Predictors and Consequences of Genomic Instability in Cancer

Jacob R. Bradley

Doctor of Philosophy
University of Edinburgh
2023

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*Jacob R. Bradley*)

To Morton...

Abstract

In this work, we employ a variety of methods, some novel, from high-dimensional statistics and machine learning to high-throughput cancer genomics data. We restrict our enquiry in the following two ways. Firstly, we are concerned with understanding genomic instability and the hypermutated phenotype, particularly in the context of its relevance to immunotherapy. Secondly, while we use a variety of 'omics data types to inform our understanding, our specific aim will be to make clinical predictions on the basis only of whole-genome, whole-exome, or targeted panel sequencing of the tumour genome. This can be motivated via both via scientific interest and clinical practicality. While somatic mutations have been understood for a long time as the instigators and drivers of tumourigenesis, it is now well known that the tumour environment and external factors play key roles in cancer development and spread. It is therefore of key importance to appreciate the nature of this balance. We formulate our interrogation of this broad question as a signal-to-noise problem, and attempt to ascertain to what extent the downstream properties of tumours can be predicted purely on the basis of the mutations they carry. This is in turn directly applicable to a growing area of clinical practice, liquid biopsy. Unlike solid biopsy, the nature of liquid biopsy means that we are only able to sequence tumour DNA, requiring any decisions based on liquid biopsy to be based purely on somatic mutations, potentially for some subset of the genome. We present a blueprint for the design, development and estimation of clinical biomarkers of response to immunotherapy based on generative models of the genome-wide landscape of molecular processes in tumour cells. From such models we can also make inferences about the underlying mechanisms leading to genomic instability, and how the hypermutated phenotype interacts with the body's natural defences against cancer.

Lay Summary

Cancer is a disease that has profound impacts on our society every day. Despite this, we feel like we know very little about cancer relative to how much there is to know. Why is this?

It's simple: cancer is not one disease. Two patients' tumours may be caused by different things, occur in different places, and have very different effects. More so than any other disease, every cancer is unique. To understand this variety, we have to look at the one thing all cancers have in common: mutations. Mutations are where DNA, the information-storing molecule in cells, has been changed by accident somewhere along its sequence. In tumours this causes cells to become detached from the normal rules that govern when cells reproduce and die. Some tumours have lots of mutations throughout their genome (complete DNA sequence), while some have very few. No two tumours share the same pattern of mutations; the human genome is so long that the chances of damage occurring in exactly the same places are minute.

In the situation we've described there are far more possible outcomes (patterns of mutation) than we will ever have samples to work with. This makes doing classical statistics, which often works on the assumption that we have a large sample size compared to the amount of data contained in each sample, very hard. Research addressing this difficulty is called high-dimensional statistics, and is the mathematical side of what I do.

In statistics there are broadly two types of questions we might like to answer. Firstly, given some data, how is that data structured and what correlations exist within it? Secondly, how does our data relate to another, separate, piece of information? We can illustrate these two types of question quite nicely in the context of cancer genomics.

As we know, tumour cells accumulate mutations. Some are important for the development of the tumour, and some are just along for the ride, caused by the same processes as the important mutations but of little effect. The difficulty is distinguishing the important from the unimportant, when the same set of mutations rarely occurs in two different tumours. To do this we need to build statistical models describing the process by which mutations accumulate, and build into these models structure that reflects our biological knowledge. Relevant knowledge might include how DNA is organised into genes, chromosomes and coding units. We hope to recover information about what locations in the genome may be important for the success or failure of a tumour, or for causing other mutations. This helps biologists refine their experiments to understand exactly what is happening – no matter how clever our method, they get the final say.

Next we need to understand how mutations interact with the busy world of a tumour. There is a decades-old debate in biology about how DNA's role is best interpreted. Some people think of it as an ingredients list for consulting whenever a specific item is needed, others as more like the recipe itself, a set of instructions for running a cell. In cancer, we might ask to what extent the properties of a tumour can be predicted just from its mutations. There is a practical motivation for this, namely that sometimes that's all the information we have. An emerging technology for sampling tumour DNA is liquid biopsy, where DNA is extracted from blood samples. This is in contrast to solid biopsy, where a tumour is surgical removed. Solid biopsy gives us access to more information, but is very invasive and sometimes impossible. In my (biological) work I try to understand the relationship between mutations and the other processes in the tumour environment. In particular I care about two types of molecules, RNAs and neoantigens, which can be directly measured by solid (but not liquid) biopsy. These molecules are important in determining how well a tumour will respond to a specific set of drugs called immunotherapies. If we can predict how they behave while only being able to see the mutations in a tumour, then we only need to use liquid biopsy when assessing patients for

immunotherapy.

Acronyms

AUPRC Area Under Precision-Recall Curve. 27, 28, 30

BMR Background Mutation Rate. 19, 21, 32

ctDNA Circulating Tumour DNA. 15

CTLA-4 Cytotoxic T Lymphocyte Associated protein 4. 15

ecTMB Estimation and Classification of Tumour Mutation Burden. 27–30

ICB Immune Checkpoint Blockade. 15, 17

LASSO Least Absolute Shrinkage and Selection Operator. 24

MCMC Monte-Carlo Markov Chain. 18

MSI Micro-Satellite Instability. 15

NSCLC Non-Small Cell Lung Cancer. 19, 20, 32

PD-L1 Programmed Death Ligand 1. 15

PV Polycythemia Vera. 13

TCGA The Cancer Genome Atlas. 14

TIB Tumour Indel Burden. 15, 19–22, 25, 29, 30, 32

TMB Tumour Mutation Burden. 15, 19, 21, 25, 27–32

WES Whole Exome Sequencing. 15, 20

Contents

Chapter 1

Introduction

1 Cancer as a disease of the genome

Cancer does not require much introduction even to the lay reader; direct or indirect experience of cancer is universal. It is consistently ranked amongst the leading causes of global mortality, and multiple subtypes of cancer are projected to increase in their ranking and share of worldwide premature deaths over the coming decades (Mathers and Loncar, 2006), including in the developing world (Kanavos, 2006). Beyond its direct death toll, cancer is responsible for the expenditure of trillions of dollars per year in cost of care and lost economic output (Wild et al., 2020). While huge gains in the understanding, prevention and treatment of cancer have been made in recent years, many challenges remains in scalably advancing each of these three categories. Modern cancer treatments in particular are often extremely expensive, with drug development costs increasing and consequently inflating the price of access to therapeutics (Howard et al., 2015). In short, there is much to be hopeful about in oncology, but it is by no means guaranteed that the current revolutions being enjoyed in scientific understanding will translate fully to equitable clinical benefit.

In order to understand the state of play in cancer research, we need to know a little about the nature of cancer itself, and a little about the nature of modern molecular biology. A key starting point is that cancer is not a unitary disease; two patients' tumours may be caused by different processes, occur in different tissues, and have very different molecular and physiological effects (Wittekind et al., 2016). More so than any other disease, every cancer is unique. It is therefore natural to ask what unifying features of all cancers justify their joint classification. The modern answer is distinct from the historical answer. Before the advent of genetics, cancers of disparate tissues of origin were grouped together under the unifying observation of malignant growths crossing over physiological boundaries. Towards the end of the 19th century, it was recognised that aberrant patterns of cell reproduction were a common feature of cancers (Weinstein and Case, 2008). By the early 1900s, with the writings of scientists such as Theodor Boveri (see Boveri, 2008, for a modern translation), an answer would be formulated foreshadowing our current understanding, although it wouldn't be until far later that this explanation was fully accepted. Boveri proposed that 'chromosomal abnormalities' gave rise to the conversion of normal cells to malignant neoplasms. In modern nomenclature, the chromosomal abnormalities to which he referred would be regarded as (a specific kind of) mutations. It is these that lay the groundwork for uncontrolled cellular reproduction and all the other associated hallmarks of cancer[1], such as avoiding detection from the body's defenses (immunosuppression/evasion), recruiting a local blood supply (angiogenesis), and invasion of separate tissues (metastasis) (Hanahan and Weinberg, 2011). Crucially, mutations convert previously normal or benign cell populations into tumours. The mutations that were observable via optical microscopy to scientists in the first half of the twentieth century were structural mutations involving large-scale chromosomal translocations or deletions. It wasn't until the identification of the structure of

[1]While some rare cancers such as Polycythemia Vera (PV) may involve uncontrolled production of cells that do not themselves harbour mutations (in this case, mature red blood cells do not contain DNA at all), this is still the downstream effect of mutations in other cell types. For example, in PV this is most commonly a mutation of the *JAK2* gene in hematopoietic stem cells (Tefferi, 2007).

DNA and its role as the primary mechanism of inheritance by Franklin, Watson and Crick (1953) and the subsequent development of molecular genetics that the discrete nature of biological information was fully appreciated. Later developments in DNA sequencing, beginning with the work of Sanger et al. (1977) allowed a fuller understanding of mutations as changes to the sequence of nucleotide bases that constitutes DNA. The science of cancer continued to progress by associating DNA mutations (errors of cellular information storage), with their mechanistic and functional consequences, in particular those that led to deregulation of normal cell-cycle control.

Now that we are armed with a general characterisation of cancer as the consequences of mutations in DNA leading to abnormal reproduction of cells, we can begin to appreciate the reasons for cancer's diversity. Since almost all cells in the body contain DNA and experience regular reproduction, cancer may occur in a wide range of tissues throughout the body[2]. Furthermore, the size of the human genome (defined as the combined total of genetic information contained in DNA, comprising of around 20,000 genes and 3 billion nucleotide base pairs) means that, even with cancer being as common a disease as it is, simple statistical reasoning allows us to say with confidence that it is almost inconceivable that two given tumours would carry exactly the same constituent mutations (even without considering complicating factors such as tumour heterogeneity). This leads us to the modern era of molecular biology. Since the completion of the human genome project (Lander et al., 2001), high-throughput sequencing, where large portions of the genome in their entirety are sequenced for a biological sample, has become ubiquitous and highly automated. We now have easy access to the precise locations of all mutations in the tumour genomes of many tens thousands of thousands of samples gathered across hundreds of studies via repositories such as The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013). This gives us an opportunity to investigate a variety of fundamental questions with regards to the progression of cancer. One of these mirrors a classic debate of nature versus nurture in developmental biology. In this case, we wish to understand the extent to which the dynamics and trajectory of a tumour are pre-determined by the genetic damage it carries. We know that cancers are defined by their mutations, but a growing field of investigation is exploring the role of the environment in which a tumour finds itself in allowing it to flourish. For the remainder of this section, we will elaborate on the balance between these two views of the tumour genome.

1.1 Nature: mutations in the driver's seat

- Discussion of the ways in which mutations drive cancer, in particular oncogenes and tumour suppressors. - Note that mutations can come from internal or external factors.

1.2 Nurture: A warm (micro-) environment

- Pull back on the importance of mutations, and look at the micro-environment, in particular hot and cold microenvironments. (Keenan et al., 2019) (Boulter et al., 2020)

[2]Note that tissues/cell types in which cancer is extremely uncommon tend to be those which experience very little reproduction, and so have little chance to accumulate mutations, e.g. neuronal cells and tissues making up the heart

2 Genomic instability

2.1 The molecular mechanics of DNA damage

2.2 The hypermutated phenotype

2.3 Mismatch repair

3 Immunotherapy, checkpoint blockade and beyond

Since the discovery of Immune Checkpoint Blockade (ICB)[3] (Ishida et al., 1992; Leach et al., 1996), there has been an explosion of interest in cancer therapies targeting immune response and ICB therapy is now widely used in clinical practice (Robert, 2020). ICB therapy works by targeting natural mechanisms (or *checkpoints*) that disengage the immune system, for example the proteins Cytotoxic T Lymphocyte Associated protein 4 (CTLA-4) and Programmed Death Ligand 1 (PD-L1) (Buchbinder and Desai, 2016). Inhibition of these checkpoints can promote a more aggressive anti-tumour immune response (Pardoll, 2012), and in some patients this leads to long-term remission (Gettinger et al., 2019). However, ICB therapy is not always effective (Nowicki et al., 2018) and may have adverse side-effects, so determining which patients will benefit in advance of treatment is vital.

4 Clinical practice and pharmacoeconomics

4.1 Biomarkers as proxies of immunogenicity

Discuss TMB and MSI.

Exome-wide prognostic biomarkers for immunotherapy are now well-established – in particular, Tumour Mutation Burden (TMB) is used to predict response to immunotherapy (Zhu et al., 2019; Cao et al., 2019). TMB is defined as the total number of non-synonymous mutations occurring throughout the tumour exome, and can be thought of as a proxy for how easily a tumour cell can be recognised as foreign by immune cells (Chan et al., 2019). However, the cost of measuring TMB using Whole Exome Sequencing (WES) (Sboner et al., 2011) currently prohibits its widespread use as standard-of-care. Sequencing costs, both financial and in terms of the time taken for results to be returned, are especially problematic in situations where high-depth sequencing is required, such as when utilising blood-based Circulating Tumour DNA (ctDNA) from liquid biopsy samples (Gandara et al., 2018). The same issues are encountered when measuring more recently proposed biomarkers such as Tumour Indel Burden (TIB) (Wu et al., 2019b; Turajlic et al., 2017), which counts the number of frameshift insertion and deletion mutations. There is, therefore, demand for cost-effective approaches to estimate these biomarkers (Fancello et al., 2019; Golkaram et al., 2020).

4.2 Liquid biopsy

(Jensen et al., 2020) (Genovese et al., 2014) (Razavi et al., 2019) (Schweizer et al., 2019) (Annala et al., 2018) (Goodall et al., 2017)

4.3 Targeted gene panels

5 Roadmap

[3]For their work on ICB, James Allison and Tasuku Honjo received the 2018 Nobel Prize for Physiology/Medicine (Ledford et al., 2018).

Chapter 2

The Generative/Predictive Blueprint for Biomarker Estimation

1 Introduction

Now armed with an understanding of the biological context of **genomic instability** in cancer, we will describe the statistical workflow underlying the following chapters. We will be concerned with a) designing models to encapsulate the signatures of genomic instability; b) understanding how these patterns of mutation interact with tumours' development in the context of ICB therapy; and c) developing practicable tests to stratify patients according to likelihood of response. This last goal, that the methods we produce must be implementable, will provide a further set of restrictions refining the scope of our efforts. In particular, we will focus on methods to maximise the informative content of targeted sequencing-based tests while minimising their cost, by identifying concise regions of genomic space that act as effective **predictors** of the genomic landscape of a tumour. Furthermore, while we incorporate other data types (such as transcriptomics data) into our models in order to understand the **consequences** of genomic instability, all resulting tests will be based purely on targeted sequencing of DNA, meaning that the predictive biomarkers we develop will be applicable to liquid biopsy technology. Finally, the philosophy behind our work will be to approach modelling the genome globally rather than locally: we will spend relatively little time discussing individual genes or loci, instead attempting to understand genome/exome-wide patterns of mutation.

We begin by representing the profile of a tumour exome with a random vector M taking values in some domain $\mathcal{X}$. The precise format of this vector, and of the space $\mathcal{X}$ is not important here and will be chosen to reflect the structures we wish to model at any given point in time. The distribution of M will be given by a density function $p_M(\mathbf{m})$. This distribution will be extraordinary complex, and we will not in general have access to it. Instead, we will propose a parameterised family $\mathcal{P}$ of *generative models*, and from this family choose a best model $\bar{p}(\mathbf{m})$. We then define a *biomarker of interest* as a function $f : \mathcal{X} \to \mathbb{R}$.

2 Generative models of mutation

2.1 The nature of generative models

Generative models attempt to capture the underlying distribution of complex data in a manner that allows new samples to be drawn from the same distribution efficiently. They come in a variety of classes and have been a particular focus of research in the machine learning community in the last decade, being utilised for data compression, representation learning, and as a means to generate new samples from complex distributions. Generative models have found particular application in in disciplines dealing with extremely high-dimensional, complex

data distributions, including image analysis and natural language processing as well as, more recently, genomics. Generative models often (but not always) attempt to learn some lower-dimensional latent representation of their high-dimensional inputs, and as such are related to the theory of dimensionality reduction, and of unsupervised learning in general.

2.2 Strategies for fitting generative models

Maximum Likelihood Estimation (Invertible Models)
Monte-Carlo Markov Chain (MCMC) Methods
Variational Inference
(Blei et al., 2017)

2.3 Discrete, high-dimensional, sparse and bursty: the challenges of mutation data

(Zhao et al., 2020)

2.4 Previous work in generative models of genome-wide mutation

Uniform Rates
(Budczies et al., 2019)
Variable Rates
(Yao et al., 2020)
Latent Variables: Random Matrix Factorisation
(Fantini et al., 2020)

2.5 A dash of impropriety: non-generative models

3 Learning from learning: biomarker prediction

This section consists, in part, of discussion adapted from 'Dimensionality and Structure in Cancer Genomics: A Statistical Learning Perspective' (Bradley, 2020). This was published as the third chapter of the book 'Artificial Intelligence in Oncology Drug Discovery and Development' (Cassidy and Taylor, 2020).

Chapter 3

Tumour Mutation Burden and Tumour Indel Burden

1 Introduction and previous work

This chapter is based on the paper 'Data-driven design of targeted gene panels for estimating immunotherapy biomarkers' (Bradley and Cannings, 2021a), and the methodology contained is implemented by the R package `ICBioMark` (Bradley and Cannings, 2021b). We demonstrate a simple example of the workflow in Chapter 2 by proposing a novel, data-driven method for estimating TMB and TIB. For our generative model, we consider mutation counts as independent Poisson variables, where the mean number of mutations depends on the gene of origin and variant type, as well as the Background Mutation Rate (BMR) of the tumour. Due to the ultrahigh-dimensional nature of sequencing data and the fact that in many genes mutations arise purely according to the BMR, we use a regularisation penalty when estimating the generative parameters. In addition, this identifies a subset of genes that are mutated above or below the background rate. We then derive a linear estimator of TMB, which is chosen to be sparse (i.e. have many entries equal to zero), so that our estimator of TMB may be calculated using only the mutation counts in a subset of genes. In particular, this allows for accurate estimation of TMB from a targeted gene panel, where the panel size (and therefore the cost) may be determined by the user. We demonstrate the practical performance of our framework using a Non-Small Cell Lung Cancer (NSCLC) dataset (Chalmers et al., 2017), and include a comparison with existing state-of-the-art approaches for estimating TMB. Moreover, since our model allows variant type-dependent mutation rates, it can be adapted easily to predict other biomarkers, such as TIB. Finally, our method may also be used in combination with an existing targeted gene panel. In particular, we can estimate a biomarker directly from the panel, or first augment the panel and then construct an estimator.

Due to its emergence as a biomarker for immunotherapy in recent years, a variety of groups have considered methods for estimating TMB. A simple and common way to estimate TMB is via the proportion of mutated codons in a targeted region. Budczies et al. (2019) investigate how the accuracy of predictions made in this way are affected by the size of the targeted region, where mutations are assumed to occur at uniform rate throughout the genome. More recently Yao et al. (2020) modelled mutations as following a negative binomial distribution while allowing for gene-dependent rates, which are inferred by comparing nonsynonymous and synonymous mutation counts. In contrast, our method does not require data including synonymous mutations. Where they are included, we do not assume that synonymous mutations occur at a uniform rate throughout the genome, giving us the flexibility to account for location-specific effects on synonymous mutation rate such as chromatin configuration (Makova and Hardison, 2015) and transcription-dependent repair mechanisms (Fong et al., 2013). Linear regression models have been used for both panel selection (Lyu et al., 2018) and for biomarker prediction (Guo et al., 2020). A review of some of the issues arising when dealing with targeted panel-based predictions of TMB biomarkers is given by Wu et al. (2019a). Finally, we are unaware of any methods for estimating TIB from targeted gene panels.

The remainder of the chapter is organised as follows. In Section 2, we introduce our data sources, and provide a detailed description of our methodological proposal. Experimental results are given in Section 3 and we conclude in Section 4.

2 Methodology

2.1 Data and terminology

Our methodology can be applied to any annotated mutation dataset obtained by WES. To demonstrate our proposal we make use of the NSCLC dataset produced by Campbell et al. (2016), which contains data from 1144 patient-derived tumours. For each sample in this dataset we have the genomic locations and variant types of all mutations identified. At the time of the study, the patients had a variety of prognoses and smoking histories, were aged between 39 and 90, 41% were female and 59% were male; see Figure 3.1. In Figure 3.2A we see that mutations counts are distributed over a very wide range, as is the case in many cancer types (Chalmers et al., 2017). For simplicity, we only consider seven nonsynonymous variant types: missense mutations (which are the most abundant), nonsense mutations, frameshift insertions/deletions, splice site mutations, in-frame insertions/deletions, nonstop mutations and translation start site mutations. We present the frequencies of these mutation types in Figure 3.2B. Frameshift insertion/deletion (also known as indel) mutations are of particular interest when predicting TIB, but contribute only a small proportion ($< 4\%$) of nonsynonymous mutations.

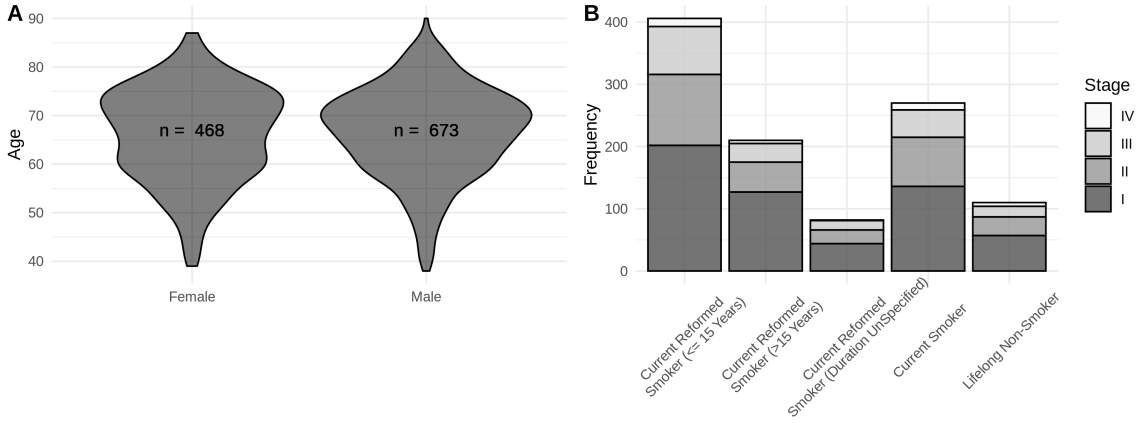

Figure 3.1: Demographic data for the clinical cohort in Campbell et al. (2016). **A**: Violin plots of age for patients, stratified by sex. **B**: Stacked bar chart of patients' smoking histories, shaded according to cancer stage diagnosis.

It is useful at this point to introduce the notation used throughout the paper. The set G denotes the collection of genes that make up the exome. For a gene $g \in G$, let ℓ_g be the length of g in nucleotide bases, defined by the maximum coding sequence[1]. A gene panel is a subset $P \subseteq G$, and we write $\ell_P := \sum_{g \in P} \ell_g$ for its total length. We let S denote the set of variant types in our data (e.g. in the dataset mentioned above, S contains the seven possible non-synonymous variants). Now, for $i = 0, 1, \ldots, n$, let M_{igs} denote the count of mutations in gene $g \in G$ of type $s \in S$ in the ith sample. Here the index $i = 0$ is used to refer to an unseen test sample for which we would like to make a prediction, while the indices $i = 1, \ldots, n$ enumerate the samples in our training data set. In order to define the exome-wide biomarker of particular interest, we specify a subset of mutation types $\bar{S} \subseteq S$, and let

$$T_{i\bar{S}} := \sum_{g \in G} \sum_{s \in \bar{S}} M_{igs}, \tag{3.1}$$

[1]The maximum coding sequence is defined as the collection of codons that may be translated for some version of a gene, even if all the codons comprising the maximum coding sequence are never simultaneously translated. Gene coding lengths are extracted from the *Ensembl* database (Yates et al., 2020).

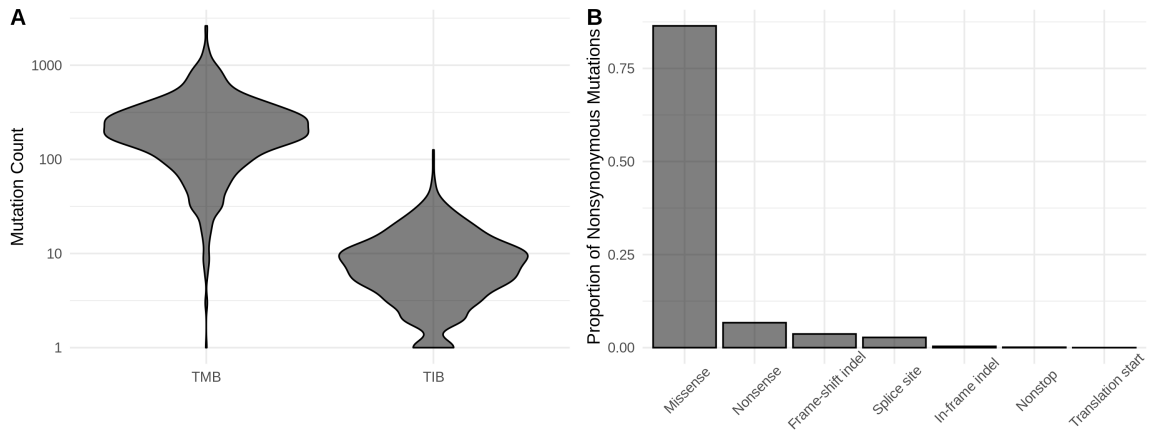

Figure 3.2: Dataset-wide distribution of mutations. **A**: Violin plot of the distribution of TMB and TIB across training samples. **B**: The relative frequency of different nonsynonymous mutation types.

for $i = 0, \ldots, n$. For example, including all non-synonymous mutation types in $\bar{S}$ specifies $T_{i\bar{S}}$ as the TMB of sample i, whereas letting $\bar{S}$ contain only indel mutations gives TIB.

Our main goal is to predict $T_{0\bar{S}}$ based on $\{M_{0gs} : g \in P, s \in S\}$, where the panel $P \subseteq G$ has length ℓ_P satisfying some upper bound. When it is clear from context that we are referring to the test sample and a specific choice of biomarker (i.e. $\bar{S}$ is fixed), we will simply write T in place of $T_{0\bar{S}}$.

2.2 Generative model

We now describe the main statistical model that underpins our methodology. In order to account for selective pressures and other factors within the tumour, we allow the rate at which mutations occur to depend on the gene and type of mutation. Our model also includes a sample-dependent parameter to account for the differing levels of mutagenic exposure of tumours, which may occur due to exogenous (e.g. UV light, cigarette smoke) or endogenous (e.g. inflammatory, free radical) factors.

We model the mutation counts M_{igs} as independent Poisson random variables with mutation rates $\phi_{igs} > 0$. More precisely, for $i = 0, 1, \ldots, n$, $g \in G$ and $s \in S$, we have

$$M_{igs} \sim \text{Poisson}(\phi_{igs}), \tag{3.2}$$

where M_{igs} and $M_{i'g's'}$ are independent for $(i, g, s) \neq (i', g', s')$. Further, to model the dependence of the mutation rate on the sample, gene and mutation type, we use a log-link function and let

$$\log(\phi_{igs}) = \mu_i + \log(\ell_g) + \lambda_g + \nu_s + \eta_{gs}, \tag{3.3}$$

for $\mu_i, \lambda_g, \nu_s, \eta_{gs} \in \mathbb{R}$, where for identifiability we set $\eta_{gs_1} = 0$, for some $s_1 \in S$ and all $g \in G$.

The terms in our model can be interpreted as follows. First, the parameter μ_i corresponds to the BMR of the ith sample. The offset $\log(\ell_g)$ accounts for a mutation rate that is proportional to the length of a gene, so that a non-zero value of λ_g corresponds to increased or decreased mutation rate relative to the BMR. The parameters ν_s and η_{gs} account for differences in frequency between mutation types for each gene.

The model in (3.2) and (3.3) (discounting the unseen test sample $i = 0$) has $n + |S| + |G||S|$ free parameters and we have $n|G||S|$ independent observations in the training data set. In principle we could attempt to fit our model directly using maximum likelihood estimation. However, we wish to exploit the fact that most genes do not play an active role in the development of a tumour, and will be mutated approximately according to the BMR. This corresponds to the parameters λ_g and η_{gs} being zero for many $g \in G$. We therefore include an ℓ_1-penalisation term applied to the parameters λ_g and η_{gs} when fitting our model. We do not penalise the

parameters ν_s or μ_i.

Writing $\mu := (\mu_1, \ldots, \mu_n)$, $\lambda := (\lambda_g : g \in G)$, $\nu := (\nu_s : s \in S)$ and $\eta := (\eta_{gs} : g \in G, s \in S)$, and given training observations $M_{igs} = m_{igs}$, we let

$$\mathcal{L}(\mu, \lambda, \nu, \eta) = \sum_{i=1}^{n} \sum_{g \in G} \sum_{s \in S} \Big(\phi_{igs} - m_{igs} \log \phi_{igs} \Big)$$

be the negative log-likelihood of the model specified by (3.2) and (3.3). We then define

$$(\hat{\mu}, \hat{\lambda}, \hat{\nu}, \hat{\eta}) = \operatorname*{arg\,min}_{\mu, \lambda, \nu, \eta} \Big\{ \mathcal{L}(\mu, \lambda, \nu, \eta) + \kappa_1 \Big(\sum_{g \in G} |\lambda_g| + \sum_{g \in G} \sum_{s \in S} |\eta_{gs}| \Big) \Big\}, \tag{3.4}$$

where $\kappa_1 \geq 0$ is a tuning parameter that controls the number of non-zero components in $\hat{\lambda}$ and $\hat{\eta}$, which we choose using cross-validation (see Section 2.5 for more detail).

2.3 Proposed estimator

We now attend to our main goal of estimating a given exome-wide biomarker for the unseen test sample. Fix $\bar{S} \subseteq S$ and recall that we write $T = T_{0\bar{S}}$. We wish to construct an estimator of T that only depends on the mutation counts in a gene panel $P \subset G$, subject to a constraint on ℓ_P. To that end, we consider estimators of the form[2]

$$T(w) := \sum_{g \in G} \sum_{s \in S} w_{gs} M_{0gs},$$

for $w \in \mathbb{R}^{|G| \times |S|}$. In the remainder of this subsection we explain how the weights w are chosen to minimise the expected squared error of $T(w)$ based on the generative model in Section 2.2.

Of course, setting $w_{gs} = 1$ for $g \in G$ and $s \in \bar{S}$ (and $w_{gs} = 0$ otherwise) will give $T(w) = T$. However, our aim is to make predictions based on a concise gene panel. If, for a given $g \in G$, we have $w_{gs} = 0$ for all $s \in S$, then $T(w)$ does not depend on the mutations in g and therefore the gene does not need to be included in the panel. In order to produce a suitable gene panel (i.e. with many $w_{gs} = 0$), we penalise non-zero components of w when minimising the expected squared error. We define our final estimator via a refitting procedure, which improves the predictive performance by reducing the bias, and is also helpful when applying our procedure to panels with predetermined genes.

To construct our estimator, note that under our model in (3.2) we have $\mathbb{E}M_{0gs} = \mathrm{Var}(M_{0gs}) = \phi_{0gs}$, and it follows that the expected squared error of $T(w)$ is

$$\begin{aligned} \mathbb{E}\big[\{T(w) - T\}^2\big] &= \mathrm{Var}(T(w)) + \mathrm{Var}(T) - 2\mathrm{Cov}(T(w), T) + \big[\mathbb{E}\{T(w) - T\}\big]^2 \\ &= \sum_{g \in G} \sum_{s \in \bar{S}} (1 - w_{gs})^2 \phi_{0gs} + \sum_{g \in G} \sum_{s \in S \setminus \bar{S}} w_{gs}^2 \phi_{0gs} \\ &\qquad + \Big(\sum_{g \in G} \sum_{s \in S} w_{gs} \phi_{0gs} - \sum_{g \in G} \sum_{s \in \bar{S}} \phi_{0gs} \Big)^2. \end{aligned} \tag{3.5}$$

This depends on the unknown parameters μ_0, λ_g, ν_s and η_{gs}, the latter three of which are replaced by their estimates given in (3.4). It is also helpful to then rescale (3.5) as follows: write $\hat{\phi}_{0gs} = \ell_g \exp(\hat{\lambda}_g + \hat{\nu}_s + \hat{\eta}_{gs})$, and define

$$p_{gs} := \frac{\hat{\phi}_{0gs}}{\sum_{g' \in G} \sum_{s' \in \bar{S}} \hat{\phi}_{0g's'}} = \frac{\ell_g \exp(\hat{\lambda}_g + \hat{\nu}_s + \hat{\eta}_{gs})}{\sum_{g' \in G} \sum_{s' \in \bar{S}} \ell_{g'} \exp(\hat{\lambda}_{g'} + \hat{\nu}_{s'} + \hat{\eta}_{g's'})}.$$

[2]Note that our estimator may use the the full set S of variant types, rather than just those in $\bar{S}$. In other words, our estimator may utilise information from every mutation type, not just those that directly constitute the biomarker of interest. This is important when estimating mutation types in $\bar{S}$ that are relatively scarce (e.g. for TIB).

Then let

$$f(w) := \sum_{g\in G}\sum_{s\in\bar{S}} p_{gs}(1-w_{gs})^2 + \sum_{g\in G}\sum_{s\in S\setminus\bar{S}} p_{gs}w_{gs}^2 + K(\mu_0)\big(1-\sum_{g\in G}\sum_{s\in S} p_{gs}w_{gs}\big)^2,$$

where $K(\mu_0) = \exp(\mu_0)\sum_{g\in G}\sum_{s\in\bar{S}}\ell_g\exp(\hat{\lambda}_g+\hat{\nu}_s+\hat{\eta}_{gs})$. Since f is a rescaled version of the error in (3.5) (with the true parameters λ,ν,η replaced by the estimates $\hat{\lambda},\hat{\nu},\hat{\eta}$), we will choose w to minimise $f(w)$.

Note that f only depends on μ_0 via the $K(\mu_0)$ term, which can be interpreted as a penalty factor controlling the bias of our estimator. For example, we may insist that the squared bias term $(1-\sum_{g\in G}\sum_{s\in S}p_{gs}w_{gs})^2$ is zero by setting $K(\mu_0)=\infty$. In practice, we propose to choose the penalty K based on the training data; see Section 2.5.

At this point $f(w)$ is minimised by choosing w to be such that $w_{gs}=1$ for all $g\in G, s\in\bar{S}$, and $w_{gs}=0$ otherwise. As mentioned above, in order to form a concise panel while optimising predictive performance, we impose a constraint on the cost of sequencing the genes used in the estimation. More precisely, for a given w, an appropriate cost is

$$\|w\|_{G,0} := \sum_{g\in G}\ell_g\mathbb{1}\{w_{gs}\neq 0 \text{ for some } s\in S\}.$$

This choice acknowledges that the cost of a panel is roughly proportional to the length of the region of genomic space sequenced, and that once a gene has been sequenced for one mutation type there is no need to sequence again for other mutation types.

Now, given a cost restriction L, our goal is to minimise $f(w)$ such that $\|w\|_{G,0}\leq L$. In practice this problem is non-convex and so computationally infeasible. As is common in high-dimensional optimisation problems, we consider a convex relaxation as follows: let $\|w\|_{G,1} := \sum_{g\in G}\ell_g\|w_g\|_2$, where $w_g=(w_{gs}:s\in S)\in\mathbb{R}^{|S|}$, for $g\in G$, and $\|\cdot\|_2$ is the Euclidean norm. Define

$$\hat{w}^{\text{first-fit}} \in \arg\min_{w}\big\{f(w)+\kappa_2\|w\|_{G,1}\big\}, \tag{3.6}$$

where $\kappa_2\geq 0$ is chosen to determine the size of the panel selected.

The final form of our estimator is obtained by a refitting procedure. First, for $P\subseteq G$, let

$$W_P := \{w\in\mathbb{R}^{|G|\times|S|} : w_g=(0,\ldots,0) \text{ for } g\in G\setminus P\}. \tag{3.7}$$

Let $\hat{P} := \{g\in G:\ \|\hat{w}_g^{\text{first-fit}}\|_2>0\}$ be the panel selected by the first-fit estimator in (3.6), and define

$$\hat{w}^{\text{refit}} \in \arg\min_{w\in W_{\hat{P}}}\big\{f(w)\big\}. \tag{3.8}$$

We then estimate T using $\hat{T} := T(\hat{w}^{\text{refit}})$, which only depends on mutations in genes contained in the selected panel $\hat{P}$. The performance of our estimator is investigated in Section 3, for comparison we also include the performance of the first-fit estimator $T(\hat{w}^{\text{first-fit}})$.

2.4 Panel augmentation

In practice, when designing gene panels a variety of factors contribute to the choice of genes included. For example, a gene may be included due to its relevance to immune response or its known association with a particular cancer type. If this is the case, measurements for these genes will be made regardless of their utility for predicting exome-wide biomarkers. When implementing our methodology, therefore, there is no additional cost to incorporate observations from these genes into our prediction if they will be helpful. Conversely researchers may wish to exclude genes from a panel, or at least from actively contributing to the estimation of a biomarker, for instance due to technical difficulties in sequencing a particular gene.

We can accommodate these restrictions by altering the structure of our regularisation penalty in (3.6). Suppose we are given (disjoint sets of genes) $P_0, Q_0\subseteq G$ to be included

and excluded from our panel, respectively. In this case, we replace $\hat{w}^{\text{first-fit}}$ in (3.6) with

$$\hat{w}^{\text{first-fit}}_{P_0,Q_0} \in \underset{w \in W_{G\setminus Q_0}}{\arg\min} \big\{ f(w) + \kappa_2 \sum_{g \in G\setminus P_0} l_g \|w_g\|_2 \big\}. \tag{3.9}$$

Excluding the elements of P_0 from the penalty term means that $\hat{w}^{\text{first-fit}}_{P_0,Q_0} \neq 0$ for the genes in P_0, while restricting our optimisation to $W_{G\setminus Q_0}$ excludes the genes in Q_0 by definition. This has the effect of augmenting the predetermined panel P_0 with additional genes selected to improve predictive performance. We then perform refitting as described above. We demonstrate this procedure by augmenting the TST-170 gene panel in Section 3.4.

2.5 Practical considerations

In this section, we discuss some practical aspects of our proposal. Our first consideration concerns the choice of the tuning parameter κ_1 in (3.4). As is common for the Least Absolute Shrinkage and Selection Operator (LASSO) estimator in generalised linear regression (see, for example, Michoel (2016) and Friedman et al. (2020)), we will use 10-fold cross-validation. To highlight one important aspect of our cross-validation procedure, recall that we consider the observations M_{igs} as independent across the sample index $i \in \{1, \dots, n\}$, the gene $g \in G$ and the mutation type $s \in S$. Our approach therefore involves splitting the entire set $\{(i, g, s) : i = 1, \dots, n, g \in G, s \in S\}$ of size $n|G||S|$ (as opposed to the sample set $\{1, \dots, n\}$) into 10 folds uniformly at random. We then apply the estimation method in (3.4) to each of the 10 folds separately on a grid of values (on the log scale) of κ_1, and select the value that results in the smallest average deviance across the folds. The model is then refitted using all the data for this value of κ_1.

The estimated coefficients in (3.6) depend on the choice of $K(\mu_0)$ and κ_2. As mentioned above, we could set $K(\mu_0) = \infty$ to give an unbiased estimator, however in practice we found that a finite choice of $K(\mu_0)$ leads to improved predictive performance. Our recommendation is to use $K(\mu_0) = K(\max_{i=1,\dots,n}\{\hat{\mu}_i\})$, where $\hat{\mu}_i = \log(T_i / \sum_{g,s} \ell_g \exp(\hat{\lambda}_g + \hat{\nu}_s + \hat{\eta}_{gs}))$ is a pseudo-MLE (in the sense of Gong and Samaniego (1981)) for μ_i, so that the penalisation is broadly in proportion with the largest values of μ_i in the training dataset. The tuning parameter κ_2 controls the size of the gene panel selected in (3.6): given a panel length L, we set $\kappa_2(L) = \max\{\kappa_2 : \ \ell_{\hat{P}} \leq L\}$ in order to produce a suitable panel.

We now comment briefly on some computational aspects of our method. The generative model fit in (3.4) can be solved via coordinate descent (see, for example, Friedman et al., 2010), which has a computational complexity of $O(N|G|^2|S|^2)$ per iteration. We fit the model 10 times, one for each fold in our cross-validation procedure. This is the most computationally demanding part of our proposal – in our experiments below, it takes approximately an hour to solve on a standard laptop – but it only needs to be carried out once for a given dataset. The convex optimisation problem in (3.6) can be solved by any method designed for the group LASSO; see, for example, Yang and Zou (2015). In our experiments in Section 3, we use the `gglasso` R package (Yang et al., 2020), which takes around 10 minutes to reproduce the plot in Figure 3.6. Note also that the solutions to (3.6) and (3.8) are unique; see, for example, Roth and Fischer (2008, Theorem 1). The last step of our proposal, namely making predictions for new test observations based on a selected panel, carries negligible computational cost.

Finally we describe a heuristic procedure for producing prediction intervals around our point estimates. In particular, for a given confidence level $\alpha \in (0, 1)$, we aim to find an interval $[\hat{T}_{\text{L}}, \hat{T}_{\text{U}}]$ such that $\mathbb{P}\big(\hat{T}_{\text{L}} \leq T \leq \hat{T}_{\text{U}}\big) \geq 1 - \alpha$. To that end, let $t_\alpha := \mathbb{E}\{(\hat{T} - T)^2\}/\alpha$, then by Markov's inequality we have that $\mathbb{P}(|\hat{T} - T|^2 \geq t_\alpha) \leq \alpha$. It follows that $[\hat{T} - t_\alpha^{1/2}, \hat{T} + t_\alpha^{1/2}]$ is a $(1 - \alpha)$-prediction interval for T. Of course, the mean squared error $\mathbb{E}\{(\hat{T} - T)^2\}$ defined in (3.5) depends on the parameters λ, η, ν and μ_0, which are unknown. Our approach is to utilise the estimates $\hat{\lambda}, \hat{\eta}, \hat{\nu}$ (see (3.4)) and replace μ_0 with $\log(\hat{T} / \sum_{g,s} \ell_g \exp(\hat{\lambda}_g + \hat{\nu}_s + \hat{\eta}_{gs}))$. While this is not an exact $(1 - \alpha)$-prediction interval for T, we will see in our experimental results in Sections 3.2 and 3.3 that in practice this approach provides intervals with valid empirical coverage.

3 Experimental results

In this section we demonstrate the practical performance of our proposal using the dataset from Campbell et al. (2016), which we introduced in Section 2.1. Our main focus is the prediction of TMB, and we show that our approach outperforms the state-of-the-art approaches. We also analyse the suitability of our generative model, consider the task of predicting the recently proposed biomarker TIB, and include a panel augmentation case study with the Foundation One gene panel.

Since we are only looking to produce estimators for TMB and TIB, we group mutations into two categories – *indel* mutations and *all other non-synonymous* mutations – so that $|S| = 2$. This simplifies the presentation of our results and reduces the computational cost of fitting the generative model. In order to assess the performance of each of the methods in this section, we randomly split the dataset into training, validation and test sets, which contain $n_{\text{train}} = n = 800$, $n_{\text{val}} = 171$ and $n_{\text{test}} = 173$ samples, respectively. Mutations are observed in $|G| = 17358$ genes. Our training set comprises samples with an average TMB of 252 and TIB of 9.25.

3.1 Generative model fit and validation

The first step in our analysis is to fit the model proposed in Section 2.2 using only the training dataset. In particular, we obtain estimates of the model parameters using equation (3.4), where the tuning parameter κ_1 is determined using 10-fold cross-validation as described in Section 2.5. The results are presented in Figure 3.3. The best choice of κ_1 produces estimates of λ and η with 44.4% and 77.8% sparsity respectively, i.e. that proportion of their components are estimated to be exactly zero. We plot $\hat{\lambda}$ and $\hat{\eta}$ for this value of κ_1 in Figures 3.4 and 3.5. Genes with $\hat{\lambda}_g = 0$ are interpreted to be mutating according to the background mutation rate, and genes with $\hat{\eta}_{\text{g,indel}} = 0$ are interpreted as having no specific selection pressure for or against indel mutations. In Figures 3.4 and 3.5 we highlight genes with large (in absolute value) parameter estimates, some of which have known biological relevance in oncology; see Section 4 for further discussion.

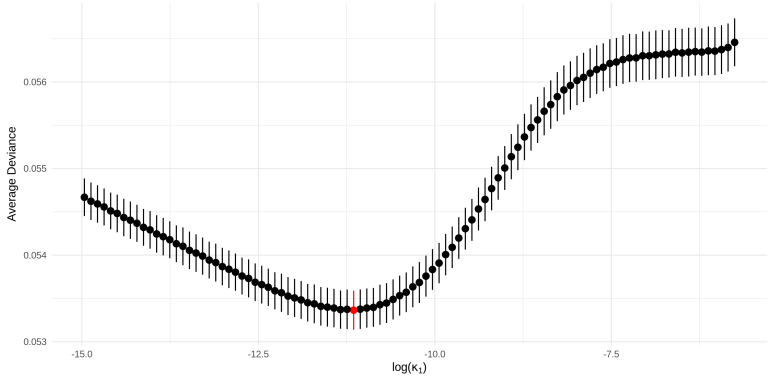

Figure 3.3: The average deviance (with one standard deviation) across the 10 folds in our cross-validation procedure plotted against $\log(\kappa_1)$. The minimum average deviance is highlighted red.

We now validate our model in (3.3) by comparing with the following alternatives:

(i) *Saturated model*: the model in (3.2), where each observation has an associated free parameter (i.e. $\phi_{igs} > 0$ is unrestricted);

(ii) *No sample-specific effects*: the model in (3.3), with $\mu_i = 0$ for all $i \in \{1, \ldots, n\}$;

(iii) *No gene-specific effects*: the model in (3.3), with $\lambda_g = \eta_{gs} = 0$ for all $g \in G$ and $s \in S$;

(iv) *No gene/mutation type interactions*: the model in (3.3), with $\eta_{gs} = 0$ for all $g \in G$ and $s \in S$.

In Table 3.1 we present the residual deviance and the residual degrees of freedom between our model and each of the models above. We see that our model is preferred over the saturated model, and all three submodels of (3.3).

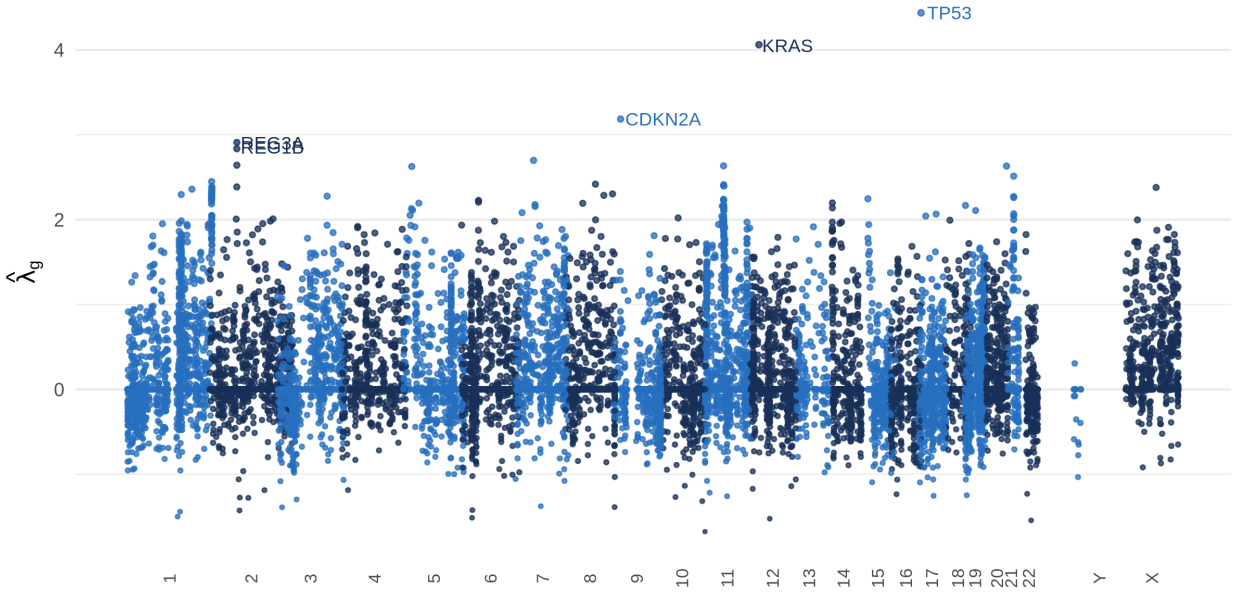

Figure 3.4: Manhattan plot of fitted parameters $\hat{\lambda}_g$ and their associated genes' chromosomal locations. The genes with the five largest positive parameter estimates are labelled.

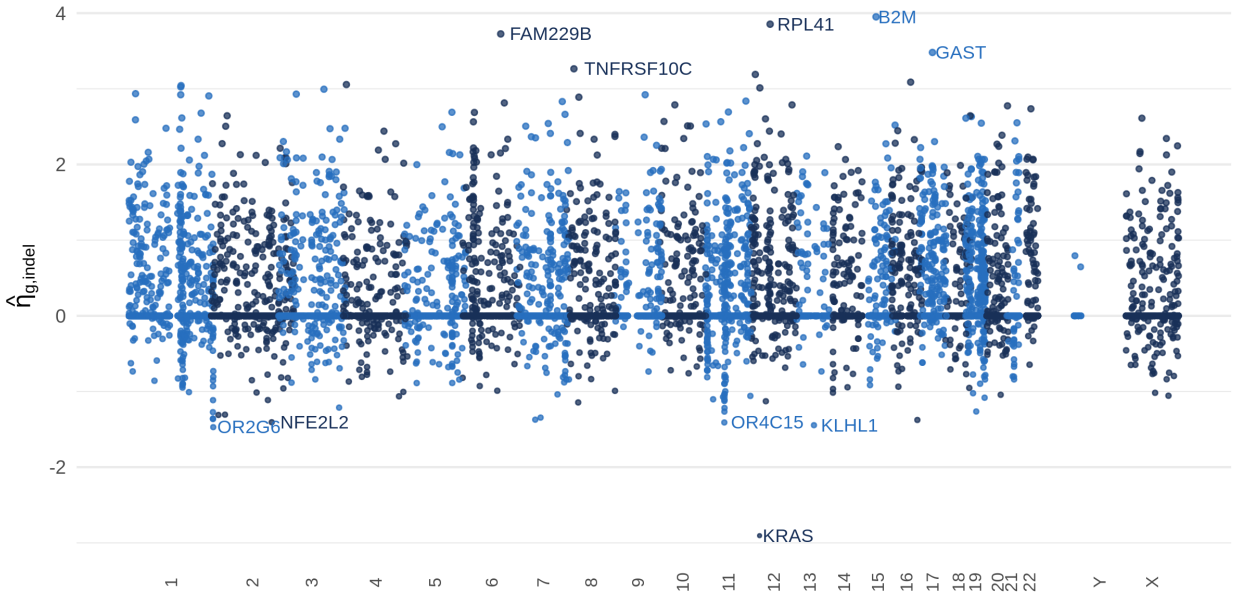

Figure 3.5: Manhattan plot of fitted parameters $\hat{\eta}_{g,\text{indel}}$ and their associated genes' chromosomal locations. The five largest positive and negative genes are labelled.

Table 3.1: Model comparisons on the basis of residual deviance statistics.

Comparison Model	Residual Deviance (dev)	Residual Degrees of Freedom (df)	dev/df	p-value
(i)	1.43×10^6	2.74×10^7	5.22×10^{-2}	1.00
(ii)	1.42×10^5	8.00×10^2	1.77×10^2	0.00
(iii)	1.10×10^5	1.33×10^4	8.24×10^0	0.00
(iv)	1.70×10^4	1.82×10^3	9.33×10^0	0.00

3.2 Predicting tumour mutation burden

We now demonstrate the excellent practical performance of our procedure for estimating TMB. First it is shown that our method can indeed select gene panels of size specified by the practitioner and that good predictions can be made even with small panel sizes (i.e. $\leq$ 1Mb). We then compare the performance of our proposal with state-of-the-art estimation procedures based on a number of widely used gene panels.

In order to evaluate the predictive performance of an estimator we calculate the R^2 score on the validation data as follows: given predictions of TMB, $\hat{t}_1, \ldots, \hat{t}_{n_{val}}$, for the observations in the validation set with true TMB values $t_1, \ldots, t_{n_{val}}$. Let $\bar{t} := \frac{1}{n_{val}} \sum_{i=1}^{n_{val}} t_i$, and define

$$R^2 := 1 - \frac{\sum_{i=1}^{n_{val}} (t_i - \hat{t}_i)^2}{\sum_{i=1}^{n_{val}} (t_i - \bar{t})^2}.$$

Other existing works have aimed to classify tumours into two groups (high TMB, low TMB); see, for example, Büttner et al. (2019) and Wu et al. (2019a). Here we also report the estimated Area Under Precision-Recall Curve (AUPRC) for a classifier based on our estimator. We define the classifier as follows: first, in line with major clinical studies (e.g. Hellmann et al., 2018; Ramalingam et al., 2018) the true class membership of a tumour is defined according to whether it has $t^* := 300$ or more exome mutations (approximately 10 Mut/Mb). In the validation set, this gives 47 (27.5%) tumours with high TMB and 124 (72.5%) with low TMB. Now, for a cutoff $t \geq 0$, we can define a classifier by assigning a tumour to the high TMB class if its estimated TMB value is greater than or equal to t. For such a classifier, we have precision and recall (estimated over the validation set) given by

$$p(t) := \frac{\sum_{i=1}^{n_{val}} \mathbb{1}_{\{\hat{t}_i \geq t,\ t_i \geq t^*\}}}{\sum_{i=1}^{n_{val}} \mathbb{1}_{\{\hat{t}_i \geq t\}}} \quad \text{and} \quad r(t) := \frac{\sum_{i=1}^{n_{val}} \mathbb{1}_{\{\hat{t}_i \geq t,\ t_i \geq t^*\}}}{\sum_{i=1}^{n_{val}} \mathbb{1}_{\{t_i \geq t^*\}}},$$

respectively. The precision-recall curve then is $\{(r(t), p(t)) : t \in [0, \infty)\}$. Note that a perfect classifier achieves a AUPRC of 1, whereas a random guess in this case would have an average AUPRC of 0.308 (the prevalence of the high TMB class).

Now recall that TMB is given by equation (3.1) with $\bar{S}$ being the set of all non-synonymous mutation types. Thus to estimate TMB we apply our procedure in Section 2.3 with $\bar{S} = S$, where the model parameters are estimated as described in Section 3.1. In Figure 3.6, we present the R^2 and AUPRC for the first-fit and refitted estimators (see (3.6) and (3.8)) as the selected panel size varies from 0Mb to 2Mb in length. We see that we obtain a more accurate prediction of TMB, both in terms of regression and classification, as the panel size increases, and that good estimation is possible even with very small panels (as low as 0.2Mb). Finally, as expected, the refitted estimator slightly outperforms the first-fit estimator.

We now compare our method with state-of-the-art estimators applied to commonly used gene panels. The three next-generation sequencing panels that we consider are chosen for their relevance to TMB. These are TST-170 (Heydt et al., 2018), Foundation One (Frampton et al., 2013) and MSK-IMPACT (Cheng et al., 2015). For each panel $P \subseteq G$, we use four different methods to predict TMB:

(i) Our refitted estimator applied to the panel P: we estimate TMB using $T(\hat{w}_P)$, where $\hat{w}_P \in \arg\min_{w \in W_P}\{f(w)\}$, and W_P is defined in (3.7).

(ii) Estimation and Classification of Tumour Mutation Burden (ecTMB): the procedure proposed by Yao et al. (2020).

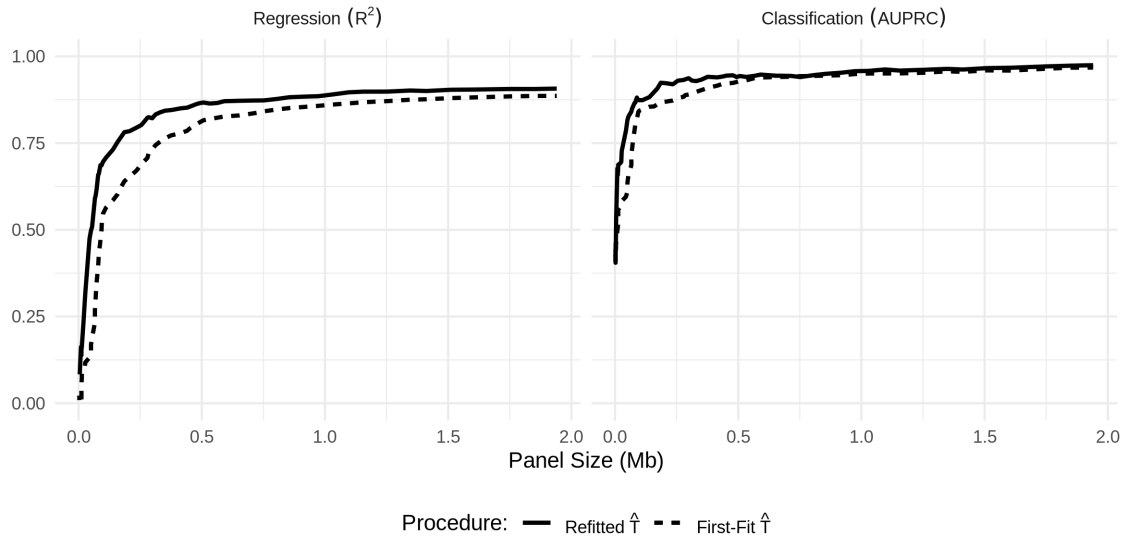

Figure 3.6: Performance of our first-fit and refitted estimators of TMB as the selected panel size varies. **Left**: R^2, **Right**: AUPRC.

(iii) A count estimator: TMB is estimated by $\frac{\ell_G}{\ell_P} \sum_{g \in P} \sum_{s \in \bar{S}} M_{0gs}$, i.e. rescaling the mutation burden in the genes of P.

(iv) A linear model: we estimate TMB via ordinary least-squares linear regression of TMB against $\left\{\sum_{s \in S} M_{0gs} : g \in P\right\}$.

The latter three comprise existing methods for estimating TMB available to practitioners. The second (ecTMB), which is based on a negative binomial model, is the state-of-the-art. The third and fourth are standard practical procedures for the estimation of TMB from targeted gene panels. The refitted estimator applied to the panel P is also included here, in order to demonstrate the utility of our approach even with a prespecified panel.

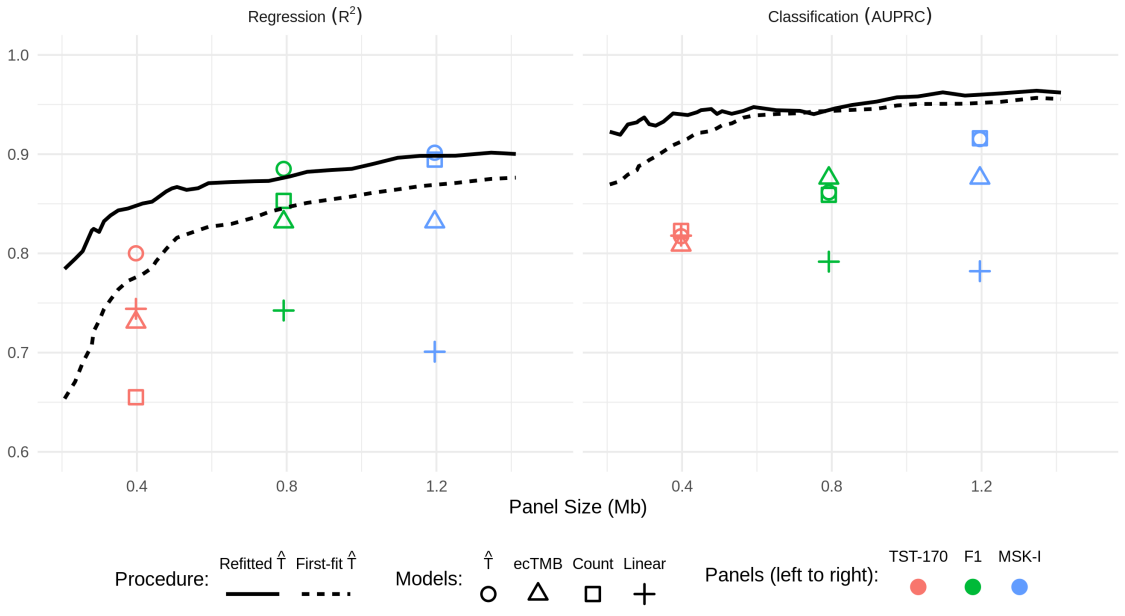

Figure 3.7: The performance of our TMB estimator in comparison to existing approaches. **Left**: R^2, **Right**: AUPRC.

We present results of these comparisons in Figure 3.7. First, for each of the three panels

considered here, we see that our refitted estimator applied to the panel outperforms all existing approaches in terms of regression performance, and that for smaller panels we are able to improve regression accuracy even further by selecting a panel based on the training data. For instance, in comparison to predictions based on the TST-170 panel, our procedure with a selected panel of the same size (0.4Mb) achieves an R^2 of 0.85. The best available existing method based on the TST-170 panel, in this case the linear estimator, has an R^2 of 0.74. Moreover, data-driven selection of panels considerably increases the classification performance for the whole range of panel sizes considered. In particular, even for the smallest panel size shown in Figure 3.7 (∼0.2Mb), the classification performance of our method outperforms the best existing methodology applied to the MSK-IMPACT panel, despite being almost a factor of six times smaller.

Finally in this section we demonstrate the practical performance of our method using the test set, which until this point has been held out. Based on the validation results above, we take the panel of size 0.6Mb selected by our procedure and use our refitted estimator on that panel to predict TMB for the 173 samples in the test set. For comparison, we also present predictions from ecTMB, the count-based estimator and the linear regression estimator applied to the same panel. In Figure 3.8 we see that our procedure performs well; we obtain an R^2 value (on the test data) of 0.85. The other methods have R^2 values of 0.67 (ecTMB), −36 (count) and 0.64 (linear regression). The count-based estimator here gives predictions which are reasonably well correlated to the true values of TMB but are positively biased. This is as expected, since our selection procedure tends to favour genes with higher overall mutation rates. We also include a red shaded region comprising all points for which heuristic 90% prediction intervals (as described in Section 2.5) include the true TMB value. We find in this case that 93.6% of the observations in the test set fall within this region, giving valid empirical coverage.

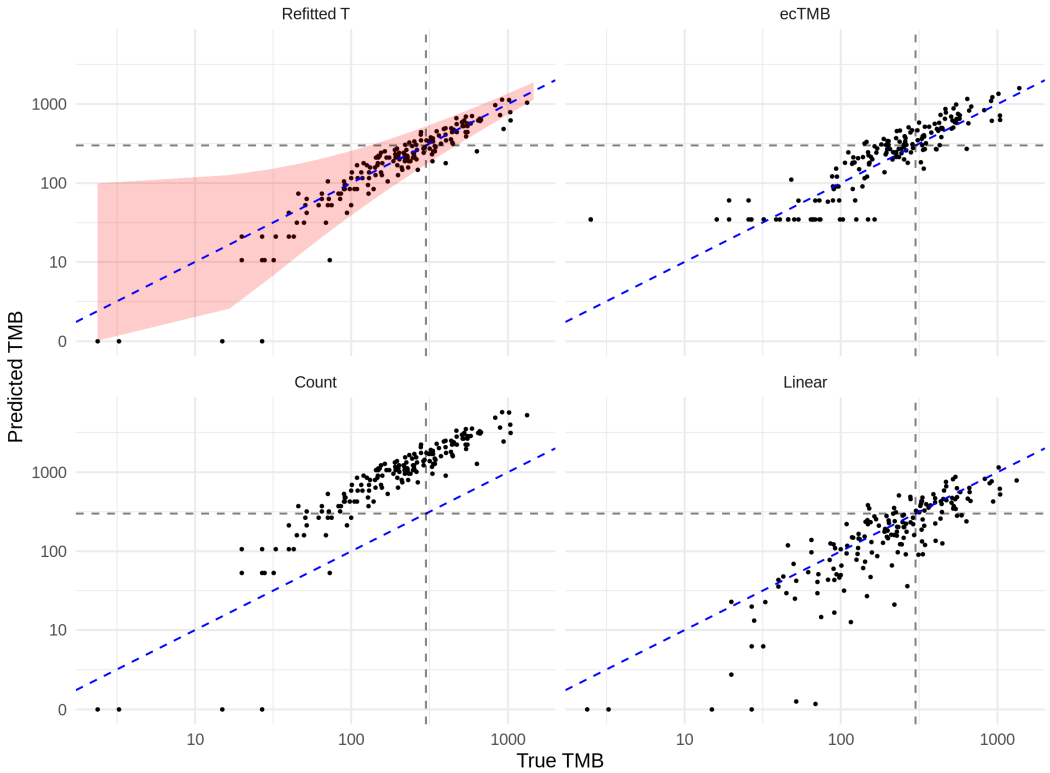

Figure 3.8: Prediction of TMB on the test dataset. Dashed blue (diagonal) line represents perfect prediction, and the black dashed lines indicate true and predicted TMB values of 300.

3.3 Predicting tumour indel burden

In this section we demonstrate how our method can be used to estimate TIB. This is more challenging than estimating TMB due to the low abundance of indel mutations relative to

other variant types (see Figure 3.2), as well as issues involved in sequencing genomic loci of repetitive nucleotide constitution (Narzisi and Schatz, 2015). Indeed, in contrast to the previous section, we are not aware of any existing methods designed to estimate TIB from targeted gene panels. We therefore investigate the performance of our method across a much wider range (0-30Mb) of panel sizes, and find that we are able to accurately predict TIB with larger panels. Our results also demonstrate that accurate classification of TIB status is possible even with small gene panels.

We let S_{indel} be the set of all frameshift insertion and deletion mutations, and apply our method introduced in Section 2.3 with $\bar{S} = S_{\text{indel}}$. As in the previous section, we assess regression and classification performance via R^2 and AUPRC, respectively, where in this case tumours are separated into two classes: high TIB (10 or more indel mutations) and low TIB (otherwise). In the validation dataset, this gives 57 (33.3%) tumours in the high TIB class.

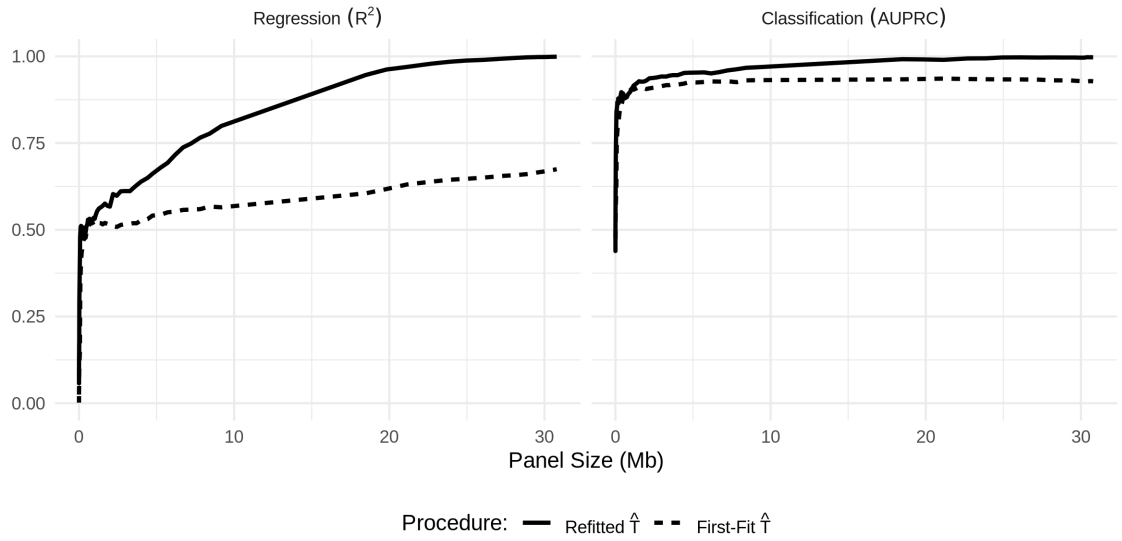

Figure 3.9: Performance of our first-fit and refitted estimators of TIB as the selected panel size varies. **Left**: R^2, **Right**: AUPRC.

The results are presented in Figure 3.9. We comment first on the regression performance: as expected, we see that the R^2 values for our first-fit and refitted estimators are much lower than what we achieved in estimating TMB. The refitted approach improves for larger panel sizes, while the first-fit estimator continues to perform relatively poorly. On the other hand, we see that the classification performance is impressive, with AUPRC values of above 0.8 for panels of less than 1Mb in size.

We now assess the performance on the test set of our refitted estimator of TIB applied to a selected panel of size 0.6Mb, and we compare with a count-based estimator and linear regression estimator. We do not compare with ecTMB here, since it is designed to estimate TMB as opposed to TIB. The count-based estimator in this case scales the total number of non-synonymous mutations across the panel by the ratio of the length of the panel to that of the entire exome, and also by the relative frequency of indel mutations versus all non-synonymous mutations in the training dataset:

$$\frac{\ell_G}{\ell_P} \frac{\sum_{i=1}^{n} \sum_{g \in G} \sum_{s \in S_{\text{indel}}} M_{igs}}{\sum_{i=1}^{n} \sum_{g \in G} \sum_{s \in S} M_{igs}} \sum_{g \in P} \sum_{s \in S} M_{0gs}.$$

In Figure 3.10 we present the predictions on the test set of our refitted estimator ($R^2 = 0.35$); the count estimator ($R^2 = -44$); and the linear regression estimator ($R^2 = 0.15$). We also include (shaded in red) the set of points for which 90% prediction intervals contain the true value. In this case we find that 97.7% of test set points fall within this region.

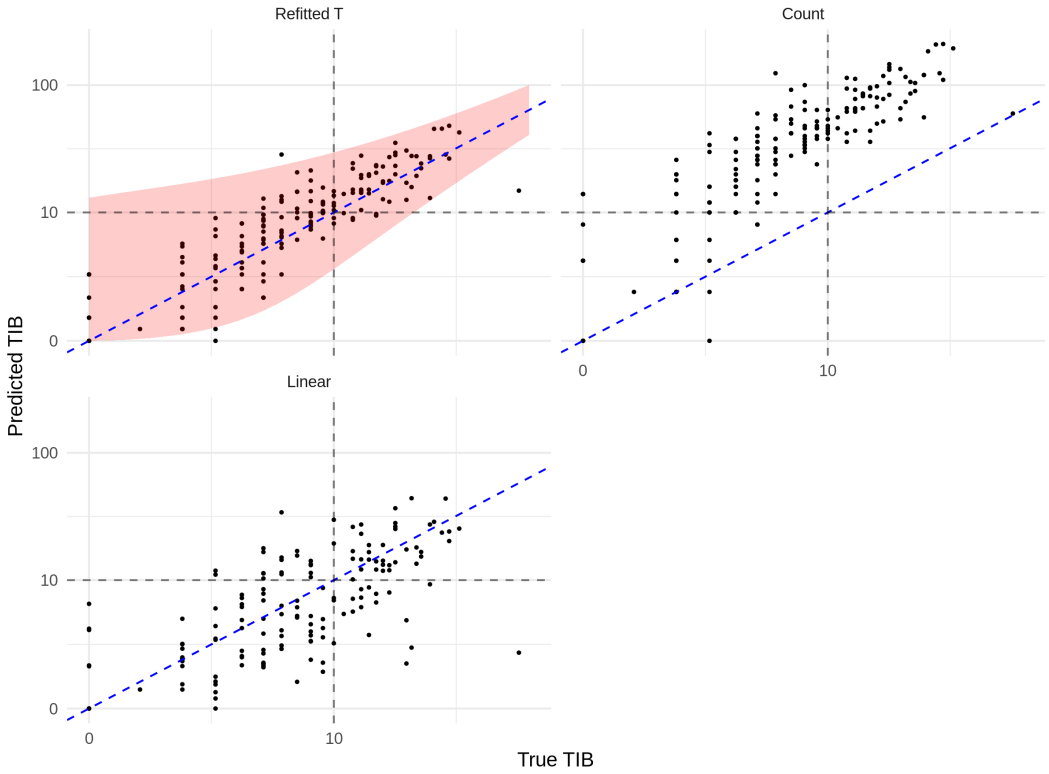

Figure 3.10: Estimation of TIB on the test dataset. Dashed blue (diagonal) line represents perfect prediction, and the grey dashed lines indicate true and predicted TIB values of 10.

3.4 A panel-augmentation case study

As discussed in Section 2.4, we may wish to include genes from a given panel, but use our methodology to augment the panel to include additional genes with goal of obtaining more accurate predictions of TMB (or other biomarkers). In this section we demonstrate how this can be done starting with the TST-170 panel (∼0.4Mb) and augmenting to 0.6Mb in length, demonstrating impressive gains in predictive performance.

We apply the augmentation method described in Section 2.4, with P_0 taken to be the set of TST-170 genes and Q_0 to be empty. The genes added to the panel are determined by the first-fit estimator in equation (3.9). To evaluate the performance, we then apply the refitted estimator on the test dataset, after selecting the augmented panel of size 0.6Mb. For comparison, we apply our refitted estimator to the TST-170 panel directly. We also present the results obtained by the other estimators described above, both before and after the panel augmentation, in Table 3.2. We find that by augmenting the panel we improve predictive performance with our refitted $\hat{T}$ estimator, both in terms of regression and classification. The refitted estimator provides better estimates than any other model on the augmented panel by both metrics.

Table 3.2: Predictive performance of models on TST-170 (0.4Mb) versus augmented TST-170 (0.6Mb) panels on the test set.

Model	Regression (R^2)		Classification (AUPRC)	
	TST-170	Aug. TST-170	TST-170	Aug. TST-170
Refitted $\hat{T}$	**0.58**	**0.84**	**0.83**	**0.94**
ecTMB	0.37	0.51	0.80	0.88
Count	0.18	0.18	**0.83**	**0.94**
Linear	0.47	0.74	0.78	0.89

4 Conclusions

We have introduced a new data-driven framework for designing targeted gene panels which allows for cost-effective estimation of exome-wide biomarkers. Using the Non-Small Cell Lung Cancer dataset from Campbell et al. (2016), we have demonstrated the excellent predictive performance of our proposal for estimating Tumour Mutation Burden and Tumour Indel Burden, and shown that it outperforms the state-of-the-art procedures. Our framework can be applied to any tumour dataset containing annotated mutations, and we provide an `R` package (Bradley and Cannings, 2021b) which implements the methodology.

Our work also has the scope to help understand mutational processes. For example, the parameters of our fitted model in Section 3.1 have interesting interpretations: of the five genes highlighted in Figure 3.4 as having the highest mutation rates relative to the BMR, three (*TP53, KRAS, CDKN2A*) are known tumour suppressors (Olivier et al., 2010; Jančík et al., 2010; Foulkes et al., 1997). Furthermore, indel mutations in *KRAS* are known to be deleterious for tumour cells (Lee et al., 2018) – in our work the *KRAS* gene has a large negative indel-specific parameter (see Figure 3.5). Our methodology identifies a number of other genes with large parameter estimates.

There are many ways in which this model can be extended. For example, it may be adapted to incorporate alternate data types (e.g. transcriptomics); we may seek to predict other features (e.g. outcomes such as survival); or we may wish to extend the method to incorporate multiple data sources (e.g. on different cancer types and tissues of origin). These developments will be discussed in following chapters.

Chapter 4

Causative Mechanisms of Genomic Instability

This chapter might be something to do with inferring causal pathways from learned generative models.

Chapter 5

Transcripts as Neoantigen Proxies: Expressed Mutation Burden

This chapter might be something to do with incorporating transcriptomic data into biomarkers for immunotherapy response.

Bibliography

M. Annala, G. Vandekerkhove, D. Khalaf, S. Taavitsainen, K. Beja, E. W. Warner, K. Sunderland, C. Kollmannsberger, B. J. Eigl, D. Finch, C. D. Oja, J. Vergidis, M. Zulfiqar, A. A. Azad, M. Nykter, M. E. Gleave, A. W. Wyatt, and K. N. Chi. Circulating Tumor DNA Genomics Correlate with Resistance to Abiraterone and Enzalutamide in Prostate Cancer. *Cancer Discovery*, 8(4):444–457, Apr. 2018. ISSN 2159-8274, 2159-8290. doi: 10.1158/2159-8290.CD-17-0937. URL http://cancerdiscovery.aacrjournals.org/lookup/doi/10.1158/2159-8290.CD-17-0937.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, Apr. 2017. ISSN 0162-1459. doi: 10.1080/01621459.2017.1285773. URL https://doi.org/10.1080/01621459.2017.1285773.

L. Boulter, E. Bullock, Z. Mabruk, and V. G. Brunton. The fibrotic and immune microenvironments as targetable drivers of metastasis. *British Journal of Cancer*, pages 1–10, Nov. 2020. ISSN 1532-1827. doi: 10.1038/s41416-020-01172-1. URL https://www.nature.com/articles/s41416-020-01172-1.

T. Boveri. Concerning the Origin of Malignant Tumours. Translated and annotated by Henry Harris. *Journal of Cell Science*, 121(Supplement 1):1–84, Jan. 2008. ISSN 0021-9533, 1477-9137. doi: 10.1242/jcs.025742. Publisher: The Company of Biologists Ltd Section: Article.

J. Bradley. Dimensionality and Structure in Cancer Genomics: A Statistical Learning Perspective. *Artificial Intelligence in Oncology Drug Discovery and Development*, Sept. 2020. doi: 10.5772/intechopen.92574. URL https://www.intechopen.com/books/artificial-intelligence-in-oncology-drug-discovery-and-development/dimensionality-and-structure-in-cancer-genomics-a-statistical-learning-perspective.

J. R. Bradley and T. I. Cannings. Data-driven design of targeted gene panels for estimating immunotherapy biomarkers. *arXiv:2102.04296 [q-bio, stat]*, Feb. 2021a. URL http://arxiv.org/abs/2102.04296. arXiv: 2102.04296.

J. R. Bradley and T. I. Cannings. ICBioMark: Data-Driven Design of Targeted Gene Panels for Estimating Immunotherapy Biomarkers (R Package), Jan. 2021b. URL https://github.com/cobrbra/ICBioMark.

E. I. Buchbinder and A. Desai. CTLA-4 and PD-1 Pathways: Similarities, Differences, and Implications of Their Inhibition. *American Journal of Clinical Oncology*, 39(1):98–106, Feb. 2016. ISSN 1537-453X. doi: 10.1097/COC.0000000000000239.

J. Budczies, M. Allgäuer, and K. Litchfield. Optimizing panel-based tumor mutational burden (TMB) measurement. *Annals of Oncology: Official Journal of the European Society for Medical Oncology*, 30(9):1496–1506, 2019. ISSN 1569-8041. doi: 10.1093/annonc/mdz205.

R. Büttner, J. W. Longshore, and F. López-Ríos. Implementing TMB measurement in clinical practice: considerations on assay requirements. *ESMO Open*, 4(1):e000442, Jan. 2019. ISSN 2059-7029. doi: 10.1136/esmoopen-2018-000442.

J. D. Campbell, A. Alexandrov, and J. Kim. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nature Genetics*, 48(6):607–616, 2016. ISSN 1546-1718. doi: 10.1038/ng.3564.

D. Cao, H. Xu, and X. Xu. High tumor mutation burden predicts better efficacy of immunotherapy: a pooled analysis of 103078 cancer patients. *Oncoimmunology*, 8(9):e1629258, 2019. ISSN 2162-4011. doi: 10.1080/2162402X.2019.1629258.

J. W. Cassidy and B. Taylor. *Artificial Intelligence in Oncology Drug Discovery and Development*. IntechOpen, Sept. 2020. ISBN 9781789858976. doi: 10.5772/intechopen.88376. URL https://www.intechopen.com/books/artificial-intelligence-in-oncology-drug-discovery-and-development.

Z. R. Chalmers, C. F. Connelly, and D. Fabrizio. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Medicine*, 9, Apr. 2017. ISSN 1756-994X. doi: 10.1186/s13073-017-0424-2.

T. A. Chan, M. Yarchoan, and E. Jaffee. Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic. *Annals of Oncology*, 30(1):44–56, Jan. 2019. ISSN 0923-7534, 1569-8041. doi: 10.1093/annonc/mdy495.

D. T. Cheng, T. N. Mitchell, and A. Zehir. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT). *The Journal of Molecular Diagnostics : JMD*, 17(3):251–264, May 2015. ISSN 1525-1578. doi: 10.1016/j.jmoldx.2014.12.006.

L. Fancello, S. Gandini, and P. G. Pelicci. Tumor mutational burden quantification from targeted gene panels: major advancements and challenges. *Journal for ImmunoTherapy of Cancer*, 7(1):183, July 2019. ISSN 2051-1426. doi: 10.1186/s40425-019-0647-4.

D. Fantini, V. Vidimar, Y. Yu, S. Condello, and J. J. Meeks. MutSignatures: An R Package for Extraction and Analysis of Cancer Mutational Signatures. *bioRxiv*, page 2020.03.15.992826, Mar. 2020. doi: 10.1101/2020.03.15.992826. URL https://www.biorxiv.org/content/10.1101/2020.03.15.992826v1.

Y. W. Fong, C. Cattoglio, and R. Tjian. The intertwined roles of transcription and repair proteins. *Molecular Cell*, 52(3):291–302, Nov. 2013. ISSN 1097-4164. doi: 10.1016/j.molcel.2013.10.018.

W. D. Foulkes, T. Y. Flanders, and P. M. Pollock. The CDKN2A (p16) gene and human cancer. *Molecular Medicine*, 3(1):5–20, Jan. 1997. ISSN 1076-1551.

G. M. Frampton, A. Fichtenholtz, and G. A. Otto. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nature Biotechnology*, 31(11):1023–1031, Nov. 2013. ISSN 1546-1696. doi: 10.1038/nbt.2696.

J. Friedman, T. Hastie, and R. Tibshirani. glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models, June 2020. URL https://CRAN.R-project.org/package=glmnet.

J. H. Friedman, T. Hastie, and R. Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22, Feb. 2010. ISSN 1548-7660. doi: 10.18637/jss.v033.i01.

D. R. Gandara, S. M. Paul, and M. Kowanetz. Blood-based tumor mutational burden as a predictor of clinical benefit in non-small-cell lung cancer patients treated with atezolizumab. *Nature Medicine*, 24(9):1441–1448, 2018. ISSN 1546-170X. doi: 10.1038/s41591-018-0134-3.

G. Genovese, A. K. Kähler, R. E. Handsaker, J. Lindberg, S. A. Rose, S. F. Bakhoum, K. Chambert, E. Mick, B. M. Neale, M. Fromer, S. M. Purcell, O. Svantesson, M. Landén, M. Höglund, S. Lehmann, S. B. Gabriel, J. L. Moran, E. S. Lander, P. F. Sullivan, P. Sklar, H. Grönberg, C. M. Hultman, and S. A. McCarroll. Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. *New England Journal of Medicine*, 371(26):2477–2487, Dec. 2014. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMoa1409405. URL http://www.nejm.org/doi/10.1056/NEJMoa1409405.

S. Gettinger, H. Borghaei, and J. Brahmer. 5-Year Outcomes From the Randomized, Phase 3 Trials CheckMate 017/057: Nivolumab vs Docetaxel in Previously Treated NSCLC. *Journal of Thoracic Oncology*, 14(10):S244–S245, Oct. 2019. ISSN 1556-0864. doi: 10.1016/j.jtho.2019.08.486.

M. Golkaram, C. Zhao, and K. Kruglyak. The interplay between cancer type, panel size and tumor mutational burden threshold in patient selection for cancer immunotherapy. *PLOS Computational Biology*, 16(11):e1008332, Nov. 2020. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1008332.

G. Gong and F. J. Samaniego. Pseudo Maximum Likelihood Estimation: Theory and Applications. *The Annals of Statistics*, 9(4):861–869, 1981. ISSN 0090-5364.

J. Goodall, J. Mateo, W. Yuan, H. Mossop, N. Porta, S. Miranda, R. Perez-Lopez, D. Dolling, D. R. Robinson, S. Sandhu, G. Fowler, B. Ebbs, P. Flohr, G. Seed, D. N. Rodrigues, G. Boysen, C. Bertan, M. Atkin, M. Clarke, M. Crespo, I. Figueiredo, R. Riisnaes, S. Sumanasuriya, P. Rescigno, Z. Zafeiriou, A. Sharp, N. Tunariu, D. Bianchini, A. Gillman, C. J. Lord, E. Hall, A. M. Chinnaiyan, S. Carreira, and J. S. de Bono. Circulating Cell-Free DNA to Guide Prostate Cancer Treatment with PARP Inhibition. *Cancer Discovery*, 7(9):1006–1017, Sept. 2017. ISSN 2159-8274, 2159-8290. doi: 10.1158/2159-8290.CD-17-0261. URL http://cancerdiscovery.aacrjournals.org/lookup/doi/10.1158/2159-8290.CD-17-0261.

W. Guo, Y. Fu, and L. Jin. An Exon Signature to Estimate the Tumor Mutational Burden of Right-sided Colon Cancer Patients. *Journal of Cancer*, 11(4):883–892, 2020. ISSN 1837-9664. doi: 10.7150/jca.34363.

D. Hanahan and R. A. Weinberg. Hallmarks of Cancer: The Next Generation. *Cell*, 144 (5):646–674, Mar. 2011. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2011.02.013. URL https://www.cell.com/cell/abstract/S0092-8674(11)00127-9.

M. D. Hellmann, T.-E. Ciuleanu, and A. Pluzanski. Nivolumab plus Ipilimumab in Lung Cancer with a High Tumor Mutational Burden. *New England Journal of Medicine*, 378(22): 2093–2104, May 2018. ISSN 0028-4793. doi: 10.1056/NEJMoa1801946.

C. Heydt, R. Pappesch, and K. Stecker. Evaluation of the TruSight Tumor 170 (TST170) assay and its value in clinical research. *Annals of Oncology*, 29:vi7–vi8, Sept. 2018. ISSN 0923-7534, 1569-8041. doi: 10.1093/annonc/mdy318.003.

D. H. Howard, P. B. Bach, E. R. Berndt, and R. M. Conti. Pricing in the Market for Anticancer Drugs. *Journal of Economic Perspectives*, 29(1):139–162, Feb. 2015. ISSN 0895-3309. doi: 10.1257/jep.29.1.139. URL https://www.aeaweb.org/articles?id=10.1257/jep.29.1.139.

Y. Ishida, Y. Agata, and K. Shibahara. Induced expression of PD-1, a novel member of the immunoglobulin gene superfamily, upon programmed cell death. *The EMBO journal*, 11(11): 3887–3895, Nov. 1992. ISSN 0261-4189.

S. Jančík, J. Drábek, and D. Radzioch. Clinical Relevance of KRAS in Human Cancers. *Journal of Biomedicine and Biotechnology*, 2010, 2010. ISSN 1110-7243. doi: 10.1155/2010/150960.

K. Jensen, E. Q. Konnick, M. T. Schweizer, A. O. Sokolova, P. Grivas, H. H. Cheng, N. M. Klemfuss, M. Beightol, E. Y. Yu, P. S. Nelson, B. Montgomery, and C. C. Pritchard. Association of Clonal Hematopoiesis in DNA Repair Genes With Prostate Cancer Plasma Cell-free DNA Testing Interference. *JAMA oncology*, Nov. 2020. ISSN 2374-2445. doi: 10.1001/jamaoncol.2020.5161.

P. Kanavos. The rising burden of cancer in the developing world, June 2006. URL https://pubmed.ncbi.nlm.nih.gov/16801335/.

T. E. Keenan, K. P. Burke, and E. M. Van Allen. Genomic correlates of response to immune checkpoint blockade. *Nature Medicine*, 25(3):389–402, Mar. 2019. ISSN 1546-170X. doi: 10.1038/s41591-019-0382-x. URL https://www.nature.com/articles/s41591-019-0382-x.

E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.-F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H.-C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. R. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. A. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S.-P. Yang, R.-F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, A. Patrinos, M. J. Morgan, International Human Genome Sequencing Consortium, C. f. G. R. Whitehead Institute for Biomedical Research, The Sanger Centre:, Washington University Genome Sequencing Center, US DOE Joint Genome Institute:, Baylor College of Medicine Human Genome Sequencing Center:, RIKEN Genomic Sciences Center:, Genoscope and CNRS UMR-8030:, I. o. M. B. Department of Genome Analysis, GTC Sequencing Center:, Beijing Genomics Institute/Human Genome Center:, T. I. f. S. B. Multimegabase Sequencing Center, Stanford Genome Technology Center:, University of Oklahoma's Advanced Center for Genome Technology:, Max Planck Institute for Molecular Genetics:, L. A. H. G. C. Cold Spring Harbor Laboratory, GBF—German Research Centre for Biotechnology:, a. i. i. l. u. o. h. *Genome Analysis Group (listed in alphabetical order, U. N. I. o. H. Scientific management: National Human Genome Research Institute, Stanford Human Genome Center:, University of Washington Genome Center:, K. U. S. o. M. Department of Molecular Biology, University of Texas Southwestern Medical Center at Dallas:, U. D. o. E. Office of Science, and The Wellcome Trust:. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb. 2001. ISSN 1476-4687. doi: 10.1038/35057062. URL https://www.nature.com/articles/35057062.

D. R. Leach, M. F. Krummel, and J. P. Allison. Enhancement of antitumor immunity by CTLA-4 blockade. *Science (New York, N.Y.)*, 271(5256):1734–1736, Mar. 1996. ISSN 0036-8075. doi: 10.1126/science.271.5256.1734.

H. Ledford, H. Else, and M. Warren. Cancer immunologists scoop medicine Nobel prize. *Nature*, 562(7725):20–21, Oct. 2018. doi: 10.1038/d41586-018-06751-0. Number: 7725 Publisher: Nature Publishing Group.

W. Lee, J. H. Lee, and S. Jun. Selective targeting of KRAS oncogenic alleles by CRISPR/Cas9 inhibits proliferation of cancer cells. *Scientific Reports*, 8(1):11879, Aug. 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-30205-2. Number: 1 Publisher: Nature Publishing Group.

G.-Y. Lyu, Y.-H. Yeh, and Y.-C. Yeh. Mutation load estimation model as a predictor of the response to cancer immunotherapy. *npj Genomic Medicine*, 3(1):1–9, Apr. 2018. ISSN 2056-7944. doi: 10.1038/s41525-018-0051-x. Number: 1 Publisher: Nature Publishing Group.

K. D. Makova and R. C. Hardison. The effects of chromatin organization on variation in mutation rates in the genome. *Nature Reviews Genetics*, 16(4):213–223, Apr. 2015. ISSN 1471-0064. doi: 10.1038/nrg3890. Number: 4 Publisher: Nature Publishing Group.

C. Mathers and D. Loncar. Projections of global mortality and burden of disease from 2002 to 2030, Nov. 2006. URL https://pubmed.ncbi.nlm.nih.gov/17132052/.

T. Michoel. Natural coordinate descent algorithm for L1-penalised regression in generalised linear models. *Computational Statistics & Data Analysis*, 97:60–70, May 2016. ISSN 0167-9473. doi: 10.1016/j.csda.2015.11.009.

G. Narzisi and M. C. Schatz. The Challenge of Small-Scale Repeats for Indel Discovery. *Frontiers in Bioengineering and Biotechnology*, 3, Jan. 2015. ISSN 2296-4185. doi: 10.3389/fbioe.2015.00008.

T. S. Nowicki, S. Hu-Lieskovan, and A. Ribas. Mechanisms of Resistance to PD-1 and PD-L1 blockade. *Cancer journal (Sudbury, Mass.)*, 24(1):47–53, 2018. ISSN 1528-9117. doi: 10.1097/PPO.0000000000000303. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5785093/.

M. Olivier, M. Hollstein, and P. Hainaut. TP53 Mutations in Human Cancers: Origins, Consequences, and Clinical Use. *Cold Spring Harbor Perspectives in Biology*, 2(1), Jan. 2010. ISSN 1943-0264. doi: 10.1101/cshperspect.a001008.

D. M. Pardoll. The blockade of immune checkpoints in cancer immunotherapy. *Nature reviews. Cancer*, 12(4):252–264, Mar. 2012. ISSN 1474-175X. doi: 10.1038/nrc3239.

S. S. Ramalingam, M. D. Hellmann, and M. M. Awad. Tumor mutational burden (TMB) as a biomarker for clinical benefit from dual immune checkpoint blockade with nivolumab (nivo) + ipilimumab (ipi) in first-line (1L) non-small cell lung cancer (NSCLC). *Cancer Research*, 78(13 Supplement):CT078–CT078, July 2018. ISSN 0008-5472, 1538-7445. doi: 10.1158/1538-7445.AM2018-CT078.

P. Razavi, B. T. Li, D. N. Brown, B. Jung, E. Hubbell, R. Shen, W. Abida, K. Juluru, I. De Bruijn, C. Hou, O. Venn, R. Lim, A. Anand, T. Maddala, S. Gnerre, R. Vijaya Satya, Q. Liu, L. Shen, N. Eattock, J. Yue, A. W. Blocker, M. Lee, A. Sehnert, H. Xu, M. P. Hall, A. Santiago-Zayas, W. F. Novotny, J. M. Isbell, V. W. Rusch, G. Plitas, A. S. Heerdt, M. Ladanyi, D. M. Hyman, D. R. Jones, M. Morrow, G. J. Riely, H. I. Scher, C. M. Rudin, M. E. Robson, L. A. Diaz, D. B. Solit, A. M. Aravanis, and J. S. Reis-Filho. High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants. *Nature Medicine*, 25(12):1928–1937, Dec. 2019. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-019-0652-7. URL http://www.nature.com/articles/s41591-019-0652-7.

C. Robert. A decade of immune-checkpoint inhibitors in cancer therapy. *Nature Communications*, 11(1):3801, July 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17670-y.

V. Roth and B. Fischer. The Group-Lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 848–855, New York, NY, USA, July 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390263.

F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12): 5463, Dec. 1977. doi: 10.1073/pnas.74.12.5463. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC431765/.

A. Sboner, X. J. Mu, and D. Greenbaum. The real cost of sequencing: higher than you think! *Genome Biology*, 12(8):125, Aug. 2011. ISSN 1474-760X. doi: 10.1186/gb-2011-12-8-125.

M. T. Schweizer, R. Gulati, M. Beightol, E. Q. Konnick, H. H. Cheng, N. Klemfuss, N. De Sarkar, E. Y. Yu, R. B. Montgomery, P. S. Nelson, and C. C. Pritchard. Clinical determinants for successful circulating tumor DNA analysis in prostate cancer. *The Prostate*, 79(7):701–708, May 2019. ISSN 0270-4137, 1097-0045. doi: 10.1002/pros.23778. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pros.23778.

A. Tefferi. JAK2 Mutations in Polycythemia Vera — Molecular Mechanisms and Clinical Applications. *New England Journal of Medicine*, 356(5):444–445, Feb. 2007. ISSN 0028-4793. doi: 10.1056/NEJMp068293. URL https://doi.org/10.1056/NEJMp068293.

S. Turajlic, K. Litchfield, and H. Xu. Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. *The Lancet. Oncology*, 18(8):1009–1021, 2017. ISSN 1474-5488. doi: 10.1016/S1470-2045(17)30516-8.

J. D. Watson and F. H. C. Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738, Apr. 1953. ISSN 1476-4687. doi: 10.1038/171737a0. URL https://www.nature.com/articles/171737a0.

I. B. Weinstein and K. Case. The History of Cancer Research: Introducing an AACR Centennial Series. *Cancer Research*, 68(17):6861–6862, Sept. 2008. ISSN 0008-5472, 1538-7445. doi: 10.1158/0008-5472.CAN-08-2827. URL https://cancerres.aacrjournals.org/content/68/17/6861.

J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10):1113–1120, Oct. 2013. ISSN 1546-1718. doi: 10.1038/ng.2764. URL https://www.nature.com/articles/ng.2764.

C. Wild, E. Weiderpass, and B. Stewart. *World Cancer Report: Cancer Research for Cancer Prevention*. 2020. ISBN 9789283204473 9789283204480. URL https://publications.iarc.fr/Non-Series-Publications/World-Cancer-Reports/World-Cancer-Report-Cancer-Research-For-Cancer-Prevention-2020.

C. Wittekind, M. K. Gospodarowicz, and J. D. Brierley. TNM Classification of Malignant Tumours, 8th Edition | Wiley, 2016. URL https://www.wiley.com/en-gb/TNM+Classification+of+Malignant+Tumours%2C+8th+Edition-p-9781119263579.

H.-X. Wu, Z.-X. Wang, and Q. Zhao. Designing gene panels for tumor mutational burden estimation: the need to shift from 'correlation' to 'accuracy'. *Journal for Immunotherapy of Cancer*, 7(1):206, 2019a. ISSN 2051-1426. doi: 10.1186/s40425-019-0681-2.

H.-X. Wu, Z.-X. Wang, and Q. Zhao. Tumor mutational and indel burden: a systematic pan-cancer evaluation as prognostic biomarkers. *Annals of Translational Medicine*, 7(22):640, Nov. 2019b. ISSN 2305-5847. doi: 10.21037/31486. Number: 22.

Y. Yang and H. Zou. A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, 25(6):1129–1141, Nov. 2015. ISSN 1573-1375. doi: 10.1007/s11222-014-9498-5.

Y. Yang, H. Zou, and S. Bhatnagar. gglasso: Group Lasso Penalized Learning Using a Unified BMD Algorithm, Mar. 2020. URL https://CRAN.R-project.org/package=gglasso.

L. Yao, Y. Fu, and M. Mohiyuddin. ecTMB: a robust method to estimate and classify tumor mutational burden. *Scientific Reports*, 10(1):1–10, Mar. 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-61575-1. Number: 1 Publisher: Nature Publishing Group.

A. D. Yates, P. Achuthan, and W. Akanni. Ensembl 2020. *Nucleic Acids Research*, 48(D1): D682–D688, Jan. 2020. ISSN 0305-1048. doi: 10.1093/nar/gkz966.

H. Zhao, P. Rai, L. Du, W. Buntine, D. Phung, and M. Zhou. Variational Autoencoders for Sparse and Overdispersed Discrete Data. In *International Conference on Artificial Intelligence and Statistics*, pages 1684–1694. PMLR, June 2020. URL http://proceedings.mlr.press/v108/zhao20c.html.

J. Zhu, T. Zhang, and J. Li. Association Between Tumor Mutation Burden (TMB) and Outcomes of Cancer Patients Treated With PD-1/PD-L1 Inhibitions: A Meta-Analysis. *Frontiers in Pharmacology*, 10, June 2019. ISSN 1663-9812. doi: 10.3389/fphar.2019.00673.

Appendix A

Supervised Dimension Reduction

This chapter might be something to do with random projection ensembles.