Statistical and Machine Learning Approaches to Genomic Medicine

THE UNIVERSITY *of* EDINBURGH

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 thesis, we develop new statistical and machine learning methods for genomic medicine, and apply them to problems in diagnostics and precision oncology. Our overall aim is to introduce techniques that inform practical decision making in the design and use of clinical tests. The work combines domain-specific context with modern advances in Bayesian hierarchical modelling, high-dimensional statistics, and causal inference.

We begin in Chapter 1 with an introduction to the concepts and methodologies that are common throughout the thesis. This includes the necessary context from molecular biology, an overview of genomics in medicine with a particular focus on cancer (the subject of Chapters 3 and 4), and a description of data-generating technologies such as DNA sequencing and gene expression profiling. We also provide an in-depth introduction to the relevant statistical learning methods and techniques. This sets the scene for the three projects presented in subsequent chapters.

In Chapter 2 we analyse the resolution of the loop-mediated isothermal amplification (LAMP) assay. LAMP is a technology that can be used in medical tests that require quantifying the presence of RNA for each of a set of gene targets. Motivated by the unmet need for statistically principled methods for guided LAMP optimisation, we show how to use data from clinical and synthetic samples to improve the resolution of a LAMP-based diagnostic test for sepsis patients. In this context, by optimisation of the assay we refer both to the selection of gene targets, and to the tuning of reactions conditions and selection of optimal primers to produce robust, high-resolution measurements of gene expression. Our analysis identifies novel quantities associated with primer design that may drive assay performance.

Chapter 3 focuses on designing gene panels to estimate tumour mutation burden (TMB) and other exome-wide biomarkers, which are used to determine which cancer patients will benefit from immunotherapy. The cost of whole-exome sequencing presently limits the widespread use of such biomarkers. In this chapter, we introduce a data-driven framework for the design of targeted gene panels for estimating a broad class of biomarkers including tumour mutation burden and tumour indel burden. The first goal is to develop a generative model for the profile of mutation across the exome, which allows for gene- and variant type-dependent mutation rates. Based on this model, we then propose a procedure for constructing biomarker estimators. Our approach allows the practitioner to select a targeted gene panel of prespecified size and construct an estimator that only depends on the selected genes. Alternatively, our method may be applied

to make predictions based on an existing gene panel, or to augment a gene panel to a given size. We demonstrate the excellent performance of our proposal using data from three non-small cell lung cancer studies, as well as data from six other cancer types.

In Chapter 4, we consider causal questions in survival analysis, and investigate the extent to which the heterogeneous treatment effects of immunotherapy vary according to patients' clinical and genomic features. Methods for identifying heterogeneous treatment effects from survival data are still in their infancy, and so in this chapter we benchmark some recently proposed strategies. In particular, we show that high-throughput targeted sequencing data may offer better understanding into which patients are likely to benefit from immunotherapy, using state-of-the art statistical learning methods based on causal survival forests and regularisation.

Lay Summary

Of the many diseases that profoundly impact our society, almost all are influenced by factors related to our genes. This can be through the genetic information we inherit from our parents (our *genome*), the changes that happen to our genome through the course of our life, or the way our genes are used throughout our body. In this century, we've seen incredible advancements in our ability to describe and measure our genomes. Despite this, we feel like we know very little about how our genomic data relates to our experience of disease compared to how much there is to find out. Even when we do have this knowledge, it can be really hard to turn this into technologies and practices that are useful for preventing and treating disease. Why is that?

There are a few reasons. One of the biggest is that our genome is massive. It consists of 20,000 genes, each comprising a sequence of molecular 'letters', dispersed throughout an overall genetic sequence around two billion characters long. It has become a lot easier, and crucially a lot cheaper – think $100,000,000 in the year 2000, less than $1,000 now – to sequence (read) a genome in its entirety for any given person. This practical advantage, however, masks a scientific nightmare. Human DNA contains more letters than there are people who've ever had their genome sequenced. With that many letters, think how many combinations of letters – how many unique possible humans – there are, and the complexity of trying to predict anything about a human from their genome alone starts to become obvious. In statistics, we call data like this *high-dimensional*, as a fancy way of saying "there's a lot of it, in fact more pieces of information for each person we look at than there are people available to study".

The 'curse of dimensionality' described above has arisen from a practical advantage (lots of data from each sample) creating a theoretical challenge (lots of data from each sample), but other challenges are themselves intrinsically practical. While genetic sequencing is now unbelievably cheap compared to two decades ago, it is not free, and for many applications sequencing the entire genome is not necessary or efficient. Additionally, clinical decision-making often has a strong time dependency: a cheap test that takes two weeks to run might just not cut it. We therefore need to think not just about how much information genomic data *can* give us in medicine, but how best to deploy the technologies that we have to make practical impact now.

A final difficulty in the translation of genomic data into medical benefit is both theoretical and practical, and concerns how data is generated and how we combine data from disparate sources. When asking a question like "can someone's genome help up decide whether to treat them with drug X?", we'll often need to compare

genomic data from patients who were treated with drug X with data from patients who weren't. These may have come from different studies, conducted by different organisations, and targeting different populations. To start to untangle all of this, we need a theoretical language to describe the interaction of all of these effects – the language of *causal inference*. In order to develop tools that are actually useful for clinicians, we'll need a lot of practical insight into the situations in which clinical and treatment decisions are being made.

Having discussed a few of the theoretical and practical problems we'll be trying to address, we'll now describe the specific use cases. In our first application study (Chapter 2), we look at trying to understand how to measure gene *expression* (how genes are used in the body) accurately, with a technology that's more difficult to get right than most but that has the highest chance of actually being employed in clinical settings. This technology, LAMP, has its advantage over other (less fiddly) gene expression measurement technologies in that it doesn't require large, expensive and slow bits of machinery, and can give results within 30 minutes. This is crucial for our application space: we're attempting to use this technology to predict whether a patient in emergency care is suffering from a bacterial or viral infection. Making this distinction is crucial in determining whether to treat with antibiotics – every hour of delay in deciding to give antibiotics to a patient suffering bacterial-driven sepsis can lead to an 8% increase in risk of death from the infection.

In the second half of this thesis, we're concerned with cancer. Cancer is a natural fit for techniques in genomic medicine as *cancer itself is a disease of the genome*. This gives rise in part to cancer's extraordinary diversity: more so than any other disease, every cancer is unique. Indeed, two patients' tumours may be caused by different things, occur in different places, and have very different effects. In Chapters 3 and 4 we address one of the practical challenges described above in the context of cancer. Namely, if a tumour is (to some extent at least) described by the portfolio of mutations in DNA sequence that it has accumulated, do we really need to know every single change in that sequence in order to sensibly guide therapeutic choices? We take two approaches. Firstly, we try and predict a tumour feature known to be associated with improved response to immunotherapy (a type of cancer drug) using as little genomic information as possible. Remember, using little genomic information hopefully means cheaper! Next, we start with a prespecified collection of genes and attempt to predict the impact of treating a given patient with immunotherapy purely on the basis of the genetic information contained in those genes.

This entire work is bound together by a few principles:

1. Genomic data contains information that is relevant for treating disease;
2. Modern techniques in data science have the capacity to unlock that data;
3. These techniques have to be applied in a way that leads to *practical benefit*.

I hope that you enjoy it!

Foreword

The work contained in this thesis constitutes the result of many partnerships, each of which I have been lucky to have been involved in. These have enabled me to continue working (to some extent) fruitfully and (to a greater extent) happily since starting my PhD at Edinburgh in 2019. In many cases, however, their roots were planted before that, and in even more cases they will hopefully continue to grow long after everyone who will ever read this document has done so[1].

To begin with some self-indulgently biographical specifics, I would not be where I am today were it not for being randomly assigned to work in the Elowitz Laboratory at Caltech on a summer placement in 2017. At that point I had a passing interest in biology, passing enough to pass over (for example) choosing to be educated in it in any way beyond what was legally required. Despite this, somehow Michael Elowitz, Yaron Antebi, and Christina Su were persuaded to babysit me in a professional laboratory for three months, and were in the process responsible for the shape of my life ever since. While I've moved into a different field of biology, I wouldn't have been on the right farm were it not for their inspiration.

Academically, the next leap for me was being encouraged to take on a Master's in Systems Biology by the likes of Anindya Sharma and my own lack of enthusiasm to join the workforce. Still a pretty clueless maths student then as much as now, I owe several people a great deal of thanks for surviving that year, including Pavel Artemov, Stephen Cole, Sean Jones, Dan Mirea, Carolina Monck, Hannah Munby, and Steve Russell.

While this was all going on, some new folks turned up, and they worked for a company called Cambridge Cancer Genomics. Shortly thereafter I also worked (sometimes) for a company called Cambridge Cancer Genomics, and this is where I first really got a sense that what I was doing there I'd be doing for quite a bit of the rest of my life. From Hannah Thompson and Belle Taylor's welcome and support (continuing to this day), to Nirmesh Patel and Harry Clifford's trusting, supportive, and empowering supervision, to Henry Farmery's lessons on the significance of unit testing, to Dobby, Dami, and Geoff's reminders that if I had a long way to go academically I had further to go as a Mario Kart driver, I learned as much during my time there as I've learned at any other time since. Oh, and they offered to fund my PhD.

All good things must come to an end, and while my friendships from CCG have certainly not ended, unfortunately the company did. At this point, I was in

[1]Current estimates inform me that I should expect this lucky few to number between zero and three (inclusive).

the market for some new (academic) partners. Luckily, a measured and thoughtful American called Michael Mayhew wandered across my mid-pandemic Zoom window and started talking to me about whacko Bayesian stuff. It turned out that he had a gaggle of his own (hello Diego Borges, Ljubomir Buturovic, Mafalda Cavaleiro, Arthur Radley, Sara Masarone, Amitesh Pratap, Melissa Remmel, and Yuan Yuan), and what was intended to be a few months of summer fun and a chance to get away from my PhD thesis became (at the time of writing, and counting in both respects) two and a bit years of fun and a pretty substantial portion of my PhD thesis. The gentle mentorship and flexibility shown by all of these folks, and especially Michael, was truly transformative for me as a researcher and person.

Getting towards the end of this unrequested (and please God unrequited) academic synopsis, I need to get to the people who've been supportive throughout the years that I've actually been doing my PhD (rather than swanning off for months at a time without explanation or, arguably, justification). Firstly the Usual Suspects: Andrew Beckett, Jamie Burke, Bella Deutsch, Linden Disney-Hogg, Jon Eugster, Josh Fogg, Augi Jacovskis (got your surname right first try), Mary Llewellyn (not so lucky). You are amongst you: my longest and most co-dependent flatmate, my principal collaborator in projects that have borne no revelance to my PhD, my favourite, and five other of my friends. I will allow you to fight it out between you who is who. Next my research group, which has grown infuriatingly from just me to me and some other goons: Louis Chislett, Aris Sionakidis, Torben Sell, Wenxing Zhou. I have never ceased resenting you all for diluting my attention, but at least one of you has a boat, so I'm willing to call it quits.

I'd thought that I might get out of mentioning my family by keeping these acknowledgements fairly work-focused. Unfortunately this darn pandemic happened, and in doing the right and proper middle-class thing by running home to mummy and daddy in the countryside, I've allowed them and my (overnumerous) siblings an unfortunate back door into my professional life. I'm therefore obliged to say: thanks Mum, thanks Dad, I couldn't have done it without you. Ruth, Sarah, Luke: I may well have been able to do it without you, but it would have been substantially less fun. You're all great.

I never mention it, but I spent a bit of time in London during the later stages of my PhD. Those who I worked with in the summer of 2022 are well-loved by myself, but unfortunately none of that stuff made the thesis, so you'll have to remain satisfied by being well-loved anonymously. If you want me to say something nice about you, I'll put it in a WhatsApp. The second half of my stay in London, at the Alan Turing Institute don't you know, brought with it many welcome new faces including Lukas Franken, Susana Garcia, and Ezra Webb. It also brought back some very welcome old ones including Seb Hickman and Sara Masarone (again), which was an absolute treat.

Eventually I made it back most of the way to Edinburgh, stopping off in the village of Auchendinny, Midlothian, much to the bemusement of my supervisor Tim. Who, pretty much alone in the western world, hasn't been mentioned so far. Grab the hanky Tim, this bit's for you!

At various points throughout my PhD, external factors and/or internal fid-

getiness have prompted me to up sticks and change direction at short notice. These changes of direction have included change of academic direction, change of city and change of time zones[2]. At each of these points, any reasonable supervisor would have been well within his rights to say "Okay, but we spent quite a while doing thing A", "Okay, but it would be nice to be in the same city", "Okay, but how about you work on this set of problems that are more relevant to my interests and expertise?", "Okay, but with two months to go do you really think now's the time?". To say that Tim's shown a fair bit of grace in allowing me to pursue my interests, needs, and fancies, would be doing it down just a touch. I have never throughout all of my forgetting and/or turning up late to meetings, redefining deadlines, bad jokes, scrapes with international law, exposure to potential defunding, or any more of an indefinite list of things, felt that Tim was anywhere other than firmly and encouragingly behind me. I hope it's not going to my head to say: I was Tim's first student, he was my first supervisor, and we just about worked it out as we went along together. From our first meeting in Edinburgh when he told me that it would never be my job to try and impress him[3], to socially distant tins of cider outside the university library after the first quelling of lockdown, to a barbecue at sea in our last week as supervisor and supervisee, I owe Tim the immense privilege it's been to have had my life for the last four years.

On that triumphant note (and contradicting myself) I can't help but list some more of the people who've touched my life in the last four years, perhaps not academically, but who've made it an absolute pleasure to be around and about and in so doing have made the world of difference: Lydie T, Naomi CS, Michael C, Esme B, Bailey B, Beth B, Kathryn W, Sula C, Elinor CA, Rowan ~~B~~H, Paul NJ, Hannah W, Vessela I, Conor C, Rob J, Eve D, Tanmay S, Freddie B, David N, Lizzie M, Lizzie J, George P, Mark A, Benji T, Daisy T, Molly S, Tommy S, Harrison F, Clem B, Lottie P, Tom D, Ollie S, Harry J, Clara D, Sophie O, Shereen S, Kai W, Hannah K, Frank D, Heather Y, Alex C, Alex M, Hannah S, Jo W, James F, Rosa H, James T, Cara W, Alex S, Lawrence (Barry) S, André V, Hawo A, Dilini K, Millie Z, Antonios PD, Jack E, Jen M, John M.

It's bizarre after all that to end with a plug, but so I shall. The work in this thesis comes in part from a variety of publications and pre-publications. Chapter 1 is based very loosely on the book chapter: "Dimensionality and structure in cancer genomics: a statistical learning perspective" (Bradley, 2020). Chapter 2 is adapted from "Hierarchical Bayesian modeling identifies key considerations in the design of loop-mediated isothermal amplification assays" (Bradley *et al.*, 2023), while Chapter 3 is adapted from "Data-driven design of targeted gene panels for estimating immunotherapy biomarkers" (Bradley and Cannings, 2021a, 2022). Chapter 4 has no associated literature, as it was being written up until the week of submission.

[2]Bizarrely, the latter and the former did not overlap.

[3]There at least I may have succeeded as a student.

Acronyms

ANN artificial neural network 25

ATE average treatment effect 44, 45, 48, 107, 127

AUPRC area under precision-recall curve 89–91, 93–95

BMR background mutation rate 78, 81, 82, 86, 100, 122, 125

CATE conditional average treatment effect 108–110, 112, 114, 116

cDNA complementary DNA 33

CSA counterfactual survival analysis 120, 127

CSF causal survival forest 118, 119, 124, 127

ctDNA circulating tumour DNA 78

CTLA-4 cytotoxic T-lymphocyte associated protein 4 77, 93, 120, 121

DNA deoxyribonucleic acid 22, 25–27, 29–35, 50–52

ecTMB estimation and classification of tumour mutation burden 90, 91, 96

gDNA genomic DNA 33, 75

HMC Hamiltonian Monte Carlo 39, 60, 62, 155

HTE heterogeneous treatment effect 102, 106–108, 118, 119, 121

ICB immune checkpoint blockade 77, 78, 102, 120, 121, 123

IPCW inverse probability of censoring weighting 119

IPM integral probability metric 117, 120, 123

IPW inverse-probability weighting 108–112, 114, 118

IQR interquartile range 64

IR image recognition 24

IVT *in vitro*-transcribed 55–57, 60, 61, 71–75

LAMP loop-mediated isothermal amplification 33, 35, 46, 47, 50–59, 62, 63, 68, 71, 73–75, 150

LASSO least absolute shrinkage and selection operator 42, 47, 50, 85

LOO-PSIS leave-one-out Pareto smoothed importance sampling 67, 68

MAF mutation annotated format 33, 47, 165, 166

MCMC Markov chain Monte Carlo 39

MLE maximum likelihood estimation 38, 39

mRNA messenger RNA 26, 27, 51, 52, 55, 56, 73

MSKCC the memorial Sloan Kettering cancer centre 121

NLP natural language processing 24

NS NanoString 55–57, 64–66

NSCLC non-small cell lung cancer 77–79, 92, 98, 99, 121

OLS ordinary least squares 40, 42

OS overall survival 101, 102, 121

PCR polymerase chain reaction 27, 33–35, 46, 52, 54, 57, 74

PD-L1 programmed death ligand 1 29, 77, 93, 120, 121

PFS progression-free survival 101

PGM probabilistic graphical model 59, 61

PKU phenylketonuria 28

PV polycythaemia vera 30

qPCR quantitative polymerase chain reaction 34, 35, 55, 57, 65

qRT-LAMP quantitative reverse-transcription loop-mediated isothermal amplification 52–57, 59, 60, 63–65, 73–75

RCT randomised controlled trial 44, 102

RdRp RNA-dependent RNA polymerase 26

RMST restricted mean survival time 119

RNA ribonucleic acid 25–27, 29, 33, 34, 46, 51, 52, 54–57, 59–61, 64–66, 71–75

ROC receiver operating characteristic 93–95

rRNA ribosomal RNA 27

SARS-CoV-2 severe acute respiratory syndrome coronavirus 2 27

SDR sufficient dimension reduction 40, 41

SVM support vector machine 25

TCGA The Cancer Genome Atlas 31, 47

TIB tumour indel burden 48, 77–82, 86, 94–96, 99, 167

TMB tumour mutation burden 48, 50, 77–81, 86, 89–94, 96–102, 120–125, 167

UV ultraviolet 32, 81

VCF variant called format 33, 47

WES whole-exome sequencing 33, 78, 79, 101, 166

WGS whole-genome sequencing 22, 24, 33

Contents

Chapter 1

Introduction

Overview

> "Maybe the universe isn't total chaos."
>
> Annie Landsberg
> Maniac, S1 E10: 'Option C'
> (2018)

Advances in technology in recent decades have enabled both the generation and computational analysis of genomic data on unprecedented scales, transforming research and clinical practice throughout medicine in the process. This effect has been especially potent in fields with a strong molecular biology component, such as oncology. While cheap abundant biological data has unlocked invaluable new insights, it has demanded complementary advances in statistical and machine learning in order to answer theoretical and practical questions. The modern researcher has access to a catalogue of tools, including methods inspired by progress in AI from disciplines such as natural language processing and image analysis. However, before employing the latest and greatest technique from predictive or generative modelling, it is worth asking a sequence of questions. What sort of data are we dealing with in genomic medicine? Given the abundance of data available for a particular problem, what level of model complexity is appropriate? If our methods do work, will we understand why? What stage of the clinical pipeline is being targeted? If our hope is for the deployment of predictive systems in clinical settings, are our tools robust enough? If so, are the technologies upon which they rely economically viable? While we will certainly not answer all of these questions, in the chapters that follow we will provide some language with which to discuss them, and a range of application cases across genomic medicine.

This introduction should be equally approachable to those with a background in statistics/machine learning and those from biology. We'll begin by providing context for genomics research as the starting point for much of modern medicine, including the discovery of diseases mechanisms, biomarkers, and therapeutics. We'll then discuss the role of predictive models in enabling 'precision medicine', and their substantial challenges. In a complementary strand, we'll describe how

the language of statistical learning can be used to phrase and interrogate biological questions, as well as those arising from modern developments in machine learning. After an introduction to the types of data encountered in sequencing-based studies along with the opportunities and problems they present, we'll provide some terminology and useful concepts from high-dimensional statistics, Bayesian statistics, and causal inference, and will discuss how these concepts arise naturally in the context of genomic medicine. This will be accompanied by some illustrative examples of how different techniques may be employed in translational scientific research, but the bulk of description of specific applications is left to the following chapters. We will conclude with a summary of the application cases considered throughout the rest of this thesis.

1.0.1 Genomics in medicine

Since the success of the Human Genome Project (Lander *et al.*, 2001), sequencing technologies have improved at an exponential rate, both in terms of cost per megabase sequenced (Wetterstrand, 2022) and the number of individuals who have had some portion of their genome sequenced. Advances have also been made in accuracy and the capacity for long-read sequencing (Goldfeder *et al.*, 2017). This has introduced an invaluable new resource into biomedical research, and initiatives such as the 10,000/100,000 Genomes Projects (Telenti *et al.*, 2016) and the UK Biobank (Bycroft *et al.*, 2018) have drastically improved the accessibility and infrastructure surrounding this data (Szustakowski *et al.*, 2021). This is beginning to show impact in clinical settings (Prokop *et al.*, 2018).

In the study of cancer, a disease of the genome, the ability to rapidly and cheaply sequence normal and tumour-derived DNA has transformed basic research, birthing the field of cancer genomics. While whole-genome sequencing is not yet standard-of-care for the generic cancer patient, access to in-depth genetic data is becoming more common. Cancer-specific repositories such as The Cancer Genome Atlas (Weinstein *et al.*, 2013) have given researchers access to large clinical datasets with a variety of accompanying genomics information, including somatic mutations, copy number alternations, and gene expression profiles.

The clinical impacts of genomic data have manifested themselves in many ways. One use of genomic data in a clincal setting is for subtyping/endotyping patients in a way that informs treatment decisions – a key tenet of so-called 'personalised medicine' or 'precision medicine'. This may involve categorisation in very broad terms, such as the separation of virally and bacterially infected patients (see Chapter 2 and Remmel *et al.*, 2022), or within cancer distinguishing tumours according to the molecular profile of their mutations (Zhao *et al.*, 2019).

Understanding the genomic landscape of disease is also critical to the field of early-stage drug discovery (Nelson *et al.*, 2015; Raja *et al.*, 2017; King *et al.*, 2019). In cancer, knowledge of the location and associated products of oncogenes (genes in which mutation can cause a cell to become cancerous) can allow for intelligent selection of druggable sites (Weinstein, 2002; Bedard *et al.*, 2020), and identification of tumour suppressor genes (genes which under normal circumstances prevent uncontrolled cell division) gives options for therapies which may replace cancer patients' defective cell cycle control mechanisms (Fang and Roth,

2003; Morris and Chan, 2015).

Alongside new drugs, it has become increasingly common for therapies to be offered alongisde genomic biomarkers which may stratify patients who are more likely to benefit from the treatment (Weber *et al.*, 2014; Awad *et al.*, 2019; Zhu *et al.*, 2019; Safarika *et al.*, 2021). This is important for a variety of reasons: to prevent unnecessary suffering for patients unlikely to benefit from drugs with adverse side-effects; to minimise unnecessary cost of wasted treatments unlikely to succeed; and to allow management of the use of therapies for reasons such as antimicrobial stewardship.

New sources and modalities of data have allowed researchers a greatly expanded toolbox with which to investigate the causes and development of cancer and other diseases. Addressing these goals, however, presents a unique set of challenges. Firstly, in clinical settings it is commonly necessary (or at least very useful) to have some sense of the uncertainty accompanying any given prediction or assignment. Secondly, the number of covariates common in 'omics datasets induces a variety of theoretical and practical problems for classical statistical analysis, a problem often referred to as the curse of dimensionality (Barbour, 2019; Bühlmann *et al.*, 2014). Finally, genomic datasets are often collected in observational settings, with limited interventional control, incomplete observation, and compiled from multiple heterogeneous data sources. In later sections we introduce some of the methods developed to address each of these methodological challenges.

1.0.2 Statistical learning and machine learning

Informally, statistical learning and machine learning attempt to address the theoretical and practical challenges associated with extracting information from data by 'fitting' (or sometimes 'training') models that can be used to achieve some goal. Often, this goal is to use a fitted model by predicting a future outcome for some new data points. This is known as *regression* when predicting continuous outcomes, and *classification* when predicting discrete outcomes. In this work we are principally concerned with predictive modelling. While we might be interested in the details of a model we've fitted (we'll refer to this as 'inference'), its subsequent application is generally the primary goal.

The distinction between machine learning and statistical learning is vague and constantly evolving. Several trends are informative but not prescriptive in describing statistical vs machine learning, including an emphasis in statistical learning on theoretical and mathematical guarantees on the future performance of predictive models, a tendency towards scale and non-linearity as features of machine learning models, and (particularly more recently) a specific focus in machine learning on neural networks and deep learning. Despite their differences in nomenclature and focus, however, statistical and machine learning share many common themes and goals. Most notably, they are both concerned with extracting useful information from data as efficiently and robustly as possible. In general, throughout the rest of this work we'll use the term statistical learning.

In order to produce performant predictive systems, extensive research exists towards understanding how best to utilise the *structure* of the complex data

types that underlie many modern common prediction problems. Examples of data types that have seen intense (sometimes frenzied) research over the last decade include images, text, and (as in our case) high-throughput genomics. Notably, none of these data types necessarily follows a standard tabular structure, and all are to some degree high-dimensional (i.e. contain a lot of information per input observation). There is currently a great deal of debate around the role of structure-informed learning algorithms in the fields of image recognition (IR) and natural language processing (NLP). Historically, huge breakthroughs have come from developing learning algorithms that 'exploited the structure' of each of these data types, including convolutional neural networks for IR (Krizhevsky *et al.*, 2012; He *et al.*, 2016; Huang *et al.*, 2018a) and word embeddings for NLP (Mikolov *et al.*, 2013; Bojanowski *et al.*, 2017; Almeida and Xexéo, 2019). However, recent years have seen a trajectory tending towards a relatively small set of model architectures dominating across surprisingly multi-disciplinary divides (Vaswani *et al.*, 2017; Dosovitskiy *et al.*, 2021), with the role of inductive bias and natural complexity regularisation speculated to contribute to this success (De Palma *et al.*, 2019; Valle-Pérez *et al.*, 2019; Goldblum *et al.*, 2023).

It has long been hoped that similar strides forward can be anticipated in biomedical science, but even with the context as laid out above it certainly remains unclear what strategies will work best for genomic and multi-omic data. A few key properties of genomic data distinguish it from other media. Firstly, its dimensionality: while images and text documents can be large, they do not come close to the scale of the human genome. The dimensionality of genomic data depends greatly on the technology being used and and with what 'resolution' the data is being viewed. For example, in Chapters 3 and 4 we take a 'gene-level' view of somatic mutation data, giving a dimensionality in the tens of thousands, while taken at its maximum dimensionality whole-genome sequencing (WGS) data includes measurements of billions of individual nucleotide bases. We'll discuss choices around technology and data dimensionality more in Sections 1.1.3 and 1.3. Secondly, data availability: as mentioned, human genomics data has exploded in abundance over the last two decades, but faces some fairly fundamental limits. Put simply, even with ultra-pervasive sequencing, it seems likely that most humans will produce far more text and images across their lifetime than they will novel genomes for sequencing[1]. Finally, prediction of emergent phenotypes from genomic data may in general simply be more complex than image and text prediction/generation tasks. Image- and text-based tasks are, by design, typically achievable by humans, and these data types are highly compressible. There is, however, no requirement for evolution to have produced a dependence structure within genomic data that is so latently simple. For all of these reasons, there is no guarantee that approaches that have borne fruit in other disciplines will necessarily do so in genomic medicine. Even if they do, we may yet be far from the required richness of data availability to make full use of these methods. Given these potential constraints, there has certainly been a great deal of work towards applying methods from the statistical and machine learning

[1]Note however that this may not be as obvious as it first seems, as advances in techniques for profiling genomic heterogeneity advance even to single-cell resolution.

toolkit to genomics, including but not limited to random forests (Breiman, 2001; Chen and Ishwaran, 2012), support vector machines (SVMs) (Cortes and Vapnik, 1995; Huang *et al.*, 2018b), and artificial neural networks (ANNs) (Avsec *et al.*, 2021; Tran *et al.*, 2021).

In each of the chapters that follow we will have to make choices about how to make best use of the structure of the data types presented, including functional reaction-curve data from gene expression assays in Chapter 2, exome-wide somatic mutation data in Chapter 3, and heterogeneous multi-study survival data in Chapter 4.

1.1 Genomics and biological data

> "En particulier, on ne saurait préciser actuellement le mécanisme selon lequel les gènes désoxyribonucléiques peuvent commander l'édification des acides ribonucléiques cytoplasmiques et le mécanisme selon lequel ces acides ribonucléiques peuvent présider à l'élaboration des enzymes..."
>
> Boivin and Vendrely (1947)

In this section we'll review the central tenets of molecular biology necessary to understand the rest of this work. In particular, we'll focus on how information flows between the three classes of molecule whose interactions determine the majority of cellular functionality. These are DNA, RNA, and proteins, and the flow of information between them is so consistent across nature that it is referred to as the 'central dogma' of molecular biology. Describing this process also allows us to exhibit the main new sources of data that have enabled the genomics (or more generally 'omics', to distinguish from DNA-only analysis) revolution, and to give context to their respective uses and limitations.

Once we've reviewed the molecular biology necessary for subsequent chapters, we'll discuss the specific role of genomics data in medicine and the study of disease. Because cancer forms the dominant thread of the second half of this thesis, we'll devote extra time to discuss some fundamental concepts from cancer genomics. We'll then go into more detail in the specifics of modern technologies for sequencing and quantification across major genomics data types.

1.1.1 The central dogma

Molecular biology has historically enabled the successful collision of two major subfields of biology concerning evolution and cells respectively. On the one hand, evolutionary biology has concerned the role of inheritance and diversity in producing the range of organisms observed in the natural world. On the other, cellular

biology has concerned the mechanistic operation of cells, the constituent units of all living beings. Molecular biology provides the link between these two disparate areas of study via the flow of information between two classes of molecule. Firstly, DNA is the carrier of genetic information and provides the mechanism for Mendelian and Darwinian inheritance. Secondly, proteins are the machines that, at the lowest level of resolution, perform the tasks that sustain and manage the operation of cells in sickness and in health.

We now have a fairly nuanced understand of the connections between the two, with RNA sat between them in the role of 'messenger'[2]. Molecular biology, via a number of key results including the discovery of the structure of DNA in the 1950s by Crick, Franklin, Watson and Wilkins (Franklin and Gosling, 1953; Watson and Crick, 1953; Wilkins *et al.*, 1953; Maddox, 2003), the role of RNA as an intermediary between DNA and proteins (Boivin and Vendrely, 1947; Crick, 1958; Brenner *et al.*, 1961; Cobb, 2015), and the 'cracking' of the genetic code (Crick *et al.*, 1961; Nirenberg and Leder, 1964; Brenner *et al.*, 1967; Tamura, 2016), enabled a conceptual and practical bridge between disciplines and a fundamental unification of modern biology.

The central dogma, a term coined by Crick, can be described as follows (see also Figure 1.1). Contained within the nucleotide sequence of DNA – often represented as a string of letters A, C, G and T, representing the nucleotide bases adenine, cytosine, guanine and thymine respectively – is the hereditary information of an organism. As is necessary for replication at the cellular or organismal level, DNA can be copied with the help of an enzyme known as a DNA polymerase. This is represented by a cyclic arrow from DNA to itself in Figure 1.1. With the aid of an enzyme known as an RNA polymerase, messenger RNA (mRNA) molecules can be can be 'transcribed' from a DNA sequence. This leaves the DNA molecule unchanged in its nucleotide information content but produces a single-stranded mRNA molecule that may be transported out of the cell nucleus where DNA is typically stored. The word 'transcription' is used to emphasise that, while chemically slightly different, DNA and RNA are essentially expressing the same language, with the nucleotides $A/C/G/T$ in DNA directly mapping to the nucleotides $A/C/G/U$ in RNA (the letter U in represents the nucleotide uracil, itself only differing from thymine by a single methyl group). Like DNA, RNA may also be used as a template for direct replication with the aid of an RNA replicase enzyme, also known as RNA-dependent RNA polymerase (RdRp).

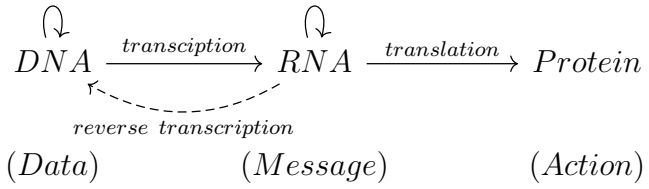

Figure 1.1: The central dogma.

For a typical gene, RNA molecules are transported to the cytoplasm of the

[2]In reality, RNA does a lot more than just this. As far as the central dogma goes, however, this is RNA's main function.

cell and, in structures known as ribosomes – themselves built out of a mixture of ribosomal RNA (rRNA) and protein –, are *translated* to proteins. Proteins are composed of sequences of amino acids, of which there are twenty types. Each amino acid is associated with a set of three-nucleotide *codons*, with special codons also associated with starting and stopping translation. Together, this collection of codon-to-protein instruction maps are referred to as the 'genetic code'.

The central dogma refers to the one-directional flow of information along the DNA → mRNA → Protein chain. Crucially, while genetic information contained in nucleic acids can be converted to protein via translation of mRNA, the reverse will never occur. It was historically unclear whether the one-directional flow of information would extend to transcription of DNA to RNA. This was settled when it was discovered that the conversion of RNA to DNA via reverse transcription, while rare in nature, is demonstrated by various retroviruses (Baltimore, 1970; Temin and Mizutani, 1970; Coffin and Fan, 2016). Nowadays, reverse transcription technology is commonly used throughout genomic research, for example as part of sample preparation for the clinical gene expression testing approach described in Chapter 2. This test also makes extensive use of DNA polymerisation at each reaction stage.

The three classes of molecule described by the central dogma are also the basis of the main new data sources that have powered the genomics revolution of the last two decades. Advances in DNA sequencing have allowed systematic and high-throughput analyses of the relationships between traits (phenotype) and DNA sequence (genotype). Where in previous decades slow and labour-intensive procedures such as Sanger sequencing (Sanger *et al.*, 1977) would have been required for any direct analysis of nucleotide sequence, today we have massively parallelised and automated systems for achieving the same analysis at far lower cost (we'll discuss these more in Section 1.1.3). In Chapters 3 and 4 we'll be working with DNA sequencing data derived from cancer patients. In fact, for each patient in the studies we discuss we'll be considering *two* sets of DNA sequencing data – one from the patient's 'normal' cells, and the other from cancerous tumour cells. We'll be particularly concerned with the points at which these two sequences differ.

The advances in sequencing described above have, in turn, also revolutionised RNA analysis, birthing the field of transcriptomics. In both genomics and transcriptomics, however, high-throughput techniques have not entirely replaced more direct approaches to detection and quantification of DNA/RNA. During the COVID-19 pandemic, for example, extensive use was made of diagnostic tests based on polymerase chain reaction (PCR) to detect the RNA sequence of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pathogen. For this application, sequencing would have been unnecessarily complex in order to detect and quantify a single RNA sequence. In Chapter 2, we'll see another application of a direct quantification technique with many similarities to PCR to a 'medium-throughput' problem, in which for different reasons both sequencing and PCR are unsatisfactory.

Advances to medium- and high-throughput study of protein abundance in cells have also been made (referred to as proteomics), but have typically relied on different underlying techniques to genomics/transcriptomics. While changes

may come with the advent of exciting new nanopore-based single-molcule protein sequencing (Afshar Bakshloo *et al.*, 2022; Motone and Nivala, 2023), this is as of yet far from commonplace. While proteomic data is certainly valuable as a snapshot of the functional state of a given sample, we don't consider proteomic data in any detail in this work precisely due to its relative lack of abundance.

1.1.2 Genomics in disease

While interest in the patterns of inheritance of common and rare diseases predates our understanding of genetics or even Mendelian inheritance (Emery, 1989), the advent of molecular biology transformed our understanding not just of patterns of disease inheritance, but also in some cases directly of the mechanisms of disease. Early examples of conditions with genetically identified causes included sickle-cell anaemia (Ingram, 1956) and Down Syndrome (Lejeune *et al.*, 1959). Notably, historic advances in our understanding of disease have often led to improvements in prognosis and diagnosis, but far less frequently to successful treatments. For example, as early as 1961 Robert Guthrie developed a method for screening infants for phenylketonuria (PKU) that depended on detection of an abnormal amino acid (whose role in PKU was itself discovered nearly thirty years earlier; Fölling, 1934). While treatable with dietary changes, however, PKU is not directly curable to this day (Mohanty, 2014). Similarly, while recent advances in gene therapy have sparked hope for a general-purpose cure to sickle-cell anaemia (currently a blood and bone marrow transplant from a close genetic match is required), this still seems some way off, re-emphasising the typically long arc from genetics to cure.

The inherent difficulty in translating advances in basic science to treatment and drug discovery has persisted into the genomics era, perhaps more so than was envisioned by many at the conclusion of the Human Genome Project (see, for example, Emilien *et al.*, 2000). In particular, end-to-end success rates for clinical trials have not improved substantially (Dowden and Munro, 2019), although there is mounting retrospective evidence that support from genomic data is a major predictor of clinical trial success (Nelson *et al.*, 2015; King *et al.*, 2019).

Outside of therapeutic and pharmacological spheres, however, there has been substantial progress in the development and deployment of genomics-based approaches to guide clinical decision making or track the course of disease. These have included the use of genomic data in diagnostic tests and as accompanying biomarkers to therapeutics. Here we'll discuss an example of each, which (conveniently) correspond to the subjects of Chapters 2 and 3 respectively.

Chapter 2 is placed in the context of acute care patients who may experience inflammation and/or sepsis/septic shock. Systemic inflammation (the erroneous or excessive application of the process by which the body fights infection) can be caused by a multitude of factors, including bacterial infection, viral infection, and non-disease related mechanism such as response to trauma (Chen *et al.*, 2017). Appropriate clinical action depends upon determining the origin of inflammation. This is particularly true when deciding whether to treat with antibodies, whose non-use can entail significant and imminent impacts to mortality (Liu *et al.*, 2017) but whose inappropriate use can contribute to the development of

antimicrobial resistance (Fitzpatrick *et al.*, 2019). Accurate diagnostic testing for bacterial infection is therefore crucial. Genomic technologies have been used to test for a multitude of bacterial infections, often through the detection of specific DNA/RNA sequences (Fournier *et al.*, 2014). However, the general success of this strategy relies upon any given bacterial infection falling within the set of genomic targets of a given detection assay. There has recently been interest (and a certain degree of success) in developing genomics-based diagnostic tests based on indirect detection via profiling *host response* rather than directly via detection of a pathogen (Safarika *et al.*, 2021; Kelly *et al.*, 2022). While pathogen detection tests are based on a set of targets (often DNA targets for bacterial infections) from the pathogens in question, host response diagnostic tests will typically comprise a set of human gene targets and will be based on profiling of gene expression (Ram-Mohan *et al.*, 2022) or protein expression (Vanderboom *et al.*, 2021).

Chapter 3 concerns the estimation of high-throughput genomic biomarkers. Very generally, a *biomarker* is a quantity that can be measured for a given patient or sample and provides some relevant clinical information. Different biomarkers target very different characteristics, for example by measuring the abundance of a given chemical (e.g. a small molecule; Qiu *et al.*, 2023, or a large protein such as programmed death ligand 1 (PD-L1); Doroshow *et al.*, 2021), the visual characteristics of a sample (Smith *et al.*, 2003), or the mutation status of a given genetic locus or gene (e.g. BRCA biomarkers in breast cancer; Walsh *et al.*, 2016). As well as measuring a wide variety of phenomena, biomarkers are also diverse in the type of task they are designed to support. These tasks may be *prognostic*, i.e. aiming to predict the likely course of a disease, *diagnostic*, i.e. aiming to establish the true nature of a disease, or *theranostic*, i.e. aiming to guide decision-making around allocation of therapies. Increasingly more complex biomarkers are being developed, including for example those depending on the mutation status of multiple genes. While complex biomarkers may increase the power of genomic medicine to support clinical decision-making, their development raises issues in reproducibility, dimensionality, cost efficiency, and interpretability.

Cancer: 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 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 areas. 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.

We now turn to the aspects of genomics specifically relevant to cancer. 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 and every tumour are 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 molecular genetics, cancers of disparate tissues of origin were grouped together because of the common 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[3], 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 DNA and its role as the primary mechanism of inheritance by Watson and Crick (1953) and others, 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[4]. Furthermore, the size of the

[3]While some rare cancers such as polycythaemia 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).

[4]Tissues/cell types in which cancer is extremely uncommon tend to be those which experience

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, 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.

Nature vs nurture: cancer beyond the genome

In the section above we described how modern sequencing technologies allow us to investigate questions in oncology. It's appropriate also to address the inherent limitations of an approach based on tumour genomes. This 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 predetermined 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.

Since in the cancer-centric chapters of this thesis we'll be focusing on what tumour genomes can tell us, it's important to highlight the information that can't be accounted for with the somatic mutation datasets we'll be using. Firstly, we'll be concerned with datasets describing the *changes* in DNA sequence between the germline (the DNA sequence a patient inherits) and somatic tumour cells. This means that we typically won't be accounting for germline variation between patients. Many cancers have ancestral components, and molecular differences in the normal functioning of cells may certainly impact the abnormal functioning of cells, such as in cancer. Tumours also do not exist in a vacuum – increasingly, oncology is beginning to acknowledge the important of the tumour *microenvironment* in cancer progression (Whiteside, 2008). The tumour microenvironment consists of non-cancer cells 'recruited' to support a tumour, in particular immune cells involved in producing an inflammatory response.

In order to gain a full picture of cancer, therefore, approaches centred purely on characteristic of tumours (such as those presented in this thesis in Chapters 3 and 4) must be augmented by investigations into the body's response to tumours. An analogy to this in a non-cancer context can be found in Chapter 2. The background to this chapter is the development of diagnostic biomarkers determining the cause of an inflammatory response (in particular, classifying bacterial versus viral infections). The approach taken, rather than attempting to detect the

very little reproduction, and so have little chance to accumulate mutations, e.g. neuronal cells and tissues making up the heart.

pathogen in question directly, is to detect the body's *response* to infection through the gene expression profile of immune cells. Chapter 2 focuses on how to make these measurements sensitive and precise.

1.1.3 Modern genomics technologies

DNA sequencing

Modern genomic medicine is underpinned by the ability to sequence DNA cheaply and quickly. DNA is organised into chromosomes, along each of which many genes are arranged, with further non-coding regions interspersed in between. As described in Section 1.1.1, the fundamental units of DNA are nucleotide bases, of which there are four varieties (labelled C, G, T, and A). These are organised in groups of length three called codons, which code for the production amino acids. Codons are arranged in sequences such that their amino acids when joined in a chain form proteins - the products of genes.

The aim of sequencing is to read, base by base, the information content of DNA. This was originally done by Sanger sequencing, a procedure to infer the base composition of a piece of DNA one base at a time (Sanger *et al.*, 1977) via electrophoresis. Short-read high-throughput sequencing automates this process via the following (Bentley *et al.*, 2008):

1. DNA is isolated from a sample and amplified (replicated many times) to ensure good signal.
2. Purified DNA is broken into many pieces of manageable length.
3. These short strands act as templates for polymerisation by fluorescently tagged nucleotide bases.
4. This fluorescnece is automatically detected by imaging to produce many short 'reads' corresponding to strands.
5. These short sequences are matched to a reference human genome to identify where the DNA in the original sample differed from that reference.

The short-read sequencing paradigm, while ideal for speed and scale, relies on the existence of a complete human 'reference genome', onto which individual sequences can be matched. While short-read technologies still dominate sequencing studies, other technologies are now increasing in capability and popularity (Jain *et al.*, 2016).

Tumour/normal variants As described in previous sections, in cancer some subset of cells accumulate mutations. These occur via random misreplication of DNA during cell division or exposure to some external mutagen (e.g. cigarette smoke or UV light). Tumour cells therefore contain DNA with a different sequence to that of the patients' typical (germline) sequence. To analyse these differences systematically with sequencing, typically two samples are collected: one from a tumour and one from normal tissue. The sequences of each of these samples are

compared to produce a list of locations at which mutations have occured: these mutations can have a variety of types (replacements, insertions, etc.) and can have vastly differing functional implications. These are often stored/provided in variant called format (VCF) or mutation annotated format (MAF) files. In our use cases we will simplify feature generation from these datasets by only considering counts of mutations of a given type per gene and sample. Finally, note the similarity in relationship *a)* between tumour genome and germline genome, and *b)* between germline genome and human reference genome. Despite this similarity, the actual computational and algorithmic steps taken to produce germline/reference versus tumour/normal comparisons differ because (at least for short-read applications) a preassembled reference genome is typically available for the former.

Targeted sequencing While WGS and whole-exome sequencing (WES) have become far cheaper in recent years, it is still often not advantageous, particularly in the presence of cost or time constraints, to sequence an entire exome. In particular, if a (relatively small) set of target regions are known to be of interest, 'gene panels' designed only to sequence those regions can be deployed fairly simply. In many cases the cost of sequencing a given region at a given depth scales fairly linearly with the total combined length of the genomic regions being targeted. Many commercial gene panels are available (for example for cancer monitoring), often comprising 100-1000 genes.

Gene expression

As described above, DNA sequencing allows us to investigate the coding sequence of DNA in normal and dysfunctional cells. The central dogma (described in Section 1.1.1), however, tells us that this is only a starting point towards understanding the full picture of cellular behaviour. The revolution in cheap DNA sequencing has, in turn, impacted our ability to quantify RNA at scale. However, in our gene expression-focused application study in Chapter 2 we don't use sequencing-based approaches. In order to give good context to the reasons for this, here we'll briefly describe RNA sequencing, PCR (another common lower-throughput method for RNA detection/quantification), and why neither are satisfactory for the aims of Chapter 2. The technology we will end up using, know as LAMP, is most similar to PCR and so we'll dedicate a reasonable amount of time to discussing how PCR looks in practice to motivate both its similarities and differences with LAMP.

RNA sequencing RNA-Seq, short for RNA Sequencing, is now a very common method for high-throughput detection and quantification of RNA (Wang *et al.*, 2009). Simplified, RNA-Seq works by converting RNA to DNA via reverse transcription (a common theme in what's to come), then applying standard high-throughput sequencing methods to the resultant complementary DNA (cDNA). This is a slight oversimplification but contains the pertinent steps. In practice, great care needs to be taken in RNA library preparation to ensure that RNA is isolated from genomic DNA (gDNA) and of sufficient quality, entailing several further steps.

While RNA-Seq has been revolutionary for exome-wide gene expression studies, many applications remain where it is unnecessary or insufficient. Commonly this is because RNA-Seq, while incredibly cheap compared to historical levels, is still more expensive than simpler techniques for low-throughput analyses, e.g. detection of a single RNA sequence. For these, it is still common to use older methods such as reverse-transcription PCR.

PCR Polymerase chain reaction (PCR) was developed by Kary Mullis while working at Cetus Corporation in 1983, a feat for which he jointly receieved the Nobel Prize in Chemistry in 1993, and has served as a component of much work manipulating DNA molecules ever since (Saiki *et al.*, 1985; Mullis *et al.*, 1986). The aim of PCR is, given an input DNA molecule, to *amplify* it *in vitro*. By *amplify*, we mean produce many identical/complementary DNA molecules. By *in vitro*, we mean in an artificial context without access to normal cellular machinery. With the basic mechanism of PCR, which we will describe below, it is fairly simple to extend to detection/quantification of RNA with a reverse-transcription preparation step.

PCR works by alternating between two steps (referred to together as a 'cycle'): a *denaturing* step and a *synthesis* step. Beginning with a double-stranded DNA sequence, a sample is heated until the DNA denature, i.e. splits into two complementary single strands. In the second step, a polymerase known as 'Taq' catalyses the synthesis of a complementary sequence onto each single strand. This is initiated by the presence of a primer, a short oligonucleotide comprising some portion of the desired sequence that hybridises to single stranded DNA and acts as a starting point for polymerisation.

A significant complication to the implementation of PCR reactions is that its two stages must occur at different temperatures. The Taq polymerase has an optimal temperature of around 70-80°C, while denaturation occurs around 94-98°C. This variation in temperature determines the progression of subsequent PCR reactions and requires a thermocycler. This restricts PCR experiments to be conducted in laboratory-like settings, rather than in field or point-of-care applications.

Up until this point we have used terminology pertaining to *detection* and *quantification* interchangeably. In this context, detection refers to an assay returning a binary outcome indicating whether the molecular sequence in question was present in a sample, whereas quantification (also sometimes referred to as *quantitation*) refers to an assay returning a continuous non-negative value denoting the level of abundance of a given molecule. Using the PCR assay for quantitation, referred to as quantitative polymerase chain reaction (qPCR), has grown in popularity since the development of automated fluorescence detection. The outputs of qPCR analyses are fluorescence curves produced by the detection of fluorescent markers attached to the nucleotide bases used in synthesis, and can track the amount of polymeristion occurring during any given cycle. These typically look as shown in Figure 1.2, consisting of relatively flat phases at the beginning and end of a reaction, with more rapid growth phase in between. Detection is typically performed via establishing a threshold (grey dashed line) above

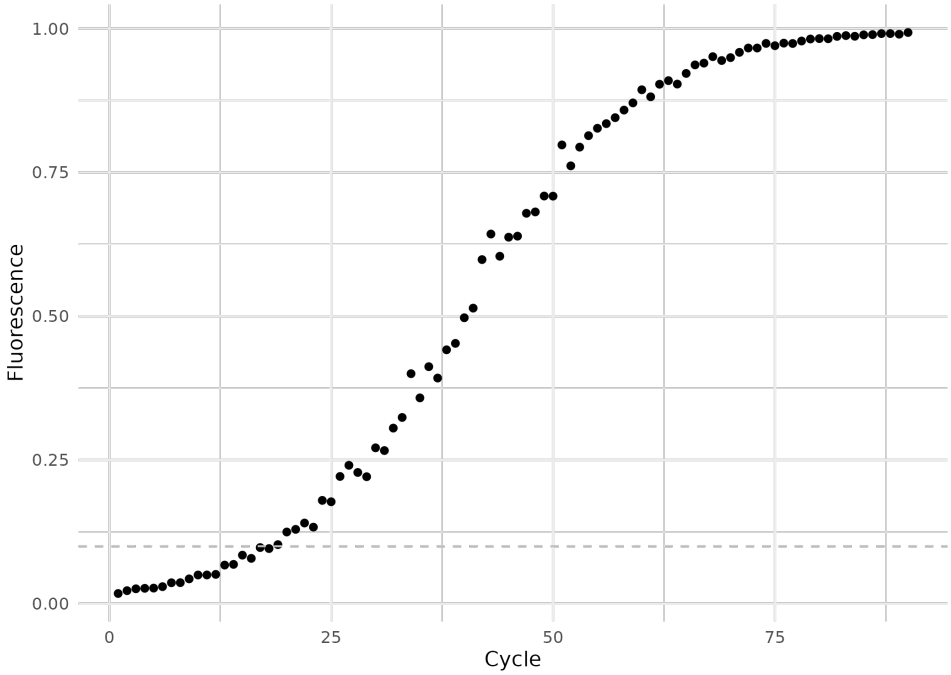

Figure 1.2: Example fluorescence curve as might be produced by a PCR reaction. Values on y-axis on arbitray scale, with example quantitation level shown with grey dashed line.

which the target is said to be present. Traditionally quantitation has proceeded in a similar fashion, by relating estimated abundance to the number of cycles before the quantitation threshold is met and disregarding all other information.

PCR and qPCR require the design/selection of a primer (actually two primers, but one is typically the complement of the other). This primer, a short oligonucleotide matching some portion of the target sequence, must be chosen to balance several ideal properties. It must be suitably specific to the target DNA sequence, while also being of an appropriate length. Designing PCR primers is both an art and a science, and analagous to one of the problems investigated in Chapter 2, albeit in the context of another assay (known as LAMP) that shares some features with PCR but, notably, depends on the design of at least six primers. This introduces a much higher degree of complexity, as does the nature of LAMP reactions, which occur far more in parallel than PCR reactions which are neatly divided into cycles. The reason to invest in understanding this more complex system is precisely that LAMP is not designed around cycles, because it happens at constant temperature. This eliminates the need for a thermocycler, enabling LAMP to be deployed cheaply and space-efficiently for point-of-care applications such as the one described in Chapter 2.

1.2 Statistical learning theory

> "We need data to construct prediction rules, often a lot of it."
>
> "For prediction purposes, [linear models] can sometimes outperform fancier nonlinear models, especially in situations with small numbers of training cases, low signal-to-noise ratio or sparse data."
>
> Elements of Statistical Learning (Hastie *et al.*, 2009)

Now familiar with the most relevant biological concepts, we turn to the mathematical theory and techniques underlying statistical learning, which has experienced a surge of interest in the last few decades, partially fuelled by explosions of data availability across multiple domains, contemporaneous increases in computational power, and the desire to provide satisfying theoretical frameworks to accompany complementary practical advances in machine learning. While so far there is no theory of statistical learning with the capacity to fully describe, explain, or validate the impressive results demonstrated in (for example) deep learning, the language of statistical learning is invaluable in specifying learning models, providing metrics by which they can be measured, and triaging where they go wrong. It is also the language with which we will be attempting to interrogate issues of inference and prediction in cancer genomics.

Firstly, we lay out some terminology and notation. We often consider a very generic setup, in which we have paired data

$$\mathcal{D}_n = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}.$$

We refer to this dataset as our 'training' data. We do not here consider settings in which components of inputs x_i or ouputs y_i are missing. We model each of these pairs as being drawn independently from a joint probability distribution $P_{X\times Y}$ describing the likelihood of observing any combination of observation x and label y. We write that $y_i \in \mathcal{Y}$ for each $i \in \{1, \ldots, n\}$. For now we make no assumptions about the nature of the $\mathcal{Y}$: labels may be continuous values for regression ($\mathcal{Y} \subset \mathbb{R}$), discrete values for classification ($\mathcal{Y} = \{0, 1\}$), or more complicated hybrid objects such as is the case in survival analysis (see Section 4.2). We assume that $x_i \in \mathcal{X} \subset \mathbb{R}^p$ for each i, so that our observed values are vectors of length p and each element is a real number (possibly restricted to some subset such as the positive reals - this is what $\mathcal{X}$ specifies). We refer to p as the dimension and n as the sample size of our data.

We wish to fit some model $\mathcal{M}$ to the data. Formally, this model $\mathcal{M}$ refers to a family of probability distributions in which we assume the true distribution $P_{X\times Y}$ lies, sometimes indexed by a discrete set of parameters (referred to as

a *parametric* model). The processing of fitting a model refers to choosing an estimated distribution that matches the true distribution as well as possible. This could be in order to make some *inference* about the true distribution, for example through estimating its parameters, which will hopefully shed light on the effect of each of the covariates contained in an observation x. Alternatively, we might be principally interested in predicting future values of y from unlabelled observations as accurately as possible. These two aims are often distinguished by the umbrella terms statistical inference and statistical learning.

1.2.1 Bayesian and frequentist approaches to statistics

In this thesis, we will use two different frameworks for describing and fitting statistical learning models. Each of these inherit from a long tradition of statistical research, and debate around the relative merits of each has been extensive and sometimes acrimonious (Bayarri and Berger, 2004; Coles, 2006; Vallverdú, 2016). These two schools of thought are *frequentist* and *Bayesian* statistics, and to give a full summary of the history of each and their interactions is well beyond the scope of this introduction. Therefore, we will restrict ourselves here to a brief description of some motivational and practical differences, enough to justify why one may select one over the other for a particular application. This will become relevant in the chapters that follow, as we adopt a Bayesian perspective in Chapter 2 and a frequentist perspective in Chapters 3 and 4.

Notions of probability

The divergence between Bayesian and frequentist statistics is in some senses deeply philosophical, and yet often also purely pragmatic. Few would nowadays hold that any working statistician need embrace only Bayesian or frequentist methods as a matter of principle (Gelman and Yao, 2021; Bon *et al.*, 2023). However, it is worth emphasising that differences in these two flavours come down to as fundamental a concept as the definition of probability. While most day-by-day decisions around which statistical framework to use would rarely necessitate considering this, it is useful to think about in order to understand where motivating differences have originated, even as an applied statistician.

In the frequentist framing, probabilities are an intrinsic property of some system. Given an experimental setup (the classic examples being a coin toss or die roll), we are invited to imagine the experiment being repeated over and over again with an identical (or at least indistinguishable) starting point. The probability of an event is defined as the limit, as the number of experiments goes towards infinity, of the the proportion of experiments in which the event occurs. So as an increasing number of, for example, fair coin flips are performed, the number of heads obtained should tend towards one half of the total number of flips. It may seem convoluted to define probability as the limit of an infinite sequence, but it is worth remembering that this is no more complex than the standard definition of a real number[5].

[5]The reals $\mathbb{R}$ are typically constructed as equivalence classes of Cauchy convergent infinite sequences of rational numbers.

In the Bayesian framing, probability is a subjective concept and the relates to a particular state of knowledge around a system or event. Bayesian probability may be formulated as a quantification of uncertainty about an event. So for example, in the Bayesian framework there is no inconsistency whatsoever about the 'probability' of an event differing between two observers even the underlying experiment is the same, if the observers are privy to different information. This freedom in fact forms the basis of Bayesian inference, where the probability distribution concerning a given quantity is updated on the basis of new data.

Finally, it is worth pointing out that these two notions of probability are simply different *interpretations* of the same axiomatic framework of probability due to Kolmogoroff (1933), or even merely that axiomatic framework being put to different uses. There is no mathematical difference in the laws of probability between the two schools; only in the machinery that probability theory is used to build. In the next section we will see Bayes' rule, a fact about probability that is equally true in any inferential framework. Bayesian statistics having derived its name from the rule[6] is simply a reflection of the central role it plays in doing Bayesian statistics.

Uncertainty quantification

The different notions of probability underlying Bayesian and frequentist statistics give rise to different methods of inference and prediction, allowing for and demanding differing interpretations. Let's return to a prediction task as described in the preceding section, and propose the following linear model for Y's dependence on X:

$$Y \mid X = x \sim \mathcal{N}(x^T\beta, \sigma^2).$$

Let us suppose for now that σ is somehow known and focus on inference for the parameter β. In the frequentist framing, we have postulated the existence of a true β and it is our objective to estimate it as well as possible. We typically do this via maximum likelihood estimation (MLE): from our training data $\mathcal{D}_n$ we calculate the likelihood of observing the values y_i that we have under our model, given the observed values of x_i and in an alternate universe where the true value of the model parameter was given by some $\hat{\beta}$. We then choose the $\hat{\beta}$ such that this likelihood is highest and use it as our estimate of the true parameter. We can show that the estimate $\hat{\beta}$ satisfies

$$\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \{\frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T\beta)^2\}.$$

We would not expect that the MLE procedure will give us the correct answer every time. Indeed, we may quantify exactly how far away the estimate will typically be from the true value under repeats of the same experiment. In this sense we are doing uncertainty quantification, but the uncertainty being quantified is due to the random noise inherent in the model, i.e. it is still strictly a frequentist uncertainty.

[6] Or at least shared the derivation of its name with the rule.

Bayesian inference takes a different approach. Because in Bayesian analysis we are permitted to define a probability distribution purely expressing our belief about a given quantity, we arrive with a predefined distribution, known as the *prior*, whose density we will refer to as $p_\beta(\beta)$. We may also (as in the frequentist case) define the likelihood of observing $Y = y$, given all other quantities according to our model. We write this as $p_{Y|X,\beta}(y; x, \beta)$. At this point we may now apply Bayes' theorem, allowing us to recast one conditional probability in terms of another, to recover:

$$p_{\beta|X,Y}(\beta; x, y) = \frac{p_{Y|X,\beta}(y; x, \beta)p_\beta(\beta)}{p_{Y|X}(y; x)}.$$

Here the denominator $p_{Y|X}(y; x)$ is the likelihood of observing $Y = y$ given $X = x$, marginalised across the prior $p_\beta(\beta)$. The converse conditional $p_{\beta|X,Y}(\beta; x, y)$ is referred to as the *posterior* distribution, and may be understood as an updated distribution reflecting our new understanding having observed data relating X and Y. Uncertainty quantification in the Bayesian paradigm is therefore far more direct: our 'current' distribution for β at each point reflects our state of knowledge about the variable, which is liable to change as and when new data becomes available.

It is worth mentioning the impact that a choice of inferential framework has on subsequent prediction tasks. When predicting outputs for future observations in a frequentist framework, typically the best estimates (e.g. MLE estimates) of a given parameter are used, with uncertainty bounds on the outcome provided due to stochastic elements of the model (in this case the variance σ^2). This is fundamentally different however in Bayesian analysis, where uncertainty in the parameters of the model can propagate through and are philosophically indistinguishable from stochasticity in outputs. In this sense, Bayesian prediction models can be thought of as continuous weighted mixtures of fixed-parameter models.

Model fitting

In performing Bayesian inference, computation of the numerator of the defining Bayes' equation in the previous section is typically computationally easy for any given β. However, the total *evidence* $p_{Y|X}(y; x)$ in the denominator can be very hard to compute, especially as the dimensionality of β increases. This necessitates the use of approximate schemes for producing empirical samples drawn from the posterior distribution of β. The most well-known of these is Markov chain Monte Carlo (MCMC) (Hastings, 1970; Gelfand and Smith, 1990). To produce empirical samples, the MCMC algorithm proceeds from a random starting point, and combines proposals of random jumps with a criterion for accepting or rejecting steps based on the ratio of posterior probabilities between the two points. This ratio is available because it does not rely on the evidence. It is possible to show that this procedure forms a Markov chain with a stationary state given by the true distribution. Alternates to MCMC include Hamiltonian Monte Carlo (HMC) (Neal, 2011), which will be used for all inferences in this work. HMC explores the state space of β in similar fashion but allows for a 'momentum' component

of the current state, allowing for more efficient overall exploration.

1.2.2 High-dimensional statistics

Informally, we may think of high-dimensional statistics as being concerned with the realm in which the dimensionality of our input data, p, is comparable to or greater than the number of training samples n we have available. In this regime the classical asymptotic theory of statistics, which generally relies on an assumption of fixed dimension and considers limiting behaviour as $n \to \infty$, may fail to apply. Results such as the law of large numbers and central limit theorem are not applicable[7].

High-dimensional statistics attempts to gauge what we can do in regimes such as these. One common approach is to assume that the data being modelled has some low-dimensional structure. For our purposes, this means that we can embed our input data into a lower dimensional space such that this smaller representation of the random variable X contains all or most relevant information about Y. This assumption may be motivated by external knowledge about the system being modelled, or may be purely a practical choice in the hope that improvements in predictive or inferential performance can be seen empirically. In order to leverage structural assumptions, we will need to translate them into some algorithmic choice for fitting models. This can be achieved in a variety of ways, including restrictions on the model family $\mathcal{M}$ and adjustments to a given model fitting method. One particularly common approach to the latter is called *regularisation*, and will form an imporant component of much of the work throughout this thesis.

Assumptions of structure

Structural assumptions for statistical learning models aim to directly tackle the problem at the heart of high-dimensional statistics: that if each component of our input variable X can have a full and independent contribution to Y, then we will not have enough data to recover the information contained in each variable. This is true even in a relatively simple class of models, e.g. ordinary least squares (OLS). We therefore assume that some lower-dimensional representation of X is responsible for determining all or most of Y's dependence on X. We can formulate this in greatest generality via the principle of sufficient dimension reduction (SDR) (Adragni and Cook, 2009). For this, we say that given random variables (X, Y) specified by a joint distribution $P_{X\times Y}$, there exists a sufficient dimension reduction of size d^* if there exists some function $\phi\colon \mathcal{X} \to \mathbb{R}^{d^*}$ with $d^* < p$ such that Y is conditionally independent of X given $\phi(X)$, i.e.

$$Y \perp\!\!\!\perp X \mid \phi(X).$$

This is a very general form, and while it includes or motivates many practically implemented strategies, in order to be useful these are typically even more restrictive. For example, one may restrict that the sufficient dimension reduction is

[7]This is not, of course, to say that they are not *true*; simply that we are working in a regime in which their conclusions are not helpful.

a linear map (Omidiran and Wainwright, 2010; Cannings and Samworth, 2017). One particularly common strengthening of SDR is an assumption of *sparsity*. Sparsity conveys that only some small subset of input covariates are important. We may formulate this by insisting that the reduction map $\phi\colon \mathcal{X} \to \mathbb{R}^{d^*}$ has the form

$$\phi\Big(\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}\Big) = \begin{pmatrix} x_{i_1} \\ x_{i_2} \\ \vdots \\ x_{i_{d^*}} \end{pmatrix},$$

where each of the i_j are distinct and $i_j \in \{1, \ldots, p\}$ for $j \in \{1, \ldots, d^*\}$. Here the subset of covariates that are important are the collection $\{x_{i_1}, x_{i_2}, \ldots, x_{i_{d^*}}\}$ of size d^*. Note that we typically do not know which set of covariates these are in advance, but even assuming that there exists such a subset can prove a surprisingly useful restriction.

It is worth at this point drawing a distinction between two phenomena in statistics and data science both referred to as 'sparsity', both of which are relevant to genomics applications. The first is sparse *data*, in which almost all observed data points have the same value (typically zero). Mutation data displays this trait – the rate at which mutations occur in the genome varies widely across and within cancer types, but rarely exceeds 100 Mut/Mb, i.e. one mutation per 10^4 nucleotide base pairs (Chalmers *et al.*, 2017). Sparse data can be exploited in many ways, including through efficient storage and specially tailored algorithms. However, here we will focus on sparse *models*, for which it is assumed that only a small subset of the input covariates are relevant.

It is also notable that sparsity has a particularly simple interpretation in the context of linear models (and some generalisations of them). Suppose that we once again consider:

$$Y \mid X = x \sim \mathcal{N}(x^T\beta, \sigma^2).$$

This does not specify an entire model for (X, Y) (we are making no claims about the marginal distribution of X), but does specify a family of distributions for Y conditional on the value taken by X, parameterised by $\beta \in \mathbb{R}^p$. We may then say that this model is d^*-sparse iff

$$|\beta|_0 \leq d^*, \text{ where } |\beta|_0 := \sum_{j=1}^{p} \mathbb{1}\{\beta_j \neq 0\}.$$

This is especially useful in that we've related the structural modelling assumption we're making to the parameters we aim to fit. This will lead naturally to model fitting approaches described in the next section. Methods that encourage sparsity as described here will naturally perform *variable selection* as part and parcel of the model fitting process.

From structure to regularisation

Now that we've shown how we may formulate a low-dimensional structural assumption, we explore how this might be translated into an algorithmic imple-

mentation or adjustment for model fitting. Picking up where we left off with sparse linear models (a good example case for exploring more general principles), we describe the least absolute shrinkage and selection operator (LASSO) method (Tibshirani, 1996). While we don't deploy this directly in the subsequent chapters, many of the methods we utilise (including the group LASSO in Chapter 3 and L_1-penalised neural networks in Chapter 4) are expansions of the concept.

Taking the definition of parametric sparsity given in the previous section, we may see that fitting an OLS model with an assumption of d^*-sparsity equates to finding $\hat{\beta}$ satisfying

$$\hat{\beta} \in \underset{\substack{\beta \in \mathbb{R}^p \\ |\beta|_0 \leq d^*}}{\arg\min} \{\frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2\}.$$

There are two further innovations required to arrive at the LASSO method. The first is to note that solving the optimisation above is equivalent, for some penalty term η indirectly determined by d^*, to solving the *Lagrangian dual*:

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \{\frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \eta |\beta|_0\}.$$

To make this equivalence valid for a small value of d^*, the 'hyperparameter' η should be chosen to be large, and vice versa. This is technically valid but computationally intractable, as the $|\cdot|_0$ norm[8] is non-convex. Practically, to solve this we would have to perform a separate optimisation for each of

$$\binom{p}{d^*} = \frac{p(p-1)\dots(p-d^*+1)}{d^*!}$$

potential covariate subsets. The second LASSO innovation is therefore to replace the L_0 norm $|\beta|_0$ with the L_1 norm $|\beta|_1 := \sum_{j=1}^{p} |\beta_j|$. This makes for a convex optimisation problem which can be solved extremely efficiently. Formally, this is known as the *convex relaxation* of the original optimisation problem (for a more thorough introduction see, for example, Wainwright, 2019, Chapter 7) and provides many appealing properties. In particular, L_1-penalised regression encourages fitting a parameter β with many covariates that are exactly zero (as opposed to, for example, L_2-penalised regression, also known as *ridge regression*). Our final LASSO optimisation is therefore to find β satisfying

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \{\frac{1}{n} (y_i - x_i^T \beta)^2 + \eta |\beta|_1\}.$$

The form of alteration that has been made to the model fitting process here, by adding an additional 'penalty' encouraging the resultant model to be fitted in a certain way, is known as *regularisation*. Several examples of regularisation will be demonstrated throughout this thesis.

We may also choose to incorporate structure into our statistical learning mod-

[8] Technically a 'psuedonorm', as it does not scale homogeneously, i.e. $|\lambda\beta|_0 \neq \lambda|\beta|_0$ in general for scalar λ.

els via the choice of model space that we allow. This approach, in which the overall 'shape' of a model is chosen to encourage it towards some structural assumption, is sometimes known as *implicit* regularisation (as opposed to the *explicit* regularisation demonstrated above). This might be, for example, via what is known as representation learning, in which a map is learnt embedding inputs into some space, the outputs of which may be used for prediction. A celebrated example of this is in autoencoder and variational autoencoder neural networks (Lecun, 1987; Kingma and Welling, 2013), in which a low-dimensional representation is fitted using a network with a 'bottleneck' consisting of relatively few nodes. In Chapter 4, we will utilise both representation learning and a tailored regularisation approach for approaches to causal inference.

1.2.3 Causal inference

In the final portion of this introduction to techniques in statistical learning, we discuss some of the foundations of causal inference. Causal inference is far from new, but has generated increased attention in recent years. Of particular interest to us in Chapter 4 will be blending flexible regression methods such as random forests and neural networks with approaches in causal inference to extract meaningful conclusions from messy, complex, and disparate data.

Causal inference concerns understanding the effect of some *treatment* on an outcome, potentially in the presence of spurious associations (known as *confounding*) between treatment and outcome due to their shared relationship with an outside factor. Estimating the impact of a treatment on an outcome is difficult because, by definition, we will never observe the outcome for a particular treated sample if we had not performed the treatment, and vice verse. To mitigate this we use a framework called *potential outcomes* due to Donald B. Rubin (Rubin, 1974, 2005).

Potential outcomes

Here we continue to consider covariates X taking values in $\mathcal{X} \in \mathbb{R}^p$ (which we may now also sometimes refer to as *confounders*), and a response variable Y taking values in $\mathcal{Y}$. We also introduce a binary treatment indicator variable A taking values in $\{0, 1\}$, such that now we have a jointly distributed triple (X, A, Y). We want to understand the effect of altering the treatment A on the outcome Y. A natural quantity to estimate might be

$$\mathbb{E}[Y|A = 1] - \mathbb{E}[Y|A = 0].$$

However, if both Y and A depend on X, we might observe some spurious dependencies. In Rubin's framework we therefore want to define reasonably comparable *potential outcomes* of Y in a hypothetical alternate universe in which we had intervened/not intervened to change A.

Let $Y^{(0)}, Y^{(1)}$ refer to potential outcomes taken by Y under intervention with $A = 0, 1$ respectively. We may therefore say[9] that $Y = Y^{(A)}$. We make two

[9]In the alternate notation of do-calculus (Pearl, 1995, 2012), we would say that $\mathbb{P}(Y^{(a)} =$

further assumptions:

1. *Positivity:* the propensity function is positive almost everywhere, i.e.

$$\pi(x) := \mathbb{P}(A = 1|X = x) \text{ is such that } \pi(X) \in (0, 1) \text{ almost surely.}$$

2. *No unmeasured confounders:* treatment is 'essentially random' given X, i.e.

$$\{Y^{(0)}, Y^{(1)}\} \perp\!\!\!\perp A|X.$$

These are considered sufficient to estimate the average treatment effect (ATE), defined below. Note that in the case of a randomised controlled trial (RCT), all the above apply except that the propensity function $\pi(x)$ is known (often constant) and unconfounded. In this case we get the two conditions above for free.

Average treatment effect

The average treatment effect (ATE) is defined via $ATE := \mathbb{E}[Y^{(1)} - Y^{(0)}]$. The assumptions described in the previous section are sufficient for the ATE to be identifiable (Stone, 1993), since

$$\begin{aligned}
ATE &= \mathbb{E}[Y^{(1)} - Y^{(0)}] \\
&= \mathbb{E}[\mathbb{E}[Y^{(1)} - Y^{(0)}|X]] && \text{(by the tower property)} \\
&= \mathbb{E}[\mathbb{E}[Y^{(1)}|X]] - \mathbb{E}[\mathbb{E}[Y^{(0)}|X]] \\
&= \mathbb{E}[\mathbb{E}[Y^{(1)}|X, A = 1]] - \mathbb{E}[\mathbb{E}[Y^{(0)}|X, A = 0]] && \text{(by Assumption 2)} \\
&= \mathbb{E}[\mathbb{E}[Y|X, A = 1] - \mathbb{E}[Y|X, A = 0]],
\end{aligned}$$

and X, Y and A are all observed quantities. Assumption 1 establishes that we can expect to observe samples with both $A = 0$ and $A = 1$ across the support of X, while Assumption 2 gives interchangeability of potential outcomes and observed outcomes conditioned on treatment and confounders. Methods for estimation of ATE have been proposed for a variety of modelling assumptions and contexts (Reiersöl, 1945; Thistlethwaite and Campbell, 1960; Rosenbaum and Rubin, 1983; Abadie, 2005; Craig *et al.*, 2017; Roth *et al.*, 2023). One common such method is propensity score matching, which relies on partitioning observations into groups

$y) := \mathbb{P}(Y = y|\text{do}(A = a))$.

with approximately equal propensity scores. Note that

$$\begin{aligned}
\mathbb{P}(A=1|\pi(X),Y^{(1)},Y^{(0)}) &= \mathbb{E}[A|\pi(X),Y^{(1)},Y^{(0)}] \\
&= \mathbb{E}[\mathbb{E}[A|X,Y^{(1)},Y^{(0)}]|\pi(X),Y^{(1)},Y^{(0)}] \\
&\qquad \text{(by the tower property)} \\
&= \mathbb{E}[\mathbb{E}[A|X]|\pi(X),Y^{(1)},Y^{(0)}] \\
&\qquad \text{(by Assumption 2)} \\
&= \mathbb{E}[\pi(X)|\pi(X),Y^{(1)},Y^{(0)}] \\
&\qquad \text{(by definition of propensity)} \\
&= \pi(X) = \mathbb{P}(A=1|\pi(X)),
\end{aligned}$$

so that $\{Y^{(0)},Y^{(1)}\} \perp\!\!\!\perp A|X \Rightarrow \{Y^{(0)},Y^{(1)}\} \perp\!\!\!\perp A|\pi(X)$ (this is due to Rosenbaum and Rubin 1983). We can then say that

$$ATE = \mathbb{E}[\mathbb{E}[Y|\pi(X),A=1] - \mathbb{E}[Y|\pi(X),A=0]].$$

To estimate the ATE we therefore need only to average over the marginal distribution of $\pi(X)$ rather than the entire marginal distribution of X, which can be particularly useful when X is high-dimensional. Methods for propensity score matching often begin by using some regression method to produce an estimate $\hat{\pi}$ of the propensity function.

Causal inference can be particularly useful when working with *observational* data, i.e. data for which we did not have direct experimental control over the treatment A. Situations where causal inference may be used include measuring the impact of administering a given drug from data in which the drug has been given to some patients but not others. Here we may not fully understand whether patterns exist in who has received the drug (e.g. those who are already very ill), and in estimating the ATE we aim to remove any bias in differences in outcomes (e.g. survival time) for those who were given or were not given the drug. While it is often useful to gauge the utility of a treatment across an entire population via the ATE, it does not necessarily make any progress towards recommending whether treatment will benefit a given patient. For this, we need heterogeneous, conditional, or individual treatment effects. This will be discussed at length in Chapter 4.

1.3 Genomic medicine questions in the language of statistical learning

> "You check the charts, and start
> to figure it out."
>
> LCD Soundsystem
> 'All My Friends' (2007)

At this point, we have covered most of the introductory molecular biology and statistical learning theory that will be necessary to understand later chapters. Here we aim to bring those two threads together by describing how, for each of a set of given application cases, we select a particular set of tools. The tools will be both biological and statistical, and we will describe our motivation in deploying them to solve or partially solve the problem at hand. Each will provide some motivational background to a particular chapter.

1.3.1 Signal and noise: understanding genomics technology

In our first application case we are confronted with the following setup: a gene expression measurement technology, known as LAMP, has been proposed as the only viable technology for transcript profiling in a given medical test. For any sample profiled by LAMP, we use a separate reaction to measure each transcript of interest. The raw data outputted from the LAMP reaction are noisy amplification curves that look fairly similar to PCR curves (see Figures 1.2 and 2.2). From each curve we want to be able to predict the true gene expression associated with that sample and transcript.

This forms the basis of a prediction problem, with input data given by the amplification curves themselves, and outputs given by the true gene expression values (for a training dataset, we have access to these via a gold standard reference technology). However, a few more factors are at play. Firstly, it is well known that LAMP has higher precision and resolution when profiling some transcripts than others. It will be important in the final implementation of the medical test which transcripts are chosen, and this decision will be made not just on the basis of their biological relevance, but also their amenabiliy to measurement with the LAMP technology. We therefore require a means of quantifying our uncertainty when using amplification curves to infer gene expression. We also have some mechanistic knowledge about typical amplification curves, and can quite easily construct models of the shape of these curves based on RNA abundance as an input. Therefore, a framework for predicting gene expression would ideally be based on the ability to invert such models.

We wish to perform inference in a manner that emphasises and enables the quantification of uncertainty; we have sufficient domain knowledge to select prior parameter distributions that match our understanding of what amplification curves should look like; and we wish in the end to be able to probabilistically invert a conditional model for amplification curves given true gene expression

into a conditional model for the opposite. All the considerations above point to using a Bayesian framework in the analysis of LAMP data. This is exactly what we do in Chapter 2, in which we do all of the analysis described, save for producing an inverted 'gene expression given amplification curve' model: while this aim was certainly the motivation, the work in this chapter is more exploratory, aiming to identify features of LAMP reactions that contribute to improved or worsened quantitative performance.

1.3.2 Gene panel selection: working to constraints

In Section 1.2.2, we discussed some of the terminology associated with high-dimensional statistics. An apt question is, when working with cancer genomics, in what manner of high-dimensional regime are we operating? This obviously depends on the specifics of the question we are trying to answer, the data we have available, and through what lens we view that data. To elaborate on this final point, we consider somatic mutation sequencing as described in Section 1.1.3 and ask how we may process it to produce 'tabular' data, for which we can think about values of n and p. Mutation data, typically stored in VCF or MAF files, contains a list of mutations, alongside information about which sample a mutation occurred in, where in the genome it occurred, and what form the mutation took. It is natural to consider the number of individual biological samples as n, but what about p? We could have one 'column' (i.e. a separate covariate of our input) devoted to each base in the genome or exome, recording whether or not a mutation was observed at that point. This data would be ultra-high dimensional – of the order $p \approx 3 \times 10^9$ for genome-wide sequencing, $p \approx 3 \times 10^6$ for exome-wide sequencing – to the extent that it would be almost impossible to use productively, with typical individual studies having on the order of magnitude of thousands of samples at best (TCGA contains around 2×10^4 samples in total across 33 cancer types; Cancer Genome Atlas Network, 2015). It is therefore often useful to take a 'gene-level' view. Here we might group exomic bases into their associated gene regions, and for each covariate column take the count of mutations in that gene. This leaves us with a resultant dimensionality of $p \approx 2 \times 10^4$, a much more manageable size.

Despite this reduced dimensionality, there are several reasons we might want to whittle down to an even smaller number as part of a prediction task (i.e. to perform variable selection). One is that the resultant selected genes might in themselves be of interest. This is one means of searching for driver genes, i.e. those that when mutated will elevate risk of the development, progression or adaptation of a tumour (Hanahan and Weinberg, 2000, 2011). Driver genes are thought to be relatively rare, and observant readers will note that this is exactly a sparsity assumption – a regularisation method such as the LASSO might be helpful.

Another justification for selecting some small set of genes/genomic loci for a prediction task is that the cost and time to perform sequencing depends (approximately linearly) on the size of the subsection of the genome to be sequenced, and the depth at which it is sequenced. This means that in many practical or clinical environments, cost is a major factor. If a future biomedical test is to be based on

sequencing then it may be essential that a concise targeted gene panel is selected, while maintaining enough accuracy that clinicians feel confident in acting upon its predictions. As mentioned above, sparsity may be well-motivated if we are predicting a phenomenon for which we can assume a small set of driver genes. But what if no such assumption is reasonable? What if we are aiming to predict an outcome that (by definition) is known to depend on all regions of the genome, and our need to select a concise set of representatives is purely practical? Will methods from high-dimensional statistics continue to be optimal? This is exactly the question investigated in Chapter 3, where we aim to develop concise gene panels for predicting tumour mutation burden (TMB) and tumour indel burden (TIB), each an exome-wide biomarker.

1.3.3 Assessing an intervention: getting the most out of data

While in this thesis we are not principally interested in analysing or accounting for missing data, a fascinating area of study in its own right, our analysis is still often shaped by what events we are and are not able to observe, and the impact this has on the data we do collect. When attempting to formulate a medical question such as "will a given patient benefit from treatment from a certain drug", we may fairly straightforwardly formulate this as a prediction task. From a historic dataset we gather some information about each patient, both clinical and genomic, and let this constitute our input values (x_i for patient i). We then observe how each patient responded (via an output y_i) and what treatment they were given (via a binary indicator a_i), and learn to predict y_i on the basis of x_i and a_i. As simple as this appears, however, reality has a way of muddying the waters. When investigating whether patients experience a survival benefit from a treatment, we are faced with two limitations to any observational dataset. Firstly, we will only observe for any patient in our historical dataset one outcome at most, any never what *would* have happened had their treatment decision been reversed. Secondly, when assessing survival outcomes we will be limited by the fact that not every patient will die within the timescale of a given study. This is obviously a very good thing, but means that for a substantial portion of our cohort we will not directly observed the outcome of interest, and this can severely impede our ability to fit models to predict it.

It should be of no surprise that, when considering unobserved alternate outcomes under a different treatment, we turn to causal inference. We will however be required to investigate several extensions to the base case laid out in Section 1.2.3, to accommodate the use of survival data and to enable answering a more personalised question than can be expressed by average treatment effect. This intersection of restrictions sits at the heart of an active field of research, and will be the focus of Chapter 4.

1.4 Summary and thesis outline

> "Now this is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning."
>
> WC (1942)

If this introduction has succeeded in its goals, it should have motivated why techniques from statistical and machine learning have value to add addressing genomics questions in medical applications. In particular, it should have provided the necessary biological and statistical background to understand each of the subsequent chapters. It should also have provided a 'high-level' view of why each technique we've selected from the statistical learning arsenal is well-matched to the application case in which we apply it. This motivation will be to some degree restated with each case study, but with more of a focus on providing specific technical context to the decisions made in each investigation.

A recurring theme in each chapter is that the complex structure of data presented in biological research can be a sticking point that, at its full potency, could severely undermine our capacity to answer the question we'd like, and to build the tools that we'd like. The restrictions we face when dealing with biological data can originate from the complexity of the underlying biological systems under investigation, or from external 'real-world' factors such as bounds on the cost and time available in order to make a solution fit for purpose.

Even at the current pace of increase of the availability of biological data (in particular from high-throughput sequencing studies), it remains to be seen whether the powerful and general machine learning techniques that have revolutionised other traditional learning problems will be at our disposal in a meaningful way. It is still firmly the case that specific biological and contextual expertise is necessary to optimally leverage the new wealth of data that is available to us. Therefore, to unlock the potential of that data, we need researchers who are able to speak the language of both statistical and biological 'camps'. It is not sufficient that researchers in cancer genomics provide data and questions to researchers in statistics/machine learning, nor that statistical researchers push forward with the development of methods without influence from the context of the problems they are attempting to address. Instead, methods need to be crafted bespokely by those who understand what features of biological data are relevant, how those features manifest themselves, and how to exploit them in a mathematically robust way.

As motivation for the challenges discussed above, it should go without saying that medical genomics in the machine learning age has potential to do a great deal of good in the long term. Yet uncovering a deeper understanding of how diseases work isn't the only worthwhile goal. Designing procedures that can work *now* to be more effective, sometimes crossing a threshold between non-pracitcality and practicality (embedded in some particular context), can have a more immediate benefit. In the clinic, the time scale and cost of data collection are not abstract

mathematical problems, so designing a predictive model that maximally leverages the data available (while acknowledging that there is never a guarantee that any particular dataset contains the information necessary to answer a particular question) can be just as enabling as uncovering a new paradigm of disease progression.

To wrap up this introduction, we'll now briefly (and with some repetition from what's come before) outline the studies discussed in each subsequent chapter. In Chapter 2 we present new analysis of the loop-mediated isothermal amplification (LAMP) assay on clinical and synthetic samples, motivated by the unmet need for statistcally principled methods for guided LAMP optimisation. To do this, we'll use Bayesian methods for prediction, where the target of prediction is not the medical outcome in question but the technical accuracy of an assay based on a given biological target. We continue in Chapter 3 with another study into the optimal, data-driven design of a biomedical test, albeit in the cancer setting. Here, we are concerned with the prediction of tumour mutation burden (TMB), a key clinical biomarker determing how likely cancer patients are to respond to immunotherapy. Our task is to select a small number of gene targets to form a targeted DNA sequencing panel from which to estimate TMB. We'll do this with an extension of LASSO-based regularisation. Finally, in Chapter 4, we extend the previous chapter's work, and investigate the extent to which panel-based genomic markers can be tailored to identify heterogeneous causal effects in immunotherapy response.

Chapter 2

Bayesian analysis for optimising LAMP gene expression assays

Overview

Loop-mediated isothermal amplification (LAMP) is a fast and cost-effective technique for detection of DNA/RNA. Its lack of reliance on complex laboratory equipment makes it ideal for use in point-of-care biomedical applications such as the diagnosis of time-critical disease. In this chapter we investigate the use of LAMP for profiling gene expression by measuring mRNA in blood samples. This is used in a test to establish whether patients with acute inflammation are bacterially or virally infected.

While LAMP has many appealing properties, it remains unclear what factors affect its quantitative resolution. A lack of model-based frameworks to characterise LAMP data presents an unmet need in enabling the development and use of the assay. At present each of the assay's multiple primers are typically optimised experimentally, presenting a major bottleneck in assay design. We present hierarchical Bayesian models of LAMP amplification based on Gompertz functions, and use these models to infer the effect of RNA variation and other factors on LAMP amplification curves derived from $\sim$100 blood samples of patients with suspected acute infection.

Our analysis uncovers associations between LAMP assay resolution and characteristics such as primer sequence composition and thermodynamic properties. In addition to correlations between RNA input abundance and time shift of the LAMP amplification curve, we also detect RNA-dependent assocations with amplification rate. We further investigate associations between primer/target properties and quantitative performance of the assay by generating a set of synthetic RNA samples with systematically varied primer sequences and applying our framework. We find evidence that the associations observed are driven by across-target rather than within-target variation. This has important implications for study design, where in the past similar analyses have relied upon datasets including only one primer configuration per target. Our findings represent first steps towards data-driven development of quantitative assays, a key need in enabling LAMP's use in the range of application spaces for which it shows promise.

2.1 Introduction

Loop-mediated isothermal amplification (LAMP) is a fast, sensitive, and precise method for detecting DNA or RNA, the latter case referred to as reverse-transcription (RT)-LAMP. Since its introduction (Notomi *et al.*, 2000), LAMP has been used for pathogen detection (Mekata *et al.*, 2009; Cao *et al.*, 2017; Thiessen *et al.*, 2018), sex identification (Hirayama *et al.*, 2013; Almasi and Almasi, 2017; Centeno-Cuadros *et al.*, 2018), and cancer monitoring (Li *et al.*, 2016; Horiuchi *et al.*, 2020; Kalofonou *et al.*, 2020). Unlike comparable assays such as PCR, LAMP is isothermal and therefore does not require a thermocycler. This makes the assay particularly amenable to point-of-care and field applications (Fu *et al.*, 2011). In particular, we study the use of LAMP in diagnostic testing for time-critical acute medicine, a setting in which lab-based PCR is unviable.

We begin by introducing some terminology. We refer to a *sample* as the material to be profiled for a given patient, in our case from a blood sample. For each sample, we aim to profile the expression of multiple RNA sequences, referred to as *targets*. Each target will typically be chosen to represent a particular gene of interest, but could be any sequence. The LAMP assay comprises an amplification reaction initiated by the target sequence and whose progress can be measured for detection/quantitation. The reaction in question relies on several *primers*. These are oligonucleotides (short sequences of DNA) that are complementary to subsequences of the target and bind to them by *hybridisation*. Once these primers have bound to the target or to subsequent products, they act as initation points for polymerisation, leading to the production of more nucelotide sequences which fold and separate to form an *amplicon*. This amplicon then serves as the basis for further amplification steps. The components of the reaction are chemically tagged such that amplification can be measured via *fluorescence*. By this we mean that the reaction emits light that can be measured and forms the basis for the data we aim to analyse.

As in PCR, LAMP specificity relies on primer-target complementarity. In its original conception, two pairs of primers (referred to as F3/B3 and FIP/BIP) were proposed, and further optional pairs of primers were introduced for loop hybridisation (Nagamine *et al.*, 2002) and amplicon enlargement (Gandelman *et al.*, 2011). In the experiments discussed here, six primers are used: the core pairs F3/B3 and FIP/BIP, and a loop hybridisation pair LF/LB. A technical overview of the LAMP reaction is given in Figure 2.1. It is not necessary to understand every step of the reaction to follow the analysis presented in this chapter; readers need only follow that the LAMP assay requires the design of six primers. The need for coordinated design of multiple primers can create a major bottleneck in LAMP assay development.

While LAMP has primarily been used in qualitative/detection applications, advances in measurement technology (Zhang *et al.*, 2014; Becherer *et al.*, 2020), in particular automated fluorescence detection for producing LAMP amplification curves, have enabled quantitative use of the assay. Previous work has extended this approach to measurement of mRNA in a given sample, a technique referred to throughout as quantitative reverse-transcription loop-mediated isothermal amplification (qRT-LAMP) (Remmel *et al.*, 2022). Quantification for qRT-LAMP

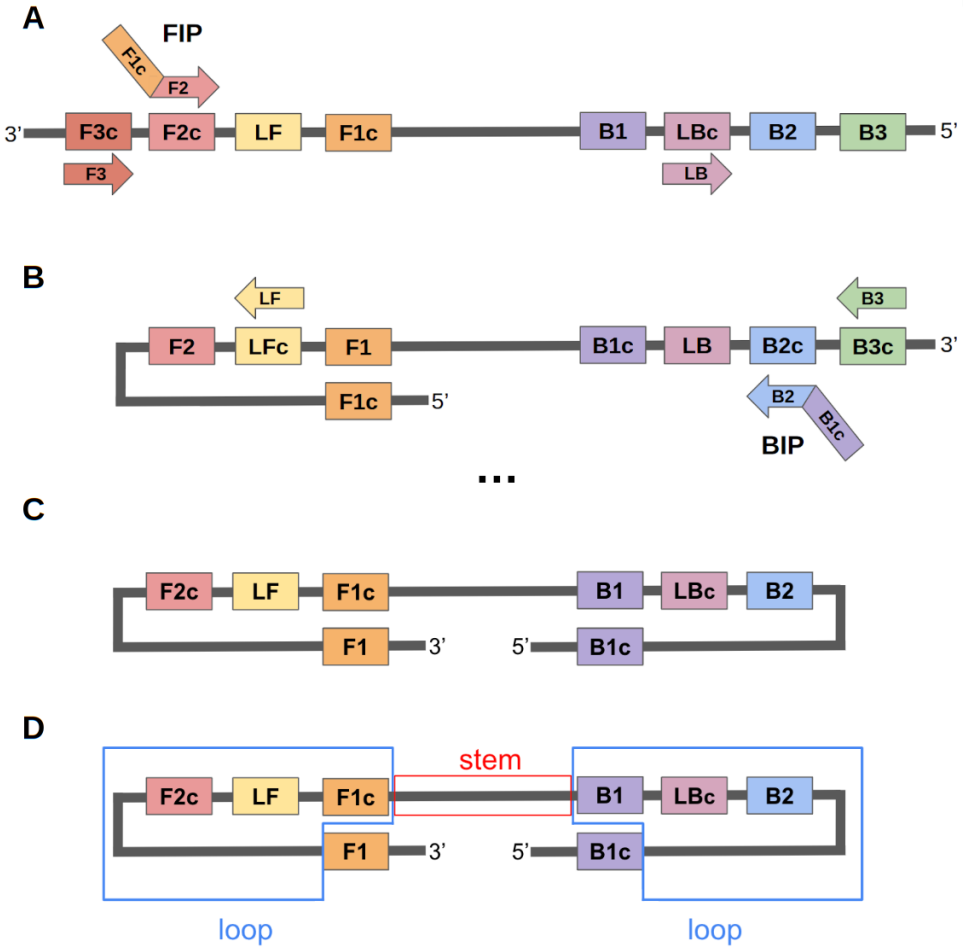

Figure 2.1: Intermediate steps in the qRT-LAMP reaction. **A**: Following reverse transcription, any of the primers F3, FIP (comprising F2 and F1c sequences), and LB may anneal to the cDNA template. Polymerisation from the bound FIP primer leads to introduction of the F1c sequence into the product. **B**: Annealing of the F1c sequence to its complement creates a loop. The primers B3, BIP (comprising B2 and B1c sequences), and LF anneal to this single-loop template. **C**: Annealing of the B1c sequence to its complement creates another loop, resulting in the canonical 'dumbbell' LAMP template, or amplicon. **D**: The stem region of the dumbbell template (the span between F1c and B1 sequence regions) contains the target sequence of interest while the loop regions (the regions spanning from F1c/B1 inclusive to F1/B1c exclusive) provide additional sites to initiate polymerisation. Crucially, any of the six primers may anneal and initiate polymerisation at any point at which a complementary region is available, leading to creation of other products (not shown for simplicity) distinct from the canonical product.

requires analysis of the same data (one amplification curve per sample per target), but rather than detecting simply whether a reaction has initiated aims to extract quantitative data reflective of the input RNA abundance.

Current understanding of the ideal properties of primers and targets to optimise qRT-LAMP's performance is incomplete and qualitative; primer design relies on a small set of guiding principles (Panno *et al.*, 2020), including restrictions on the GC nucleotide content, melting temperature, and presence of GG repeats in primer sequences. In practice, primer design proceeds in trial-and-error fashion, as these guidelines can often be in conflict and are far easier to simultaneously satisfy for some RNA targets than others. Furthermore, there has been little work to assess the impact on performance of sequence-based and thermodynamic properties of the LAMP primers and amplicon.

Statistical modelling frameworks provide one approach to link assay properties and characteristics of amplification curves. Extensive work exists for modelling both PCR amplification curves (Spiess *et al.*, 2008; Matz *et al.*, 2013; Subramanian and Gomez, 2014; Nguyen *et al.*, 2020) and the effect of primer/target properties on PCR assay performance (Mallona *et al.*, 2011; Wright *et al.*, 2014; Döring *et al.*, 2019). Previous work on primer design for PCR has identified thermodynamic properties of primer/target interactions (Mann *et al.*, 2009; Li and Brownley, 2010; Döring *et al.*, 2019) and sequence length (Huang *et al.*, 2022) as key quantities associated with assay specificity. Thermodynamic properties (e.g. free energy of primer-target hybridisation) have also been associated with the speed of isothermal reactions (Kimura *et al.*, 2011). In contrast, the study of LAMP-based quantitation is still under active development, and methods for qRT-LAMP data analysis are lacking.

Gompertz functions (Gompertz, 1825) have proved a useful tool for empirical modelling of growth curves in varied biological settings (Tjørve and Tjørve, 2017). One formulation of the Gompertz curve is parameterised by an 'amplitude' (the maximum value reached at asymptote), a 'rate', and an 'offset' (the horizontal shift of a curve along its time axis). Gompertz functions are similar to logistic functions, but allow for steeper increases from baseline and slower approaches to asymptote than logistic functions. In Section 2.2.2 we define Gompertz functions formally (Equation 2.1) and give some examples alongside real amplification curves to demonstrate their properties (Figure 2.2). Bayesian and hierarchical Bayesian formulations of Gompertz curves have appeared in biomedical applications (Wiper *et al.*, 2010; Demirhan and Ata Tutkun, 2015; Sasaki and Kondo, 2016; Gotuzzo *et al.*, 2019; Vaghi *et al.*, 2020; Berihuete *et al.*, 2021) and have started to appear in LAMP reaction modelling for detection (Carvalho *et al.*, 2021). However, no Bayesian or hierarchical methods have yet been applied to characterise qRT-LAMP dynamics or to model the effects of different properties of LAMP targets/primers on quantitative assay performance.

Here we present a model-based analysis of the quantitative characteristics of the LAMP assay. We first develop a Bayesian modelling framework for inferring the impact of changes in RNA input on amplification rate and offset, and we apply this framework to a dataset of qRT-LAMP curves derived from 109 patients suspected of acute infection. Then, we compile a collection of primer/target features, and extend our hierarchical framework to detect associations between

these features and resolution (or the ability to detect differences in RNA input) of the qRT-LAMP assay. In order to further investigate the effects of primer/-target features on assay resolution, we utilise a novel dataset from synthetic *in vitro*-transcribed (IVT) RNA templates assayed with systematically varied LAMP primer sequences. This dataset allows us to apply our framework in a context where we apply multiple primer sets to each target, and in so doing more directly interrogate the effects of changes to primer design. Our model-based analysis highlights several key considerations in the design and optimisation of the qRT-LAMP assay, including the use of amplification rate for quantitative inference and the inclusion of thermodynamic properties into decisions around primer design. We provide a description of our implementation of this work via the R package `LAMPPrimerFeatures` and a `targets` analysis workflow in Appendix A.

2.2 Methods

2.2.1 Datasets and technologies

We utilised two datasets, one derived from clinical samples and one from synthetic IVT RNA templates. We summarise the data availability for each of the two experimental setups in Table 2.1.

Our clinical dataset comprised blood samples collected from 109 patients, each adjudicated to be either bacterially infected, virally infected, or non-infected (with these three labels occurring in approximately equal proportions). For each sample, we analysed gene expression measured from 28 target RNA transcripts: 25 genes previously identified as targets of interest for bacterial/viral/non-infected classification (He *et al.*, 2021) as well as 3 *housekeeping* genes (KPNA6, RREB1, and YWHAB). By a housekeeping gene, we mean one whose expression is not expected to vary substantially between samples, and so can act as reference for normalisation. We tracked fluorescence for 20 minutes or 60 'cycles' (1 cycle = 20 seconds)[1]. For each sample/target, two technical replicates were performing using separate subportions (*aliquots*) of the given sample. Another aliquot was used for mRNA quantification using the NanoString (NS) nCounter system. The NS nCounter system directly counts RNA transcript number in a single step by measurement of fluorescence via hybridisation of capture and reporter probes. As such, NS measurements are a singular value (unlike other comparison technologies such as qPCR) and very sensitive even at low expression levels (Geiss *et al.*, 2008). While NS is unsuitable for point-of-care usage, it is an ideal technology to provide 'gold standard' benchmark measurements. Higher NS values correspond to higher target expression.

Our IVT RNA dataset (preparation described in Remmel *et al.*, 2022) comprised measurements for three targets: IFI27, JUP, and OASL. This dataset was generated by assaying different concentrations of synthetic RNA templates corresponding to each of the the targets. To test the effects of varying different primer properties on qRT-LAMP performance, FIP and BIP primer sequences were varied for each target. These sequence alterations consisted of variations in the

[1]Note that these are not strictly cycles since the LAMP reaction is isothermal, but the cycle nomenclature is still common in the LAMP literature.

Summary	Clinical Dataset	IVT Dataset
Number of samples	109	3
Number of targets (non-housekeeper)	25	3
Number of targets (housekeeper)	3	0
Replicates per sample/target (qRT-LAMP)	2	≥ 2
Primer sets per sample/target	1	35-63
Replicates per sample/target (NS)	1	0
Cycles per run	60	90

Table 2.1: Specification of clinical and IVT datasets.

lengths of annealing sequence for either the F1c (B1c) or F2 (B2) subsequences of the FIP (BIP) primer (Figure 2.1). F3, B3, LF, and LB primer sequences were not varied. For each target/primer set combination, qRT-LAMP measurements at three RNA concentrations (10^1, 10^3, and 10^5 copies/μL; 20μL per reaction) were collected, with at least two replicates per concentration. We investigated effects of changes in FIP/BIP primer sequence in 35, 48 and 63 primer combinations for IFI27, JUP, and OASL respectively. For the IVT samples, fluorescence was tracked for 30 minutes, or 90 cycles.

We now describe technical aspects of sample preparation and collection for each of the datasets used. Blood samples from our clinical dataset were collected using the PAXgene RNA system, as described in Ram-Mohan *et al.* (2022). These patients samples were drawn from patients involved in two clinical trials: "HostDx Sepsis in the Diagnosis and Prognosis of Emergency Department Patients With Suspected Infections: a Multicenter Pilot Study" (clinicaltrial.gov/NCT03744741), and "Efficiency in Management of Organ Dysfunction Associated with Infection by the Novel SARS-CoV-2 Virus Through a Personalised Immunotherapy Approach: The ESCAPE Clinical Trial: The 'ESCAPE' Trial (a.k.a ESCAPE)" (EudraCT/2020-001039-29). At enrollment, patient blood was collected and immediately stored with PAXgene RNA stabilisation. Two independent aliquots of each sample underwent RNA extraction for profiling by qRT-LAMP and were dispensed along with primers and reagents and run on a 96-well plate with a QuantStudio™ 5 Real-Time PCR System for Human Identification (Applied Biosystems; Waltham, MA USA). We normalised NS measurements by first rescaling counts such that housekeeping genes had a geometric mean of 1000 and then applying a $\log_2$ transformation to the rescaled counts. To measure mRNA expression with qRT-LAMP, a set of six LAMP primers (FIP/BIP, F3/B3, and LF/LB; Figure 2.1) for each target were designed using PrimerExplorer v5 (Eiken, 2019) or via a third-party service provider (Primer Digital, Ltd; Helsinki, Finland). QRT-LAMP reactions occurred at 65°C, in the presence of 90mM KCl and 8mM $MgSO_4$. Originally 29 non-housekeeper targets were measured, but we excluded 4 for technical reasons: FURIN and S100A12 due to the use of different primer concentrations; CTSB and DEFA4 as LF/LB

primers did not anneal to their target sites. Reaction conditions for qRT-LAMP with IVT RNA templates were identical to those used for the clinical samples. Replicate samples were dispensed along with primers and reagents and run on a 384-well plate with the QuantStudio™ 5 instrument.

2.2.2 Amplification curve pre-processing and normalisation

To prepare qRT-LAMP curves for analysis, we performed a series of pre-processing steps. We first applied baseline correction to set the initial fluorescence of each reaction approximately to zero. We then used a nonlinear least-squares approach to fit Gompertz functions to each amplification curve. The Gompertz function is parameterised by three quantities (a, b, and c in Equation 2.1) associated with the amplitude, rate, and cycle offset, respectively, of the amplification curve. One can think of the parameter c as a time-to-threshold measurement (as used for qPCR) in which the threshold is specific to each reaction curve. Higher cycle offset (a right-shifted curve) would indicate lower amounts of RNA, and vice versa. The Gompertz function is defined as follows:

$$y = a \exp\Big(- \exp\big(- b(t - c)\big)\Big). \tag{2.1}$$

Here, y corresponds to baseline-corrected fluorescence and t is time in cycles. Our procedure estimates parameters $(\hat{a}^{\text{gomp}}, \hat{b}^{\text{gomp}}, \hat{c}^{\text{gomp}})$ for a given amplification curve. Anecdotally and experimentally, LAMP amplification curves often become very noisy around asymptote, and for reasons of technical calibration the absolute level of asymptotic fluorescence (a) is relatively unimportant (as with qRT-PCR) to quantitation compared to cycle offset (c). The impact of input RNA abundance on amplification rate (b) is, by contrast, fairly unexplored. Motivated by this and after noting some misfit to the asymptote in early versions of our curve-fitting procedure, we developed and applied an iterative procedure, described in Algorithm 1, to remove observations exceeding the fitted asymptote at each iteration. We then retained the fitted asymptote ($\hat{a}^{\text{gomp}}$) at the last iteration (when no points of the curve exceeded the asymptote) to rescale the curve. A selection of example outputs from this procedure appear in Figure 2.2.

2.2.3 Single-target model

To investigate the effects of RNA abundance on qRT-LAMP amplification curves for individual targets, we model the Gompertz curve parameter estimates produced by our nonlinear fitting procedure. For samples $i = 1, \ldots, N$, replicates $j = 1, \ldots, J$, with NS-measured gene expression x_i (a proxy for true RNA input in clinical samples), we model estimated parameters for rate ($\hat{b}_{ij}^{\text{gomp}}$) and cycle offset ($\hat{c}_{ij}^{\text{gomp}}$) of an individual target as realisations of independent random variables

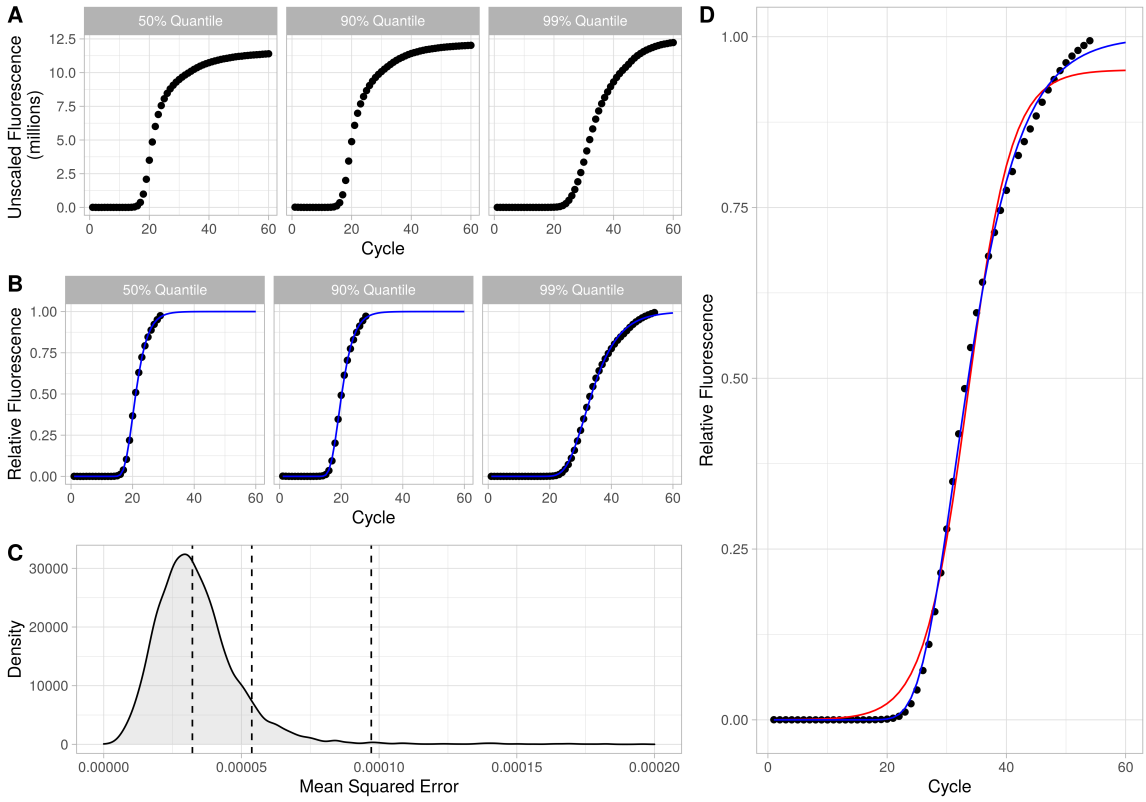

Figure 2.2: The normalisation procedure described in 2.2.2, with examples from our clinical dataset. **A**: Baseline-corrected LAMP curves with points corresponding to individual fluorescence measurements (arbitrary units) over 20s cycles. **B**: Rescaled and pruned LAMP amplification curves. Points show scaled measurements while solid blue lines correspond to Gompertz curve fits from our iterative procedure. **C**: Distribution of goodness-of-fit (measured by mean squared error) for our fitted Gompertz curves, with 50%, 90% and 99% quantiles indicated by dashed vertical lines. Representative fits from these quantiles are shown in A and B. **D**: Reproduction of the third figure of panel B, with a logistic curve fit for comparison (red).

Algorithm 1: Iterative least-squares Gompertz curve-fitting procedure

Data: The set $\{(t, D_t) : t = 1, \ldots, T\}$, where T is the number of cycles observed and D_t is the fluorescence measured by the qRT-LAMP assay at cycle t.

Output: A triple $(\hat{a}^{\text{gomp}}, \hat{b}^{\text{gomp}}, \hat{c}^{\text{gomp}})$ of Gompertz parameter estimates, and a pruned set $\{(t, D_t) : t = 1, \ldots, n^*\}$.

Set $n = T$;
Set $\delta = 0$;
while $(\delta \geq 0)$ **do**
  Set $(\hat{a}, \hat{b}, \hat{c}) = \underset{(a,b,c)\in\mathbb{R}^3}{\arg\min} \left\{ \sum_{t=1}^{n} \left(D_t - a \exp\left(- \exp\left(- b(t-c) \right) \right) \right)^2 \right\}$;
  Set $\delta = \max\{D_t : \ t \in \{1, \ldots, n\}\} > \hat{a}$;
  Set $n = \min\{t : D_{t+1} > \hat{a}\}$;
end
Set $(\hat{a}^{\text{gomp}}, \hat{b}^{\text{gomp}}, \hat{c}^{\text{gomp}}) = (\hat{a}, \hat{b}, \hat{c})$;
Set $n^* = n$.

$\hat{B}_{ij}$ and $\hat{C}_{ij}$ distributed according to:

$$\hat{B}_{ij} \sim \text{Student-t}(\mu_B + \lambda_B x_i + \rho_B \hat{h}_i^{\text{gomp}}, \ \nu_B, \ \sigma_B), \tag{2.2}$$

$$\hat{C}_{ij} \sim \text{Student-t}(\mu_C + \lambda_C x_i + \rho_C \hat{h}_i^{\text{gomp}}, \ \nu_C, \ \sigma_C). \tag{2.3}$$

Probabilistic graphical model (PGM) plate diagrams for these models are given in Figure 2.3. We use a $\text{Student-t}(\mu, \nu, \sigma)$ likelihood to provide robustness to outliers, with μ, ν, and σ corresponding to the location, shape, and scale of the distribution respectively. We refer to λ_B and λ_C as resolution parameters since a larger absolute value of λ_B or λ_C would reflect greater sensitivity, on average, to changes in RNA input for a given target. Note that these properties are with respect to absolute changes in RNA input, not changes in RNA input relative to the overall variation of a given target. We are therefore specifically attempting to measure LAMP's technical performance measuring a biochemical target, not the total discriminatory power of the range of a target's expression. When, for example, selecting targets for a diagnostic test, other considerations such as clinical significance and variation in expression will have to be weighed against technical resolution of the assay.

The input $\hat{h}_i^{\text{gomp}}$ is the geometric mean of the values $\hat{c}_{ijk}^{\text{gomp}}$ across all replicates j and all housekeeping genes $k \in \{\text{KPNA6, RREB1, YWHAB}\}$, included to normalise for sample-to-sample variation in RNA yield. The parameters σ_B and σ_C capture variation in rate and cycle offset that is not attributable to variation in RNA input. Bayesian inference is performed with the R package `brms` (Bürkner, 2017). We specify priors as follows. For the rate model (eq. 2.2): $\mu_B \sim \mathcal{N}(0.5, 0.5^2)$, $\lambda_B \sim \mathcal{N}(0, 0.1^2)$, $\rho_B \sim \mathcal{N}(0, 0.1^2)$, $\nu_B \sim \Gamma(2, 0, 1)$, $\sigma_B \sim \text{Student-t}(3, 0, 2.5^2)$; for the cycle offset model (eq. 2.3): $\mu_C \sim \mathcal{N}(30, 5^2)$, $\lambda_C \sim \mathcal{N}(0, 1^2)$, $\rho_C \sim \mathcal{N}(0, 1^2)$, and identical priors to the rate model for ν_C, σ_C. We

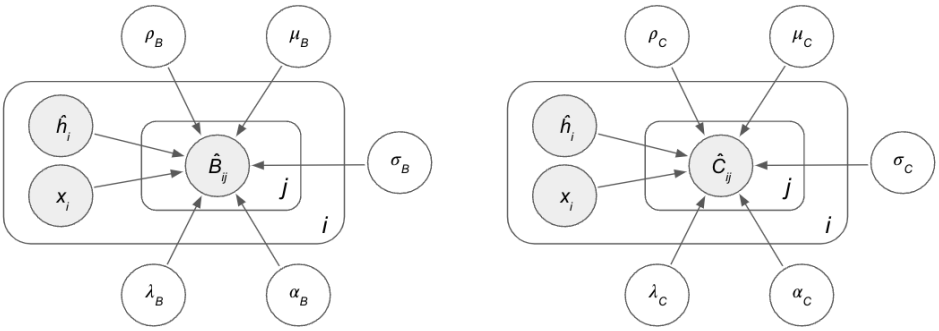

Figure 2.3: PGM plate diagrams for single-target models of $\hat{B}$ (left) and $\hat{C}$ (right).

approximate model posteriors with 4000 samples (1000 warm-up) from 3 HMC chains, assessed for convergence with all parameter posteriors achieving $\hat{R} \leq 1.01$ and bulk and tail effective sample sizes above 1000.

2.2.4 Multi-target model

To evaluate associations between primer/target properties and assay resolution, and to model qRT-LAMP amplification curves across all targets, we propose hierarchical models of the estimated rate and cycle offset. We derive estimates of the rate $\hat{b}_{ijkl}^{\text{gomp}}$ and offset $\hat{c}_{ijkl}^{\text{gomp}}$ for sample i, replicate j, target k, and primer set l (only varied in the case of IVT RNA templates) from our non-linear least squares fitting procedure. We then model each of the $\hat{b}_{ijkl}^{\text{gomp}}$ and $\hat{c}_{ijkl}^{\text{gomp}}$ as being independent realisations of random variables $\hat{B}_{ijkl}$, $\hat{C}_{ijkl}$, where:

$$\hat{B}_{ijkl} \sim \text{Student-t}(\mu_B(\boldsymbol{z}_{kl};\boldsymbol{\beta}_B) + \lambda_B(\boldsymbol{z}_{kl};\boldsymbol{\gamma}_B)x_i + \rho_B\hat{h}_i^{\text{gomp}},\ \nu_B,\ \sigma_B), \tag{2.4}$$

$$\hat{C}_{ijkl} \sim \text{Student-t}(\mu_C(\boldsymbol{z}_{kl};\boldsymbol{\beta}_C) + \lambda_C(\boldsymbol{z}_{kl};\boldsymbol{\gamma}_C)x_i + \rho_C\hat{h}_i^{\text{gomp}},\ \nu_C,\ \sigma_C). \tag{2.5}$$

Here, for each model the location of the likelihood comprises RNA-independent baseline offset μ_B ($/\mu_C$) and RNA-dependent resolution λ_B ($/\lambda_C$) functions, parameterised by $\boldsymbol{\beta}_B$ ($/\boldsymbol{\beta}_C$) and $\boldsymbol{\gamma}_B$ ($/\boldsymbol{\gamma}_C$), respectively. These functions (which we will refer to as μ and λ when speaking interchangeably about either model) take as an input a vector of sequence-based and thermodynamic features, $\boldsymbol{z}_{kl}$, associated with the given primer set l and target k. In our clinical dataset, $\boldsymbol{z}$ has no dependence on l as each target corresponds to a single primer set. Also, for our clinical dataset, we let baseline offset μ vary across targets k by specifying it as a hierarchical random effects term. We also remove dependence of μ on $\boldsymbol{z}_k$:

$$\mu(\boldsymbol{z}_k;\boldsymbol{\beta}) = \beta_{0,k}, \tag{2.6}$$

$$\beta_{0,k} \sim \mathcal{N}(\mu_{\beta_0}, \sigma_{\beta_0}^2). \tag{2.7}$$

This random term models target-specific effects unrelated to assay characteristics. Our models of baseline offset in clinical samples did not have slopes (β_z) varying by target as we had only one primer set per target.

For our IVT RNA dataset, we note that: *a)* x_i represents the log RNA copy

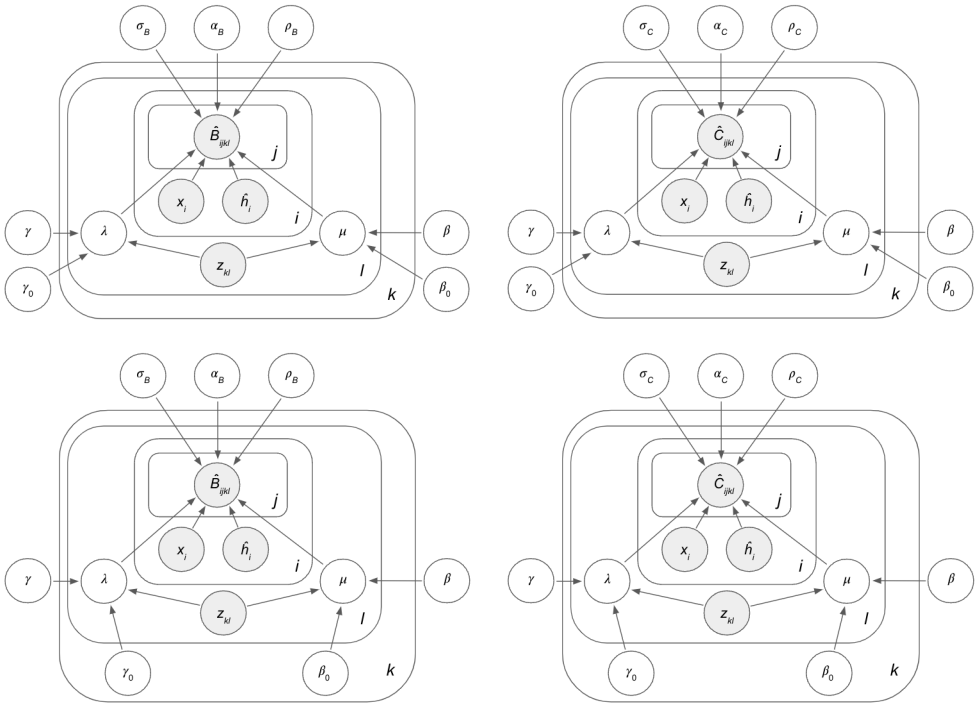

Figure 2.4: PGM plate diagrams for $\hat{B}$ (left) and $\hat{C}$ (right). Specifications are given for multi-target models with across-target resolution functions (top row) and within-target resolution functions (bottom row).

number at concentration i, and *b)* we drop the housekeeping term ($\rho\hat{h}_i^{\text{comp}}$) as input RNA concentrations are known. For both our IVT RNA and clinical data, we use λ to model effects of assay features on (RNA-dependent) assay resolution (e.g. $\lambda(\boldsymbol{z}_{kl};\gamma) = \gamma_0 + \gamma_{L_{\text{stem}}} L_{\text{stem},k,l}$).

For analysis of our IVT RNA dataset, we specify two types of models, which we refer to as *across-target* and *within-target* models respectively. For both model types, the RNA-independent offset function, μ, includes random intercepts and random slopes to model dependence on assay properties. For example, when using stem length as an input: $\mu(\boldsymbol{z}_{kl};\boldsymbol{\beta}) := \beta_{0,k} + \beta_{L_{\text{stem},k}} L_{\text{stem},k,l}$. This is to allow maximum flexibility in the RNA-independent offset to focus on changes in resolution (via λ) due to primer set features. The form of the resolution function λ for our *across-target* model is identical to that applied to clinical data, while for the *within-target* model, we include random target-specific intercepts in the resolution function λ. Again, using stem length as an example, the *within-target* model λ function would be:

$$\lambda(\boldsymbol{z}_{kl};\boldsymbol{\gamma}) = \gamma_{0,k} + \gamma_{L_{\text{stem}}} L_{\text{stem},k,l}, \tag{2.8}$$

$$\gamma_{0,k} \sim \mathcal{N}(\mu_{\gamma_0}, \sigma^2_{\gamma_0}). \tag{2.9}$$

We give PGM plate diagrams for both across- and within-target models for $\hat{B}$ and $\hat{C}$ in Figure 2.4. Note that the only difference is whether the parameter γ_0 is 'inside' the plate associated with target k, i.e. is allowed to vary between targets. When the λ function parameter $\boldsymbol{\gamma}$ is a 'global' parameter (i.e. not specified

Property type	Property	Symbol	Unit
Amplicon-specific	Stem length	L_s^{amp}	bp
	Loop length	L_l^{amp}	bp
	GC Content	GC^{amp}	-
Specific to primer p	4-complexity	C^p	-
	Free energy of annealing	ΔG_a^p	kJmol^{-1}
	Free energy of self-dimerisation	ΔG_{sd}^p	kJmol^{-1}
	Free energy of secondary structure formation	ΔG_{ss}^p	kJmol^{-1}

Table 2.2: Properties of targets/primers used in hierarchical modelling of LAMP performance. Here $p \in \{F3, B3, FIP, BIP, LF, LB\}$.

as a hierarchical term), we use independent $\mathcal{N}(0, 2^2)$ priors for its component parameters. When $\boldsymbol{\gamma}$ includes hierarchical terms for the intercept, we specify priors $\mu_{\gamma_0} \sim \mathcal{N}(0, 2^2)$, $\sigma_{\gamma_0} \sim \text{Student-t}(3, 0, 2.5^2)$. Priors for the baseline offset parameters $\boldsymbol{\beta}$ are set similarly, aside from the intercept term β_0: $\mu_{\beta_0} \sim \mathcal{N}(1, 1^2)$ for our rate model and $\mu_{\beta_0} \sim \mathcal{N}(23, 5^2)$ for our cycle offset model. In each case the parameters ν and σ are given priors $\nu \sim \Gamma(2, 0.1), \sigma \sim \text{Student-t}(3, 0, 2.5^2)$. For multi-target model parameter inference, we use 3 HMC chains of 4000 samples (1000 warm-up), again assessing convergence by confirming that all chains have $\hat{R} \leq 1.01$ and effective bulk and tail sample sizes above 1000.

2.2.5 Properties of primers and targets

Multiple properties are known to impact efficacy of LAMP primers, including melting temperature, GC content and GG repeat abundance (Eiken, 2019). In practice, many of these properties are tightly controlled in primer design workflows, resulting in little variability with respect to these quantities. We therefore focused on three classes of LAMP properties: *a)* sequence features of the LAMP target or amplicon; *b)* thermodynamic properties of primers and their targets; and *c)* measures of primer sequence complexity. The full list of properties appears in Table 2.2. Stem length (L_{stem}) and loop length (L_{loop}) measure the total length in nucleotide bases of the amplicon stem and loop regions, respectively (Figure 2.1D). GC content (GC) is the proportion of bases of the entire amplification region (F3 to B3, inclusive) that are either G or C.

Each free energy property measures the strength of some binding process via Gibbs free energy. Free energy of annealing (ΔG_a^p) measures the binding affinity of a primer with its complementary target region, while free energy of self-dimerisation (ΔG_{sd}^p) quantifies a primer's propensity to anneal to another copy of the same primer. Finally, ΔG_{ss}^p measures a single primer's tendency to form secondary structure within itself. We follow the approach of Döring *et al.* (2019), using UNAFold (Markham and Zuker, 2008) for free energy calculations and applying ion concentration adjustments as described in Peyret (2000). For FIP (BIP) primers, we calculated separate free energies of annealing for F1c (B1c) and F2 (B2) primer subsequences.

To measure the extent to which a primer sequence contains unexpected k-mer sequences (i.e. sequences of length k) compared to a reference sequence, we also computed sequence complexity (C^p) (Nielsen *et al.*, 2003; Xia *et al.*, 2010). This measure is based on the information-theoretic notion of entropy (Shannon, 1948). We measure complexity with respect to 4-mers and use the entire human exome as contained in the *Ensembl* database (Yates *et al.*, 2020) as our set of reference sequences. We use a simplified formulation of 4-complexity:

$$C_4^p := 1 - \frac{\sum_{\ell=1}^{L-3} n_4^p(\ell) \log(n_4^p(\ell))}{Z}.$$

Here, $n_4^p(\ell)$ refers to the count of 4-mers found in the reference sequence set that match the 4-mer in primer p's sequence starting at position ℓ. The normalising constant Z is set to ensure that complexity varies between 0 and 1. A low complexity score indicates that the majority of a primer's constituent k-mers occur commonly in the reference set.

2.2.6 Bayesian stacking for model comparison and validation

To compare multi-target models and identify which LAMP properties are most robustly associated with resolution, we use Bayesian model stacking based on Pareto-smoothed importance sampling approximations of each model's performance in leave-one-out cross-validation (LOO-PSIS; implemented by the R package `loo`; Vehtari *et al.*, 2020).

In traditional model stacking, predictions from multiple models are ensembled via a linear combination in order to minimise the average cross-validation prediction error (Wolpert, 1992). The same process is applied in Bayesian model stacking, but rather than using the squared error of models' predictions, one calculates log predictive densities for left-out observed values (i.e. those allocated to a validataion set; Yao *et al.*, 2018). The outputs of stacking are model weights associated with each individual model in the ensemble. These weights will sum to one, and a larger weight indicates a more substantial contribution for the corresponding model (relative to other models in the ensemble) to the optimal model combination. Thus, we take large stacking weights to be indicative of stronger associations of the corresponding model's features with qRT-LAMP assay resolution. We use Pareto-smoothed importance sampling to approximate (Vehtari *et al.*, 2017, 2021) the log posterior predictive density of each left out sample. Pareto k diagnostic values > 0.5 would indicate a poor approximation of the LOO posterior predictive density for some observations. We did not observe any such problematic Pareto k-estimates in our analyses.

2.3 Results

2.3.1 RNA dependence for qRT-LAMP cycle offset and amplification rate varies widely between targets

Our first aim was to infer the resolution (λ_B, λ_C) and precision (σ_B, σ_C) parameters associated with RNA input's effect on qRT-LAMP curve characteristics in clinical samples. We used our single-target models (Equations 2.2 and 2.3, Section 2.2.3) to decompose variation in cycle offset ($\hat{c}_{ij}$; median 22.2 across all targets, interquartile range (IQR) 5.03 across all targets) and rate ($\hat{b}_{ij}$; median 0.334 across all targets, IQR 0.076 across all targets) into separate RNA-dependent and RNA-independent components. We fitted these models for each of our 25 targets of interest (He *et al.*, 2021). Summaries of the posterior distributions for rate and cycle offset parameters appear in Tables 2.3 and 2.4. More in-depth summaries are given in Appendix B.

We see a wide range of inferred posterior means for both λ_C and σ_C. For the parameter λ_C, which denotes resolution with respect to cycle offset, larger negative values correspond to improved resolution, and suggest that for a given change in RNA input a larger average change in amplification curve cycle offset will be observed. Offset precision σ_C is always positive, with lower values corresponding to a more precise assay (i.e. less variation in amplification curve cycle offset between samples of the same RNA input). For example, targets CTSL1 and ZDHHC19 both show good resolution (posterior means of -1.86 and -2.20, 95% credible intervals of (-2.09, -1.64) and (-2.59, -1.81), respectively), but different levels of precision (posterior means of 1.57 and 6.15, 95% credible intervals of (1.42, 1.74) and (5.57, 6.80), respectively). Inferences of poor resolution and precision for some targets (e.g. KIAA1370) indicate the potential need for further optimisation of assay properties and/or conditions. We also note that, as expected and as reflected by the posteriors for ρ_C, the geometric mean of the housekeeping cycle offsets is generally correlated with the cycle offset of the target (i.e., lower sample yield should correspond to higher cycle offsets for all targets).

The amplification rate parameter λ_B also varied between targets, with a larger positive value indicating greater average sensitivity to changes in RNA input. While some reactions, including for example CD163 and RAPGEF, showed almost no detectable variation in rate due to RNA input (posterior mean for λ_B of 0.00, 95% credible intervals of (0.00, 0.01) in both cases), others such as PER1 and ZDHHC19 did show a dependence on NS-derived measurements of RNA (posterior means for λ_B of 0.04 in both cases, 95% credible intervals of (0.03, 0.05) and (0.04, 0.04) respectively). Posterior summaries for the precision in amplification rate (σ_B) were roughly similar across targets. We also found that housekeeping levels showed a negative association with the rate of amplification (for example, 95% posterior credible intervals for ρ_B of (-0.03, -0.02) in ARG1 and (-0.02, -0.01) in BATF).

To further elucidate the range of behaviours captured by these single-target models, we simulated from the posterior predictive distributions of $\hat{B}$ and $\hat{C}$ for four targets (CEACAM1, BATF, PER1, and ZDHHC19), to parameterise scaled Gompertz curves. We used these samples to simulate scaled Gompertz curves

Target	$\boldsymbol{\lambda_C}$	$\boldsymbol{\sigma_C}$	$\boldsymbol{\rho_C}$
ARG1	-0.71 (-0.78, -0.65)	1.14 (1.02, 1.26)	1.14 (1.02, 1.26)
BATF	-1.02 (-1.16, -0.88)	0.73 (0.66, 0.81)	0.78 (0.70, 0.86)
C3AR1	-0.90 (-1.00, -0.79)	0.88 (0.79, 0.97)	0.54 (0.43, 0.65)
C9orf95	-0.64 (-0.85, -0.44)	0.86 (0.78, 0.96)	0.51 (0.41, 0.60)
CD163	-0.61 (-0.68, -0.53)	0.51 (0.46, 0.56)	0.36 (0.30, 0.43)
CEACAM1	-0.64 (-0.73, -0.56)	0.97 (0.88, 1.08)	0.42 (0.31, 0.54)
CTSL1	-1.86 (-2.09, -1.64)	1.57 (1.42, 1.74)	1.03 (0.86, 1.21)
GADD45A	-0.68 (-0.80, -0.56)	1.03 (0.94, 1.15)	0.29 (0.17, 0.40)
GNA15	-0.77 (-0.93, -0.61)	0.67 (0.60, 0.73)	0.95 (0.85, 1.04)
HK3	-0.70 (-0.76, -0.63)	0.44 (0.40, 0.49)	0.71 (0.66, 0.76)
HLA-DMB	-0.95 (-1.41, -0.51)	3.25 (2.94, 3.61)	0.67 (0.35, 0.98)
IFI27	-0.91 (-1.01, -0.80)	2.10 (1.89, 2.34)	0.69 (0.48, 0.89)
ISG15	-0.97 (-1.06, -0.87)	2.10 (1.89, 2.34)	0.84 (0.63, 1.06)
JUP	-1.57 (-1.78, -1.36)	2.43 (2.20, 2.69)	1.51 (1.25, 1.77)
KCNJ2	-0.70 (-0.81, -0.60)	0.75 (0.68, 0.83)	0.18 (0.09, 0.27)
KIAA1370	-0.34 (-0.57, -0.10)	0.99 (0.89, 1.10)	0.42 (0.31, 0.53)
LY86	-0.71 (-0.85, -0.58)	0.66 (0.60, 0.74)	0.78 (0.70, 0.86)
OASL	-0.83 (-0.91, -0.76)	0.89 (0.80, 0.99)	0.83 (0.74, 0.93)
OLFM4	-1.01 (-1.12, -0.90)	1.58 (1.43, 1.77)	0.52 (0.35, 0.68)
PDE4B	-0.95 (-1.21, -0.69)	1.35 (1.23, 1.49)	0.90 (0.74, 1.05)
PER1	-1.41 (-1.74, -1.10)	2.61 (2.36, 2.88)	0.39 (0.20, 0.58)
PSMB9	-0.94 (-1.06, -0.83)	0.66 (0.60, 0.73)	0.88 (0.79, 0.96)
RAPGEF1	-0.60 (-0.95, -0.26)	1.79 (1.62, 1.98)	0.68 (0.51, 0.86)
TGFBI	-0.71 (-0.79, -0.63)	0.58 (0.53, 0.65)	0.79 (0.72, 0.87)
ZDHHC19	-2.20 (-2.59, -1.81)	6.15 (5.57, 6.80)	0.89 (0.53, 1.26)

Table 2.3: Posterior mean estimates of cycle offset resolution, precision and housekeeping normalisation parameters for single-target models based on each target (values in brackets = 95% credible intervals).

from targets showing different types of association between RNA input and amplification rate/cycle offset. We simulated these curves at three NanoString-derived RNA input values (results shown in Figure 2.5). CEACAM1 displays negligible rate dependence on RNA input, whereas all three others appear qualitatively different in shape across the gene expression range. *PER1* and *ZDHHC19* exhibit the greatest offset resolution, but this is accompanied by substantially worse precision. A key point to emphasise is that, unlike in the case of qPCR, opting for an approach based purely on time-to-threshold will under-utilise the full information content of qRT-LAMP curves. As rates of amplification also appear important for characterising qRT-LAMP reactions, parameter estimates of the rates could be used in tandem with cycle offsets for target quantitation.

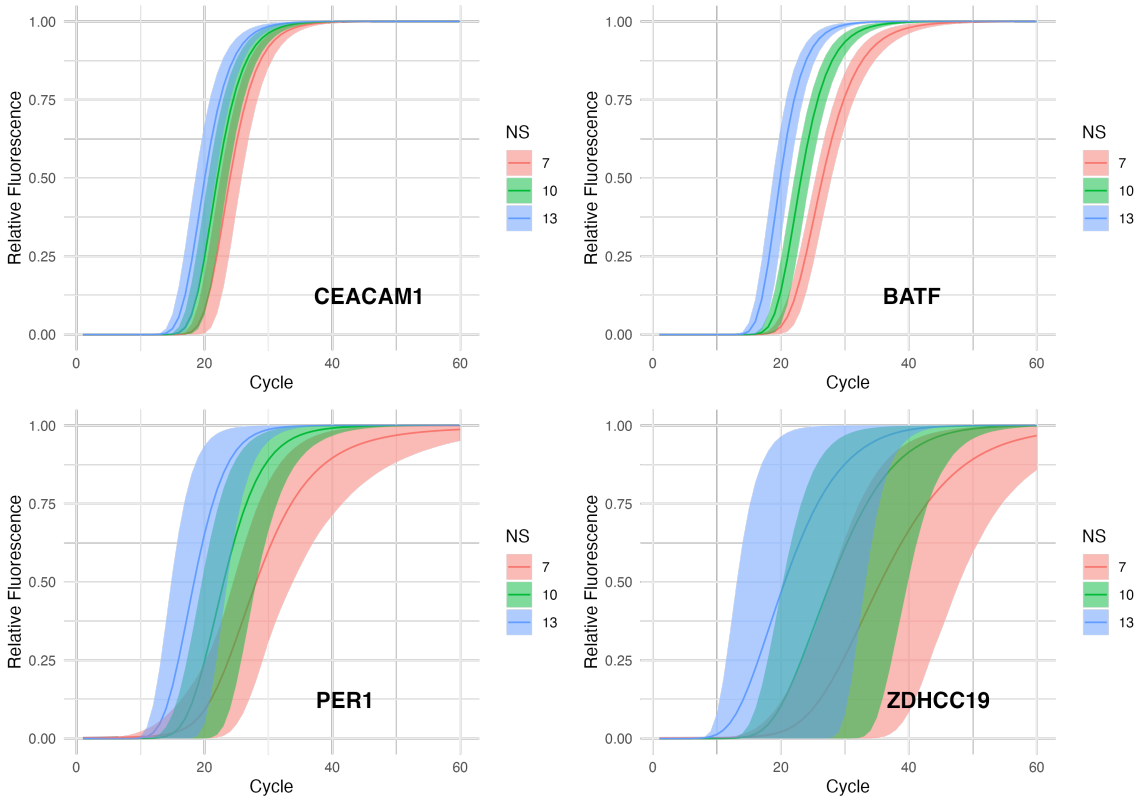

Figure 2.5: Simulated amplification curves for four primers (CEACAM1, RAPGEF1, BATF, and ZDHHC19). Posterior predictive samples of Gompertz curves of scaled fluorescence based on posterior predictive samples from models 2.2 and 2.3. For each target, curves are simulated at three different RNA input (NS) levels – note that higher NS values correspond to higher expression and, generally, to lower cycle offsets (left-shifted amplification curves). Solid lines depict posterior means, while highlighted regions correspond to 95% credible intervals.

Target	$\boldsymbol{\lambda_B}$	$\boldsymbol{\sigma_B}$	$\boldsymbol{\rho_B}$
ARG1	0.02 (0.02, 0.02)	0.03 (0.03, 0.03)	-0.03 (-0.03, -0.03)
BATF	0.03 (0.02, 0.03)	0.03 (0.03, 0.03)	-0.02 (-0.02, -0.01)
C3AR1	0.02 (0.02, 0.02)	0.02 (0.02, 0.03)	-0.01 (-0.02, -0.01)
C9orf95	0.02 (0.01, 0.02)	0.03 (0.02, 0.03)	-0.01 (-0.02, -0.01)
CD163	0.00 (0.00, 0.01)	0.03 (0.02, 0.03)	-0.01 (-0.01, -0.01)
CEACAM1	0.00 (0.00, 0.01)	0.02 (0.02, 0.02)	-0.01 (-0.02, -0.01)
CTSL1	0.03 (0.02, 0.03)	0.03 (0.03, 0.03)	-0.01 (-0.02, -0.01)
GADD45A	0.02 (0.01, 0.02)	0.03 (0.03, 0.03)	-0.02 (-0.02, -0.01)
GNA15	0.02 (0.01, 0.02)	0.02 (0.02, 0.02)	-0.02 (-0.02, -0.02)
HK3	0.00 (0.00, 0.00)	0.02 (0.02, 0.02)	-0.01 (-0.02, -0.01)
HLA-DMB	0.01 (0.00, 0.01)	0.03 (0.03, 0.03)	-0.01 (-0.01, -0.01)
IFI27	0.01 (0.00, 0.01)	0.03 (0.02, 0.03)	-0.01 (-0.02, -0.01)
ISG15	0.01 (0.00, 0.01)	0.02 (0.02, 0.02)	-0.02 (-0.02, -0.01)
JUP	0.02 (0.02, 0.02)	0.03 (0.03, 0.03)	-0.02 (-0.02, -0.01)
KCNJ2	0.00 (0.00, 0.00)	0.02 (0.02, 0.02)	-0.01 (-0.01, 0.00)
KIAA1370	0.02 (0.01, 0.03)	0.04 (0.03, 0.04)	-0.01 (-0.02, -0.01)
LY86	0.02 (0.01, 0.02)	0.02 (0.02, 0.02)	-0.02 (-0.02, -0.01)
OASL	0.01 (0.01, 0.02)	0.02 (0.02, 0.03)	-0.02 (-0.02, -0.01)
OLFM4	0.01 (0.01, 0.01)	0.03 (0.03, 0.03)	-0.01 (-0.01, -0.01)
PDE4B	0.01 (0.00, 0.01)	0.02 (0.02, 0.02)	-0.01 (-0.02, -0.01)
PER1	0.04 (0.03, 0.05)	0.05 (0.05, 0.06)	-0.01 (-0.02, -0.01)
PSMB9	0.01 (0.00, 0.01)	0.02 (0.01, 0.02)	-0.01 (-0.02, -0.01)
RAPGEF1	0.00 (0.00, 0.01)	0.02 (0.02, 0.02)	-0.02 (-0.02, -0.02)
TGFBI	0.01 (0.00, 0.01)	0.02 (0.02, 0.03)	-0.01 (-0.02, -0.01)
ZDHHC19	0.04 (0.04, 0.04)	0.05 (0.04, 0.05)	-0.02 (-0.02, -0.01)

Table 2.4: Posterior mean estimates of rate resolution, precision and housekeeping normalisation parameters for single-target models based on each target (values in brackets = 95% credible intervals).

2.3.2 Multi-target model identifies assay properties associated with LAMP resolution in patient data

We fitted the multi-target models described in Equations 2.4 and 2.5 to our clinical dataset using different combinations of primer/target features in the resolution function $\lambda(\boldsymbol{z}_k;\boldsymbol{\gamma})$. We considered the following forms for λ in each analysis: *a)* an intercept-only model $\lambda(\boldsymbol{\gamma}) = \gamma_0$; *b)* for each of the three amplicon properties, a univariate linear function (e.g. for stem length L_{stem}, $\lambda(z_k;\boldsymbol{\gamma}) = \gamma_0 + \gamma_{L_{\text{stem}}} L_{\text{stem,k}}$); and *c)* for each primer-specific quantity and primer pair, a bivariate model (e.g. in the case of free energy of annealing for the F1/B1 pair, $\lambda(\boldsymbol{z}_k;\boldsymbol{\gamma}) = \gamma_0 + \gamma_{\Delta G_a^{\text{F1}}} \Delta G_{a,k}^{\text{F1}} + \gamma_{\Delta G_a^{\text{B1}}} \Delta G_{a,k}^{\text{B1}}$). Note again that in this setting we only have one candidate primer set for each target and so no dependence on primer set l.

Posterior summaries of the resolution function coefficients as well as the leave-one-out Pareto smoothed importance sampling (LOO-PSIS) stacking weights for each of these models appear in Figure 2.6 for models based on estimated rate

and in Figure 2.7 for those based on estimated cycle offsets. In each case, for the stacking weights (left-hand panels of Figures 2.6/2.7), the higher the stacking weight, the higher quality predictions the model was judged to have made for a left-out sample in comparison to alternative models. In the right-hand panel of Figures 2.6/2.7, we show posterior summaries of effect sizes (i.e. $\boldsymbol{\gamma}_B$ and $\boldsymbol{\gamma}_C$ coefficients) associated with each of the primer/target properties used as features in the resolution function, λ.

We found quantities associated with cycle offset resolution in both positive and negative directions whose corresponding models were assigned strong weightings by the LOO-PSIS procedure. Increased stem length, complexity in the F3/B3 primer pair, free energy of annealing in F2/B2 regions of the FIP/BIP primers, and free energy of secondary structure formation in F3/B3 were associated with improved resolution (negative effect sizes), and all corresponding models received relatively high stacking weights. Immediately striking is the observation that, despite the widespread assumption that design of the FIP/BIP primers is most influential in determining LAMP performance, we detect surprisingly strong associations with properties of other primer pairs. This is a strong signal suggesting the need for future follow-up work. These associations, particularly for F3, may reflect the importance of the primer to the synthesis of a full-length template, resulting in more sites for primer annealing and amplification (Figure 2.1). Increases in complexity of LF/LB primers, free energy of self-dimerisation for FIP/BIP, and free energy of secondary structure formation in LF/LB were associated with worse resolution (positive effect sizes), with corresponding models also receiving high stacking weights. We do observe some primer pairs for which posterior mean effect sizes were in opposing directions. For example, free energy of self-dimerisation shows a positive effect for the B3 primer and a negative effect for F3. We see that no such models were assigned large stacking weights. We might expect these different associations with resolution due to stochasticity of the LAMP assay in the production of intermediates from either strand of the target.

We also observe associations between primer/amplicon properties and rate resolution (Figure 2.6). Note that in this case a positive effect size corresponds to improved resolution. Some properties appear to be associated with rate resolution in the same way as offset resolution: for example, 4-complexity of F3/B3 are associated with improved resolution in each case. Others, such as stem length, are associated with worsened rate resolution but improved offset resolution. Others still show strong effects in one model but not in another, including free energy of self-dimerisation of the F3/B3 pair. Unlike for offset models, one of the most strongly identified contributors to rate resolution (free energy of annealing of the F1/B1 regions) exhibits distinct reversed directions of effect between F1 and B1 regions.

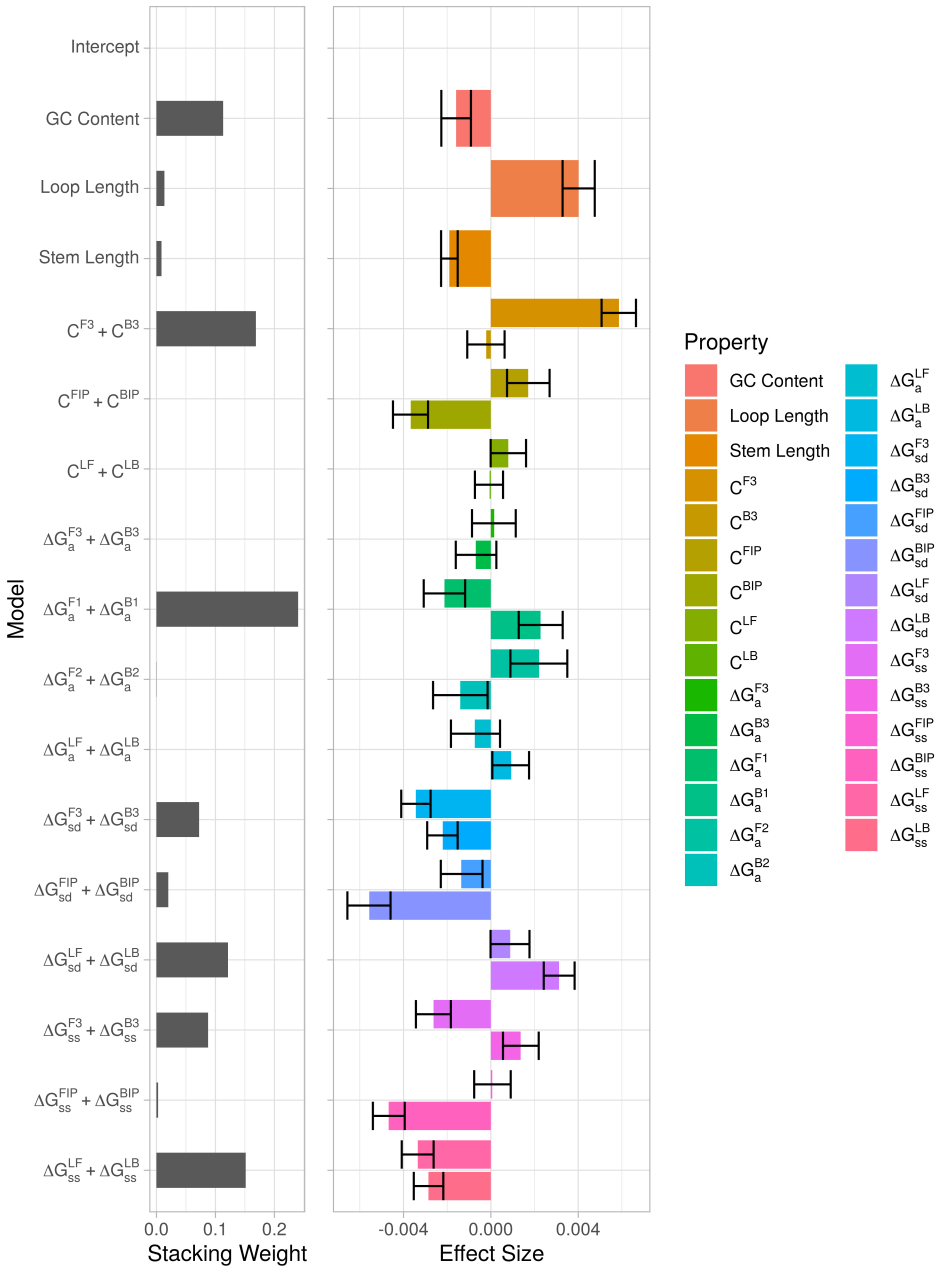

Figure 2.6: Posterior summaries of multi-target models for rate resolution specified in Sections 2.2.4 and 2.3.2, based on combinations of primer/target properties, with their associated Bayesian model stacking weights (left) and γ_B-coefficient effects sizes (right). Effect sizes are posterior means, with error bars delineating 95% credible intervals. Notation as given in Table 2.2.

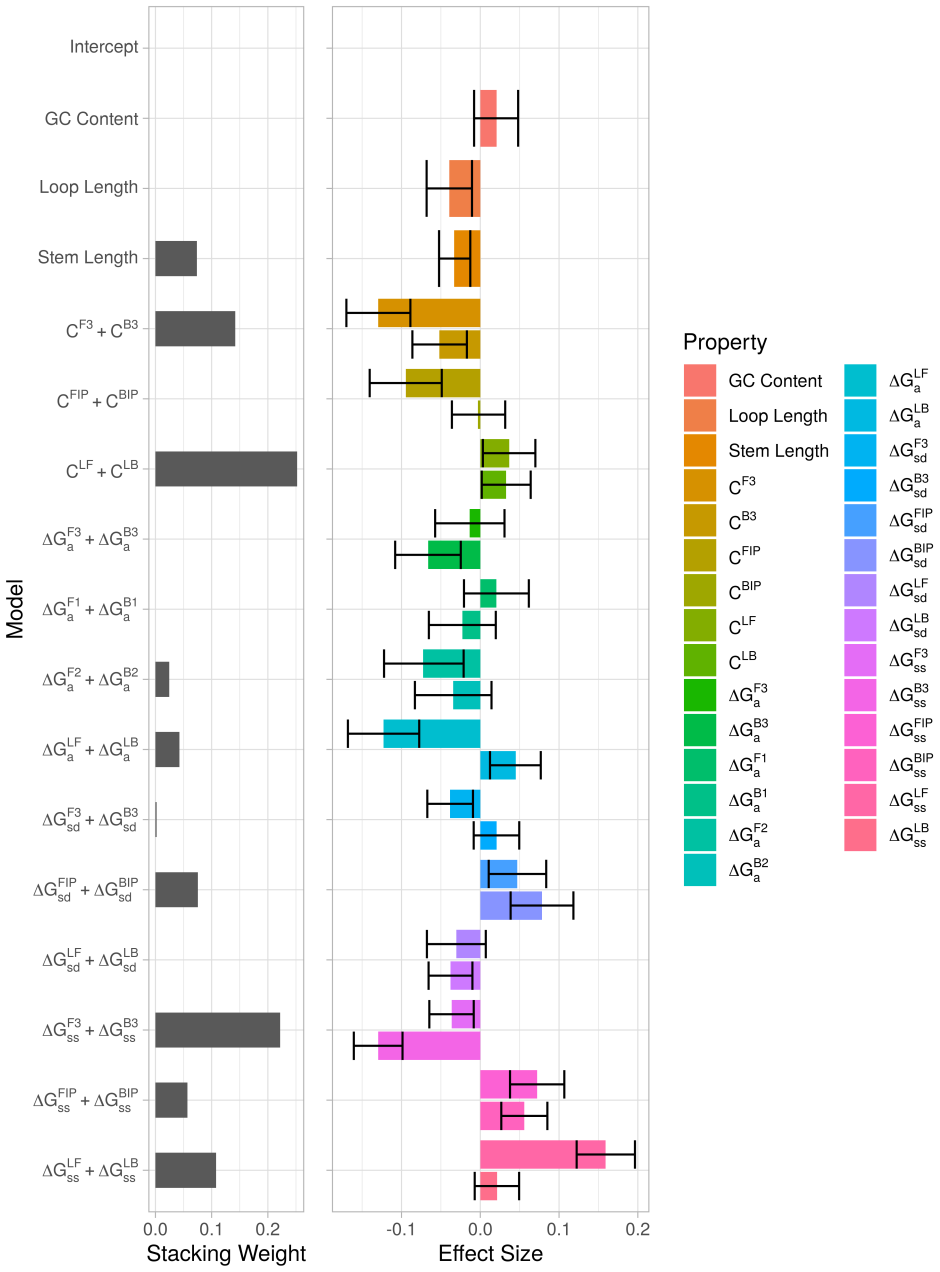

Figure 2.7: Posterior summaries of multi-target models for cycle offset resolution specified in Sections 2.2.4 and 2.3.2, based on combinations of primer/target properties, with their associated Bayesian model stacking weights (left) and γ_C-coefficient effects sizes (right). Effect sizes are posterior means, with error bars delineating 95% credible intervals. Notation as given in Table 2.2.

2.3.3 Varying qRT-LAMP primer sequences identifies distinct across-target and within-target patterns of assay resolution

In this section we describe our use of a distinct, novel synthetic dataset (preparation described in Section 2.2.1 and Remmel *et al.*, 2022) to address with more granularity the direct effects of changes to primer sequences on assay resolution. While in Section 2.3.2 we identified interesting and unexpected associations with other primer pairs, here we focused on FIP and BIP primers – future work will address comparable questions for F3/B3/LF/LB. We also focus here on offset resolution only. In the previous analysis, we pooled our data across targets to estimate associations of assay properties with resolution. However, while we showed associations of these properties with resolution across targets, we could not determine for any *individual* target whether variation in these properties was associated with variation in resolution. To address the latter objective (arguably more important for LAMP optimisation), we generated a IVT RNA dataset comprising multiple primer sets (containing multiple combinations of the FIP and BIP primers) for each target. We fitted models with two different forms of resolution function. In the first form, for each combination of assay properties we fitted a model based on those properties alone with no target-level grouping terms (*across-target* models). In this way, the resolution function λ did not depend on target (k) other than through the primer set (l) features $\boldsymbol{z}_{kl}$ (e.g., $\lambda(z_{kl};\boldsymbol{\gamma}) = \gamma_0 + \gamma_{L_{\text{stem}}} L_{\text{stem},k,l}$).

In the second model form (*within-target*), we specified a hierarchical random effects term (varying by target) for the resolution intercept, γ_0. In this way, we were able to infer the dependence of resolution on primer set properties while accounting for intrinsic target-to-target variation in resolution. Summarised inferences and model stacking weights after fitting of the across-target and within-target models to the IVT RNA data appear in Figure 2.8. Note again that these models are not a subset of those shown in Figure 2.7, but rather similar models applied to a novel synthetic IVT dataset with systematically varied FIP and BIP primers.

We note that the across-target models gave far more inferences than the within-target models of high-confidence, high-magnitude effect sizes of assay properties on resolution. Relatively few properties were inferred to have effect sizes of high magnitude (many 95% credible intervals contain 0) for within-target models. Importantly, these results indicate that much of the variation in resolution explained in previous models may be due to intrinsic target-to-target variation not attributable to the assay properties we considered. Of the within-target models, those based on complexity of the FIP primer, free energy of self-dimerisation of the FIP primer, and free energy of secondary structure formation of the FIP primer had significant posterior mean effect sizes (i.e. 95% posterior credible intervals not containing zero). Notably, several of these within-target inferences differ in directionality from their across-target counterparts, indicating that not accounting for intrinsic target-to-target variation could be masking the nature of the association between assay properties and resolution. Furthermore, these

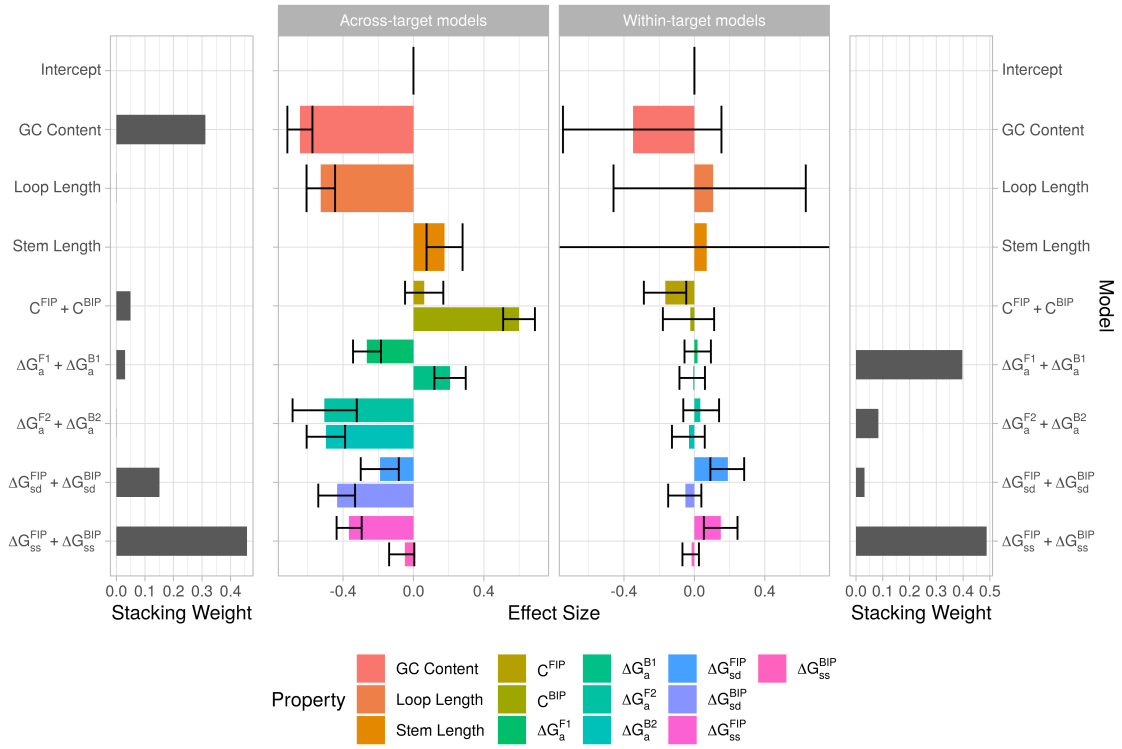

Figure 2.8: Posterior summaries of multi-target models as specified in Sections 2.2.4 and 2.3.3 applied to IVT RNA data. Effect sizes for $\boldsymbol{\gamma}$-coefficients from across-target (center left) and within-target (center right) models, as well as associated model stacking weights (far left/right respectively). Error bars delineate 95% credible intervals.

results highlight the importance of utilising multiple primer sets per target in characterising this variation. Interestingly, we do note that both within- and across-target models agreed in allocation of the largest stacking weight to free energy of secondary structure formation for FIP/BIP, a novel quantity for consideration in the design of qRT-LAMP assays.

2.4 Discussion

We have presented and applied a novel model-based framework to characterise quantitative performance of the qRT-LAMP assay. We first introduced a non-linear least-squares fitting procedure based on Gompertz curves that allows for fast, accurate characterisation of qRT-LAMP reactions. Based on LAMP reaction curve characteristics derived from this Gompertz curve-based fitting procedure, our Bayesian hierarchical models have generated new insights into the qRT-LAMP reaction and its properties at the single- and multi-target levels. At the single-target level, our analysis identified RNA-independent and RNA-dependent associations with qualitative characteristics and quantitative performance of qRT-LAMP in a dataset of patients with suspected acute infection. Our investigation of the effects of various sequence-based and thermodynamic primer/target properties on qRT-LAMP performance identified novel associations between some of these properties and LAMP assay resolution. We've not only adapted our framework to single-target and multi-target settings but also to different qRT-LAMP data modalities (from IVT RNA and clinical samples). Finally, we have used our IVT RNA data and modelling framework to investigate the extent to which associations observed between primer/target properties and LAMP performance will be borne out by artificially altering primers as part of an assay optimisation workflow.

We highlight several advantages of our approach. The single-target models introduced in Section 2.2.3 and applied in Section 2.3.1 provide a formal framework for inference with uncertainty of qRT-LAMP performance measures such as resolution and precision. This approach allows assay developers to ascertain the quantitative performance of an predesigned LAMP target. Such inferences may in turn motivate optimisation of the given target or, in light of poor target performance, alternative target selection as part of a broader panel of target biomarkers. Furthermore, both our single-target and multi-target models allow decomposition of variation in LAMP reaction characteristics into components useful for mRNA quantitation (i.e. amplification rate and cycle offset). By incorporating different features of the assay into model components for RNA-dependent and RNA-independent variation, we can uncover potentially tunable determinants of LAMP assay performance. Indeed, our framework applied to IVT RNA data revealed significant associations between primer and target features and assay resolution. While IVT RNA data are useful for assay development due to their cost and ease of handling, they are not biologically representative and would likely not recapitulate assay performance seen on clinical samples. As we could adapt our models to both IVT RNA and clinical samples, we were able to compare our inferences across different data types and to better anticipate clinical

performance of a given assay. We see our framework and analysis as important first steps towards a system for LAMP assay design and optimisation.

However, the results of our hierarchical modelling (Section 2.3.3) highlight the complexities of optimising qRT-LAMP assay performance. While FIP and BIP primers are known to play a central role in LAMP-based amplification, our model-based analysis (Figures 2.6 and 2.7) provides evidence that other primer characteristics may also affect assay performance. Models based on loop amplification and inner primer pairs were identified by model stacking as producing some of the most robust and predictive associations. Furthermore, our findings suggest that increasing a given property (for example, free energy of self-dimerisation) may be associated with improved resolution for one primer pair (FIP/BIP) but the opposite for another (LF/LB). As changing one oligo property (e.g. secondary structure formation) may alter another (propensity to self-dimerise), our results support joint consideration of all primer pair characteristics as part of assay optimisation efforts. We also observe, for several properties, a difference in magnitude of effect sizes between forward and backward primers (Figure 2.8). Subject to further investigation, we hypothesise that this difference in association strength between members of a primer pair could arise due to the orientation of initial RNA inputs to the LAMP assay. For example, B3 primers could anneal to RNA inputs before reverse transcription while F3 primers – relying on the generation of an F3c region – could not anneal until after reverse transcription, and this could explain some degree of within-primer pair disparity. However, this trend is far from clear and consistent across different quantities and primer pairs, so much further investigation is required.

Further work is required to accurately predict performance in a proposed assay for a potentially novel target. As part of our efforts towards a more predictive model, we analyzed multiple primer sets per target with our IVT RNA dataset. Indeed, we found for multiple models that the effects of varying properties of primers and amplicons differ in their strength and direction when viewed across-target vs within-target (Figure 2.8). Put another way, accounting for target-to-target differences in resolution not due to the assay features attenuated the strength of association between the features and resolution. These results suggest that other assay features may be explaining target-to-target differences in resolution and that predictive models based on pooling all targets together (i.e. across-target models) may not generalise well to new targets. Interestingly, recent work on predictive modelling of PCR amplification from primer/target features showed good performance in held-out test sets (Döring *et al.*, 2019). This performance could be due to both a focus on a highly related set of targets which might be expected to share assay performance properties as well as due to random data subsampling to create training, validation and test data splits for model development. In this setting, all splits would likely include observations for a given target (a kind of information leakage), complicating interpretation of the model's generalisability. However, if we consider a highly related family of targets to be similar to a single target, the results of Döring *et al.* (2019) are consistent with our findings that within-target trends (rather than general trends shared across targets) in the relationship between LAMP assay performance and primer properties are important for optimisation of a given target or target family. Reliably

determining the optimal properties of a primer set to achieve a pre-specified level of quantitative performance would allow for less trial-and-error in the development process, and ongoing and future work will be required to add both more targets and primer sets to our IVT RNA dataset.

To extend and more comprehensively validate the results presented here, a follow-up study will be required to profile many more targets with – crucially – multiple combinations of primers for each target. In particular, we note that while our IVT RNA data analysis included multiple primer sets per target, we were only able to assay three targets due to time/cost constraints. Further data collection will also permit investigation of additional candidate assay properties, including gDNA interference, sensitivity to reagents and other sequence-based/thermodynamic quantities. In addition, while IVT RNA data are useful for assay development due to their cost and ease of handling, they are not biologically representative. The degree to which conclusions drawn from IVT RNA translate to clinical samples is the subject of ongoing and future investigation. Also, while in our multi-target modelling work we have investigated variation in resolution with respect to cycle offset, we have not provided a complementary analysis for the other properties shown to be important in our single-target modelling section. These include resolution with respect to amplification rate, and precision. Extending our analysis to identify associations between these performance characteristics and assay properties is the subject of ongoing and future work. In addition, for simplicity, we assumed a linear relationship between assay properties and resolution. Further investigations will involve nonlinear relationships between these quantities.

2.5 Conclusion

This work provides an extensible platform to address the significant knowledge gap in enabling reliable quantitative design and use of the qRT-LAMP assay. Our analysis has uncovered important insights of RNA-dependent and RNA-independent effects on qRT-LAMP assay behaviour as well as assay properties associated with improved quantitative resolution of LAMP. Our efforts have also raised important considerations for constructing truly predictive systems of assay performance. With further data collection, expansion of our set of assay features, and extension of our modelling framework, we hope to take another important step towards principled design of robust, clinically performant qRT-LAMP assays.

Chapter 3

Data-driven design of targeted gene panels for estimating immunotherapy biomarkers

Overview

In this chapter, we introduce a novel data-driven framework for the design of targeted gene panels for estimating exome-wide biomarkers in cancer immunotherapy. Our first goal is to develop a generative model for the profile of mutation across the exome, which allows for gene- and variant type-dependent mutation rates. Based on this model, we then propose a new procedure for estimating biomarkers such as tumour mutation burden and tumour indel burden. Our approach allows the practitioner to select a targeted gene panel of a prespecified size, and then construct an estimator that only depends on the selected genes. Alternatively, the practitioner may apply our method to make predictions based on an existing gene panel, or to augment a gene panel to a given size. We demonstrate the excellent performance of our proposal using data from three non-small cell lung cancer studies, as well as data from six other cancer types.

3.1 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). Since the discovery of immune checkpoint blockade (ICB)[1] (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 PD-L1 (Buchbinder and Desai, 2016). Inhibition of these checkpoints can promote

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

a more aggressive anti-tumour immune response (Pardoll, 2012), and in some patients this leads to long-term remission (Borghaei *et al.*, 2021). 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.

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 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).

In this chapter 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 high-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 parameters of the model. In addition, this identifies a subset of genes that are mutated above or below the background rate. Our 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. Targeted gene panels have found use throughout cancer and principles behind their design and implementation are of substantial general interest (Bewicke-Copley *et al.*, 2019).

We demonstrate the excellent practical performance of our framework using a non-small cell lung cancer (NSCLC) dataset (Campbell *et al.*, 2016), and include a comparison with existing state-of-the-art approaches for estimating TMB. We further validate these results by testing the performance on data from two more NSCLC studies (Hellmann *et al.*, 2018a; Rizvi *et al.*, 2015). Moreover, since our model allows variant type-dependent mutation rates, it can be adapted easily to predict other biomarkers, such as TIB. 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. Finally, in order to further investigate the utility of our proposal across a range of mutation profiles, we use it to select targeted gene panels and estimate TMB in six other cancer types.

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 as follows. In Section 3.2, we introduce our NSCLC data sources, and provide a detailed description of our methodological proposal. The full demonstration of our method using the NSCLC dataset is given in Section 3.3. Section 3.4 provides several further analyses to investigate the robustness of our proposal in other cancer types and we conclude in Section 3.5. We also provide an `R` package `ICBioMark` (Bradley and Cannings, 2021b) which implements the methodology and reproduces our experimental results. This is described in Appendix C.

3.2 Methodology

3.2.1 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 muta-

tions, 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.

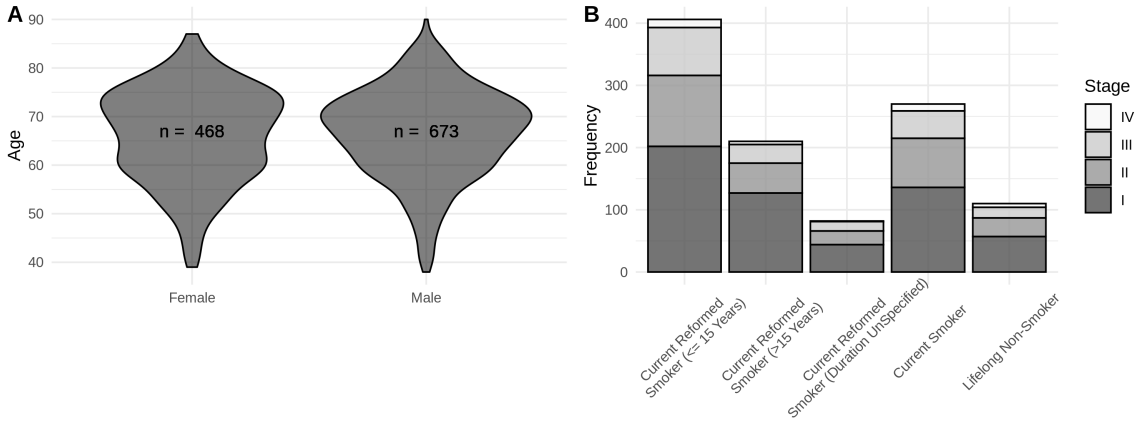

Figure 3.1: Demographic data for the clinical cohort in Campbell *et al.* (2016). **A**: Violin plots of age for patients, stratified by sex. **B**: Stacked bar chart of patients' smoking histories, shaded according to cancer stage diagnosis.

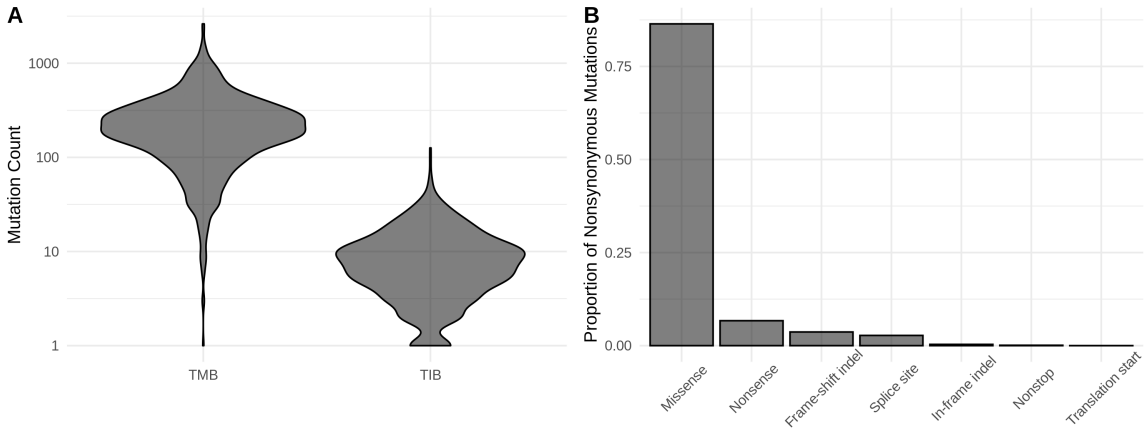

Figure 3.2: Dataset-wide distribution of mutations. **A**: Violin plot of the distribution of TMB and TIB across training samples. **B**: The relative frequency of different nonsynonymous mutation types.

It is useful at this point to introduce the notation used throughout this chapter. 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[2]. A gene panel is a subset $P \subseteq G$, and we write $\ell_P := \sum_{g \in P} \ell_g$

[2]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).

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

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

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 $T_{0\bar{S}}$.

3.2.2 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 since we expect that different mutation types occur at different rates, and that the BMR is different in each sample.

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 3.2.5 for more detail).

3.2.3 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[3]

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

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 3.2.2.

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 $w_{gs} = 0$),

[3]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).

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

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

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 3.2.5.

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 3.3, for comparison we also include the performance of the first-fit estimator $T(\hat{w}^{\text{first-fit}})$.

3.2.4 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

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

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 3.3.5.

3.2.5 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 LASSO estimator in generalised linear regression (see, for example, Michoel (2016) and Friedman *et al.* (2021)), 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 3.3, 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 $[\hat{T}_{\mathrm{L}}, \hat{T}_{\mathrm{U}}]$ 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, 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 3.3.2 and 3.3.4 that in practice this approach provides intervals with valid empirical coverage.

3.3 Demonstration using an NSCLC dataset

In this section we demonstrate the practical performance of our proposal using the dataset from Campbell *et al.* (2016), which we introduced in Section 3.2.1. 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 TST-170 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.

3.3.1 Generative model fit and validation

The first step in our analysis is to fit the model proposed in Section 3.2.2 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 3.2.5. 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 3.5 for further discussion. Finally, note that the average μ_i among current smokers is 5.40 (with standard deviation 0.76), amongst reformed smokers is 5.26 (0.84), and among lifelong non-smokers is 4.04 (1.12). This suggest that smokers may have higher BMRs, which is as we expect.

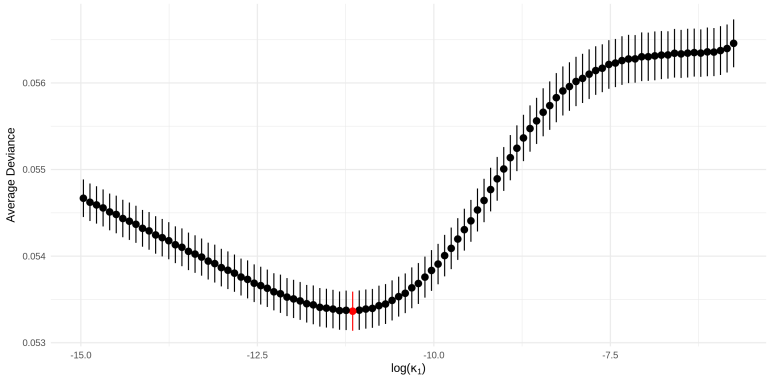

Figure 3.3: The average deviance (with one standard deviation) across the 10 folds in our cross-validation procedure plotted against $\log(\kappa_1)$. The minimum average deviance is highlighted red.

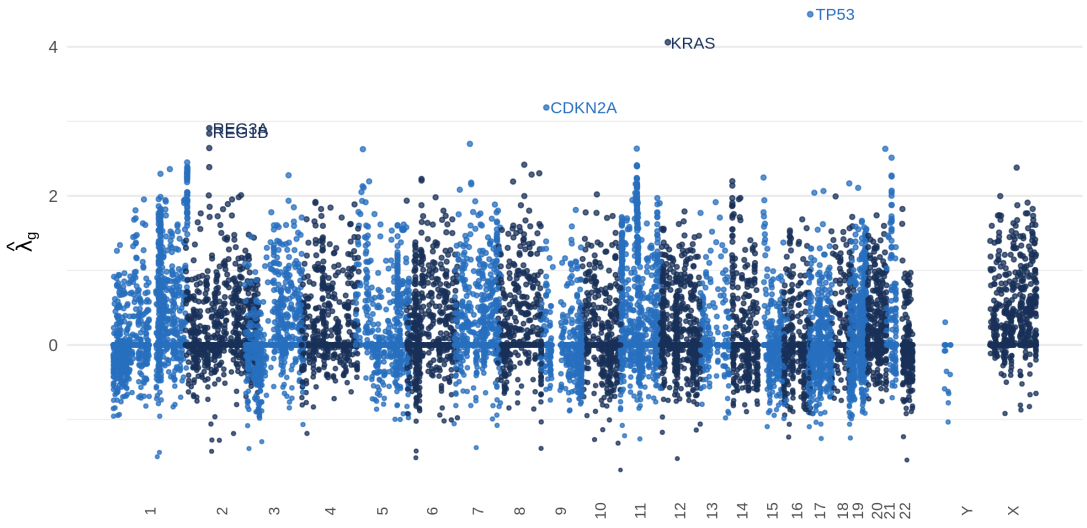

Figure 3.4: Manhattan plot of fitted parameters $\hat{\lambda}_g$ and their associated genes' chromosomal locations. The genes with the five largest positive parameter estimates are labelled.

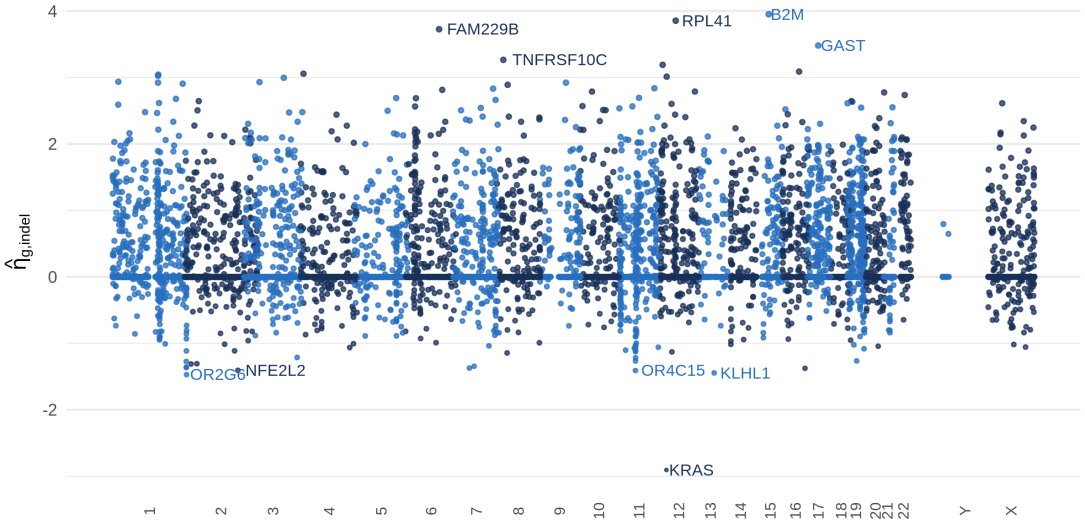

Figure 3.5: Manhattan plot of fitted parameters $\hat{\eta}_{g,\text{indel}}$ and their associated genes' chromosomal locations. The five largest positive and negative genes are labelled.

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).

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

3.3.2 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.*, 2018b; 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 0.275 (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 3.2.3 with $\bar{S} = S$, where the model parameters are estimated as described in Section 3.3.1. 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, as well as a panel selected by the proposal of Lyu *et al.*

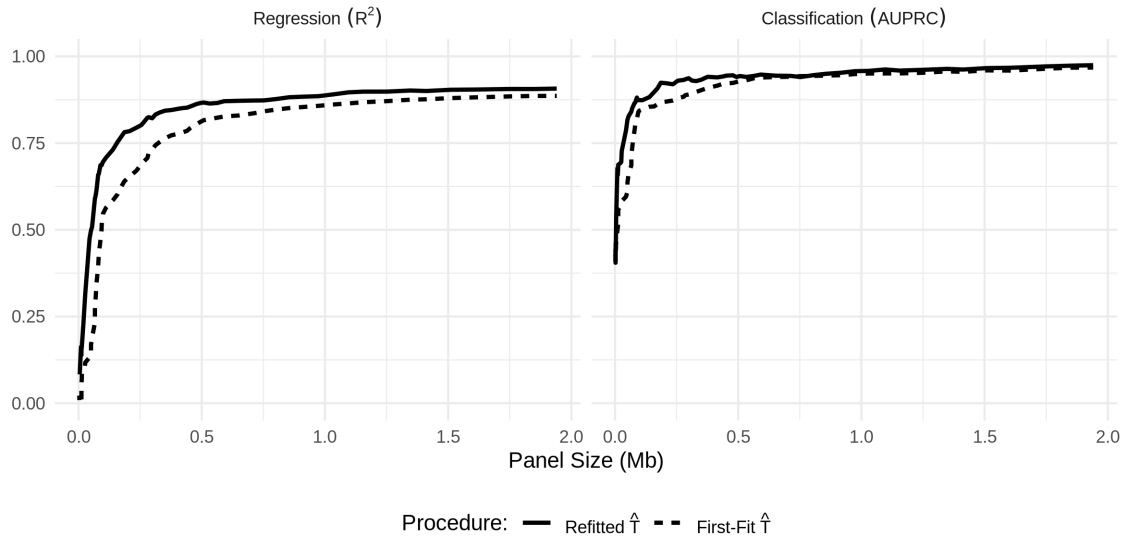

Figure 3.6: Performance of our first-fit and refitted estimators of TMB as the selected panel size varies. **Left**: R^2, **Right**: AUPRC.

(2018). 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). Further, the panel selected by the approach in Lyu *et al.* (2018) consists of the genes that are mutated more than 10% of the time, that are less than 0.015Mb in length and for which the presence of a mutation in the gene is significantly associated with higher TMB values. 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).

(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 is a standard practical procedure for the estimation of TMB from targeted gene panels. The fourth is the approach proposed by Lyu *et al.* (2018). 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.

We present results of these comparisons in Figure 3.7. First, for each of the four panels considered here, we see that our refitted estimator applied to the panel

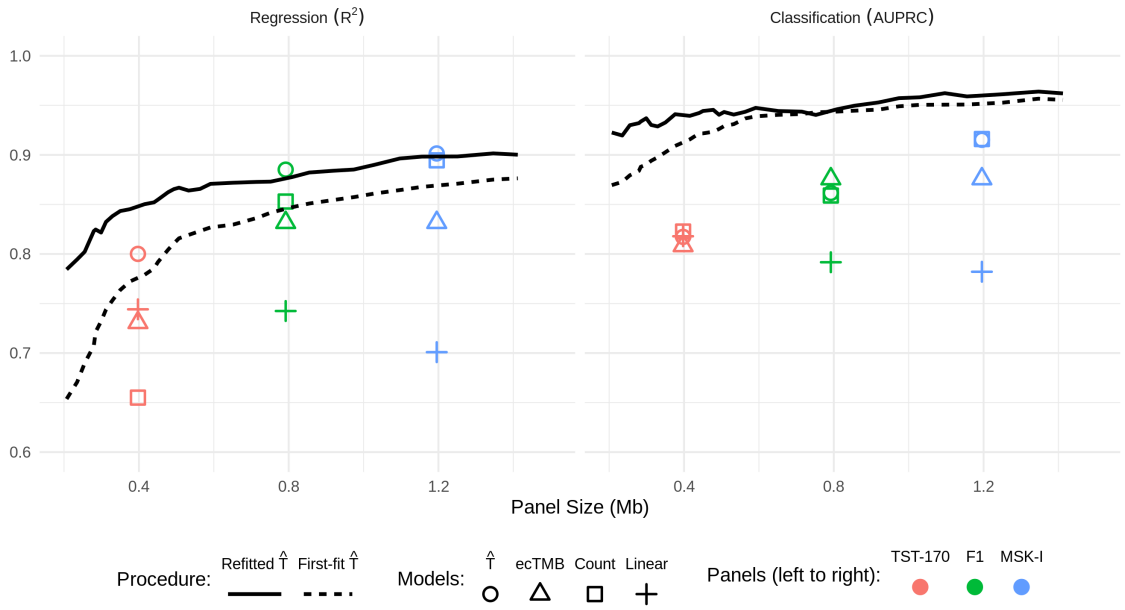

Figure 3.7: The performance of our TMB estimator in comparison to existing approaches. **Left**: R^2, **Right**: AUPRC.

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 (perhaps even of smaller size) based on the training data. For instance, in comparison to predictions based on the TST-170 panel, our procedure can achieve higher R^2 with a selected panel of half the size (with 0.2Mb we obtain an R^2 of 0.78). 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 ($\sim$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. The full proposal of Lyu *et al.* (2018), which involves applying the linear regression estimate to the panel selected as described above, also performs well here.

Finally in this subsection 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 because our selection procedure tends to favour genes with higher overall mutation rates and thus a count estimator based

on the highly mutated genes will overestimate the total number of mutations. We also include a red shaded region comprising all points for which heuristic 90% prediction intervals (as described in Section 3.2.5) 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.

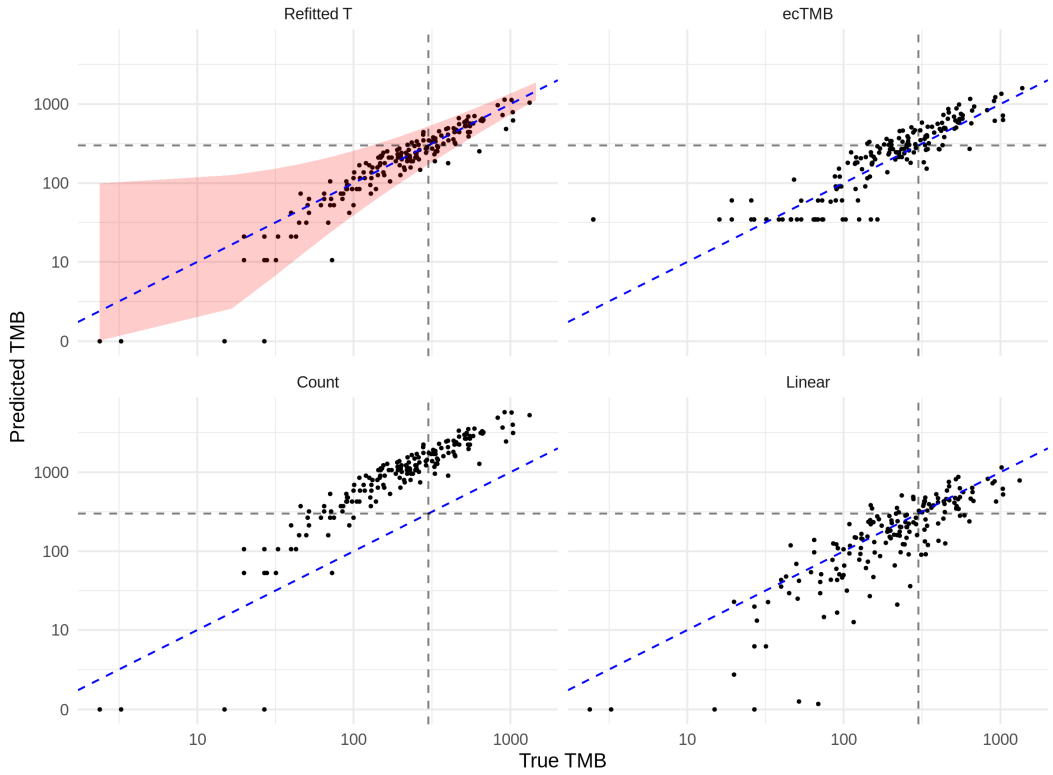

Figure 3.8: Prediction of TMB on the test dataset. Dashed blue (diagonal) line represents perfect prediction, and the black dashed lines indicate true and predicted TMB values of 300.

Robustness to different training datasets

Here we investigate the robustness of our proposal to changes in the training dataset. We conduct an experiment that first involves splitting the training data set of $n = 800$ observations into four disjoint datasets of 200 observations. We then retrain our model and estimator given in Sections 3.2.2 and 3.2.3 based on the four possible datasets that combine three of the four subsets. We then evaluate the predictive performance on the validation dataset similarly to in the previous subsection. The results are given in Figure 3.9; we see that our proposal is has very similar regression performance on the four different subsets.

3.3.3 External testing and classification for immunotherapy

The aim of this section is to further test our proposed estimator of TMB by making use of two external NSCLC datasets for which the response to immunotherapy

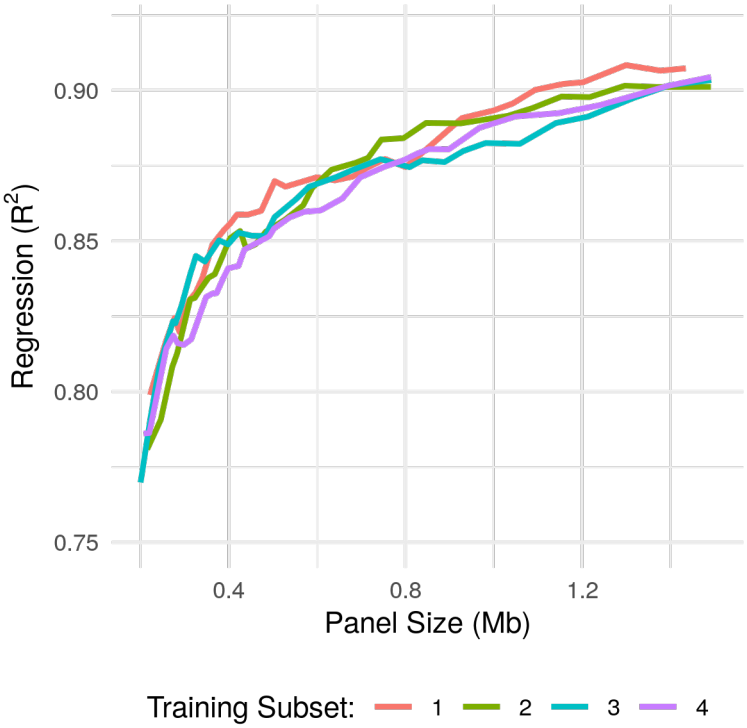

Figure 3.9: The regression performance on the validation dataset for the four different training subsets of 600 observations.

is available: Hellmann *et al.* (2018a), which contains 75 samples with an average TMB of 261; and Rizvi *et al.* (2015), which contains 34 samples with an average TMB of 258.

We first use our refitted estimator trained on the same data as in Section 3.3.2 to predict TMB for the samples in the new datasets using the selected panel of size 0.6Mb. The predictions are given in Figure 3.10; the corresponding regression performance is $R^2 = 0.70$ across the two datasets, with a joint AUPRC for classifying tumours to high or low TMB classes of 0.91.

These datasets also allow us to assess the practical utility of using our estimated TMB values to predict response to immunotherapy. Of the 75 samples in the Hellmann *et al.* (2018a) study, 37 were identified as having a *Durable Clinical Benefit* (Class 1) in response to immunotherapy (PD-L1+CTLA-4 blockade), and the remaining 38 were deemed to have *No Benefit* (Class 0). Of the 34 samples in the Rizvi *et al.* (2015) study, 14 were identified as having a *Durable clinical benefit beyond 6 months* (Class 1) in response to immunotherapy (Pembrolizumab), while the remaining 20 were deemed not to have such benefit (Class 0). Since the treatment and outcome definition differ between studies, we separate them for analysis of response. We construct two simple classifiers for comparison, the first assigning a sample to Class 1 if the true TMB value is greater than some threshold t, and the second using our estimated value of TMB in the same way. In Figure 3.11, we plot the receiver operating characteristic (ROC) curve (that is the false positive rate against the true positive rate as the classification threshold

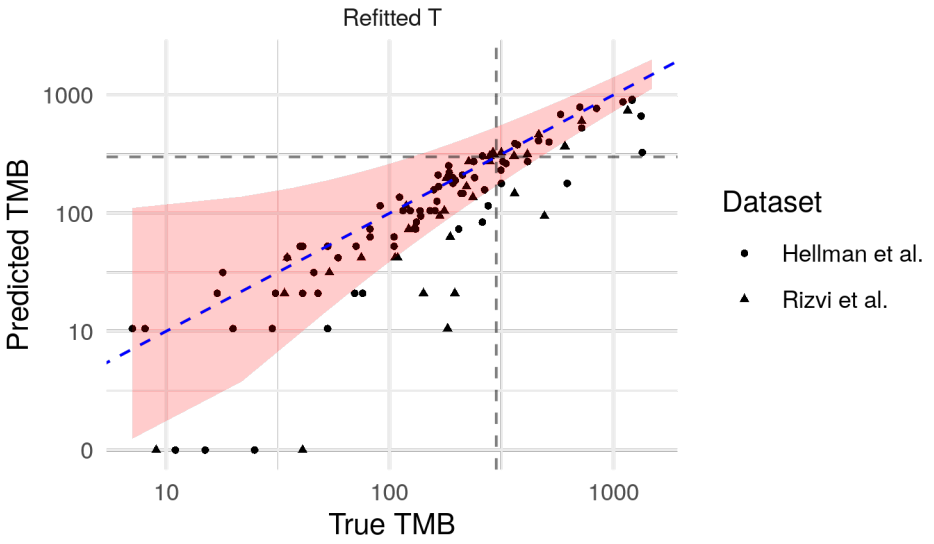

Figure 3.10: Performance of our model trained on the Campbell *et al.* (2016) dataset used to predict TMB based on the panel of size 0.6Mb selected by our method on the external test datasets of Hellmann *et al.* (2018a) and Rizvi *et al.* (2015)

t varies). The area under the ROC curve is 0.68 for the Hellmann *et al.* (2018a) dataset when using the true TMB value and is 0.64 when using the estimated TMB value. The Rizvi *et al.* (2015) has an area under the ROC curve of 0.79 using true TMB values and 0.76 using estimated TMB values. We see that, in both cases, very little is lost in terms of predicting response to immunotherapy when using our estimated value of TMB.

3.3.4 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 3.2.3 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.

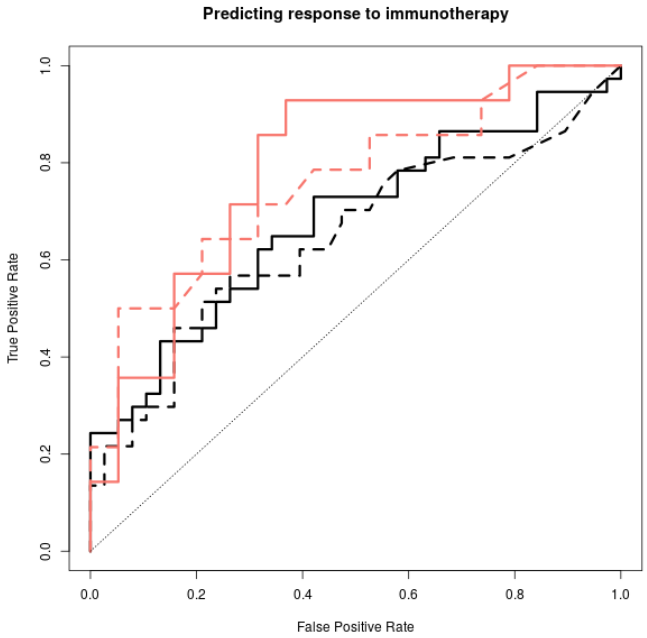

Figure 3.11: ROC curves for classifying the response to immunotherapy in the Hellmann *et al.* (2018a) (black) and Rizvi *et al.* (2015) (red) datasets using the true TMB values (solid) and estimated TMB values (dashed) based on the panel of size 0.6Mb selected by our method.

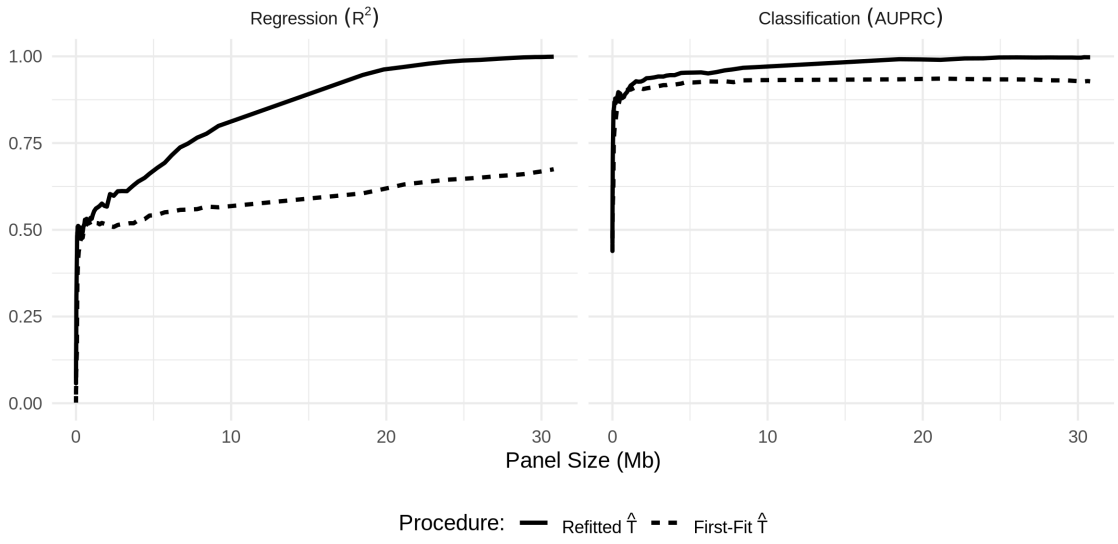

Figure 3.12: Performance of our first-fit and refitted estimators of TIB as the selected panel size varies. **Left**: R^2, **Right**: AUPRC.

The results are presented in Figure 3.12. 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 3.13 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.

3.3.5 A panel-augmentation case study

As discussed in Section 3.2.4, 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 ($\sim$0.4Mb) and augmenting to 0.6Mb in length, demonstrating impressive gains in predictive performance.

We apply the augmentation method described in Section 3.2.4, 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.

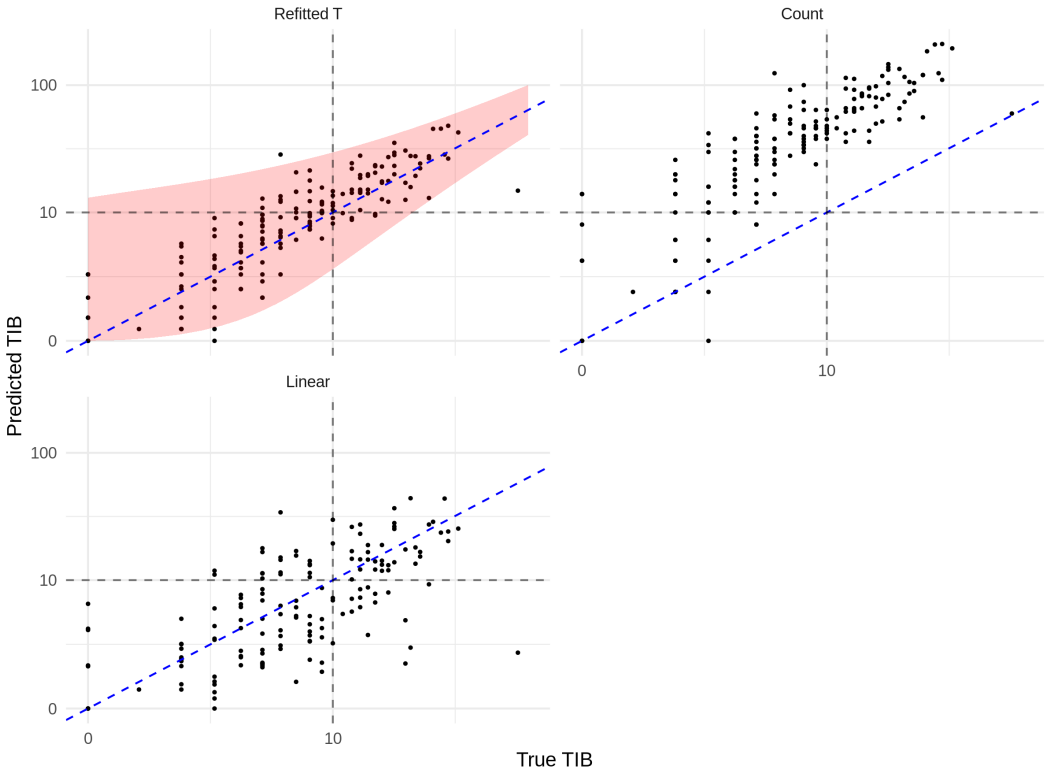

Figure 3.13: Estimation of TIB on the test dataset. Dashed blue (diagonal) line represents perfect prediction, and the grey dashed lines indicate true and predicted TIB values of 10.

3.4 Further testing in other cancer types

The aim of this section is to further demonstrate the performance of our proposed framework in a number of other cancer types. We apply our method for estimating TMB in six more cancer types, namely bladder cancer, breast cancer, colorectal cancer, melanoma, prostate cancer and renal cell cancer. For each cancer, data from two studies are used. Data from the first study is (randomly) split into a training and validation set; the training data is used to construct our estimator for a range of panel sizes, we then evaluate the predictive performance on the validation set (note that in contrast to our analysis in Section 3.3, we do not require a separate test set since the panel size is not selected based on the data). Further, in order to test the robustness of our approach to study effects, for each cancer type, we will also apply our fitted estimator (trained using data from the first study) to predict TMB values for tumours from the second study.

The datasets used (with training, validation and external test sample sizes in parentheses) are from the following studies:

- Bladder cancer: the bladder cancer dataset from the TCGA Pan-Cancer Atlas[4] ($n_{\text{train}} = 300$, $n_{\text{val}} = 109$) and Guo *et al.* (2013) ($n_{\text{test}} = 99$);
- Breast cancer: the breast cancer dataset from the TCGA Pan-Cancer Atlas

[4]data available at https://www.cbioportal.org/study/clinicalData?id=blca_tcga_pan_can_atlas_2018

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

($n_{\text{train}} = 700$, $n_{\text{val}} = 300$) and Kan *et al.* (2018) ($n_{\text{test}} = 187$);

- Colorectal cancer: Giannakis *et al.* (2016) ($n_{\text{train}} = 500$, $n_{\text{val}} = 119$) and Seshagiri *et al.* (2012) ($n_{\text{test}} = 72$);
- Melanoma: Cancer Genome Atlas Network (2015) ($n_{\text{train}} = 250$, $n_{\text{val}} = 96$) and Krauthammer *et al.* (2012) ($n_{\text{test}} = 91$);
- Prostate cancer: Armenia *et al.* (2018) ($n_{\text{train}} = 700$, $n_{\text{val}} = 312$) and Kumar *et al.* (2016) ($n_{\text{test}} = 141$);
- Renal cell cancer: the renal cell cancer dataset from TCGA Firehose[5] ($n_{\text{train}} = 350$, $n_{\text{val}} = 101$) and Guo *et al.* (2011) ($n_{\text{test}} = 98$).

These datasets have a range of mutation rates, specifically the average TMB values in the training datasets are 247 (bladder cancer), 91 (breast cancer), 339 (colorectal cancer), 568 (melanoma), 63 (prostate cancer) and 77 (renal cell cancer).

In Figure 3.14, the black lines plot the R^2 values obtained on the internal validation set from the first study for the six cancer types as the panel size varies from 0.25Mb to 1.25Mb. The blue lines show the R^2 values obtained when predicting TMB for tumours in the external test set from the second study. We see that the performance on the internal validation set is very good and broadly in line with the performance we obtained for the NSCLC dataset (with the exception of the renal cell cancer). The main factor effecting the performance appears to be the overall mutation rate; our method performs very well in cancer types with large mutation rates (colorectal cancer and melanoma), but less well in the cancers with lower overall mutation rates (prostate and renal cell). The performance on the renal cell dataset is particularly poor due to the combination of low sample size and the low average mutation rate.

The results on the external test datasets are more mixed; there is a drop off in performance in comparison with the internal validation results for breast cancer and melanoma, but apparent improvement for prostate cancer. This highlights that study effects, such as differences in patient demographics and clinical profiles, as well as variations in sequencing technologies need to be considered carefully. In practice, one should ensure that the patients in the training data used to fit the model have similar characteristics to the intended test cohort.

[5]data available at https://www.cbioportal.org/study/summary?id=kirc_tcga

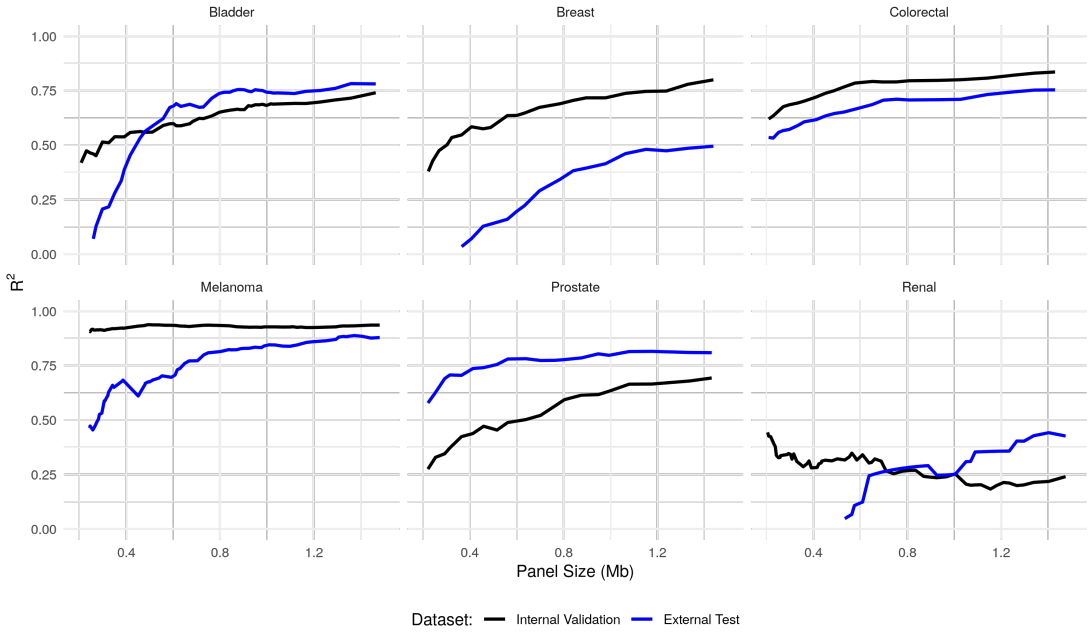

Figure 3.14: The performance of our refitted TMB estimator in the six further cancer types.

3.5 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 datasets from Campbell *et al.* (2016), Hellmann *et al.* (2018a), and Rizvi *et al.* (2015), 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. We further tested the applicability and robustness of our method, by applying it to datasets on several other cancer types. 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.

The main use of TMB is to help identify patients that are more likely to respond to immunotherapy. While TMB is a good single predictor of response (Cao *et al.*, 2019; Zhu *et al.*, 2019), it is of course desirable to improve the predictive performance by including other factors. For instance these may include cancer type (and subtype), specific mutational signatures, aneuploidy and tumor histology, as well as other variables, such as gender, age and exogenous factors. Indeed, Litchfield *et al.* (2021) show that, by including markers of T-cell infiltration and other factors, a multivariate predictor of response to immunotherapy significantly improves the classification performance in comparison to using TMB alone. Nevertheless, one would certainly like to include TMB (or a closely related measure) as a factor in any classifier of response.

Our work also has the scope to help understand mutational processes. For example, the parameters of our fitted model in Section 3.3.1 have interesting

interpretations: of the five genes highlighted in Figure 3.4 as having the highest mutation rates relative to the BMR, two (*TP53, CDKN2A*) are known tumour suppressors (Olivier *et al.*, 2010; Foulkes *et al.*, 1997) and *KRAS* is an oncogene (Jancík *et al.*, 2010). 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. Of course, any such associations need to be carefully investigated in follow up studies.

While there are means by which our framework could be extended, we believe that future work in the field should move from further optimising the prediction of TMB (from which marginal benefit can be gained beyond what has been achieved here and elsewhere) and towards critically evaluating TMB's role as a predictor of response to immunotherapy. In particular, while TMB serves as a good proxy of the impact of somatic mutations on immunotherapy response, more flexible predictions may be made based on specific mutational signatures that can be related directly to survival (as opposed to being related to TMB as an intermediate). The need for this work forms the basis of the investigation in Chapter 4.

Chapter 4

Causal survival analysis in oncology

Overview

In this final chapter, we aim to target the questions motivating Chapter 3 in their broader context. In particular, previously we have proceeded from an established body of literature verifying the utility of TMB as a proxy biomarker for response to immunotherapy. This has been advantageous in that is has allowed us not to concern ourselves with directly showing our derived mutation signatures are associated with improved response to immunotherapy (aside from in some particular cases, such as in Figure 3.11). Indeed, in Chapter 3 we have been able to leverage datasets that did not provide information on response to treatment with immunotherapy. Instead, we've simply aimed to predict TMB as well as we can from targeted panel data (often 'artificial' panel data derived from WES). This has been especially convenient because TMB can be directly calculated from WES data and does not need to be provided by a study as a separate covariate. While aiming to predict TMB has been a helpful simplification, broader methods are required if we wish to more directly interrogate the relationship between genomic signatures (including but not limited to TMB) derived from targeted sequencing and clinical outcomes. Here we discuss in detail strategies for causal estimation of treatment effect on survival from observational studies, the combination of which is an active and evolving field of research. We review some basics of survival analysis and causal inference, describe some recent developments in combining the two, and demonstrate two very different methodological approaches on targeted sequencing data.

4.1 Introduction

In order to more fully understand the relationship between somatic biomarkers such as TMB and clinical outcomes, two substantial hurdles are presented. Firstly, in Chapter 3 we were able to utilise two small datasets for which binary clinical outcomes (corresponding to 'durable clinical benefit') were available. In general the measurement of clinical outcomes is not as simple. By far the most common means of reporting is via survival-like clinical outcomes, including endpoints such as overall survival (OS) or progression-free survival (PFS). These

outcomes are fundamentally restricted by the timespan of a given study period such that not every sample has an associated endpoint observed. In survival analysis this problem is known as *censoring*, and is discussed in Section 4.2. Secondly, our reliance on TMB was motivated by extensive work establishing TMB's clinical utility in randomised treatment scenarios. We often don't have the resources to conduct a large randomised controlled trial (RCT), and so will have to proceed by combining multiple *observational* datasets. This requires a framework for causal inference, and the one we choose is known as heterogeneous treatment effect (HTE).

In this chapter, after describing the fundamentals of survival analysis (Section 4.2) and HTE (Section 4.3) we review recent literature on methods to combine these two techniques in Section 4.4. The main aim of this chapter is to demonstrate an application of these methods to immunotherapy. This is given in Section 4.5, where we combine two somatic mutation datasets to produce flexible estimators of benefit from ICB therapy.

4.2 Survival analysis

4.2.1 Survival times

Survival analysis concerns modelling the amount of time elapsing before a given event occurs. In medical settings, this is often time-to-death analysis, but the same framework is used to approach other events, both medical (e.g. relapse or visit to hospital) and non-medical (machine lifetime, customer churn, etc.). Since our main application cases are concerned with overall survival (OS), we will refer here to a generic 'event' as death. For a patient i, we will refer to survival times with either a lower case t_i (for observations) or upper case T_i (for random variables). As a starting point for motivating techniques in survival analysis, we note the following properties we might expect from survival times:

1. Survival times are *continuous*: $t_i \in \mathbb{R}$. This is intuitively true in many situations, but we typically only measure data in discrete bins (e.g. months). This can introduce complexities such as exact ties in survival time.

2. Survival times are *positive*: $t_i > 0$. This is a consequence of survival times always being relative to some other time point. The choice of this baseline can often vary between (and even within) studies.

The second point above emphasises the importance of an appropriate definition of starting time for interpreting survival times. Defining suitable end points can also present nuanced issues, particularly accounting for, for example, deaths due to competing risks. While this is studied extensively in the literature, in this work we assume a fairly simple modelling scenario with no patients removed from the dataset from external causes. One form of removal/missingness that we certainly do care about, however, is *censoring*.

4.2.2 Censoring

Censoring refers to a structured pattern of non-observance of a given outcome, in this case, the death of a patient. In the studies we will discuss in this chapter, this is typically because the period of observation for the study ended. This is know as *right*-censoring. Because we have not observed the death of the patient, we only have access to a one-sided interval in which this would have occurred. Addressing censoring is one of the core missions of survival analysis.

At this point we introduce some notation. For a patient i, we let $c_i \in \mathbb{R}^+$ (or C_i for a random variable) denote the time at which the patient is removed from further observation, relative to the baseline observation time point for that patient. We therefore observe only the following: $y_i \in \mathbb{R}^+$ (or Y_i for random variable) is the time of final follow-up, i.e. $y_i = \min(c_i, t_i)$; and δ_i (or Δ_i) for censorship status, i.e. $\delta_i = \mathbb{1}\{t_i \leq c_i\}$.

Our general aim, therefore, is to perform inference about the behaviour of the random variables T_i from observation only of a given set of n pairs $\{(y_i, \delta_i)\}_{i=1}^n$. Here we assume that these tuples of observations (and potential future observations) are realisations of independent and identically distributed random variables, and so refer to this underlying joint distribution without reference to sample i, i.e. we may consider the joint variable (T, C, Δ). We discuss in the next section some common practices for performing inference on these targets.

One further concept worth mentioning is that of *informative* vs *non-informative censoring*. In the latter, it is assumed that censoring is independent of survival time, given the covariates. Throughout this chapter, we assume we are considering non-informative censoring for simplicity, although some of the methods presented have mechanisms for adjusting for informative censoring (see, for example, the CSA-INFO method of Chapfuwa *et al.*, 2021).

4.2.3 Modelling approaches

We now describe some common approaches to statistical modelling of survival data. We begin by defining *hazard*. While it is possible to characterise the distribution of survival times T in a variety of ways, hazard is common as it is both interpretable and easy to work with. For the distribution T, we define hazard $h(t)$ as a function of time t as the instantaneous rate of death at time t, given survival up until time t, via:

$$h(t) := \lim_{\delta t \to 0} \frac{\mathbb{P}(T \in [t, t + \delta t] | T > t)}{\delta t}.$$

We also define the *survival* function $S(t)$ as $\mathbb{P}(T > t)$. We may then observe that hazard is expressible as

$$\begin{aligned} h(t) &= \lim_{\delta t \to 0} \frac{\mathbb{P}(T > t) - \mathbb{P}(T > t + \delta t)}{\mathbb{P}(T > t)\delta t} \\ &= \frac{1}{S(t)} \lim_{\delta \to 0} \frac{S(t) - S(t + \delta t)}{\delta t} = -\frac{1}{S(t)} \frac{d}{dt} S(t) \\ &= -\frac{d}{dt} \log S(t). \end{aligned}$$

Finally note that $\frac{dS}{dt} = -f(t)$, where $f(t)$ is the probability density function for T, so we may also write $h(t) = \frac{f(t)}{S(t)}$. This is consistent with the intuition that hazard denotes rate of death, scaled by likelihood of survival so far. Often when fitting predictive models, we use hazard as the target of prediction given inputs x, aiming to estimate $h(t|x)$. Before we move on to this case, however, we'll discuss a few more properties of hazard. It can be useful to define the cumulative hazard function $H(t) = \int_0^t h(u)du$. We may then observe further useful relations between hazard, cumulative hazard, and density. Firstly note that

$$\begin{aligned} H(t) &:= \int_0^t h(u)du \\ &= -\int_0^t \frac{d}{du}\big(\log S(u)\big)du = -\big[\log S(u)\big]_0^t = -\log S(t), \end{aligned}$$

so we can also express $H(t)$ in terms of the survival function. Likewise we can express $f(t)$ simply in terms of hazard via

$$f(t) = h(t)S(t) = h(t)\exp(-H(t)).$$

From this last relationship we can quickly see that under an assumption of *constant hazard* $h(t) = \lambda$, survival times will follow an exponential distribution, with density function $f(t) = \lambda \exp(-\lambda t)$, with mean λ^{-1} and variance λ^{-2}.

However, as mentioned above, a single model for all survival times is not sufficient for most purposes. In general we want to learn about the effect of some covariates on survival time. We therefore introduce a random (potentially vector-valued) variable X taking values in $\mathbb{R}^p$, and define the following conditional versions of survival and hazard functions:

$$\begin{aligned} h(t|x) &:= \lim_{\delta t \to 0} \frac{\mathbb{P}(T \in [t, t + \delta t] | T > t, X = x)}{\delta t} \\ S(t|x) &:= \mathbb{P}(T > t | X = x). \end{aligned}$$

Note that the same relationships as above hold. When making decisions around modelling, assumptions about the behaviour of these two functions are typically the starting point. We'll here evaluate two common assumptions, based on restricting the functional form of each of these two functions respectively.

The proportional hazards assumption

Proportional hazards models are based on the assumption that changes in a given covariate have the effect of increasing hazard by a constant factor across all time points. We state this formally via the requirement that there exist functions $h_0 : \mathbb{R}^+ \to \mathbb{R}^+$ and $\theta : \mathbb{R}^p \to \mathbb{R}^+$ such that

$$h(t|x) = h_0(t)\theta(x).$$

In practice the function θ is often chosen to be $\theta(x) = \exp(-\beta^T x)$ for parameter $\beta \in \mathbb{R}^p$ to be fitted. In this case $h_0(t)$ is left to be fitted flexibly (although this is not necessary for inferring β, making the resulting model (the Cox model; Cox, 1972) semi-parameteric. Note that we can derive the resultant form of the survival function as $S(t|x) = S_0(t)^{\theta(x)}$, where $S_0(t) = \exp(-\int_0^t h_0(u)du)$.

The accelerated failure time assumption

Accelerated failure time models (Wei, 1992) are based on the assumption that changes in a given covariate have the effect of scaling the entire lifetime of a patient by a constant factor. This is formalised by position that there exist functions $S_0 : \mathbb{R}^+ \to [0, 1]$ and $\theta : \mathbb{R}^p \to \mathbb{R}^+$ such that

$$S(t|x) = S_0(\theta(x)t).$$

As for proportional hazards, the function θ is often chosen to be $\theta(x) = \exp(-\beta^T x)$. Again as above, we may use this restriction to infer the form of the hazard function as $\theta(x)h_0(\theta(x)t)$, where $h_0(t) = -\frac{d}{dt}\log S_0(t)$. The function $S_0(t)$ is often chosen to be the exponential survival function e^{-t}.

Note that the proportional hazards and accelerated failure time assumptions are not mutually exclusive. Consider for example the following setting: T is Weibull distributed with parameters λ, k, i.e. has cumulative distribution function $1 - \exp((t/\lambda)^k$. Now suppose that the parameter λ is given by $\lambda = \lambda_0 \exp(\beta^T x)$. We then have that

$$S(t|x) = \exp\big(-t^k \lambda_0^{-k} \exp(-\beta^T x)\big),$$

which satisfies the correct form for the accelerated failure time assumption with $S_0(t) = \exp(-t^k)$ and $\theta(x) = \lambda_0^{-k} \exp(\beta^T x)$. We can also show that the hazard is given by

$$h(t|x) = kt^{k-1}\lambda_0^{-k} \exp(-k\beta^T x),$$

satisfying the proportional hazards assumption with $h_0(t) = k\lambda_0^{-k}t^{k-1}$ and $\theta(x) = \exp(-k\beta^T x)$.

4.2.4 Fitting survival models

The Cox proportional-hazards model The Cox proportional-hazards models (Cox, 1972) remains by far the most commonly employment survival analysis techniques, and likely represents once of the most significant contributions to

statistics of the twentieth century. As would be expected, Cox regression employs the proportional hazards assumption described in the previous section in a semi-parametric linear model of the hazard function. The Cox model can be stated as follows:

$$h(t|x) = h_0(t)\exp(x^T\beta).$$

In the language of Section 4.2.3, we refer to $\theta(x) = \exp(x^T\beta)$ as the 'parametric' portion of the model (parameterised by the vector $\beta \in \mathbb{R}^p$), while the *baseline hazard* function $h_(t)$, when needed, is typically fitted with non-parametric methods. The term semi-parametric is used to describe the combination of these two components. However, the power of Cox regression is that the parameter β can be inferred without the need to fit the non-parametric baseline hazard $h_0(t)$. This is achieved via minimisation of the partial likelihood:

$$\mathcal{L}(\beta) = \prod_{i:\delta_i=1} \frac{\exp(x_i^T\beta)}{\sum_{j:t_j \geq t_i} \exp(x_j^T\beta)},$$

which can be separated from all non-parametric estimates of the model and used as an objective for minimisation with respect to β.

Non-parametric methods While Cox proportional hazards models are by far the most commonly applied methodology underlying the majority of survival analyses, the specific nature of the Cox partial likelihood is not always easy to adapt to scenarios where more flexible survival distributions are sought. This is not to say that the regression function underlying a Cox model cannot be more general; indeed, methods such as *DeepSurv* (Katzman *et al.*, 2018) have made use of deep neural networks to specify a proportional hazards function. However, even in this scenario the proportional hazards assumption can be quite restrictive. Later in this chapter we demonstrate a method for non-parametric survival analysis due to Chapfuwa *et al.* (2018). Their strategy relies on applying a flexible regression function to a known source of simple randomness (e.g. a neural network applied to a multivariate normal source). To fit such models, they propose a loss function that includes an optimisation term to ensure that predicted survival times are *a)* close to true survival times in the case of uncensored data and *b)* greater than the observed censoring time in the case of censored data. To be specific, for a predicted survival time t, observed endpoint time y and censoring status δ they penalise their loss function with

$$\delta|t-y| + (1-\delta)\max\{0, y-t\}.$$

Here we see the first term (applied to uncensored data) penalises estimated survival times far from the observed value, and the second term (applied to censored data) penalises estimated survival times less than the observed censoring time. In their original paper Chapfuwa *et al.* (2018) propose weighting these terms in a tunable manner, while in later work (such as applied to HTE estimation in Chapfuwa *et al.*, 2021) they present them without tuning parameters.

4.3 Causal inference and heterogeneous treatment effects

Heterogeneous treatment effect (HTE) estimation refers to the discovery of interactions between the effect of some intervention, or *treatment*, and some other covariates. In particular we aim to predict, from some input covariates, the change in outcome for a given sample upon applying a given intervention, while by definition never being able to observe a sample when the given treatment has both been applied and not been applied (referred to as the *fundamental problem for causal inference*; Holland, 1986). We take as a starting point the content of introductory Section 1.2.3 on modelling assumptions for causal inference and average treatment effect (ATE). In particular, we continue with the notation of A for a binary treatment variable. In this section we will assume that the space $\mathcal{Y}$ is simple (e.g. the space of binary outcomes $\mathcal{Y} = \{0, 1\}$), while later we will deal with situations in which Y takes composite values as in survival analysis. We first restate the two core propositions of causal inference:

1. *Positivity:* the propensity function is positive almost everywhere, i.e.

$$\pi(x) := \mathbb{P}(A = 1|X = x) \text{ is such that } \pi(X) \in (0, 1) \text{ almost surely.}$$

2. *No unmeasured confounders:* treatment is 'essentially random' given X, i.e.

$$\{Y^{(0)}, Y^{(1)}\} \perp\!\!\!\perp A|X.$$

   This is also sometimes referred to as *strong ignorability*.

4.3.1 Treatment with covariates

We now define HTEs, which will be our tool to make predictions at the individual level. We define the treatment function $\tau : \mathcal{X} \to \mathcal{Y}$ as follows:

$$\begin{aligned}\tau(x) &:= \mathbb{E}[Y^{(1)}|X = x] - \mathbb{E}[Y^{(0)}|X = x] \\ &= \mu_1(x) - \mu_0(x)\end{aligned}$$

where $\mu_a(x) := \mathbb{E}[Y^{(a)}|X = x]$ is the conditional regression function for potential outcome $A = a$. Note that the logic demonstrated in the derivation of the identifiability of the ATE applies to conditional treamtent effect and we can also say that (under Assumptions 1 and 2) we have that $\mu_a(x) = \mathbb{E}[Y|X = x, A = a]$. This means that one viable strategy for estimation of HTEs is simply to estimate μ_0, μ_1 with $\hat{\mu}_0, \hat{\mu}_1$ respectively, and τ with $\hat{\tau}(x) := \hat{\mu}_1(x) - \hat{\mu}_0(x)$. We will discuss this strategy, known as a T-Learner, and its potential shortcomings in the next section.

While we've introduced HTEs in order to be able to make decisions at the individual level, it is worth pointing out that they are in general distinct from individual treatment effects (Vegetabile, 2021). In essence, since heterogeneous effect only requires conditioning on some set of covariates satisfying the strong

ignorability assumption, it need not contain all relevant information about Y. In particular, suppose we are given a variable X satisfying strong ignorability. Then we can form $X' = (X, Z)$ for any other covariate Z and estimate an identifiable HTEs on the basis of X'. To emphasise the difference between truly individual treatment effects and those based merely on conditioning on some strongly ignorable set of variables, some authors use the term *conditional average treatment effect (CATE)* (Vegetabile, 2021).

4.3.2 Meta-learner strategies

As mentioned in the above, we divide strategies for estimating HTEs into general classes of *meta-learner* (Künzel *et al.*, 2019; Curth and Schaar, 2021). Each of these relies on some baseline strategy for fitting a regression function given a training dataset of inputs and outputs (e.g. linear regression, random forests, a neural network), and applies it to some combination of target functions. We describe a few below.

1. T-Learners: this base strategy centred on the identifiability of the conditional regression functions μ_0, μ_1. We simply split our dataset into two portions such that, according to observed values, the treatment variable takes the values $A = 0$ or $A = 1$. We then fit $\hat{\mu}_0(x), \hat{\mu}_1(x)$ separately on these two subsets, and define:

$$\hat{\tau}(x) := \hat{\mu}_1(x) - \hat{\mu}_0(x).$$

2. S-Learners: a small variation on T-Learners. With this approach we jointly fit $\hat{\mu}(x, a)$ such that $\hat{\mu}(x, 0) \approx \mu_0(x)$ and that $\hat{\mu}(x, 1) \approx \mu_1(x)$, then define:

$$\hat{\tau}(x) := \hat{\mu}(x, 1) - \hat{\mu}(x, 0).$$

3. Methods based on *pseudo-outcomes* define some intermediate statistic that can be estimated based on observed data. The simplest is based on the Horvitz-Thompson transformation (Horvitz and Thompson, 1952), also known as inverse-probability weighting (IPW). We define:

$$\tilde{Y}_{IPW} := \Big(\frac{A}{\hat{\pi}(X)} - \frac{1 - A}{1 - \hat{\pi}(X)}\Big) Y,$$

   relying on the observation that, if $\hat{\pi}(x) = \pi(x)$, then $\mathbb{E}[\tilde{Y}_{IPW} | X = x] = \tau(x)$. The advantage of this approach is that the quantity $\tilde{Y}_{IPW}$ only depends on observable quantities, and so can be estimated from the full dataset.

4. Hybrid methods comprise some combination of strategies based on fitting potential regression functions (for example via T-Learners) and pseudo-outcomes. These include X-Learners as proposed by Künzel *et al.* (2019) – for which special cases have been investigated further by Curth and Schaar (2021) – and generalised 'doubly robust' estimators proposed by Kennedy

(2020). In this context, double robustness refers to a learner that produces consistent estimates even if only one of $\hat{\mu}$ and $\hat{\pi}$ are correctly specified.

5. Nie and Wager (2017) proposed an alternate hybrid method, termed an R-Learner, that instead of estimates of propensity $\hat{\pi}(x)$ and conditional regression functions $\hat{\mu}_0(x), \hat{\mu}_1(x)$, uses estimates of $\hat{\pi}(x)$ and a marginal regression function $\hat{\mu}(x) := \hat{\mathbb{E}}[Y|X = x]$. Their approach relies on fitting a function $\hat{\tau}$ minimising the objective

$$\sum_{i=1}^{n} \Big(\big(Y_i - \hat{\mu}(X_i)\big) - \big(A_i - \hat{\pi}(X_i)\big)\hat{\tau}(X_i) \Big)^2$$

where i ranges across the training dataset. This approach later served as the basis for causal forests as developed by Athey *et al.* (2019).

6. Shalit *et al.* (2017) proposed a method based on distributional regularisation such that, by construction, information is shared between estimated regression functions $\hat{\mu}_0$ and $\hat{\mu}_1$. This level of information sharing is tunable via a regularisation parameter that may itself be selected by cross-validation (more on this later).

4.3.3 Double robustness

In settings in which the CATE demonstrates complex structure, it is an important challenge to provide estimators that are demonstrably robust and flexible. This ideally includes robustness in cases of various forms of model misspecification. Before we formally define the doubly robust (DR)-Learner proposed by Kennedy (2022), we heuristically motivate the cases of model misspecification it attempts to address by considering two example simulation settings. As well as motivating the discussion of double robustness that is to come, they also serve to develop our intuition as to when some of the strategies described in Section 4.3.2, in particular T-Learners and estimators based on IPW. They are as follows.

Example 4.3.1. (Simulation Setting 1) We model observations of the triple (X, A, Y), with X and Y taking values in $\mathbb{R}$ and A taking values in $\{0, 1\}$, as independent triplet realisations of the joint distribution described as follows:

$$\begin{aligned} X &\sim \text{Norm}(0, 1) \\ A \mid (X = x) &\sim \text{Bern}\Big(\frac{1}{1 + \exp(-x)}\Big) \\ Y \mid (X = x) &\sim \begin{cases} x^2 + x & \text{where } A = 1 \\ x^2 & \text{where } A = 0 \end{cases} \end{aligned}$$

Example 4.3.1 aims to model a setting in which estimation of the individual conditional regression functions $\mu_a(x)$ is difficult, but estimation of both the propensity function $\pi(x)$ and the CATE $\tau(x)$ is simple. Note that here the functions $\mu_a(x)$ are not intrinsically complex (we could easily fit them with polynomial

regression), but we will attempt to model them with linear regression models than cannot adequately capture them. This is therefore a case of model misspecification rather than one of functional complexity, but we will demonstrate the same points. Kennedy (2022) gives similar examples in which more general nonparametric models are used and the issue is genuinely one of functional complexity without model misspecification, but the underlying phenomena and results are the same.

Example 4.3.2. (Simulation Setting 2) We model observations of the triple (X, A, Y), with X and Y taking values in $\mathbb{R}$ and A taking values in $\{0, 1\}$, as independent triplet realisations of the joint distribution described as follows:

$$\begin{aligned} X &\sim \text{Norm}(0, 1) \\ A \mid (X = x) &\sim \text{Bern}\Big(\frac{1}{1 + \exp(-\cos(2x))}\Big) \\ Y \mid (X = x) &\sim \begin{cases} 3x/4 & \text{where } A = 1 \\ -x/4 & \text{where } A = 0 \end{cases} \end{aligned}$$

Example 4.3.2, by constrast with Example 4.3.1, aims to model a setting in which estimation of the propensity function $\pi(x)$ is difficult, but estimation of the conditional regression functions $\mu_a(x)$ is easy.

These example settings are by design such that for Example 4.3.1 the IPW estimator defined in (3.) will outperform the T-Learner defined in (1.), and vice versa for Example 4.3.2. We find that this does indeed bear out in Figure 4.1. Here we have generated 100 samples from each of the models specified, applied each of the T-Learner and IPW strategies with linear regression models providing each of the estimation procedures (and logistic regression applied to produce an estimated propensity function $\hat{\pi}(x)$), and applied the resultant estimated CATE function to each of the training set data points. The true CATE function (in each case $\tau(x) = x$) is given for comparison. We do indeed see that our expectations are borne out in each of these settings, with the IPW method performing very well on Simulation Setting 1 and the T-Learner performing very well on Simulation Setting 2.

These illustrative cases are representative of two potential sources of error via model misspecification. The aim of double robustness is to create estimators of CATE that work in each of the regimes we've illustrated. Formally, we aim to produce consistent estimates even if one of $\pi(x)$ or $\mu_a(x)$ are misspecified. Practically, this tends to work via weighted combinations of IPW and T-Learner estimates. Kennedy (2022), building upon work from Nie and Wager (2021) and others, define the DR-Learner via the following two steps[1] (again depending upon an arbitrary regression method).

1. Nuisance function training:

[1] Kennedy (2022) define an optional third to incorporate sample splitting and cross-validation, but for simplicity we omit this here.

- Construct estimates $\hat{\pi}$ of the propensity scores π.
- Construct estimates $(\hat{\mu}_0, \hat{\mu}_1)$ of the regression functions (μ_0, μ_1).

2. Pseudo-outcome regression: construct the pseudo-outcome

$$\tilde{Y}_{DR} := \frac{A - \hat{\pi}(X)}{\hat{\pi}(X)(1 - \hat{\pi}(X))}\Big(Y - \hat{\mu}_A(X)\Big) + \hat{\mu}_1(X) - \hat{\mu}_0(X),$$

and regress it on covariates X, yielding

$$\hat{\tau}_{DR}(x) = \hat{\mathbb{E}}[\tilde{Y}_{DR}|X = x],$$

where the symbol $\hat{\mathbb{E}}$ denotes estimated expectation under the chosen regression function.

Under certain conditions, the DR-Learner can be shown to produce optimal convergence rates amongst classes of models based on composition of IPW and conditional regression functions (Kennedy, 2022). We won't go into these conditions in detail here, but can begin to appreciate the underlying ideas by investigating the two extreme cases we've discussed. Firstly, let's assume in Step (1) of the DR-Learner, we correctly estimate propensity (i.e. $\hat{\pi} = \pi$) but do not correctly estimate our conditional regression functions (i.e. $(\hat{\mu}_0, \hat{\mu}_1) \neq (\mu_0, \mu_1)$). We can then write that

$$\begin{aligned}
\tilde{Y}_{DR} &= \frac{A - \pi(X)}{\pi(X)(1 - \pi(X))}\Big(Y - \hat{\mu}_A(X)\Big) + \hat{\mu}_1(X) - \hat{\mu}_0(X) \\
&= \frac{A - \pi(X)}{\pi(X)(1 - \pi(X))}\Big(A\big(Y^{(1)} - \hat{\mu}_1(X)\big) + (1 - A)\big(Y^{(0)} - \hat{\mu}_0(X)\big)\Big) \\
&\qquad + \hat{\mu}_1(X) - \hat{\mu}_0(X)
\end{aligned}$$

and therefore that

$$\begin{aligned}
\mathbb{E}[\tilde{Y}_{DR}|X = x] &= \frac{1}{\pi(x)(1 - \pi(x))}\Big(\mathbb{E}[AY^{(1)} - A\hat{\mu}_1(x) - A\pi(x)Y^{(1)} + A\pi(x)\hat{\mu}_1(x) - \\
&\qquad (1 - A)\pi(x)Y^{(0)} + (1 - A)\pi(x)\hat{\mu}_0(x) \mid X = x]\Big) + \hat{\mu}_1(x) - \hat{\mu}_0(x) \\
&= \frac{1}{\pi(x)(1 - \pi(x))}\Big(\pi(x)\mu_1(x) - \pi(x)\hat{\mu}_1(x) - \pi(x)^2\mu_1(x) + \\
&\qquad \pi(x)^2\hat{\mu}_1(x) - (1 - \pi(x))\pi(x)\mu_0(x) + (1 - \pi(x))\pi(x)\hat{\mu}_0(x)\Big) + \\
&\qquad \hat{\mu}_1(x) - \hat{\mu}_0(x) \\
&= \frac{1}{\pi(x)(1 - \pi(x))]}\Big(\pi(x)\big(\mu_1(x) - \hat{\mu}_1(x) - \pi(x)(\mu_1(x) - \hat{\mu}_1(x))\big) - \\
&\qquad \pi(x)(1 - \pi(x))\big(\mu_0(x) - \hat{\mu}_0(x)\big)\Big) + \hat{\mu}_1(x) - \hat{\mu}_0(x) \\
&= (\mu_1(x) - \hat{\mu}_1(x)) - (\mu_0(x) - \hat{\mu}_0(x)) + \hat{\mu}_1(x) - \hat{\mu}_0(x) \\
&= \mu_1(x) - \mu_0(x) = \tau(x),
\end{aligned}$$

where here in the first equality we are leveraging that $A^2 = A$ and (equivalently)

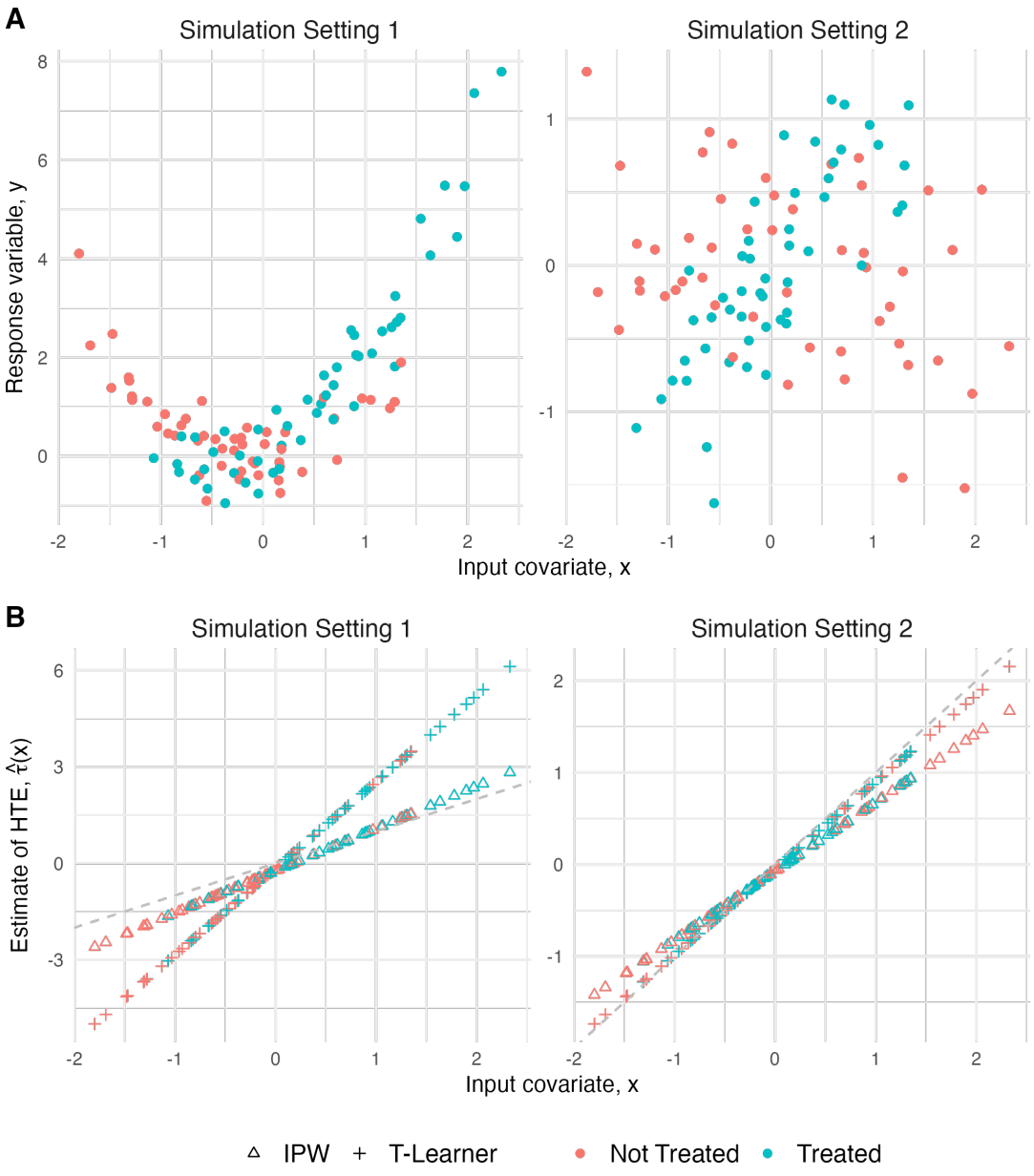

Figure 4.1: Simulation Settings 1 and 2 with **A**: the response variable y, and **B**: estimates of $\hat{\tau}(x)$ for each simulated data point according to the T-Learner (cross) and IPW (triangle) strategies. The true CATE, $\tau(x)$, is shown with a grey dashed line.

that $A(1-A) = 0$, and in the second equality we make use of the fact that $Y \perp A \mid X$ and that $\mu_a(x) := \mathbb{E}[Y^{(a)}|X = x]$.

Secondly, let's now assume we've estimated the conditional regression functions well (so that $(\hat{\mu}_1, \hat{\mu}_0) = (\mu_1, \mu_0)$). Then we can say that

$$\begin{aligned}
\mathbb{E}[\tilde{Y}_{DR}|X = x] &= \mathbb{E}[\frac{A - \hat{\pi}(x)}{\hat{\pi}(x)(1-\hat{\pi}(x)}\Big(A(Y^{(1)} - \mu_1(x)) + \\
&\qquad (1-A)(Y^{(0)} - \mu_0(x))\Big) \mid X = x] \; + \mu_1(x) - \mu_0(x) \\
&= \frac{1}{\hat{\pi}(x)(1-\hat{\pi}(x))}\mathbb{E}[\Big(A(Y^{(1)} - \mu_1(x)) - \hat{\pi}(x)A(Y^{(1)} - \mu_1(x)) - \\
&\qquad \hat{\pi}(x)(1-A)(Y^{(0)} - \mu_0(x))\Big) \mid X = x] + \mu_1(x) - \mu_0(x) \\
&= \frac{1}{\hat{\pi}(x)(1-\hat{\pi}(x))}\mathbb{E}[\Big(\pi(x)(\mu_1(x) - \mu_1(x)) - \\
&\qquad \pi(x)\hat{\pi}(x)(\mu_1(x) - \mu_1(x)) - \\
&\qquad (1-\pi(x))\hat{\pi}(x)(\mu_0(x) - \mu_0(x))\Big) \mid X = x] + \mu_1(x) - \mu_0(x) \\
&= \mu_1(x) - \mu_0(x) = \tau(x),
\end{aligned}$$

where again we've used that $A^2 = A$, $A(1-A) = 0$, the conditional independence of A and $Y^{(a)}$, and the definition of the conditional regression functions $Y^{(a)}$. Without elaborating further on the asymptotic properties of the doubly robust estimator, we can see from the above where robustness to each type of model misspecification arises. We reconsider Examples 4.3.1 and 4.3.2, with the DR-Learner also included (Figure 4.2). While we don't see the DR-Learner perform the outright best in either scenario, we do see it lives up to its doubly robust credentials, performing well in both cases. In scenarios where both propensity and conditional regression estimates are misspecified, we cannot expect faithful prediction of treatment effect.

4.3.4 Causal forests

Before we describe causal forests, which were developed and implemented as a specific instance of so-called 'generalised random forests' (Athey *et al.*, 2019; Athey and Wager, 2019), we first describe briefly the process behind fitting (or 'growing') random forests in general. Random forests were introduced by Breiman (2001) and have been in the standard repertoire of machine learning methods ever since, noted in particular for their strong out-of-the-box performance with little parameter tuning, even when applied to high-dimensional data. Here we'll discuss only the application of random forests to regression problems, although classification forests proceed in a very similar fashion.

Random forest models are *ensembles* (weighted combinations) of *decision trees*. Decision trees rely on a recursive partition of the domain space (in our case $\mathcal{X}$) into rectangular subsets known as 'branches' (see Figure 4.3 for a visual example), with the elements of the final partition referred to as 'leaves'. Within

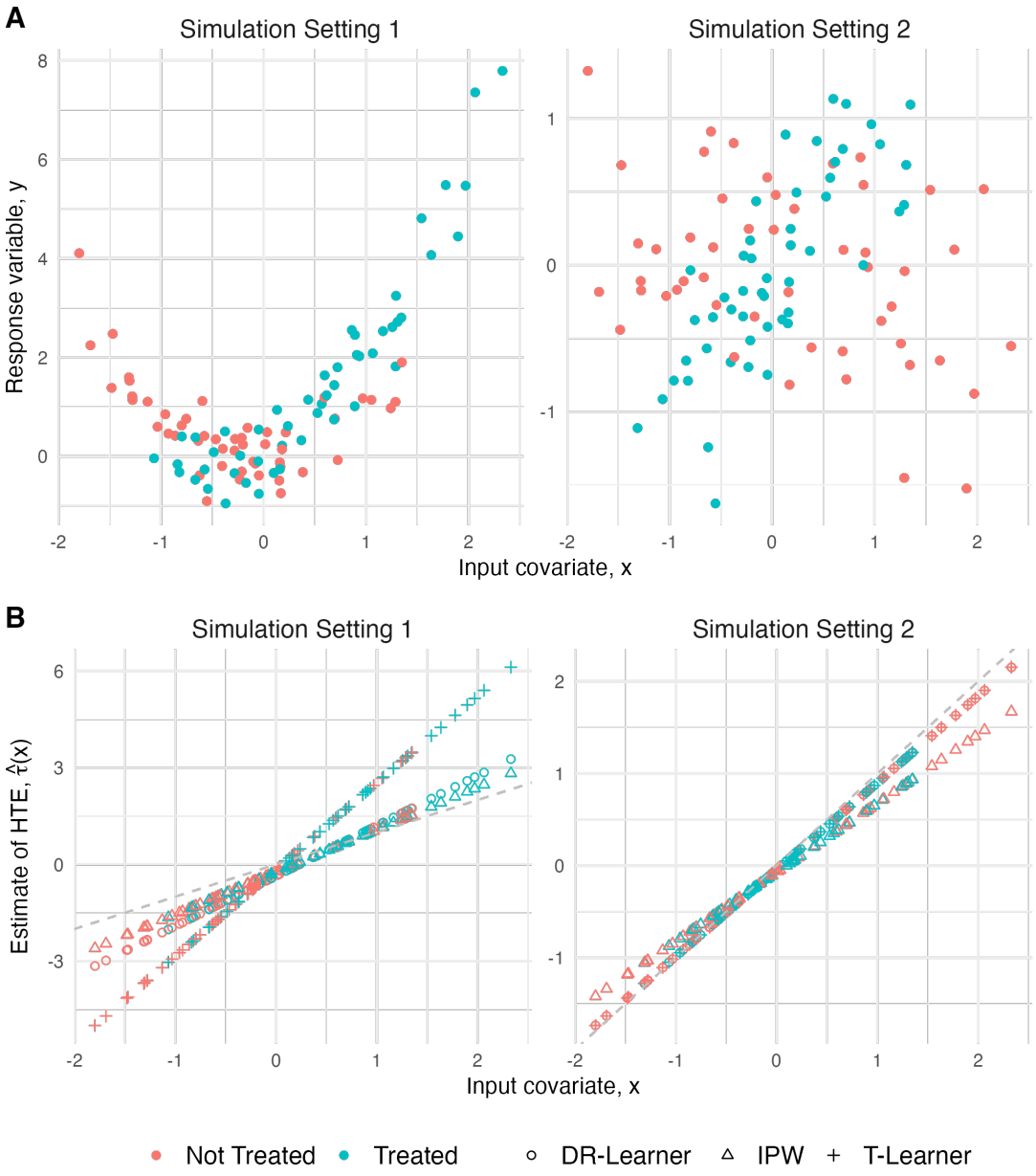

Figure 4.2: Simulation Settings 1 and 2 with **A**: the response variable y, and **B**: estimates of $\hat{\tau}(x)$ for each simulated data point according to the T-Learner (cross), IPW (triangle), and DR-Learner (circle) strategies. The true CATE, $\tau(x)$, is shown with a grey dashed line.

Figure 4.3: Example of a recursive partition of a space into 'leaves' in two dimensions, reproduced from Reeve *et al.* (2021).

each leaf, a representative value is used as the regression output, typically the mean of all training sample output values within that leaf. Branching of the tree is determined by choosing a threshold at which to subdivide the covariate space. This is often chosen so as to minimise the variance of the resultant within-branch training sample outputs. The set of candidate covariates on which to threshold is chosen out of a random subset whose size is a hyperparameter of the algorithm. Other hyperparameters determine how deep and wide a decision tree is allowed to grow. Random forests, as the name suggest, average over the predictions of many trees, each trained on a randomly selected subset of the training data. For a given number of trees B, each tree (generically labelled $b \in \{1, \dots, B\}$) naturally produces a partition function $L_b \colon \mathcal{X} \to 2^{\mathcal{X}}$ which, given a point $x \in \mathcal{X}$, maps to the leaf of the tree b in which x resides, so that for all $x \in \mathcal{X}$ it holds that $x \in L_b(x)$. A prediction for a point x from tree b is given by the weighted average

$$\sum_{i=1}^{n} \frac{\mathbb{1}\{X_i \in L_b(x), i \in \mathcal{S}_b\} Y_i}{|\{j \colon X_j \in L_b(x), j \in \mathcal{S}_b\}|},$$

where $\mathcal{S}_b$ is the random subsample of the training set chosen for tree b. We can then say that a prediction from the entire random forest for a point x is given by

$$\frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{n} \frac{\mathbb{1}\{X_i \in L_b(x), i \in \mathcal{S}_b\} Y_i}{|\{j \colon X_j \in L_b(x), j \in \mathcal{S}_b\}|} = \sum_{i=1}^{n} \alpha_i(x) Y_i,$$

$$\text{where } \alpha_i(x) := \frac{1}{B} \sum_{b=1}^{B} \frac{\mathbb{1}\{X_i \in L_b(x), i \in \mathcal{S}_b\}}{|\{j \colon X_j \in L_b(x), j \in \mathcal{S}_b\}|}.$$

This framing, in which we reverse the order of summation and consider prediction with a random forest as an x-dependently weighted mean of outcomes, will be useful in Section 4.4.1.

Causal forests, produced by an adaptation of the random forests algorithm, are inspired by the R-Learner approach of Nie and Wager (2017). To motivate

this approach we note that, under the standard causal assumptions,

$$\begin{aligned} Y - \mu(X) &= Y^{(1)} + (1-A)Y^{(0)} - \pi(X)\mathbb{E}[Y^{(1)}|X] - (1-\pi(X))\mathbb{E}[Y^{(0)}|X] \\ &= AY^{(1)} + (1-A)Y^{(0)} - \mathbb{E}[Y^{(0)}|X] - \pi(X)\tau(X) \\ &= A(Y^{(1)} - \mathbb{E}[Y^{(1)}|X]) + (1-A)(Y^{(0)} - \mathbb{E}[Y^{(0)}|X]) + (A - \pi(X))\tau(X) \\ &= (Y - \mathbb{E}[Y|X,A]) + (A - \pi(X))\tau(X). \end{aligned}$$

Here the first term $(Y - \mathbb{E}[Y|X,A])$ is the intrinsic variation due to randomness in Y after all controls. Nie and Wager (2017) therefore propose first estimating the propensity function $\hat{\pi}(x)$ and (marginal) regression function $\hat{\mu}(x)$. These are then combined into a target for optimisation, such that $\hat{\tau}$ is selected satisfying[2]

$$\hat{\tau} \in \arg\min_{\tau} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\big(y_i - \hat{\mu}(x_i)\big) - \big(a_i - \hat{\pi}(x_i)\big)\tau(x_i) \right)^2 \right\}.$$

To adapt this method to the regression forest setting, Athey *et al.* (2019) first fit forests for $\hat{\mu}$ and $\hat{\pi}$, and in growing their output forest select for each leaf a constant value of τ minimising the expression above. To eliminate bias, at each point they replace $\hat{\mu}(x)$ and $\hat{\pi}(x)$ with their *out-of-bag* estimates $\hat{\mu}^{(-i)}(x_i)$ and $\hat{\pi}^{(-i)}(x_i)$, where in the given regression forests only leaves not containing the given training sample are used for prediction[3].

4.3.5 Regularisation

Another view of the two heuristic extremes described in Section 4.3.3 is as a balance between 'information sharing' across the estimated conditional regression functions $\hat{\mu}_0$ and $\hat{\mu}_1$, and allowing them to be as flexible as possible in order to capture nuances to the difference between them (i.e. the treatment effect). Shalit *et al.* (2017) proposed a method based on representation learning, in which three functions are learned: $\Phi(x), m_0(x), m_1(x)$, such that $\hat{\mu}_a(x) := m_a(\Phi(x))$. In this way information is shared between $\hat{\mu}_0$ and $\hat{\mu}_1$ via the embedding Φ. The strength of this sharing is determining by applying a regularisation term to the loss function when fitting all three functions simultaneously, which penalises distributional divergence between $\Phi(X) \mid A = 1$ and $\Phi(X) \mid A = 0$.

We let $\tau_{\Phi,m_0,m_1}(x) := m_1(\Phi(x)) - m_0(\Phi(x))$ to refer to the CATE estimate derived from these functions. Here we measure the success of this estimate with expected *precision in estimation of heterogeneous treatment effect*, as defined in Hill (2011):

$$\epsilon_{PEHE}(\Phi, m_0, m_1) := \int_{\mathcal{X}} \big(\hat{\tau}_{\Phi,m_0,m_1}(x) - \tau(x)\big)^2 p(x)dx,$$

where $p(x)$ is the marginal density of the variable X over its domain $\mathcal{X}$. Here we assume that $\Phi\colon \mathcal{X} \to \mathcal{R}$ is a one-to-one twice differentiable function with image $\mathcal{R}$

[2]The original authors also included a regularisation term, which we omit for simplicity.

[3]If not in the context of random forests, 'out-of-bag' may equally be replaced by 'out-of-fold' for cross-validated models.

and inverse $\Psi\colon \mathcal{R} \to \mathcal{X}$. We define $p_\Phi^{A=a}(r) := p_\Phi(r|A=a)$ to be the conditional density of the image $R = \Phi(X)$ under treatment $A = a$.

As mentioned above, our strategy will be to regularise the distributional distance between the distributions $R|A=1$ and $R|A=0$. To do so, we will need an appropriate distance measure. For this, we employ the integral probability metric (IPM) family. These are defined for two density functions p, q over the domain $\mathcal{R}$. For a family G of functions $g\colon \mathcal{R} \to \mathbb{R}$, let

$$\mathrm{IPM}_G(p,q) := \sup_{g\in G} \Big| \int_{\mathcal{R}} g(s)\big(p(s) - q(s)\big)ds \Big|.$$

The IPM metrics include, for an appropriate choice of function family G, the maximum mean discrepancy (Gretton *et al.*, 2008) and the Wasserstein distance (Villani, 2009; Sriperumbudur *et al.*, 2012; Cuturi, 2013). The aim of regularising the distributional distance between the distributions of our two conditional representations is to bound the divergence betwwen 'factual' and 'counterfactual' loss. To do so, we let $L\colon \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$ be a loss function and $\ell_{\Phi,m_0,m_1}\colon \mathcal{X}\times\{0,1\} \to \mathbb{R}$ be the expected loss at the point (x,a), i.e. $\ell_{\Phi,m_0,m_1}(x,a) := \int_{\mathcal{Y}} L(y^{(a)}, m_a(\Phi(x)))p(y^{(a)}|X=x)dy^{(a)}$. We then define factual and counterfactual losses $\epsilon_F(\Phi,m_0,m_1)$ and $\epsilon_{CF}(\Phi,m_0,m_1)$ respectively via

$$\begin{aligned}
\epsilon_F^{a=1}(\Phi,m_0,m_1) &:= \int_{\mathcal{X}} \ell_{\Phi,m_0,m_1}(x,1)p^{a=1}(x)dx,\\
\epsilon_F^{a=0}(\Phi,m_0,m_1) &:= \int_{\mathcal{X}} \ell_{\Phi,m_0,m_1}(x,0)p^{a=0}(x)dx,\\
\epsilon_{CF}^{a=1}(\Phi,m_0,m_1) &:= \int_{\mathcal{X}} \ell_{\Phi,m_0,m_1}(x,0)p^{a=1}(x)dx,\\
\epsilon_{CF}^{a=0}(\Phi,m_0,m_1) &:= \int_{\mathcal{X}} \ell_{\Phi,m_0,m_1}(x,1)p^{a=0}(x)dx,\\
\epsilon_F(\Phi,m_0,m_1) &:= u\epsilon_F^{a=1}(\Phi,m_0,m_1) + (1-u)\epsilon_F^{a=0}(\Phi,m_0,m_1),\\
\epsilon_{CF}(\Phi,m_0,m_1) &:= (1-u)\epsilon_{CF}^{a=1}(\Phi,m_0,m_1) + u\epsilon_{CF}^{a=0}(\Phi,m_0,m_1),
\end{aligned}$$

where $u := \mathbb{P}(A=1)$. Shalit *et al.* (2017) prove that under certain conditions on the likelihood ℓ_{Φ,m_0,m_1} and the function family G, there exists a constant B_Φ such that the counterfactual and factual losses defined above can be related via

$$\begin{aligned}
\epsilon_{CF}(\Phi,m_0,m_1) \le (1-u)\epsilon_F^{a=1}(\Phi,m_0,m_1) + u\epsilon_F^{a=0}(\Phi,m_0,m_1)\\
+ B_\Phi \mathrm{IPM}_G(p_\Phi^{a=1}, p_\Phi^{a=0}).
\end{aligned}$$

Their proof relies on Φ having a well-defined inverse Ψ, although in practice this is not always chosen to be the case. The key result here is that we have now bounded the counterfactual loss (by definition unobserved) with quantities that can be directly estimated from observed data; namely, the estimated average loss on treated patients, the estimated average loss on untreated patients, and the observed distributional distance between the two conditional representations. As

a direct consequence of this, Shalit *et al.* (2017) are able to show that

$$\epsilon_{PEHE}(\Phi, m_0, m_1) \leq 2\Big(\epsilon_F^{a=1}(\Phi, m_0, m_1)+\epsilon_F^{a=0}(\Phi, m_0, m_1)+B_\Phi \text{IPM}_G\big(p_\Phi^{a=1}, p_\Phi^{a=0}\big)-K\Big),$$

where K does not depend on the choices of functions Φ, m_0, m_1. They therefore propose empirically minimising the objective

$$\frac{1}{n}\sum_{i=1}^{n} w_i L(m_{a_i}(\Phi(x_i)), y_i) + \alpha \text{IPM}_G\big(\{\Phi(x_i)\}_{i:a_i=0}, \{\Phi(x_i)\}_{i:a_i=1}\big),$$

$$\text{where } w_i = \frac{a_i}{2u} + \frac{1-a_i}{2(1-u)}, u = \frac{1}{n}\sum_{i=1}^{n} a_i.$$

Here α is now a tunable parameter to be optimised by (for example) cross-validation.

4.4 Heterogeneous treatment effects applied to survival analysis

Having summarised several concepts and approaches for HTE estimation, we now turn to two specific (and very different) examples from recent literature on adapting estimation methods to the prediction of discrepancies in survival. These follow on from the last two examples from the previous section, namely causal forests and neural network-based distributional regularisation. As well as using different underlying methodologies, they produce starkly different outputs. Causal survival forests, like causal forests in general, learn only the target function $\hat{\tau}$, whereas the regularisation approach learns two distinct stochastic functions for survival outcomes. Notably, neither are based on a parametric statistical model for survival such as proportional hazards or accelerated failure time.

4.4.1 Forests again

Cui *et al.* (2022) propose causal survival forests (CSFs) for HTE estimation in the presence of censored data. Based on causal forests and implemented via the same R package `grf` (Tibshirani *et al.*, 2023), their first major contribution is to propose a weighting scheme similar to the IPW approach introduced in Section 4.3.2. They then proceed to describe a correction for double robustness under misspecification of the learned censoring function. This corrective term is too technically verbose to describe in full here, so having introduced the concept of double robustness in Section 4.3.3 we limit ourselves here to an outline of what is achieved by this correction.

Firstly, we note that in most medical (or indeed any other) survival analyses, some maximum time horizon for uncensored data is often present. For example, in a clinical study spanning ten years, any time points greater than ten years must be censored. We may not, then, expect to be able to recover any meaningful treatment effect for this range of values. Cui *et al.* (2022) therefore introduce

a *time horizon* h, a deterministic transformation γ, and attempt to estimate an altered HTE funtion

$$\tau^{\gamma,h}(x) := \mathbb{E}[\gamma(T^{(1)}, h) - \gamma(T^{(0)}, h)|X = x].$$

Here the transformation γ is restricted to satisfy $\gamma(t, h) \geq \gamma(h, h)$ for all $t \geq h > 0$. The choice of the function γ induces variations in the effect being measured, with $\gamma(t, h) = \min\{t, h\}$ inducing estimation of difference in restricted mean survival time (RMST) and $\gamma(t, h) = \mathbb{1}\{t \geq h\}$ inducing estimation of differences in survival probability. In our later analyses we investigate the application of CSFs to estimation heterogeneous effects on survival probabilities with time horizons of one and two years.

In its simplest form the method of Cui *et al.* (2022) draws inspiration from inverse probability of censoring weighting (IPCW), the core tenet of which is to first estimate the *censoring function*

$$S_a^C(s|x) = \mathbb{P}(C \geq s|A = a, X = x).$$

We refer to this estimate as $\hat{s}_a(t|x)$. We may do this using only censored data. We could then use uncensored data to fit a causal forest for τ. In the language introduced in Section 4.3.4, we could express this forest as a weighted sample-wise sum with weightings

$$\alpha_i(x) = \frac{1}{B}\sum_{b=1}^{B} \frac{\mathbb{1}\{X_i \in L_b(x), i \in \mathcal{S}_b\}}{|\{i : X_i \in L_b(x), i \in \mathcal{S}_b\}|}\}.$$

We may then simply replace this weighting $\alpha_i(x)$ with $\alpha_i(x)/\hat{s}_{A_i}(\gamma(T_i, h)|X_i)$. This has the effect of more highly prioritising observations which were unlikely to have been uncensored, and has been shown empirically to be broadly successful in eliminating censoring bias (Van Der Laan and Robins, 2003).

While the IPCW strategy may be a good baseline, it suffers from two key drawbacks: firstly, each of its censoring and treatment functions can only be learned on a subset of the available training data. Secondly, it is vulnerable to misspecification/poor fit of the censoring function $\hat{s}_a(t|x)$. The solution proposed by Cui *et al.* (2022) entails the joint estimation of several more 'nuisance' functions detailing different aspects of the censoring distribution. These can be estimated using all the available data, and taken together are shown to provide double robustness – an analogue to the double robustness demonstrated in Section 4.3.3. Here the two potential sources of misspecification are the censoring function and the propensity function.

4.4.2 Regularisation again

Our second example of a modern strategy for enabling survival HTE estimation returns to the distributional regularisation method proposed by Shalit *et al.* (2017) and discussed in Section 4.3.5. Chapfuwa *et al.* (2021) proposed an ex-

tension, counterfactual survival analysis (CSA)[4], applicable to survival data that retains the same overall structure: learning a representation $\Phi\colon \mathcal{X} \to \mathcal{R}$ of input data X taking values in $\mathcal{X}$ that serves to 'share information' between two conditional regression functions defined by compositions of the representation function Φ with treatment-specific functions m_0, m_1. The same regularisation term, based on integral probability metric (IPM) families which include the Wasserstein distance, is used to force the learned representations $R = \Phi(X)$ to be distributionally similar regardless of the assigned treatment.

The required extension to survival analysis is provided by replacing the deterministic functions m_0, m_1 with flexible regression functions (in this case provided by neural networks) dependent on a separate random input ε. This allows for a non-parametric specification of the distribution of survival times, as described in Section 4.2.4. The adapted survival loss from Chapfuwa *et al.* (2018) is used, giving a strategy defined by minimising the objective

$$\frac{1}{n}\sum_{i=1}^{n} w_i L(m_{a_i}(\Phi(x_i), \varepsilon), y_i, \delta_i) + \alpha \mathrm{IPM}_G\big(\{\Phi(x_i)\}_{i:a_i=0}, \{\Phi(x_i)\}_{i:a_i=1}\big),$$

$$\text{with } L(t, y, \delta) := \delta|y - t| + (1 - \delta)\max\{0, y - t\}$$

and w_i, u defined as in Section 4.3.5. In the experiments presented in later sections, we use an adapted PyTorch (Paszke *et al.*, 2019) implementation of these metods provided by (Chapfuwa *et al.*, 2021).

4.5 Application to immunotherapy

We now put the methods described across previous sections into practice by demonstrating the use of two techniques for causal survival analysis in an oncology setting. To do so, we use publicly available clinical and somatic mutation data combining studies in which immunotherapy treatment (ICB therapy based on PD-L1 and CTLA-4 checkpoints) has been used and in which it hasn't. This is used more as a platform to assess the strengths and vulnerabilities of each method than as a demonstration of novel findings. Indeed, the limitations of these datasets (discussed in detail later in this section) serve as good talking points for the difficulty in applying these methods in practice. We begin by describing our data sources, then proceed to give results firstly of validating the role of TMB as a biomarker and then of producing more general somatic mutation signatures associated with immune response. We condlude with a discussion of the pros and cons of each method and their aforementioned shared limitations.

4.5.1 Data sources

We compare two datasets comprising mutation and clinical data derived from late stage cancer patients. The first dataset, produced by Zehir *et al.* (2017), is

[4]They also provide a further extension, CSA-INFO, to deal with informative censoring; we omit this for simplicity.

derived from a large cohort treated at the memorial Sloan Kettering cancer centre (MSKCC). It was produced in order to inform decisions around experimental treatment regimes for patients whose tumours had been refractory to standard care. The second, produced by Hellmann *et al.* (2018a), describes a smaller (but still substantial) cohort of patients at similar disease stages who were then treated with ICB. While there are differences between the two datasets in both the availability of clinical data and the specifics of patient trajectory, the two studies provide usefully comparable data sources for patients that were and were not treated by immunotherapy. Crucially, since both analyses were performed using the MSK-IMPACT targeted sequencing panel, their mutation profiling is very comparable. We describe further characteristics of each dataset below.

Zehir et al.

This study by MSKCC provides somatic mutation data, as well as some clinical covariates, for over 10,000 patients with late stage cancer of a variety of types. These patients were, by entry criteria, those who would be eligible to participate in clinical trials for experimental drugs. After the sequencing study was conducted, a minority (11%; 527 patients) did indeed go on to be matched to further trials on the basis of their mutation profile. OS statistics are provided relative to a baseline given by the time point at which the procedure for recovering the sequencing sample was performed. Many samples were taken from metastatic sites, and some patients had multiple samples recovered. The fact that all tumours sequenced in the study were those of late stage cancers is emphasised by the authors as allowing a fuller picture of the mutations associated with cancer development and response to first-line treatment. Clinical features available include detailed cancer type information, metastatic status, sex, and smoking history.

Hellman et al.

This dataset describes advanced cancer patients who did receive ICB treatment based on inhibition of PD-L1, CTLA-4, or both. For these patients, OS was measured from the date of first ICB treatment. Clinical features available include sex, age, and drug type.

4.5.2 Validation with TMB

We begin by investigating the extent to which the methods we've described for HTE estimation can recapitulate the known relationship between TMB and benefit from immunotherapy. This is a valuable benchmark for any modelling approach, because basing predictions on TMB alone (alongside any available general clinical features) means working with an input of low dimension, which should present a smalelr challenge to any estimation method.

We begin by applying the causal survival forests method of Cui *et al.* (2022). To restrict the focus of this demonstrative application, we consider only NSCLC patients from each of the datasets described above. To simplify matters further,

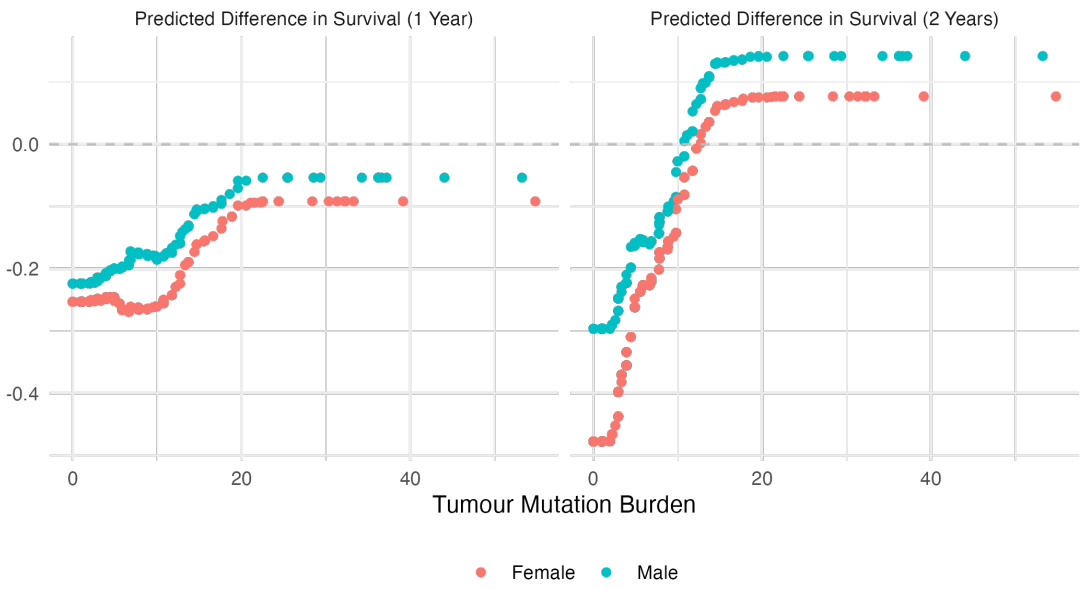

Figure 4.4: Out-of-bag CSF predictions of difference of probability of one- and two-year survival (left and right panels respectively) after treatment with immunotherapy for non-small cell lung cancer patients.

we consider only patients without missing data for time-to-event, censoring status, sex, or TMB values, and where multiple samples are available for a given patient we take only the sample with the highest sequencing coverage. After these restrictions we are left with samples from 129 immunotherapy-treated patients, and 603 not treated with immunotherapy. We take 75% of these samples (96 immunotherapy, 452 non-immunotherapy) as a training dataset. We apply causal survival forests to predict differences in probability of survival after one and two years, as described in Section 4.4.1.

We present out-of-bag predictions of difference in survival probability for survival after one and two years for each patient in the training set using this method in Figure 4.4. For two year survival we see a substantial improvement in expected benefit from immunotherapy, with increased TMB at low BMR, before a plateau at around 20mut/Mb. Sex, the only clinical covariate available across both datasets (more on this later), appears to play a role fairly independently from TMB. Within this trend, male patients appear to be consistently benefiting more from immunotherapy across TMB values.

Interestingly, while a strong dependence on TMB is observed, the highest level of estimated benefit from immunotherapy is fairly low, and indeed for the majority of patients with low TMB we estimate negative benefit from immunotherapy treatment. We hypothesise that this is due to uncaptured differences in typical clinical features between the two datasets. As mentioned above, to perform this analysis we were only able to utilise clinical factors that were recorded across both studies. As a conglomeration of multiple studies itself, the Zehir *et al.* (2017) dataset in particular did not make available several pieces of data that would be expected to impact disease trajectories significantly. The two most important of

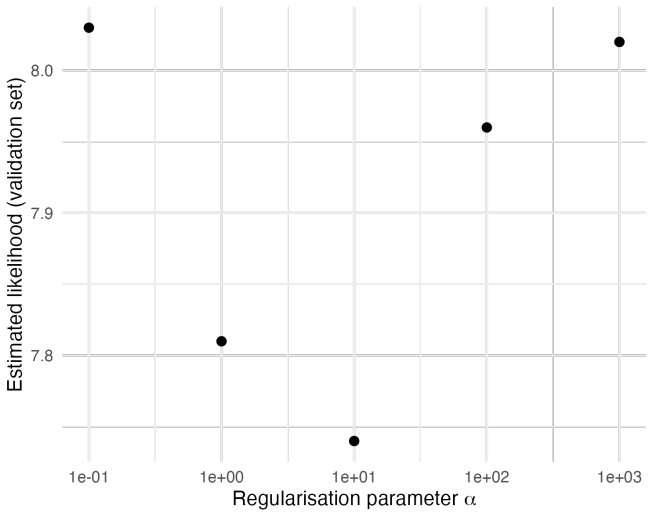

Figure 4.5: Selection of α penalisation parameter for CSA model. A value of $\alpha = 10^1$ was selected.

these, which would certainly be necessary to strengthen any associations learned here, would be patient age, and disease stage. Systematic variation in either would be likely to impact the difference in average rate of two-year survival between the two data sources and skew overall estimation of immunotherapy benefit (even if the relationship with TMB was accurately captured). This same trajectory is observed for one-year survival but the lack of overall ascribed benefit to immunotherapy is even clearer; in this case for no values of TMB does the inferred survival benefit exceed net neutrality. The overall variation in benefit with increased TMB is also lower, despite the similar shape of the inferred response curve. It is not clear why the overall (presumed) underestimate of benefit from immunotherapy is so much more pronounced after twelve months than twenty four. It has been noted that ICB-type immunotherapy can particularly effective at producing long-term survival benefits to the subset of patients who receive some benefit (Putzu *et al.*, 2023). This, if the case here, may explain to some degree this qualitative difference.

We also fitted neural network-based conditional regression functions as described in Section 4.4.2. This was done performed using a neural network with a single hidden layer of dimension 10 and an output also of dimension 10 for $\Phi(x)$. The functions m_0, m_1 were also neural networks with a single hidden layer of dimension 10. The random input ε was of size ten, with each element independently distributed according to $U[0, 1]$. The Sinkhorn distance (Cuturi, 2013) was used for IPM regularisation. We performed selection of the regularisation parameter α by likelihood minimisation on the validation set across the range $\alpha \in \{10^{-1}, 10^0, 10^1, 10^2, 10^3\}$. We report the output of these runs in Figure 4.5, and selected $\alpha = 10^1$. We then used this fitted model to sample 200 estimated survival times under each treatment for covariates corresponding to each patient in the training set, and from these calculated the estimated differences in survival probability after one and two years (as for the causal forests example above). These predictions are given in Figure 4.6.

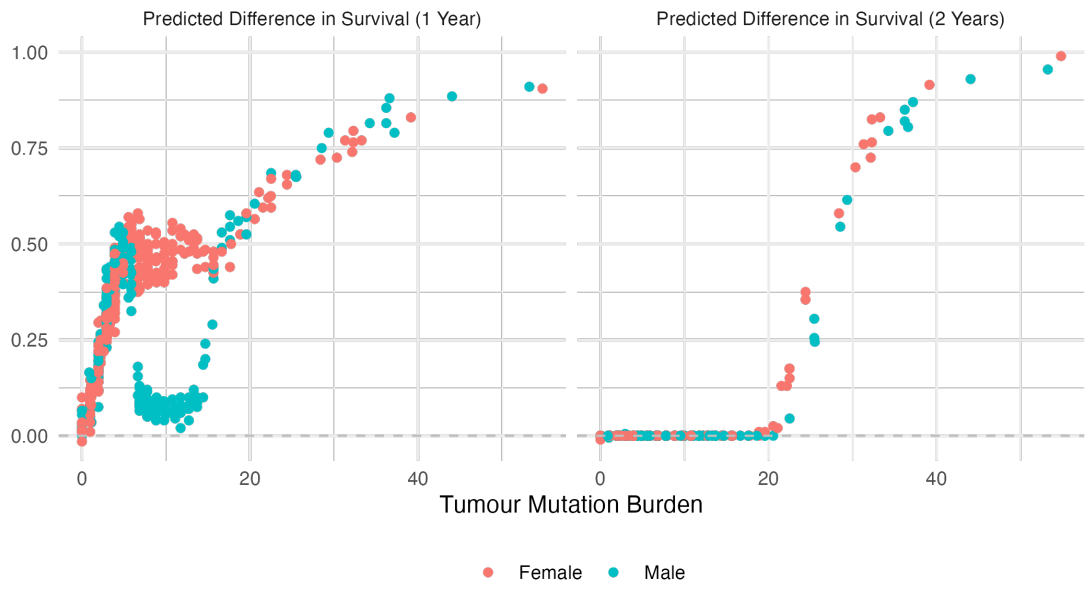

Figure 4.6: CSA neural network predictions of difference of probability of 1- and 2-year survival (left and right panels respectively) after treatment with immunotherapy for non-small cell lung cancer patients.

Firstly, we see a clearly similar overall picture of response to increased TMB in estimated treatment benefit as in the previous analysis. In both cases benefit rises with TMB, with a plateau after around 20mut/MB in the case of one-year survival. Strikingly, the predicted range of benefit is much higher than in the causal forests analysis; here for both one- and two-year survival there are very few samples with a predicted negative treatment effect. This may well be principally driven by more pessimistic predictions of overall survival time. Particularly in the two-year survival case, many input data points are assigned zero probability of survival beyond two years under either treatment regime. This is difficult to compare directly with causal survival forests, since the latter method does not provide predictions of survival, instead providing simply an prediction of the difference in survival probabilities.

4.5.3 Exploring more general markers

In order to explore signatures of response depending on more than just on clinical factors and TMB, we include nonsynonymous mutation counts for each gene in the MSK-IMPACT gene panel. This panel consists of 341 gene targets. Some samples in the given datasets were profiled with more extensive gene panels, so we restricted our analysis only to the genes covered by all panels. We fitted CSFs based on these mutation counts, patient sex, and TMB. For this application case forests were chosen in order to leverage easily interpretable measures of covariate importance. We computed these variable importance measures for each gene and clinical covariate included in the joint datasets, with all non-zero values shown in Figure 4.7.

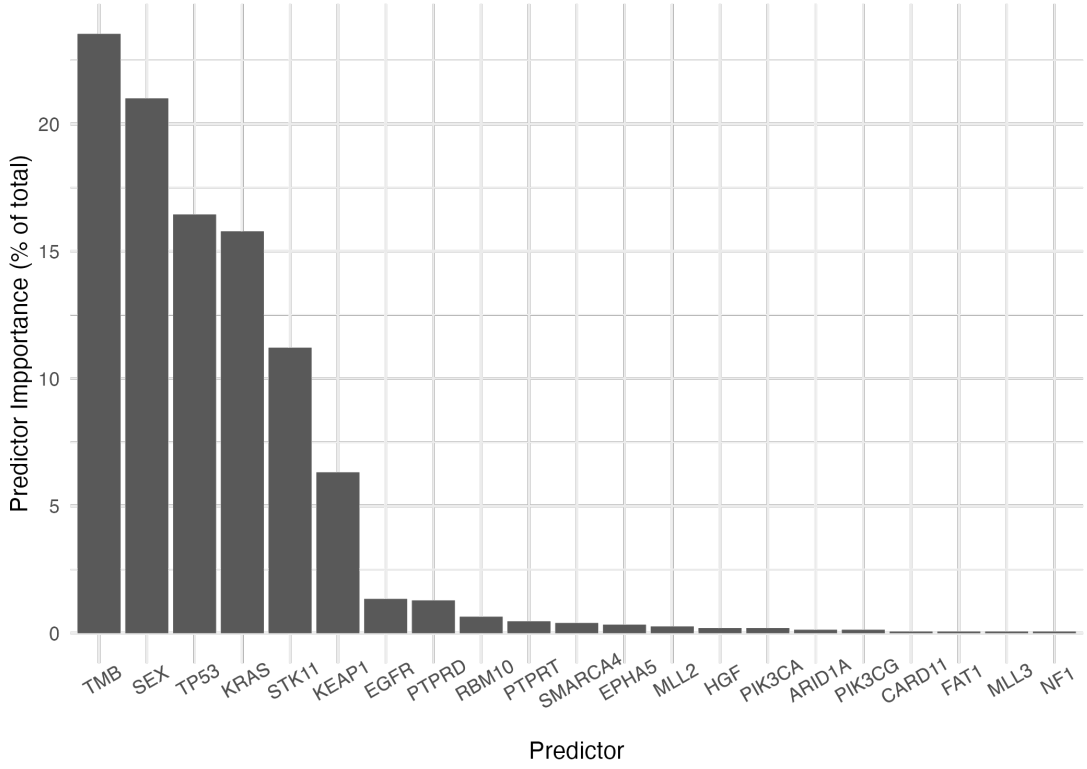

Figure 4.7: Top ∼20 genomic and clinical predictors of heterogeneous survival treatment effect, alongside their estimated within-forest importance metrics.

We see that, while no individual gene target is of the predictive importance of TMB or sex in this larger combined model, but several come close. These include several usual suspects such as *TP53*, *KRAS*, *STK11*, and *KEAP1*. In order to understand the relationship between a gene's influence on the causal survival forest model and its BMR, we compared variable importance scores with total mutation counts across the training dataset for each gene in the MSK-IMPACT panel, with outputs shown in Figure 4.8. While we observe some anomalies, such as *EGFR*, we do see broadly that all highly weighted genes are those with elevated mutation rate.

4.6 Conclusions

In this work we've utilised recent developments for causal survival analysis in oncology. Our first aim was to validate these methods by recapitulating the known causal dependence of immunotherapeutic treatment effect on TMB. This we have done successfully; for each method and prediction target employed we have clearly seen a positive trend in the relationship between immunotherapy benefit and TMB, notwithstanding differences in calibration of overall benefit. We've then taken first steps towards exploring the utility of more generic mutation signatures than TMB in theranostic decision settings. This, similarly, has garnered results than pass a 'sniff check' for viability, recapitulating the impact of several known

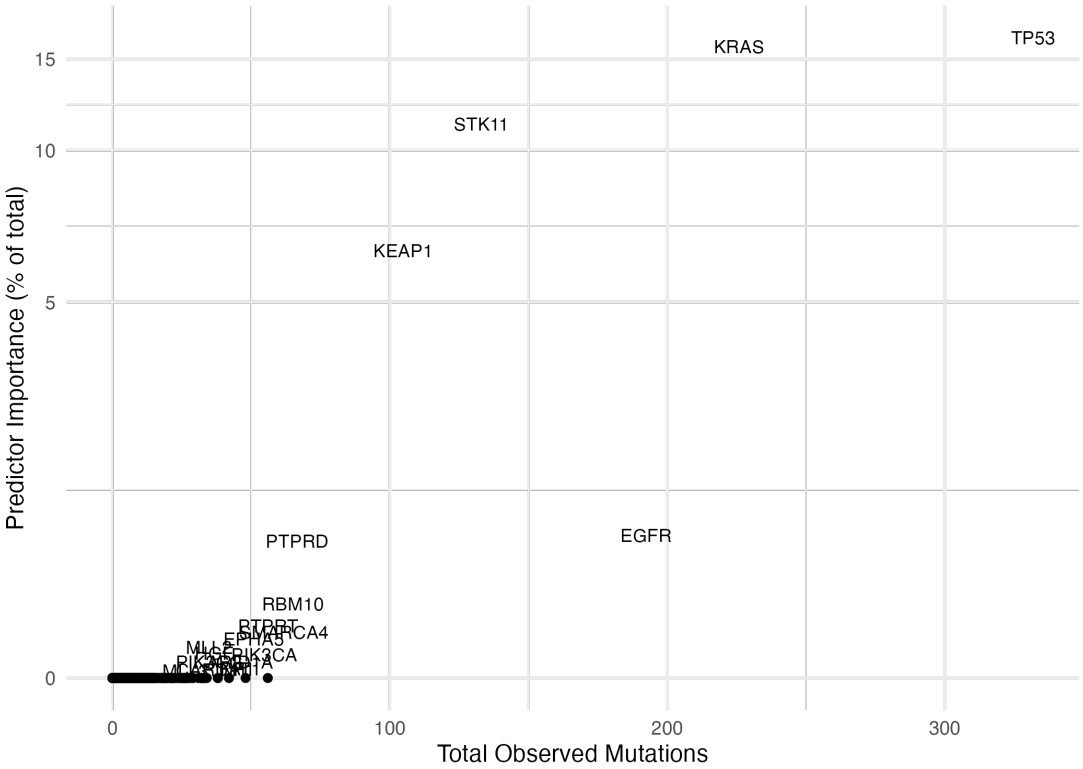

Figure 4.8: Total number of mutations observed in each gene target across the training set, vs percentage of predictive importance in the resultant causal survival forest ascribed to each target (targets with importance > 0 labelled).

driver genes. It remains for further investigation to see how well this method can detect high-impact, low-frequency mutations.

The principal limitation of this pilot study, which is common amongst attempts to bring together disparate clinical datasets, is the poor consistency of available clinical features. Here we were able to use patient sex to demonstrate the role of a shared, potentially confounding, covariate, but by itself this is clearly inadequate. It is hard to imagine that none of age, cancer progression stage, or prior treatment would play a confounding role in the propensity of patients to belong to one treatment arm or the other. It seems likely that patients enrolled in immunotherapy trials are in general more sick, have recieved more previous lines of treatment, and have been more refractory to the treatments that they have received. This may to some extent explain an overall underestimate of immunotherapy benefit. Intriguingly, this seems to be shown far more clearly in CSF analysis than CSA. Further work is required to understand exactly what leads to these discrepancies in estimation of overall average treatment effect, and to what extent it may be representative of an increased sensitivity to confounding by the random forests approach.

In attempting to decompose the reasons for prediction differences between the two methods exhibited, it is worth noting their very different properties with respects to interpretability and adaptability. While on the whole random forests are typically considered more interpretable than neural networks (particular with a subtly controlled joint representation layer), causal survival forests predict only the given treatment effect whereas Chapfuwa *et al.* (2021)'s method effectively produces generative models for survival under each of the treatment arms. In the case of the CSA-INFO method (omitted in this chapter for simplicity), we attain further treatment-specific generative models of censoring status. In practice this means that it is far easier to untangle these model's behaviours in different setting. It also greatly simplifies the procedure of model fitting for a particular target (e.g. difference in two-year survival probability). While CSFs require refitting for each potential time horizon h and transform γ, all of this analysis can be performed via sampling after fitting a CSA model.

In the context of causal prediction models, validation in general provides a very difficult challenge. While there are natural ways to validate the performance of each of these model types, they are natural attuned to the performance of their own model type and may translate poorly to the other. In general two strategies have emerged for cross-model comparison: firstly, simulated or semi-simulated analyses; and secondly, analyses geared towards recapitulating a known phenomenon or causal association. In our case we have chosen the latter, although more extensive simulation studies comparing these two methods alongside more simple procedures described earlier in the chapter would also be beneficial.

As a final note, in general the more complex the mechanisms underpinning a model's training, prediction, and validation, the more difficult it is for such a model to attain useful real-world implementation. It is out of scope for our work here to give a full analysis of difficulties in translating clinical prediction models beyond single studies all the way to deployment. However, these difficulties are reflected in the specific (and in our case largely unmet) demands on data type and quality to enable truly reliable usage of the models described. This comes

even though the stated purpose of such models is, to some extent, to ease the need for expensive and arduous collection of controlled data tailored to the desired prediction task. It remains to be seen whether there is a place for models complex enough to deal with messy and only partially cohesive 'real-world data', or whether their complexity will render them fragile enough that they would only become practically useful when sufficient data collection has occurred to enable simpler analyses regardless.

Bibliography

Abadie, A. (2005). Semiparametric Difference-in-Differences Estimators. *The Review of Economic Studies*, **72**(1), 1–19.

Adragni, K. P. and Cook, R. D. (2009). Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **367**(1906), 4385–4405.

Afshar Bakshloo, M. *et al.* (2022). Nanopore-Based Protein Identification. *Journal of the American Chemical Society*, **144**(6), 2716–2725.

Almasi, M. A. and Almasi, G. (2017). Loop Mediated Isothermal Amplification (LAMP) for Embryo Sex Determination in Pregnant Women at Eight Weeks of Pregnancy. *Journal of Reproduction & Infertility*, **18**(1), 197–204.

Almeida, F. and Xexéo, G. (2019). Word Embeddings: A Survey. arXiv:1901.09069 [cs, stat].

Armenia, J. *et al.* (2018). The long tail of oncogenic drivers in prostate cancer. *Nature Genetics*, **50**(5), 645–651.

Athey, S. and Wager, S. (2019). Estimating Treatment Effects with Causal Forests: An Application. arXiv:1902.07409 [stat].

Athey, S. *et al.* (2019). Generalized random forests. *The Annals of Statistics*, **47**(2), 1148–1178.

Avsec, Z. *et al.* (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, **18**(10), 1196–1203.

Awad, K. *et al.* (2019). The Precision Medicine Approach to Cancer Therapy: Part 1 — Solid Tumours. *Pharmaceutical Journal*.

Baltimore, D. (1970). RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature*, **226**(5252), 1209–1211.

Barbour, D. L. (2019). Precision medicine and the cursed dimensions. *npj Digital Medicine*, **2**(1), 1–2.

Bayarri, M. J. and Berger, J. O. (2004). The Interplay of Bayesian and Frequentist Analysis. *Statistical Science*, **19**(1), 58–80.

Becherer, L. *et al.* (2020). Loop-mediated isothermal amplification (LAMP) – review and classification of methods for sequence-specific detection. *Analytical Methods*, **12**(6), 717–746.

Bedard, P. L. *et al.* (2020). Small molecules, big impact: 20 years of targeted therapy in oncology. *The Lancet*, **395**(10229), 1078–1088.

Bentley, D. R. *et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**(7218), 53–59.

Berihuete, A. *et al.* (2021). A Bayesian Model of COVID-19 Cases Based on the Gompertz Curve. *Mathematics*, **9**(3), 228.

Bewicke-Copley, F. *et al.* (2019). Applications and analysis of targeted genomic sequencing in cancer studies. *Computational and Structural Biotechnology Journal*, **17**, 1348–1359.

Boivin, A. and Vendrely, R. (1947). Sur le rôle possible des deux acides nucléiques dans la cellule vivante. *Experientia*, **3**(1), 32–34.

Bojanowski, P. *et al.* (2017). Enriching Word Vectors with Subword Information. arXiv:1607.04606 [cs].

Bon, J. J. *et al.* (2023). Being Bayesian in the 2020s: opportunities and challenges in the practice of modern applied Bayesian statistics. arXiv:2211.10029 [stat].

Borghaei, H. *et al.* (2021). Five-Year Outcomes From the Randomized, Phase III Trials CheckMate 017 and 057: Nivolumab Versus Docetaxel in Previously Treated Non-Small-Cell Lung Cancer. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, **39**(7), 723–733.

Boveri, T. (2008). Concerning the Origin of Malignant Tumours. Translated and annotated by Henry Harris. *Journal of Cell Science*, **121**(Supplement 1), 1–84. Publisher: The Company of Biologists Ltd Section: Article.

Bradley, J. (2020). Dimensionality and Structure in Cancer Genomics: A Statistical Learning Perspective. *Artificial Intelligence in Oncology Drug Discovery and Development*.

Bradley, J. R. and Cannings, T. I. (2021a). Data-driven design of targeted gene panels for estimating immunotherapy biomarkers. *arXiv:2102.04296 [q-bio, stat]*. arXiv: 2102.04296.

Bradley, J. R. and Cannings, T. I. (2021b). ICBioMark: Data-Driven Design of Targeted Gene Panels for Estimating immunotherapy Biomarkers.

Bradley, J. R. and Cannings, T. I. (2022). Data-driven design of targeted gene panels for estimating immunotherapy biomarkers. *Communications Biology*, **5**(1), 1–12.

Bradley, J. R. *et al.* (2023). Hierarchical Bayesian modeling identifies key considerations in the development of quantitative loop-mediated isothermal amplification assays.

Breiman, L. (2001). Random Forests. *Machine Learning*, **45**(1), 5–32.

Brenner, S. *et al.* (1961). An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature*, **190**, 576–581.

Brenner, S. *et al.* (1967). UGA: A Third Nonsense Triplet in the Genetic Code. *Nature*, **213**(5075), 449–450.

Buchbinder, E. I. and Desai, A. (2016). CTLA-4 and PD-1 Pathways: Similarities, Differences, and Implications of Their Inhibition. *American Journal of Clinical Oncology*, **39**(1), 98–106.

Budczies, J. *et al.* (2019). Optimizing panel-based tumor mutational burden (TMB) measurement. *Annals of Oncology: Official Journal of the European Society for Medical Oncology*, **30**(9), 1496–1506.

Bycroft, C. *et al.* (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, **562**(7726), 203–209.

Bühlmann, P. *et al.* (2014). High-Dimensional Statistics with a View Toward Applications in Biology. *Annual Review of Statistics and Its Application*, **1**(1), 255–278. _eprint: https://doi.org/10.1146/annurev-statistics-022513-115545.

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, **80**, 1–28.

Büttner, R. *et al.* (2019). Implementing TMB measurement in clinical practice: considerations on assay requirements. *ESMO open*, **4**(1), e000442.

Campbell, J. D. *et al.* (2016). Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nature Genetics*, **48**(6), 607–616.

Cancer Genome Atlas Network (2015). Genomic Classification of Cutaneous Melanoma. *Cell*, **161**(7), 1681–1696.

Cannings, T. I. and Samworth, R. J. (2017). Random-projection ensemble classification. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **79**(4), 959–1035.

Cao, D. *et al.* (2019). High tumor mutation burden predicts better efficacy of immunotherapy: a pooled analysis of 103078 cancer patients. *Oncoimmunology*, **8**(9), e1629258.

Cao, Y. *et al.* (2017). Development of a real-time fluorescence loop-mediated isothermal amplification assay for rapid and quantitative detection of Ustilago maydis. *Scientific Reports*, **7**(1), 13394.

Carvalho, J. *et al.* (2021). Faster monitoring of the invasive alien species (IAS) Dreissena polymorpha in river basins through isothermal amplification. *Scientific Reports*, **11**(1), 10175.

Centeno-Cuadros, A. *et al.* (2018). Validation of loop-mediated isothermal amplification for fast and portable sex determination across the phylogeny of birds. *Molecular Ecology Resources*, **18**(2), 251–263.

Chalmers, Z. R. *et al.* (2017). Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Medicine*, **9**(1), 34.

Chan, T. A. *et al.* (2019). Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic. *Annals of Oncology: Official Journal of the European Society for Medical Oncology*, **30**(1), 44–56.

Chapfuwa, P. *et al.* (2018). Adversarial Time-to-Event Modeling. *Proceedings of machine learning research*, **80**, 735–744.

Chapfuwa, P. *et al.* (2021). Enabling counterfactual survival analysis with balanced representations. In *Proceedings of the Conference on Health, Inference, and Learning*, CHIL '21, pages 133–145, New York, NY, USA. Association for Computing Machinery.

Chen, L. *et al.* (2017). Inflammatory responses and inflammation-associated diseases in organs. *Oncotarget*, **9**(6), 7204–7218.

Chen, X. and Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics*, **99**(6), 323–329.

Cheng, D. T. *et al.* (2015). Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. *The Journal of molecular diagnostics: JMD*, **17**(3), 251–264.

Cobb, M. (2015). Who discovered messenger RNA? *Current Biology*, **25**(13), R526–R532.

Coffin, J. M. and Fan, H. (2016). The Discovery of Reverse Transcriptase. *Annual Review of Virology*, **3**(1), 29–51.

Coles, P. (2006). Bayesians versus Frequentists. In P. Coles, editor, *From Cosmos to Chaos: The Science of Unpredictability*, page 0. Oxford University Press.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, **20**(3), 273–297.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, **34**(2), 187–220.

Craig, P. *et al.* (2017). Natural Experiments: An Overview of Methods, Approaches, and Contributions to Public Health Intervention Research. *Annual Review of Public Health*, **38**(1), 39–56.

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

Crick, F. H. C. *et al.* (1961). General Nature of the Genetic Code for Proteins. *Nature*, **192**(4809), 1227–1232.

Cui, Y. *et al.* (2022). Estimating heterogeneous treatment effects with right-censored data via causal survival forests. arXiv:2001.09887 [cs, stat].

Curth, A. and Schaar, M. (2021). Nonparametric Estimation of Heterogeneous Treatment Effects: From Theory to Learning Algorithms. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*.

Cuturi, M. (2013). Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances. arXiv:1306.0895 [stat].

De Palma, G. *et al.* (2019). Random deep neural networks are biased towards simple functions. arXiv:1812.10156 [cond-mat, physics:math-ph, physics:quant-ph, stat].

Demirhan, H. and Ata Tutkun, N. (2015). A Bayesian approach to Cox-Gompertz model. *Hacettepe Journal of Mathematics and Statistics*, **45**, 1–1.

Doroshow, D. B. *et al.* (2021). PD-L1 as a biomarker of response to immune-checkpoint inhibitors. *Nature Reviews Clinical Oncology*, **18**(6), 345–362.

Dosovitskiy, A. *et al.* (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 [cs].

Dowden, H. and Munro, J. (2019). Trends in clinical success rates and therapeutic focus. *Nature Reviews Drug Discovery*, **18**(7), 495–496.

Döring, M. *et al.* (2019). Modeling the Amplification of Immunoglobulins through Machine Learning on Sequence-Specific Features. *Scientific Reports*, **9**(1), 10748.

Eiken (2019). PrimerExplorer V5 Manual, https://primerexplorer.jp/e/index.html.

Emery, A. E. (1989). Joseph Adams (1756-1818). *Journal of Medical Genetics*, **26**(2), 116–118.

Emilien, G. *et al.* (2000). Impact of genomics on drug discovery and clinical medicine. *QJM: An International Journal of Medicine*, **93**(7), 391–423.

Fancello, L. *et al.* (2019). Tumor mutational burden quantification from targeted gene panels: major advancements and challenges. *Journal for Immunotherapy of Cancer*, **7**(1), 183.

Fang, B. and Roth, J. A. (2003). Tumor-suppressing gene therapy. *Cancer Biology & Therapy*, **2**(4 Suppl 1), S115–121.

Fitzpatrick, F. *et al.* (2019). Sepsis and antimicrobial stewardship: two sides of the same coin. *BMJ Quality & Safety*, **28**(9), 758–761.

Fong, Y. W. *et al.* (2013). The intertwined roles of transcription and repair proteins. *Molecular Cell*, **52**(3), 291–302.

Foulkes, W. D. *et al.* (1997). The CDKN2A (p16) gene and human cancer. *Molecular Medicine (Cambridge, Mass.)*, **3**(1), 5–20.

Fournier, P.-E. *et al.* (2014). Clinical detection and characterization of bacterial pathogens in the genomics era. *Genome Medicine*, **6**(11), 114.

Frampton, G. M. *et al.* (2013). Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nature Biotechnology*, **31**(11), 1023–1031.

Franklin, R. E. and Gosling, R. G. (1953). Molecular Configuration in Sodium Thymonucleate. *Nature*, **171**(4356), 740–741.

Friedman, J. *et al.* (2021). glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models.

Friedman, J. H. *et al.* (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, **33**(1), 1–22.

Fu, S. *et al.* (2011). Applications of loop-mediated isothermal DNA amplification. *Applied Biochemistry and Biotechnology*, **163**(7), 845–850.

Fölling, A. (1934). Über Ausscheidung von Phenylbrenztraubensäure in den Harn als Stoffwechselanomalie in Verbindung mit Imbezillität. **227**(1-4), 169–181.

Gandara, D. R. *et al.* (2018). 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.

Gandelman, O. *et al.* (2011). Loop-mediated amplification accelerated by stem primers. *International Journal of Molecular Sciences*, **12**(12), 9108–9124.

Geiss, G. K. *et al.* (2008). Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nature Biotechnology*, **26**(3), 317–325.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, **85**(410), 398–409.

Gelman, A. and Yao, Y. (2021). Holes in Bayesian Statistics. *Journal of Physics G: Nuclear and Particle Physics*, **48**(1), 014002. arXiv:2002.06467 [math, stat].

Giannakis, M. *et al.* (2016). Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma. *Cell Reports*, **15**(4), 857–865.

Goldblum, M. *et al.* (2023). The No Free Lunch Theorem, Kolmogorov Complexity, and the Role of Inductive Biases in Machine Learning. arXiv:2304.05366 [cs, stat].

Goldfeder, R. L. *et al.* (2017). Human Genome Sequencing at the Population Scale: A Primer on High-Throughput dna Sequencing and Analysis. *American Journal of Epidemiology*, **186**(8), 1000–1009.

Golkaram, M. *et al.* (2020). The interplay between cancer type, panel size and tumor mutational burden threshold in patient selection for cancer immunotherapy. *PLOS Computational Biology*, **16**(11), e1008332.

Gompertz, B. (1825). XXIV. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. In a letter to Francis Baily, Esq. F. R. S. &c. *Philosophical Transactions of the Royal Society of London*, **115**, 513–583.

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

Gotuzzo, A. G. *et al.* (2019). Bayesian hierarchical model for comparison of different nonlinear function and genetic parameter estimates of meat quails. *Poultry Science*, **98**(4), 1601–1609.

Gretton, A. *et al.* (2008). Covariate Shift by Kernel Mean Matching. In J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, editors, *Dataset Shift in Machine Learning*, page 0. The MIT Press.

Guo, G. *et al.* (2011). Frequent mutations of genes encoding ubiquitin-mediated proteolysis pathway components in clear cell renal cell carcinoma. *Nature Genetics*, **44**(1), 17–19.

Guo, G. *et al.* (2013). Whole-genome and whole-exome sequencing of bladder cancer identifies frequent alterations in genes involved in sister chromatid cohesion and segregation. *Nature Genetics*, **45**(12), 1459–1463.

Guo, W. *et al.* (2020). An Exon Signature to Estimate the Tumor Mutational Burden of Right-sided Colon Cancer Patients. *Journal of Cancer*, **11**(4), 883–892.

Hanahan, D. and Weinberg, R. A. (2000). The Hallmarks of Cancer. *Cell*, **100**(1), 57–70.

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

Hastie, T. *et al.* (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY.

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, **57**(1), 97–109.

He, K. *et al.* (2016). Deep Residual Learning for Image Recognition. pages 770–778.

He, Y. D. *et al.* (2021). The Optimization and Biological Significance of a 29-Host-Immune-mRNA Panel for the Diagnosis of Acute Infections and Sepsis. *Journal of Personalized Medicine*, **11**(8), 735.

Hellmann, M. D. *et al.* (2018a). Genomic Features of Response to Combination Immunotherapy in Patients with Advanced Non-Small-Cell Lung Cancer. *Cancer Cell*, **33**(5), 843–852.e4.

Hellmann, M. D. *et al.* (2018b). Nivolumab plus Ipilimumab in Lung Cancer with a High Tumor Mutational Burden. *The New England Journal of Medicine*, **378**(22), 2093–2104.

Heydt, C. *et al.* (2018). Evaluation of the TruSight Tumor 170 (TST170) assay and its value in clinical research. *Annals of Oncology*, **29**, vi7–vi8.

Hill, J. L. (2011). Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, **20**(1), 217–240.

Hirayama, H. *et al.* (2013). Embryo Sexing and Sex Chromosomal Chimerism Analysis by Loop-Mediated Isothermal Amplification in Cattle and Water Buffaloes. *The Journal of reproduction and development*, **59**, 321–6.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, **81**(396), 945–960.

Horiuchi, S. *et al.* (2020). A novel loop-mediated isothermal amplification method for efficient and robust detection of EGFR mutations. *International Journal of Oncology*, **56**(3), 743–749.

Horvitz, D. G. and Thompson, D. J. (1952). A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association*, **47**(260), 663–685.

Howard, D. H. *et al.* (2015). Pricing in the Market for Anticancer Drugs. *Journal of Economic Perspectives*, **29**(1), 139–162.

Huang, G. *et al.* (2018a). Densely Connected Convolutional Networks. arXiv:1608.06993 [cs].

Huang, S. *et al.* (2018b). Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer Genomics & Proteomics*, **15**(1), 41–51.

Huang, X. *et al.* (2022). Developing RT-LAMP assays for rapid diagnosis of SARS-CoV-2 in saliva. *EBioMedicine*, **75**.

Ingram, V. M. (1956). A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin. *Nature*, **178**(4537), 792–794.

Ishida, Y. *et al.* (1992). Induced expression of PD-1, a novel member of the immunoglobulin gene superfamily, upon programmed cell death. *The EMBO journal*, **11**(11), 3887–3895.

Jain, M. *et al.* (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, **17**(1), 239.

Jancík, S. *et al.* (2010). Clinical relevance of KRAS in human cancers. *Journal of Biomedicine & Biotechnology*, **2010**, 150960.

Kalofonou, M. *et al.* (2020). A novel hotspot specific isothermal amplification method for detection of the common PIK3CA p.H1047R breast cancer mutation. *Scientific Reports*, **10**(1), 4553.

Kan, Z. *et al.* (2018). Multi-omics profiling of younger Asian breast cancers reveals distinctive molecular signatures. *Nature Communications*, **9**(1), 1725.

Kanavos, P. (2006). The rising burden of cancer in the developing world.

Katzman, J. L. *et al.* (2018). DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, **18**(1), 24.

Kelly, E. *et al.* (2022). Systematic review of host genomic biomarkers of invasive bacterial disease: Distinguishing bacterial from non-bacterial causes of acute febrile illness. *eBioMedicine*, **81**, 104110.

Kennedy, E. H. (2020). Optimal doubly robust estimation of heterogeneous causal effects. *arXiv: Statistics Theory*.

Kennedy, E. H. (2022). Towards optimal doubly robust estimation of heterogeneous causal effects. arXiv:2004.14497 [math, stat].

Kimura, Y. *et al.* (2011). Optimization of turn-back primers in isothermal amplification. *Nucleic Acids Research*, **39**(9), e59.

King, E. A. *et al.* (2019). Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS genetics*, **15**(12), e1008489.

Kingma, D. P. and Welling, M. (2013). Auto-Encoding Variational Bayes.

Kolmogoroff, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin, Heidelberg.

Krauthammer, M. *et al.* (2012). Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma. *Nature Genetics*, **44**(9), 1006–1014.

Krizhevsky, A. *et al.* (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Kumar, A. *et al.* (2016). Substantial interindividual and limited intraindividual genomic diversity among tumors from men with metastatic prostate cancer. *Nature Medicine*, **22**(4), 369–378.

Künzel, S. R. *et al.* (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, **116**(10), 4156–4165.

Landau, W. M. *et al.* (2023). targets: Dynamic Function-Oriented 'Make'-Like Declarative Pipelines.

Lander, E. S. *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921.

Leach, D. R. *et al.* (1996). Enhancement of antitumor immunity by CTLA-4 blockade. *Science (New York, N.Y.)*, **271**(5256), 1734–1736.

Lecun, Y. (1987). *PhD thesis: Modeles connexionnistes de l'apprentissage (connectionist learning models)*. Universite P. et M. Curie (Paris 6).

Ledford, H. *et al.* (2018). Cancer immunologists scoop medicine Nobel prize. *Nature*, **562**(7725), 20–21. Number: 7725 Publisher: Nature Publishing Group.

Lee, W. *et al.* (2018). Selective targeting of KRAS oncogenic alleles by CRISPR/-Cas9 inhibits proliferation of cancer cells. *Scientific Reports*, **8**(1), 11879.

Lejeune, J. *et al.* (1959). [Human chromosomes in tissue cultures]. *Comptes Rendus Hebdomadaires Des Seances De l'Academie Des Sciences*, **248**(4), 602–603.

Li, K. and Brownley, A. (2010). Primer Design for RT-PCR. In N. King, editor, *RT-PCR Protocols: Second Edition*, Methods in Molecular Biology, pages 271–299. Humana Press, Totowa, NJ.

Li, S. *et al.* (2016). Loop-mediated isothermal amplification (LAMP): real-time methods for the detection of the survivin gene in cancer cells. *Analytical Methods*, **8**(33), 6277–6283.

Litchfield, K. *et al.* (2021). Meta-analysis of tumor- and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition. *Cell*, **184**(3), 596–614.e14.

Liu, V. X. *et al.* (2017). The Timing of Early Antibiotics and Hospital Mortality in Sepsis. *American Journal of Respiratory and Critical Care Medicine*, **196**(7), 856–863.

Lyu, G.-Y. *et al.* (2018). Mutation load estimation model as a predictor of the response to cancer immunotherapy. *NPJ genomic medicine*, **3**, 12.

Maddox, B. (2003). The double helix and the 'wronged heroine'. *Nature*, **421**(6921), 407–408.

Makova, K. D. and Hardison, R. C. (2015). The effects of chromatin organization on variation in mutation rates in the genome. *Nature Reviews. Genetics*, **16**(4), 213–223.

Mallona, I. *et al.* (2011). pcrEfficiency: a Web tool for PCR amplification efficiency prediction. *BMC Bioinformatics*, **12**(1), 404.

Mann, T. *et al.* (2009). A thermodynamic approach to PCR primer design. *Nucleic Acids Research*, **37**(13), e95.

Markham, N. R. and Zuker, M. (2008). UNAFold: software for nucleic acid folding and hybridization. *Methods in Molecular Biology (Clifton, N.J.)*, **453**, 3–31.

Mathers, C. and Loncar, D. (2006). Projections of global mortality and burden of disease from 2002 to 2030.

Matz, M. V. *et al.* (2013). No Control Genes Required: Bayesian Analysis of qRT-PCR Data. *PLOS ONE*, **8**(8), e71448.

Mekata, T. *et al.* (2009). Real-time quantitative loop-mediated isothermal amplification as a simple method for detecting white spot syndrome virus. *Letters in Applied Microbiology*, **48**(1), 25–32.

Michoel, T. (2016). Natural coordinate descent algorithm for L1-penalised regression in generalised linear models. *Computational Statistics & Data Analysis*, **97**, 60–70.

Mikolov, T. *et al.* (2013). Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs].

Mohanty, D. (2014). A century after discovery of sickle cell disease: keeping hope alive! *The Indian Journal of Medical Research*, **139**(6), 793–795.

Morris, L. G. T. and Chan, T. A. (2015). Therapeutic Targeting of Tumor Suppressor Genes. *Cancer*, **121**(9), 1357–1368.

Motone, K. and Nivala, J. (2023). Not if but when nanopore protein sequencing meets single-cell proteomics. *Nature Methods*, **20**(3), 336–338.

Mullis, K. *et al.* (1986). Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor Symposia on Quantitative Biology*, **51 Pt 1**, 263–273.

Nagamine, K. *et al.* (2002). Accelerated reaction by loop-mediated isothermal amplification using loop primers. *Molecular and Cellular Probes*, **16**(3), 223–229.

Narzisi, G. and Schatz, M. C. (2015). The Challenge of Small-Scale Repeats for Indel Discovery. *Frontiers in Bioengineering and Biotechnology*, **3**.

Neal, R. M. (2011). *MCMC using Hamiltonian dynamics*. arXiv:1206.1901 [physics, stat].

Nelson, M. R. *et al.* (2015). The support of human genetic evidence for approved drug indications. *Nature Genetics*, **47**(8), 856–860.

Nguyen, H. Q. *et al.* (2020). Quantification of colorimetric isothermal amplification on the smartphone and its open-source app for point-of-care pathogen detection. *Scientific Reports*, **10**(1), 15123.

Nie, X. and Wager, S. (2017). Learning Objectives for Treatment Effect Estimation.

Nie, X. and Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, **108**(2), 299–319.

Nielsen, H. B. *et al.* (2003). Design of oligonucleotides for microarrays and perspectives for design of multi-transcriptome arrays. *Nucleic Acids Research*, **31**(13), 3491–3496.

Nirenberg, M. and Leder, P. (1964). RNA codewords and protein synthesis: the effect of trinucleotides upon binding of sRNA to ribosomes. *Science (New York, N.Y.)*, **145**(3639), 1399–1407.

Notomi, T. *et al.* (2000). Loop-mediated isothermal amplification of DNA. *Nucleic Acids Research*, **28**(12), e63–e63.

Nowicki, T. S. *et al.* (2018). Mechanisms of Resistance to PD-1 and PD-L1 blockade. *Cancer journal (Sudbury, Mass.)*, **24**(1), 47–53.

Olivier, M. *et al.* (2010). TP53 Mutations in Human Cancers: Origins, Consequences, and Clinical Use. *Cold Spring Harbor Perspectives in Biology*, **2**(1).

Omidiran, D. and Wainwright, M. J. (2010). High-dimensional Variable Selection with Sparse Random Projections: Measurement Sparsity and Statistical Efficiency. *Journal of Machine Learning Research*, **11**(82), 2361–2386.

Panno, S. *et al.* (2020). Loop Mediated Isothermal Amplification: Principles and Applications in Plant Virology. *Plants*, **9**(4), 461.

Pardoll, D. M. (2012). The blockade of immune checkpoints in cancer immunotherapy. *Nature reviews. Cancer*, **12**(4), 252–264.

Paszke, A. *et al.* (2019). PyTorch: an imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, number 721, pages 8026–8037. Curran Associates Inc., Red Hook, NY, USA.

Pearl, J. (1995). Causal Diagrams for Empirical Research. *Biometrika*, **82**(4), 669–688.

Pearl, J. (2012). The do-calculus revisited. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI'12, pages 3–11, Arlington, Virginia, USA. AUAI Press.

Peyret, N. (2000). Prediction of nucleic acid hybridization : Parameters and Algorithms. *Wayne State University Dissertations*.

Prokop, J. W. *et al.* (2018). Genome sequencing in the clinic: the past, present, and future of genomic medicine. *Physiological Genomics*, **50**(8), 563–579.

Putzu, C. *et al.* (2023). Duration of Immunotherapy in Non-Small Cell Lung Cancer Survivors: A Lifelong Commitment? *Cancers*, **15**(3), 689.

Qiu, S. *et al.* (2023). Small molecule metabolites: discovery of biomarkers and therapeutic targets. *Signal Transduction and Targeted Therapy*, **8**(1), 1–37.

Raja, R. *et al.* (2017). Integrating Genomics into Drug Discovery and Development: Challenges and Aspirations. *Pharmaceutical Medicine*, **31**(4), 217–233.

Ram-Mohan, N. *et al.* (2022). Using a 29-mRNA Host Response Classifier To Detect Bacterial Coinfections and Predict Outcomes in COVID-19 Patients Presenting to the Emergency Department. *Microbiology Spectrum*, **10**(6), e02305–22.

Ramalingam, S. S. *et al.* (2018). 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.

Reeve, H. W. J. *et al.* (2021). Adaptive transfer learning. *The Annals of Statistics*, **49**(6), 3618–3649.

Reiersöl, O. (1945). Confluence analysis by means of instrumental sets of variables. In *Issue 4 of Arkiv for matematik, astronomi och fysik*.

Remmel, M. C. *et al.* (2022). Diagnostic Host Gene Expression Analysis by Quantitative Reverse Transcription Loop-Mediated Isothermal Amplification to Discriminate between Bacterial and Viral Infections. *Clinical Chemistry*, page hvab275.

Rizvi, N. A. *et al.* (2015). Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. *Science (New York, N.Y.)*, **348**(6230), 124–128.

Robert, C. (2020). A decade of immune-checkpoint inhibitors in cancer therapy. *Nature Communications*, **11**(1), 3801.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**(1), 41–55.

Roth, J. *et al.* (2023). What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature. arXiv:2201.01194 [econ, stat].

Roth, V. and Fischer, B. (2008). 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. Association for Computing Machinery.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**(5), 688–701.

Rubin, D. B. (2005). Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, **100**(469), 322–331.

Safarika, A. *et al.* (2021). A 29-mRNA host response test from blood accurately distinguishes bacterial and viral infections among emergency department patients. *Intensive Care Medicine Experimental*, **9**(1), 31.

Saiki, R. K. *et al.* (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science (New York, N.Y.)*, **230**(4732), 1350–1354.

Sanger, F. *et al.* (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**(12), 5463.

Sasaki, T. and Kondo, O. (2016). An informative prior probability distribution of the gompertz parameters for bayesian approaches in paleodemography. *American Journal of Physical Anthropology*, **159**(3), 523–533.

Sboner, A. *et al.* (2011). The real cost of sequencing: higher than you think! *Genome Biology*, **12**(8), 125.

Seshagiri, S. *et al.* (2012). Recurrent R-spondin fusions in colon cancer. *Nature*, **488**(7413), 660–664.

Shalit, U. *et al.* (2017). Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 3076–3085, Sydney, NSW, Australia. JMLR.org.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, **27**(3), 379–423.

Smith, J. J. *et al.* (2003). Biomarkers in Imaging: Realizing Radiology's Future. *Radiology*, **227**(3), 633–638.

Spiess, A.-N. *et al.* (2008). Highly accurate sigmoidal fitting of real-time PCR data by introducing a parameter for asymmetry. *BMC Bioinformatics*, **9**(1), 221.

Sriperumbudur, B. K. *et al.* (2012). On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, **6**(none), 1550–1599.

Stone, R. (1993). The Assumptions on Which Causal Inferences Rest. *Journal of the Royal Statistical Society. Series B (Methodological)*, **55**(2), 455–466.

Subramanian, S. and Gomez, R. D. (2014). An Empirical Approach for Quantifying Loop-Mediated Isothermal Amplification (LAMP) Using Escherichia coli as a Model System. *PLOS ONE*, **9**(6), e100596.

Szustakowski, J. D. *et al.* (2021). Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. *Nature Genetics*, **53**(7), 942–948.

Tamura, K. (2016). The Genetic Code: Francis Crick's Legacy and Beyond. *Life*, **6**(3), 36.

Tefferi, A. (2007). JAK2 Mutations in Polycythemia Vera — Molecular Mechanisms and Clinical Applications. *New England Journal of Medicine*, **356**(5), 444–445.

Telenti, A. *et al.* (2016). Deep sequencing of 10,000 human genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **113**(42), 11901–11906.

Temin, H. M. and Mizutani, S. (1970). RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature*, **226**(5252), 1211–1213.

Thiessen, L. D. *et al.* (2018). Development of a quantitative loop-mediated isothermal amplification assay for the field detection of Erysiphe necator. *PeerJ*, **6**, e4639.

Thistlethwaite, D. L. and Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, **51**, 309–317.

Tibshirani, J. *et al.* (2023). grf: Generalized Random Forests.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(1), 267–288.

Tjørve, K. M. C. and Tjørve, E. (2017). The use of Gompertz models in growth analyses, and new Gompertz-model approach: An addition to the Unified-Richards family. *PLOS ONE*, **12**(6), e0178691.

Tran, K. A. *et al.* (2021). Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Medicine*, **13**(1), 152.

Turajlic, S. *et al.* (2017). Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. *The Lancet. Oncology*, **18**(8), 1009–1021.

Vaghi, C. *et al.* (2020). Population modeling of tumor growth curves and the reduced Gompertz model improve prediction of the age of experimental tumors. *PLOS Computational Biology*, **16**(2), e1007178.

Valle-Pérez, G. *et al.* (2019). Deep learning generalizes because the parameter-function map is biased towards simple functions. arXiv:1805.08522 [cs, stat].

Vallverdú, J. (2016). *Bayesians Versus Frequentists*. SpringerBriefs in Statistics. Springer, Berlin, Heidelberg.

Van Der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer Series in Statistics. Springer, New York, NY.

Vanderboom, P. M. *et al.* (2021). Proteomic Signature of Host Response to SARS-CoV-2 Infection in the Nasopharynx. *Molecular & Cellular Proteomics : MCP*, **20**, 100134.

Vaswani, A. *et al.* (2017). Attention Is All You Need. arXiv:1706.03762 [cs].

Vegetabile, B. G. (2021). On the Distinction Between "Conditional Average Treatment Effects" (CATE) and "Individual Treatment Effects" (ITE) Under Ignorability Assumptions. arXiv:2108.04939 [cs, stat].

Vehtari, A. *et al.* (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, **27**(5), 1413–1432.

Vehtari, A. *et al.* (2020). loo: Efficient Leave-One-Out Cross-Validation and WAIC for Bayesian models.

Vehtari, A. *et al.* (2021). Pareto Smoothed Importance Sampling. *arXiv:1507.02646 [stat]*. arXiv: 1507.02646.

Villani, C. (2009). *Optimal Transport*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press. Google-Books-ID: 8C8nuQEACAAJ.

Walsh, M. F. *et al.* (2016). Genomic Biomarkers for Breast Cancer Risk. *Advances in experimental medicine and biology*, **882**, 1–32.

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

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

Weber, B. *et al.* (2014). EGFR mutation frequency and effectiveness of erlotinib: a prospective observational study in Danish patients with non-small cell lung cancer. *Lung Cancer (Amsterdam, Netherlands)*, **83**(2), 224–230.

Wei, L. J. (1992). The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. *Statistics in Medicine*, **11**(14-15), 1871–1879.

Weinstein, I. B. (2002). Cancer. Addiction to oncogenes–the Achilles heal of cancer. *Science (New York, N.Y.)*, **297**(5578), 63–64.

Weinstein, I. B. and Case, K. (2008). The History of Cancer Research: Introducing an AACR Centennial Series. *Cancer Research*, **68**(17), 6861–6862.

Weinstein, J. N. *et al.* (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, **45**(10), 1113–1120.

Wetterstrand, K. (2022). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).

Whiteside, T. (2008). The tumor microenvironment and its role in promoting tumor growth. *Oncogene*, **27**(45), 5904–5912.

Wild, C. *et al.* (2020). *World Cancer Report: Cancer Research for Cancer Prevention*. IARC Publications.

Wilkins, M. H. F. *et al.* (1953). Molecular Structure of Nucleic Acids: Molecular Structure of Deoxypentose Nucleic Acids. *Nature*, **171**(4356), 738–740.

Wiper, M. P. *et al.* (2010). Bayesian hierarchical modelling of bacteria growth. Technical Report ws102109, Universidad Carlos III de Madrid. Departamento de Estadística.

Wittekind, C. *et al.* (2016). TNM Classification of Malignant Tumours, 8th Edition | Wiley.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, **5**(2), 241–259.

Wright, E. S. *et al.* (2014). Exploiting extension bias in polymerase chain reaction to improve primer specificity in ensembles of nearly identical DNA templates. *Environmental Microbiology*, **16**(5), 1354–1365.

Wu, H.-X. *et al.* (2019a). Designing gene panels for tumor mutational burden estimation: the need to shift from 'correlation' to 'accuracy'. *Journal for Immunotherapy of Cancer*, **7**(1), 206.

Wu, H.-X. *et al.* (2019b). Tumor mutational and indel burden: a systematic pan-cancer evaluation as prognostic biomarkers. *Annals of Translational Medicine*, **7**(22), 640.

Xia, X.-Q. *et al.* (2010). Evaluating oligonucleotide properties for DNA microarray probe design. *Nucleic Acids Research*, **38**(11), e121.

Yang, Y. and Zou, H. (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, **25**(6), 1129–1141.

Yang, Y. *et al.* (2020). gglasso: Group Lasso Penalized Learning Using a Unified BMD Algorithm.

Yao, L. *et al.* (2020). ecTMB: a robust method to estimate and classify tumor mutational burden. *Scientific Reports*, **10**(1), 4983.

Yao, Y. *et al.* (2018). Using Stacking to Average Bayesian Predictive Distributions (with Discussion). *Bayesian Analysis*, **13**(3), 917–1007.

Yates, A. D. *et al.* (2020). Ensembl 2020. *Nucleic Acids Research*, **48**(D1), D682–D688.

Zehir, A. *et al.* (2017). Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nature Medicine*, **23**(6), 703–713.

Zhang, X. *et al.* (2014). Brief review of monitoring methods for loop-mediated isothermal amplification (LAMP). *Biosensors and Bioelectronics*, **61**, 491–499.

Zhao, L. *et al.* (2019). Molecular subtyping of cancer: current status and moving toward clinical applications. *Briefings in Bioinformatics*, **20**(2), 572–584.

Zhu, J. *et al.* (2019). 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**, 673.

Appendix A

Computational workflow for LAMP assay analysis

In this appendix we discuss software and computational aspects of the analysis presented in Chapter 2. This is divided into two sections.

In Chapter 2 we discuss the generation of several primer/amplicon *features*, each a function of a given primer/amplicon pair's sequences. To ensure that this is reproducible and usable by other researchers, we produced an R package, `LAMPPrimerFeatures`, encapsulating the above functionality. In section A.1 we discuss the design and use of this package, working with a very small example dataset provided alongside the package.

Beyond from the core functionality of producing primer features, we wanted the code-level specifics of all analysis presented in Chapter 2 to be available, understandable, and reproducible. In Section A.2 we demonstrate how we made use of the `targets` framework in R to achieve these goals.

Regrettably, at the time of submission neither of these codebases have been made available publicly (while both are intended to be), due to the need for final approval from industrial collaborators. We have therefore decided to include this appendix, and interested readers should contact the author if they wish to receive special access to the two repositories associated with each section.

A.1 R package `LAMPPrimerFeatures`

A.1.1 Goals

We developed the R package `LAMPPrimerFeatures` (to be released publicly in the near future) with three goals. Firstly, we aimed to make consistent the process for mapping from amplicon and primer sequences to the features described in Table 2.2. This allows for simple and concise code to perform the subsequent modelling analysis, and easy integration of extra features in future analyses should they be required. Secondly, we were required to make our feature generation pipeline amenable to analysis of multiple data formats generated by different experimental setups. In particular, we need to be able to deal both with primer specification datasets in which FIP/BIPP primer sets are varied for each target,

and for which each of F3/B3, FIP/BIP, and LF/LB are varied once per target. Thirdly, we set out to automate several aspects of manual work common to primer analysis, including generation of complement/reverse complement sequences and identification of F1/F2 (B1/B2) subsequences of FIP (BIP) primers.

A.1.2 Implementation and dependencies

The typical analysis workflow for use of `LAMPPrimerFeatures` comprises three steps. Firstly, primer sequences are aligned with their respective amplicon sequences. This necessitates the identification of amplicon regions associated with F3, B3, LF, LB, F1, B1, F2, and B2 sequences, where F1/F2 and B1/B2 may have to be identified within given FIP/BIP primer sequences. Secondly, where necessary comparison sequences are generated, such as reverses, complements, and reverse complements for each relevant sequence. Finally, these sets of processed sequences are provided to functions producing each of the required features. We provide wrappers to each of these functions such that they can be applied at scale across a dataset.

The `LAMPPrimerFeatures` package has various dependencies, including several packages in the `tidyverse` ecosystem such as `dplyr` for data manipulation, `stringr` for string processing, and `purrr` for efficient application of functions to multiple inputs. For string alignment algorithms, `Biostrings` is used from the Bioconductor project. Finally, UNAFold (Markham and Zuker, 2008) is a command line tool for approximate free energy calculations. It is available free for researchers, and can be purchased for commercial use.

A.1.3 Example use case

The development version of `LAMPPrimerFeatures` can be installed (with correct permissions) from GitHub with:

```
# install.packages("devtools")
devtools::install_github("cobrbra/LAMPPrimerFeatures")
```

An example dataset is provided to demonstrate a typical workflow (output in Table A.1):

```
devtools::load_all()
library(tidyverse)
library(knitr)

kable(example_primer_data)
```

primer	target	primer_set_id	name	sequence	amplicon_raw
F3	TP53	1	TP53_F3_1	AGT	CCGACTCCC
B3	TP53	1	TP53_B3_1	AGTC	CCGACTCCC
FIP	KRAS	1	KRAS_FIP_1	AACC	AGCCTTGA

Table A.1: Dataframe example_primer_data.

Firstly, we generate alignments of primer sequences with their amplicon reference, and use these to identify subregions and create new F1/F2/B1/B2 primers (output in Table A.2):

```
alignments <- get_alignments(example_primer_data)
#> aligning sequences
#> finding start/end points
#> done!
subprimers <- generate_subprimers(alignments)

kable(subprimers)
```

primer	target	name	sequence	amplicon_raw
F3	TP53	TP53_F3_1	AGT	CCGACTCCC
B3	TP53	TP53_B3_1	AGTC	CCGACTCCC
FIP	KRAS	KRAS_FIP_1	AACC	AGCCTTGA
F1	KRAS	KRAS_FIP_1R1	TT	AGCCTTGA
F2	KRAS	KRAS_FIP_1R2	CC	AGCCTTGA

Table A.2: Example primers with relevant subsequences identified. Note the sequence AA in FIP has been complemented to TT in F1, as technically FIP is formed of F1c $\rightarrow$ F2.

We then generate amplicon and stem endpoints. These are only relevant for F1/F2/B1/B2, and are used downstream to ascertain stem length and loop length for an individual experimental setup. Output is shown in Table A.3.

```
amplicon_endpoints <- get_amplicon_endpoints(subprimers)
kable(amplicon_endpoints)
```

primer	target	...	sequence	amplicon_raw	stem_end	amplicon_end
F3	TP53	...	AGT	CCGACTCCC	NA	NA
B3	TP53	...	AGTC	CCGACTCCC	NA	NA
FIP	KRAS	...	AACC	AGCCTTGA	NA	NA
F1	KRAS	...	TT	AGCCTTGA	6	NA
F2	KRAS	...	CC	AGCCTTGA	NA	3

Table A.3: Example primers with amplicon and stem endpoints identified.

We may then write to file comparison sequences (complements, reverses, reverse complements) and use them to calculate free energies. We do this in a temporary directory, as we will not need these written files after free energy properties are calculated. For this step UNAFold must be installed.

```
# Run with UNAFold installed

# .old_wd <- setwd(tempdir())
# write_sequences(example_primer_data)
# free_energies <- amplicon_endpoints %>%
```

```
#    calculate_free_energies()
# setwd(.old_wd)
#
# kable(free_energies)
```

Finally we can add k-mer complexities. Here we analyse 2-mers as we only have very short example primer sequences (output in Table A.4).

```
complexities <- amplicon_endpoints %>%
calculate_complexities(len = 2, reference = "AGTAGTCAACCTTCCGAGAG")

kable(complexities)
```

primer	target	...	sequence	amplicon_raw	...	complexity_2
F3	TP53	...	AGT	CCGACTCCC	...	0.3750
B3	TP53	...	AGTC	CCGACTCCC	...	0.5000
FIP	KRAS	...	AACC	AGCCTTGA	...	0.9166
F1	KRAS	...	TT	AGCCTTGA	...	1.0000
F2	KRAS	...	CC	AGCCTTGA	...	0.7500

Table A.4: Example primers with complexities provided.

A.2 Implementation of LAMP analysis with `targets`

A.2.1 Goals

In enabling a reproducible, modular and extensible piece of computational analysis in Chapter 2, we aimed to produce a codebase such that readers would be able to re-run the entire analysis (from raw data) with little or no manual work. While simple execution and reproduction was key, we also wanted it to be easy to understand the tools applied at each processing step without having to run the entire analysis themselves. To do this, we used the R package `targets` (Landau *et al.*, 2023), a workflow management system with which users define a dependency graph tracking the relationships between quantities of interest in an analysis. This is useful for external readers/users to build an intuitive understand of the processed underlying the analysis, but also during development and computation – the `targets` framework is able to track which sub-analyses have and haven't been run, as well as the impact on downstream quantities of changing functions/input data earlier in the analysis, e.g. in preprocessing.

A.2.2 Specifying a dependency graph

At the core of the `targets` approach is the specification of a *dependency graph*. This directed acyclic graph relates quantities of interest throughout the analysis. For example, in the workflow accompanying our LAMP analysis we have targets

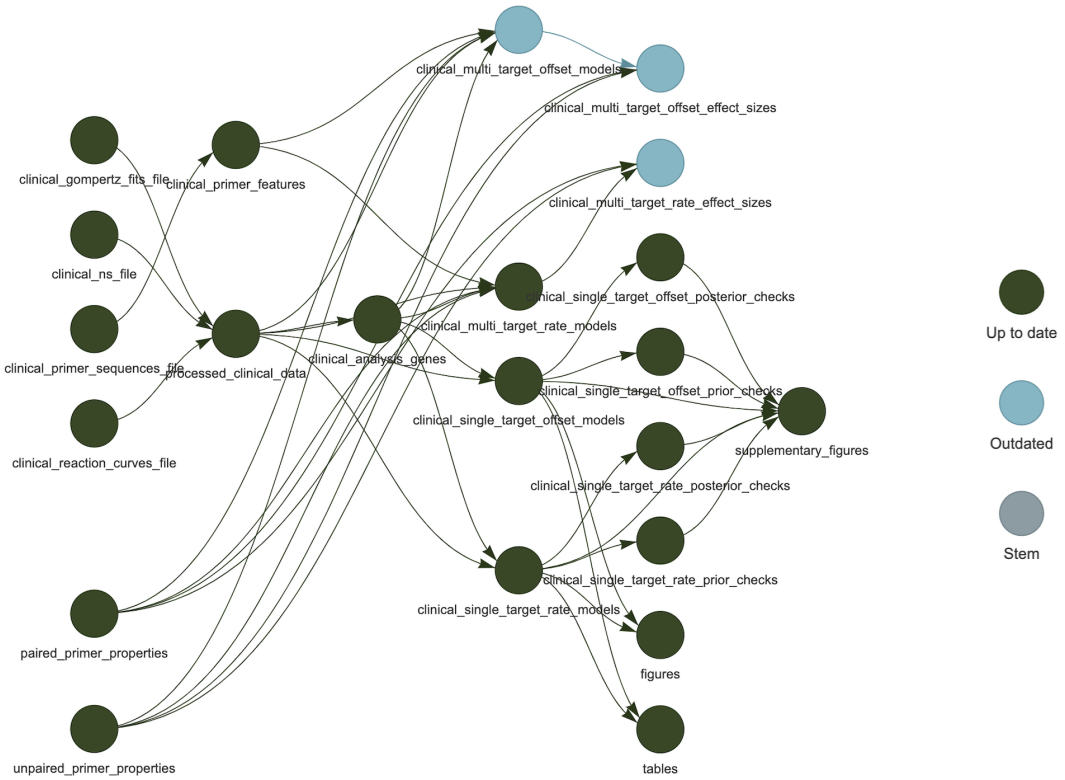

Figure A.1: An example (partially complete) targets dependency graph for the LAMP analysis project in Chapter 2.

for raw and processed clinical data, single-target models, multi-target models, and figures. A dependency graph is specified in a file "targets.R", in which the relationships between targets are defined in terms of functions within targets project's scope. See Figure A.1 for an example dependency graph from this project.

A.2.3 Running and tracking analyses

The dependency graph as described is useful for ensuring an analysis is modular and flexible (any function for moving from one target to another can be investigated, tweaked or replaced), but also entirely reproducible. At any point, all targets can be completely updated by running targets :: tar_make(). Furthermore, `targets` is a fantastic resource for tracking the current status of a project (particularly when some constituent processes take a long time). Note that in Figure **??**, some nodes are blue while others are grey. Blue nodes here indicate a portion of the analysis that has not yet been performed. We are able to automatically track whether any targets need re-running (or running for the first time) via a system of hashes of objects and relating functions. For example, upon altered a preprocessing function for clinical data, we arrive at the dependency graph shown in Figure A.2. The node corresponding to this altered preprocessed data and all its downstream children are now due for updating the next time targets :: tar_make() is run.

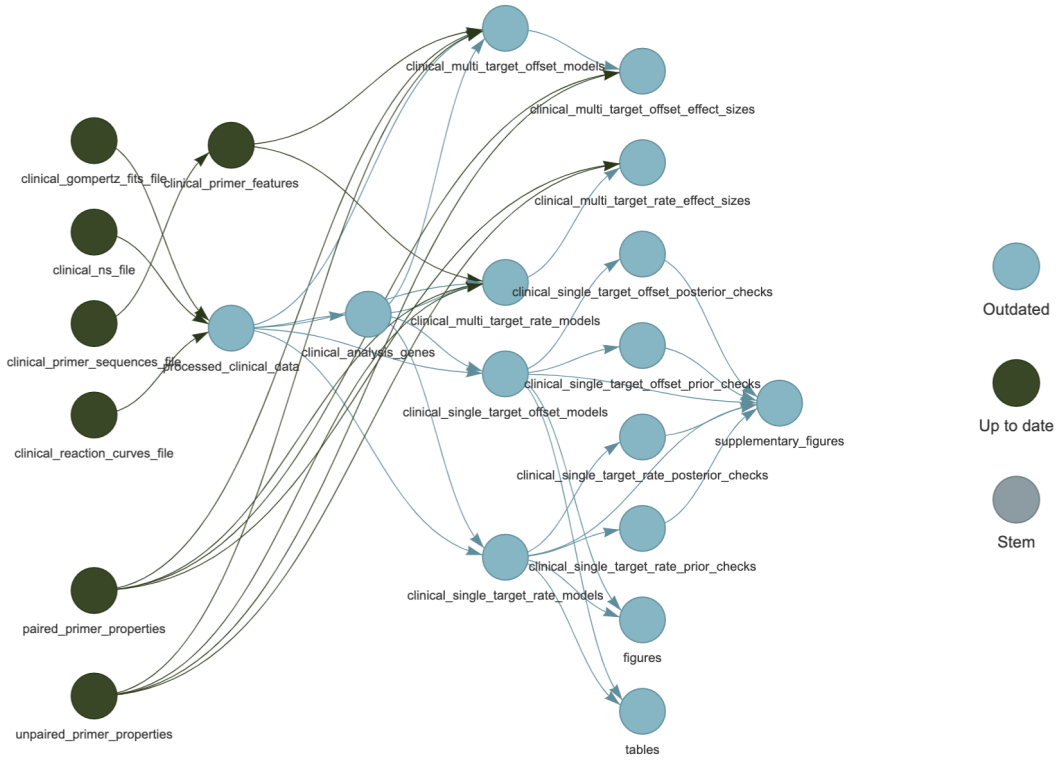

Figure A.2: An example (partially complete) targets dependency graph for the LAMP analysis project in Chapter 2, after changes have been made to a preprocessing step.

Appendix B

Extended model summaries for LAMP assay analysis

In this appendix we give more thorough model summaries than were provided in the text of Chapter 2.

B.1 Prior checks

We begin by displaying prior predictive check plots for single-target models specified in Section 2.2.3. Prior predictive checks are used to make sure that the distribution of outcomes implicitly postulated by the choices of priors suitably covers the true observed distribution. We show these for models of rate (Figure B.1) and of cycle offset (Figure B.2).

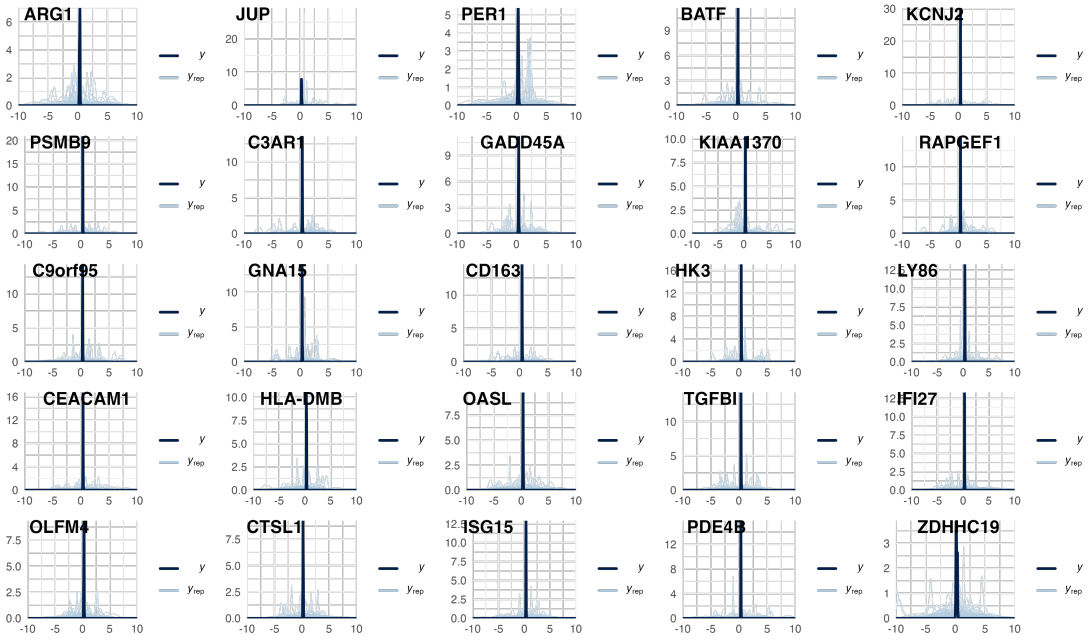

Figure B.1: Prior predictive plots for each of the 25 single-target models fitted of rate $\hat{B}_{ij}$, summarised in Table 2.4 in the main text.

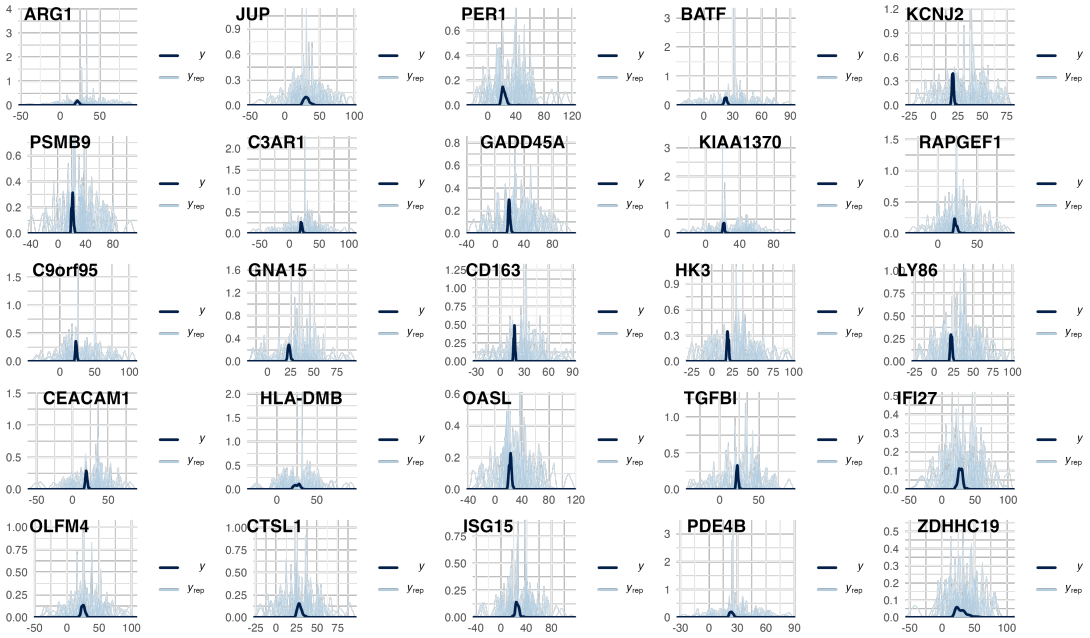

Figure B.2: Prior predictive plots for each of the 25 single-target models fitted of offset $\hat{C}_{ij}$, summarised in Table 2.3 in the main text.

B.2 Posterior Checks

We next give posterior predictive check plots for the same set of 25 models. These, rather than simply aiming to show a distribution encompassing the range of the observed distribution, aim to show a fitted distribution that closely matches it. Once again, we show posterior summaries for models of rate (Figure B.3) and of cycle offset (Figure B.4).

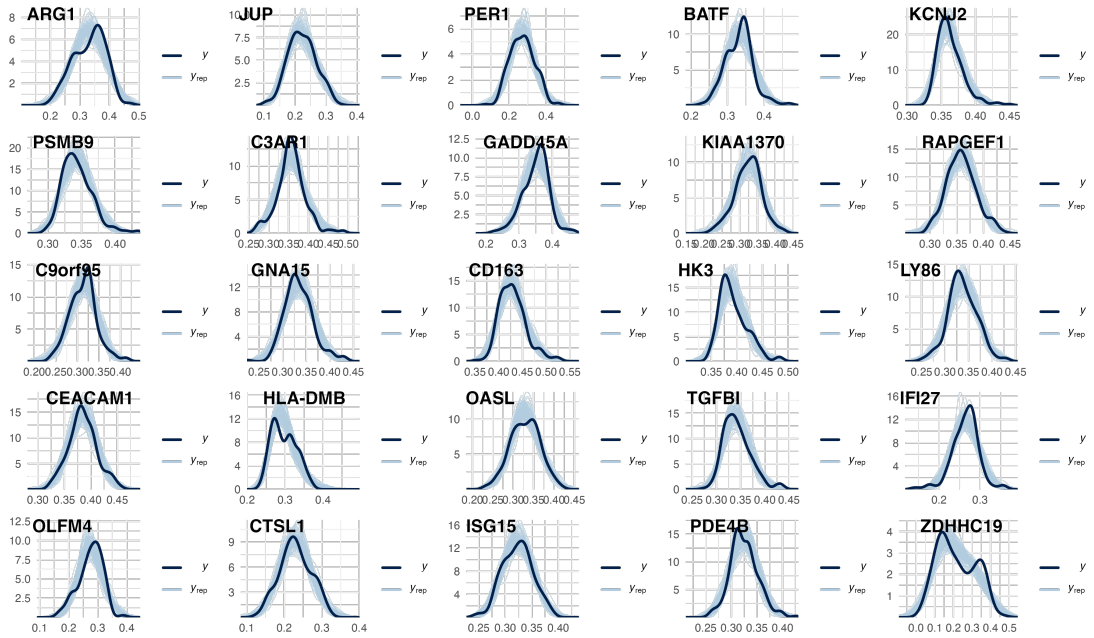

Figure B.3: Posterior predictive plots for each of the 25 single-target models fitted of offset $\hat{C}_{ij}$, summarised in Table 2.4 in the main text.

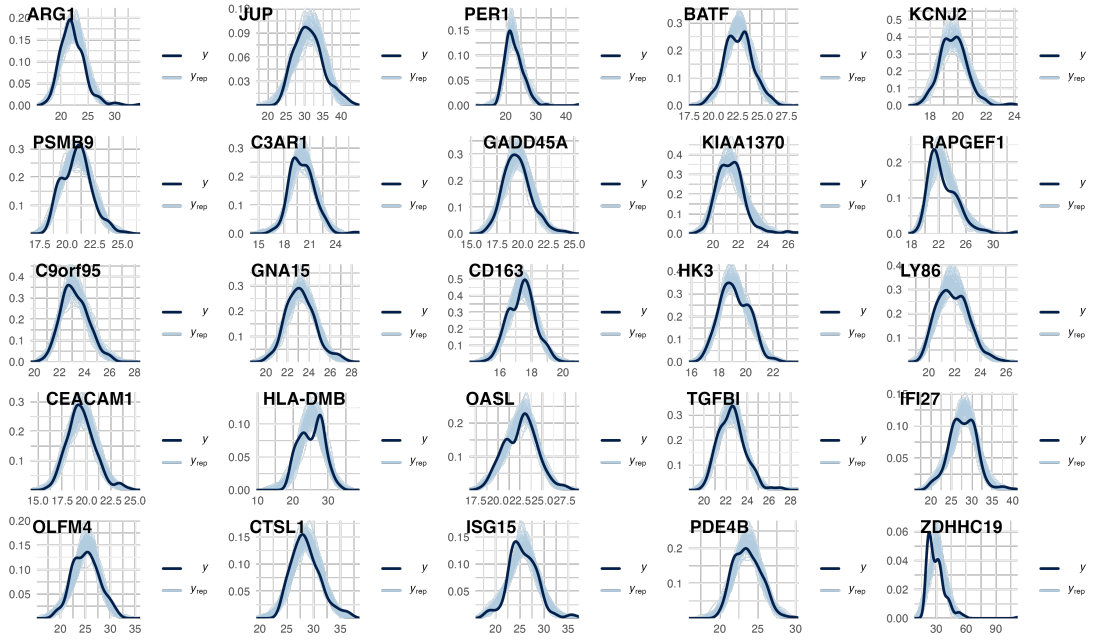

Figure B.4: Posterior predictive plots for each of the 25 single-target models fitted of offset $\hat{C}_{ij}$, summarised in Table 2.3 in the main text.

B.3 Model Convergence

Finally, we present HMC convergence summaries for each of the 25 single-target models of both rate and cycle offset.

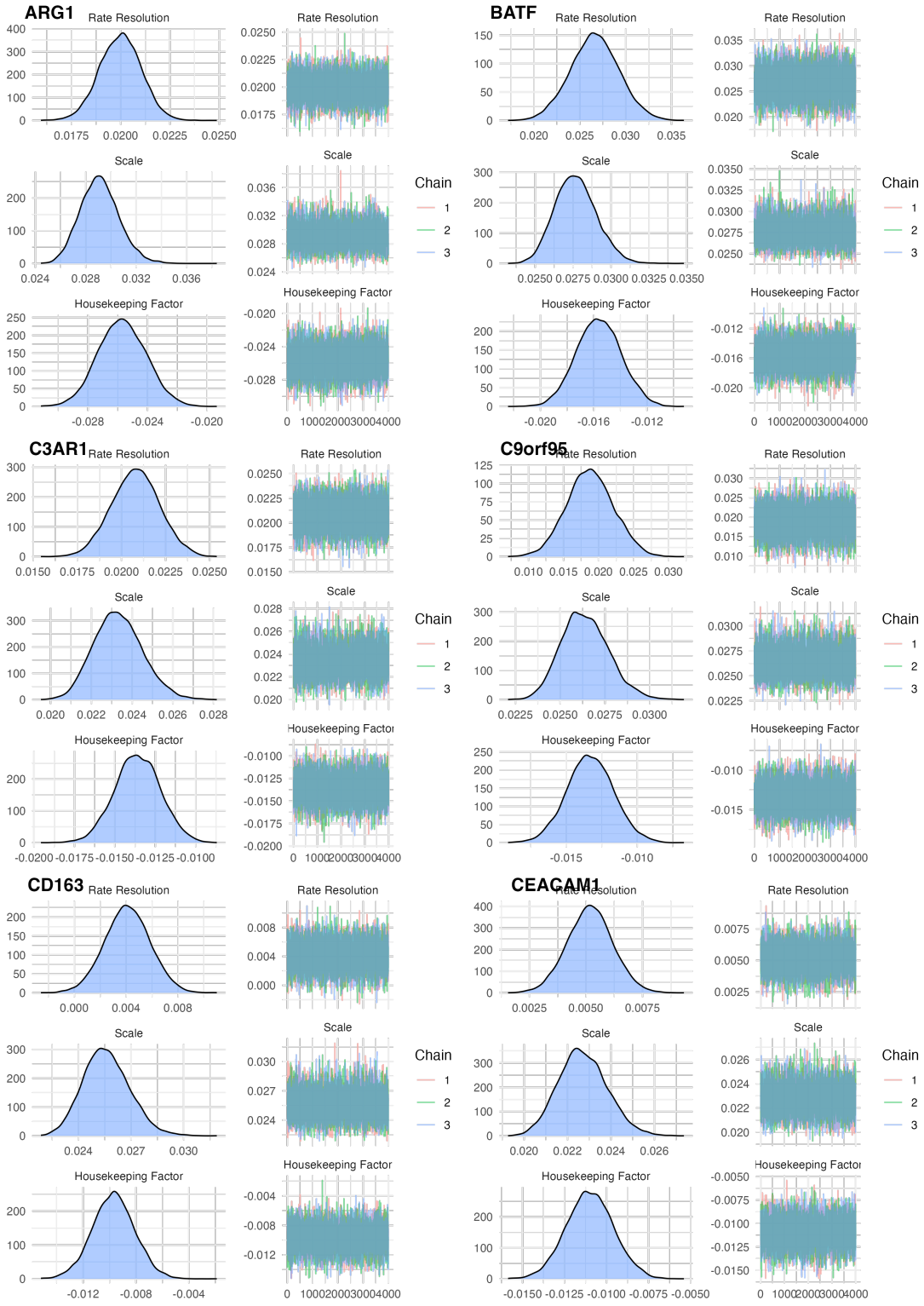

Figure B.5: Model convergence summaries for each of the 25 single-target rate models summarised in Table 2.4.

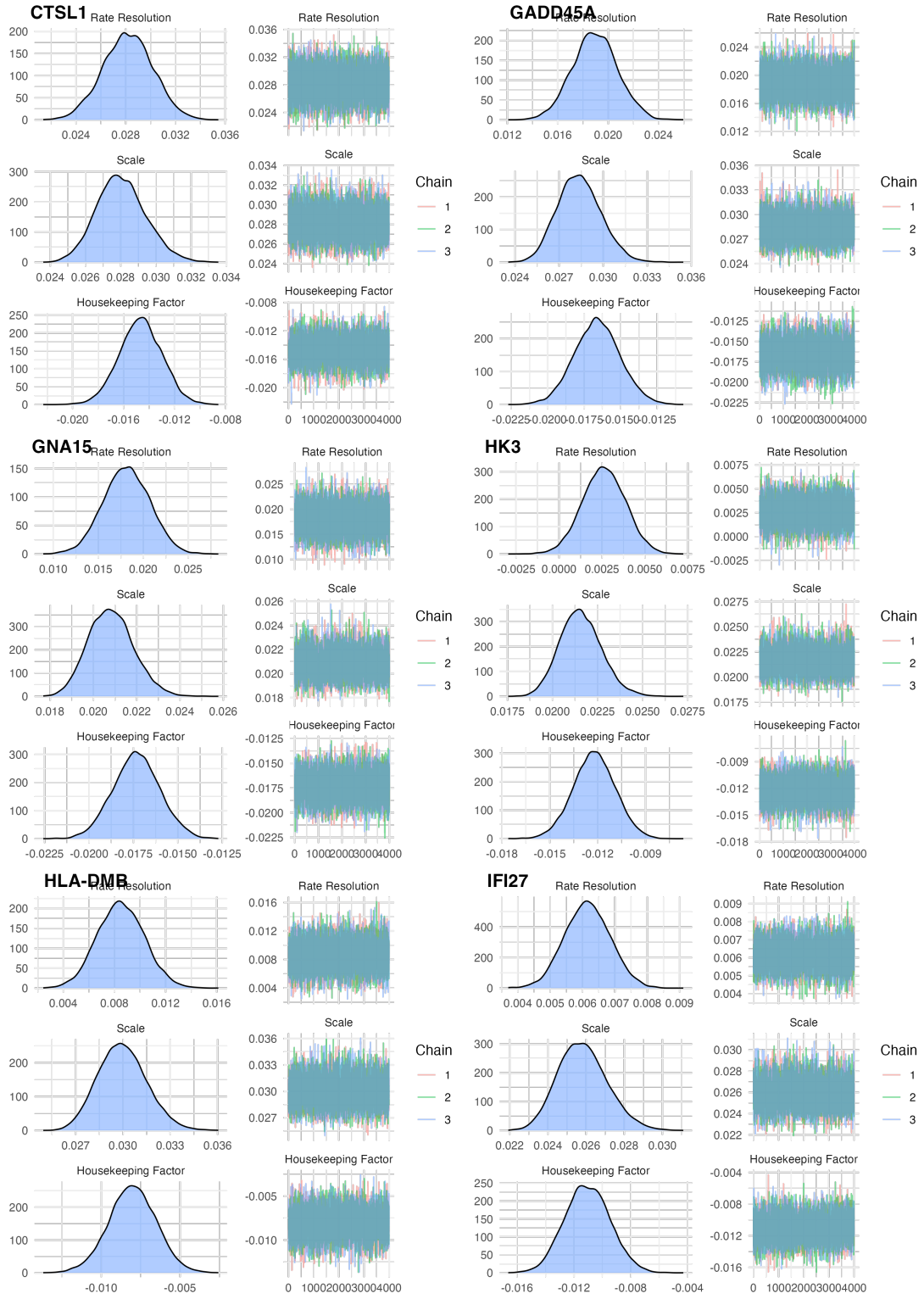

Figure B.6: Model convergence summaries for each of the 25 single-target rate models summarised in Table 2.4 (continued).

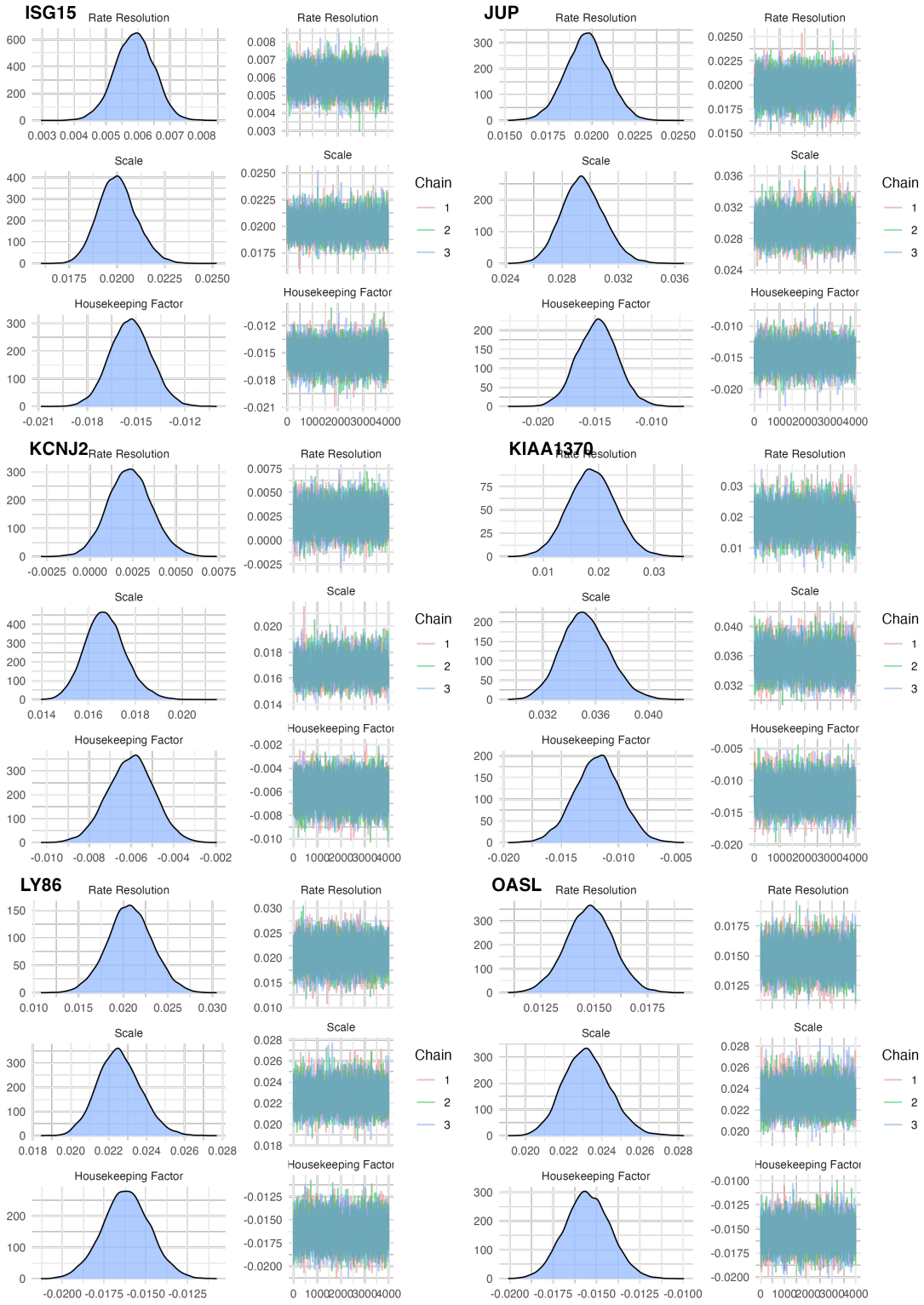

Figure B.7: Model convergence summaries for each of the 25 single-target rate models summarised in Table 2.4 (continued).

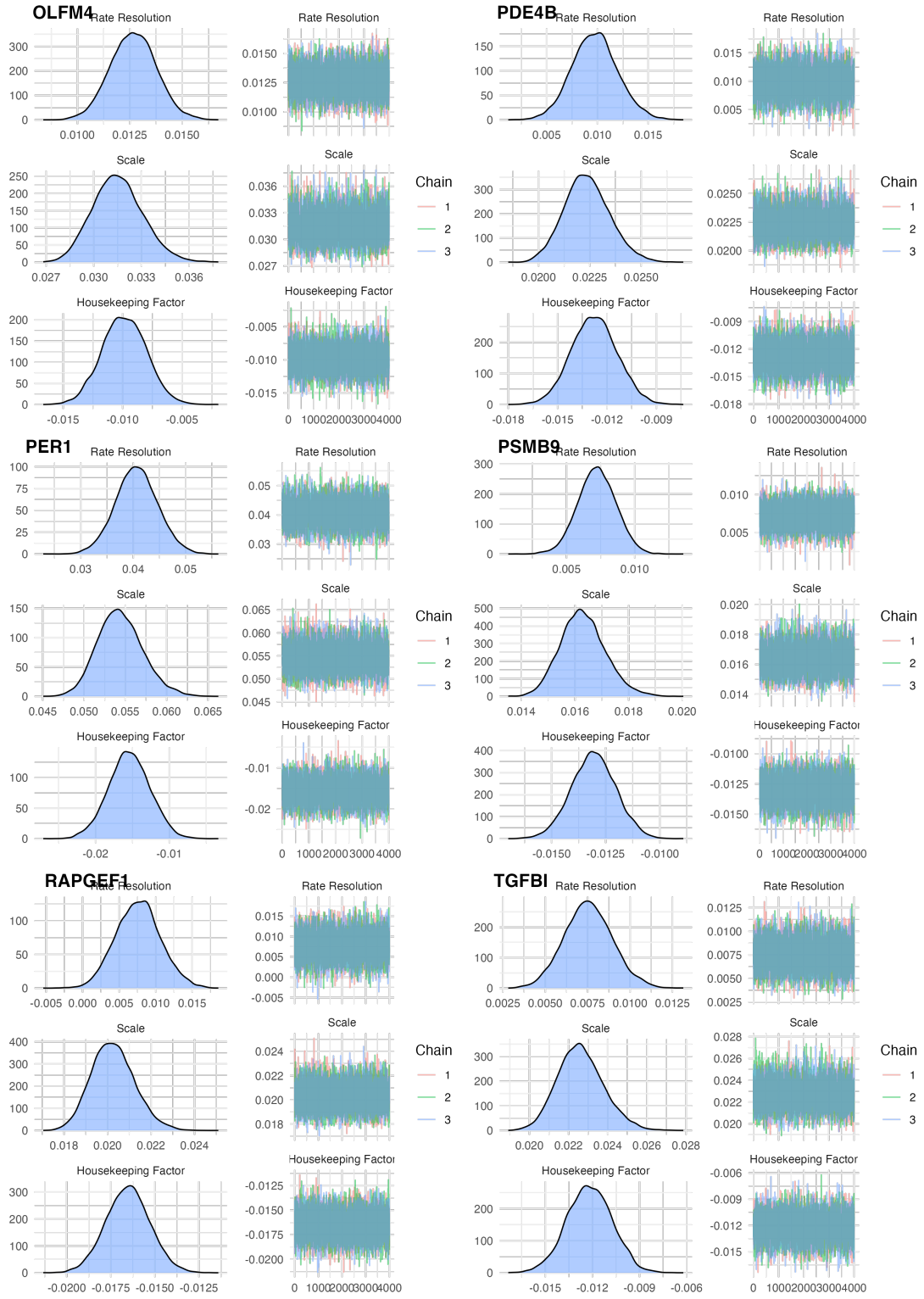

Figure B.8: Model convergence summaries for each of the 25 single-target rate models summarised in Table 2.4 (continued).

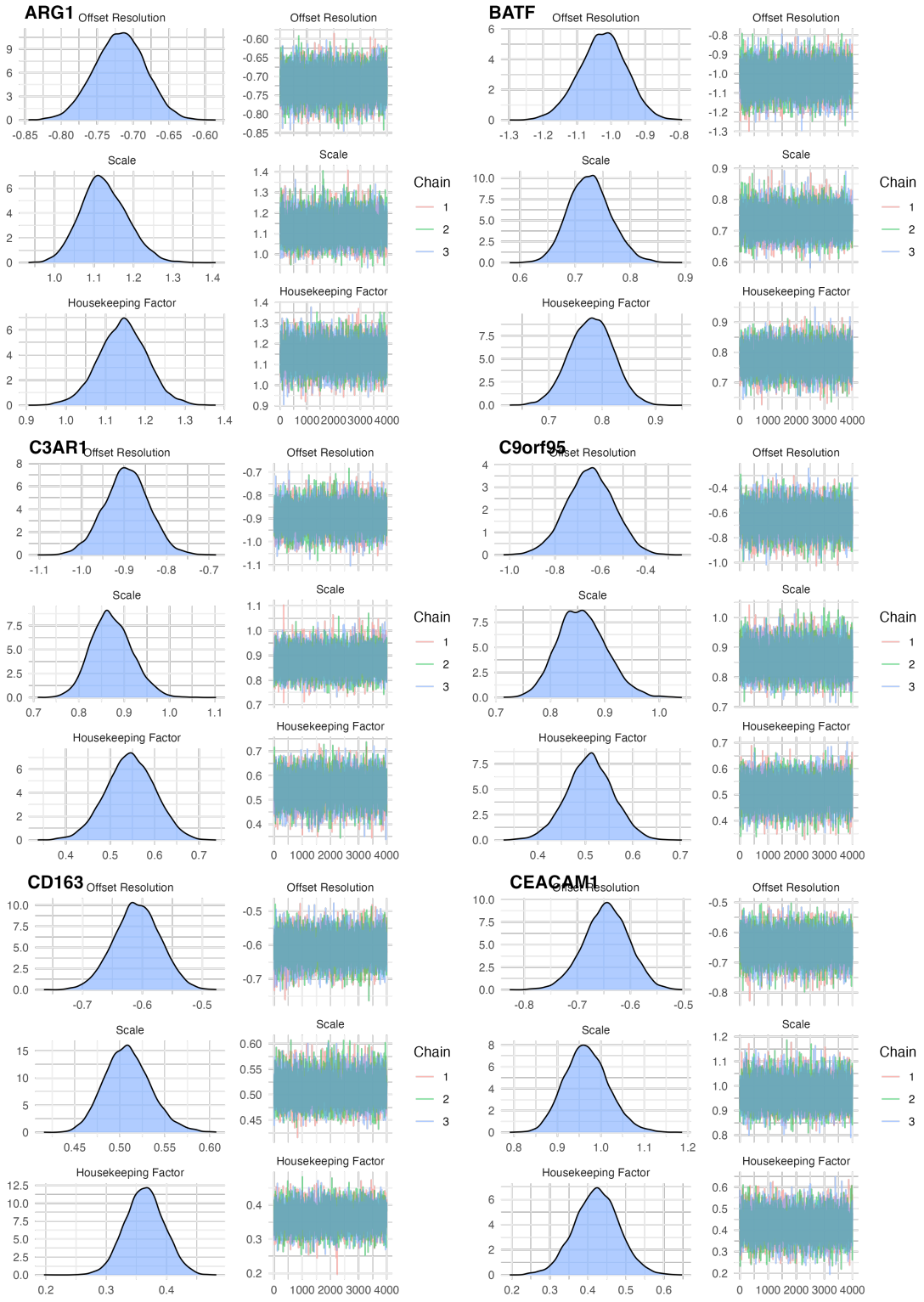

Figure B.9: Model convergence summaries for each of the 25 single-target offset models summarised in Table 2.3.

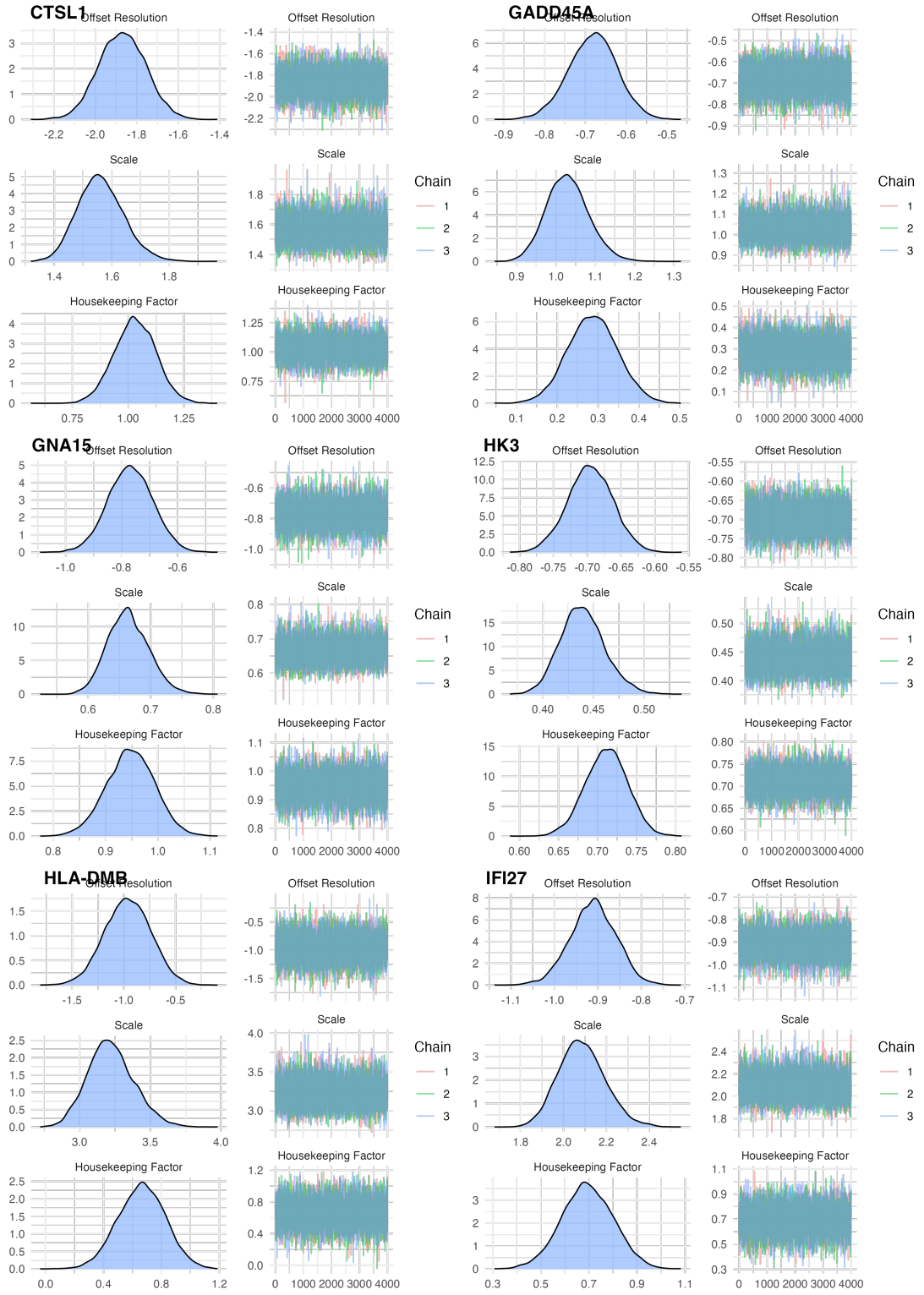

Figure B.10: Model convergence summaries for each of the 25 single-target offset models summarised in Table 2.3 (continued).

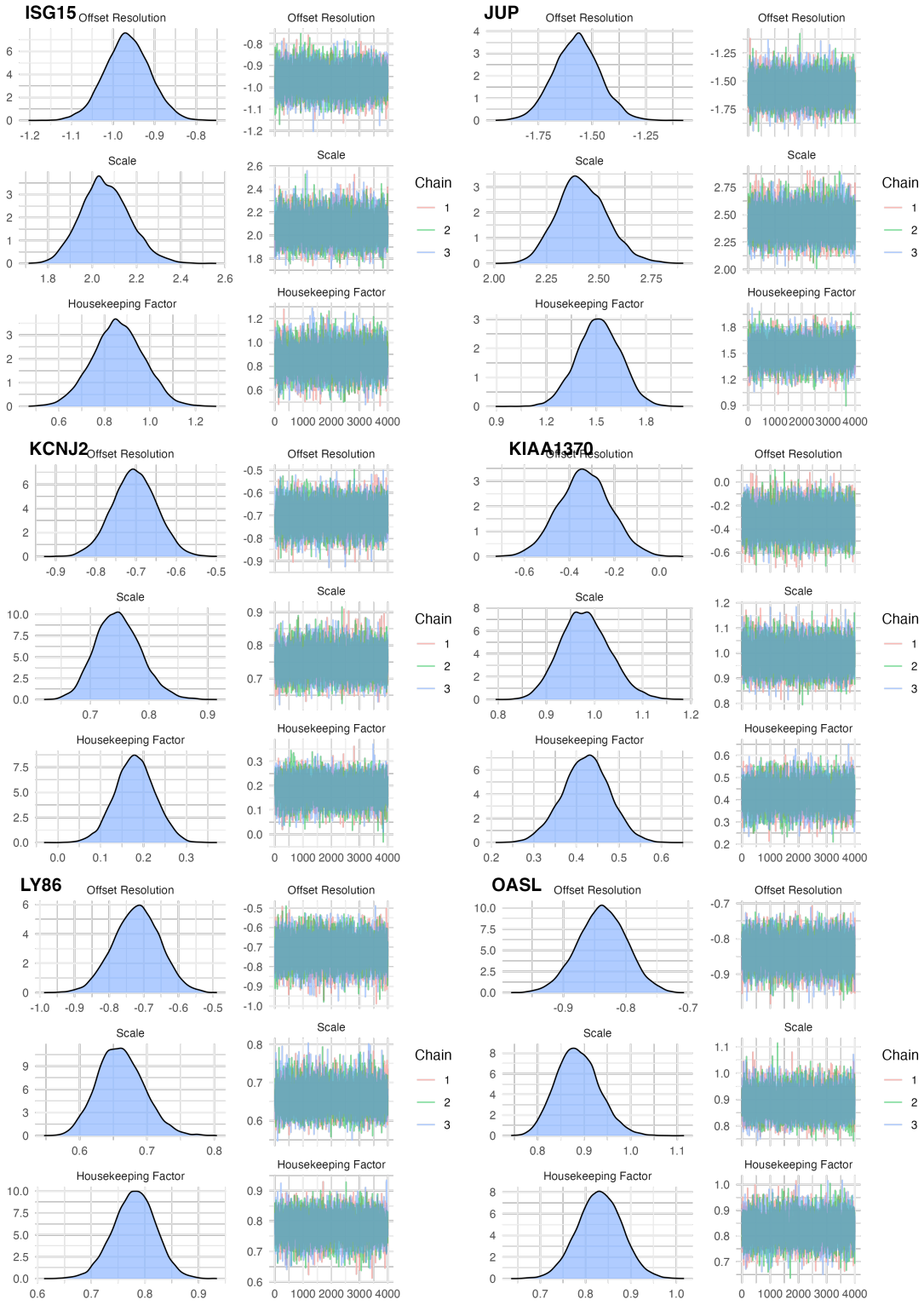

Figure B.11: Model convergence summaries for each of the 25 single-target offset models summarised in Table 2.3 (continued).

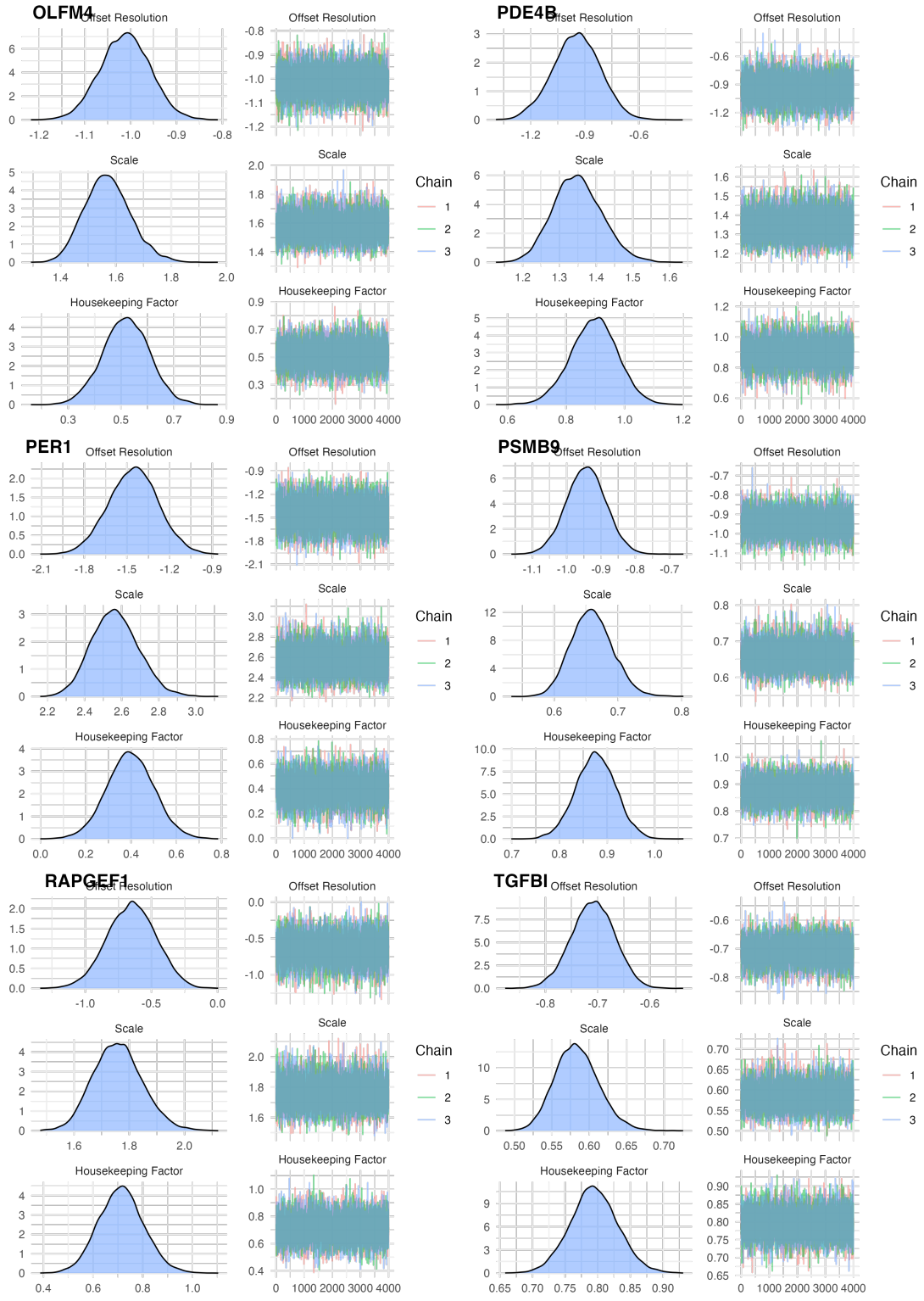

Figure B.12: Model convergence summaries for each of the 25 single-target offset models summarised in Table 2.3 (continued).

Appendix C

Open-source software accompanying TMB estimation

In this appendix we discuss software and computational aspects of the analysis presented in Chapter 3.

C.1 R package `ICBioMark`

In this section we'll demonstrate use of the R package `ICBioMark` (Bradley and Cannings, 2021b) on a toy example containing simulated data. The same core workflow, applied to real data, will form almost all of the results described throughout the remained of the chapter.

The package `ICBioMark` is available via the R repository CRAN and so can be installed and loaded with:

```
install.packages('ICBioMark')
library(ICBioMark)
```

or from GitHub with:

```
install.packages('devtools')
devtools::install_github('cobrbra/ICBioMark')
library(ICBioMark)
```

(note that the GitHub version is development and may not be as stable as the CRAN release, but may contain more features). The toy dataset we use here comes pre-loaded with the package, but just comes from the data simulation function generate_maf_data(), which allows for a variety of choices of dataset size and shape.

Our example dataset, called example_maf_data, is a list with two elements: maf and gene_lengths. These two pieces of data are required for most of the tasks performed by `ICBioMark`. They are organised as follows:

1. The dataset maf is a data frame in mutation annotated format. For a set of sequenced tumour/normal pairs, this means a table with a row for every mutation identified, with columns corresponding to properties such as the sample ID for the tumour of origin, the gene, chromosome and nucelotide

location of the mutation, and the type of mutation observed. In the real world, MAF datasets often have lots of extra information beyond this, but for our example we only include sample, gene and mutation type. See the first five rows in Table C.1 – generating code below:

```
head(example_maf_data$maf, 5)
```

Tumor_Sample_Barcode	Hugo_Symbol	Variant_Classification
SAMPLE_96	GENE_14	Missense_Mutation
SAMPLE_73	GENE_14	Frame_Shift_Ins
SAMPLE_55	GENE_4	Missense_Mutation
SAMPLE_96	GENE_3	Missense_Mutation
SAMPLE_38	GENE_7	Missense_Mutation

Table C.1: First five rows of example_maf_data$maf.

2. The data frame gene_lengths contains the names (referred to by Hugo Symbol) of genes to be included in downstream modelling, alongside their length. As discussed above, gene length is a complex and subtle quantity to define – we advise using coding length as defined in the *Ensembl* database (Yates *et al.*, 2020). For this example, however, gene lengths are again randomly chosen by the simulating function. Example rows are given in Table C.2 with accompanying code snippet below:

```
head(example_maf_data$gene_lengths, 5)
```

Tumor_Sample_Barcode	Hugo_Symbol
GENE_1	961
GENE_2	1009
GENE_3	1011
GENE_4	976
GENE_5	1016

Table C.2: First five rows of example_maf_data$gene_lengths.

User-provided gene length datasets, which we recommend extracting from the *Ensembl* database (Yates *et al.*, 2020), are permitted to contain values for more genes than are observed in the accompanying MAF dataset. In general, if a few genes covered by a WES experiment are missing gene length information it will not drastically impact model performance, but lots of missing values will begin to cause issues with model accuracy. Later versions of this `ICBioMark` will address missing gene length data.

In our next step we specify training/validation/test split and build model matries. The MAF format is widely used and standardised, but not especially helpful for sample(/patient)-level prediction. The ideal format for our data is a matrix in which every row corresponds to a sample, every column corresponds to a gene/mutation type combination, and each entry corresponds to how many mutations in that sample, gene and type were identified by sequencing. At the

same time as this, we'd like to separate our training data from separately reserved validation and test data. We do this using the function get_mutation_tables().

Our procedure, described in Section 3.2.2, models different mutation types separately, so in theory one could have separate parameters for each mutation type (e.g. `'Missense_Mutation'` or `'Frame_Shift_Ins'`). However, doing so would vastly increase the computational complexity of fitting a generative model. It is also not particularly informative to fit parameters to extremely scarce mutation types. Mutations types are grouped and filtered by the function get_mutation_dictionary(). In general we recommend separately modelling indel mutations (so that we can predict TIB later), synonymous mutations (as these don't count towards TMB or TIB), and grouping together all other non-synonymous mutation types. The function get_mutation_dictionary() allows for the production of a list of mutation types, with labels for their groupings.

Next we produce our training, validation and test sets. Again, for this example workflow these are availbale pre-loaded, but can also be produced with the 'get_mutation_tables' function. The object produced has three elements: 'train', 'val' and 'test'. Each of these contains a sparse mutation matrix ('matrix') and other information describing the contents of the matrix ('sample_list', 'gene_list', 'mut_types_list', and 'col_names'). See example code below.

```
example_tables <- get_mutation_tables(
    example_maf_data$maf,
    sample_list = paste0("SAMPLE_", 1:100)
)
print(head(example_tables$train$col_names, 10))
>  [1] 'GENE_1_NS' 'GENE_1_I'  'GENE_1_I'  'GENE_2_NS'
>  [5] 'GENE_2_I'  'GENE_2_S'  'GENE_3_NS' 'GENE_3_I'
>  [9] 'GENE_3_S'  'GENE_4_NS'
```

At this point we are ready to fit a generative model, for which we need only to provide gene lengths data and training data to the function fit_gen_model(). We can visualise output of our model with vis_model_fit() (see Figure C.1). Since this is a small example, we don't get a particularly strong signal, but we do see an optimum level of penalisation.

```
example_gen_model <- fit_gen_model(
    gene_lengths = example_maf_data$gene_lengths,
    table = example_tables$train
)

print(vis_model_fit(example_gen_model))
```

We now construct a first-fit predictive model. The parameter lambda controls the sparsity of each iteration, so it may take some experimentation to get a good range of panel lengths. From this, we can construct a range of refit estimators.

```
example_first_pred_tmb <- pred_first_fit(
    gen_model = example_gen_model,
    lambda = exp(seq(-9, -14, length.out = 100)),
    training_matrix = example_tables$train$matrix,
```

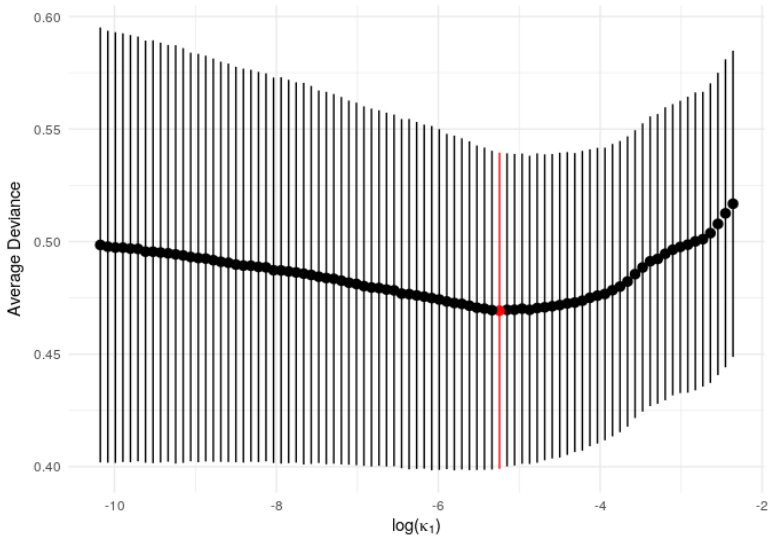

Figure C.1: Output example showing the variation in cross-validated generative model fit for a model fitted to example simulated data.

```
        gene_lengths = example_maf_data$gene_lengths
    )
    example_refit_range <- pred_refit_range(
        pred_first = example_first_pred_tmb,
        gene_lengths = example_maf_data$gene_lengths
    )
```

With a predictive model fitted, we can use the function get_predictions() along with a new (validation or test) dataset to produce predictions on that dataset. We then provide several functions including get_stats() to analyse the output compared to true values (example results in Figure C.2).

```
example_refit_range %>%
    get_predictions(new_data = example_tables$val) %>%
    get_stats(
        biomarker_values = example_tmb_tables$val,
        model = "T",
        threshold = 10
    ) %>%
    ggplot(aes(x = panel_length, y = stat)) +
        geom_line() +
        facet_wrap(~metric) +
        theme_minimal() +
        labs(
            x = "Panel Length",
            y = "Predictive Performance"
        )
```

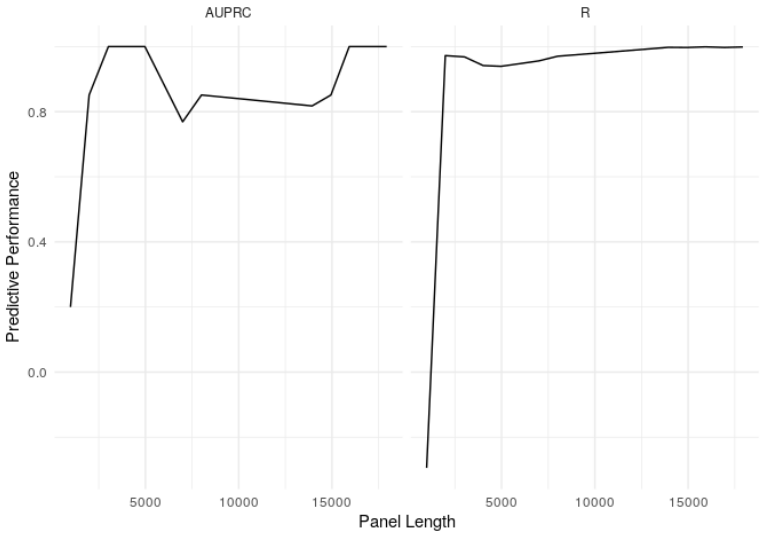

Figure C.2: Example prediction summary statistics from `ICBioMark` estimators applied to simulated data.