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. 27–29, 31

BMR Background Mutation Rate. 19, 22

ctDNA Circulating Tumour DNA. 19

CTLA-4 Cytotoxic T Lymphocyte Associated protein 4. 19

ecTMB Estimation and Classification of Tumour Mutation Burden. 28, 29

ICB Immune Checkpoint Blockade. 19

ICI Immune Checkpoint Inhibitor. 19

LASSO Least Absolute Shrinkage and Selection Operator. 20, 24, 25

MSI Micro-Satellite Instability. 15

PD-L1 Programmed Death Ligand 1. 19

PV Polycythemia Vera. 13

TCGA The Cancer Genome Atlas. 14

TIB Tumour Indel Burden. 19–22, 25, 29, 31

TMB Tumour Mutation Burden. 15, 19–22, 25, 27–29, 31, 32

WES Whole Exome Sequencing. 19

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

4 Clinical practice and pharmacoeconomics

4.1 Biomarkers as proxies of immunogenicity

Discuss TMB and MSI.

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

Chapter 2

The Generative/Predictive Blueprint for Biomarker Estimation

This chapter consists, in part, of discussion adapted from 'Dimensionality and Structure in Cancer Genomics: A Statistical Learning Perspective' (Bradley, 2020) (chapter 3 of 'Artificial Intelligence in Oncology Drug Discovery and Development' (Cassidy and Taylor, 2020)).

1 High-dimensional, sparse and bursty: generative models of mutation

2 Learning from learning: biomarker prediction

3 The double role of regularisation

Chapter 3

A Simple Example: Tumour Mutation Burden

1 Introduction

It has been understood for a long time that cancer, a disease occurring in many distinct tissues of the body and giving rise to a wide range of presentations, is initiated and driven by the accumulation of mutations in a subset of a person's cells (Boveri, 2008). In the last decade there has been an explosion of interest in cancer therapies targeting immune response, with immune checkpoint blockade therapy winning the Nobel Prize in Physiology or Medicine in 2018 (Ledford et al., 2018). Immune checkpoint inhibitors/blockades (ICI/ICBs) are antibodies that bind to proteins on the surface of immune cells. These surface proteins are known as immune checkpoints, the two most widely targeted being Cytotoxic T Lymphocyte Associated protein 4 (CTLA-4) and Programmed Death Ligand 1 (PD-L1) (Buchbinder and Desai, 2016). Inhibition of immune checkpoints with antibodies promotes a more aggressive immune response to cancer in some patients (Pardoll, 2012).

Exome-wide prognostic biomarkers for immunotherapy are becoming increasingly well-established, with Tumour Mutation Burden (TMB) (Zhu et al., 2019; Cao et al., 2019) in particularly being widely employed in clinical practice (Stenzinger et al., 2019). The TMB of a tumour is the total number of (non-synonymous) mutations occurring throughout the exome, and can be thought of as a proxy for how easily a tumour cell can be recognised as foreign by immune cells. However, the cost of measuring TMB using Whole Exome Sequencing (WES) (Sboner et al., 2011) currently prohibits its widespread use as standard-of-care, particularly given the small proportion of patients for whom ICB is beneficial in many cancer types (Nowicki et al., 2018). The cost of sequencing, both financial and in terms of the time taken for results to be returned, are particularly problematic in situations where high-depth sequencing is required, such as when utilising blood-based Circulating Tumour DNA (ctDNA) from liquid biopsy (Gandara et al., 2018). These issues are also encountered when measuring more recently proposed biomarkers such as Tumour Indel Burden (TIB) (Wu et al., 2019a; Turajlic et al., 2017) which only include mutations of strong immunogenicity. There is, therefore, a demand for cost-effective approaches for estimating these biomarkers (Fancello et al., 2019).

In this paper we propose a novel, data-driven method for biomarker estimation, based on a generative model of how mutations arise in the tumour exome. More precisely, we model 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, we use an l_1 regularisation penalty when estimating the model parameters, in order to reflect the fact that in many genes' mutations arise purely according to the BMR. In addition this identifies a set of genes that are mutated above or below the background rate, which are often of clinical interest. The flexibility of this model facilitates the construction of a new estimator of TMB, based on a weighted linear combination of the number of mutations in each gene. The vector of weights 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 excellent practical performance of our framework using non-small cell lung cancer dataset (Chalmers et al., 2017), including a comparison with the existing state-of-the-art data-driven approaches for estimating TMB, as well as estimators based on commercially available targeted gene panels.

Aside from its excellent performance at predicting TMB, our method has a number of secondary benefits. It can be adapted very easily to predict TIB, due to our separately modelling mutations of different variant type. For estimating TMB from a targeted gene panel, we are able to either select an optimal set of genes of a given total length, produce predictions from a pre-defined set of observed genes, or augment a pre-defined set of genes by a given amount of genomic space. Inferring background mutation rate by applying regularisation to our generative model rather than by comparing synonymous and non-synonymous mutation rates (as has been proposed elsewhere) allows us to take into account known non-selective effects on mutation rate, including chromatin configuration (Makova and Hardison, 2015) and transcription-dependent repair mechanisms (Fong et al., 2013). This also allows our method to be applied to datasets for which synonymous mutation counts are not available. The learned parameters of our model are interpretable and correspond to tangible biological effects and processes. We are also able to provide heuristic error bounds.

Due to the emergence of TMB as a biomarker in recent years, a variety of groups have considered methods for its estimation. 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 prediction made in this way are affected by the size of the targeted region, when mutations are assumed to occur at uniform rate throughout the genome. Yao et al. (2020) modelled mutations as following a negative binomial distribution while allowing for gene-dependent rates, and estimate TMB using maximum likelihood. 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. (2019b). Finally, we are unaware of any research dedicated to methods for efficiently estimating TIB, a more difficult task due to the low relative abundance of indel mutations and known reliability issues for sequencing genomic loci of highly repetitive nucleotide constitution (Narzisi and Schatz, 2015).

The remainder of the paper is as follows. After a brief discussion of our data sources and terminology in Section 2.1 we describe our generative model of mutation in Section 2.2, and show how regularised model fitting can allow us to infer a background mutation rate. In Section 2.3 we show how to produce a sparse linear estimator via a convex optimisation based on the group LASSO. We present experimental results of our procedures in Section 3, comparing our predictive performance to that of various state of the art statistical learning methods and commercially available gene panels. We conclude in Section 4 with some remarks about our method's utility in the context of a wide span of literature around biomarker estimation, and future avenues for research.

2 Methodology

2.1 Data and terminology

To demonstrate our proposal we make use of the non-small cell lung cancer dataset produced by Campbell et al. (2016), which contains data from 1144 patient-derived tumours. The patients had a variety of prognoses and were mixed in age, sex and smoking history (see Figure 3.1). For each sample we have the genomic locations and variant types of all mutations identified, as well as other covariates such as gender, age and smoking history. In Figure 3.2**A** 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 use nonsynonymous variants, which we group into seven types: the four most common are missense mutations, nonsense mutations, frameshift insertions/deletions and splice site mutations. The final three variant types, which are in-frame insertions/deletions, nonstop mutations and translation start site mutations, are

much rarer. We show the respective frequencies of the non-synonymous mutations described in Figure 3.2**B**. Frameshift insertion/deletion (also known as indel) mutations are of particular interest to us, but contribute only a small proportion ($< 4\%$) of mutations, which will have important consequences when trying to predict their abundance.

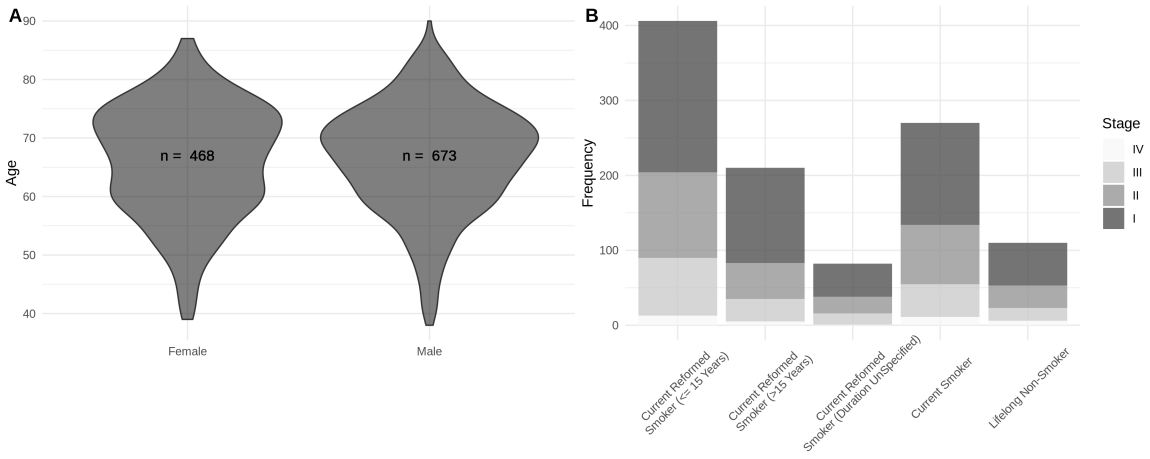

Figure 3.1: Demongraphic data for the clinical cohort in Campbell et al. (2016). **A**: Violin plots of age for patients, stratefied by sex. **B**: Stacked bar chart of patients' smoking histories. Shade corresponds to cancer stage diagnosis at time of measurement.

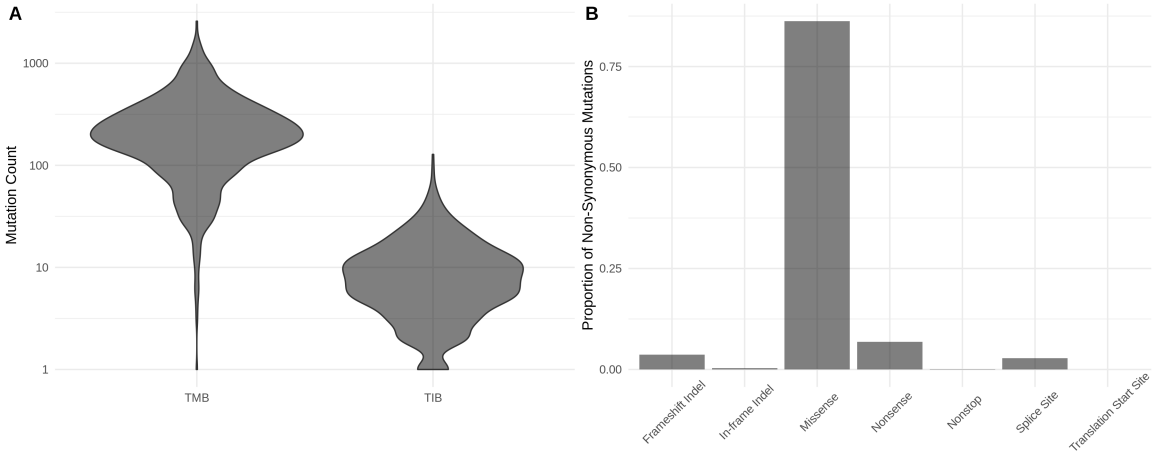

Figure 3.2: Dataset-wide distribution of mutations. **A**: Violin plot of the distribution of TMB and TIB across training samples. **B**: The relative frequency of different mutation types.

It is useful at this point to introduce the notation used throughout the paper. We let the set G denote a relevant collection of genes. For our purposes G will be considered to enumerate the entire 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 P is a subset $P \subseteq G$, and we may write $\ell_P := \sum_{g \in P} \ell_g$ for the total length of a panel P. Let S denote the possible variant types in our data (e.g. in the dataset mentioned above, S contains the seven possible non-synonymous variants). Now, for $i \in \{0, 1, \ldots, n\}$, let M_{igs} denote the count of mutations in gene $g \in G$ of type $s \in S$ in the ith sample. Here n refers to the number of samples in the training dataset, and the label $i = 0$ will be used to refer to an unseen test sample. In order to define the exome-wide biomarker of particular interest, we specify a subset of mutation types $\bar{S} \subset S$, and

[1]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.

let

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

For example, including all non-synonymous mutation types in $\bar{S}$ specifies $T_{i\bar{S}}$ as TMB for sample i, while letting $\bar{S}$ contain only indel mutations gives TIB. Our main goal is to predict $T_{0\bar{S}}$ after only observing $\{M_{0gs} : g \in P, s \in S\}$, 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 a test sample and a specific choice of biomarker, we will simply write T for $T_{0\bar{S}}$.

2.2 Generative model

We now describe the main statistical model that underpins our methodology. To account for selective pressures within the tumour and differences in transcription rate between different genes, we allow the rate at which mutations occur to depend on the gene and type of mutation. We also include a sample-dependent parameter, to account for differing levels of mutagenic exposure between tumours that may occur due to exogenous (e.g. UV light, cigarette smoke) or endogenous (e.g. inflammatory, free radical) factors. More precisely, for $i \in \{0, 1, \ldots, n\}$, $g \in G$ and $s \in S$, let $\phi_{igs} > 0$ denote the mutation rate, and suppose that

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

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, μ_i is a sample-dependent parameter which corresponds to the BMR of the sample. The offset $\log(\ell_g)$ accounts for a baseline 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. Other work has taken into account gene length in panel selection by rejecting genes beneath a given length threshold (for example in Lyu et al. (2018)). The parameters ν_s account for differences in frequency between mutation types.

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 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 so will experience mutations purely 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 (van de Geer, 2008). We do not penalise the parameters ν_s or μ_i, since we expect differences in mutation rate between variant types and between samples. More precisely, 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 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

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

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.

2.3 Proposed estimator

We now attend to our main goal of estimating a given exome-wide biomarker for an unseen sample. Fix $\bar{S} \subseteq S$ and write $T = T_{0\bar{S}}$ to simplify notation. Recall that 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 $w \in \mathbb{R}^{|G|\times|S|}$. We choose w to minimise the expected squared error of $T(w)$ based on the generative model above. Note that setting $w_{gs} = 1$ for $g \in G$ and $s \in \bar{S}$ and $w_{gs} = 0$ otherwise will give $T(w) = T$. However if, for a given g, we have $w_{gs} = 0$ for all $s \in S$, then our estimator does not depend on g and therefore the gene does not need to be included in the panel. In order to produce a concise targeted gene panel (i.e. with many $w_{gs} = 0$), we penalise non-zero components of w. In order to improve predictive performance by eliminating shrinkage bias we define our final estimator via a refitting procedure. This will also be helpful when extending our procedure to accommodate panel design with predetermined genes.

Now, by (3.2), we have that $\mathbb{E}M_{gs} = \mathrm{Var}(M_{gs}) = \phi_{0gs}$, and it follows that the marginal expected squared error of $T(w)$ is

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

This error depends on the unknown parameters μ_0, λ_g, ν_s and η_{gs}, the latter three of which are estimated as described in (3.4). To highlight the dependence on μ_0, write $\hat{\phi}_{0gs} = \ell_g \exp(\hat{\lambda}_g + \hat{\nu}_s + \hat{\eta}_{gs})$, and define

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

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 $K(\mu_0) = \exp(\mu_0)\sum_{g\in G}\sum_{s\in\bar{S}} l_g \exp(\hat{\lambda}_g + \hat{\nu}_s + \hat{\eta}_{gs})$. Notice that f is a rescaled version of the error $\mathbb{E}[\{T(w)-T\}^2]$, where the true parameters (λ, ν, η) have been replaced by the estimates $(\hat{\lambda}, \hat{\nu}, \hat{\eta})$. Moreover, f only depends on μ_0 via the $K(\mu_0)$ term. The quantity $K(\mu_0)$ 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 $g \in G, s \in \bar{S}$. As mentioned above, in order to form a concise panel while optimising predictive performance we solve this minimisation while constraining the cost of sequencing the genes used in the estimation. For a given w, an appropriate cost is given by

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

[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, not just those that directly constitute the biomarker being predicted. This is important when estimating the mutation types in $\bar{S}$ are relatively scarce (e.g. tumour indel burden).

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. Given a cost restriction L, our optimisation problem 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 of this problem. More precisely, let $\|w\|_{G,1} = \sum_{g \in G} \ell_g \|w_g\|_2$, where w_g here refers to the vector $(w_{g1}, \ldots, w_{g|S|})^T$ and $\|.\|_2$ is the Euclidean norm. Let[3]

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

where $\kappa_2 \geq 0$ is chosen to specify the level of cost penalisation, and will determine the size of panel selected.

Now, we could use $T(\hat{w}^{\text{first-fit}})$ as our estimator of T, and indeed we will include the performance of this approach in our numerical comparison in Section 3. However our final proposed estimator is motivated by the refitted LASSO procedure (see Chzhen et al., 2019), which in the case of standard linear regression can improve predictive performance by reducing bias. Indeed, let $\hat{P} := \{g \in G : \ \|\hat{w}_g^{\text{first-fit}}\|_2 > 0\}$ be the panel selected by our procedure above, and for $P \subseteq G$ let

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

Finally, then, our proposed estimator of T is given by $\hat{T} := T(\hat{w}^{\text{refit}})$, where

$$\hat{w}^{\text{refit}} \in \arg\min_{w \in W_{\hat{P}}} \big\{ f(w) \big\}. \tag{3.7}$$

2.4 Panel augmentation

When designing gene panels in practice, there are often a variety of factors contributing to the choice of which genes are 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). More precisely, given (disjoint) $P_0, Q_0 \subseteq G$ that have been separately identified for inclusion and exclusion from a panel respectively, we can replace $\hat{w}^{\text{first-fit}}$ in (3.6) with

$$\hat{w}^{\text{first-fit}}_{P_0,Q_0} \in \arg\min_{w \in W_{G \setminus Q_0}} \big\{ f(w) + \kappa_2 \sum_{g \in G \setminus P_0} l_g \|w_g\|_2 \big\}. \tag{3.8}$$

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

2.5 Practical considerations

Our first consideration here concerns the choice of the tuning parameter κ_1 in (3.4). We use 10-fold cross-validation to select κ_1, as is common for the LASSO estimator in generalised linear regression and for which there exists a rich theory and computational toolkit (Michoel, 2016; Friedman et al., 2020). To highlight one important aspect of our cross-validation procedure, recall that we consider the observations M_{igs} as independent across the sample index $i \in$

[3] We will assume that $\hat{w}^{\text{first-fit}}$, and the later defined $\hat{w}^{\text{refit}}$, are unique; see Section 2.5.

$\{1, \ldots, n\}$, the gene $g \in G$ and the mutation type $s \in S$. Our cross-validation method therefore involves splitting the entire set $\{(i, g, s) : i = 1, \ldots, n, g \in G, s \in S\}$ of size $n|G||S|$ (as opposed to simply the sample set $\{1, \ldots, n\}$) into 10 folds uniformly at random.

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. In particular, we recommend using $K(\mu_0) = K(\max_{i=1,\ldots,n}\{\hat{\mu}_i\})$, with $\hat{\mu}_i = \log(\hat{T}_i / \sum_{g,s} \ell_g \exp(\hat{\lambda}_g + \hat{\nu}_s + \hat{\eta}_{gs}))$, so that the penalisation is broadly in proportion with the values observed in the training dataset. The tuning parameter κ_2 controls the size of the gene panel selected in (3.6): for example, we set $\kappa_2 = \max\{\kappa : \ \ell_{\hat{P}} \leq L\}$ in order to produce a panel of length less than or equal to L.

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 below we used the `gglasso` R package (Yang et al., 2020). Note also that the solutions to (3.6) and (3.7) are unique; see Roth and Fischer (2008, Theorem 1).

Finally we describe a heuristic procedure for producing prediction interval around our estimates. In particular, for a given confidence level α, we aim to find an interval $[\hat{T}_{\mathrm{L}}, \hat{T}_{\mathrm{U}}]$ such that $\mathbb{P}\big(\hat{T}_{\mathrm{L}} \leq T \leq \hat{T}_{\mathrm{U}}\big) \geq 1 - \alpha$. 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, t_α is unknown since it depends on the mean squared error $\mathbb{E}\{(\hat{T} - T)^2\}$. For our refitted estimator, this is determined by (3.5) with $w = \hat{w}^{\mathrm{refit}}$. Here, the parameters λ, η, ν and μ_0 are unknown. Our heuristic approach is to plug-in the fitted estimates $\hat{\lambda}, \hat{\eta}, \hat{\nu}$ and estimate μ_0 with $\log(\hat{T} / \sum_{g,s} \ell_g \exp(\hat{\lambda}_g + \hat{\nu}_s + \hat{\eta}_{gs}))$. This is not, therefore, an exact $(1-\alpha)$-prediction interval for T; however, we will see in our experimental results in Sections 3.2 and 3.3 that in practice this approach does provide valid 90% prediction intervals. JB: Come back to this!.

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. We focus on prediction of TMB and TIB, due to their stature as established and recently proposed biomarkers respectively. We aim to show state-of-the-art performance for prediction of TMB, and to provide proof-of-principle results for estimating TIB, for which there is little prior art.

Since we are only looking to produce estimators for TMB and TIB, we group mutations in two categories so that $|S| = 2$. These are *a)* indel mutations and *b)* all other non-synonymous mutations. This has a simplifying effect, both in terms of the presentation of our results and the computational power required to fit the generative model. We split the dataset into training, validation and test sets containing $n = 800, n_{val} = 172$ and $n_{test} = 172$ samples respectively. We have mutation information for $|G| = 17681$ genes. For the samples forming our training set, the average TMB is 267 and the average TIB is 9.91.

In the following sections, we describe the fitted generative model and investigate the extent to its structure is reasonable, then separately address the performance of derived estimators of TMB and TIB. We conclude with a case study into augmenting the Foundation 1 panel to predict TMB and TIB.

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 of the cross-validation procedure are presented in Figure 3.3. The best choice of κ_1 produces estimates of λ and η with 40.2% and 76.2% 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 as mutating according to the background mutation rate, and genes with $\hat{\eta}_{\mathrm{g,indel}} = 0$ 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 parameter estimates that also have known biological relevance in oncology; see Section 4 for further discussion.

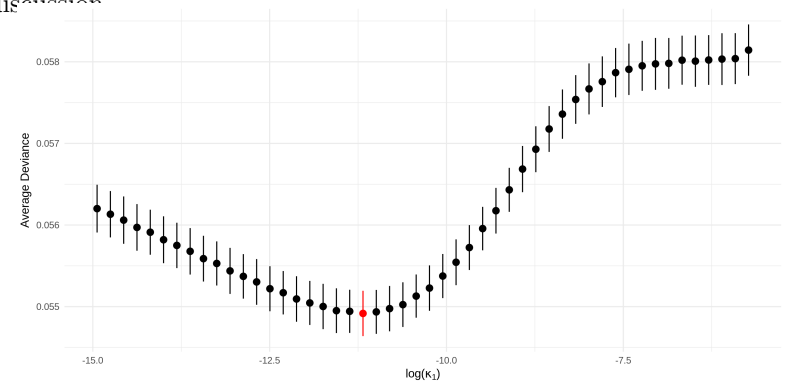

Figure 3.3: The average deviance (with one standard deviation) across the 10 folds in our cross-validation procedure plotted against $\log(\kappa_1)$. The minimum deviance is highlighted red.

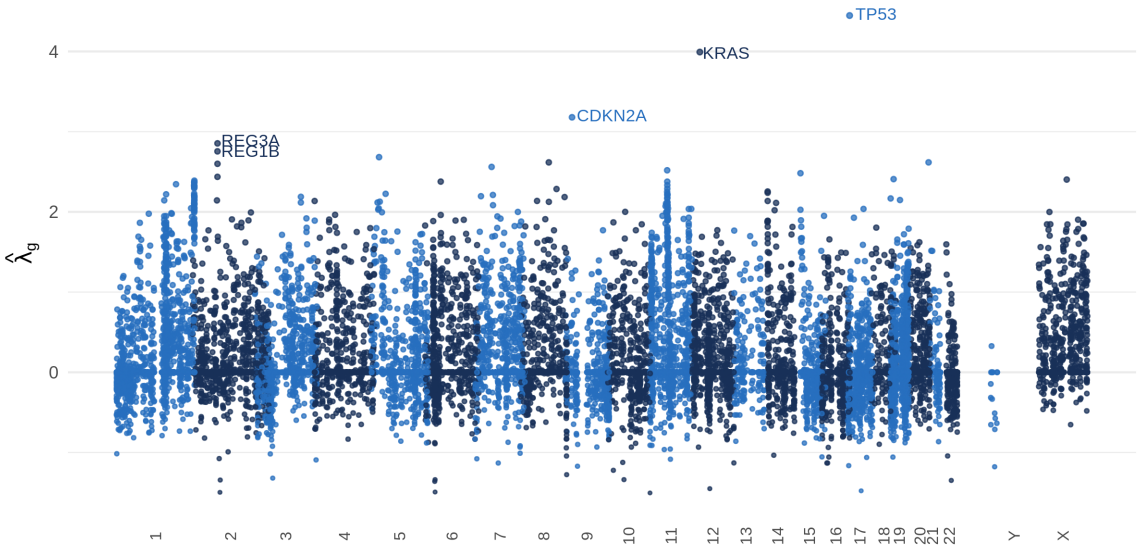

Figure 3.4: Manhattan plot of fitted parameters $\hat{\lambda}_g$ and their associated genes' chromosomal locations. The genes with the five largest positive parameter estimates are labelled.

In order to provide some validation of the model formulation in (3.3), we compare our model with the following alternatives:

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

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

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

(iv) No gene/mutation type interactions: $\eta_{gs} = 0$ 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. First note that the decrease in deviance between our model and model (i) is very small relative to the additional degrees of freedom associated with the saturated model; this suggests that our model is more appropriate. On the other hand, when comparing our model with the alternatives in (ii)-(iv), the decrease in deviance large relative to the increase in degrees of freedom, and our model is therefore preferred.

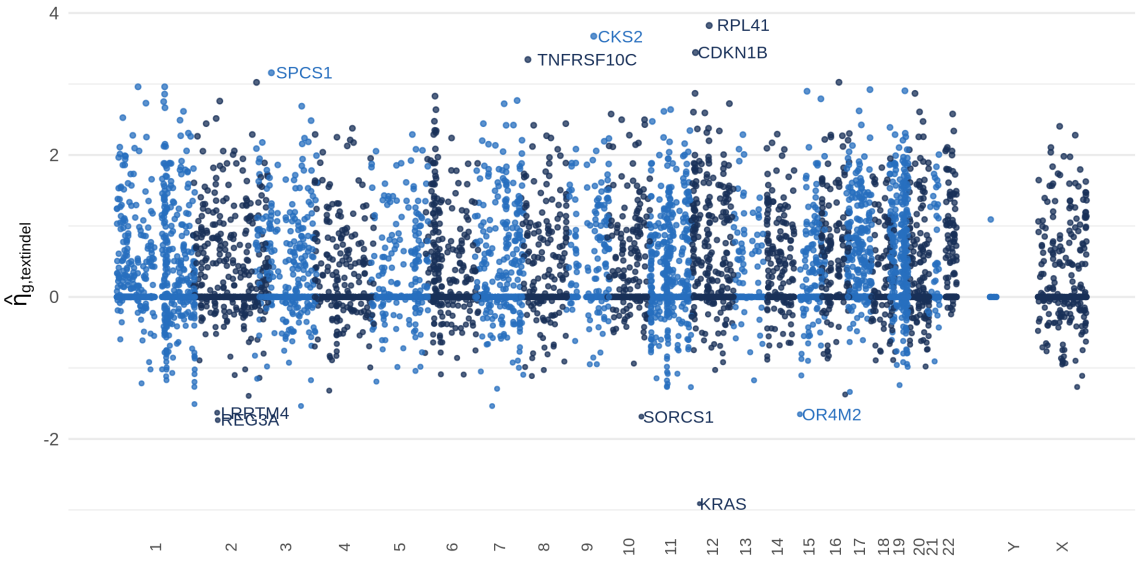

Figure 3.5: Manhattan plot of fitted parameters $\hat{\eta}_{g,\text{indel}}$ and their associated genes' chromosomal locations. The five largest positive and negative genes are labelled.

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)	7.50×10^5	2.80×10^7	2.68×10^{-2}	1.00
(ii)	1.61×10^5	8.00×10^2	2.01×10^2	0.00
(iii)	2.00×10^5	3.49×10^4	5.74×10^0	0.00
(iv)	1.09×10^5	1.75×10^4	1.09×10^1	0.00

3.2 Predicting tumour mutation burden

We now demonstrate the excellent practical performance of our procedure for estimating TMB. Recall that TMB is given by equation (3.1) with $\bar{S}$ being the set of all non-synonymous mutation types. 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. More precisely, given predictions of TMB, $\hat{t}_1, \ldots, \hat{t}_{n_{val}}$, for the observations in the validation set with true TMB scores $t_1, \ldots, t_{n_{val}}$. Let $\bar{t} := \frac{1}{n_{val}} \sum_{i=1}^{n_{val}} t_i$, and let

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

While it's not the focus of our approach, other works have been interested in separating tumours into two classes (High TMB, Low TMB) (Büttner et al., 2019; Wu et al., 2019b). We therefore also report the estimated area under the precision-recall curve (AUPRC) for a classifier based on our approach. The AUPRC is calculated as follows: first, in line with major clinical studies (see, for example, 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). This gives 53 (30.8%) tumours with high TMB and 119 (69.2%) with low TMB in the validation set. Now, for a cutoff $t \geq 0$, let

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

be the estimated (over the validation set) precision and recall, respectively, of a classifier that assigns a tumour to the High TMB class if its estimated TMB is greater than or equal to t. The precision-recall curve 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).

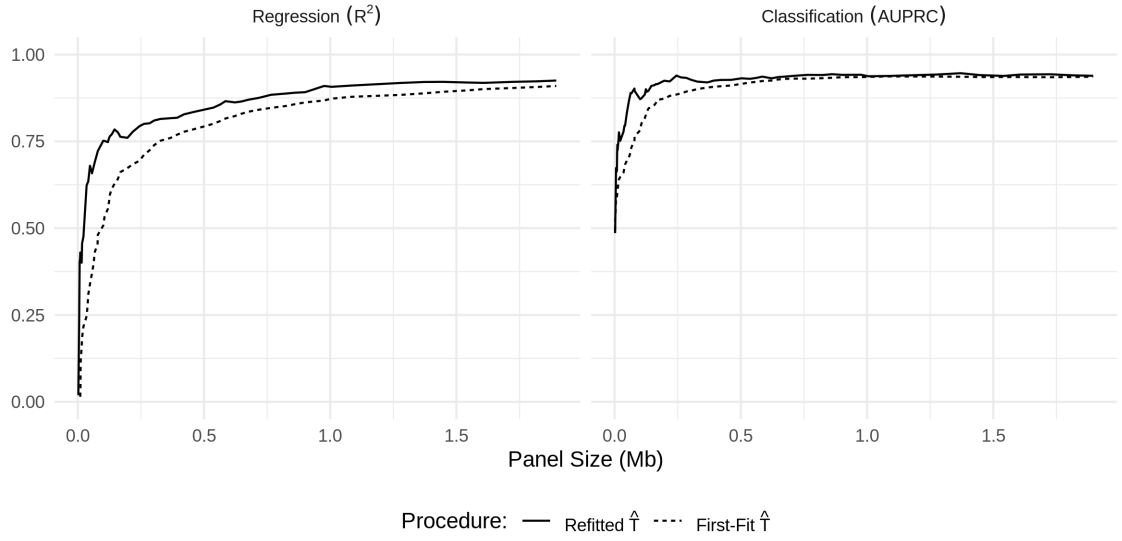

Figure 3.6: Performance of our first-fit and refitted estimators of TMB as the selected panel size varies. **Left**: R^2, **Right**: AUPRC.

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 (3.7), respectively) 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 0.25Mb). 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 four next generation sequencing panels, Foundation One (Frampton et al., 2013), MSK-IMPACT (Cheng et al., 2015), TST-170 (Heydt et al., 2018) and TSO-500 (Pestinger et al., 2020), are chosen for their relevance to TMB. 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 $\hat{w}_P \in \arg\min_{w \in W_P}\{f(w)\}$.

(ii) ecTMB: the procedure proposed by Yao et al. (2020).

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

(iv) A linear model: we estimate TMB with ordinary linear regression of TMB against $\{M_{0gs} : g \in P, s \in S\}$.

We include our refitted estimator applied to the panel P in order to demonstrate the utility of our approach, even with a prespecified panel. The second approach is a recent method, based on a negative binomial model. The third and fourth are simple and standard practical procedures for inferring TMB from targeted gene panels.

We present results of these comparisons in Figure 3.7. For each of the four panels considered here, we see that our refitted estimator applied to the panel outperforms all existing approaches in terms of regression performance, demonstrating that making predictions based on our generative model has tangible benefit. Furthermore, we see that by first selecting a panel based on

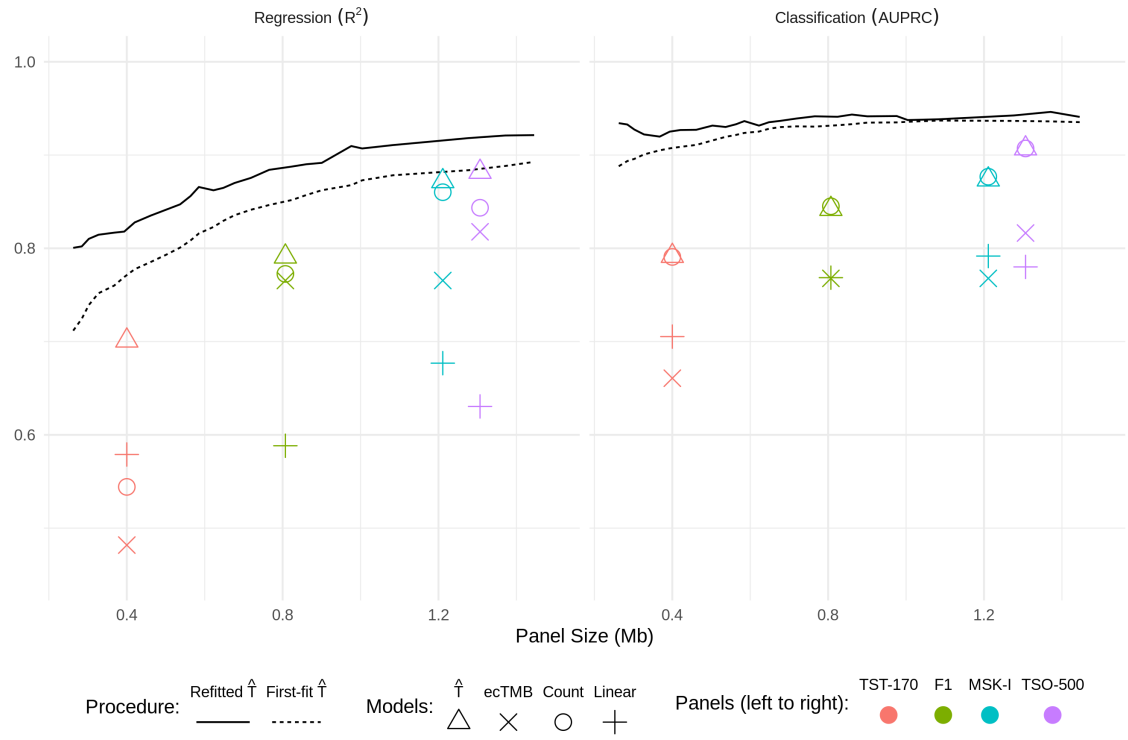

Figure 3.7: The performance of our TMB estimator in comparison to existing approaches. **Left**: R^2, **Right**: AUPRC.

the data, we are able to improve predictive accuracy even further. For instance, in comparison to predictions based on the Foundation 1 panel, our procedure with a selected panel of the same size achieves an R^2 of 0.88. The best available existing method based on the panel, in this case the count estimator, has an R^2 of 0.77. Similar results are seen for all panels included here. In terms of classification performance, we see that good predictions based on given panels are possible using the count estimator and our refitted estimator applied to the panel. However, we see that by selecting a panel our approach is more accurate. In particular, even with the smallest panel size presented in Figure 3.7, our method is competitive with those based on the large panels.

Finally in this section we demonstrate the practical performance of our method using the test set, which until this point has been reserved. Based on the validation results above, we take the panel of size 1Mb selected by our procedure and use our refitted estimator on that panel to predict TMB for the samples in the test set. For comparison, we also present predictions from ecTMB, count-based and linear regression estimators applied to the same panel. In Figure 3.8 we see that our procedure performs very well; we obtain an R^2 value (on the test data) 0.89. The other methods perform less well, with R^2 values of 0.72 (ecTMB), -19.2 (count) and 0.67 (linear). The count 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 plot the region for which our procedure for generating heuristic prediction intervals (as described in Section 2.5) is successful. More precisely, we shade all true TMB/predicted TMB pairs for which a 90% prediction interval generated around the estimate TMB value would include the true value. We find in this case that 160 (93.0%) of the points in the test set fall within this region, a good sign that our prediction intervals are performing as expected.

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

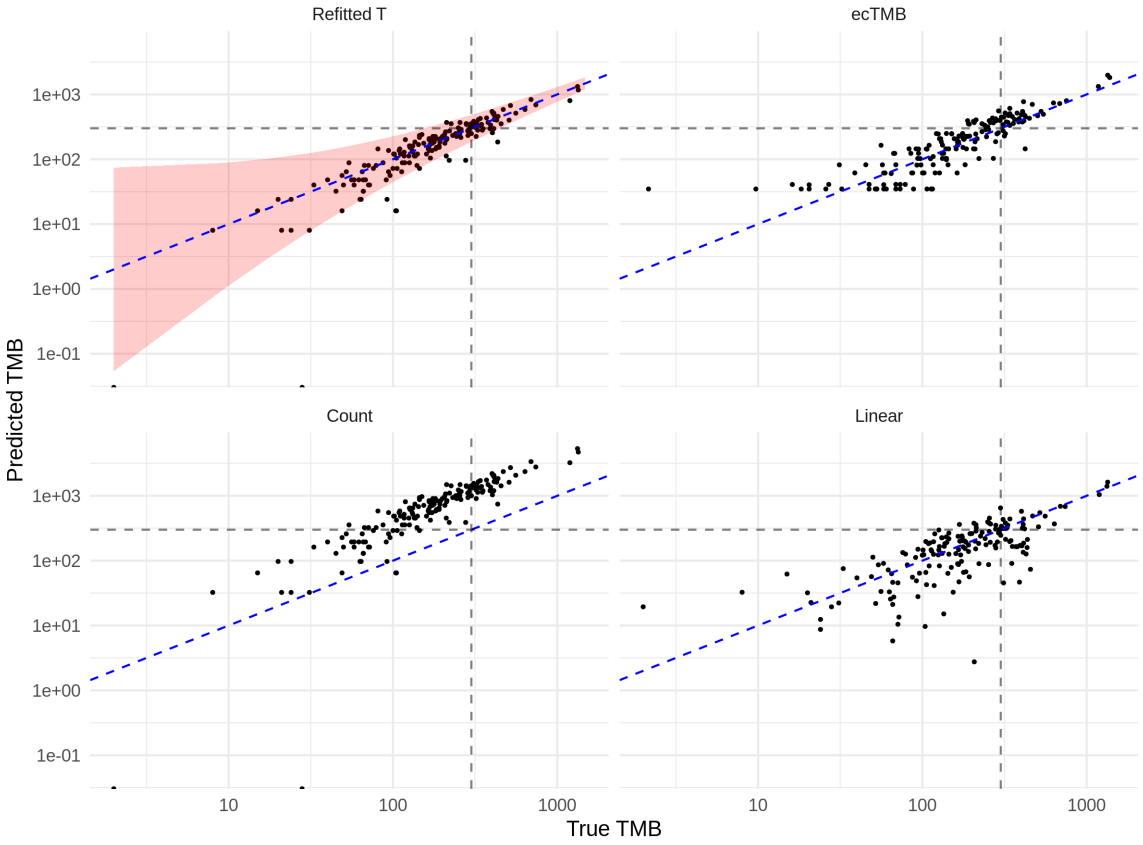

Figure 3.8: Prediction of TMB on the test dataset. Dashed blue (diagonal) line represents perfect prediction, and the black dashed lines indicate true and predicted TMB values of 300. Red shaded area delineates region for which 90% prediction intervals contain truth.

other variant types. 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 of panel sizes in order to ascertain the feasibility of targeted TIB estimation. Our results demonstrate that accurate classification of TIB status is still possible, even though estimating the precise value of TIB is difficult with small 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 (more than 11 indel mutations) and Low TIB (otherwise). This gives 43 (25.0%) tumours in the High TIB class in the validation dataset.

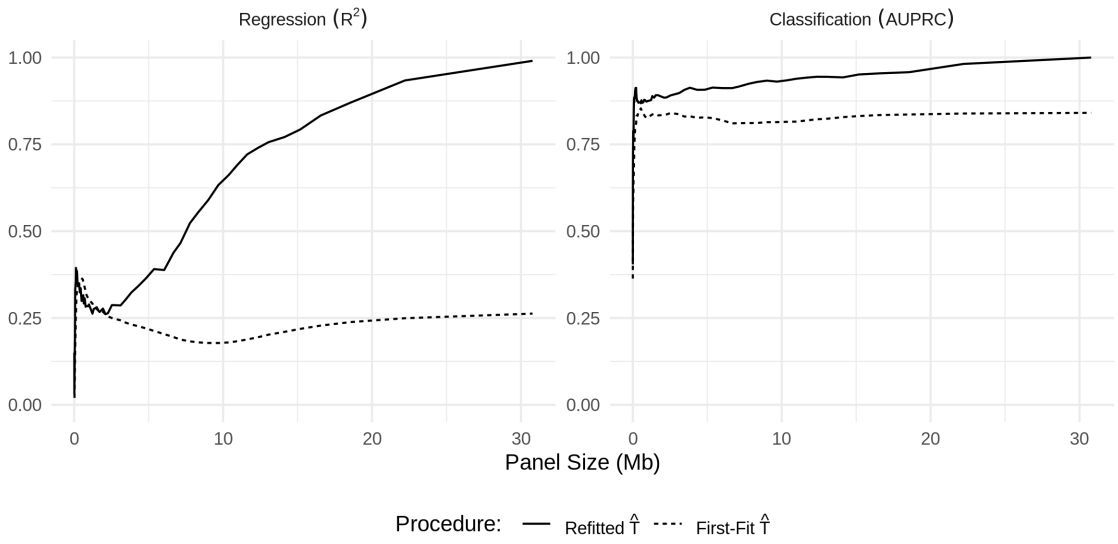

Figure 3.9: Performance of our first-fit and refitted estimators of TIB as the selected panel size varies. **Left**: R^2, **Right**: AUPRC.

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 in this case are much lower than what we achieved in estimating TMB. Perhaps surprisingly, the predictive performance of both estimators very slightly decreases as the panel size increases in the range 0.5-3.0 Mb; the refitted approach then improves for larger panel sizes. On the other hand, we see that the classification performance is impressive; our refitted method achieves an AUPRC of 0.87 even with panel of size of 1Mb.

As in the previous section, we now asses the performance on the test set of our refitted estimator of TIB applied to a selected panel of size 1Mb. In this case, we compare with a count estimator, which scales the total non-synonymous mutation count 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; i.e.

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

The results are presented in Figure 3.10. Our refitted $\hat{T}$ estimates give an R^2 value of 0.47 on the test set, while the count estimator again suffers from positive bias and has an R^2 value of -9.9. Again we include a 90% prediction interval success region. In this case we find that 166 (96.5%) of test set points fall within the success region.

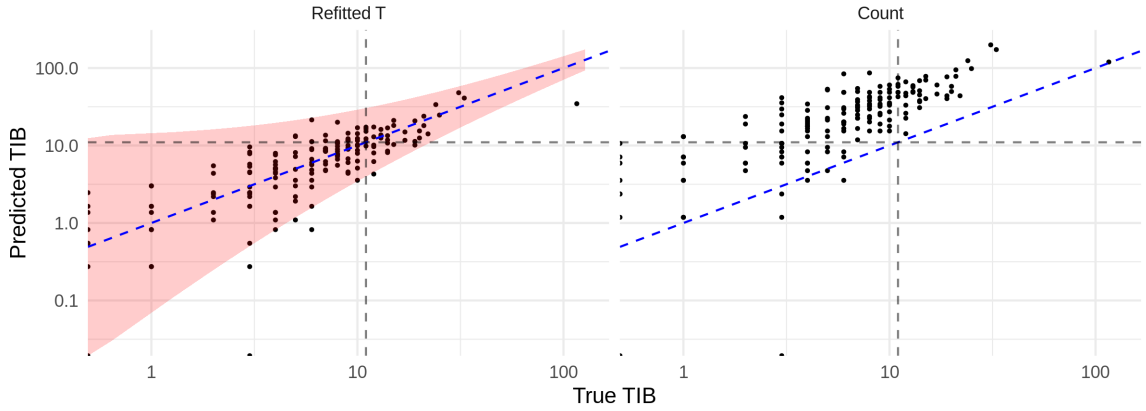

Figure 3.10: Prediction of TIB on the test dataset. Dashed blue (diagonal) line represents perfect prediction, and the grey dashed lines indicate true and predicted TIB values of 11. Red shaded area delineates region for which 90% prediction intervals contain truth.

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 a small set of genes in order to obtain better predictions of TMB. In this section we demonstrate how this can be done starting with the Foundation 1 panel. The Foundation 1 panel contains 324 genes (approximately 0.8Mb) - we will see that, by increasing the panel to 1Mb in size using our proposal, we can significantly improve the predictive performance.

Table 3.2: Predictive performance of models on Foundation 1 (0.8Mb) versus augmented Foundation 1 (1.0Mb) panels.

Model	Regression (R^2)		Classification (AUPRC)	
	Foundation	Augmented	Foundation	Augmented
Refitted $\hat{T}$	0.69	0.82	0.80	0.91
ecTMB	0.76	0.80	0.82	0.79
Count	0.61	-0.54	0.79	0.91
Linear	0.58	0.62	0.73	0.84

We apply the augmentation method described in Section 2.4, with P_0 taken to be Foundation 1 genes and set Q_0 to be empty. The genes added to the panel are determined by the first-fit estimator in equation (3.8). To evaluate the performance, we then apply the refitted estimator on the test dataset, after selecting the augmented panel of size 1Mb. For comparison, we apply our refitted estimator to the Foundation 1 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 the augmented panel produces enhanced predictive performance under the refitted $\hat{T}$ estimator in regression and classification. The refitted estimator provides better estimates than any other model on the augmented panel by both metrics. Notably, the ecTMB and count estimators suffer weakened results on the augmented panel (the count estimator at regression, ecTMB at classification). While augmenting a gene panel can produce enhanced results, it appears that this (perhaps unsurprisingly) may only be the case when employing an estimator designed in conjunction with the panel selection procedure.

4 Conclusions

We have proposed a new methodology for designing targeted gene panel which outperforms state-of-the-art procedures for the estimation of tumour mutation burden. We also show promise for estimating other biomarkers such as tumour indel burden, a problem for which no substantial

body of work currently exists. These biomarkers are useful for identifying patients who might benefit from immunotherapy, but as of yet techniques for their estimation that are efficient enough to be viable have not been shown to be sufficiently reliable to make the transition to the clinic and a key concern of practitioners remains the confidence they can expect to have in the predictions being made. By grounding our predictive architecture in a generative model of mutation we are able to provide justification for including genes in predictive panels that have previously been excluded, as well as useful practical bounds for the error of our predictions.

A further advantage of our method is the level of biological interpretation it affords. We were able through optimised sparsity regularisation of our generative model to identify a large subset of genes following approximately the same mutation rate. The notion of a background mutation rate has typically been used as a starting assumption for modelling work, rather than a quantified inference. We were also able to identify genes showing particularly strong positive and negative selection for indel mutations, which in at least one case corresponded directly to an observation at the center of ongoing work on potential cancer therapies.

We see the key contribution of our work not as the performance of our method on the dataset with which we have demonstrated it, but rather its flexibility and broad applicability. There continues to be a large amount of research into immunosurveilling targeted gene panels, with many individual proposal panels presented each year. We have not produced such a panel beyond as a demonstration, but instead have designed our workflow to be as applicable to as broad a range of situations as possible. Our methodology can be used by those with a predefined gene panel or by those designing a new panel, either from scratch or augmenting on a starting set of targets. It is flexible in the sense that it can be used to predict biomarkers based on any set of mutation types. This will hopefully allow for further investigation into the utility of tumour indel burden, as well as being robust to changing standards of what genes and mutation types to include when reporting the more established biomarker of tumour mutation burden.

Chapter 4

Causative Mechanisms of Genomic Instability

Chapter 5

Transcripts as Neoantigen Proxies: Expressed Mutation Burden

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.

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 by Theodor Boveri. 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. URL https://jcs.biologists.org/content/121/Supplement_1/1. 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.

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, K. Litchfield, E. Rempel, P. Christopoulos, D. Kazdal, V. Endris, M. Thomas, S. Fröhling, S. Peters, C. Swanton, P. Schirmacher, and A. Stenzinger. 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, F. López-Ríos, S. Merkelbach-Bruse, N. Normanno, E. Rouleau, and F. Penault-Llorca. 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. URL https://esmoopen.bmj.com/content/4/1/e000442.

J. D. Campbell, A. Alexandrov, J. Kim, J. Wala, A. H. Berger, C. S. Pedamallu, S. A. Shukla, G. Guo, A. N. Brooks, B. A. Murray, M. Imielinski, X. Hu, S. Ling, R. Akbani, M. Rosenberg, C. Cibulskis, A. Ramachandran, E. A. Collisson, D. J. Kwiatkowski, M. S. Lawrence, J. N. Weinstein, R. G. W. Verhaak, C. J. Wu, P. S. Hammerman, A. D. Cherniack, G. Getz, Cancer Genome Atlas Research Network, M. N. Artyomov, R. Schreiber, R. Govindan, and M. Meyerson. 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, X. Xu, T. Guo, and W. Ge. 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, D. Fabrizio, L. Gay, S. M. Ali, R. Ennis, A. Schrock, B. Campbell, A. Shlien, J. Chmielecki, F. Huang, Y. He, J. Sun, U. Tabori, M. Kennedy, D. S. Lieber, S. Roels, J. White, G. A. Otto, J. S. Ross, L. Garraway, V. A. Miller, P. J. Stephens, and G. M. Frampton. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Medicine*, 9, Apr. 2017. ISSN 1756-994X. doi: 10.1186/s13073-017-0424-2. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5395719/.

D. T. Cheng, T. N. Mitchell, A. Zehir, R. H. Shah, R. Benayed, A. Syed, R. Chandramohan, Z. Y. Liu, H. H. Won, S. N. Scott, A. R. Brannon, C. O'Reilly, J. Sadowska, J. Casanova, A. Yannes, J. F. Hechtman, J. Yao, W. Song, D. S. Ross, A. Oultache, S. Dogan, L. Borsu, M. Hameed, K. Nafa, M. E. Arcila, M. Ladanyi, and M. F. Berger. 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. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5808190/.

E. Chzhen, M. Hebiri, and J. Salmon. On Lasso refitting strategies. *Bernoulli*, 25(4A):3175–3200, Nov. 2019. ISSN 1350-7265. doi: 10.3150/18-BEJ1085. URL https://projecteuclid.org/euclid.bj/1568362056. Publisher: Bernoulli Society for Mathematical Statistics and Probability.

L. Fancello, S. Gandini, P. G. Pelicci, and L. Mazzarella. 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. URL https://doi.org/10.1186/s40425-019-0647-4.

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.

G. M. Frampton, A. Fichtenholtz, G. A. Otto, K. Wang, S. R. Downing, J. He, M. Schnall-Levin, J. White, E. M. Sanford, P. An, J. Sun, F. Juhn, K. Brennan, K. Iwanik, A. Maillet, J. Buell, E. White, M. Zhao, S. Balasubramanian, S. Terzic, T. Richards, V. Banning, L. Garcia, K. Mahoney, Z. Zwirko, A. Donahue, H. Beltran, J. M. Mosquera, M. A. Rubin, S. Dogan, C. V. Hedvat, M. F. Berger, L. Pusztai, M. Lechner, C. Boshoff, M. Jarosz, C. Vietz, A. Parker, V. A. Miller, J. S. Ross, J. Curran, M. T. Cronin, P. J. Stephens, D. Lipson, and R. Yelensky. 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, R. Tibshirani, B. Narasimhan, K. Tay, N. Simon, and J. Qian. glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models, June 2020. URL https://CRAN.R-project.org/package=glmnet.

D. R. Gandara, S. M. Paul, M. Kowanetz, E. Schleifman, W. Zou, Y. Li, A. Rittmeyer, L. Fehrenbacher, G. Otto, C. Malboeuf, D. S. Lieber, D. Lipson, J. Silterra, L. Amler, T. Riehl, C. A. Cummings, P. S. Hegde, A. Sandler, M. Ballinger, D. Fabrizio, T. Mok, and D. S. Shames. 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.

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, L. Jin, K. Song, R. Yu, T. Li, L. Qi, Y. Gu, W. Zhao, and Z. Guo. 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, A. Pluzanski, J. S. Lee, G. A. Otterson, C. Audigier-Valette, E. Minenza, H. Linardou, S. Burgers, P. Salman, H. Borghaei, S. S. Ramalingam, J. Brahmer, M. Reck, K. J. O'Byrne, W. J. Geese, G. Green, H. Chang, J. Szustakowski, P. Bhagavatheeswaran, D. Healey, Y. Fu, F. Nathan, and L. Paz-Ares. 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. URL https://doi.org/10.1056/NEJMoa1801946.

C. Heydt, R. Pappesch, K. Stecker, J. Neumann, R. Buettner, and S. Merkelbach-Bruse. 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. URL https://www.annalsofoncology.org/article/S0923-7534(19)32372-5/abstract.

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.

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.

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. URL https://www.nature.com/articles/d41586-018-06751-0. Number: 7725 Publisher: Nature Publishing Group.

G.-Y. Lyu, Y.-H. Yeh, Y.-C. Yeh, and Y.-C. Wang. 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. URL https://www.nature.com/articles/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. URL https://www.nature.com/articles/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. URL http://www.sciencedirect.com/science/article/pii/S0167947315002923.

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. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4306302/.

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/.

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. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4856023/.

V. Pestinger, M. Smith, T. Sillo, J. M. Findlay, J.-F. Laes, G. Martin, G. Middleton, P. Taniere, and A. D. Beggs. Use of an Integrated Pan-Cancer Oncology Enrichment Next-Generation Sequencing Assay to Measure Tumour Mutational Burden and Detect Clinically Actionable Variants. *Molecular Diagnosis & Therapy*, 24(3):339–349, 2020. ISSN 1177-1062. doi: 10.1007/s40291-020-00462-x. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7264086/.

S. S. Ramalingam, M. D. Hellmann, M. M. Awad, H. Borghaei, J. Gainor, J. Brahmer, D. R. Spigel, M. Reck, K. J. O'Byrne, L. Paz-Ares, K. Zerba, X. Li, W. J. Geese, G. Green, B. Lestini, J. D. Szustakowski, H. Chang, and N. Ready. Abstract CT078: 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): identification of TMB cutoff from CheckMate 568. *Cancer Research*, 78 (13 Supplement):CT078–CT078, July 2018. ISSN 0008-5472, 1538-7445. doi: 10.1158/1538-7445.AM2018-CT078. URL https://cancerres.aacrjournals.org/content/78/13_Supplement/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.

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. URL https://doi.org/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, D. Greenbaum, R. K. Auerbach, and M. B. Gerstein. 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. URL https://doi.org/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. Stenzinger, J. D. Allen, J. Maas, M. D. Stewart, D. M. Merino, M. M. Wempe, and M. Dietel. Tumor mutational burden standardization initiatives: Recommendations for consistent tumor mutational burden assessment in clinical samples to guide immunotherapy treatment decisions. *Genes, Chromosomes & Cancer*, 58(8):578–588, 2019. ISSN 1098-2264. doi: 10.1002/gcc.22733.

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, H. Xu, R. Rosenthal, N. McGranahan, J. L. Reading, Y. N. S. Wong, A. Rowan, N. Kanu, M. Al Bakir, T. Chambers, R. Salgado, P. Savas, S. Loi, N. J. Birkbak, L. Sansregret, M. Gore, J. Larkin, S. A. Quezada, and C. Swanton. 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.

S. A. van de Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36(2):614–645, Apr. 2008. ISSN 0090-5364, 2168-8966. doi: 10.1214/009053607000000929. URL https://projecteuclid.org/euclid.aos/1205420513. Publisher: Institute of Mathematical Statistics.

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, Q. Zhao, D.-L. Chen, M.-M. He, L.-P. Yang, Y.-N. Wang, Y. Jin, C. Ren, H.-Y. Luo, Z.-Q. Wang, and F. Wang. Tumor mutational and indel burden: a systematic pan-cancer evaluation as prognostic biomarkers. *Annals of Translational Medicine*, 7(22):640, Nov. 2019a. ISSN 2305-5847. doi: 10.21037/31486. URL http://atm.amegroups.com/article/view/31486. Number: 22.

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

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. URL https://doi.org/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, M. Mohiyuddin, and H. Y. K. Lam. 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. URL https://www.nature.com/articles/s41598-020-61575-1. Number: 1 Publisher: Nature Publishing Group.

J. Zhu, T. Zhang, J. Li, J. Lin, W. Liang, W. Huang, N. Wan, and J. Jiang. 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. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6587434/.

Appendix A

High-Dimensional Causal Inference

Appendix B

Supervised Dimension Reduction