1	0.01
Brain Putamen basal ganglia	1	9.5	0.5	0.5	0.5	1	0
Brain Putamen basal ganglia	1	10	0.5	0.5	0.5	1	0
Brain Spinal cord cervical c-1	1	7.5	0.5	0.5	0.5	1	0.01
Brain Spinal cord cervical c-1	1	8	0.5	0.5	0.5	1	0.01
Brain Spinal cord cervical c-1	1	8.5	0.5	0.5	0.5	1	0.01
Brain Spinal cord cervical c-1	1	9	0.5	0.5	0.5	1	0
Brain Spinal cord cervical c-1	1	9.5	0.5	0.5	0.5	1	0
Brain Spinal cord cervical c-1	1	10	0.5	0.5	0.5	1	0
Brain Spinal cord cervical c-1	2	10	0.5	0.5	0.5	1	0.01
Brain Substantia nigra	1	7.5	0.5	0.5	0.5	1	0.01
Brain Substantia nigra	1	8	0.5	0.5	0.5	1	0.01
Brain Substantia nigra	1	8.5	0.5	0.5	0.5	1	0.01
Brain Substantia nigra	1	9	0.5	0.5	0.5	1	0.01

Brain Substantia nigra	1	9.5	0.5	0.5	0.5	1	0
Brain Substantia nigra	1	10	0.5	0.5	0.5	1	0
Brain Substantia nigra	2	10	0.5	0.5	0.5	1	0.01
Breast Mammary Tissue	1	8.5	0.5	0.5	0.5	1	0.01
Breast Mammary Tissue	1	9	0.5	0.5	0.5	1	0.01
Breast Mammary Tissue	1	9.5	0.5	0.5	0.5	1	0.01
Breast Mammary Tissue	1	10	0.5	0.5	0.5	1	0.01
Cells Cultured fibroblasts	1	8.5	0.5	0.5	0.5	1	0.01
Cells Cultured fibroblasts	1	9	0.5	0.5	0.5	1	0.01
Cells Cultured fibroblasts	1	9.5	0.5	0.5	0.5	1	0
Cells Cultured fibroblasts	1	10	0.5	0.5	0.5	1	0
Cells EBV transformed lymphocytes	1	7.5	0.5	0.5	0.5	1	0.01
Cells EBV transformed lymphocytes	1	8	0.5	0.5	0.5	1	0.01
Cells EBV transformed lymphocytes	1	8.5	0.5	0.5	0.5	1	0.01
Cells EBV transformed lymphocytes	1	9	0.5	0.5	0.5	1	0.01
Cells EBV transformed lymphocytes	1	9.5	0.5	0.5	0.5	1	0
Cells EBV transformed lymphocytes	1	10	0.5	0.5	0.5	1	0
Colon Sigmoid	1	8.5	0.5	0.5	0.5	1	0.01
Colon Sigmoid	1	9	0.5	0.5	0.5	1	0.01
Colon Sigmoid	1	9.5	0.5	0.5	0.5	1	0.01
Colon Sigmoid	1	10	0.5	0.5	0.5	1	0

Colon Transverse	1	8	0.5	0.5	0.5	1	0.01
Colon Transverse	1	8.5	0.5	0.5	0.5	1	0.01
Colon Transverse	1	9	0.5	0.5	0.5	1	0.01
Colon Transverse	1	9.5	0.5	0.5	0.5	1	0.01
Colon Transverse	1	10	0.5	0.5	0.5	1	0
Esophagus Gastroesophageal Junction	1	8	0.5	0.5	0.5	1	0.01
Esophagus Gastroesophageal Junction	1	8.5	0.5	0.5	0.5	1	0.01
Esophagus Gastroesophageal Junction	1	9	0.5	0.5	0.5	1	0.01
Esophagus Gastroesophageal Junction	1	9.5	0.5	0.5	0.5	1	0
Esophagus Gastroesophageal Junction	1	10	0.5	0.5	0.5	1	0
Esophagus Mucosa	1	8	0.5	0.5	0.5	1	0.01
Esophagus Mucosa	1	8.5	0.5	0.5	0.5	1	0.01
Esophagus Mucosa	1	9	0.5	0.5	0.5	1	0.01
Esophagus Mucosa	1	9.5	0.5	0.5	0.5	1	0
Esophagus Mucosa	1	10	0.5	0.5	0.5	1	0
Esophagus Muscularis	1	8.5	0.5	0.5	0.5	1	0.01
Esophagus Muscularis	1	9	0.5	0.5	0.5	1	0.01
Esophagus Muscularis	1	9.5	0.5	0.5	0.5	1	0
Esophagus Muscularis	1	10	0.5	0.5	0.5	1	0
Heart Atrial Appendage	1	8	0.5	0.5	0.5	1	0.01
Heart Atrial Appendage	1	8.5	0.5	0.5	0.5	1	0.01

Heart Atrial Appendage	1	9	0.5	0.5	0.5	1	0.01
Heart Atrial Appendage	1	9.5	0.5	0.5	0.5	1	0
Heart Atrial Appendage	1	10	0.5	0.5	0.5	1	0
Heart Left Ventricle	1	8	0.5	0.5	0.5	1	0.01
Heart Left Ventricle	1	8.5	0.5	0.5	0.5	1	0.01
Heart Left Ventricle	1	9	0.5	0.5	0.5	1	0.01
Heart Left Ventricle	1	9.5	0.5	0.5	0.5	1	0.01
Heart Left Ventricle	1	10	0.5	0.5	0.5	1	0
Kidney Cortex	1	7.5	0.5	0.5	0.5	1	0.01
Kidney Cortex	1	8	0.5	0.5	0.5	1	0.01
Kidney Cortex	1	8.5	0.5	0.5	0.5	1	0.01
Kidney Cortex	1	9	0.5	0.5	0.5	1	0
Kidney Cortex	1	9.5	0.5	0.5	0.5	1	0
Kidney Cortex	1	10	0.5	0.5	0.5	1	0
Kidney Cortex	2	10	0.5	0.5	0.5	1	0.01
Kidney Cortex	5	9.5	0.5	0.5	0.51	1	0.01
Kidney Cortex	5	10	0.5	0.5	0.5	1	0
Kidney Cortex	6	9	0.5	0.5	0.5	1	0
Kidney Cortex	6	9.5	0.5	0.5	0.5	1	0
Kidney Cortex	6	10	0.5	0.5	0.5	1	0
Liver	1	7.5	0.5	0.5	0.5	1	0.01
Liver	1	8	0.5	0.5	0.5	1	0.01
Liver	1	8.5	0.5	0.5	0.5	1	0.01
Liver	1	9	0.5	0.5	0.5	1	0
Liver	1	9.5	0.5	0.5	0.5	1	0
Liver	1	10	0.5	0.5	0.5	1	0
Lung	1	8	0.5	0.5	0.5	1	0.01
Lung	1	8.5	0.5	0.5	0.5	1	0.01
Lung	1	9	0.5	0.5	0.5	1	0.01
Lung	1	9.5	0.5	0.5	0.5	1	0.01
Lung	1	10	0.5	0.5	0.5	1	0
Minor Salivary Gland	1	8.5	0.5	0.5	0.5	1	0.01
Minor Salivary Gland	1	9	0.5	0.5	0.5	1	0.01
Minor Salivary Gland	1	9.5	0.5	0.5	0.5	1	0.01
Minor Salivary Gland	1	10	0.5	0.5	0.5	1	0
Muscle Skeletal	1	8	0.5	0.5	0.5	1	0.01

Muscle Skeletal	1	8.5	0.5	0.5	0.5	1	0.01
Muscle Skeletal	1	9	0.5	0.5	0.5	1	0.01
Muscle Skeletal	1	9.5	0.5	0.5	0.5	1	0
Muscle Skeletal	1	10	0.5	0.5	0.5	1	0
Nerve Tibial	1	8	0.5	0.5	0.5	1	0.01
Nerve Tibial	1	8.5	0.5	0.5	0.5	1	0.01
Nerve Tibial	1	9	0.5	0.5	0.5	1	0.01
Nerve Tibial	1	9.5	0.5	0.5	0.5	1	0
Nerve Tibial	1	10	0.5	0.5	0.5	1	0
Nerve Tibial	2	10	0.5	0.5	0.5	1	0.01
Ovary	1	8	0.5	0.5	0.5	1	0.01
Ovary	1	8.5	0.5	0.5	0.5	1	0.01
Ovary	1	9	0.5	0.5	0.5	1	0.01
Ovary	1	9.5	0.5	0.5	0.5	1	0
Ovary	1	10	0.5	0.5	0.5	1	0
Pancreas	1	8	0.5	0.5	0.5	1	0.01
Pancreas	1	8.5	0.5	0.5	0.5	1	0.01
Pancreas	1	9	0.5	0.5	0.5	1	0.01
Pancreas	1	9.5	0.5	0.5	0.5	1	0
Pancreas	1	10	0.5	0.5	0.5	1	0
Pituitary	1	8	0.5	0.5	0.5	1	0.01
Pituitary	1	8.5	0.5	0.5	0.5	1	0.01
Pituitary	1	9	0.5	0.5	0.5	1	0.01
Pituitary	1	9.5	0.5	0.5	0.5	1	0
Pituitary	1	10	0.5	0.5	0.5	1	0
Pituitary	2	9.5	0.5	0.5	0.5	1	0.01
Pituitary	2	10	0.5	0.5	0.5	1	0.01
Prostate	1	8	0.5	0.5	0.5	1	0.01
Prostate	1	8.5	0.5	0.5	0.5	1	0.01
Prostate	1	9	0.5	0.5	0.5	1	0.01
Prostate	1	9.5	0.5	0.5	0.5	1	0
Prostate	1	10	0.5	0.5	0.5	1	0
Skin Not Sun Exposed Suprapubic	1	8.5	0.5	0.5	0.5	1	0.01
Skin Not Sun Exposed Suprapubic	1	9	0.5	0.5	0.5	1	0.01
Skin Not Sun Exposed Suprapubic	1	9.5	0.5	0.5	0.5	1	0
Skin Not Sun Exposed Suprapubic	1	10	0.5	0.5	0.5	1	0
Skin Sun Exposed Lower leg	1	8	0.5	0.5	0.5	1	0.01

Skin Sun Exposed Lower leg	1	8.5	0.5	0.5	0.5	1	0.01
Skin Sun Exposed Lower leg	1	9	0.5	0.5	0.5	1	0.01
Skin Sun Exposed Lower leg	1	9.5	0.5	0.5	0.5	1	0
Skin Sun Exposed Lower leg	1	10	0.5	0.5	0.5	1	0
Skin Sun Exposed Lower leg	2	10	0.5	0.5	0.5	1	0.01
Small Intestine Terminal Ileum	1	8.5	0.5	0.5	0.5	1	0.01
Small Intestine Terminal Ileum	1	9	0.5	0.5	0.5	1	0.01
Small Intestine Terminal Ileum	1	9.5	0.5	0.5	0.5	1	0.01
Small Intestine Terminal Ileum	1	10	0.5	0.5	0.5	1	0
Spleen	1	7.5	0.5	0.5	0.5	1	0.01
Spleen	1	8	0.5	0.5	0.5	1	0.01
Spleen	1	8.5	0.5	0.5	0.5	1	0.01
Spleen	1	9	0.5	0.5	0.5	1	0
Spleen	1	9.5	0.5	0.5	0.5	1	0
Spleen	1	10	0.5	0.5	0.5	1	0
Spleen	2	10	0.5	0.5	0.5	1	0.01
Stomach	1	8.5	0.5	0.5	0.5	1	0.01
Stomach	1	9	0.5	0.5	0.5	1	0.01
Stomach	1	9.5	0.5	0.5	0.5	1	0.01
Stomach	1	10	0.5	0.5	0.5	1	0.01
Testis	1	7.5	0.5	0.5	0.5	1	0.01
Testis	1	8	0.5	0.5	0.5	1	0.01
Testis	1	8.5	0.5	0.5	0.5	1	0.01
Testis	1	9	0.5	0.5	0.5	1	0
Testis	1	9.5	0.5	0.5	0.5	1	0
Testis	1	10	0.5	0.5	0.5	1	0
Testis	2	9.5	0.5	0.5	0.5	1	0.01
Testis	2	10	0.5	0.5	0.5	1	0.01
Thyroid	1	8	0.5	0.5	0.5	1	0.01
Thyroid	1	8.5	0.5	0.5	0.5	1	0.01
Thyroid	1	9	0.5	0.5	0.5	1	0.01
Thyroid	1	9.5	0.5	0.5	0.5	1	0
Thyroid	1	10	0.5	0.5	0.5	1	0
Thyroid	2	10	0.5	0.5	0.5	1	0.01

Uterus	1	7.5	0.5	0.5	0.5	1	0.01
Uterus	1	8	0.5	0.5	0.5	1	0.01
Uterus	1	8.5	0.5	0.5	0.5	1	0.01
Uterus	1	9	0.5	0.5	0.5	1	0.01
Uterus	1	9.5	0.5	0.5	0.5	1	0
Uterus	1	10	0.5	0.5	0.5	1	0
Vagina	1	7.5	0.5	0.5	0.5	1	0.01
Vagina	1	8	0.5	0.5	0.5	1	0.01
Vagina	1	8.5	0.5	0.5	0.5	1	0.01
Vagina	1	9	0.5	0.5	0.5	1	0.01
Vagina	1	9.5	0.5	0.5	0.5	1	0
Vagina	1	10	0.5	0.5	0.5	1	0
Whole Blood	1	8	0.5	0.5	0.5	1	0.01
Whole Blood	1	8.5	0.5	0.5	0.5	1	0.01
Whole Blood	1	9	0.5	0.5	0.5	1	0.01
Whole Blood	1	9.5	0.5	0.5	0.5	1	0
Whole Blood	1	10	0.5	0.5	0.5	1	0

Table 5.3: **Top 5% worst AUC predictions.** CI= Confidence Interval, Sens=Sensitivity, Spec=Specificity

Gene	**Gene Symbol**	**Total ASE Samples**	**Expressed Allele**
ENSG00000109320	NFKB1	3	
ENSG00000181649	PHLDA2	10	Maternal
ENSG00000078900	TP73	16	Maternal
ENSG00000184937	WT1	17	Paternal
ENSG00000105825	TFPI2	24	Maternal
ENSG00000139687	RB1	33	Maternal
ENSG00000166619	BLCAP	36	Isoform Dependent
ENSG00000140009	ESR2	57	
ENSG00000162595	DIRAS3	320	Paternal
ENSG00000185559	DLK1	1111	Paternal
ENSG00000182636	NDN	3123	Paternal
ENSG00000198300	PEG3	6050	Paternal
ENSG00000118495	PLAGL1	7517	Paternal
ENSG00000214548	MEG3	13022	Maternal

Table 5.4: **Number of samples showing allele specific expression in imprinted tumor suppressor genes in the phASER results.**

Gene	**Location**	**Expressed Allele**
ADTRP	6p24.1 AS	Maternal
AIM1	6q21	Paternal
ANO1	11q13.3	Maternal
ATP10A	15q11.2 AS	Maternal
ATP5F1EP2	13q12.2	Maternal
BLCAP	20q11.2-q12 AS	Isoform Dependent

CCDC71L	7q22.3 AS	Paternal
CDKN1C	11p15.5 AS	Maternal
CMTM1	16q21	Paternal
COPG2IT1	7q32	Paternal
CPA4	7q32	Maternal
DDC	7p12.2 AS	Isoform Dependent
DGCR6	22q11.21	Random
DGCR6L	22q11.21 AS	Random
DIO3	14q32	Paternal
DIO3OS	14q32.31 AS	Maternal
DIRAS3	1p31 AS	Paternal
DLGAP2	8p23	Paternal
DLK1	14q32.2	Paternal
DLX5	7q22 AS	Maternal
DNMT1	19p13.2 AS	Paternal
DSCAM	21q22.2 AS	Paternal
ERAP2	5q15	Paternal
ESR2	14q23.2-q23.3 AS	
FAM50B	6p25.2	Paternal
GDAP1L1	20q12	Paternal
GLI3	7p13 AS	Paternal
GLIS3	9p24.2 AS	Paternal
GNAS	20q13.3	Isoform Dependent
GNASAS	20q13.32 AS	Paternal
GPR1	2q33.3 AS	Paternal
GRB10	7p12-p11.2 AS	Isoform Dependent
H19	11p15.5 AS	Maternal
HECW1	7p14.1-p13	Paternal
HNF1A	12q24.31	
HOXA4	7p15-p14 AS	Maternal
HYMAI	6q24.2 AS	Paternal
IGF2	11p15.5 AS	Paternal
IGF2AS	11p15.5	Paternal
INPP5F V2	10q26.11	Paternal
INS	11p15.5 AS	Paternal
IRAIN	15q26.3 AS	Paternal
KCNK9	8q24.3 AS	Maternal
KCNQ1	11p15.5	Maternal
KCNQ1DN	11p15.4	Maternal
KCNQ1OT1	11p15	Paternal
KLF14	7q32.3 AS	Maternal
L3MBTL1	20q13.12	Paternal
LIN28B	6q21	Paternal
LRRTM1	2p12 AS	Paternal
MAGEL2	15q11-q12 AS	Paternal
MAGI2	7q21 AS	Maternal
MCTS2	20q11.21	Paternal
MEG3	14q32	Maternal

MEG8	14q32.2-q32.31	Maternal
MEST	7q32	Paternal
MESTIT1	7q32.2 AS	Paternal
MIMT1	19q13.4	Paternal
MIR296	20q13.32 AS	Paternal
MIR298	20q13.32 AS	Paternal
MIR371A	19q13.42	Paternal
MKRN3	15q11-q13	Paternal
NAA60	16p13.3	Maternal
NAP1L5	4q22.1 AS	Paternal
NDN	15q11.2-q12 AS	Paternal
NFKB1	4q24	
NLRP2	19q13.42	Maternal
NNAT	20q11.2-q12	Paternal
NPAP1	15q11.2	Paternal
NTM	11q25	Maternal
OSBPL5	11p15.4 AS	Maternal
PARD6G	18q23 AS	Maternal
PEG10	7q21	Paternal
PEG13	8q24.22	Paternal
PEG3-AS1	19q13.43	Paternal
PEG3	19q13.4 AS	Paternal
PHLDA2	11p15.5 AS	Maternal
PLAGL1	6q24-q25 AS	Paternal
PPP1R9A	7q21.3	Maternal
PRR25	16p13.3	Paternal
PSIMCT-1	20q11.2	Paternal
PWAR6	15q11.2	Paternal
PWCR1	15q11.2	Paternal
PXDC1	6p25.2 AS	Paternal
RAC1	7p22.1	
RASGRF1	15q24.2 AS	Paternal
RB1	13q14.2	Maternal
RBP5	12p13.31 AS	Maternal
RHOBTB3	5q15	Paternal
RNU5D-1	1p34.1 AS	Paternal
RTL1	14q32.31 AS	Paternal
SANG	20q13.32	Paternal
SGCE	7q21-q22 AS	Paternal
SGK2	20q13.2	Paternal
SLC22A18	11p15.5	Maternal
SLC22A2*	6q26 AS	Maternal
SLC22A3*	6q26-q27	Maternal
SMOC1	14q24.2	Maternal
SNORD107	15q11.2	Paternal
SNORD108	15q11.2	Paternal
SNORD109A	15q11.2	Paternal
SNORD109B	15q11.2	Paternal

SNORD113-1	14q32.31	Maternal
SNORD114-1	14q32.31	Maternal
SNORD115-48	15q11.2	Paternal
SNORD115@	15q11.2	Paternal
SNORD116	15q11.2	Paternal
SNORD64	15q12	Paternal
SNRPN	15q11.2	Paternal
SNURF	15q12	Paternal
ST8SIA1	12p12.1 AS	Paternal
SVOPL	7q34 AS	Maternal
TCEB3C	18q21.1 AS	Maternal
TFPI2	7q22 AS	Maternal
TP53	17p13.1 AS	
TP73	1p36.3	Maternal
UBE3A	15q11-q13 AS	Maternal
VTRNA2-1	5q31.1 AS	Paternal
WT1-AS	11p13	Paternal
WT1	11p13 AS	Paternal
ZC3H12C	11q22.3	Paternal
ZDBF2	2q33.3	Paternal
ZFAT-AS1	8q24.22	Paternal
ZFAT	8q24.22 AS	Paternal
ZFP90	16q22.1	Paternal
ZIM2	19q13.4 AS	Paternal
ZNF396	18q12.2 AS	Paternal

Table 5.5: **List of imprinted genes obtained from geneimprint.**

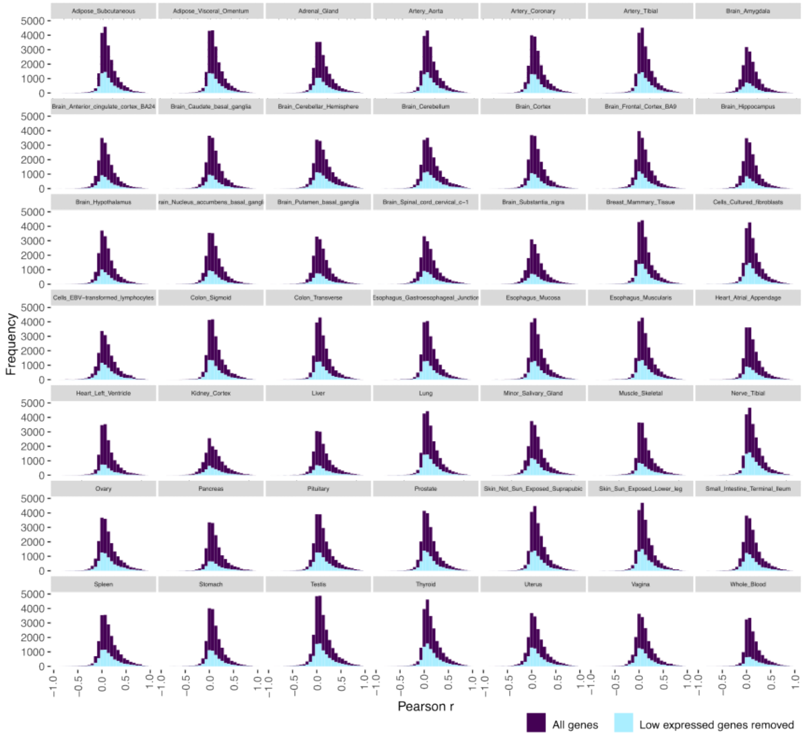


Figure 5.8: **Comparison of predicted allele specific expression using Predixcan versus allelic fold change from phASER results for all genes within a sample with all genes and with highly expressed genes only.** Histogram of Pearson r values for comparing PrediXcan z-score ratios to aFC for all samples within a tissue type for a particular gene. Plots are split by tissue type. Purple= distribution of all genes. Light blue= distribution of genes with lowly expressed (TPM <10) removed.

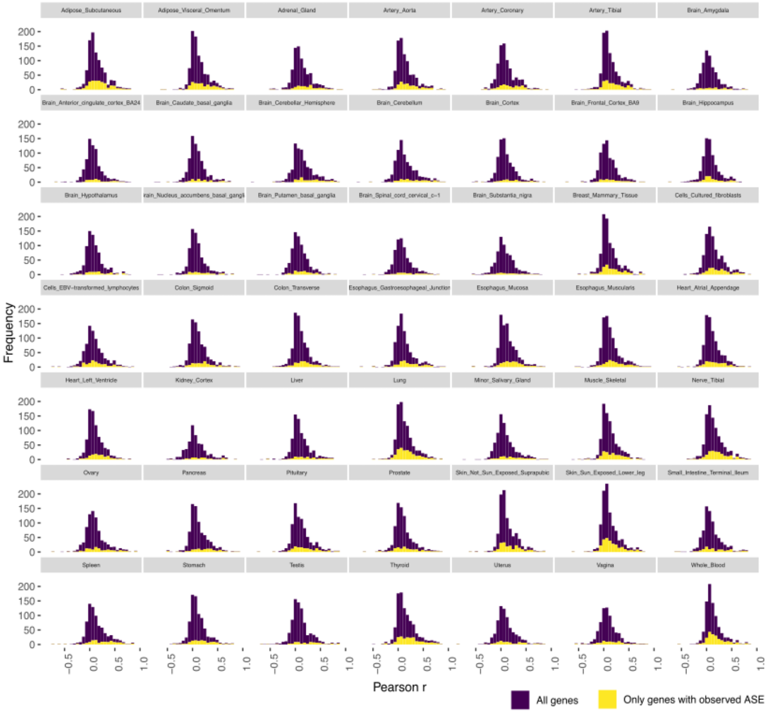


Figure 5.9: **Comparison of predicted allele specific expression using Predixcan versus allelic fold change from phASER results for all genes within a sample with all genes and with genes limited to those that have observed allele specific expression in the phASER results.** Histogram of Pearson r values for comparing PrediXcan z-score ratios to aFC for all samples within a tissue type for a particular gene. Plots are split by tissue type. Purple= distribution of all genes. Yellow= distribution of genes that have observed ASE in the phASER results.

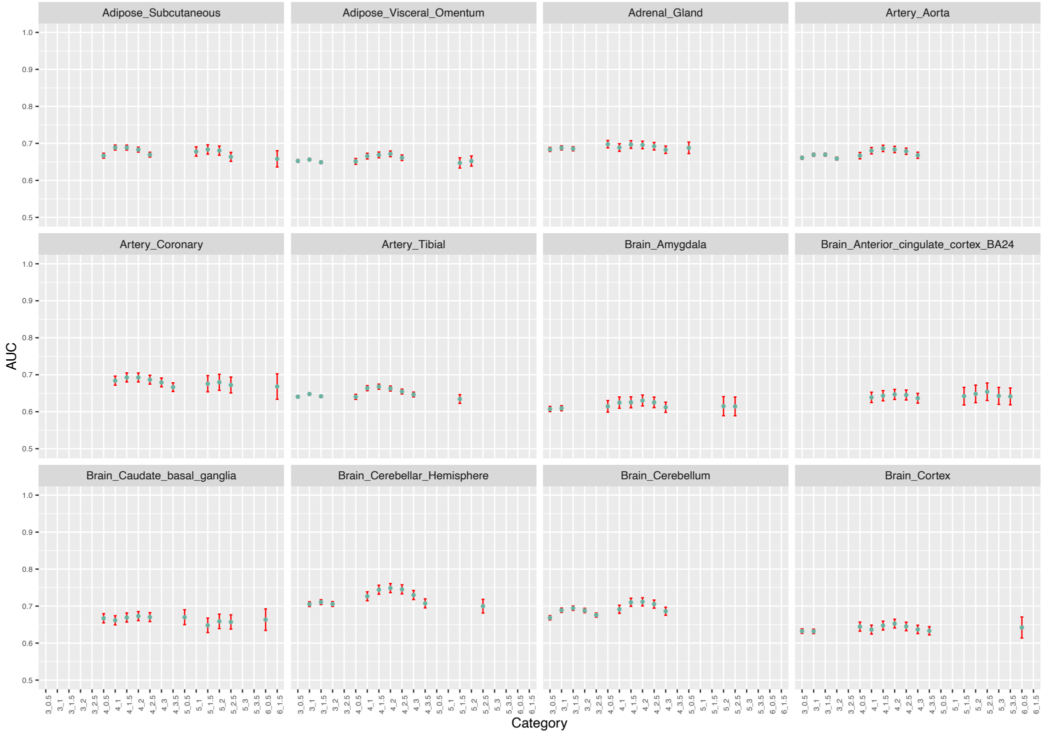


Figure 5.10: **The top 10 AUC scores for tissue types 1 to 12 alphabetically.** X axis Category= phASER aFC threshold PrediXcan z-score ratio. The results only show the combination of thresholds applied for the top 10 AUC results and therefore are blank for the combinations that are in the top 10 for a different tissue type. Red lines= delong confidence intervals. Y axis = AUC score.

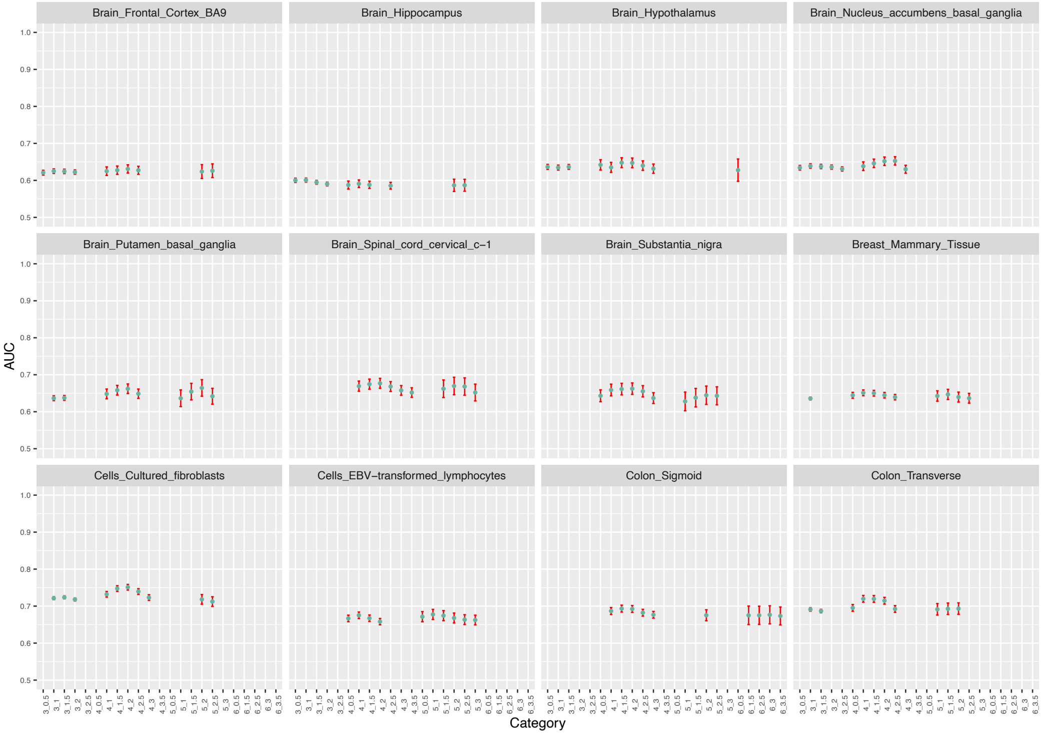


Figure 5.11: **The top 10 AUC scores for tissue types 13 to 24 alphabetically.** X axis Category= phASER aFC threshold PrediXcan z- score ratio. The results only show the combination of thresholds applied for the top 10 AUC results and therefore are blank for the combinations that are in the top 10 for a different tissue type. Red lines= delong confidence intervals. Y axis = AUC score.

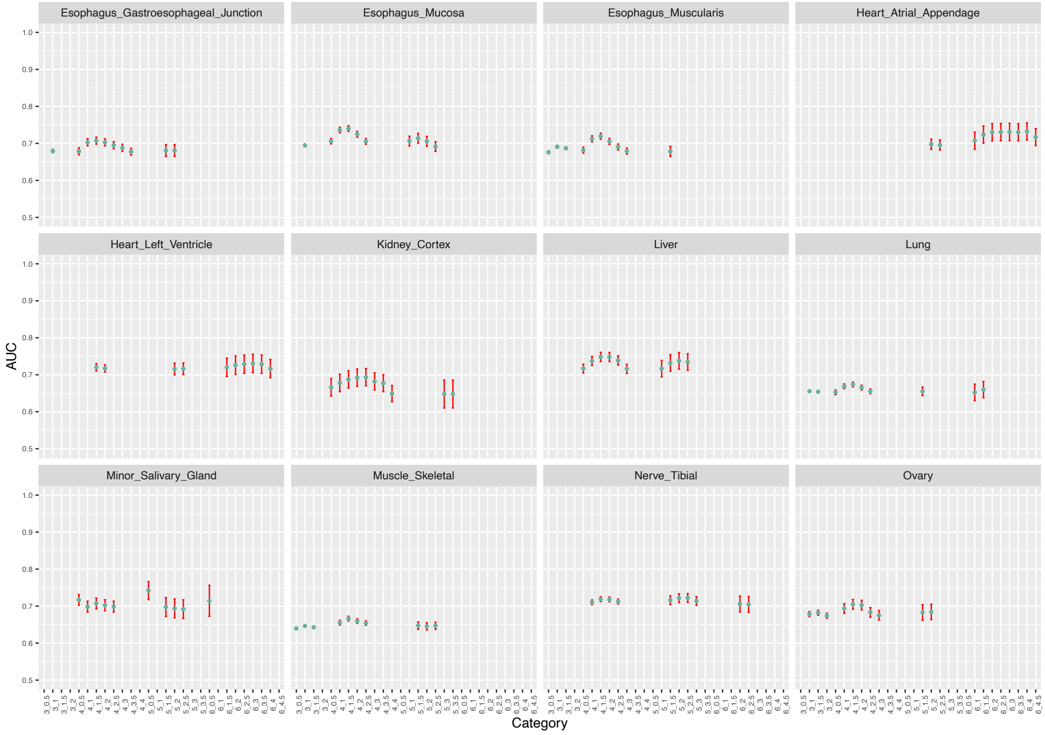


Figure 5.12: **The top 10 AUC scores for tissue types 25 to 36 alphabetically.** X axis Category= phASER aFC threshold PrediXcan z-score ratio. The results only show the combination of thresholds applied for the top 10 AUC results and therefore are blank for the combinations that are in the top 10 for a different tissue type. Red lines= delong confidence intervals. Y axis = AUC score.

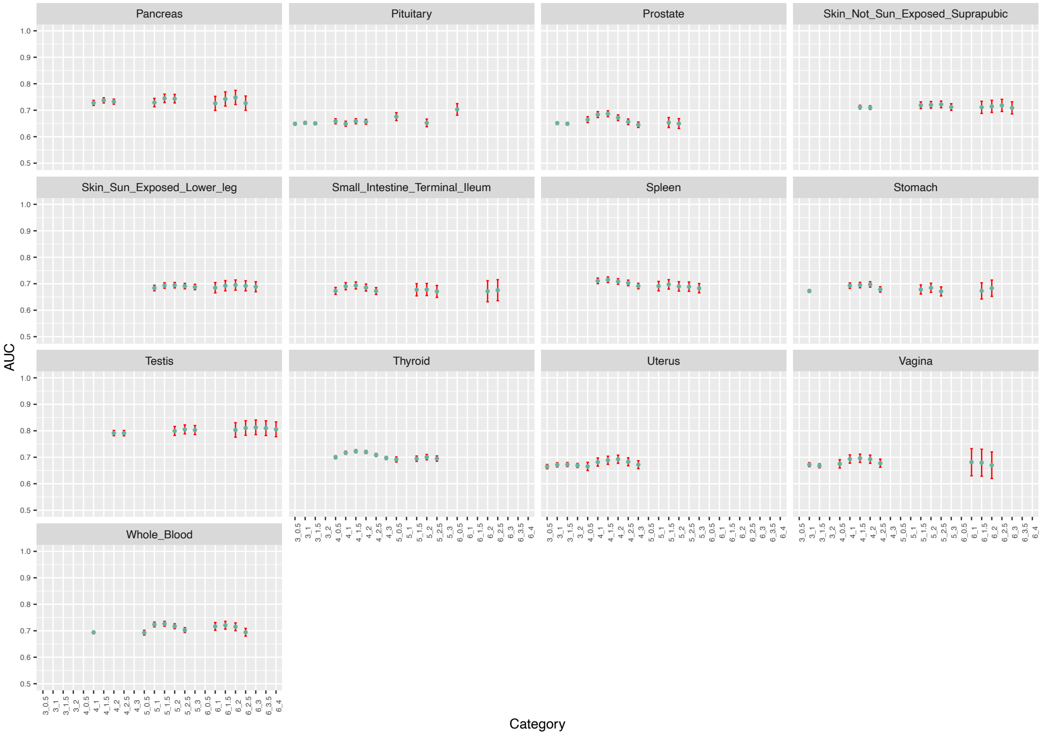


Figure 5.13: **The top 10 AUC scores for tissue types 37 to 49 alphabetically.** X axis Category= phASER aFC threshold PrediXcan z score ratio. The results only show the combination of thresholds applied for the top 10 AUC results and therefore are blank for the combinations that are in the top 10 for a different tissue type. Red lines= delong confidence intervals. Y axis = AUC score.