The University Of Sheffield.

Quality and clinical utility of genomic variants in complex diseases

By:

Timothy Martin Freeman

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield

Faculty of Medicine, Dentistry and Health.

Department of Neuroscience

Submission Date: May 2021

Abstract

Continuous improvements in high-throughput genomic sequencing over the past two decades have made it exponentially faster and cheaper, enabling its routine use in the clinic and scientific research. Genomic prognostic tools make use of personalised genomic data to aid clinical decision making and inform patients of disease outcomes, allowing enhanced tailoring of treatment beyond traditional prognostic tools, which are insufficient for understanding the nuances of individual complex disease cases. This relies upon accurate sequencing data and effective quality control. We have developed improved genomic prognostic tools for use in the clinic and demonstrate a novel method for quality control of genomic sequencing data with broad applicability.

Non-small-cell lung cancer (NSCLC) is the second most common cancer type in both males and females globally. Previous attempts to predict survival time for cancer patients have used genomic prognostic tools based on the burden of tumour mutations and neoantigens, but with limited success. We developed greatly improved classifiers of tumour mutation and neoantigenic burden showing strong 5-year survival differences between early-stage NSCLC patients. By using these together, we showed additional increases in prognostic efficacy, with the best survival group displaying a ~92% decreased risk of death in a 5-year period compared to the worst survival group.

To improve the accuracy of sequencing data for uses such as this, we developed the first tool for automatically cataloguing systematic sequencing biases for a sequencing pipeline, and we demonstrated its value in human and SARS-CoV-2 sequencing quality control with Illumina and Oxford Nanopore sequencing. We discovered and blacklisted a range of false positive variants, and investigated the causes of these. Identifying these errors contributed to multiple studies, altering research conclusions. We share these tools to provide continued improvements to genomic prognostics and sequencing accuracy affecting a wide range of fields.

Publications and manuscripts

Wang, Dennis, Nhu-An Pham, Timothy M. Freeman, Vibha Raghavan, Roya Navab, Jonathan Chang, Chang-Qi Zhu, Dalam Ly, Jiefei Tong, Bradly G. Wouters, Melania Pintilie, Michael F. Moran, Geoffrey Liu, Frances A. Shepherd and Ming-Sound Tsao. 2019. “Somatic Alteration Burden Involving Non-Cancer Genes Predicts Prognosis in Early-Stage Non-Small Cell Lung Cancer.” *Cancers*. https://doi.org/10.3390/cancers11071009.

Freeman, Timothy M., Genomics England Research Consortium, Dennis Wang, and Jason Harris. 2020. “Genomic Loci Susceptible to Systematic Sequencing Bias in Clinical Whole Genomes.” *Genome Research* 30 (3): 415–26. https://doi.org/10.1101/gr.255349.119

Korber, Bette, Will M. Fischer, Sandrasegaram Gnanakaran, Hyejin Yoon, James Theiler, Werner Abfalterer, Nick Hengartner, Elena E. Giorgi, Tanmoy Bhattacharya, Brian Foley, Kathryn M. Hastie, Matthew D. Parker, David G. Partridge, Cariad M. Evans, Timothy M. Freeman, Thushan I. de Silva, and The Sheffield COVID-19 Genomics Group. 2020. “Tracking Changes in SARS-CoV-2 Spike: Evidence That D614G Increases Infectivity of the COVID-19 Virus.” *Cell* 182 (4): 812–27.e19. https://doi.org/10.1016/j.cell.2020.06.043

Freeman, Timothy M., Matthew Wyles, Adrienn Angyal, Luke R Green, Paul Parsons, Rachel M Tucker, Rebecca Brown, Danielle Groves, Katie Johnson, Laura Carrilero, Joe Heffer, David Partridge, Cariad Evans, Mohammad Raza, Alexander J Keeley, Nikki Smith, Thushan I de Silva, Dennis Wang, The COVID-19 Genomics UK (COG-UK) Consortium and Matthew D. Parker. 2021. “Benchmarking SARS-CoV-2 Oxford Nanopore Sequencing Pipelines and Generation of a Variant Blacklist.” *Prepared manuscript*.

Acknowledgements

First and foremost, I would like to thank my supervisor, Dr Dennis Wang, for his continuous, generous support throughout my PhD. I have known Dennis since working with him during my MPhil dissertation at the University of Cambridge, and knowing that he would be my supervisor was a key reason for my decision to undertake this PhD in the first place. Dennis has gone above and beyond to further my career and enabled me to undertake multiple collaborations, with access to additional data, funding and resources, to make the most out of my PhD. Dennis' wise advice has helped me develop both as a scientist and as a person.

I am also very grateful to Dr Jason Harris at Personalis Inc. and Dr Matt Parker at the University of Sheffield, for their supervision and feedback on the publications I completed with them as senior authors. Their assistance and patience has been invaluable in developing my academic writing and experimental independence. Furthermore, I thank Personalis Inc., Genomics England, and COG-UK for sharing sequencing data with me on which my research was based, and collaborating with me on the publication of results.

I have many other inspiring colleagues at the University of Sheffield, past and present, who have contributed towards this thesis by providing feedback at lab meetings on my experimental work and presentations for conferences and interviews. I would particularly like to thank Szen for his detailed advice, helping me to settle into Sheffield and getting me involved as the Life Sciences Lead in the MedTech society, as well as Claire who introduced me to the Medical School PGR forum, inspiring me to become the Chair of the Faculty PGR Society. The soft skills I gained organising hackathons, conferences and events, working with wonderful groups of people and regularly speaking in public as part of these societies greatly boosted my confidence outside of my field and helped me to feel I was part of a larger Sheffield scientific community.

During my PhD journey, my family, greyhounds and friends have lifted my spirits and provided support outside of academia. I have enjoyed fun times with the Cambridge Space Dwarves and my bouldering friends. Last but not least, I am grateful to my amazing partner Martha, who has provided endless support in the form of humour, fine cooking, music, artistry, holiday planning and hugs. Martha has tolerated many weekends and nights of me working during crunch periods and has been the best lockdown partner anyone could ever ask for. I look forward to many more fun adventures together in the next steps of our journey.

Table of Contents

List of figures

List of tables

List of abbreviations

A	Adenine
ABL	ABL proto-oncogene 1, non-receptor tyrosine kinase (gene)
ACE2	Angiotensin-converting enzyme 2 (gene)
AF	Allelic fraction
ALK	Anaplastic lymphoma kinase (gene)
ALT	Alternative (non-reference allele)
AML	Acute myeloid leukaemia
ANN	Artificial neural network
APC	Antigen-presenting cell
BAM	Binary Alignment Map (file format)
BCR	Breakpoint cluster region (gene)
BED	Browser Extensible Data (file format)
BLAST	Basic Local Alignment Search Tool (software)
bp	Base pairs
BRAF	B-rapidly accelerated fibrosarcoma (gene)
BWA	Burrows-Wheeler algorithm
BWA-MEM	BWA-Maximal Exact Matches
CAR-T	Chimeric antigen receptor T (cell)
C	Cytosine
CD4+	Helper T lymphocytes
CD8+	Cytotoxic T lymphocytes
CFTR	Cystic fibrosis transmembrane conductance regulator (gene)
CI	Confidence interval
CNV	Copy number variant
COG-UK	COVID-19 Genomics UK (consortium)
COVID-19	Coronavirus disease 2019

DEL	Deletion
DFTD	Devil facial tumour disease
EGFR	Epidermal growth factor receptor (gene)
ER	Endoplasmic reticulum
ERGIC	ER-to-Golgi intermediate compartment
FASTA/FASTQ	FAST-All/FAST-Quality (file format)
GATK	Genome Analysis Toolkit (software)
GeCIP	Genomics England Clinical Partnership (consortium)
G	Guanine
GIAB	Genome in a Bottle (consortium)
GISAID	Global Initiative for Sharing All Influenza Data (repository)
gnomAD	Genome Aggregation Database
GRCh37/38	Genome Reference Consortium Human genome build 37/38
GWAS	Genome-wide association study
GZMB	Granzyme B (gene)
HAC	High-accuracy (Guppy basecalling model)
HER2	Human epidermal growth factor receptor 2 (gene)
HIV-1	Human immunodeficiency virus 1
HLA	Human leukocyte antigen
HPV	Human papillomavirus
HR	Hazard ratio
HTS	High-throughput sequencing
HTT	Huntingtin (gene)
IGV	Integrative Genomics Viewer (software)
IncDB	Incremental Database
INS	Insertion
KIT	KIT proto-oncogene, receptor tyrosine kinase (gene)
LIPT2	Lipoyl(octanoyl) transferase 2 (gene)

LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MAF	Mutant allelic fraction
MATLAB	Matrix Laboratory (Programming language)
MERS	Middle-East respiratory syndrome
MHC	Major histocompatibility complex
MPL	Mutation-prone locus
MPS	Massively-parallel sequencing
NAG	Number of altered genes (among a specific 865 gene panel)
NGS	Next-generation sequencing
NHS	(UK) National Health Service
NIST	The National Institute of Standards and Technology
NonDEL20	Non-deletion variant with >20% reads supporting deletions
NPRL2	Nitrogen permease regulator 2-like (gene)
NSCLC	Non-small cell lung cancer
ONT	Oxford Nanopore technology
ORF	Open reading frame
OR	Odds ratio
OS	Overall survival
PacBio	Pacific Biosciences
PCR	Polymerase chain reaction
PDGFRA	Platelet-derived growth factor receptor A (gene)
PDX	Patient-derived xenograft
PFS	Progression-free survival
PRF1	Perforin (gene)
Q	Quality
RdRp	RNA-dependent RNA polymerase (protein)
RLE	Run-length-encoded (Guppy basecalling model)

RNA-seq	RNA-sequencing
RSEM	RNA-seq by expectation maximisation
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
SARS	Severe acute respiratory syndrome
SD	Standard deviation
SMM	Stabilised matrix method
SNV	Single nucleotide variant
TAP	Transporter associated with antigen processing (protein)
TCGA	The Cancer Genome Atlas
TCR	T-cell receptor (protein)
TDDS	Targeted drug delivery system
TIMER	Tumour immune estimation resource
TMB	Tumour mutation burden
TNM	Tumour-Nodes-Metastases (staging system)
TSV	Tab-separated values (file format)
T	Thymine
T-Vec	Talimogene laherparepvec
UCSC	University of California, Santa Cruz
U	Uracil
UV	Ultraviolet
VAF	Variant allelic fraction
VCF	Variant Call Format (file format)
VQSR	Variant Quality Score Recalibration
vt	Variant Tool (software)
WGS	Whole Genome Sequencing
WHO	World Health Organisation

Chapter I: Introduction

Genomics, sequencing and its applications

Brief history of genomics

A genome is defined as the complete set of genetic material of a human, animal, plant, or other living thing and includes all DNA, or RNA in RNA viruses, found in chromosomes, mitochondria and chloroplasts within an organism, and is further subdivided into categories such as coding and non-coding regions. Genomics is the study of genomes and includes numerous areas, ranging from structural and functional genomics: the study of the physical structure and biological function of genomes and their individual components such as genes; through to epigenomics: how these regulate and are regulated by each other and chemical changes such as histone methylation; and comparative genomics: the study of the evolutionary relationships between the genes and genomes of different species or lineages of the same species. These genomics approaches are often used in combination to study how genomes function and how their dysfunction can lead to disease. This informs genomic diagnostics: how to diagnose disease using sequencing information; and pharmacogenomics: the development of new drugs and treatments in order to optimise patient health.

The first genome sequenced was of bacteriophage MS2, a single-stranded RNA virus which infects many bacteria such as *E. coli*, at the University of Ghent in Belgium, in 1976 (Fiers et al. 1976). Bacteriophage MS2 has one of the smallest genomes known, at 3,569 nucleotides long, about 850,000x smaller than the human genome, and does not contain repetitive sequence regions, making it much easier to sequence. Nearly 20 years later, the first bacterial genome, *H. influenzae*, was sequenced at the Johns Hopkins University School of Medicine in 1995, with a genome size of 1.8 million nucleotides (Fleischmann et al. 1995), and the following five years saw the first genomes sequenced of a fungus, *S. cerevisiae* (Goffeau et al. 1996); animal, *C. elegans* (The C. elegans Sequencing Consortium* 1998); and plant, *A. thaliana* (Initiative and The Arabidopsis Genome Initiative 2000) respectively. All of these eukaryotes are common model organisms for studying genetics and the sequencing of their genomes allowed scientists to greatly improve their general understanding of genomics across all fungi, plants and animals, since many of their genomic components are highly conserved within these groups. The successful completion of the Human Genome Project was announced on the 14th April, 2003 (Collins et al. 2003) and the resulting human genome was described in 2004 (International Human Genome Sequencing Consortium 2004), with more

than 99% coverage, laying the groundwork for genomics to become technically possible as a discipline. This included the development of much faster sequencing methods to determine the sequence of the four bases adenine, cytosine, guanine and thymine (A, C, G, T) across whole genomes rather than just small sections of DNA, known as whole genome sequencing (WGS).

Mutation and genomic diversity

DNA sequences encode the genetic instructions required for the essential biological systems in all cell-based organisms, contained within the genome. Differences in the sequence of bases across the genome give rise to genetic diversity between species and within different individuals of the same species. A *mutation* is a change in the sequence of bases within an individual organism's genome, which can in turn lead to the existence of genomic *variants*. Mutations can be caused by a range of factors, including environmental factors such as exposure to mutagenic chemicals and ultraviolet light, but can also occur spontaneously at a low level due to the chemical instability of the DNA bases, or from errors occurring during replication of DNA and cell division (Lodish, Berk, and Zipursky 2000). In the field of genomic sequencing, a variant is defined as one or more consecutive nucleotides that are different from the reference genome at that position. Examples include single nucleotide variants (SNVs), such as where a G is substituted by a T, and indels, which are variants that include insertions and/or deletions, such as if the reference sequence ACCC changes to GGGGTTT (Bohannan and Mitrofanova 2019). Variants can be further subdivided into coding and non-coding variants based on whether they occur within genes encoding a protein. Variants in coding regions which alter the amino acid sequence of the resulting protein are known as nonsynonymous, while some SNVs, termed synonymous, do not alter the amino acid sequence due to redundancy in the genetic code. For example, CCA and CCG codons both encode the amino acid proline, so an A>G SNV in this codon is synonymous. Nonsynonymous variants are classed as either missense, if the codon change causes an amino acid substitution, or nonsense, if the codon change results in a STOP codon causing the premature termination of translation, and also as frameshift variants, if an indel results in the length of the coding sequence changing by a number that is not a multiple of three. SNVs make up roughly 80% of all genomic variants between individual humans due to a higher rate of occurrence and lower evolutionary selection against them, compared to larger variants which are more likely to be deleterious (Katsonis et al. 2014). SNVs are therefore the main focus of this thesis, alongside indels.

In some cases multiple nucleotide variants occur between samples which are the same length, such as ACG to TTT. Because the process of mutation is not generally observed, in these cases we cannot infer whether such a variant is the product of multiple consecutive substitutions, balanced insertions and deletions, or a mix of both. In addition to these types of variants, larger-scale variants can occur as a result of the loss, duplication or movement of large sections of DNA. Humans typically have two copies of most sections of DNA that are not within an X or Y chromosome, and so are termed diploid, and changes in the number of copies of a section of DNA are termed copy number variants (CNVs). Genomic variants that were caused by mutations occurring after conception in an individual's lifetime are known as somatic variants, and are only present in cells descended from the original cell in which that mutation occurred, such as tumour cells sharing an oncogenic variant not found in healthy cells of the same individual, while variants present at conception, either inherited or caused by *de novo* mutations in egg or sperm cells, are known as germline variants. In the field of virology a variant is instead defined as the entire genetically-distinct genome of a virus, rather than the precise nucleotide changes that cause this (Kuhn et al. 2013). Where I use the term *variant* in this thesis, I am referring to the genomics definition.

Common applications of genomics

Targeted and whole genome sequencing have many applications across medicine and biological research. Genetic variants in the DNA that cause disease can be identified in patients using sequencing approaches, even before the resulting disease presents any symptoms. This can be used to diagnose inherited diseases such as Huntington's disorder, caused by high numbers of CAG repeats — usually 40+ — in the *HTT* gene (Lee et al. 2012), or cystic fibrosis, usually caused by one or more mutations in the gene encoding the CFTR protein (Klimova et al. 2017), but also genetic diseases resulting from *de novo* mutations in germline cells, such as Down syndrome, which is caused by trisomy of chromosome 21, and in somatic cells, including most cancers (Goldmann, Veltman, and Gilissen 2019). In addition to this, most non-inherited diseases have a mix of genetic and environmental components that affect how likely individuals are to develop that disease, and with what severity. By correlating the genomic data of large groups of people with their medical records, genomic variants can be found that correlate with disease risk, using approaches such as genome-wide association studies (GWAS), which may suggest genes and pathways that are important in the biology of a disease, helping to discover their function (M. Chang, He, and Cai 2018). For example, the 100,000 Genomes Project in the UK completed sequencing of 100,000 genomes in December 2018, and provides scientists with access to this anonymised sequencing data along with

patients' corresponding phenotypic and occupational data, providing insights into risk factors for a wide range of diseases alongside other academic resources (*Genomics England Press Releases* 2018). Commercial companies such as 23andMe have also sequenced large numbers of human genomes and often provide personalised estimates of disease risk factors directly to any member of the public without the involvement of health professionals, along with genealogy information, although the accuracy of these tests is often poorly regulated and sometimes different tests will give starkly different estimates for disease risk (Kim et al. 2014).

In addition to these areas, sequencing is used in many molecular biology techniques that are not covered in this thesis, in research and the clinic, including targeted sequencing to confirm successful gene editing (Montaño et al. 2018), bisulphite sequencing to identify epigenomic markers (Kernaleguen et al. 2018), chromatin conformation capture-based methods to study the spatial organisation of chromatin within cells (Kempfer and Pombo 2020) and RNA-seq to study gene expression levels and splicing isoforms in cells (Chambers et al. 2018; Burgess 2019). These molecular biology techniques can be used in combination to study many complex biological research questions and are vital to understanding how cells, tissues and biochemical processes work on a fundamental level, which in turn benefits not just human medicine, but also innovations in other biological areas such as agricultural productivity and sustainability (Gupta, Kulwal, and Jaiswal 2019; K. Chen et al. 2019) and other areas such as human history (Nielsen et al. 2017) and palaeontology (Mitchell and Rawlence 2021).

Genomic biomarkers of diagnosis

Genomic diagnostics is the process of sequencing and analysing patient DNA, using the information gained to provide valuable disease prognostic information to patients and clinicians, and to optimise treatment decisions according to patient genotype. Together with clinical factors that are not genomic, such as patient age, sex, disease symptoms and health status, genomic biomarkers such as somatic or germline variants in DNA sequences, tissue mRNA expression levels and epigenetic markers can be used to identify appropriate treatments for diseases in specific individuals, referred to as "personalised genomic medicine" (Kerr et al. 2021). Genomic diagnostics has the potential to revolutionise healthcare with the increase in genomic data becoming available, since it can potentially allow fast, accurate diagnosis of health conditions before a patient presents any symptoms, inform them on the best way to treat these, and optimise their lifestyle to minimise deterioration of health.

Genomic diagnostics improves decisions on treatment choices for patients who present similar physical symptoms in the clinic, but have a different genomic background that would respond better to some treatments in a genome-specific manner. This is particularly common in cancer, where a patient's tumour may be sequenced to check for a specific panel of mutations that indicate whether the cancer will be susceptible or resistant to treatments such as Trastuzumab for oesophageal and breast cancer, which is only effective in tumours where the *HER2* gene is overexpressed (Mao et al. 2021). In addition, hereditary disease gene panels can be used to diagnose patients with these conditions also, and even allows sub-types of the same disease to be diagnosed that cannot be distinguished without genomic sequence data, but which may respond to treatments differently, such as different types of cystic fibrosis (Sharma and Cutting 2020) and Duchenne muscular dystrophy (Verhaart and Aartsma-Rus 2019; Schneider and Aartsma-Rus 2021). Genomic diagnostics also includes using sequencing to detect and diagnose infectious diseases in patients and can benefit in understanding disease transmission during pandemics to guide public health policy internationally. For example, SARS-CoV-2 genome sequencing has been used extensively to study the structure and function of its genes (Michel et al. 2020) and to confirm this infection in patients showing clinical symptoms, and can also identify the specific mutations present in each sample, allowing any variants of concern to be monitored if these are seen to be rapidly increasing in the population, as well as revealing patterns by which viruses spread between countries and communities (Korber et al. 2020). Genomic diagnostics can also reveal whether a bacterial infection can be treated with certain antibiotics, by showing which antibiotic resistance genes are present in a patient sample (Gurwitz 2019).

Genomic biomarkers of prognosis

Similar to genomic diagnostics, *genomic prognostics* is the process of using this same genomic data, including DNA sequencing data, mRNA expression levels and epigenetic data, to calculate prognoses for individual patients, indicating the likely course of their diseases with a range of different treatments, or none, in a personalised manner, instead of relying on blunt disease statistics that do not take into account the exact condition of the patient (Kratz and Jablons 2009). In order to develop genomic prognostic models, scientists and clinicians need to track metrics related to the health and survival of patients in the clinic alongside genomic and treatment information and identify causative correlations between these. Kaplan-Meier curves are a common way of following the survival of patients grouped into different categories (Ranstam and Cook 2017; Rich et al. 2010). A Kaplan-Meier curve shows time along the x-axis and patient survival on the y-axis, starting from 100%. Depending on the statistic of

interest, either *overall survival* (OS) or — in the case of progressive diseases such as cancer — *progression-free survival* (PFS) are usually used for the survival statistic, although *disease specific survival* — where only deaths from the disease under investigation cause a drop in the y-value — is also sometimes used. If the status of a patient who has not died is no longer known then they are censored from the graph and further patient deaths will result in a proportionally larger drop in the y-value of the curve. It is common for these graphs to have two or more different patient categories plotted on them and a statistic known as the *hazard ratio* can be calculated between any two curves for any period of time along the x-axis. For OS, the hazard ratio reflects the odds of a patient dying in a given time period compared to a patient in another group, and is greater than one if the patient has an increased risk of death or less than one for a decreased risk of death. The hazard ratio is usually calculated using a log-rank statistical test, along with a P-value showing whether the difference between the groups is statistically significant.

In order to determine how to subdivide patients into categories for a Kaplan-Meier curve, statistical and biological considerations must be taken into account. For example, if some categories of patients are too tightly defined, then the numbers of patients within those categories may be too low for identifying any meaningful statistically significant difference in prognosis from other categories (Yung and Liu 2020). At the same time, patient grouping needs to make biological sense and categories should not be based entirely on whichever groupings give the most statistically significant differences, since this will lead to overfitting of prognostic models that only apply within the training set and cannot be used for correctly predicting prognoses in other patients. Scientists developing deep learning models of patient prognostics need to be particularly wary of this, since the outputs of these models are often extremely complex and difficult to link back to what is known about the disease biology (Jiménez-Luna, Grisoni, and Schneider 2020).

Sequencing technologies

Since the completion of the Human Genome Project, sequencing has become increasingly faster and more accurate, with sequencing costs decreasing at an even faster rate than the cost of storing sequencing data (November 2018), despite the latter being exponential itself. This has resulted in the cost of sequencing a human genome hovering around $1000 for several years (**Figure 1.1**), with staff and analysis costs being much higher (Schwarze et al. 2020). Modern sequencing technologies demonstrating this very fast sequencing speed are known as high-throughput sequencing (HTS) technologies. The main HTS technologies in use today can be classified as short-read (Illumina, IonTorrent, Beijing Genomics Institute) and

long-read (Oxford Nanopore Technology (ONT), Pacific Biosciences (PacBio)) technologies and use a range of different methods with varying advantages and disadvantages (Kumar, Cowley, and Davis 2019). Reads are defined as the fragments of DNA or RNA that are sequenced in each case. In this thesis I worked with Illumina and ONT sequencing, which currently dominate the market for short-read and long-read sequencing respectively.

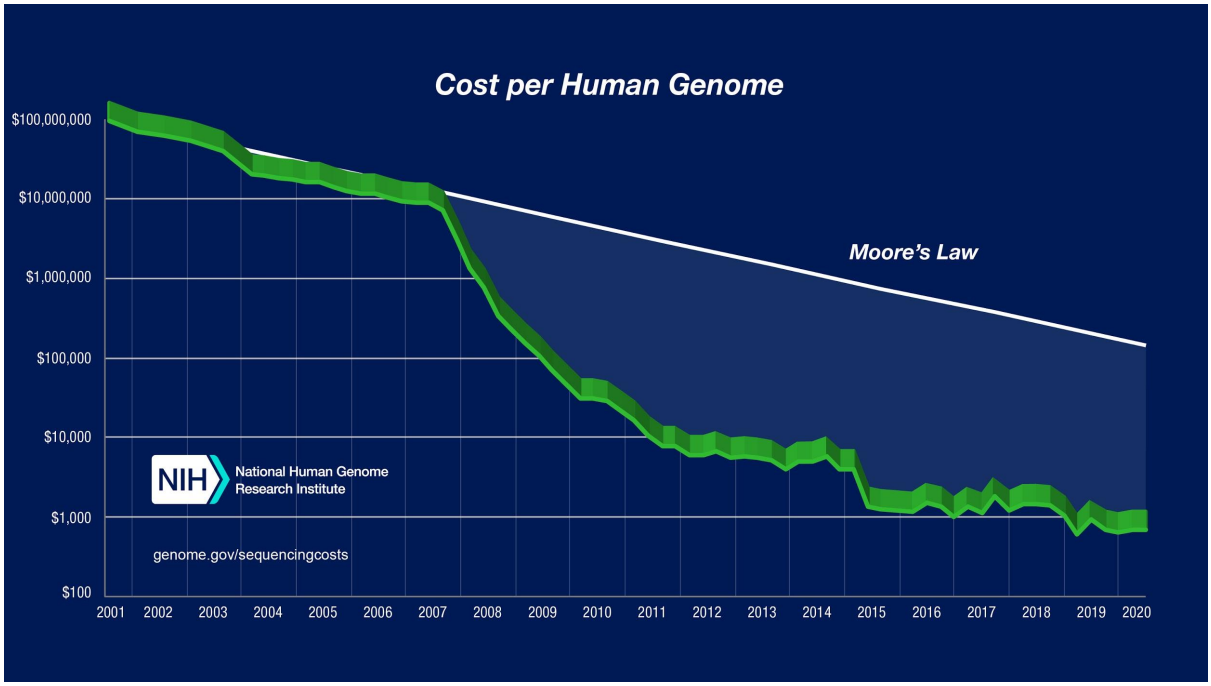


Figure 1.1: Twenty-year trend in sequencing costs — August 2000-2020 (Wetterstrand 2021). The cost of sequencing a single human genome, plotted in green, over the last two decades has decreased by more than half every two years, outstripping the rate of computational technology advances predicted by Moore's Law, shown as a white line.

Illumina sequencing uses sequencing-by-synthesis combined with camera imaging of large numbers of short DNA fragments simultaneously, one base at a time, known as massive parallel sequencing (**Figure 1.2**). DNA fragments are prepared by ligating adapter sequences to them which enable the DNA fragments to attach to complementary oligonucleotides on the surface of an acrylamide-coated glass flow cell (Quail et al. 2012). DNA fragments are spaced out in nanowells that prevent overcrowding, so that no two distinct DNA fragments are close to each other (Clark, Pazdernik, and McGehee 2018). The free ends of each individual DNA fragment loop around and the adapters bind to nearby oligonucleotides within the nanowell, forming a bridge structure which allows a primer to attach to that end, enabling DNA polymerase to make a copy of the DNA fragment such that forward and reverse strands of the same DNA fragment occur next to each other within the same nanowell. This process, named bridge amplification PCR, continues with the DNA fragments being repeatedly copied until

about 1,000 copies are present. Once this is complete, the sequencing step begins. Primers are washed onto the chip along with modified nucleotides that have a reversible 3' fluorescent blocker. The fluorescent blocker ensures that only one nucleotide can be added at a time to synthesise the beginning of the complementary strands for each DNA fragment, after which a camera images all of the nanowells simultaneously. Because there are large numbers of identical DNA fragments clustered next to each other in each nanowell, each cluster is visible as a single coloured dot at a fixed location on the image that corresponds to the identity of the nucleotide that was added (A, C, G, T), which is recorded by the machine for each different DNA fragment nanowell cluster in parallel. The flow cell is then washed to remove the fluorescent blocker on the incorporated nucleotide so that the previous step can be repeated, but instead incorporating the next nucleotide in the DNA sequence, which is again imaged and recorded by the machine. This step is repeated until the full DNA molecule for each nanowell has been sequenced, after which the sequence is provided for each separate DNA molecule cluster. In addition to this, the machine predicts a likelihood of its sequence prediction being correct at each individual base along the DNA fragments, on a logarithmic scale using the formula $Q = -10\log_{10}(e)$ where e is the probability of the sequence being incorrect at that position and Q is known as the *quality score*. Higher Q scores indicate a lower probability of a sequencing error. The process of identifying the sequence of bases and assigning quality scores to them is referred to as *basecalling*. There are different basecalling algorithms available for different types of sequencing which are often built into the sequencing machines themselves and give different sequencing results from the same raw sequencing data.

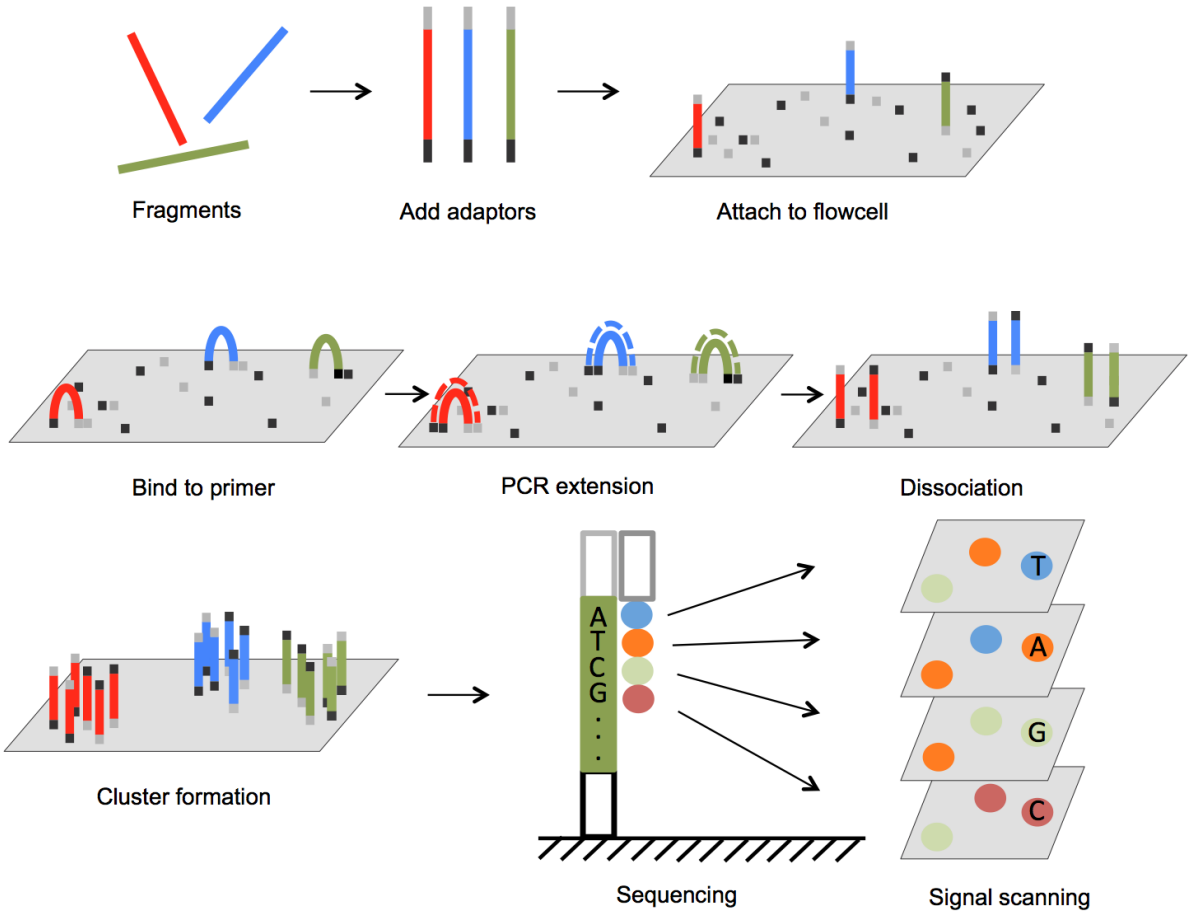


Figure 1.2: Illumina sequencing steps (Lu et al. 2016). Adapters anneal to the end of the DNA fragments, allowing them to attach to the flow cell and form a bridge structure, followed by cycles of PCR extension and dissociation to form clusters of identical DNA fragments. Sequencing by synthesis is carried out with one nucleotide read at a time by the addition of modified fluorescent bases that are detected by optical scanning. High throughput is achieved by simultaneously sequencing large numbers of different DNA fragment clusters in parallel on a single flow cell.

These individual short sequences are known as reads. Because each read is short, it only covers a small portion of the original genome being sequenced. Short-read alignment algorithms are used to recognise sequence overlap between the ends of reads, and join these up into contiguous sequences named contigs (**Figure 1.3**). Contigs are then mapped to a reference genome if one is available. If a reference genome is not available, then sequencing is considered *de novo.* This can leave gaps in the genome generated where repetitive sequences occur (Mandelker et al. 2016), since the reads may be too short to span these repetitive sequences or may have identical sequences to multiple different parts of the genome — a phenomenon known as homology — leading to many identical reads being present with different sources, without a clear indication of where to map these on the genome. Although this type of sequencing has a high coverage and basecalling accuracy at a low price (Kumar, Cowley, and Davis 2019), it is therefore more prone to gaps and mapping errors.

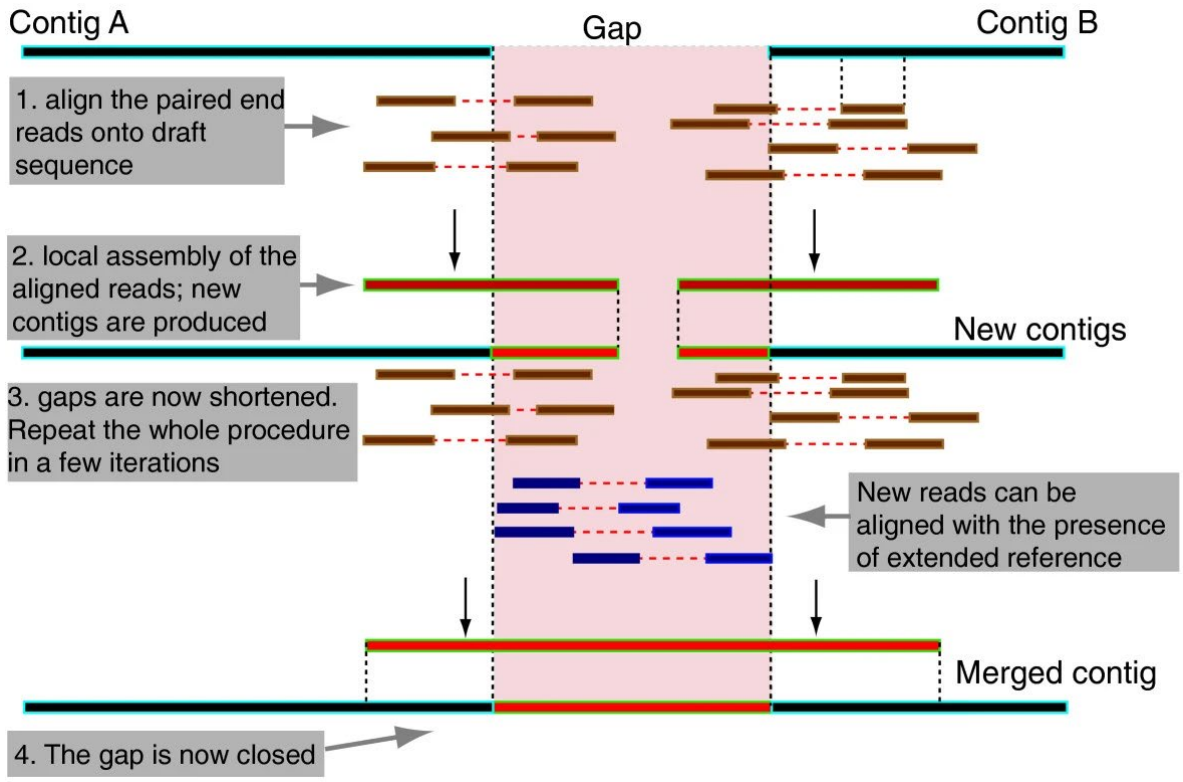


Figure 1.3: Genome sequence assembly using contigs (Tsai, Otto, and Berriman 2010). Short reads are aligned against the initial sequence assembly, with a small amount of overhang allowing its extension. The alignment process is repeated with reads that did not successfully align to the previous assembly, with some now aligning to the new contigs allowing further extension. This cycle is repeated until all remaining reads are aligned and gaps are closed, if possible.

ONT sequencing uses transmembrane protein nanopores embedded in a synthetic membrane to sequence long single DNA strands in real time (Oxford Nanopore Technologies 2021). Double-stranded DNA fragments are treated to ligate the end with a sequencing adapter and motor protein which allow the complex to dock at a nanopore, with the motor protein separating the DNA into two strands, allowing a single DNA strand to enter the nanopore protein and cross through it while the complementary single DNA strand does not cross. An electrical current is passed over the synthetic membrane and the crossing of the DNA strand through the nanopore causes fluctuations in this current that are different depending on the sequence of bases in the DNA (**Figure 1.4**). The resulting fluctuations in the electrical current are measured and can be read in real time by the machine, providing the sequence of bases that have passed through it. ONT technology does not have a theoretical read length limit (in practice reads are rarely longer than 50,000bp due to DNA shearing during preparation) and will continuously sequence any reads passed to it, allowing very long reads to be sequenced (Leggett and Clark 2017). The real time sequencing allows low quality DNA

sequences to be detected at an early stage, so the experiment can be stopped early if this is the case without needing to wait for the DNA to be fully sequenced (Xu et al. 2020). This allows experimental errors such as contamination to be detected and fixed rapidly. Nanopore sequencing has a higher basecalling error rate than Illumina sequencing, but reads are much easier to map due to their longer lengths (Amarasinghe et al. 2020), allowing gaps to be resolved more easily across the genome, and enabling the length measurement of genomic regions that are composed of long repeated sections of DNA subunits.

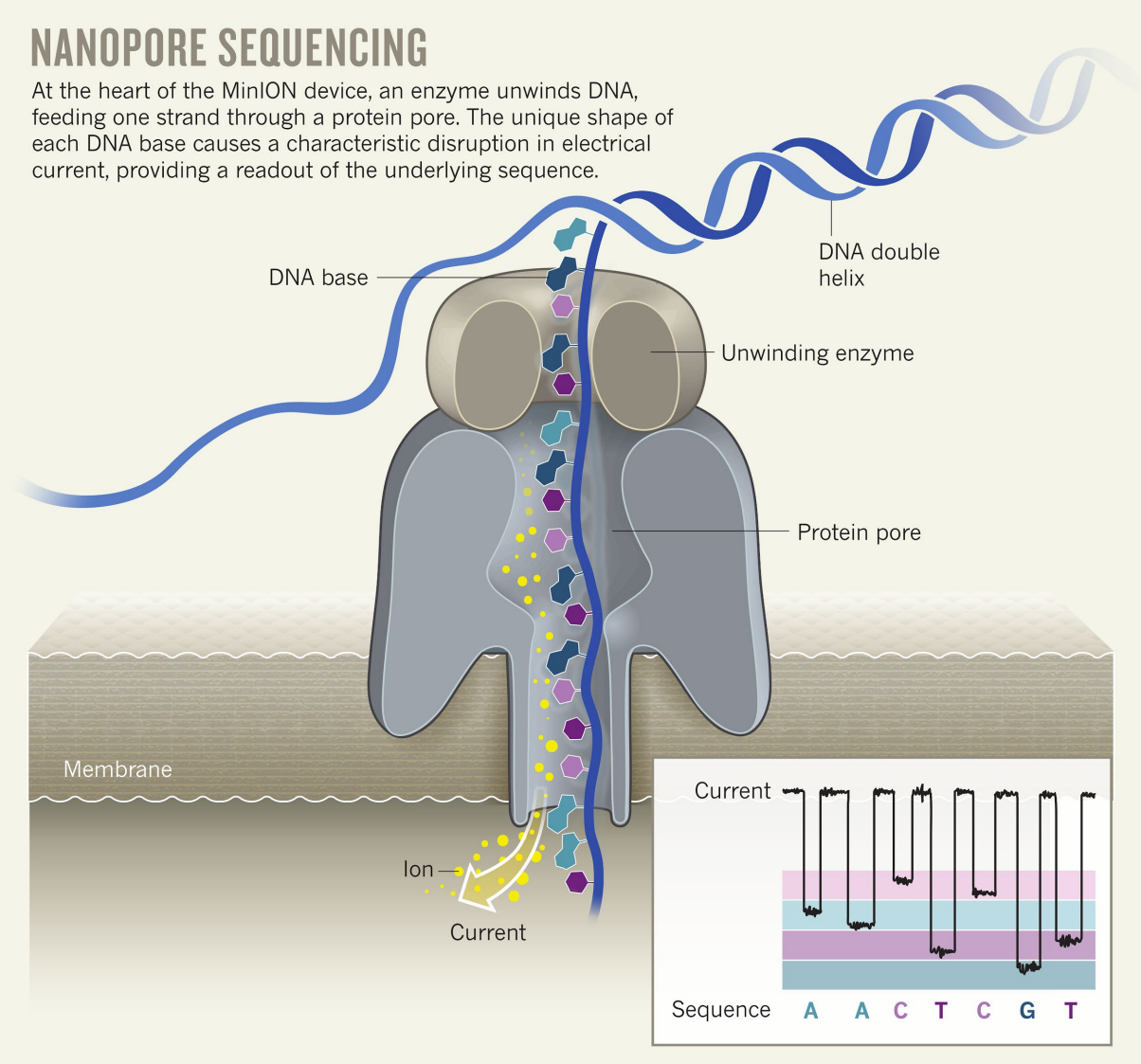


Figure 1.4: ONT sequencing (Eisenstein 2017). DNA is unwound at a protein nanopore and a single strand passes through. The change in the sequence of bases as it passes through the nanopore disrupts the electrical current through the synthetic membrane, which is measured. The current signal is converted in real time to a sequence of bases using the knowledge of how different bases affect this current.

Sequence alignment and variant calling

In addition to the physical sequencing processes carried out by sequencing technologies like Illumina and ONT, bioinformatics methods are usually required for post-processing of the read sequences generated in order to acquire finished sequences, with annotated markers of sequence quality and coverage that can be used to ascertain confidence in the accuracy of a genome sequence, especially with short-read sequencing. Two data processing steps of particular importance are sequence alignment and variant calling.

Sequence alignment is the process of identifying overlap between different DNA or RNA sequences (Mount 2004). This is used within the context of sequencing for joining sequencing reads up into contigs, for assembling genomes *de novo*, and for aligning reads and/or contigs to a reference genome, as detailed in the previous section. In these cases, reads require a high level of sequence identity, and longer sequence alignments are more likely to be accurate since there is a lower risk of larger sequences occurring elsewhere in the genome by chance. Sequence alignment is also used outside of sequencing for comparative genomic analysis such as determining sequence phylogeny between species, where the evolutionary distance between organisms can be computed by aligning gene and protein sequences against each other and measuring how close matches are to perfect (Ortet and Bastien 2010). Genes and proteins that show the closest matches over evolutionary time are considered to be more conserved, and this can indicate that they have a more important functional role, while genes showing heavy sequence divergence may be less important, or even pseudogenes that are an evolutionary relic of an ancestral species and no longer functional. In addition, sequence alignment is used to discover if genes present in one species are present in another species, and if their function is known in one case, then a similar function can be predicted in other species that the gene is found in. Many different sequence alignment tools are available, with methods for this constantly evolving, and with each tool usually having a specific specialisation for which it is optimised. For example, tools such as Bowtie (Langmead et al. 2009) and BWA (Li and Durbin 2009) are commonly used for short-read sequence alignment and make use of a computational technique known as the Burrows-Wheeler transform to minimise memory usage for aligning large numbers of short reads, while other tools such as BLAST (Altschul et al. 1990) are more specialised for the alignment of individual long reads against a database of multiple genomes.

Variant calling is the process of identifying whether a given variant is present at a specific position on the genome, based on the sequencing data provided. This is carried out using bioinformatics tools named variant callers, which analyse all of the reads that cover a

genomic position, checking what base is being called in each, and at what quality, as well as other factors, such as the overall quality of the reads and already-established information about sequencing performance or repetitiveness at a given position. Coverage thresholds are usually applied, below which a variant is not called, and reads with poor quality may be discarded. The context of the sequencing may also be used to make variant calling decisions. For example, in haploid organisms a real variant would be expected to be supported by close to 100% of reads, while in diploid organisms (**Figure 1.5**) a variant could be either heterozygous, in around 50% of reads — one copy out of two — or homozygous, close to 100% of reads — two copies — while in a tumour sample variants could theoretically occur with any percentage of supporting reads since the sample would likely contain a mixture of genetically heterogeneous cells that may also have copy number changes (Koboldt 2020). Many different variant callers exist that use a range of different methods tailored to the type of sequencing technology and sequencing context, and may annotate variants with further information and scores, allowing the end user to make the final decisions about which variants they believe are true positives.

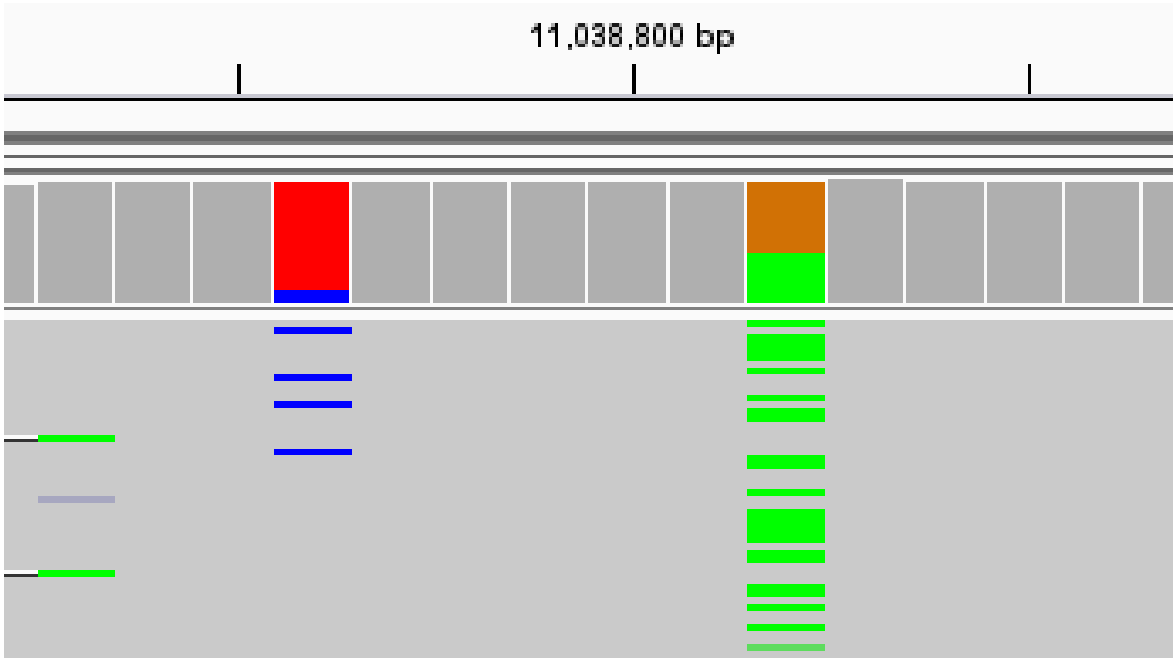


Figure 1.5: Example read coverage for a heterozygous variant. Sequenced reads are visualised with Integrative Genomics Viewer, showing possible variants in the human genome NA12878 chromosome 21 at positions 11,038,796 and 11,038,802. A, C, G, T basecalls are colour-coded green, blue, brown and red respectively. Alleles that match the reference are shown as grey. The sample shown is likely to be heterozygous for A and G at 11,038,802, but the low level of reads supporting a C variant at 11,038,796 indicates that the C basecalls are likely to be the result of sequencing error, with the position being homozygous for the reference T allele.

Improving the quality of variant identification

Challenges to accurate sequencing

In order to draw useful conclusions from sequencing data, it is important to ensure that it is as accurate as possible. This is of particular importance in clinical settings where inaccurate sequencing could result in patients' genetic diseases being misdiagnosed, resulting in incorrect advice or treatments being offered, with life-changing effects, but also within many research settings such as the identification of SARS-CoV-2 variants of concern which could have dramatic effects on national and international health policy. There are two types of sequencing errors that can occur: *false positive* and *false negative* variant calls. *False positive* variant calls falsely suggest the presence of genomic variants that are in reality absent in the sample, while *false negative* variant calls suggest the absence of a variant that is in reality present in the sample. These can be caused by factors at every stage of sequencing (**Figure 1.6**), such as contamination of a patient sample with microbial DNA — where part of the microbial DNA sequence is genetically very similar to the human reference genome with a small number of changes, such as at highly evolutionarily-conserved genes, leading to microbial DNA reads mapping to the human genome and causing inter-species sequence differences to be detected as false positive variant calls — or even DNA from the staff member taking the sample (Schmieder and Edwards 2011); limitations in the way samples are processed and prepared for sequencing, such as lack of DNA fragments from tightly-bound centromeric regions or chemical degradation of DNA fragments introducing base changes that were not present in the original sample (Costello et al. 2013); limitations of the sequencing technology itself, such as GC sequencing bias with Illumina technology — in which GC-rich DNA fragments have higher sequencing coverage than GC-poor fragments (Benjamini and Speed 2012) — or lower basecalling accuracy with ONT sequencing (Petersen et al. 2019); or the bioinformatics methods used to process and call variants from the sequencing data, such as mapping variants to homologous genomic regions where they do not occur (Shringarpure et al. 2017). Sequencing errors can be further categorised into systematic sequencing errors, which occur consistently across samples sequenced using the same sequencing protocol and techniques (Freeman et al. 2020), such as consistent mismapping of variants to the wrong genomic location by an aligner, and non-systematic errors, which only occur in a small number of samples, often in an irregular manner, such as individual contaminated samples.

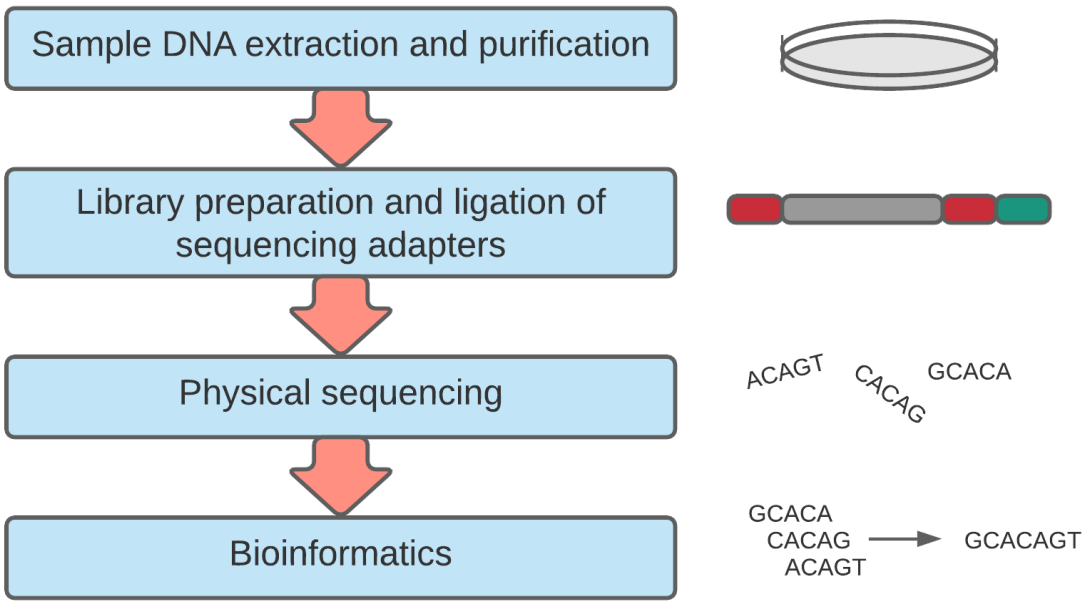


Figure 1.6: Generic workflow for NGS. DNA is extracted and purified from a physical sample. Library preparation involves shearing DNA into fragments of known length, with appropriate sequencing adapters annealed to the ends compatible with the sequencing technology used. This is followed by the physical sequencing process itself using machines that generate the raw read sequences observed. Computational methods such as read alignment are then used to combine these short read sequences into full genome sequences and carry out post-sequencing quality control and analysis. Sometimes some bioinformatics steps are built into the sequencing machinery itself, so there is some overlap. Errors and/or biases at any stage of sequencing can result in a loss of sequencing accuracy.

In addition to this, there is usually a compromise between false positive and false negative variant calls, since using stricter criteria for calling a variant at a position — such as higher coverage or quality thresholds for reads supporting a variant — generally minimises the number of false positive variant calls at the cost of a greater number of false negatives. The number of true positive variant calls detected divided by the total number of variants in a sample (true positives + false negatives) is referred to as the *sensitivity* of sequencing, while the number of true negative variant calls divided by (true negatives + false positives) is referred to as the *specificity* of sequencing. Depending on the exact context of sequencing, the user may prioritise one of these statistics over the other to different extents. For example, a test for diagnosing a genetic disease will usually prioritise sensitivity, since further testing with different methods can be carried out to confirm the result for a patient initially testing positive for a variant, while patients that initially test negative would not usually undergo further testing. Another important factor in the practicality of achieving high sequencing accuracy is cost: Sequencing samples for longer, with more amplification of DNA fragments costs more, but

results in a greater number of reads supporting each variant call, increasing the statistical power and confidence of variant calling; and Sanger sequencing, which is usually considered the gold-standard for validating the accuracy of other sequencing methods (De Cario et al. 2020), is extremely expensive and not used for whole genome sequencing outside of the original Human Genome Project.

Existing quality control approaches

In order to ensure that sequencing is accurate, it is important to carry out post-sequencing analyses of the results in order to determine whether the presence or absence of variant calls may have been caused by error at any stage of sequencing, and whether any steps can be taken to minimise this. The process of estimating sequencing accuracy and putting measures in place to counteract any sequencing errors or limitations is known as *quality control*.

There are a multitude of methods that can be used to identify and control for sequencing inaccuracies, the most common of which I summarise in this section. Most standard sequencing pipelines measure the depth of coverage and quality of basecalls across the genome and analyse their distributions. Genomic positions or regions showing significantly lower coverage are more likely to be affected by noise in genomic sequencing, especially if the basecalling quality is also poor, so variants that are called with fewer than ten genomic reads supporting them are usually filtered out. On the other hand, genomic positions with very high coverage are also suspicious, since this may be the result of reads from elsewhere in the genome mapping to a similar genomic location. For this reason, repeat regions often have high coverages. This could lead to variants from elsewhere in the genome being called at the incorrect location, so these are also often filtered out. Irregular coverage distributions can also indicate the presence of contaminant DNA that was sequenced along with the sample, which may come from a microbial source (Goig et al. 2020) or even from the staff member preparing the DNA sample for sequencing (Peyrégne and Prüfer 2020). Common contaminating sequences can be identified if they are in a database of known microbial or other contaminants, but contaminating DNA sequences that originate from similar species to the one being sequenced are harder to remove in this way, since this would filter out sequences from the sample itself. In addition, primer and adapter sequences that are ligated to the ends of DNA reads can act as contaminants, but these have known sequences and can usually be identified and trimmed by alignment algorithms, and also because they rarely occur in the middle of sequencing reads. In order to avoid primer and adapter sequences interfering with sequencing results in the first place, these can be selected based upon choosing sequences

that have minimal homologous sequences within the genome of interest, with the exception of any sites to which the primers need to bind for sequencing to occur.

Allelic biases may also occur to different extents in different sequencing contexts. For example, a GC content bias can be identified if reads with a higher GC content show higher coverage than lower GC content reads, by calculating the correlation between read GC content and coverage, as is the case with many Illumina pipelines (Benjamini and Speed 2012). Statistical and mathematical methods can also be utilised to identify patterns in sequencing data that are suggestive of sequencing errors, such as systematic biases. Systematic biases are systematic effects on sequencing that occur in all samples sequenced using the same protocols, at any stage from sample preparation, to the sequencing itself, to the processes of base and variant calling, and which may lead to consistent false-positive detection of genomic variants at some positions. Other quality control issues are commonly identified by analysing the proportion of reads that support each variant called, known as the *variant allelic fraction* (VAF). Variant calls with a low allelic fraction of reads supporting them are difficult to distinguish from low-level sequencing error, especially in cancer, and it can be difficult to determine what allelic fraction cutoffs are reasonable to use. To help with this, the likelihood of individual variant calls being accurate can be estimated directly from the numbers of reads supporting them or other alleles using statistical methods such as Bayesian inference. This determines the probability of the observed result occurring given different genotypes and assumed error rates and is used in common variant callers such as FreeBayes (Garrison and Marth 2012) and GATK (DePristo et al. 2011) to achieve a more accurate cutoff for variant calling than using a fixed variant allelic fraction cutoff (Sandmann et al. 2017), in particular increasing the sensitivity for detecting low allelic fraction variants from high-coverage tumour sequencing data. In general quality control can be improved by comparing sequencing results against those produced with a range of different bioinformatics tools, including read mappers, basecallers and variant callers and checking how the results from these compare, because using a range of tools will include more different quality control checks automatically, since these are often built into these tools, and variants that are consistently called across tools are more likely to be accurate.

Benchmarking sequencing methods

Benchmarking sequencing methods is the process of measuring the performance of different sequencing protocols and pipelines by directly comparing them against each other. This enables the continual optimisation of sequencing accuracy, speed and cost, by providing best

practice guidelines showing the optimal combinations of sequencing protocols and computational tools such as base and variant callers for different sequencing contexts as new tools are released and updated. Benchmarking also reveals weaknesses of sequencing pipelines, and can help their creators to identify the source of these and develop better wet lab protocols and bioinformatics software in response (**Figure 1.7**). In order to draw fair comparisons it is important to run benchmarking tests using the same input data for each sequencing pipeline or process being used, and to change minimal parameters between the different pipelines being tested, so that any resulting changes in sequencing outputs can be directly traced back to the exact parameter changes responsible, such as changes in individual basecalling algorithms or settings.

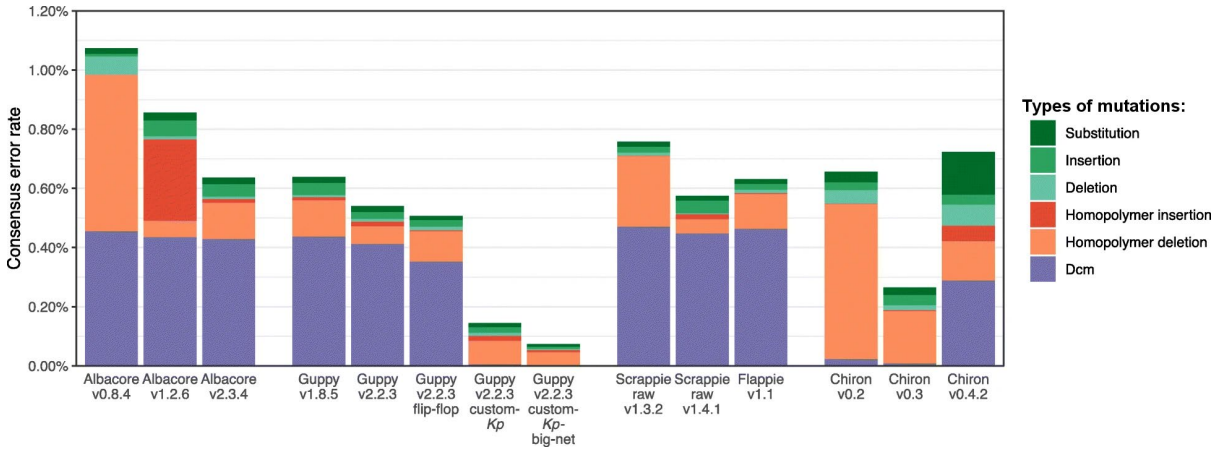


Figure 1.7: Benchmarking basecalling consensus error rate for four ONT basecalling algorithms — Albacore, Guppy, Scrappie/Flappie and Chiron — on the K. pneumoniae bacterial genome (Adapted from Wick, Judd, and Holt 2019). Consensus error rate bars are colour coded to show the type of error occurring. This reveals the best performing basecallers and their main strengths and weaknesses. The CCAGG/CCTGG Dcm motif is the main source of error, but is removed in the custom Guppy models and earlier versions of Chiron tested.

The National Institute of Standards and Technology Genome in a Bottle Consortium (NIST GIAB) has developed a set of gold-standard reference human genomes which have been sequenced with multiple sequencing pipelines, providing standardised human genome sequencing benchmarking results for the larger scientific community (Zook et al. 2014, 2019), in addition to sets of cell lines corresponding to these gold-standard reference genomes which it provides for benchmarking new pipelines. By comparing the exact sets of variant calls given by different sequencing pipelines NIST GIAB has proposed a “truth set” of variant calls which show consensus between these, covering roughly 75% of positions in the human genome, using the logic that if many different sequencing methods agree on a variant call, then it is highly unlikely to be inaccurate. There are some limitations, however, to the NIST GIAB “truth-

set"-based approach to annotating positions at which variant calls are expected to be accurate. Concordance between sequencing methods is not necessarily an indicator of a variant call being accurate, since this also occurs at genomic positions that are difficult to map reads to accurately, which can only be sequenced by a small number of similar methods, usually long- or linked-read-based methods, which are likely to give similar results regardless of whether these are correct or not. In addition, the number of reference genomes used by NIST GIAB is low, and structural variants that are unique to specific samples used by NIST GIAB therefore result in the affected genomic regions being absent from the 'high confidence' list despite having accurate sequencing in most other samples. A list of all of these positions, which they term "high-confidence" loci, is available to download (Zook et al. 2019), and scientists can use this for masking variant calls at genomic loci that do not show consensus, if they seek to maximise the specificity of their called variant sets, minimising false positives, at the cost of losing the ability to call variants outside of these regions. This is not an acceptable compromise for most sequencing however, since it discards many important genomic regions, so in practice scientists will typically still call variants outside of the high-confidence regions, with the caveat that they may apply higher thresholds for evidence supporting variants at these positions, such as high coverage or quality cutoffs. For genomic positions of clinical importance that show discrepancies in variant calls between different pipelines, the Sanger sequencing results are usually assumed to be the most accurate, if these are available, but even these are not guaranteed to be correct 100% of the time (De Cario et al. 2020). Caution is required when calling variants at positions where different sequencing pipelines do not agree on a variant call, since there is no way to definitively prove which pipeline is correct, and even if variant calls are supported by all sequencing pipelines available, it is still possible for them to be false positives. The human genome is 3.1 billion bp long, so even very low sequencing error rates would be expected to result in large numbers of false positive and false negative variant calls.

Cancer biology and the role of the immune system

Cancer biology overview and tumour evolution mechanisms

Cancer is the second most common cause of death globally after ischaemic heart disease, killing an estimated 10 million people in 2020 (Sung et al. 2021), but if current trends continue it may become the leading cause of death as global life expectancy increases and many countries face aging populations (Mattiuzzi and Lippi 2019). Cancer is a genetic disease typically characterised by the accumulation of specific sets of somatic mutations in cells that increase their ability to evade programmed cell death, leading to those cells dividing rapidly,

forming a tumour and eventually spreading to other tissues and organs around the body, causing them to fail and leading to the death of the patient. These driver cancer-causing mutations are termed *oncogenic*, but many non-oncogenic mutations also accumulate in tumours due to their genomic instability, which are termed *passenger* mutations and have a neutral or restrictive effect on tumour progression.

There are six main biological capabilities acquired by cancer cells during the formation of a tumour, known as the hallmarks of cancer (Hanahan and Weinberg 2011): These include "sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion and metastasis". Proliferative signaling and growth suppressor evasion are respectively the processes by which mutations either upregulate cell growth and division, or downregulate repression of these, leading to tumour growth. Resisting cell death is achieved through mutations that inactivate the programmed cell death pathways that would usually lead to the deaths of cells acquiring oncogenic mutations to maintain genomic integrity. Enabling replicative immortality is achieved through mutations that remove limits on the number of times a cell can divide, including mutations that activate telomerase, a specialised DNA polymerase that lengthens telomeres but is normally not expressed in differentiated cells — this extends the length of DNA molecules at both ends, compensating for the fact that regular DNA polymerases cannot copy the entire DNA molecule and generate shortened telomeres on the replicated genome with each round of cell division. Angiogenesis is the process by which tumour cells signal to their local environment to increase the flow of blood and nutrients to support tumour growth, and increase the number of blood vessels around the tumour. Invasion and metastasis are the final stages of cancer, in which tumour cells escape containment by the immune system, enter the bloodstream and migrate to other parts of the body where they divide to form secondary tumours. Tumours gradually acquire these hallmarks of cancer in steps as they grow and acquire more oncogenic mutations. A further four hallmarks are sometimes separated out as additional categories to the other six, including promoting inflammation, deregulating cellular energetics, promoting genomic instability and mutation, and avoiding immune destruction (**Figure 1.8**).

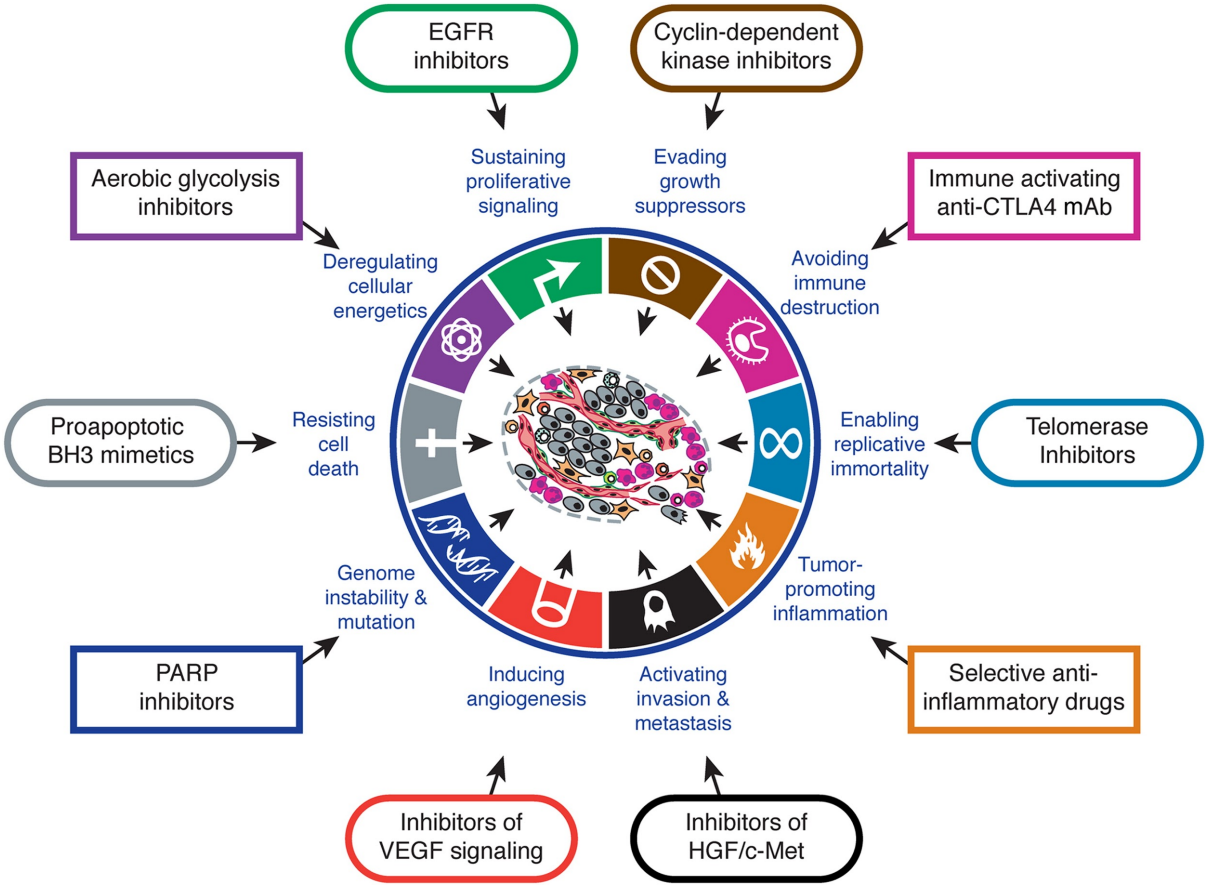


Figure 1.8: Hallmarks of cancer (Hanahan and Weinberg 2011). For each hallmark, an example treatment is shown in a coloured box that could be used to target it, often by inhibiting/promoting an upregulated/downregulated pathway or process, reversing the effects of mutations that cause that hallmark.

Patient tumours in the clinic are classified based on the scales of *stage* and *grade*, to determine the severity of an individual tumour, and aid in providing patient prognostic and treatment information. There are multiple different staging and grading systems that can be used based upon the tissue of origin of the cancer, such as the TNM staging system (Hortobagyi, Edge, and Giuliano 2018) and the Nottingham grading system (J. M. Chang et al. 2015) for breast cancer, which rely on information from clinical patient examinations and microscopic observation of tumour samples. *Stage* refers to how much a tumour has grown and spread, with a low stage, such as stage I, indicating that the tumour is small and well-contained within the tissue of origin, while a high stage, such as stage IV, indicates that the tumour has metastasised to distant parts of the body. In the TNM staging system the T metric represents the extent of tumour growth and invasion, the N metric describes the degree of tumour cell spread to regional lymph nodes and the M metric describes whether metastases are found elsewhere in the body. *Grade* refers to how differentiated tumour cells appear, with a low grade indicating high differentiation and a better patient prognosis and a high grade

indicating poorly differentiated tumour cells that are more stem-cell-like and do not resemble their tissue of origin, with a poorer prognosis. In the Nottingham grading system breast carcinomas are scored on a scale of one to three for the formation of tubules, irregular distribution and shape of cell nuclei, and the density of dividing cells, with scores being added up to give an overall score (Elston 1984, J. M. Chang et al. 2015).

Tumours are made up of large numbers of cells and have a high mutation rate due to their genomic instability that results in high genetic diversity between different populations of tumour cells, known as intratumour genetic *heterogeneity* (Meeks et al. 2020). Since tumour cells are descended from a single progenitor cell that gradually acquired mutations along with its descendants over time, the tumour cell population displays branching evolution, with tumour cells divided into groups based on their evolutionary branch. This is known as a *clonal* structure, with nearby cells within the same subclonal population sharing more mutations in common, while separate subclones show more mutational divergence. The dominant subclone with the most favourable mutations for division and survival typically grows faster and therefore the tumour becomes spatially centred around this subclone over time (McGranahan and Swanton 2017). As the number of cells in a tumour increases, known as *clonal expansion*, the number of new mutations that can occur scales with this, leading to greater genetic heterogeneity in an accelerating manner. This can lead to separate subclones evolving different sets of mutations that cause the development of the hallmarks of cancer in different ways. This results in tumours becoming more difficult to treat at later stages of cancer in two different ways: Firstly, late-stage tumours have acquired more driver mutations that protect tumour cells and enable them to resist treatment more aggressively. Secondly, the presence of genetically and functionally different tumour subclones means that treatments may not be able to target all tumour subclones effectively, preventing them from working long-term. This is commonly seen in cases where a cancer treatment will initially work becauses it targets the dominant subclone of a tumour, killing most of the tumour cells and leading to loss of disease symptoms in the patient, termed *remission*, but subsequently another subclone that resists the treatment becomes the new dominant subclone — a process termed *clonal selection* (**Figure 1.9**) — and continues to grow and divide, with the remaining cells becoming more resistant to that treatment and disease symptoms re-emerging, termed *relapse*. As a result, treatment for later stage cancers is often more focussed on delaying tumour progression and providing palliative care, since attempts to fully destroy all tumour cells are unlikely to succeed.

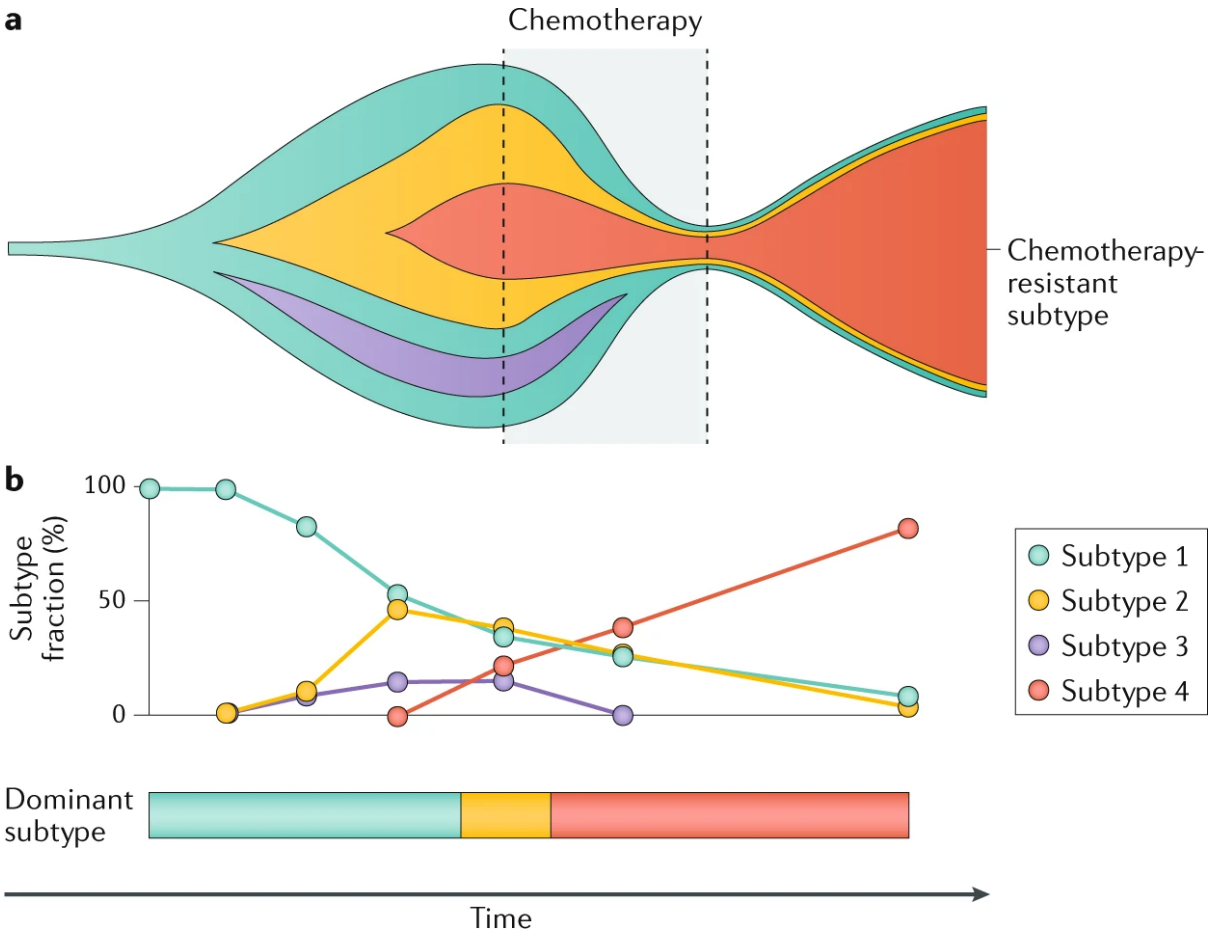


Figure 1.9: Tumour heterogeneity over time and clonal selection in response to treatment (Meeks et al. 2020). The number (a) and subclonal fraction (b) of tumour cells is colour-coded according to the subtype, with time along the x-axis. The number of subtypes increases initially as tumour cells acquire more mutations. The effect of chemotherapy is shown in the centre of the graph, where the number and heterogeneity of tumour cells reduces and the chemotherapy-resistant subtype makes up a greater fraction of the remaining cells, leading to relapse as the tumour becomes more resistant overall to chemotherapy.

Cancers can arise in almost any tissue, with different types of mutations occurring in each case and prognoses and treatments differing between these cancer subtypes. Tissues that are exposed to higher rates of mutational processes or chemical mutagens, such as skin exposed to UV light or lung epithelial cells in smokers, are usually affected by cancer at higher rates (Wu et al. 2018). Lung cancer and melanoma of the skin are respectively the second and fifth most common cancers in both males and females, reflecting their high mutational exposure (Siegel, Miller, and Jemal 2019), and typically exhibit far higher mutational loads than other types of tumours. Some rarer cancers, especially childhood cancers such as heritable retinoblastoma (AlAli et al. 2018), are caused primarily by germline mutations instead of somatic mutations. Germline mutations are inherited and present at birth rather than

acquired during the lifetime of an individual, explaining why cancers arising from these usually present at much earlier ages in the clinic.

Common immune mechanisms in cancer, infection and autoimmunity

Most types of cancer are not infectious. This is because cancer cells from one individual, if inserted into another, genetically-different individual, are not recognised by the immune system of the latter individual as their own cells and are rejected. The process by which the immune system recognises non-self cells is known as *allorecognition* and is mediated by short peptides on the surface of cells named *antigens* that are presented to immune cells by a protein complex named the *major histocompatibility complex* (MHC) (Kotsias, Cebrian, and Alloatti 2019). There are three classes of MHC (**Figure 1.10**), with MHC-I being present on the surface of all nucleated cells, including all cancer cell types, and they show great genetic diversity between individuals, who have specific versions of each MHC class, termed *human leukocyte antigen* (HLA) types. Each different HLA-type has a different antigen-binding profile, although some are more similar to each other than others. Cells without nuclei, such as red blood cells, cannot form cancers because they are unable to divide. Cytosolic proteins in the cell are gradually degraded by the proteasome and this gives rise to smaller peptide fragments, which are actively loaded into the endoplasmic reticulum through TAP channel protein complexes. Depending on their size — a length of 8-11 amino acids is optimal — and sequence, some of these peptides are able to bind to MHC-I within the endoplasmic reticulum, in which case they are termed *antigens*. The antigen/MHC-I complexes are shuttled to the surface of the cell where the antigens are presented to immune cells named cytotoxic T lymphocytes (CD8+ T-cells).

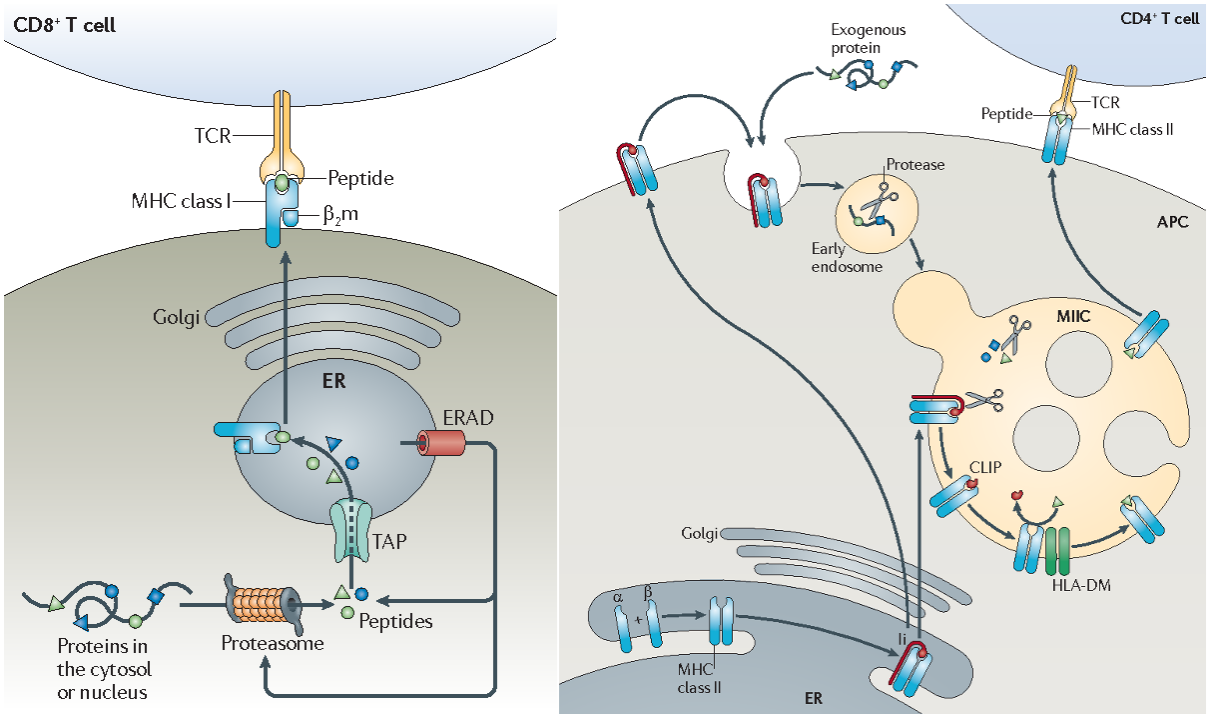


Figure 1.10: Antigen presentation pathways for MHC-I and II. Figure and caption adapted from (Neefjes et al. 2011). In both pathways proteins are degraded in the cell, leading to small peptide fragments that bind to the MHC-I (left) and MHC-II (right) transmembrane proteins, which are shuttled to the surface of the cell to present them to immune cells. Presentation of non-self peptides by MHC-I enables the direct destruction of non-self, tumour and virus-infected cells by CD8+, while MHC-II regulates the broader immune system response and memory via CD4+. **Left:** Antigens are degraded by the proteasome and transporter-associated-with-antigen-presentation (TAP) translocates the resulting peptides into the endoplasmic reticulum (ER) lumen for loading onto MHC-I. Peptide–MHC-I complexes are released from the ER and transported via the Golgi to the plasma membrane for antigen presentation to CD8+ T cells. **Right:** α- and β-chains assemble in the ER together with the invariant chain (Ii), forming the MHC-II receptor, which is transported through the Golgi to the MHC-II compartment (MIIC), directly and/or via the plasma membrane. Endocytosed proteins and Ii are degraded by resident proteases in the MIIC. The class II-associated Ii peptide (CLIP) fragment of Ii remains in the peptide-binding groove of the MHC-II dimer and is exchanged for an antigenic peptide with the help of the dedicated chaperone HLA-DM. MHC-II molecules are then transported to the plasma membrane to present antigenic peptides to CD4+ T cells.

Immature CD8+ T-cells are exposed to a wide variety of self-antigens in the thymus, and any CD8+ T-cells that bind to self-antigens in the thymus during maturation undergo negative selection and apoptosis, so only CD8+ T-cells that are not self-reactive enter the bloodstream (Cohn 2015). CD8+ T-cells have *T-cell receptors* (TCRs) on their surface that bind to foreign antigen/MHC-I complexes presented to them on the *antigen-presenting cell* (APC). Foreign antigens may occur if a cell has been infected by a virus, or has acquired

mutations that lead to novel protein sequences being present that bind MHC-I and do not resemble any self-proteins, termed *neoantigens*. Combined with co-stimulation via other molecules on the surface of the APC and CD8+ T-cell, this triggers the CD8+ T-cell to fully mature and to release cytotoxins into the cytoplasm of the APC, causing a caspase cascade that leads to apoptosis in the APC, thus destroying virus-infected cells and tumour cells, to protect against viral infection and cancer respectively (Rudd-Schmidt et al. 2019). Some viruses, such as HIV-1, are able to evade immune detection by interfering with this process, although this may depend on the HLA-type of the human host (Pymm et al. 2017). Antigen presentation also causes organ rejection when patients receive transplants from a donor who is not a close genetic match, unless the immune system is suppressed artificially. If errors occur in CD8+ negative selection in the thymus and self-reactive CD8+ cells do not undergo apoptosis, then this can lead to CD8+ cells attacking healthy host cells, causing damage to tissues, a process known as *autoimmunity*. For example, the autoimmune diseases Behçet's disease (Leccese and Alpsoy 2019; Giza et al. 2018), ankylosing spondylitis (Lorente et al. 2019), and HLA-C*0602-associated skin psoriasis (Schön 2019) are all associated with MHC-I dysfunction associated with specific HLA types of MHC-I.

MHC-II also presents antigens to immune cells, but targets helper T lymphocytes (CD4+) instead of CD8+ cells, and is only present on the surface of specific, antigen-presenting immune cells: Dendritic cells, B lymphocytes, some endothelial cells, mononuclear phagocytes, and thymus epithelial cells (Unanue, Turk, and Neefjes 2016). These cells acquire non-self antigens through phagocytosis or endocytosis of pathogens directly, or via other immune cells, and provide an immunological memory of these antigens by triggering the CD4+ cells to mature into a range of T cell types, that coordinate the immune response to pathogenic antigens or immune tolerance of antigens from a benign source. MHC-III proteins by contrast are not involved in antigen presentation, but include a range of different immune-signaling proteins (Yau, Tuncel, and Holmdahl 2017).

Because of the genetic diversity between humans, allorecognition prevents human cancers from being infectious, with the exception of transmission between identical twins, who are genetically identical. However, even in this case transmission generally does not occur because tumour cells are contained within the body and are unlikely to enter other people. Most identified cases of cancer transmission between identical twins are leukaemias transmitted during gestation via a shared placenta, reflecting the increased mobility of white blood cells compared to other tissues and the physical bridge for transmission between the foetuses (Greaves and Hughes 2018), with the earliest known case recorded in Germany in 1882 (Senator 1882). Transmissible cancers are more common in some other mammals

however. Tasmanian devil populations are genetically very close, and as such they are able to transmit *devil facial tumour disease* (DFTD) through biting, which is a common feature of Tasmanian devil behaviour (Peel and Belov 2018). Although not directly transmissible, some human cancers are indirectly transmissible through infectious pathogens that predispose those individuals to cancer, such as human papillomavirus (HPV), which is sexually transmissible and commonly causes cervical cancer (Araldi et al. 2018).

Tumour and neoantigen burden

Apart from preventing infection from foreign cancer cells and pathogens, the MHC-I antigen presentation pathway is capable of destroying cancer cells originating in the host itself. As tumour cells acquire more mutations in protein-coding regions during cancer progression, there is an increased likelihood that one of these novel mutations will lead to a novel, non-self peptide sequence that is able to bind to MHC-I as an antigen, and which is also recognised as non-self by a CD8+ cell. Novel tumour antigens that trigger an immune response in this way are termed *neoantigens*, and the existence of these within tumours has a large effect on tumour progression, patient prognoses and the efficacy of some cancer treatments (Schumacher, Scheper, and Kvistborg 2019).

Cancer cells typically need to acquire large numbers of specific driver mutations in order for tumour progression to occur, but they are unable to control which mutations they gain over time, so they must acquire very large numbers of mutations in general through mutations that disable the DNA replication and DNA-damage repair pathways and promote genomic instability. Due to this genomic instability and/or mutagen exposure, tumours often also include passenger mutations that reduce the viability of tumour cells and hinder tumour progression (McFarland et al. 2017). If the combined survival benefits of pro-cancer, oncogenic mutations in a tumour cell are greater than the effect of the anti-cancer mutations, then that cell will still have a survival advantage and may divide to become the dominant subclone in a tumour, with most of the tumour cells in that patient descended from that cell and therefore also carrying the anti-cancer mutations. If the number of anti-cancer mutations is particularly high in every tumour subclone, then this can eventually lead to the tumour growing more slowly or even stalling completely, despite it acquiring many of the hallmarks of cancer, due to poor tumour cell viability and little evolutionary means of selecting against the anti-cancer mutations, since all the tumour cells have them. Because of this, the total number of clonal mutations in a tumour, referred to as the *tumour mutation burden* (TMB), is — perhaps counterintuitively — correlated with improved survival in some cancer types, especially when controlling for the stage and grade of a tumour. Cancer types with very high

mutagen exposure in particular, such as lung cancers and skin melanoma are most likely to show this correlation, while cancer types that are characterised by higher genomic stability, such as acute myeloid leukaemia (AML), are more likely to show no correlation or a negative correlation between TMB and patient survival (Chalmers et al. 2017).

Subsets of cancer passenger mutations occur within protein-coding regions, and result in the expression of neoantigenic peptides. In a similar manner therefore, the tumour *neoantigen burden* can be defined as the total number of clonal neoantigenic mutations that occur within a tumour, which is a subset of the TMB, and has a similar correlation with patient survival. For this definition, neoantigenic mutations are usually only included within the neoantigen burden score if their corresponding genes show signs of expression within the tumour (Rosenthal et al. 2019; Miller et al. 2017; Wood et al. 2020), since they would not be expected to have a functional effect on tumour cell viability if their corresponding neoantigenic peptide were not actually present within the tumour cells.

TMB and neoantigen burden can both be used to classify cancer patients into separate groups with separate survival outcomes in many cancer types, enabling their use as genomic prognostic tools. In order for these tools to produce an accurate result, it is important that variant calls which would contribute to TMB or cause antigenic changes in peptides are accurate, so good quality control is essential. This can be difficult in tumour samples, where mutations might only show up in a small proportion of reads if the number of affected tumour cells is low and contamination from healthy cells is high, since it is difficult to distinguish between these and sequencing artefacts. However, these tools are less sensitive to inaccurate variant calls than genomic prognostic tools which rely on much smaller, specific panels of mutations or genes, in which a single error is more likely to affect the outcome of the prognostic tool (McGranahan et al. 2016; Kuo et al. 2017).

Understanding for how long an individual patient is likely to survive with cancer, or whether they have a chance of completely removing the tumour, allows doctors and patients to plan their therapy with more care for the patient, deciding whether to use curative or palliative treatments, calculating the likely costs of treatment and helping patients to make any final plans (Simmons et al. 2017). In addition, prognostic tools can be used to compare survival odds between patient cohorts receiving different treatments, in order to determine how beneficial different options are so that doctors can select the optimal treatment for each separate patient (Ou, Nagasaka, and Zhu 2018). Incorporating genomic information such as TMB and neoantigen burden into these prognostic tools further increases their accuracy and allows personalised, targeted therapy (Wirth and Kühnel 2017). This ensures that patients,

doctors and health bodies such as the NHS — which control what treatments are available — are able to make the best decisions.

Cancer therapies targeting the immune system

There are many different types of treatments available or in development for cancer, with various advantages and disadvantages in how they are targeted and applied. Some of these, termed *immunotherapies*, work by strengthening the immune system's anti-cancer response (Pan et al. 2020). There are five classes of immunotherapies, with some overlap between them (Billan, Kaidar-Person, and Gil 2020; van den Bulk, Verdegaal, and de Miranda 2018): Cell-based immunotherapies, which provide additional, modified immune cells to target to the tumour, such as chimeric antigen receptor T (CAR-T) cells (Shah and Fry 2019); immunomodulators, such as cytokines and checkpoint inhibitors including ipilimumab (Reck, Borghaei, and O'Byrne 2019), which decrease restraints on the immune response; antibody-based targeted therapies, such as rituximab (Focosi, Tuccori, and Maggi 2019), which target tumour cells via cancer-specific antigens; oncolytic viruses, such as talimogene laherparepvec (T-Vec) for advanced melanoma (Bommareddy et al. 2017), which preferentially infect and destroy tumour cells, as well as triggering anti-cancer immune recognition and activity; and anti-cancer vaccines, such as Sipuleucel-T for prostate cancer patients (Madan et al. 2020), which provide an immune memory of tumour cells so that they are more easily recognised and destroyed by the immune system. These are all examples of targeted cancer treatments, and display fewer unpleasant side effects for patients due to inflicting lower levels of damage to healthy cells and tissues across the body.

Other, general cancer treatments include therapies such as surgery to remove a tumour, which is typically performed whenever it is possible and safe to do so alongside other treatments; radiotherapy, which involves using radiation to kill cells on the premise that tumour cells are more susceptible to DNA damage; and non-targeted chemotherapy, where chemicals that are damaging to cancer cells, but also to normal cells to a lesser extent, are injected or ingested. In addition, targeted chemotherapy drugs also exist that specifically target cancer cells and cause minimal harm to non-cancer cells. For example, treatments can be targeted to only affect tumour cells with specific mutations or proteins not shared by healthy cells, such as trastuzumab for HER2-positive oesophageal and breast cancer (Mao et al. 2021), or they can be targeted at tumours via their delivery method, using targeted drug delivery systems (TDDSs) such as tumour-specific antibodies (Qin, Zhang, and Zhang 2018). Some cancer types have specific forms of treatment based on their disease biology. This includes cancers which have their growth regulated by hormones, such as prostate cancer and breast cancer,

which are often treated with hormone therapy alongside other treatments (Teo, Rathkopf, and Kantoff 2019), and some blood cancers which can be treated by destroying all of the patient's blood cells of that type with chemotherapy or radiotherapy and replacing them using stem cell therapy, such as in acute myeloid leukaemia (Kassim and Savani 2017).

For many cancer types there are multiple types of general and targeted treatments that are approved for use, and it can be difficult for clinicians to determine what treatment combinations, if any, would be optimal for a specific patient, so powerful genomic prognostic tools are valuable in this regard. Patient health status, cancer type, stage, grade, driver mutation profile, TMB and neoantigen burden are all important factors for selecting treatments, and understanding how the immune system interacts with cancer biology is of great importance. For example, many immunotherapies show advantages in treating later stage cancers that are not effectively treated by more traditional methods, but some of these, such as ipilimumab, are dependent on the patient exhibiting a high neoantigen burden (van den Bulk, Verdegaal, and de Miranda 2018; Reck, Borghaei, and O'Byrne 2019), since their mechanism of action relies on the tumour cells displaying neoantigens which allow targeted tumour destruction by immune cells, while oncolytic virus therapy does not require neoantigen presence (Bommareddy et al. 2017). At the same time, the vaccine Sipuleucel-T is effective against cancer types such as prostate cancer, in which the microenvironment around the tumour shows low-levels of inflammation that would usually hinder immune cell infiltration and reduce the efficacy of other immunotherapies (Madan et al. 2020). These immune features, which can be measured through combining tumour genomic sequencing with other clinical metrics, therefore offer a pathway to optimising cancer treatments and patient prognoses.

SARS-CoV-2 and genomic surveillance during the pandemic

SARS-CoV-2 biology

COVID-19 is a respiratory disease causing an ongoing pandemic which has resulted in hundreds of millions of cases globally, millions of deaths, trillions of US dollars lost from the global economy, severe suffering and unemployment, social, work and travel restrictions and damage to education and mental and physical health (Nicola et al. 2020; Siddik 2020), first reported in China in December 2019 (Zhou et al. 2020), and declared a global pandemic by the WHO on the 11th March 2020 (Cucinotta and Vanelli 2020). COVID-19 is commonly characterised by symptoms including a dry cough, fever, headache, nausea, fatigue and anosmia, and in severe cases disease progression can lead to hypoxemia, lung damage, cardiac failure, and less commonly injury to other organs such as the kidneys and liver, all of which can cause the death of the patient (Berlin, Gulick, and Martinez 2020). Age is the most important risk factor for disease progression leading to death (Williamson et al. 2020). Other risk factors include health conditions such as cardiovascular disease, obesity, immunosuppression, asthma and diabetes mellitus, along with pregnancy, being male, and belonging to a black or south Asian ethnic group. COVID-19 is caused by the betacoronavirus severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (Coronaviridae Study Group of the International Committee on Taxonomy of Viruses 2020), which is thought to have originated in bats, with an intermediate host, possibly pangolins sold at the Wuhan seafood wholesale market where the first cluster of cases was identified, responsible for the first transmission to humans (N. Chen et al. 2020). SARS-CoV-2 transmission has been primarily from human-to-human spread during the pandemic, although other mammals, such as mink, can act as a reservoir for the disease (Oude Munnink et al. 2021). Two similar bat-origin coronaviruses, severe acute respiratory syndrome (SARS) and Middle-East respiratory syndrome (MERS), have been recently transmitted to humans from an intermediate host, suspected to be palm civets and dromedary camels respectively, causing outbreaks in 2002 and 2012, but were contained with fewer than 10,000 cases observed (Petrosillo et al. 2020).

SARS-CoV-2 has a 29.9kb single-stranded RNA-based genome encoding four canonical 3′ structural proteins — the spike (S), envelope (E), membrane (M) and nucleocapsid (N) proteins — as well as a 5′ frameshifted polyprotein (ORF1a/ORF1ab) found in all coronaviruses and other accessory proteins (Ashour et al. 2020). Individual virus particles are formed of the positive-stranded RNA genome bound within a helical nucleocapsid, primarily formed of the N protein. This is coated in a spherical lipid bilayer membrane containing the E and M proteins, with the spike glycoproteins protruding from this (Acter et al.

2020). The virus infects cells by using its spike proteins to bind to a host transmembrane protein named angiotensin-converting enzyme 2 (ACE2) which is most abundant in type II alveolar cells in the lungs, causing the virus to enter the cell and the uncoating of the viral RNA (**Figure 1.11**). The viral RNA is translated by the host ribosome to form the long polyprotein ORF1a/ORF1ab, which is cleaved into nonstructural proteins that combine into a replicase-transcriptase complex, named RNA-dependent RNA polymerase (RdRp). This replicates the entire RNA genome into additional positive-stranded RNAs, and also transcribes smaller, subgenomic RNA transcripts which are translated into the structural and accessory viral proteins in the host endoplasmic reticulum (ER). The structural proteins are embedded in the ER membrane, assemble together and transit through the ER-to-Golgi intermediate compartment (ERGIC), where they combine with the N-encapsidated viral RNA copies, resulting in membrane budding to form additional, fully-assembled virus particles named virions. The virions leave the cell via exocytosis to repeat the viral infection cycle (V'kovski et al. 2021).

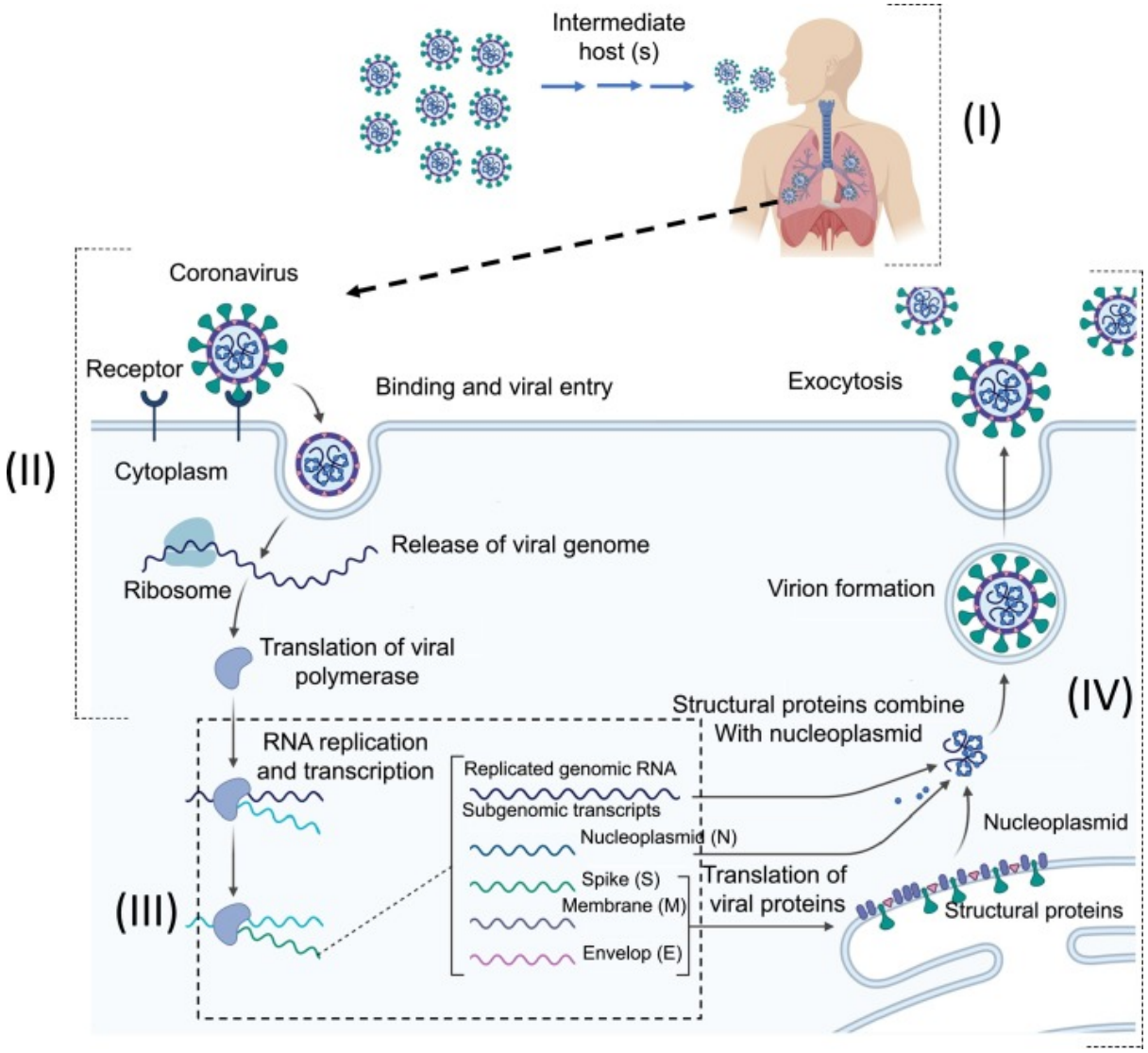


Figure 1.11: SARS-CoV-2 viral infection cycle (Acter et al. 2020). After viral transmission, SARS-CoV-2 binds to receptors on the surface of cells (primarily ACE2 on type II alveolar cells in the lungs) using its spike protein. This leads to viral entry and translation of the viral RdRp polymerase protein by the host ribosomes. RdRp transcribes the remaining subgenomic viral transcripts which are translated into the viral structural proteins, as well as replicating the viral RNA. The structural proteins assemble around the viral RNA to form completed virions, which exit the cell by exocytosis to repeat the viral infection cycle.

Importance of SARS-CoV-2 sequencing

Accurate genomic sequencing of viruses is important for identifying them, tracking chains of transmission and emergence of new mutations (Korber et al. 2020), understanding their underlying biology, correctly diagnosing disease in the clinic, and developing vaccines and treatments (Michel et al. 2020). All of these reasons apply to SARS-CoV-2, as they have historically for viral epidemics such as HIV (Bbosa, Kaleebu, and Ssemwanga 2019) and

Ebola (Holmes et al. 2016), which have been heavily sequenced. Tracking the transmission of viruses enables scientists to identify how different cases are genetically related (**Figure 1.12**), and can identify effective ways for public health policy to minimise transmission and prevent the introduction of specific viral lineages into a country, and identify when a breach has occurred. For example, identification of the increased transmission rates of the B.1.1.7 lineage of SARS-CoV-2 in the Kent region of the UK informed international decisions about locking down international travel from the UK from 20th December 2020 (Michaels and Douglas 2020), and the introduction of tighter "tier four" coronavirus restrictions in London and the South-East of England from 19th December 2020. As of 31st March 2021, there have been five SARS-CoV-2 lineages of concern (Tegally et al. 2021; Galloway et al. 2021; Faria et al. 2021; Zhang et al. 2021; Rambaut Group 2021) — visible at https://www.gisaid.org/hcov19-variants/ — which have exhibited large recent increases in population frequency, in association with the possible functional effects of their key spike protein amino acid changes: N501Y, E484K and L452R. Sequencing SARS-CoV-2 samples from hospitals also allows clinicians to determine whether these are closely related — suggesting a possible hospital source and a need to test staff — or from distinct sources, in which case the infections likely occurred in the community outside of the hospital setting (Stirrup et al. 2020; Løvestad et al. 2021).

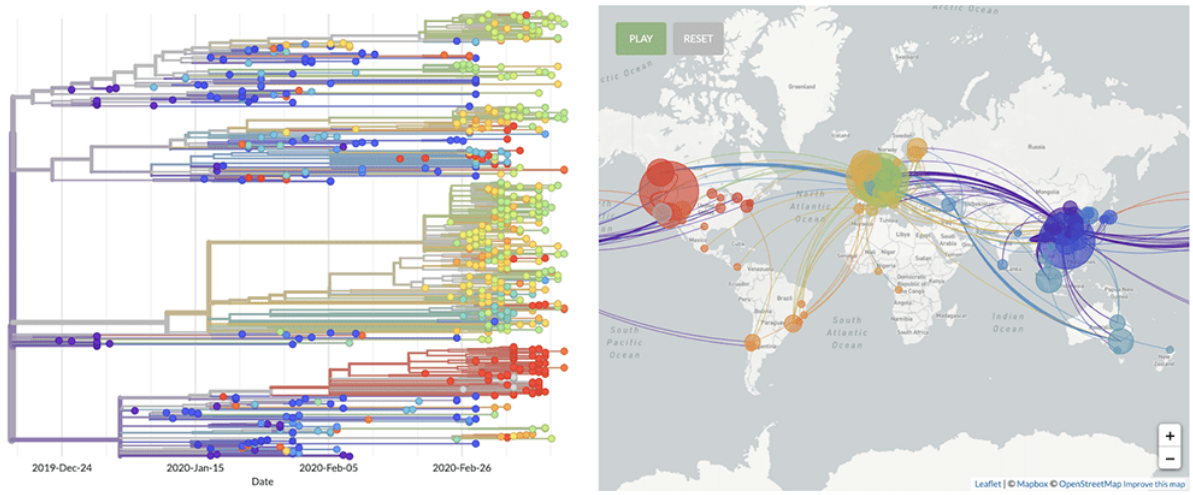


Figure 1.12: Tracking the transmission of the SARS-CoV-2 pandemic using NextStrain GISAID sequencing data (UKRI 2020). Left: SARS-CoV-2 phylogeny shows when new mutations arise, and how these are transmitted to other samples, with branching showing ancestral relations between samples and the x-axis showing time from first emergence of SARS-CoV-2. Right: Worldwide spread of different pandemic lineages colour-coded.

Using sequencing to study the underlying biology of viruses allows scientists to identify the function of genes and proteins and develop effective treatments targeting these, as well

as identifying how a new mutation might affect treatment. For example, the D614G amino acid change in the SARS-CoV-2 spike protein can be detected using sequencing, and structural biochemistry analyses were used to show that this increased infectivity via the host ACE2 protein across a range of mammals (Yurkovetskiy et al. 2020). Furthermore, the D614G amino acid change was found to occur outside of the parts of the spike protein that act as antigens for the antiviral immune response, suggesting that this was unlikely to interfere with vaccine efficacy against these antigens (Groves, Rowland-Jones, and Angyal 2021). Sequencing has also revealed SARS-CoV-2 viral evolution in individual patients as a response to treatment, showing possible routes to the virus acquiring resistance to these treatments. For example, immunocompromised, chronically-infected individuals treated with convalescent antibodies against SARS-CoV-2 have been sequenced at multiple time points to check how the viral sequence diversity changes over time and in response to treatment, revealing mutations that become dominant under the selective pressure imposed by the treatment, reducing the ability of the antibodies to bind and neutralise virus particles (Kemp et al. 2021). This suggests that treating immunocompromised individuals with convalescent therapy carries a risk of introducing antibody-resistant SARS-CoV-2 lineages into the population, so should be undertaken with caution or avoided altogether, since these could potentially re-infect recovered SARS-CoV-2 patients.

In order to carry out large-scale SARS-CoV-2 sequencing and analysis, national consortiums of sequencing centres have been set up, along with repositories for the resulting viral sequences, allowing free access for researchers. This includes the COVID-19 Genomics UK (COG-UK) consortium, which was launched in March 2020 initially to sequence SARS-CoV-2 samples from up to 230,000 patients, health-care workers, and other essential workers in the UK (COVID-19 Genomics UK (COG-UK) consortiumcontact@cogconsortium.uk 2020) — a figure which it has since exceeded. COG-UK deposits the sequences it generates into multiple public online repositories, including the Global Initiative for Sharing All Influenza Data repository (GISAID) (Elbe and Buckland-Merrett 2017; Shu and McCauley 2017) which contains over 900,000 SARS-CoV-2 viral sequences (accessed 31/03/2021), and NextStrain (Hadfield et al. 2018). COG-UK has used a range of sequencing technologies at different sequencing centres, but most sequencing has been done using the ARTIC protocol (Loman, Rowe, and Rambaut 2020) to prepare SARS-CoV-2 samples for sequencing, followed by either Illumina (Charre et al. 2020) or ONT (Tyson et al. 2020) sequencing pipelines.

The ARTIC protocol is a combined laboratory and bioinformatics protocol for SARS-CoV-2 sample preparation and sequencing with ONT, developed by the ARTIC network. It consists of a range of steps. Firstly, cDNA is prepared by reverse transcription of the viral RNA

template, followed by preparation of a specific pool of primer pairs at 10μM concentration, which are used for PCR amplification of 400bp amplicons of the SARS-CoV-2 sequence. DNA quantification steps are then used to normalise the concentration of SARS-CoV-2 cDNA output from PCR amplification to a fixed standard, after which samples are barcoded and sequenced with an ONT MinION machine, using the MinION's integrated sequencing bioinformatics. The output sequencing data from the ARTIC protocol require basecalling and variant calling separate from the ARTIC protocol steps. Multiple different bioinformatics tools can be used for this in combination with the ARTIC protocol, including those benchmarked in chapter IV of this thesis.

SARS-CoV-2 sequencing pipelines and their challenges

As described earlier in this chapter, Illumina and ONT have various advantages and disadvantages. Illumina sequencing is inexpensive and has a low basecalling error, but is more prone to effects such as GC-bias (Benjamini and Speed 2012), while ONT sequencing has low upfront costs to purchase basic sequencing machines, a simple sample preparation process and real-time sequencing, but suffers from a higher basecalling error rate, especially at homopolymers (Petersen et al. 2019). ONT sequencing is also capable of using much longer reads than Illumina, but the ARTIC protocol generates amplicons that are only 400bp long, so this is not an important advantage for ONT in this scenario (Tyson et al. 2020). The SARS-CoV-2 genome does not contain large repetitive sequences, so mapping reads to the reference genome is straightforward and longer read lengths are not required, and using 400bp reads decreases the ability of degraded reads to reduce sequencing accuracy.

SARS-CoV-2 was discovered less than two years ago, and neither the Illumina nor ONT sequencing technologies were originally developed with a specific focus on SARS-CoV-2 sequencing. It is therefore likely that SARS-CoV-2 sequencing methods are not fully optimised yet, and further sequencing inaccuracies will be found over time. Especially earlier in the pandemic, there were higher levels of false positive variant calls that went undiscovered (De Maio et al. 2020a). High-accuracy SARS-CoV-2 sequencing is essential for accurately tracking viral transmission and new mutations of concern, among other important uses listed in this section. Quality control measures to help ensure sequencing is highly accurate include blacklists flagging up positions with poor sequencing accuracy, or false positive variant calls (Bull et al. 2020; De Maio et al. 2020b). For example, Bull et al. have blacklisted regions with homopolymeric or repetitive content at which ONT and Illumina sequencing found different consensus sequences. De Maio et al. have blacklisted many other genomic variants, including poor quality variants found at few sequencing centres, variants showing high *homoplasy* — i.e. that appeared to emerge *de novo* at high rates inconsistent with standard mutation rates,

suggesting that they were sequencing artefacts — as well as variants removed from sequences after recalling, variants confirmed as false positives arising from contamination from other species or nanopore adapter sequences, and variants linked with any other blacklisted variants.

Thesis outlook

Exponential advances in next-generation sequencing (NGS) technologies in the last two decades have made sequencing considerably cheaper and more accessible, and will continue to do so in the coming years. As a result, genomic sequencing on an individual basis has recently become affordable and is beginning to be implemented into patient treatment. However, the computational tools to analyse this sequencing data and obtain useful information for clinical treatment of patients are lagging behind this, and significant work will be required in the coming decades to take full advantage of the sequencing data available and translate this into improvements in individual treatment (Muir et al. 2016).

My PhD research revolves around designing and improving computational methods used in genome variant analysis to improve the diagnosis of cancer and infectious disease. I therefore had two broad objectives, into which most of my research fell:

1) The usage of large-scale genomic sequencing data to identify sources of genetic diversity that have important consequences in medicine and/or epidemiology. E.g. Mutational markers of survival odds from tumours in cancer patients — Chapter II — or the presence of novel variants in SARS-CoV-2 genomes that could lead to viral fitness differences — Chapter IV.

2) Improving the accuracy of genomic variant detection — Chapters III and IV. This in turn also aids the former objective by means of improved accuracy of the input data.

All of the research in this publication format thesis links back to the objectives above, with an overall aim to design and improve genomic diagnostics using computational methods. In this chapter I have presented a general background to my research. The following three chapters, with associated publications of which I am a co-author, demonstrate how I have made advances in this field and show their research impact:

Chapter II: This chapter describes my work developing a pipeline to classify non-small-cell lung cancer (NSCLC) patients into groups based upon the burden of somatic mutations and immunogenic peptides present within their tumours. TMB has previously been suggested as a surrogate measure for predicting neoantigenicity in the clinic (Klempner et al. 2020; Chan et al. 2019; Hendriks, Rouleau, and Besse 2018), since neoantigen burden is not independent of TMB and it is not clear to what extent using both measures provides additional value to genomic prognostics. In this chapter, I present the publication "*Somatic Alteration Burden*

Involving Non-Cancer Genes Predicts Prognosis in Early-Stage Non-Small Cell Lung Cancer" (Wang et al. 2019), which includes the key findings of my work developing and combining classifiers of both mutation and neoantigen burden to achieve stronger estimation of patient survival odds than using either measure on its own. This includes the identification of a subgroup of NSCLC patients that are 85% less likely to die over a 5-year period than other NSCLC patients, as well as a corresponding high-risk subgroup, greatly outperforming previously-published neoantigen-based genomic prognostic classifiers for NSCLC and other cancer types which do not achieve survival differences of this magnitude between groups (Rosenthal et al. 2019; Miller et al. 2017; Wood et al. 2020). This publication also details and explains the methods I used to categorise these patients so that similar methods could be applied to other cancers. This would allow potential therapeutic benefits to patients, helping them to plan their treatment with clinicians on a personalised level and improve scientists' understanding of their cancers and the associated mutational and immune effects present in individual tumours.

Chapter III: The success of my research in the area of NSCLC relied upon my access to accurate genomic sequencing data, without false positive genomic variants. After reflecting on this, I sought to identify and address gaps among the current quality control methods for identifying false positive variants so that these could be removed from NGS data more effectively, to improve the accuracy of genomic sequencing outputs. In this chapter I present my publication "*Genomic loci susceptible to systematic sequencing bias in clinical whole genomes*" (Freeman et al. 2020). In this publication, I identify systematic bias in genomic sequencing and alignment as an important source of error in whole genome sequencing using the Illumina sequencing platform, describe a novel method for identifying and removing variants prone to systematic bias, and share a catalogue of suspect positions across the human genome that are prone to systematic biases with a variety of Illumina-based pipelines. I explain the background behind this method, which models the expected standard deviation of allelic fractions across patients in a cohort sequenced using the same pipeline, and compares this with the actual observed allelic fraction values, to identify loci that show a consistent sequencing bias towards a specific allele catalogued across the cohort. Furthermore, I show which areas of the genome are more prone to systematic bias and investigate possible causes for this.

Chapter IV: In this chapter I adapt the methods I developed in the previous chapter in order to apply these to 884 SARS-CoV-2 genome sequences from various different ONT sequencing pipelines. It is likely that there are SARS-CoV-2 sequencing inaccuracies that have not been characterised in previously-generated blacklists (Bull et al. 2020; De Maio et

al. 2020b), especially those arising from systematic bias, since no general measures existed to identify systematic biases and minimise their effect before my publication in chapter III. I used a modified version of the approach I invented in chapter III to carry out quality control of variants for my third publication *"Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus"* (Korber et al. 2020). By creating an Incremental Database (IncDB) of allelic fractions across the cohort of viruses, I discovered that one of the variants which was of great interest to our research group (S943P in the spike protein) was in fact a systematic error, and tracked the source of this error back to the ONT adapter sequence used for sequencing. This variant had been observed in much of the SARS-CoV-2 sequencing data at the time and was thought to potentially be increasing in frequency in the population due to fitness advantages over the reference SARS-CoV-2 virus. My research confirmed that its supposed prevalence was in fact due to systematic bias causing its widespread false-positive detection, rather than any real sequence variation, and resulted in its exclusion from the list of variants of concern published in *Cell*. In addition to the work relating to this publication, I also carried out sequencing benchmarking tests to provide basecalling and variant calling recommendations for an optimal ONT-based sequencing pipeline, along with blacklists of variants I have shown to be prone to higher sequencing error rates, which should be treated with caution. By providing these recommendations and resources for highly-accurate SARS-CoV-2 sequencing using ONT with the ARTIC protocol, I aimed to address these gaps in sequencing performance and avoid sequencing errors that could have large impacts on the global efforts against the COVID-19 pandemic.

Chapter V. Conclusion: The body of work in the previous chapters demonstrates the value of accurate genomic sequencing data to human medicine and epidemiology, with novel impacts in the areas of cancer and SARS-CoV-2 sequencing, and describes the creation of new tools to improve the accuracy of genomic sequencing data, with examples of how we have applied these to advance research. In this section I link back to the previous chapters to demonstrate how they fulfil the aims of this PhD thesis, explaining how the outputs of each chapter complement each other and exploring remaining unmet needs and potential future directions for this body of research. I finish by concluding the impact on science and medicine of the results found in this PhD thesis.

References

Acter, Thamina, Nizam Uddin, Jagotamoy Das, Afroza Akhter, Tasrina Rabia Choudhury, and Sunghwan Kim. 2020. “Evolution of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) as Coronavirus Disease 2019 (COVID-19) Pandemic: A Global Health Emergency.” *The Science of the Total Environment* 730 (August): 138996.

AlAli, Alaa, Stephanie Kletke, Brenda Gallie, and Wai-Ching Lam. 2018. “Retinoblastoma for Pediatric Ophthalmologists.” *Asia-Pacific Journal of Ophthalmology (Philadelphia, Pa.)* 7 (3): 160–68.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology*. https://doi.org/10.1016/s0022-2836(05)80360-2.

Amarasinghe, Shanika L., Shian Su, Xueyi Dong, Luke Zappia, Matthew E. Ritchie, and Quentin Gouil. 2020. “Opportunities and Challenges in Long-Read Sequencing Data Analysis.” *Genome Biology* 21 (1): 30.

Araldi, Rodrigo Pinheiro, Thalita Araujo Sant’Ana, Diego Grando Módolo, Thatiana Correa de Melo, Diva Denelle Spadacci-Morena, Rita de Cassia Stocco, Janete Maria Cerutti, and Edislane Barreiros de Souza. 2018. “The Human Papillomavirus (HPV)-Related Cancer Biology: An Overview.” *Biomedicine & Pharmacotherapy = Biomedecine & Pharmacotherapie* 106 (October): 1537–56.

Ashour, Hossam M., Walid F. Elkhatib, Md Masudur Rahman, and Hatem A. Elshabrawy. 2020. “Insights into the Recent 2019 Novel Coronavirus (SARS-CoV-2) in Light of Past Human Coronavirus Outbreaks.” *Pathogens* 9 (3). https://doi.org/10.3390/pathogens9030186.

Bbosa, Nicholas, Pontiano Kaleebu, and Deogratius Ssemwanga. 2019. “HIV Subtype Diversity Worldwide.” *Current Opinion in HIV and AIDS* 14 (3): 153–60.

Benjamini, Yuval, and Terence P. Speed. 2012. “Summarizing and Correcting the GC Content Bias in High-Throughput Sequencing.” *Nucleic Acids Research* 40 (10): e72.

Berlin, David A., Roy M. Gulick, and Fernando J. Martinez. 2020. “Severe Covid-19.” *The New England Journal of Medicine* 383 (25): 2451–60.

Billan, Salem, Orit Kaidar-Person, and Ziv Gil. 2020. “Treatment after Progression in the Era of Immunotherapy.” *The Lancet Oncology* 21 (10): e463–76.

Bohannan, Zachary S., and Antonina Mitrofanova. 2019. “Calling Variants in the Clinic: Informed Variant Calling Decisions Based on Biological, Clinical, and Laboratory Variables.” *Computational and Structural Biotechnology Journal* 17 (April): 561–69.

Bommareddy, Praveen K., Anand Patel, Saamia Hossain, and Howard L. Kaufman. 2017. “Talimogene Laherparepvec (T-VEC) and Other Oncolytic Viruses for the Treatment of Melanoma.” *American Journal of Clinical Dermatology* 18 (1): 1–15.

Bulk, Jitske van den, Els Me Verdegaal, and Noel Fcc de Miranda. 2018. “Cancer Immunotherapy: Broadening the Scope of Targetable Tumours.” *Open Biology* 8 (6). https://doi.org/10.1098/rsob.180037.

Bull, Rowena A., Thiruni N. Adikari, James M. Ferguson, Jillian M. Hammond, Igor Stevanovski, Alicia G. Beukers, Zin Naing, et al. 2020. “Analytical Validity of Nanopore Sequencing for Rapid SARS-CoV-2 Genome Analysis.” *Nature Communications* 11 (1): 6272.

Burgess, Darren J. 2019. “Spatial Transcriptomics Coming of Age.” *Nature Reviews. Genetics*.

Chalmers, Zachary R., Caitlin F. Connelly, David Fabrizio, Laurie Gay, Siraj M. Ali, Riley Ennis, Alexa Schrock, et al. 2017. “Analysis of 100,000 Human Cancer Genomes Reveals the Landscape of Tumor Mutational Burden.” *Genome Medicine* 9 (1): 34.

Chambers, Daniel C., Alan M. Carew, Samuel W. Lukowski, and Joseph E. Powell. 2018. “Transcriptomics and Single-cell RNA-sequencing.” *Respirology* 24 (1): 29–36.

Chang, James M., Ann E. McCullough, Amylou C. Dueck, Heidi E. Kosiorek, Idris T. Ocal,

Thomas K. Lidner, Richard J. Gray, et al. 2015. “Back to Basics: Traditional Nottingham Grade Mitotic Counts Alone Are Significant in Predicting Survival in Invasive Breast Carcinoma.” *Annals of Surgical Oncology* 22 Suppl 3 (December): S509–15.

Chang, Michelle, Lin He, and Lei Cai. 2018. “An Overview of Genome-Wide Association Studies.” *Methods in Molecular Biology* 1754: 97–108.

Chan, T. A., M. Yarchoan, E. Jaffee, C. Swanton, S. A. Quezada, A. Stenzinger, and S. Peters. 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 / ESMO* 30 (1): 44–56.

Charre, Caroline, Christophe Ginevra, Marina Sabatier, Hadrien Regue, Grégory Destras, Solenne Brun, Gwendolyne Burfin, et al. 2020. “Evaluation of NGS-Based Approaches for SARS-CoV-2 Whole Genome Characterisation.” *Virus Evolution* 6 (2): veaa075.

Chen, Kunling, Yanpeng Wang, Rui Zhang, Huawei Zhang, and Caixia Gao. 2019. “CRISPR/Cas Genome Editing and Precision Plant Breeding in Agriculture.” *Annual Review of Plant Biology* 70 (April): 667–97.

Chen, Nanshan, Min Zhou, Xuan Dong, Jieming Qu, Fengyun Gong, Yang Han, Yang Qiu, et al. 2020. “Epidemiological and Clinical Characteristics of 99 Cases of 2019 Novel Coronavirus Pneumonia in Wuhan, China: A Descriptive Study.” *The Lancet*. https://doi.org/10.1016/s0140-6736(20)30211-7.

Clark, David P., Nanette J. Pazdernik, and Michelle R. McGehee, eds. 2018. “DNA Sequencing.” In *Molecular Biology (3rd Edition)*, 253–55. Academic Cell.

Cohn, Melvin. 2015. “Rationalizing Thymic Selection for Functional T-Cells: A Commentary.” *Cellular Immunology* 298 (1-2): 83–87.

Collins, Francis S., Eric D. Green, Alan E. Guttmacher, and Mark S. Guyer. 2003. “A Vision for the Future of Genomics Research.” *Nature* 422 (6934): 835–47.

Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. 2020. “The Species Severe Acute Respiratory Syndrome-Related Coronavirus: Classifying 2019-nCoV and Naming It SARS-CoV-2.” *Nature Microbiology* 5 (4): 536–44.

Costello, Maura, Trevor J. Pugh, Timothy J. Fennell, Chip Stewart, Lee Lichtenstein, James C. Meldrim, Jennifer L. Fostel, et al. 2013. “Discovery and Characterization of Artifactual Mutations in Deep Coverage Targeted Capture Sequencing Data due to Oxidative DNA Damage during Sample Preparation.” *Nucleic Acids Research* 41 (6): e67.

COVID-19 Genomics UK (COG-UK) consortiumcontact@cogconsortium.uk. 2020. “An Integrated National Scale SARS-CoV-2 Genomic Surveillance Network.” *The Lancet. Microbe* 1 (3): e99–100.

Cucinotta, Domenico, and Maurizio Vanelli. 2020. “WHO Declares COVID-19 a Pandemic.” *Acta Bio-Medica: Atenei Parmensis* 91 (1): 157–60.

De Cario, Rosina, Ada Kura, Samuele Suraci, Alberto Magi, Andrea Volta, Rossella Marcucci, Anna Maria Gori, Guglielmina Pepe, Betti Giusti, and Elena Sticchi. 2020. “Sanger Validation of High-Throughput Sequencing in Genetic Diagnosis: Still the Best Practice?” *Frontiers in Genetics* 11 (December): 592588.

De Maio, Nicola, Conor Walker, Rui Borges, Lukas Weilguny, Greg Slodkowicz, and Nick Goldman. 2020a. “Issues with SARS-CoV-2 Sequencing Data.” *Virological.org*. https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473.

———. 2020b. “Masking Strategies for SARS-CoV-2 Alignments.” Virological. July 29, 2020. https://virological.org/t/masking-strategies-for-sars-cov-2-alignments/480.

DePristo, Mark A., Eric Banks, Ryan Poplin, Kiran V. Garimella, Jared R. Maguire, Christopher Hartl, Anthony A. Philippakis, et al. 2011. “A Framework for Variation Discovery and Genotyping Using next-Generation DNA Sequencing Data.” *Nature Genetics* 43 (5): 491–98.

Eisenstein, Michael. 2017. “An Ace in the Hole for DNA Sequencing.” *Nature* 550 (7675): 285–88.

Elbe, Stefan, and Gemma Buckland-Merrett. 2017. “Data, Disease and Diplomacy: GISAID’s Innovative Contribution to Global Health.” *Global Challenges (Hoboken, NJ)* 1 (1): 33–46.

Elston C. W. 1984. The assessment of histological differentiation in breast cancer. The Australian and New Zealand journal of surgery, 54(1), 11–15. https://doi.org/10.1111/j.1445-2197.1984.tb06677.x

Faria, Nuno R., Thomas A. Mellan, Charles Whittaker, Ingra M. Claro, Darlan da S. Candido, Swapnil Mishra, Myuki A. E. Crispim, et al. 2021. “Genomics and Epidemiology of a Novel SARS-CoV-2 Lineage in Manaus, Brazil.” *medRxiv : The Preprint Server for Health Sciences*, March. https://doi.org/10.1101/2021.02.26.21252554.

Fiers, W., R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W. Min Jou, et al. 1976. “Complete Nucleotide Sequence of Bacteriophage MS2 RNA: Primary and Secondary Structure of the Replicase Gene.” *Nature* 260 (5551): 500–507.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick. 1995. “Whole-Genome Random Sequencing and Assembly of Haemophilus Influenzae Rd.” *Science* 269 (5223): 496–512.

Focosi, Daniele, Marco Tuccori, and Fabrizio Maggi. 2019. “Progressive Multifocal Leukoencephalopathy and Anti-CD20 Monoclonal Antibodies: What Do We Know after 20 Years of Rituximab.” *Reviews in Medical Virology* 29 (6): e2077.

Freeman, Timothy M., Genomics England Research Consortium, Dennis Wang, and Jason Harris. 2020. “Genomic Loci Susceptible to Systematic Sequencing Bias in Clinical Whole Genomes.” *Genome Research* 30 (3): 415–26.

Galloway, Summer E., Prabasaj Paul, Duncan R. MacCannell, Michael A. Johansson, John T. Brooks, Adam MacNeil, Rachel B. Slayton, et al. 2021. “Emergence of SARS-CoV-2 B.1.1.7 Lineage - United States, December 29, 2020-January 12, 2021.” *MMWR. Morbidity and Mortality Weekly Report* 70 (3): 95–99.

Garrison, E., and G. Marth. 2012. “Haplotype-Based Variant Detection from Short-Read Sequencing.” *arXiv Preprint*. https://arxiv.org/pdf/1207.3907.pdf.

Genomics England Press Releases. 2018. “The UK Has Sequenced 100,000 Whole Genomes in the NHS,” December 5, 2018. https://www.genomicsengland.co.uk/the-uk-has-sequenced-100000-whole-genomes-in-the-nhs/.

Giza, M., D. Koftori, L. Chen, and P. Bowness. 2018. “Is Behçet’s Disease a ‘Class 1-Opathy’? The Role of HLA-B*51 in the Pathogenesis of Behçet’s Disease.” *Clinical and Experimental Immunology* 191 (1): 11–18.

Goffeau, A., B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, et al. 1996. “Life with 6000 Genes.” *Science* 274 (5287): 546–67.

Goig, Galo A., Silvia Blanco, Alberto L. Garcia-Basteiro, and Iñaki Comas. 2020. “Contaminant DNA in Bacterial Sequencing Experiments Is a Major Source of False Genetic Variability.” *BMC Biology* 18 (1): 24.

Goldmann, J. M., J. A. Veltman, and C. Gilissen. 2019. “De Novo Mutations Reflect Development and Aging of the Human Germline.” *Trends in Genetics: TIG* 35 (11): 828–39.

Greaves, Mel, and William Hughes. 2018. “Cancer Cell Transmission via the Placenta.” *Evolution, Medicine, and Public Health* 2018 (1): 106–15.

Groves, Danielle C., Sarah L. Rowland-Jones, and Adrienn Angyal. 2021. “The D614G Mutations in the SARS-CoV-2 Spike Protein: Implications for Viral Infectivity, Disease Severity and Vaccine Design.” *Biochemical and Biophysical Research Communications* 538 (January): 104–7.

Gupta, Pushpendra K., Pawan L. Kulwal, and Vandana Jaiswal. 2019. “Association Mapping in Plants in the Post-GWAS Genomics Era.” *Advances in Genetics* 104 (January): 75–154.

Gurwitz, David. 2019. “Whole-Genome Sequencing for Combatting Antibiotic Resistance.” *Drug Development Research* 80 (1): 3–5.

Hadfield, James, Colin Megill, Sidney M. Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A. Neher. 2018. “Nextstrain: Real-Time Tracking of Pathogen Evolution.” *Bioinformatics* 34 (23): 4121–23.

Hanahan, Douglas, and Robert A. Weinberg. 2011. “Hallmarks of Cancer: The next

Generation.” *Cell* 144 (5): 646–74.
Hendriks, Lizza E., Etienne Rouleau, and Benjamin Besse. 2018. “Clinical Utility of Tumor Mutational Burden in Patients with Non-Small Cell Lung Cancer Treated with Immunotherapy.” *Translational Lung Cancer Research* 7 (6): 647–60.
Holmes, Edward C., Gytis Dudas, Andrew Rambaut, and Kristian G. Andersen. 2016. “The Evolution of Ebola Virus: Insights from the 2013-2016 Epidemic.” *Nature* 538 (7624): 193–200.
Hortobagyi, Gabriel N., Stephen B. Edge, and Armando Giuliano. 2018. “New and Important Changes in the TNM Staging System for Breast Cancer.” *American Society of Clinical Oncology Educational Book. American Society of Clinical Oncology. Annual Meeting* 38 (May): 457–67.
Initiative, The Arabidopsis Genome, and The Arabidopsis Genome Initiative. 2000. “Analysis of the Genome Sequence of the Flowering Plant Arabidopsis Thaliana.” *Nature*. https://doi.org/10.1038/35048692.
International Human Genome Sequencing Consortium. 2004. “Finishing the Euchromatic Sequence of the Human Genome.” *Nature* 431 (7011): 931–45.
Jiménez-Luna, J., F. Grisoni, and G. Schneider. 2020. “Drug Discovery with Explainable Artificial Intelligence.” *Nat Mach Intell*, no. 2 (October): 573–84.
Kassim, Adetola A., and Bipib N. Savani. 2017. “Hematopoietic Stem Cell Transplantation for Acute Myeloid Leukemia: A Review.” *Hematology/oncology and Stem Cell Therapy* 10 (4): 245–51.
Katsonis, Panagiotis, Amanda Koire, Stephen Joseph Wilson, Teng-Kuei Hsu, Rhonald C. Lua, Angela Dawn Wilkins, and Olivier Lichtarge. 2014. “Single Nucleotide Variations: Biological Impact and Theoretical Interpretation.” *Protein Science: A Publication of the Protein Society* 23 (12): 1650–66.
Kempfer, Rieke, and Ana Pombo. 2020. “Methods for Mapping 3D Chromosome Architecture.” *Nature Reviews. Genetics* 21 (4): 207–26.
Kemp, Steven A., Dami A. Collier, Rawlings P. Datir, Isabella A. T. M. Ferreira, Salma Gayed, Aminu Jahun, Myra Hosmillo, et al. 2021. “SARS-CoV-2 Evolution during Treatment of Chronic Infection.” *Nature*, February. https://doi.org/10.1038/s41586-021-03291-y.
Kernaleguen, Magali, Christian Daviaud, Yimin Shen, Eric Bonnet, Victor Renault, Jean-François Deleuze, Florence Mauger, and Jörg Tost. 2018. “Whole-Genome Bisulfite Sequencing for the Analysis of Genome-Wide DNA Methylation and Hydroxymethylation Patterns at Single-Nucleotide Resolution.” *Methods in Molecular Biology* 1767: 311–49.
Kerr, Anne, Choon Key Chekar, Emily Ross, Julia Swallow, and Sarah Cunningham-Burley. 2021. *Personalised Cancer Medicine: Future Crafting in the Genomic Era*. Manchester (UK): Manchester University Press.
Kim, Sollip, Ki-Won Eom, Chong-Rae Cho, and Tae Hyun Um. 2014. “Comparison of Commercial Genetic-Testing Services in Korea with 23andMe Service.” *BioMed Research International* 2014 (June): 539151.
Klempner, Samuel J., David Fabrizio, Shalmali Bane, Marcia Reinhart, Tim Peoples, Siraj M. Ali, Ethan S. Sokol, et al. 2020. “Tumor Mutational Burden as a Predictive Biomarker for Response to Immune Checkpoint Inhibitors: A Review of Current Evidence.” *The Oncologist* 25 (1): e147–59.
Klimova, Blanka, Kamil Kuca, Michal Novotny, and Petra Maresova. 2017. “Cystic Fibrosis Revisited - a Review Study.” *Medicinal Chemistry* 13 (2): 102–9.
Koboldt, Daniel C. 2020. “Best Practices for Variant Calling in Clinical Sequencing.” *Genome Medicine* 12 (1): 91.
Korber, Bette, Will M. Fischer, Sandrasegaram Gnanakaran, Hyejin Yoon, James Theiler, Werner Abfalterer, Nick Hengartner, et al. 2020. “Tracking Changes in SARS-CoV-2 Spike: Evidence That D614G Increases Infectivity of the COVID-19 Virus.” *Cell* 182 (4): 812–27.e19.
Kotsias, Fiorella, Ignacio Cebrian, and Andrés Alloatti. 2019. “Antigen Processing and Presentation.” *International Review of Cell and Molecular Biology* 348 (August): 69–121.

Kratz, Johannes R., and David M. Jablons. 2009. “Genomic Prognostic Models in Early-Stage Lung Cancer.” *Clinical Lung Cancer*. https://doi.org/10.3816/clc.2009.n.021.

Kuhn, Jens H., Yiming Bao, Sina Bavari, Stephan Becker, Steven Bradfute, J. Rodney Brister, Alexander A. Bukreyev, et al. 2013. “Virus Nomenclature below the Species Level: A Standardized Nomenclature for Natural Variants of Viruses Assigned to the Family Filoviridae.” *Archives of Virology*. https://doi.org/10.1007/s00705-012-1454-0.

Kumar, Kishore R., Mark J. Cowley, and Ryan L. Davis. 2019. “Next-Generation Sequencing and Emerging Technologies.” *Seminars in Thrombosis and Hemostasis* 45 (07): 661–73.

Kuo, Frank C., Brenton G. Mar, R. Coleman Lindsley, and Neal I. Lindeman. 2017. “The Relative Utilities of Genome-Wide, Gene Panel, and Individual Gene Sequencing in Clinical Practice.” Blood 130 (4): 433–39.

Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. 2009. “Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome.” *Genome Biology* 10 (3): R25.

Leccese, Pietro, and Erkan Alpsoy. 2019. “Behçet’s Disease: An Overview of Etiopathogenesis.” *Frontiers in Immunology* 10 (May): 1067.

Lee, J-M, E. M. Ramos, J-H Lee, T. Gillis, J. S. Mysore, M. R. Hayden, S. C. Warby, et al. 2012. “CAG Repeat Expansion in Huntington Disease Determines Age at Onset in a Fully Dominant Fashion.” *Neurology* 78 (10): 690–95.

Leggett, Richard M., and Matthew D. Clark. 2017. “A World of Opportunities with Nanopore Sequencing.” *Journal of Experimental Botany* 68 (20): 5419–29.

Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics* 25 (14): 1754–60.

Lodish, H., A. Berk, and S. L. Zipursky, eds. 2000. “Section 8.1, Mutations: Types and Causes.” In *Molecular Cell Biology. 4th Edition.* New York: W. H. Freeman.

Loman, Nick, Will Rowe, and Andrew Rambaut. 2020. “nCoV-2019 Novel Coronavirus Bioinformatics Protocol.” https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html.

Lorente, Elena, Jennifer Redondo-Antón, Adrian Martín-Esteban, Pablo Guasp, Eilon Barnea, Pilar Lauzurica, Arie Admon, and José A. López de Castro. 2019. “Substantial Influence of ERAP2 on the HLA-B*40:02 Peptidome: Implications for HLA-B*27-Negative Ankylosing Spondylitis*.” *Molecular & Cellular Proteomics*. https://doi.org/10.1074/mcp.ra119.001710.

Løvestad, A. H., S. B. Jørgensen, N. Handal, O. H. Ambur, and H. V. Aamot. 2021. “Investigation of Intra-Hospital SARS-CoV-2 Transmission Using Nanopore Whole-Genome Sequencing.” *The Journal of Hospital Infection*, February. https://doi.org/10.1016/j.jhin.2021.02.022.

Lu, Yuan, Yingjia Shen, Wesley Warren, and Ronald Walter. 2016. “Next Generation Sequencing in Aquatic Models.” *Next Generation Sequencing - Advances, Applications and Challenges*. https://doi.org/10.5772/61657.

Madan, Ravi A., Emmanuel S. Antonarakis, Charles G. Drake, Lawrence Fong, Evan Y. Yu, Douglas G. McNeel, Daniel W. Lin, Nancy N. Chang, Nadeem A. Sheikh, and James L. Gulley. 2020. “Putting the Pieces Together: Completing the Mechanism of Action Jigsaw for Sipuleucel-T.” *Journal of the National Cancer Institute* 112 (6): 562–73.

Mandelker, Diana, Ryan J. Schmidt, Arunkanth Ankala, Kristin Mcdonald Gibson, Mark Bowser, Himanshu Sharma, Elizabeth Duffy, et al. 2016. “Navigating Highly Homologous Genes in a Molecular Diagnostic Setting: A Resource for Clinical next-Generation Sequencing.” *Genetics in Medicine*. https://doi.org/10.1038/gim.2016.58.

Mao, Chengyi, Xiaoxi Zeng, Chao Zhang, Yushang Yang, Xin Xiao, Siyuan Luan, Yonggang Zhang, and Yong Yuan. 2021. “Mechanisms of Pharmaceutical Therapy and Drug Resistance in Esophageal Cancer.” *Frontiers in Cell and Developmental Biology* 9 (February): 612451.

Mattiuzzi, Camilla, and Giuseppe Lippi. 2019. “Current Cancer Epidemiology.” *Journal of Epidemiology and Global Health*. https://doi.org/10.2991/jegh.k.191008.001.

McFarland, Christopher D., Julia A. Yaglom, Jonathan W. Wojtkowiak, Jacob G. Scott, David L. Morse, Michael Y. Sherman, and Leonid A. Mirny. 2017. “The Damaging Effect of Passenger Mutations on Cancer Progression.” *Cancer Research* 77 (18): 4763–72.
McGranahan, Nicholas, Andrew J. S. Furness, Rachel Rosenthal, Sofie Ramskov, Rikke Lyngaa, Sunil Kumar Saini, Mariam Jamal-Hanjani, et al. 2016. “Clonal Neoantigens Elicit T Cell Immunoreactivity and Sensitivity to Immune Checkpoint Blockade.” Science 351 (6280): 1463–69.
McGranahan, Nicholas, and Charles Swanton. 2017. “Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future.” *Cell* 168 (4): 613–28.
Meeks, Joshua J., Hikmat Al-Ahmadie, Bishoy M. Faltas, John A. Taylor 3rd, Thomas W. Flaig, David J. DeGraff, Emil Christensen, Benjamin L. Woolbright, David J. McConkey, and Lars Dyrskjøt. 2020. “Genomic Heterogeneity in Bladder Cancer: Challenges and Possible Solutions to Improve Outcomes.” *Nature Reviews. Urology* 17 (5): 259–70.
Michaels, Daniel, and Jason Douglas. 2020. “Countries Ban Travel From U.K. in Race to Block New Covid-19 Strain.” *WSJ*, December 20, 2020.
Michel, Christian Jean, Claudine Mayer, Olivier Poch, and Julie Dawn Thompson. 2020. “Characterization of Accessory Genes in Coronavirus Genomes.” *Virology Journal* 17 (1): 131.
Miller, A., Y. Asmann, L. Cattaneo, E. Braggio, J. Keats, D. Auclair, S. Lonial, MMRF CoMMpass Network, S. J. Russell, and A. K. Stewart. 2017. “High Somatic Mutation and Neoantigen Burden Are Correlated with Decreased Progression-Free Survival in Multiple Myeloma.” *Blood Cancer Journal* 7 (9): e612.
Mitchell, Kieren J., and Nicolas J. Rawlence. 2021. “Examining Natural History through the Lens of Palaeogenomics.” *Trends in Ecology & Evolution* 36 (3): 258–67.
Montaño, Adrián, Maribel Forero-Castro, Jesús-María Hernández-Rivas, Ignacio García-Tuñón, and Rocío Benito. 2018. “Targeted Genome Editing in Acute Lymphoblastic Leukemia: A Review.” *BMC Biotechnology* 18 (1): 45.
Mount, David W. 2004. *Bioinformatics: Sequence and Genome Analysis*. CSHL Press.
Muir, Paul, Shantao Li, Shaoke Lou, Daifeng Wang, Daniel J. Spakowicz, Leonidas Salichos, Jing Zhang, et al. 2016. “The Real Cost of Sequencing: Scaling Computation to Keep Pace with Data Generation.” *Genome Biology* 17 (March): 53.
Neefjes, Jacques, Marlieke L. M. Jongsma, Petra Paul, and Oddmund Bakke. 2011. “Towards a Systems Understanding of MHC Class I and MHC Class II Antigen Presentation.” *Nature Reviews. Immunology* 11 (12): 823–36.
Nicola, Maria, Zaid Alsafi, Catrin Sohrabi, Ahmed Kerwan, Ahmed Al-Jabir, Christos Iosifidis, Maliha Agha, and Riaz Agha. 2020. “The Socio-Economic Implications of the Coronavirus Pandemic (COVID-19): A Review.” *International Journal of Surgery* 78 (June): 185–93.
Nielsen, Rasmus, Joshua M. Akey, Mattias Jakobsson, Jonathan K. Pritchard, Sarah Tishkoff, and Eske Willerslev. 2017. “Tracing the Peopling of the World through Genomics.” *Nature* 541 (7637): 302–10.
November, Joseph. 2018. “More than Moore’s Mores: Computers, Genomics, and the Embrace of Innovation.” *Journal of the History of Biology*. https://doi.org/10.1007/s10739-018-9539-6.
Ortet, Philippe, and Olivier Bastien. 2010. “Where Does the Alignment Score Distribution Shape Come From?” *Evolutionary Bioinformatics Online* 6 (December): 159–87.
Oude Munnink, Bas B., Reina S. Sikkema, David F. Nieuwenhuijse, Robert Jan Molenaar, Emmanuelle Munger, Richard Molenkamp, Arco van der Spek, et al. 2021. “Transmission of SARS-CoV-2 on Mink Farms between Humans and Mink and back to Humans.” *Science* 371 (6525): 172–77.
Ou, Sai-Hong Ignatius, Misako Nagasaka, and Viola W. Zhu. 2018. “Liquid Biopsy to Identify Actionable Genomic Alterations.” *American Society of Clinical Oncology Educational Book. American Society of Clinical Oncology. Annual Meeting* 38 (May): 978–97.
Oxford Nanopore Technologies. 2021. “How Does Nanopore DNA/RNA Sequencing Work?” GB. 2021. https://nanoporetech.com/how-it-works#nanopore-scaling&.

Pan, Chongxian, Hongtao Liu, Elizabeth Robins, Wenru Song, Delong Liu, Zihai Li, and Lei Zheng. 2020. “Next-Generation Immuno-Oncology Agents: Current Momentum Shifts in Cancer Immunotherapy.” *Journal of Hematology & Oncology* 13 (1): 29.

Peel, Emma, and Katherine Belov. 2018. “Lessons Learnt from the Tasmanian Devil Facial Tumour Regarding Immune Function in Cancer.” *Mammalian Genome: Official Journal of the International Mammalian Genome Society* 29 (11-12): 731–38.

Petersen, Lauren M., Isabella W. Martin, Wayne E. Moschetti, Colleen M. Kershaw, and Gregory J. Tsongalis. 2019. “Third-Generation Sequencing in the Clinical Laboratory: Exploring the Advantages and Challenges of Nanopore Sequencing.” *Journal of Clinical Microbiology* 58 (1). https://doi.org/10.1128/JCM.01315-19.

Petrosillo, N., G. Viceconte, O. Ergonul, G. Ippolito, and E. Petersen. 2020. “COVID-19, SARS and MERS: Are They Closely Related?” *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 26 (6): 729–34.

Peyrégne, Stéphane, and Kay Prüfer. 2020. “Present-Day DNA Contamination in Ancient DNA Datasets.” *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 42 (9): e2000081.

Pymm, Phillip, Patricia T. Illing, Sri H. Ramarathinam, Geraldine M. O’Connor, Victoria A. Hughes, Corinne Hitchen, David A. Price, et al. 2017. “MHC-I Peptides Get out of the Groove and Enable a Novel Mechanism of HIV-1 Escape.” *Nature Structural & Molecular Biology* 24 (4): 387–94.

Qin, Si-Yong, Ai-Qing Zhang, and Xian-Zheng Zhang. 2018. “Recent Advances in Targeted Tumor Chemotherapy Based on Smart Nanomedicines.” *Small* 14 (45): e1802417.

Quail, Michael A., Miriam Smith, Paul Coupland, Thomas D. Otto, Simon R. Harris, Thomas R. Connor, Anna Bertoni, Harold P. Swerdlow, and Yong Gu. 2012. “A Tale of Three next Generation Sequencing Platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq Sequencers.” *BMC Genomics* 13 (July): 341.

Rambaut Group. 2021. “B.1.525.” PANGO Lineages. February 20, 2021. https://cov-lineages.org/global_report_B.1.525.html.

Ranstam, J., and J. A. Cook. 2017. “Kaplan-Meier Curve.” *The British Journal of Surgery* 104 (4): 442.

Reck, Martin, Hossein Borghaei, and Kenneth J. O’Byrne. 2019. “Nivolumab plus Ipilimumab in Non-Small-Cell Lung Cancer.” *Future Oncology* 15 (19): 2287–2302.

Rich, Jason T., J. Gail Neely, Randal C. Paniello, Courtney C. J. Voelker, Brian Nussenbaum, and Eric W. Wang. 2010. “A Practical Guide to Understanding Kaplan-Meier Curves.” *Otolaryngology--Head and Neck Surgery: Official Journal of American Academy of Otolaryngology-Head and Neck Surgery* 143 (3): 331–36.

Rosenthal, Rachel, Elizabeth Larose Cadieux, Roberto Salgado, Maise Al Bakir, David A. Moore, Crispin T. Hiley, Tom Lund, et al. 2019. “Neoantigen-Directed Immune Escape in Lung Cancer Evolution.” *Nature* 567 (7749): 479–85.

Rudd-Schmidt, Jesse A., Adrian W. Hodel, Tahereh Noori, Jamie A. Lopez, Hyun-Jung Cho, Sandra Verschoor, Annette Ciccone, Joseph A. Trapani, Bart W. Hoogenboom, and Ilia Voskoboinik. 2019. “Lipid Order and Charge Protect Killer T Cells from Accidental Death.” *Nature Communications* 10 (1): 5396.

Sandmann, Sarah, Aniek O. de Graaf, Mohsen Karimi, Bert A. van der Reijden, Eva Hellström-Lindberg, Joop H. Jansen, and Martin Dugas. 2017. “Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data.” *Scientific Reports* 7 (February): 43169.

Schmieder, Robert, and Robert Edwards. 2011. “Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets.” *PloS One* 6 (3): e17288.

Schneider, Anne-Fleur E., and Annemieke Aartsma-Rus. 2021. “Developments in Reading Frame Restoring Therapy Approaches for Duchenne Muscular Dystrophy.” *Expert Opinion on Biological Therapy* 21 (3): 343–59.

Schön, Michael P. 2019. “Adaptive and Innate Immunity in Psoriasis and Other Inflammatory

Disorders." *Frontiers in Immunology* 10 (July): 1764.
Schumacher, Ton N., Wouter Scheper, and Pia Kvistborg. 2019. "Cancer Neoantigens." *Annual Review of Immunology* 37 (April): 173–200.
Schwarze, Katharina, James Buchanan, Jilles M. Fermont, Helene Dreau, Mark W. Tilley, John M. Taylor, Pavlos Antoniou, et al. 2020. "The Complete Costs of Genome Sequencing: A Microcosting Study in Cancer and Rare Diseases from a Single Center in the United Kingdom." *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 22 (1): 85–94.
Senator, Hermann. 1882. Zur Kenntniss der Leukämie und Pseudoleukämie im Kindesalter. *Berliner Klinische Wochenschrift.* 1882;35: 533-536.
Shah, Nirali N., and Terry J. Fry. 2019. "Mechanisms of Resistance to CAR T Cell Therapy." *Nature Reviews. Clinical Oncology* 16 (6): 372–85.
Sharma, N., and G. R. Cutting. 2020. "The Genetics and Genomics of Cystic Fibrosis." *Journal of Cystic Fibrosis*. https://doi.org/10.1016/j.jcf.2019.11.003.
Shringarpure, Suyash S., Rasika A. Mathias, Ryan D. Hernandez, Timothy D. O'Connor, Zachary A. Szpiech, Raul Torres, Francisco M. De La Vega, et al. 2017. "Using Genotype Array Data to Compare Multi- and Single-Sample Variant Calls and Improve Variant Call Sets from Deep Coverage Whole-Genome Sequencing Data." *Bioinformatics* 33 (8): 1147–53.
Shu, Yuelong, and John McCauley. 2017. "GISAID: Global Initiative on Sharing All Influenza Data - from Vision to Reality." *Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin* 22 (13). https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494.
Siddik, Md Nur Alam. 2020. "Economic Stimulus for COVID-19 Pandemic and Its Determinants: Evidence from Cross-Country Analysis." *Heliyon* 6 (12): e05634.
Siegel, Rebecca L., Kimberly D. Miller, and Ahmedin Jemal. 2019. "Cancer Statistics, 2019." *CA: A Cancer Journal for Clinicians* 69 (1): 7–34.
Simmons, Claribel P. L., Donald C. McMillan, Kerry McWilliams, Tonje A. Sande, Kenneth C. Fearon, Sharon Tuck, Marie T. Fallon, and Barry J. Laird. 2017. "Prognostic Tools in Patients With Advanced Cancer: A Systematic Review." *Journal of Pain and Symptom Management* 53 (5): 962–70.e10.
Stirrup, Oliver T., Joseph Hughes, Matthew Parker, David G. Partridge, James G. Shepherd, James Blackstone, Francesc Coll, et al. 2020. "Rapid Feedback on Hospital Onset SARS-CoV-2 Infections Combining Epidemiological and Sequencing Data." *medRxiv*. https://doi.org/10.1101/2020.11.12.20230326.
Sung, Hyuna, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. 2021. "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries." *CA: A Cancer Journal for Clinicians*. https://doi.org/10.3322/caac.21660.
Tegally, Houriiyah, Eduan Wilkinson, Marta Giovanetti, Arash Iranzadeh, Vagner Fonseca, Jennifer Giandhari, Deelan Doolabh, et al. 2021. "Emergence of a SARS-CoV-2 Variant of Concern with Mutations in Spike Glycoprotein." *Nature*, March. https://doi.org/10.1038/s41586-021-03402-9.
Teo, Min Yuen, Dana E. Rathkopf, and Philip Kantoff. 2019. "Treatment of Advanced Prostate Cancer." *Annual Review of Medicine* 70 (January): 479–99.
The C. elegans Sequencing Consortium*. 1998. "Genome Sequence of the Nematode C. Elegans: A Platform for Investigating Biology." *Science* 282 (5396): 2012–18.
Tsai, Isheng J., Thomas D. Otto, and Matthew Berriman. 2010. "Improving Draft Assemblies by Iterative Mapping and Assembly of Short Reads to Eliminate Gaps." *Genome Biology* 11 (4): R41.
Tyson, John R., Phillip James, David Stoddart, Natalie Sparks, Arthur Wickenhagen, Grant Hall, Ji Hyun Choi, et al. 2020. "Improvements to the ARTIC Multiplex PCR Method for SARS-CoV-2 Genome Sequencing Using Nanopore." *bioRxiv : The Preprint Server for Biology*, September. https://doi.org/10.1101/2020.09.04.283077.
UKRI. 2020. "How Does Virus Genome Sequencing Help the Response to COVID-19?"

March 25, 2020. https://coronavirusexplained.ukri.org/en/article/und0001/.
Unanue, Emil R., Vito Turk, and Jacques Neefjes. 2016. “Variations in MHC Class II Antigen Processing and Presentation in Health and Disease.” *Annual Review of Immunology* 34 (May): 265–97.
Verhaart, Ingrid E. C., and Annemieke Aartsma-Rus. 2019. “Therapeutic Developments for Duchenne Muscular Dystrophy.” *Nature Reviews. Neurology* 15 (7): 373–86.
V’kovski, Philip, Annika Kratzel, Silvio Steiner, Hanspeter Stalder, and Volker Thiel. 2021. “Coronavirus Biology and Replication: Implications for SARS-CoV-2.” *Nature Reviews. Microbiology* 19 (3): 155–70.
Wang, Dennis, Nhu-An Pham, Timothy M. Freeman, Vibha Raghavan, Roya Navab, Jonathan Chang, Chang-Qi Zhu, et al. 2019. “Somatic Alteration Burden Involving Non-Cancer Genes Predicts Prognosis in Early-Stage Non-Small Cell Lung Cancer.” *Cancers*. https://doi.org/10.3390/cancers11071009.
Wetterstrand, K. A. 2021. “DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).” https://www.genome.gov/sequencingcostsdata.
Wick, Ryan R., Louise M. Judd, and Kathryn E. Holt. 2019. “Performance of Neural Network Basecalling Tools for Oxford Nanopore Sequencing.” *Genome Biology* 20 (1): 129.
Williamson, Elizabeth J., Alex J. Walker, Krishnan Bhaskaran, Seb Bacon, Chris Bates, Caroline E. Morton, Helen J. Curtis, et al. 2020. “Factors Associated with COVID-19-Related Death Using OpenSAFELY.” *Nature* 584 (7821): 430–36.
Wirth, Thomas C., and Florian Kühnel. 2017. “Neoantigen Targeting-Dawn of a New Era in Cancer Immunotherapy?” *Frontiers in Immunology* 8 (December): 1848.
Wood, Mary A., Benjamin R. Weeder, Julianne K. David, Abhinav Nellore, and Reid F. Thompson. 2020. “Burden of Tumor Mutations, Neoepitopes, and Other Variants Are Weak Predictors of Cancer Immunotherapy Response and Overall Survival.” *Genome Medicine*. https://doi.org/10.1186/s13073-020-00729-2.
Wu, Song, Wei Zhu, Patricia Thompson, and Yusuf A. Hannun. 2018. “Evaluating Intrinsic and Non-Intrinsic Cancer Risk Factors.” *Nature Communications* 9 (1): 3490.
Xu, Feng, Chongtao Ge, Hao Luo, Shaoting Li, Martin Wiedmann, Xiangyu Deng, Guangtao Zhang, Abigail Stevenson, Robert C. Baker, and Silin Tang. 2020. “Evaluation of Real-Time Nanopore Sequencing for Salmonella Serotype Prediction.” *Food Microbiology* 89 (August): 103452.
Yau, Anthony C. Y., Jonatan Tuncel, and Rikard Holmdahl. 2017. “The Major Histocompatibility Complex Class III Haplotype Ltab-Ncr3 Regulates Adjuvant-Induced but Not Antigen-Induced Autoimmunity.” *The American Journal of Pathology* 187 (5): 987–98.
Yung, Godwin, and Yi Liu. 2020. “Sample Size and Power for the Weighted Log-Rank Test and Kaplan-Meier Based Tests with Allowance for Nonproportional Hazards.” *Biometrics* 76 (3): 939–50.
Yurkovetskiy, Leonid, Xue Wang, Kristen E. Pascal, Christopher Tomkins-Tinch, Thomas P. Nyalile, Yetao Wang, Alina Baum, et al. 2020. “Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein Variant.” *Cell* 183 (3): 739–51.e8.
Zhang, Wenjuan, Brian D. Davis, Stephanie S. Chen, Jorge M. Sincuir Martinez, Jasmine T. Plummer, and Eric Vail. 2021. “Emergence of a Novel SARS-CoV-2 Variant in Southern California.” *JAMA: The Journal of the American Medical Association*, February. https://doi.org/10.1001/jama.2021.1612.
Zhou, Peng, Xing-Lou Yang, Xian-Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, Hao-Rui Si, et al. 2020. “A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin.” *Nature* 579 (7798): 270–73.
Zook, Justin M., Brad Chapman, Jason Wang, David Mittelman, Oliver Hofmann, Winston Hide, and Marc Salit. 2014. “Integrating Human Sequence Data Sets Provides a Resource of Benchmark SNP and Indel Genotype Calls.” *Nature Biotechnology* 32 (3): 246–51.
Zook, Justin M., Jennifer McDaniel, Nathan D. Olson, Justin Wagner, Hemang Parikh, Haynes Heaton, Sean A. Irvine, et al. 2019. “An Open Resource for Accurately

Benchmarking Small Variant and Reference Calls.” *Nature Biotechnology* 37 (5): 561–66.

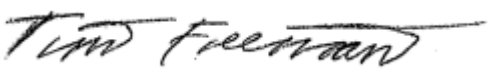
Chapter II: Somatic Alteration Burden Involving Non-Cancer Genes Predicts Prognosis in Early-Stage Non-Small Cell Lung Cancer

As part of my PhD thesis, I am including work from my published paper *"Somatic Alteration Burden Involving Non-Cancer Genes Predicts Prognosis in Early-Stage Non-Small Cell Lung Cancer"*, which was published in *Cancers*, DOI: 10.3390/cancers11071009

This work was done in collaboration with the Princess Margaret Cancer Centre and the University of Toronto, who provided access to the PDX data used in the study and carried out the wet lab work and much of the data collection involved with the paper. For my role, I contributed a large proportion of the experimental analyses, including much of the bioinformatics and statistical analyses and stratification of patients for the main results of the paper, comprising roughly 50% of the figures and tables in this paper, 10 of which I produced. My PhD supervisor Dr Dennis Wang, who is also the first author, helped guide the study design for the experimental work and organised the drafting of the manuscript.

I produced figures 3A (2.2A), 3B (2.2B), S2A (2.1), S3 (2.5) and tables 1 (2.2), 2 (2.3) and S5 (2.4) in this paper, which represent experimental work and analyses that I performed as part of this publication. In addition, my experimental work and analyses contributed towards figures 1, 2B, 2C, 3C (2.2C), 3D (2.2D), S2B (2.4A), and S2C (2.4B) and table S6 (2.1) although I did not produce those figures myself. I wrote parts of the results, discussion and materials/methods that related to my work stratifying patients by neoantigen presence and various other genomic and clinical factors (especially within sections 2.3, 2.4, 4.2, 4.3 and 4.5), and also provided feedback on other sections, which was used to edit the final manuscript.

Yours sincerely,



Timothy Freeman (PhD candidate)

Dr Dennis Wang (first author and corresponding author)

The definitive version of this work was published in Cancers, DOI: 10.3390/cancers11071009

This version is adapted to only include the sections, figures and tables of the publication which included my experimental work and analyses. I contributed to the writing of these sections and edited the final draft. I have re-written the introduction, discussion and conclusion extensively to frame my contribution to this publication and provide further background and interpretation, including additional references to recent studies that were unpublished before the published manuscript was written. I have also added in supplementary figures and tables I generated all or part of, that were not included in the text of the original manuscript. Figures and tables are therefore numbered differently from the published version. Where parts of results, figures or tables were co-authored, I have indicated the contributions.

Introduction

Motivation

This chapter provides an example of a computational genomic diagnostic tool that explores the utility of a genomic feature named “neoantigenic burden” in classifying non-small cell lung cancer (NSCLC) patients, in order to predict patient survival rates and tumour immune status, and better understand the mechanisms by which immune activity regulates tumour progression.

In the US alone it is estimated that there will be 235,760 new cases of non-small cell lung cancers (lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), which make up 80-85% of all lung cancer diagnoses) and 131,880 deaths in 2021, with the majority of patients failing to survive longer than a year after diagnosis (Siegel et al. 2021). It is difficult to accurately predict survival outcomes for patients at diagnosis (Pellegrino et al. 2021), which can obstruct clinicians from selecting the appropriate treatment regimen on a case-by-case basis, leading to reduced patient survival outcomes, quality of life and financial losses (Qu et al. 2021; Duma, Santana-Davila, and Molina 2019). It is therefore vitally important that clinicians have access to prognostic tools that can give a personalised view of survival outcomes for individual patients. In addition, NSCLC is a comparatively good cancer for developing genomic prognostics because it has a relatively high number of mutations (Siegel et al. 2021), so there are a greater number of genomic variants available to examine. For these reasons, I have chosen to develop two separate genomic prognostic tools, that classify mutational burden and neoantigenic burden, for the purpose of identifying NSCLC patients’ survival odds, and to provide insight into the biology behind their condition, potentially paving the way for enhanced therapies and improved treatment decisions in the clinic on a personalised basis. We define mutational burden as a measure of the extent to which mutations occur in a specific panel of 865 genes, and neoantigen burden as a measure of the extent to which mutations in this gene panel lead to the expression of mutant peptides that trigger an immune response. Because these are general signatures, rather than specific markers in a single gene, they can be detected in a much larger cohort of patients and have the potential to provide utility to many more people.

Mutational and neoantigenic burden as genomic prognostic tools

Mutational burden is a measure of the number of mutations in a tumour (in this study referring to single nucleotide variants as well as copy number variants that are detected), and their

prevalence. Later stage, more pathogenic tumours typically have more mutations since they have had more time to accumulate these, so mutational burden often negatively correlates with patient survival. However, there is evidence that this is not the case for all cancer types, since a positive correlation has been found with patient survival in NSCLC (Nan et al. 2021; Greillier, Tomasini, and Barlesi 2018), melanoma (Buder-Bakhaya and Hassel 2018), breast cancer (Noguchi, Shien, and Iwata 2021) and multiple myeloma patients (Amber Miller et al. 2016) with a range of treatments, especially immune checkpoint inhibitors. Mutational burden may therefore help to predict patient prognosis and inform treatment.

Neoantigenic burden is the sum total of all novel immunogenic peptides present in a tumour that are expressed, presented to the immune system and which elicit an immune response. Neoantigenic burden has a positive correlation with patient survival outcome in NSCLC, but requires many factors to score accurately (Richters et al. 2019). There are a variety of different computational algorithms, trained using in vitro validated antigenic peptides, to predict the binding of peptides to receptors which present them to the immune system. These include artificial neural network (ANN) based methods, such as NetMHC (Nielsen and Andreatta 2016), NetMHCpan (Reynisson et al. 2020) and MHCflurry (O'Donnell, Rubinsteyn, and Laserson 2020), stabilised matrix methods (SMMs) (Peters and Sette 2005) and more simplistic methods that do not take into account specific amino acid sequences (Calis et al. 2013). By considering a variety of biological factors along with the binding affinity of the peptide to major histocompatibility complex I (MHC I) (De Mattos-Arruda et al. 2020), we can predict neoantigenic burden using these algorithms, which has been shown to be a major factor in NSCLC patient survival odds with immunotherapy treatments, but had not been conclusively shown in patients undergoing other treatments or untreated patients before the study presented in this chapter (Wang et al. 2019). In work I carried out prior to my PhD (Freeman 2016), I compared a range of ANN-based methods for calculating MHC-I binding affinities, including those listed above. By measuring their predictions for a balanced set of experimentally-validated binding and non-binding peptides I was able to calculate the sensitivity and specificity for correctly predicting whether a given peptide sequence would bind MHC-I at a range of different binding affinity thresholds, from which I plotted a receiver operating characteristic (ROC) curve for each algorithm. NetMHCpan performed the best in this comparison with the highest area under the curve (AUC), showing that it had the most accurate MHC-I binding affinity predictions. In this chapter, I therefore decided to use NetMHCpan for MHC-I binding affinity prediction.

Clinical usage of genomic prognostic markers in cancer

Diagnostics profiling the patient's RNA or DNA can be used to identify prognostic factors and aid treatment selection. Earlier successes include analyses of gene expression from a single gene, such as the use of *EGFR* expression to predict patients who would benefit most from cetuximab treatment in combination with chemotherapy vs. chemotherapy alone (Pirker et al. 2012; Ellis et al. 2011), in addition to examples where specific genomic alterations were used as prognostic factors, in both untreated and treated patients, such as *ALK* rearrangements in lung cancer (Pikor et al. 2013; Ellis et al. 2011), *BRAF* mutations in melanomas, *KIT* and *PDGFRA* mutations in gastrointestinal stromal tumours and *BCR-ABL* kinase fusions in chronic myelogenous leukaemia (Heuckmann and Thomas 2015). More recent examples of genomic prognostics include the use of panels containing many genes, rather than a small number of specific genes, for assessing mutational and expression changes, along with complex multi-omic scores, including the use of various different panels for assessing TMB (Heydt et al. 2020) and different methods of calculating immunogenicity that use neoantigen scoring, genetic and epigenetic signatures in combination with immune cell counts (Darvin et al. 2018).

Models for identifying genomic prognostic markers

Genomic prognostic markers can be identified either directly from patient tumours, or indirectly from human tumour samples inserted into immunodeficient mice, known as patient-derived xenografts (PDXs). PDXs have been found to closely model patient primary tumours in predicting aggressive tumour phenotypes, and are particularly effective compared to cell lines (Wang et al. 2017; Wu et al. 2017; John et al. 2011; Zhang et al. 2013). Some concerns have been raised about their heterogeneity and lack of host immune interaction not accurately reflecting real tumours (Aparicio, Hidalgo, and Kung 2015; Hidalgo et al. 2014). Despite this, PDX tumour behaviours, such as engraftment in mouse hosts, have been found to be valuable prognostic tools for the patients from which they are derived (Whittle et al. 2015; Yoshida 2020). Before our study (Wang et al. 2019), PDXs had not been used to identify genomic prognostic features from tumour DNA or RNA in untreated NSCLC patients. Prior to the work presented in this chapter, Wang et al. had used PDXs to develop a panel of 865 genes, which had non-synonymous mutations in at least two of 36 total PDX tumours and in which mutations correlated with 5-year overall survival (OS). They used an ElasticNet model from the R package "glmnet", with a regression fitting method using an alpha value of 0.1, to do this. The best fitting model was then selected from the maximum of the deviance ratio. The threshold of ≥2 PDX tumours was chosen in order to remove mutations that only occurred in a single

sample, and the remaining genes after applying this threshold which had non-zero coefficients under the ElasticNet model composed the panel of 865 genes correlated with OS (Wang et al. 2019). I used this gene panel for the mutation and neoantigen burden genomic prognostic classifiers I developed in this study.

Aims

In this chapter I aimed to establish the value of genomic prognostic measures using mutational and neoantigenic burden in establishing 5-year OS in NSCLC patients, particularly those sequenced at an early stage of cancer. As part of this, I sought to establish whether using both of these measures in combination boosted prognostic ability beyond using just one on its own, and also whether using a specific panel of genes offered improved prognostic ability above classifying neoantigens for all protein-coding genes. By carrying out these comparisons, I aimed to develop and benchmark optimal genomic prognostic methods for use in the clinic and among the scientific community.

Methods

TCGA patients

Published mutation, copy number, RNA-seq and clinical data were downloaded for LUAD (lung adenocarcinoma) (The Cancer Genome Atlas Research Network 2014) and LUSC (lung squamous cell carcinoma) (Cancer Genome Atlas Research Network 2012) patients on cBioPortal (https://www.cbioportal.org/) (Cerami et al. 2012; Gao et al. 2013), using the R package cgdsr v1.3.0 (Jacobsen 2017), and clinical data were downloaded using the FireBrowseR R package (Deng et al. 2017). Copy number gains and losses were determined using an LRR cut-off of ±0.5. RNA expression data in the RNA-seq by expectation maximisation (RSEM) scale from RNA-seq were transformed by the asinh function for survival analysis.

Stratification by high-level mutation burden

The number of somatic alterations among 865 genes (NAG) was used as the risk score for each TCGA patient. NAG scores above the median value for all patients (≥87) were considered high NAG, otherwise patients were considered to have low NAG. This “high NAG” threshold was calculated as the median value within the 36 patient PDX tumour cohort for the

number of genes with mutations, among the 865 genes. The proportions of OS were calculated using the Kaplan–Meier method, and the difference between curves was tested using the log-rank test. The OS is the time between the date of diagnosis to the date of death or last follow-up. All reported hazard ratio (HR) scores and p-values used the Kaplan–Meier method unless reported as multivariate analyses. A Cox proportional hazards model was used to fit survival times to the number of altered genes while adjusting for other clinical factors including age, stage and smoking status. The significance for the Cox proportional hazards model was based on the Wald test.

Estimating immunogenicity

Non-synonymous mutations causing amino acid changes were identified among the 865-gene panel for TCGA patients. Protein sequences were acquired from the UniProt reviewed canonical human proteome UP000005640 FASTA file. Windows of 8–11 amino acids in length were derived by applying non-synonymous SNVs to their respective protein sequence and using all 8–11mer amino acid peptides containing the altered amino acid. The 8–11mer altered peptides for each patient were input into NetMHCpan v3.0 (Nielsen and Andreatta 2016) along with both of the patient's supertyped HLA-A alleles to calculate their predicted HLA-A binding affinity values for each altered peptide. The 4-digit HLA types for the patients were sourced from The Cancer Immunome Atlas (Charoentong et al. 2017), which used Optitype (Szolek et al. 2014) to call the HLA-A alleles from RNA-seq FASTQ files. The HLA types were supertyped into standard categories prior to carrying out the HLA-A binding affinity prediction analysis (Sidney et al. 2008).

We defined peptides as antigenic if they had an HLA-A binding affinity below 500 nM with both HLA-A alleles. In addition, antigenic peptides' genes had to be expressed above the median expression level for that gene calculated across all patients in order for the peptides to be considered immunogenic, and the patient HLA-A gene needed to be expressed above the median HLA-A expression level across all patients. Patients were considered to have immunogenic peptides that could potentially trigger an immune response by fulfilling these three conditions: strong binding affinity, high expression of the antigenic peptide and high expression of the MHC-I receptor. Patients were stratified based on whether or not they had above the median number of immunogenic peptides in the 865-gene panel (1 or more) as outlined above, either alone, or in combination with the NAG score. A Cox proportional hazards model was used to calculate the effect size/risk factor associated with the presence/absence of these features, along with the p-values for the differences between patient groups. In order to justify the use of a Cox proportional hazards model, two

assumptions must be met by the underlying data. Firstly, the survival curves must have hazard functions (i.e. rates of patient death) that are proportional over time. Secondly, the relationship between the log hazard and each covariate needed to be linear. These assumptions were tested using the coxph() and cox.zph() functions from the R package *survival* (https://github.com/therneau/survival) on both the TCGA LUAD and LUSC datasets in this study, with no significant p-values from the cox.zph test showing deviance from these assumptions. In addition, visual inspection of the residual plots derived from the output of these tests confirmed that the proportional hazards assumption was met.

Data & Software Availability

This study was conducted using publicly-available TCGA NSCLC data downloaded via cBioPortal (https://www.cbioportal.org/). A list of the 865 genes used in the NAG panel is available at https://github.com/TransAnalytics/cancer_prog_prediction , developed by Dennis Wang, along with the code used to download the mutation, expression and clinical data from cBioPortal and Firehose, stratify patients based on this data and carry out statistical analyses.

Results

Validation of PDX gene panel in TCGA Datasets

Tumour mutation burden (TMB) was classified based on the number of genes in the PDX-derived 865-gene panel with nonsynonymous mutations and copy number alterations, termed the number of altered genes (NAG). To test the mutation burden together with the expression signature as a prognostic classifier for NSCLC, we attempted to predict the overall survival of 221 lung adenocarcinoma (LUAD) and 173 lung squamous cell carcinoma (LUSC) cases from TCGA. The focus of this study was on NSCLC more broadly rather than its histopathology-defined subclasses LUAD and LUSC. This focus reflected the fact that current immunotherapeutic treatments being trialled are for NSCLC as a whole rather than a specific histopathological subclass, so prognostic value needed to be established more generally. This has the added value that histopathological subclasses do not need to be known for the genomic prognostics to be applied. While the LUAD and LUSC TCGA cohorts showed different distributions for the proportions of samples at different NAG classifiers (100% of stage I LUAD samples had high NAG vs. 66% of stage I LUSC samples, **Table 2.1**, Chi-squared $p = 1.49 \times 10^{-11}$), multivariate statistical analyses did not show any significant differences in survival between the histopathological groups (Wald test $p = 0.1615$) and there were no

significant differences in the stage distribution between histopathologies (54.5/56.6% Stage I in LUAD/LUSC respectively, Chi-squared p = 0.681), and survival analyses separated patients based upon tumour stage and NAG in any case. However, given the differences in clinical parameters between LUAD and LUSC tumours, future studies could benefit from analysing genomic prognostics in a LUAD- or LUSC-specific manner with an aim towards providing more precise prognostic information to NSCLC patients whose histopathological subclass is already known. We counted the NAG among the 865 genes for each TCGA patient tumour, then classified the patients as having a high NAG if they had at least 87 altered genes, and low NAG otherwise. Dennis Wang classified the 36 patient tumours from which the PDXs were derived into two groups, according to their expression of 23 genes that were differentially expressed between the high NAG and low NAG group and which correlated with OS, using a binary cut-off corresponding to the highest risk quartile of tumours. To derive these 23 genes, differential expression analysis was used to identify 79 genes that were differentially expressed with a >2 absolute fold change between the two groups in the cohort of 36 patients, which was further reduced to 23 differentially-expressed genes when only considering those which had non-zero coefficients under penalised regression fitted to OS in this cohort. The expression profiles of the 79 genes were fitted to the OS by Dennis Wang using a separate ElasticNet model with alpha = 0.1 and a lambda parameter selected from 5-fold cross-validated likelihood. The coefficients from this ElasticNet model, which formed the 23-gene prognostic classifier, were summed to create an expression risk score for each patient. We used this mRNA expression signature, with the same cut-offs, to classify TCGA patient tumours into two groups (Wang et al. 2019). Patients with a high NAG or a low expression risk score had a relatively better prognosis (HR = 0.575, 95% CI = 0.382–0.870, p = 0.0075); this also held true for stage I patients (HR = 0.391, 95% CI = 0.222–0.685, $p = 6.9 \times 10^{-4}$). Multivariate survival analyses of stage I patients adjusted for age, sex, smoking history and histology confirmed that NAG was an independent predictor of good OS (HR = 0.407, 95% CI = 0.211–0.787, $p = 7.7 \times 10^{-3}$, **Table 2.2**).

Variation	Low NAG	High NAG	High NAG + Neoantigen	All
Number of Patients	33	147	37	217
Age (Median)	71	68	66	68
Female	8 (24%)	77 (52%)	17 (46%)	102 (47%)
Male	25 (76%)	70 (48%)	20 (54%)	115 (53%)
Stage IA	7 (21%)	59 (40%)	12 (32%)	78 (36%)
Stage IB	26 (79%)	87 (59%)	24 (65%)	137 (63%)
Adenocarcinoma	0 (0%)	95 (65%)	25 (68%)	120 (55%)
Squamous cell	33 (100%)	52 (35%)	12 (32%)	97 (45%)
Non-smoker (Lifelong)	0 (0%)	17 (12%)	3 (8%)	20 (9%)
Adjuvant Therapy	0 (0%)	5 (3%)	3 (8%)	8 (4%)
Survival past 5 years	16 (48%)	100 (68%)	34 (92%)	150 (69%)

Table 2.1: Clinical data from TCGA NSCLC stage I patients across subpopulations.

Variation	**HR**	**95% CI**	**Wald Test *p*-Value**
Age (>65 vs. ≤65)	1.04	0.62–1.73	0.89
Sex (F vs. M)	0.85	0.51–1.43	0.54
Tobacco (smoker vs. never)	1.75	0.74–4.12	0.2
Histology (adeno vs. squamous)	0.96	0.52–1.79	0.91
Overall NAG score (high vs. low burden)	0.407	0.211–0.787	0.0077 *

Table 2.2: Multivariate survival model of NAG score with clinical-pathological factors of TCGA NSCLC stage I patients. NAG, number of altered genes; HR, hazard ratio; CI, confidence interval; * designates significance at $p \leq 0.05$.

Immunogenic mutations correlate with the best survival and cytotoxic T-Cell signature

Given that TMB is associated with good survival outcome, we further hypothesized that some of the mutations in the 865 genes may induce an immune response towards the tumour through the expression of immunogenic peptides or neoantigens. We examined whether the somatic variants within the 865 genes found in TCGA NSCLC patients could result in tumour-specific antigens presented by MHC class I molecules. A subset of 86 patients had at least one expressed mutated peptide with high MHC I binding affinity against both HLA-A alleles, and this immunogenic group had significantly better OS (**Figure 2.1**; HR = 0.536, CI = 0.341–0.8419, p = 0.0068). Among NSCLC stage I patients, 47 were classified as having immunogenic mutations and had significantly better OS (**Figure 2.2A**; HR = 0.266, CI = 0.1068–0.6619, p = 0.0044), even after adjusting for other clinical factors (**Table 2.3**). In comparison, patients stratified using all possible neoantigens, beyond those in the panel of 865 genes, had no significant difference ($p < 0.01$) in overall survival (**Table 2.4**). By combining this immunogenicity classifier with the TMB based on NAG, we were able to identify patients with extremely good prognosis. Stage I patients with both immunogenic and high NAG tumours had significantly better survival than patients with either low or high NAG alone (**Figure 2.2B**; HR = 0.0807, CI = 0.02315–0.2813, $p = 7.8 \times 10^{-5}$). In order to exclude the possibility that the improved survival in patients with immunogenic tumours in the high NAG group, compared to those who were in the high NAG group without immunogenic tumours, was not due to the former group being composed of a subset of the high NAG group with even higher numbers of altered genes, the distributions of NAG values in tumours from both groups were compared with a Mann-Whitney U test, using the wilcox.test() function in R. This test was appropriate because the NAG values were not normally-distributed in either of the two groups and were integer values, so ties in NAG values were common. The Mann-Whitney U test did not find any significant differences in the NAG values between these groups (p = 0.119) which could have accounted for differences in survival.

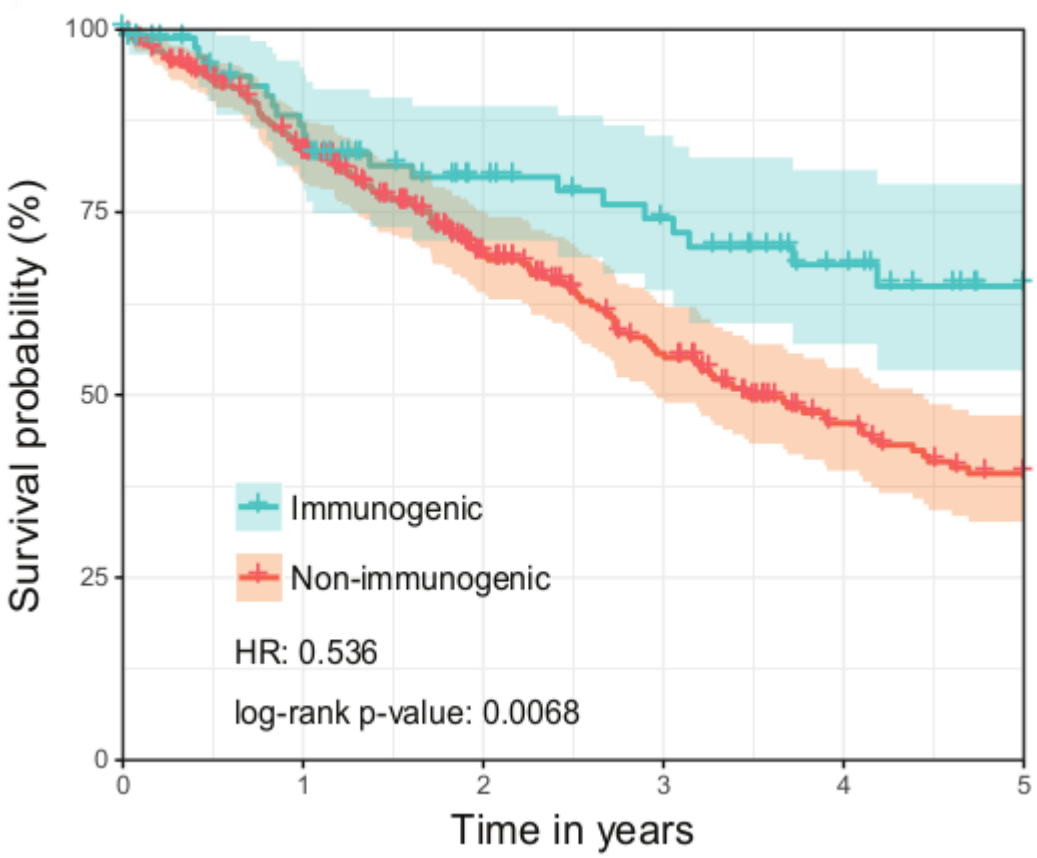


Figure 2.1: Effect of immunogenic peptides in patients (all stages). Overall survival of 86 TCGA NSCLC patients with immunogenic neoantigens (teal) and 310 patients without immunogenic peptides (red).

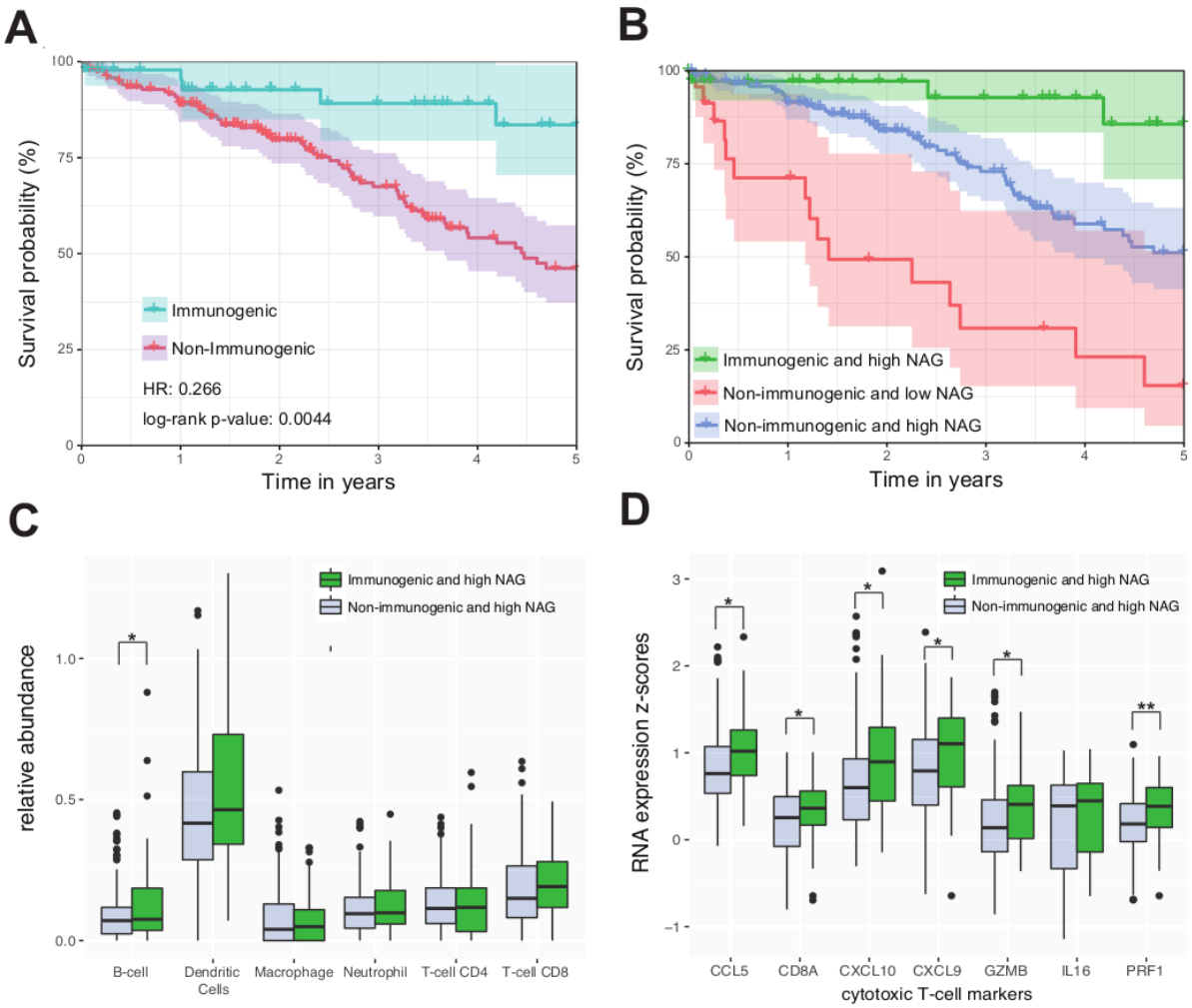


Figure 2.2: TCGA NSCLC patients stratified by immunogenic neoantigens. (**A**) Stage I patients (47 with neoantigens, 170 without neoantigens) are grouped into those with immunogenic neoantigens and those with none. Hazard ratios and log-rank *p*-values compare immunogenic patient tumours to non-immunogenic patient tumours. (**B**) The overall survival differences of stage I patients were classified based on the presence of immunogenic neoantigens and the number of altered genes. Immunogenic and high-NAG patient tumours (green, 37 patients) vs. non-immunogenic and low-NAG patient tumours (red, 23 patients) show HR = 0.0807, $p = 7.8 \times 10^{-5}$. Immunogenic and high-NAG tumours (green) vs. non-immunogenic and high-NAG patient tumours (blue, 147 patients) show HR = 0.229, $p = 0.013$. Non-immunogenic and high-NAG tumours (blue) vs. non-immunogenic and low-NAG patient tumours (red) show HR = 0.309, $p = 7.9 \times 10^{-5}$. High NAG patients with and without neoantigens were contrasted based on the relative abundance of immune cell types, estimated using the TIMER algorithm (Li et al. 2017) (**C**), and the RNA expression of cytotoxic T-cell markers (**D**). Significant differences for each component are marked (*t*-test *p*-value * <0.05; ** <0.01). I stratified the patients by immunogenicity for figures 2.2C and 2.2D into the two groups tested, but the figures were generated by Dennis Wang, along with the comparison of the cytotoxic T-cell markers examined.

Variation	HR	95% CI	Wald Test *p*-Value
Age of diagnosis (≤65 y vs. >65 y)	0.929	0.545–1.583	0.7865
Gender (male vs. female)	0.871	0.524–1.447	0.5929
Smoking history (last 15 y; yes vs. no)	0.751	0.434–1.302	0.3085
Histology (adeno vs. squamous cell)	0.677	0.392–1.169	0.1615
Immunogenicity (≥1 neoantigen vs. 0 neoantigens)	0.296	0.119–0.740	0.00919 *

Table 2.3: Multivariate survival model of the immunogenicity factor with clinical-pathological factors of TCGA NSCLC stage I patients. * designates significance at p ≤ 0.05.

Threshold	Cut-Off for Number of Neoantigen	Patient Number < Cut-Off	Patient Number ≥ Cut-Off	Hazard Ratio (95% Confidence Interval)	Log-Rank p Value
Median	1	132	86	0.487 (0.281–0.844)	0.0104
75%	3	169	49	0.369 (0.169–0.809)	0.0128
90%	8	196	22	0.545 (0.171–1.74)	0.305

Table 2.4: Survival analysis of neoantigens estimated from all coding mutations.

The RNA expression profiles from these two groups of patients were analysed further for immune components. Immune cell population estimates from TIMER (Li et al. 2017) revealed moderately, but not significantly higher proportions of B-cells, dendritic cells, and CD8 T-cells in high NAG patient tumours with immunogenic peptides compared to those without neoantigens (**Figure 2.2C**). In addition, LUAD patients with higher levels of B-cells had significantly improved 5-year OS (HR = 0.569, CI = 0.339 - 0.954, p = 0.0326) (**Figure 2.3**). Focusing specifically on markers of cytotoxic T-cells in **Figure 2.2D**, Dennis Wang found significantly higher (p < 0.05) expression of CD8A, granzyme B (*GZMB*) and perforin (*PRF1*) in high NAG tumours with immunogenic peptides. Previously, cytokines CCL5, CXCL9, CXCL10 and IL16 were found to be associated with cytotoxic T-cells (McGrail et al. 2018). We found CCL5, CXCL9 and CXCL10 to be significantly more highly expressed in our group with

immunogenic peptides. In contrast, there were no consistent differences in immune cell populations or cytotoxic T-cell markers between high and low NAG patients without immunogenic peptides (**Figure 2.4**).

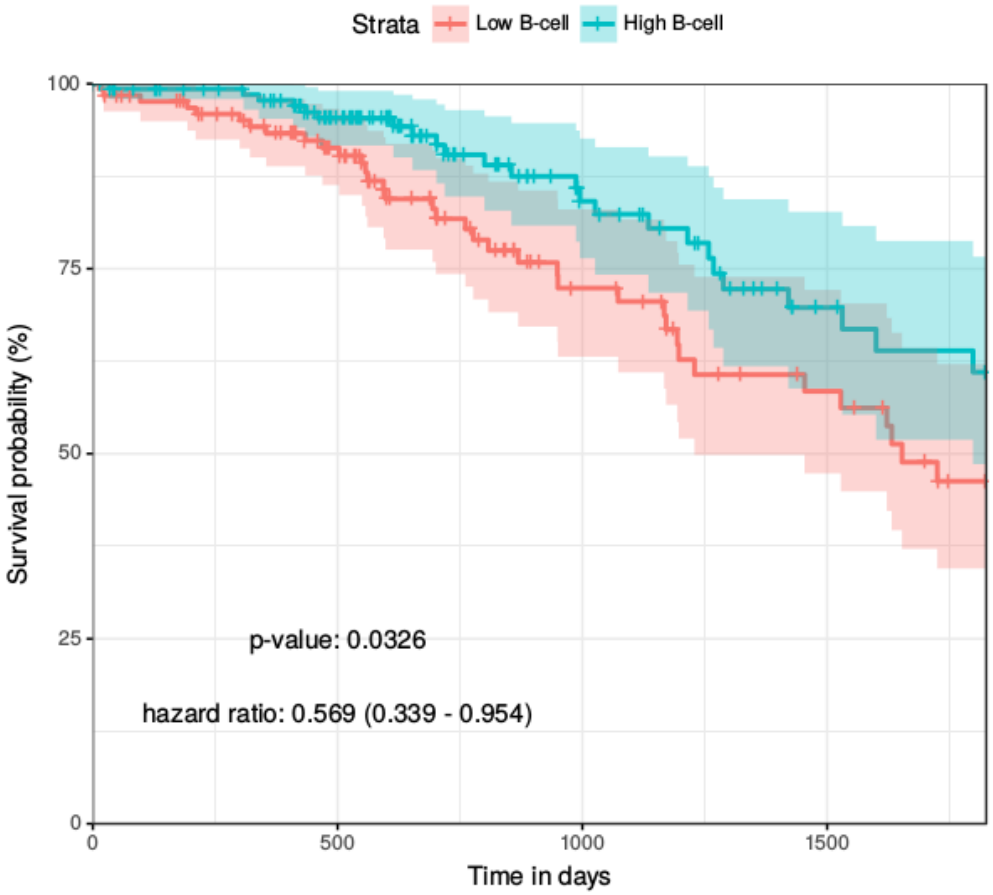


Figure 2.3: TCGA stage I LUAD patient 5-year OS stratified by B-cell abundance. Patients (n=271) are grouped into those with greater than or equal to the median B-cell abundance, and those with less than this value, showing improved survival outcomes for the high B-cell group.

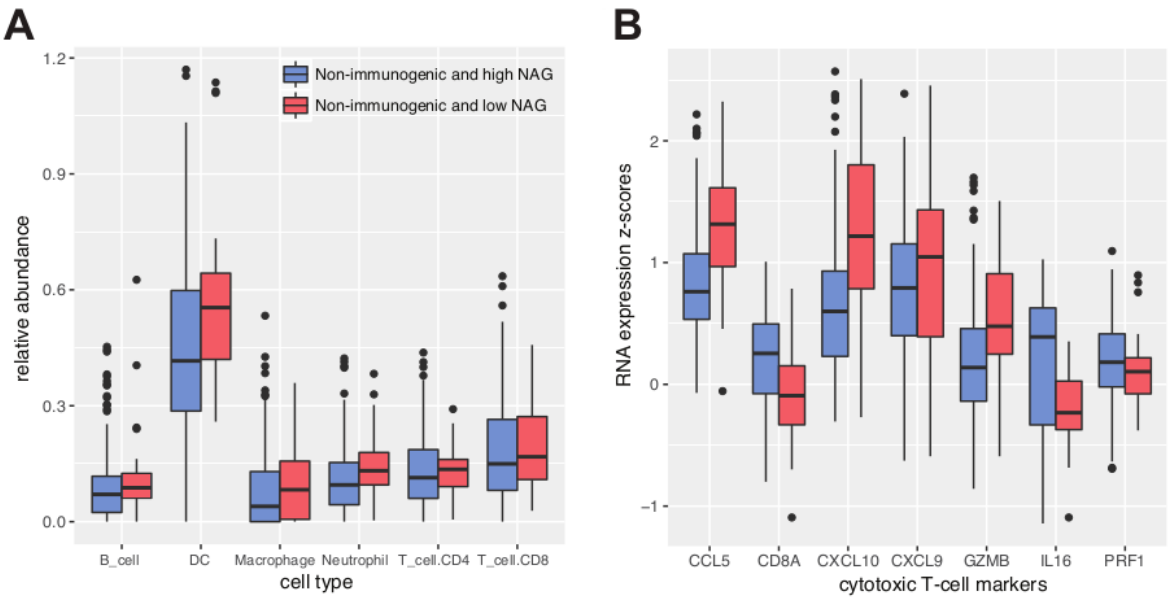


Figure 2.4: Immune cell abundance and markers. Patients without immunogenic peptides stratified into high and low NAG groups are contrasted by their relative abundance of immune cell types estimated using the TIMER algorithm (Li et al. 2017) (**A**), and the RNA expression of cytotoxic T-cell markers (**B**).

Discussion

We have developed and demonstrated two separate prognostic classifiers that are strongly associated with survival in early stage NSCLC patients, based upon mutation and neoantigenic burden in a novel panel of 865 genes. The prognostic classifiers use non-synonymous mutations as well as copy number alterations, gene expression and markers of immunogenicity such as presence and expression of neoantigenic peptides within this gene panel. Patients were classified into three distinct subpopulations with low NAG, high NAG and high NAG with neoantigens, and showed significant survival differences between all groups. These subpopulations were distributed equally in their frequency of clinical features (**Table 2.1**).

Previous investigations on prognostic classifiers of mutation and neoantigenic burden in early stage or untreated lung cancers have disputed the strength and directionality of the effect of mutation and neoantigenic burden on survival (Devarakonda et al. 2018; Owada-Ozaki et al. 2018; Yu et al. 2019). Our study utilised alternative methods, including counting altered genes in a specific panel rather than all point mutations across the genome, making it more robust against variant calling errors and differences in the distribution of variants across genes. This may have contributed to the greater survival differences found between prognostic groups in this study, improving its utility in the clinic and research. Median thresholds were chosen for dividing patients into subgroups with low/high NAG and neoantigens, showing that this approach had value even when using non-optimised thresholds. This was following standard practice employed by previous studies in this field which also used median thresholds (Miller et al. 2017; Owada-Ozaki et al. 2018; Yu et al. 2019). We also considered using machine learning approaches to optimise the thresholds for separating patients into groups with different outcomes but decided against this for two main reasons. Firstly, the sample sizes in this study were too low to get any meaningful results from machine learning, especially when considering that some subgroups of tumours with specific features were particularly uncommon, and this could have led to overfitting of thresholds since the number of samples was low to begin with, even before dividing up into training and test sets. Secondly, fitting thresholds to achieve the maximum differences between survival groups may have led to extreme thresholds for NAG and neoantigen burden resulting in very few patients being categorised in the best and worst survival groups, making the prognostic classifiers less generally useful to the broader diversity of patients. Despite this, we believe that machine

learning approaches have a lot of potential to offer in genomic diagnostics, especially in the future as sample sizes increase and more precise prognostic tools can be trained.

The clinical value of small, targeted panels of previously-established cancer driver genes for assessing TMB has been demonstrated before (Devarakonda et al. 2018; Chalmers et al. 2017). However, the panel of 865 genes in this study also incorporates non-driver genes, and our observations indicate that these contribute to patient survival outcomes. Increased TMB can be prognostic of better or worse survival depending on the tissue of origin of a tumour — cancer types with higher genomic stability, such as acute myeloid leukaemia (AML), are more likely to show no correlation or a negative correlation with patient survival (Chalmers et al. 2017) — but high TMB appears to improve survival in early-stage NSCLC patients, as well as other cancer types affecting heavily-mutated tissues, such as melanoma. The mechanism by which this occurs could be due to large numbers of mutations causing genomic instability, resulting in tumour cell death and reduced growth (Burrell et al. 2013; Belvedere et al. 2012; McFarland et al. 2017), especially since NSCLC is already characterised by a very high number of mutations compared to most other cancers, linked to the exposure of the lung tissue to higher levels of mutagenic chemicals than other internal tissues, via inhalation (Siegel, Miller, and Jemal 2019). Many of the 865 genes in our panel have cellular functions that are important for tumour cell viability and growth (Petitjean et al. 2007; Martincorena and Campbell 2015), which may explain why mutations within these decreased the rate of cancer progression to patient death.

Increased TMB raises the number of non-self mutant peptides that may bind to MHC-I, which in turn can lead to greater numbers of neoantigens, and correlates with T-cell infiltration in NSCLC and melanoma (McGranahan et al. 2016; A. Miller et al. 2017; Campesato et al. 2015; Kinkead et al. 2018). However, the relationship between TMB and T-cell infiltration has not been shown in early-stage, untreated cancers until this study, which demonstrates that TMB and neoantigen burden are both correlated with T-cell activity. TMB is sometimes used as a proxy for neoantigen burden because they are correlated and classifiers based on each give similar prognostic groupings. Both classifiers are currently used to select candidates for immunotherapy treatments such as immune checkpoint inhibitors (Nan et al. 2021; Greillier, Tomasini, and Barlesi 2018), but this study shows that combining both measures can potentially boost their individual prognostic ability. The survival differences found in this study between patients with high mutation and neoantigen burden compared to patients with low mutation and neoantigen burden ($HR = 0.0807$, $p = 7.8 \times 10^{-5}$) are far stronger than those found in previous studies, and stronger than using either classifier on its own, ranging from >85% survival over five years, to a median survival of two years. The

enhanced prognostic value of the neoantigen classifier used in this study is perhaps in part due to stricter requirements for mutations to be classified as immunogenic. Other studies (Rosenthal et al. 2019; Wood et al. 2020; A. Miller et al. 2017) have found far larger numbers of neoantigens per patient, but used weak filters, or none at all, for filtering out antigenic peptides that were unlikely to have functional effects, be expressed, or which resembled self-peptides. While the resulting neoantigenic classifier exhibited greater survival differences because of these filters, it may have resulted in some clonal or subclonal neoantigens being missed, especially those occurring in genes outside of the 865-gene panel. The mutations analysed in this study appeared unlikely to be subclonal, since they displayed high variant allele fractions (**Figure 2.5**), compared to the subclonal mutations reported to be prognostic in other studies (McGranahan et al. 2016, 2017), perhaps due in part to our investigation primarily examining tumours sequenced at an early stage, with less corresponding tumour heterogeneity.

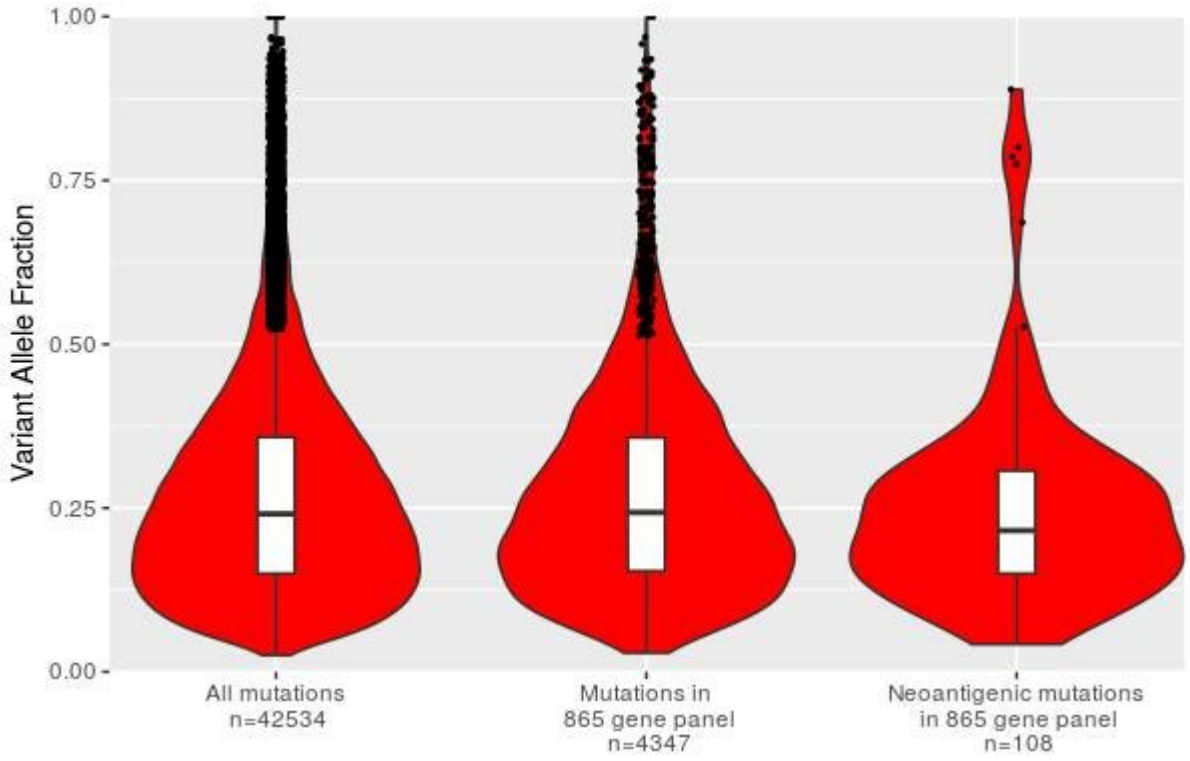


Figure 2.5: Distributions of variant allele frequencies (VAFs) of missense mutations used to stratify the TCGA NSCLC stage I patients. Three categories of mutations are included: all missense mutations, missense mutations in the 865 gene panel only, and immunogenic missense mutations in the 865 gene panel. Boxplots show the interquartile ranges and outliers are plotted as separate points. VAFs for all 3 groups were not significantly different between the three groups of mutations.

Conclusion

Developing powerful prognostic tools for cancer progression is a difficult task (Pellegrino et al. 2021), but valuable to clinicians seeking to inform and treat patients optimally. In particular, the combined value of using TMB alongside neoantigenic burden has not always been clear (McGranahan et al. 2016; Chalmers et al. 2017; Amber Miller et al. 2016). We examined prognostic classifiers in early stage, untreated NSCLC patients since this would allow earlier prognosis and response prediction, and potentially improved clinical interventions for the patient. This study improves upon methods for establishing both TMB and neoantigenic burden as prognostic tools using a novel 865 gene panel. We have demonstrated additional improvements from using both TMB and neoantigenic burden in combination, including large survival differences between the groups with the best and worst survival rates (HR = 0.0807, $p = 7.8 \times 10^{-5}$), which have not been achieved by other, similar studies (Rosenthal et al. 2019; Wood et al. 2020; A. Miller et al. 2017). The methods used here to develop genomic prognostic classifiers can be tested on other cancer types and diseases, to improve survival prediction in the clinic, and to provide further utility to researchers trying to functionally classify tumours according to their immune status, such as for optimising cancer drug targeting (van den Bulk, Verdegaal, and de Miranda 2018; Reck, Borghaei, and O'Byrne 2019).

References

Aparicio, Samuel, Manuel Hidalgo, and Andrew L. Kung. 2015. "Examining the Utility of Patient-Derived Xenograft Mouse Models." *Nature Reviews. Cancer* 15 (5): 311–16.

Belvedere, Ornella, Stefano Berri, Rebecca Chalkley, Caroline Conway, Fabio Barbone, Federica Pisa, Kenneth MacLennan, et al. 2012. "A Computational Index Derived from Whole-Genome Copy Number Analysis Is a Novel Tool for Prognosis in Early Stage Lung Squamous Cell Carcinoma." *Genomics* 99 (1): 18–24.

Buder-Bakhaya, Kristina, and Jessica C. Hassel. 2018. "Biomarkers for Clinical Benefit of Immune Checkpoint Inhibitor Treatment-A Review From the Melanoma Perspective and Beyond." *Frontiers in Immunology* 9 (June): 1474.

Bulk, Jitske van den, Els Me Verdegaal, and Noel Fcc de Miranda. 2018. "Cancer Immunotherapy: Broadening the Scope of Targetable Tumours." *Open Biology* 8 (6). https://doi.org/10.1098/rsob.180037.

Burrell, Rebecca A., Nicholas McGranahan, Jiri Bartek, and Charles Swanton. 2013. "The Causes and Consequences of Genetic Heterogeneity in Cancer Evolution." *Nature* 501 (7467): 338–45.

Calis, Jorg J. A., Matt Maybeno, Jason A. Greenbaum, Daniela Weiskopf, Aruna D. De Silva, Alessandro Sette, Can Keşmir, and Bjoern Peters. 2013. "Properties of MHC Class I Presented Peptides That Enhance Immunogenicity." *PLoS Computational Biology* 9 (10): e1003266.

Campesato, Luís Felipe, Romualdo Barroso-Sousa, Leandro Jimenez, Bruna R. Correa, Jorge Sabbaga, Paulo M. Hoff, Luiz F. L. Reis, Pedro Alexandre F. Galante, and Anamaria A. Camargo. 2015. "Comprehensive Cancer-Gene Panels Can Be Used to Estimate Mutational Load and Predict Clinical Benefit to PD-1 Blockade in Clinical Practice." *Oncotarget* 6 (33): 34221–27.

Cancer Genome Atlas Research Network. 2012. "Comprehensive Genomic Characterization

of Squamous Cell Lung Cancers." *Nature* 489 (7417): 519–25.
Cerami, Ethan, Jianjiong Gao, Ugur Dogrusoz, Benjamin E. Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, et al. 2012. "The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data." *Cancer Discovery* 2 (5): 401–4.
Chalmers, Zachary R., Caitlin F. Connelly, David Fabrizio, Laurie Gay, Siraj M. Ali, Riley Ennis, Alexa Schrock, et al. 2017. "Analysis of 100,000 Human Cancer Genomes Reveals the Landscape of Tumor Mutational Burden." *Genome Medicine* 9 (1): 34.
Charoentong, Pornpimol, Francesca Finotello, Mihaela Angelova, Clemens Mayer, Mirjana Efremova, Dietmar Rieder, Hubert Hackl, and Zlatko Trajanoski. 2017. "Pan-Cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade." *Cell Reports* 18 (1): 248–62.
Darvin, Pramod, Salman M. Toor, Varun Sasidharan Nair, and Eyad Elkord. 2018. "Immune Checkpoint Inhibitors: Recent Progress and Potential Biomarkers." *Experimental & Molecular Medicine* 50 (12): 1–11.
De Mattos-Arruda, L., M. Vazquez, F. Finotello, R. Lepore, E. Porta, J. Hundal, P. Amengual-Rigo, et al. 2020. "Neoantigen Prediction and Computational Perspectives towards Clinical Benefit: Recommendations from the ESMO Precision Medicine Working Group." *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO* 31 (8): 978–90.
Deng, Mario, Johannes Brägelmann, Ivan Kryukov, Nuno Saraiva-Agostinho, and Sven Perner. 2017. "FirebrowseR: An R Client to the Broad Institute's Firehose Pipeline." *Database: The Journal of Biological Databases and Curation* 2017 (January). https://doi.org/10.1093/database/baw160.
Devarakonda, Siddhartha, Federico Rotolo, Ming-Sound Tsao, Irena Lanc, Elisabeth Brambilla, Ashiq Masood, Ken A. Olaussen, et al. 2018. "Tumor Mutation Burden as a Biomarker in Resected Non-Small-Cell Lung Cancer." *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 36 (30): 2995–3006.
Duma, Narjust, Rafael Santana-Davila, and Julian R. Molina. 2019. "Non-Small Cell Lung Cancer: Epidemiology, Screening, Diagnosis, and Treatment." *Mayo Clinic Proceedings. Mayo Clinic* 94 (8): 1623–40.
Ellis, Peter M., Normand Blais, Dennis Soulieres, Diana N. Ionescu, Meenakshi Kashyap, Geoff Liu, Barb Melosky, et al. 2011. "A Systematic Review and Canadian Consensus Recommendations on the Use of Biomarkers in the Treatment of Non-Small Cell Lung Cancer." *Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer* 6 (8): 1379–91.
Freeman, Timothy M. 2016. "Neoantigen prediction for the development of immunotherapies." *University of Cambridge Computational Biology MPhil thesis*. https://app.box.com/s/dw8n7w7dy5olf8fqr0uu3rmv7my8dg3s
Gao, Jianjiong, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S. Onur Sumer, Yichao Sun, et al. 2013. "Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal." *Science Signaling* 6 (269): l1.
Greillier, Laurent, Pascale Tomasini, and Fabrice Barlesi. 2018. "The Clinical Utility of Tumor Mutational Burden in Non-Small Cell Lung Cancer." *Translational Lung Cancer Research* 7 (6): 639–46.
Heuckmann, J. M., and R. K. Thomas. 2015. "A New Generation of Cancer Genome Diagnostics for Routine Clinical Use: Overcoming the Roadblocks to Personalized Cancer Medicine." *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO* 26 (9): 1830–37.
Heydt, Carina, Jan Rehker, Roberto Pappesch, Theresa Buhl, Markus Ball, Udo Siebolts, Anja Haak, et al. 2020. "Analysis of Tumor Mutational Burden: Correlation of Five Large Gene Panels with Whole Exome Sequencing." *Scientific Reports* 10 (1): 11387.
Hidalgo, Manuel, Frederic Amant, Andrew V. Biankin, Eva Budinská, Annette T. Byrne, Carlos Caldas, Robert B. Clarke, et al. 2014. "Patient-Derived Xenograft Models: An Emerging Platform for Translational Cancer Research." *Cancer Discovery* 4 (9): 998–

1013.
Jacobsen, A. 2017. *Cgdsr: R-Based API for Accessing the MSKCC Cancer Genomics Data Server (CGDS)* (version version 1.3.0). CRAN: Taipei City, Taiwan. https://github.com/cBioPortal/cgdsr.
John, Thomas, Derek Kohler, Melania Pintilie, Naoki Yanagawa, Nhu-An Pham, Ming Li, Devang Panchal, et al. 2011. "The Ability to Form Primary Tumor Xenografts Is Predictive of Increased Risk of Disease Recurrence in Early-Stage Non-Small Cell Lung Cancer." *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 17 (1): 134–41.
Kinkead, Heather L., Alexander Hopkins, Eric Lutz, Annie A. Wu, Mark Yarchoan, Kayla Cruz, Skylar Woolman, et al. 2018. "Combining STING-Based Neoantigen-Targeted Vaccine with Checkpoint Modulators Enhances Antitumor Immunity in Murine Pancreatic Cancer." *JCI Insight* 3 (20). https://doi.org/10.1172/jci.insight.122857.
Li, Taiwen, Jingyu Fan, Binbin Wang, Nicole Traugh, Qianming Chen, Jun S. Liu, Bo Li, and X. Shirley Liu. 2017. "TIMER: A Web Server for Comprehensive Analysis of Tumor-Infiltrating Immune Cells." *Cancer Research* 77 (21): e108–10.
Martincorena, Iñigo, and Peter J. Campbell. 2015. "Somatic Mutation in Cancer and Normal Cells." *Science* 349 (6255): 1483–89.
McFarland, Christopher D., Julia A. Yaglom, Jonathan W. Wojtkowiak, Jacob G. Scott, David L. Morse, Michael Y. Sherman, and Leonid A. Mirny. 2017. "The Damaging Effect of Passenger Mutations on Cancer Progression." *Cancer Research* 77 (18): 4763–72.
McGrail, Daniel J., Lorenzo Federico, Yongsheng Li, Hui Dai, Yiling Lu, Gordon B. Mills, Song Yi, Shiaw-Yih Lin, and Nidhi Sahni. 2018. "Multi-Omics Analysis Reveals Neoantigen-Independent Immune Cell Infiltration in Copy-Number Driven Cancers." *Nature Communications* 9 (1): 1317.
McGranahan, Nicholas, Andrew J. S. Furness, Rachel Rosenthal, Sofie Ramskov, Rikke Lyngaa, Sunil Kumar Saini, Mariam Jamal-Hanjani, et al. 2016. "Clonal Neoantigens Elicit T Cell Immunoreactivity and Sensitivity to Immune Checkpoint Blockade." *Science* 351 (6280): 1463–69.
McGranahan, Nicholas, Rachel Rosenthal, Crispin T. Hiley, Andrew J. Rowan, Thomas B. K. Watkins, Gareth A. Wilson, Nicolai J. Birkbak, et al. 2017. "Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution." *Cell* 171 (6): 1259–71.e11.
Miller, A., Y. Asmann, L. Cattaneo, E. Braggio, J. Keats, D. Auclair, S. Lonial, MMRF CoMMpass Network, S. J. Russell, and A. K. Stewart. 2017. "High Somatic Mutation and Neoantigen Burden Are Correlated with Decreased Progression-Free Survival in Multiple Myeloma." *Blood Cancer Journal* 7 (9): e612.
Miller, Amber, Laura Cattaneo, Yan W. Asmann, Esteban Braggio, Jonathan J. Keats, Daniel Auclair, Sagar Lonial, Stephen J. Russell, and A. Keith Stewart. 2016. "Correlation Between Somatic Mutation Burden, Neoantigen Load and Progression Free Survival in Multiple Myeloma: Analysis of MMRF CoMMpass Study." In .
Nan, Zhang, Wang Guoqing, Yu Xiaoxu, Mi Yin, He Xin, Li Xue, and Wang Rong. 2021. "The Predictive Efficacy of Tumor Mutation Burden (TMB) on Nonsmall Cell Lung Cancer Treated by Immune Checkpoint Inhibitors: A Systematic Review and Meta-Analysis." *BioMed Research International* 2021 (March): 1780860.
Nielsen, Morten, and Massimo Andreatta. 2016. "NetMHCpan-3.0; Improved Prediction of Binding to MHC Class I Molecules Integrating Information from Multiple Receptor and Peptide Length Datasets." *Genome Medicine* 8 (1). https://doi.org/10.1186/s13073-016-0288-x.
Noguchi, Emi, Tadahiko Shien, and Hiroji Iwata. 2021. "Current Status of PD-1/PD-L1 Blockade Immunotherapy in Breast Cancer." *Japanese Journal of Clinical Oncology* 51 (3): 321–32.
O'Donnell, Timothy J., Alex Rubinsteyn, and Uri Laserson. 2020. "MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing." *Cell Systems* 11 (1): 42–48.e7.
Owada-Ozaki, Yuki, Satoshi Muto, Hironori Takagi, Takuya Inoue, Yuzuru Watanabe,

Mitsuro Fukuhara, Takumi Yamaura, et al. 2018. “Prognostic Impact of Tumor Mutation Burden in Patients With Completely Resected Non-Small Cell Lung Cancer: Brief Report.” *Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer* 13 (8): 1217–21.

Pellegrino, Sara, Rosa Fonti, Alessandro Pulcrano, and Silvana Del Vecchio. 2021. “PET-Based Volumetric Biomarkers for Risk Stratification of Non-Small Cell Lung Cancer Patients.” *Diagnostics (Basel, Switzerland)* 11 (2). https://doi.org/10.3390/diagnostics11020210.

Peters, Bjoern, and Alessandro Sette. 2005. “Generating Quantitative Models Describing the Sequence Specificity of Biological Processes with the Stabilized Matrix Method.” *BMC Bioinformatics* 6 (May): 132.

Petitjean, Audrey, Ewy Mathe, Shunsuke Kato, Chikashi Ishioka, Sean V. Tavtigian, Pierre Hainaut, and Magali Olivier. 2007. “Impact of Mutant p53 Functional Properties on TP53 Mutation Patterns and Tumor Phenotype: Lessons from Recent Developments in the IARC TP53 Database.” *Human Mutation* 28 (6): 622–29.

Pikor, Larissa A., Varune R. Ramnarine, Stephen Lam, and Wan L. Lam. 2013. “Genetic Alterations Defining NSCLC Subtypes and Their Therapeutic Implications.” *Lung Cancer* 82 (2): 179–89.

Pirker, Robert, Jose R. Pereira, Joachim von Pawel, Maciej Krzakowski, Rodryg Ramlau, Keunchil Park, Filippo de Marinis, et al. 2012. “EGFR Expression as a Predictor of Survival for First-Line Chemotherapy plus Cetuximab in Patients with Advanced Non-Small-Cell Lung Cancer: Analysis of Data from the Phase 3 FLEX Study.” *The Lancet Oncology* 13 (1): 33–42.

Qu, Jingjing, Quanhui Mei, Li Liu, Tianli Cheng, Peng Wang, Lijun Chen, and Jianying Zhou. 2021. “The Progress and Challenge of Anti-PD-1/PD-L1 Immunotherapy in Treating Non-Small Cell Lung Cancer.” *Therapeutic Advances in Medical Oncology* 13 (February): 1758835921992968.

Reck, Martin, Hossein Borghaei, and Kenneth J. O’Byrne. 2019. “Nivolumab plus Ipilimumab in Non-Small-Cell Lung Cancer.” *Future Oncology* 15 (19): 2287–2302.

Reynisson, Birkir, Bruno Alvarez, Sinu Paul, Bjoern Peters, and Morten Nielsen. 2020. “NetMHCpan-4.1 and NetMHCIIpan-4.0: Improved Predictions of MHC Antigen Presentation by Concurrent Motif Deconvolution and Integration of MS MHC Eluted Ligand Data.” *Nucleic Acids Research* 48 (W1): W449–54.

Richters, Megan M., Huiming Xia, Katie M. Campbell, William E. Gillanders, Obi L. Griffith, and Malachi Griffith. 2019. “Best Practices for Bioinformatic Characterization of Neoantigens for Clinical Utility.” *Genome Medicine* 11 (1): 56.

Rosenthal, Rachel, Elizabeth Larose Cadieux, Roberto Salgado, Maise Al Bakir, David A. Moore, Crispin T. Hiley, Tom Lund, et al. 2019. “Neoantigen-Directed Immune Escape in Lung Cancer Evolution.” *Nature* 567 (7749): 479–85.

Sidney, John, Bjoern Peters, Nicole Frahm, Christian Brander, and Alessandro Sette. 2008. “HLA Class I Supertypes: A Revised and Updated Classification.” *BMC Immunology* 9 (January): 1.

Siegel, Rebecca L., Kimberly D. Miller, Hannah E. Fuchs, and Ahmedin Jemal. 2021. “Cancer Statistics, 2021.” *CA: A Cancer Journal for Clinicians* 71 (1): 7–33.

Siegel, Rebecca L., Kimberly D. Miller, and Ahmedin Jemal. 2019. “Cancer Statistics, 2019.” *CA: A Cancer Journal for Clinicians* 69 (1): 7–34.

Szolek, András, Benjamin Schubert, Christopher Mohr, Marc Sturm, Magdalena Feldhahn, and Oliver Kohlbacher. 2014. “OptiType: Precision HLA Typing from next-Generation Sequencing Data.” *Bioinformatics* 30 (23): 3310–16.

The Cancer Genome Atlas Research Network. 2014. “Comprehensive Molecular Profiling of Lung Adenocarcinoma.” *Nature* 511 (7511): 543–50.

Wang, Dennis, Nhu-An Pham, Timothy M. Freeman, Vibha Raghavan, Roya Navab, Jonathan Chang, Chang-Qi Zhu, et al. 2019. “Somatic Alteration Burden Involving Non-Cancer Genes Predicts Prognosis in Early-Stage Non-Small Cell Lung Cancer.” *Cancers* 11 (7). https://doi.org/10.3390/cancers11071009.

Wang, Dennis, Nhu-An Pham, Jiefei Tong, Shingo Sakashita, Ghassan Allo, Lucia Kim, Naoki Yanagawa, et al. 2017. “Molecular Heterogeneity of Non-Small Cell Lung Carcinoma Patient-Derived Xenografts Closely Reflect Their Primary Tumors.” *International Journal of Cancer. Journal International Du Cancer* 140 (3): 662–73.

Whittle, James R., Michael T. Lewis, Geoffrey J. Lindeman, and Jane E. Visvader. 2015. “Patient-Derived Xenograft Models of Breast Cancer and Their Predictive Power.” *Breast Cancer Research: BCR* 17 (February): 17.

Wood, Mary A., Benjamin R. Weeder, Julianne K. David, Abhinav Nellore, and Reid F. Thompson. 2020. “Burden of Tumor Mutations, Neoepitopes, and Other Variants Are Weak Predictors of Cancer Immunotherapy Response and Overall Survival.” *Genome Medicine*. https://doi.org/10.1186/s13073-020-00729-2.

Wu, Licun, Ghassan Allo, Thomas John, Ming Li, Tetsuzo Tagawa, Isabelle Opitz, Masaki Anraku, et al. 2017. “Patient-Derived Xenograft Establishment from Human Malignant Pleural Mesothelioma.” *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 23 (4): 1060–67.

Yoshida, Go J. 2020. “Applications of Patient-Derived Tumor Xenograft Models and Tumor Organoids.” *Journal of Hematology & Oncology* 13 (1): 4.

Yu, Hui, Zhengming Chen, Karla V. Ballman, Mark A. Watson, Ramaswamy Govindan, Irena Lanc, David G. Beer, et al. 2019. “Correlation of PD-L1 Expression with Tumor Mutation Burden and Gene Signatures for Prognosis in Early-Stage Squamous Cell Lung Carcinoma.” *Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer* 14 (1): 25–36.

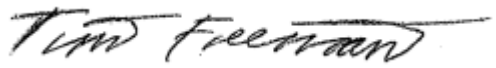
Zhang, Xiaomei, Sofie Claerhout, Aleix Prat, Lacey E. Dobrolecki, Ivana Petrovic, Qing Lai, Melissa D. Landis, et al. 2013. “A Renewable Tissue Resource of Phenotypically Stable, Biologically and Ethnically Diverse, Patient-Derived Human Breast Cancer Xenograft Models.” *Cancer Research* 73 (15): 4885–97.

Chapter III: Genomic loci susceptible to systematic sequencing bias in clinical whole genomes

As part of my PhD thesis, I am including my published paper *"Genomic loci susceptible to systematic sequencing bias in clinical whole genomes"*, which was published in *Genome Research*, DOI: 10.1101/gr.255349.119 .

Data for this analysis was provided by Personalis Inc. and the Genomics England Research Consortium. I produced all of the figures and tables in this paper, with the exception of Figure 1 (Figure 3.4 in the thesis), which was produced by Dr Dennis Wang. All of these figures represent experimental work and analyses that were performed by myself. I drafted all sections of the paper, after which feedback was provided by senior authors Dr Jason Harris and Dr Dennis Wang, which was used to edit the final manuscript.

Yours sincerely,



Timothy Freeman (PhD candidate)

Dr Dennis Wang (corresponding author)

Introduction

DNA sequencing is an imperfect process and although error rates are low, mistakes in identifying genomic variants can still occur. While the sources of random sequencing errors are relatively well understood (Benjamini and Speed 2012; Ma et al. 2019), identifying systematic errors in whole genomes sequenced in a clinical or commercial setting is not always possible due to restrictions in gathering information about the samples and sequencing processes. These errors could cause incorrect decisions on the presence or absence of disease-relevant variants in the genome and influence clinical and research decisions (Goldfeder et al. 2016).

One of the major challenges to improving variant detection is that certain regions of the genome are prone to higher rates of systematic sequencing or alignment errors, which can result in the false identification of variants at a low allelic fraction. In the case of diploid genotype calls, variants are expected to be around a 50% or 100% allelic fraction, corresponding to heterozygous and homozygous loci. In this thesis the terms locus and loci are defined as individual single-nucleotide genomic positions that may display single-nucleotide variants (SNVs), rather than larger chromosomal segments. However, real variants sometimes occur at low allelic fractions, such as somatic variants in tumors and in cases of mosaicism, where nearby cells sampled together can show genetic heterogeneity within the sample (Vattathil and Scheet 2016; King et al. 2017). In these cases, the ability to identify loci that systematically exhibit a low allelic fraction across individuals becomes critical, since these artifacts may be misidentified as variant alleles.

Lists of ‘high confidence’ loci from gold-standard reference genomes are sometimes used for quality control purposes in clinical and commercial sequencing laboratories, since they leave out regions which cannot be sequenced reliably by any technology, although they are not designed to reflect high sequencing accuracy. For example, The National Institute of Standards and Technology Genome in a Bottle Consortium (NIST GIAB) has proposed a list of ‘high confidence’ genomic regions to be used for benchmarking different sequencing methods, developed using a top-down approach, by analyzing the consensus between different sequencing technologies and variant callers for the same genomic samples to develop a ‘truth set’ of variant calls (Zook et al. 2014, 2019). However, because genomic regions only require at least one sequencing method showing no evidence of systematic error to be included in the ‘high confidence’ list, the sequencing pipeline being used by any one scientist could be a different method that is affected by systematic error, so filtering out genomic regions not in the ‘high confidence’ list does not guarantee high sequencing accuracy

in all remaining regions. This disparity between the ‘high confidence’ set and regions with high sequencing accuracy for any one sequencing pipeline is likely to increase over time as more genomic regions are included in the ‘high confidence’ set which can be sequenced by long- and linked-read-based methods but which show systematic biases with short-read-based sequencing. Another drawback is that clinically collected samples can vary in quality, and contamination may introduce variants with low allelic fractions not seen in reference genomes. Furthermore, the number of reference genomes used may be quite small, so sample-specific structural variants, which are not representative of the diversity of clinically sequenced genomes, can cause genomic regions to be missing from the ‘high confidence’ list despite having accurate sequencing for most samples.

Other top-down approaches of evaluating thresholds for allelic fraction or read quality may differ depending on the variant calling pipelines used (Sandmann et al. 2017). Benchmarking these different approaches on cohorts of genomes may be insightful for research but impractical for clinical applications, and it risks leaking sensitive genetic information. Furthermore, standard quality control measures for variant calling can often be overly simplistic, such as fixed read depth thresholds for calling variants across the genome, which are not tailored to wide regional differences in systematic biases. A reliance on high read depth for accurate variant calling increases the costs of sequencing studies, which are forced to compromise between the number of genomes sequenced and the depth of coverage achieved (1000 Genomes Project Consortium et al. 2010).

A ‘bottom-up’ approach to learning about sequencing error looks at different cohorts of clinically sequenced genomes independently and does not rely on consensus between multiple sequencing technologies. This study aims to address the limitations affecting other quality control methods described above, by developing a ‘bottom-up’ method to evaluate position-dependent systematic bias in detected allele fractions across whole genomes. We also seek to quantify its utility using results from its application to five small whole genome sequencing (WGS) noncancer cohorts, to detect how common these systematic biases are and the extent to which they affect different genomic regions and to gauge whether the systematic bias predictions from our method are supported by gold-standard reference data.

Methods

Data set 1 (Personalis, Inc.)

WGS data were obtained by Personalis, Inc. using Illumina HiSeq and the standard library prep and sequencing protocol. Paired-end reads of 100-bp length were mapped with BWA-MEM (H. Li 2013) to align reads against the GRCh37 reference human genome (**Table 3.1**). The mean depth of coverage across patients was 45×. There were 150 noncancer individuals in the cohort, including trios (infant and two parents) recruited from hospitals in the USA and a mix of ethnicities, but specific age or ethnic data were not available.

Data set	Source	Individual genomes	Genome build	Sequencing	Alignment
1	Personalis Inc. (USA)	150	GRCh37	Illumina HiSeqX	BWA-MEM 0.7.12
2	100,000 Genomes Project (UK)	215	GRCh37	Illumina HiSeqX	Isaac-aligner (SAAC00776.15.01.27)
3-5	100,000 Genomes Project (UK)	215 in each data set, with no patient overlap between data sets	GRCh38	Illumina HiSeqX	Isaac-aligner (iSAAC-03.16.02.19)

Table 3.1: Overview of cohorts whole genome sequenced and analysed.

Data set 2 (100,000 Genomes Project)

Blood samples were taken from 215 distinct, genetically-unrelated patients of mixed ethnicities with noncancer neurological diseases in each cohort, recruited from hospitals in the UK. 215 was chosen as the sample size because this gave the largest possible sample size such that the same number of samples could be achieved for each of data sets 2-5 without repeating any samples across data sets. A larger sample size would not have allowed this due to insufficient data available. The libraries were prepared using Illumina TrueSeq DNA PCR-Free Library Prep for the majority of the samples. For a small proportion of samples, when the

concentration was <20 ng/μL at Illumina, the Nano DNA library prep method was used. All main program samples were sequenced using 150-bp length paired-end sequencing on HiSeq X, and the mean depth of coverage across patients was 30×. WGS data were obtained by Genomics England's 100,000 Genomes Project, using Illumina HiSeq sequencing and mapped with the Illumina Isaac-aligner (version SAAC00776.15.01.27) (Raczy et al. 2013) to align reads against the GRCh37 reference human genome. All variants were called using Isaac Variant Caller (Starling, v2.1.4.2) (Raczy et al. 2013) and annotated using Ensembl database (v72) (Cunningham et al. 2019). The median age of patients was 13 yr (**Fig. 3.1**).

Data sets 3, 4, and 5 (100,000 Genomes Project)

No patients were in multiple data sets. Blood samples were collected and library-prepped in the same way as data set 2. WGS data were obtained by Genomics England's 100,000 Genomes Project, using Illumina HiSeq sequencing and mapped with the Illumina Isaac-aligner (iSAAC-03.16.02.19) to align reads against the GRCh38 reference human genome. The reads were aligned by Illumina with the Illumina Isaac-aligner (v03.15.09.04) and variants called by Isaac Variant Caller (Starling, v2.3.13) (Raczy et al. 2013) with annotation by Ensembl (v81) (Cunningham et al. 2019). Data sets 3, 4, and 5 all used 215 distinct, genetically-unrelated patient samples from the noncancer neurological diseases cohort, but the patients in data sets 3 and 4 were randomly sampled, so these data sets can be treated as technical replicates. For data set 5, we used data from the same cohort but selected all of the oldest patients available since the patients in data sets 3 and 4 were generally very young. The age distributions of patients in cohorts 2–5 are illustrated in **Figure 3.1** and had median ages of 13, 16, 13, and 64 yr, respectively. The numbers of patients across different ethnic categories in data sets 2–5 were recorded in **Table 3.2**.

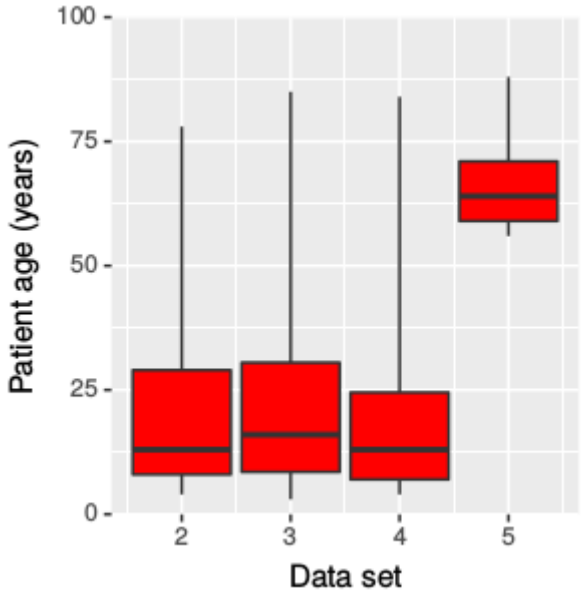


Figure 3.1: Boxplots showing the distribution (quartiles) of patient ages in data sets 2-5. This information was not available for data set 1, but we were told that the patients in that data set were all trios including two parents and an infant, so the age spread for data set 1 was ⅓ very young patients and ⅔ medium age patients. For data set 5 we intentionally selected older patients from the 100,000 Genomes Project than data sets 2-4, to identify if age had an effect on suspect locus annotation, but did not find any evidence for this.

	Data Set				
Ethnicity	**1**	**2**	**3**	**4**	**5**
Asian or Asian British: Bangladeshi	NA	0	0	1	0
Asian or Asian British: Indian	NA	6	8	6	2
Asian or Asian British: Pakistani	NA	7	7	12	4
Asian or Asian British: Any other Asian background	NA	1	2	3	2
Black or Black British: African	NA	1	0	3	1
Black or Black British: Caribbean	NA	1	2	1	3
Black or Black British: Any other Black background	NA	0	0	0	2
Mixed: White and Asian	NA	0	0	4	3
Mixed: White and Black African	NA	1	2	3	0
Mixed: White and Black Caribbean	NA	2	7	1	0
Mixed: Any other mixed background	NA	1	1	2	0
White: British	NA	149	174	152	174
White: Irish	NA	1	1	0	5
White: Any other White background	NA	10	8	12	5
Other Ethnic Groups: Any other ethnic group	NA	0	1	1	2
Not stated	NA	35	2	14	12
Total	150	215	215	215	215

Table 3.2: Distribution of ethnicities. A table showing the distribution of patient ethnicities in data sets 2-5 (this data was not available for data set 1). The number of patients for each of the 16 ethnic groups is listed along with the total number of patients for the data set. All categories are distinct and no patient is in more than one category.

Incremental Database Generation

The read depth values for each allele (A, C, G, T) at every autosomal genomic locus were calculated from aligned BAMs and divided by the total read depth at the corresponding loci — including reads that supported indels rather than A, C, G or T — to get the allelic coverage fraction, x_p, for each allele at each locus in each patient. Allelic fraction systematic biases for indels were not examined in this study. Individual IncDBs were created for each data set from the aggregate allelic fraction and standard deviation values for each allele at each locus across the entire cohort, which were calculated from x_p as described below:

Aggregate allelic fraction = $\frac{1}{N}\sum_{p=1}^{N} x_p$

Standard deviation = $\sqrt{\frac{1}{N}\sum_{p=1}^{N}(x_p - \bar{x})^2} = \sqrt{\frac{1}{N}\sum_{p=1}^{N}({x_p}^2) - (\frac{1}{N}\sum_{p=1}^{N} x_p)^2}$

N is the number of patients, *p* is the patient identifier, x_p is the allelic coverage fraction for a specific allele in patient *p*, and $\bar{x}$ is the mean of all of the allelic coverage fractions for that same allele across all patients (aggregate allelic fraction). Notice that to compute the aggregate allelic fraction, we do not store each individual's x_p values, but the sum of x_p across all individuals. Similarly, we can compute the standard deviation across individuals by storing the sum of x_p , as well as the sum of ${x_p}^2$. This approach not only removes all individual-specific genomic information, but also allows the IncDB to grow indefinitely, as more samples are sequenced and analyzed: They can simply contribute to the running sums of x_p and ${x_p}^2$. Also note that the sums in the equations above do not take up more size on disk as the number of samples increases, so the overall IncDB file size does not increase as new samples are added.

Identifying loci affected by systematic bias (suspect loci)

For each locus in all autosomal chromosomes, the standard deviation and aggregate allelic fraction values were taken from the IncDB and plotted against each other in a density plot using MATLAB 9.6, R2019a (www.mathworks.com/downloads). The main bow-shaped feature of this plot (**Fig. 3.5A–C**) was the expected result of Mendelian alleles in the human population, present at a variety of population frequencies, while observed loci with standard deviations below the 99.9% expected confidence interval for a given nucleotide were defined as suspect loci for that allele. Autosomal positions that displayed at least one suspect allele at that position were termed unique suspect loci. The total count of unique suspect loci was therefore lower than the total count of allele-specific suspect loci, since some positions had multiple suspect alleles.

The 99.9% confidence interval was estimated using Monte Carlo sampling as detailed in the pseudocode below. Monte Carlo sampling used three nested loops which respectively simulated the standard deviation at a single genomic locus between *n* individual patient allelic fractions (loop 3), 1000 times to calculate the upper and lower 99.9% confidence intervals (loop 2), for each aggregate allelic fraction from 0 to 1 in intervals of 0.01 (loop 1). The standard deviation values were recorded and used to classify suspect loci as visually illustrated (**Fig. 3.5A–C**); *n* = 150 for data set 1, and *n* = 215 for data sets 2/3. The model assumed an error rate of 0.01, corresponding to an approximation of the error rate of Illumina WGS (Wall et al. 2014). Approximately 90% of genomic reads in data set 1 had a quality score of 20 or above, corresponding to this error rate. Decreasing the assumed error rate increases the numbers of sites with very low systematic allelic biases that get annotated as suspect loci by increasing the standard deviation threshold for classification as suspect (**Fig. 3.2**). The distribution of the suspect locus allelic fractions at an assumed error rate of 0 are illustrated in **Figure 3.3**, for comparison with the allelic fractions at an assumed error rate of 0.01 (**Fig. 3.7**).

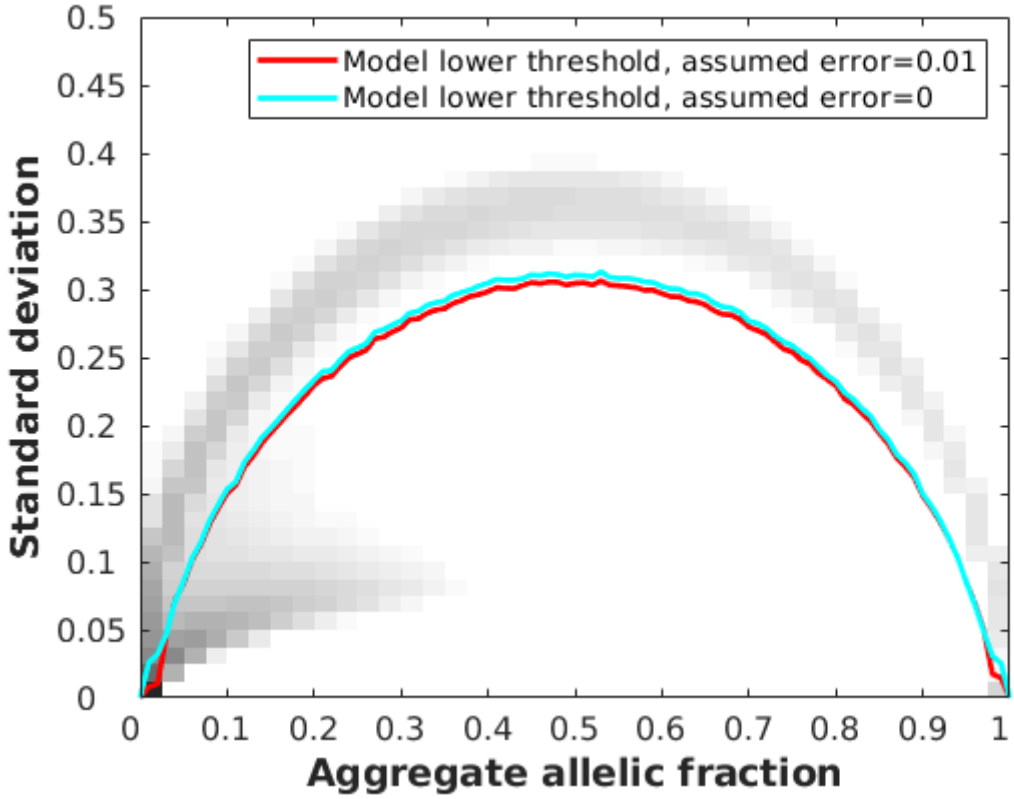


Figure 3.2: Density plot showing the distribution of aggregate allelic fraction and standard deviation values for all autosomal chromosomes in the Incremental Database for data set 1, alongside the maximum standard deviation limits used to annotate suspect loci assuming error rates of 0.01 (red) and 0 (cyan) respectively, which were obtained using Monte Carlo simulations at a 99.9% confidence interval. The greatest difference between these thresholds can be observed at aggregate allelic fractions of <0.03, which results in an increase in suspect loci from 38.7 million to 137.9 million in data set 1. However, loci at these low allelic fractions are rarely called due to low number of supporting reads, and there is only a small resulting increase (200,265 to 214,522) in the number of suspect loci with called SNVs when we reduce the error rate.

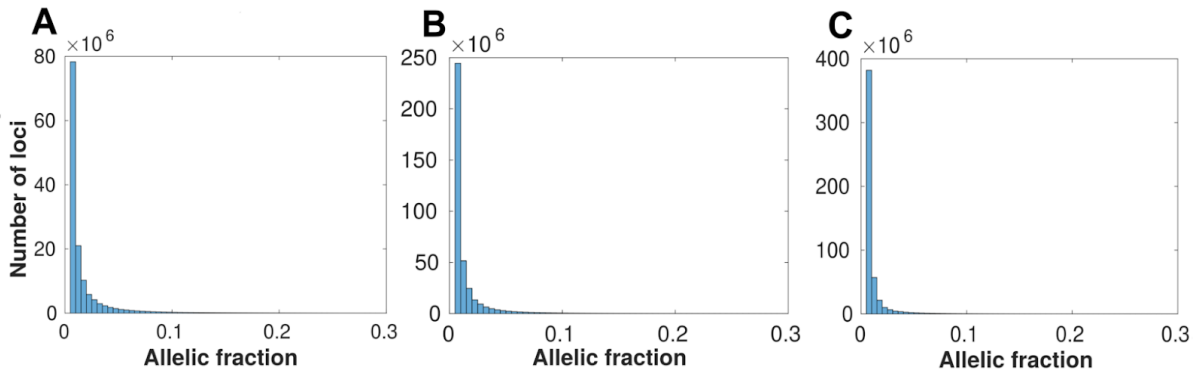


Figure 3.3: Distribution of allelic fractions observed at suspect loci in data sets 1-3 respectively (A-C), using a less conservative threshold for suspect locus classification than the default parameters used in this study (confidence interval=99.9%, assumed error rate=0). Each bin has a width of 0.005 allelic fraction and the scale is cut off at allelic fraction=0.3 because the number of suspect loci decreases to near-zero. Compared to Fig. 3.7, the number of suspect loci at very low allelic fractions (0.5-1.5%) is much higher since decreasing the assumed error rate to zero raises the maximum standard deviation limit at very low allelic fractions greatly, causing many more of these loci to fall under it, but does not greatly affect the numbers of suspect loci annotated at >3% allelic fraction.

Monte Carlo simulation of standard deviation (pseudocode)

1. For aggregate allelic fractions, AAF, from 0 to 1 in intervals of 0.01 (each representing a simulated single autosomal genomic position with that aggregate allelic fraction across all patients) do {

2. repeat 1000 times {

3. repeat for *n* simulated patients {

Randomly generate diploid genotype for each simulated patient using the binomial distribution (assuming two alleles) at the given AAF value;

Assuming a sequencing error rate of 0.01, randomly draw *c* reads from the binomial distribution to simulate observed major/minor allelic reads for the simulated biallelic diploid genotype. No positions were modelled as multiallelic;

Divide by total read depth, *c*, to get the individual allelic fractions for each patient;

}

Calculate the standard deviations between the individual allelic fractions for all *n* patients at the simulated genomic position;

}

Maximum and minimum values of 1000 repetitions mark upper and lower 99.9% confidence intervals for standard deviation at given AAF;

}

Analysis of regional enrichment of unique suspect loci

Histograms of suspect locus density across Chromosome 1 were plotted in MATLAB alongside chromosome ideograms taken from the UCSC Genome Browser (Kent et al. 2002) (GRCh37, http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/cytoBandIdeo.txt.gz; and GRCh38, http://hgdownload.cse.ucsc.edu/goldenpath/hg38/database/cytoBandIdeo.txt.gz) in order to show suspect locus density in comparison with chromosomal banding patterns.

BED files for 18 different types of genomic region were analyzed to check enrichment of unique suspect loci using a Fisher's exact test to calculate the exact significance values. Full contingency tables for all regions in data sets 1–3 (autosomal chromosomes only) are available in **Table 3.3** to show how these were calculated.

The regions tested were the NIST GIAB high-confidence regions, *Alu* repeats, GCgt70 (>70% GC content) regions, NonUnique100 regions (defined as all regions where a single 100-bp read could not map uniquely), segmental duplications, small/large homopolymers, flanking regions of small/large homopolymers, the RepeatMasker region, introns, exons, genes, intergenic regions, ClinVar short variants, and three neurological clinical panels (see next section for full list and details of BED files used).

"All sequenced regions" referred to all genomic loci where the number of aligned reads was greater than zero. There were no suspect loci outside of this region by definition, so the odds ratio was 1 by default.

Genomic region BED file sources

NIST GIAB high-confidence region (Xiao et al. 2014; Zook et al. 2014, 2019) — a selection of genomic loci covering the majority of the human genome that are considered to have high-confidence calls (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/latest/GRCh37/HG001_GRCh37_GIAB_highconf_CG-IllFB-IllGATKHC-Ion-10X-SOLID_CHROM1-X_v.3.3.2_highconf_nosomaticdel.bed).

NonUnique100—all regions where a single 100-bp read cannot map uniquely (so all stretches on the reference that are 100 bp or longer that are repeated on the GRCh37 reference).

Segmental duplication—long DNA sequences (>10 kb) that are found in multiple locations across the human genome as a result of duplications.

GCgt70, small/large homopolymers and their 100-bp flanking regions were all calculated as described below, and BED files (available at https://doi.org/10.15131/shef.data.9927053) were generated in-house rather than downloaded from another source.

GCgt70 (GC content > 70%)—regions with >70% GC content. Loci were annotated as within this region if the surrounding 100 bp around each locus had >70% GC content.

Small/large homopolymer—region of DNA containing a single nucleotide (9–19 bp for small homopolymers, ≥ 20 bp for large homopolymers).

Small/large homopolymer flanks—100-bp flanks surrounding the small/large homopolymer regions, respectively.

RepeatMasker region—a BED file containing a variety of different types of repeats (Smit, AFA, Hubley, R & Green, P 2013-2015). The open-3-2-7 version of RepeatMasker was downloaded from the UCSC Table Browser (https://genome.ucsc.edu/cgi-bin/hgTables) (Karolchik et al. 2004).

Alu repeats (Hasler and Strub 2007)—the most common type of transposable element in the human genome, of which there are over one million copies. The BED file was composed of all RepeatMasker Regions downloaded from the UCSC Table Browser (https://genome.ucsc.edu/cgi-bin/hgTables) that were annotated as *Alu* repeats in the repName column.

BED files were also downloaded for genic regions, intergenic regions, exonic regions, intronic regions (March 22, 2019), and ClinVar short variants (June 12, 2019), acquired from the UCSC Table Browser (https://genome.ucsc.edu/cgi-bin/hgTables).

Clinical panel BED files were also downloaded for the three most reviewed neurological clinical panels on PanelApp (Martin et al. 2019) (for intellectual disability [October 19, 2018], genetic epilepsy syndromes [February 7, 2019], and hereditary spastic paraplegia [February 7, 2019], respectively; https://panelapp.genomicsengland.co.uk/panels/). These clinical panels contain whole genes and are not limited to exonic regions only. Since the density of suspect loci is similar in exons and introns, the impact of looking at whole genes vs. just introns or just exons would not be expected to have much effect on suspect loci density.

Calculating allelic fractions at suspect loci in NA12878

An indexed BAM file for NA12878 Chromosome 1, sequenced using the same pipeline used in data set 1, was provided by Personalis, Inc. We used the Integrative Genomics Viewer (IGV) (Robinson et al. 2017) to examine NA12878's read pileup at data set 1 suspect locus positions at which Chromosome 1 SNVs had previously been called, to confirm that NA12878 exhibited low-fraction alleles at these positions. NA12878 was not part of any cohorts used to build the

IncDBs in this study. Chromosome 1 SNVs in NA12878 were extracted from a VCF file corresponding to the sequencing pipeline used for data set 1, which was also provided by Personalis, Inc. SNVs were classified as suspect if they corresponded to the same alleles at the same positions as suspect loci calculated for data set 1. The allelic fractions for these variants were calculated from the NA12878 BAM file using SAMtools v1.9 (http://www.htslib.org/download/) mpileup (Heng Li et al. 2009; Heng Li 2011). Variants with fewer than 10 supporting reads were deemed to have insufficient coverage and were not included.

Analyzing the proportion of gnomAD SNVs that are suspect

A list of all gnomAD variants (Karczewski et al. 2019), along with their allelic fraction and annotation as PASS-flagged or not, was obtained from a TSV file (available at https://doi.org/10.15131/shef.data.9927062). This was filtered to only include autosomal SNVs. These were classified as suspect and nonsuspect SNVs as above. Variants in gnomAD were annotated as PASS variants if they were marked this way in gnomAD v2.1 (https://macarthurlab.org/2018/10/17/gnomad-v2-1/).

Comparing sequencing quality between suspect and nonsuspect loci

The coverage of different allelic reads across all loci/nucleotide combinations on Chromosome 1 was available for data set 1 (Cov), along with the corresponding coverage of allelic reads filtered to only include reads with sequencing and mapping quality scores greater than 20 (Cov20). The filtered coverage values of allelic reads (Cov20) were divided by the corresponding unfiltered coverage values (Cov) to get the proportion of allelic reads with sequencing and mapping quality scores both greater than 20 at each locus/nucleotide combination. The cumulative distributions of these values were calculated separately for locus/nucleotide combinations that were annotated as suspect loci or nonsuspect loci.

Using suspect loci to check the quality of your own sequenced samples

To address the limitations of existing quality control procedures used in WGS and variant calling pipelines discussed in this study, we have created resources for researchers and clinicians to carry out quality control of suspect variants occurring at positions that show a consistent systematic allelic bias (see “Data access”). We have provided BED files containing all of the suspect loci that we have identified in all five data sets used for every allele, along with their corresponding aggregate allelic fractions, standard deviation between individuals’ allelic fractions, and allelic fractions corresponding to z values of −2, −1, +1, and +2. We

suggest that researchers and clinicians using sequencing pipelines that are similar to any of the three pipelines used in this study identify the corresponding BED files and check the allelic fractions of any variants that they have previously called that are present in these BED files. If the allelic fraction of a called variant is not significantly greater than would be obtained by systematic bias alone, i.e., if it is lower than the suspect allelic fraction at a *z* value of +1 for a milder filter or +2 for a stricter filter, then we would advise annotating or even removing that called variant. Researchers can also design custom filters based on their own preference if they wish to use different *z* value thresholds or other combinations of information. Researchers who are using Illumina WGS pipelines that have less in common with the pipelines shown here may also use a more conservative BED file of suspect loci to filter their data, containing only suspect loci found in both data sets 1 and 2, which is also included in the data set.

The suspect loci BED files provided by this study were only generated for specific WGS pipelines using Illumina HiSeq, and it is unlikely that these results could be used reliably for quality control of sequencing and alignment pipelines that are not similar to these. We would therefore recommend that researchers who have access to sequence data from a large cohort of patients should develop their own IncDB and calculate suspect loci for their sequencing pipelines based on that. The drawback of this, as opposed to using the suspect loci BED files above, is that this is computationally demanding and requires that researchers have access to large sequence data sets for their pipeline.

Data access

Links to the sources of all nonconfidential data used in this article are referenced in the text at first mention and also below. We do not provide public links to download the raw sequence data we used in this study to generate the IncDBs for individual genome sequences since these are confidential. Data set 1 individuals were sequenced for commercial purposes. While raw sequence reads are no longer accessible for data set 1, we provide summary statistics from the BAM files stored in the Incremental Database (link below). In order to protect participants in data sets 2 and 3, the 100,000 Genomes Project data can only be accessed through a secure research environment. To access the data for research, you must be a member of the Genomics England Clinical Interpretation Partnership (GeCIP), the research community set up to analyze the Project data. You can apply to join the GeCIP (https://www.genomicsengland.co.uk/join-a-gecip-domain/). To be eligible for data access, you must have your affiliation verified at an institution that has signed the GeCIP Participation Agreement and have your application to join a valid GeCIP domain accepted by the domain lead or a member of the GeCIP team. The Incremental Database containing the variant summary statistics at all loci for each data set is available for download at

https://doi.org/10.15131/shef.data.9995945, and the gnomAD TSV file generated in this study is available for download at https://doi.org/10.15131/shef.data.9927062. The code used to generate and analyze the IncDBs used in this study is publicly available at GitHub (https://github.com/tmfreeman400/IncDB_code) and as Supplemental Code.

Results

Biased variant allelic fractions occur across the genome and are persistent across individuals

Five separate ethnically mixed patient cohorts in the USA and UK (**Table 3.1**) were previously sequenced under confidential conditions using distinct WGS pipelines prior to this study by Personalis, Inc. and the 100,000 Genomes Project. Incremental Databases (IncDBs) were generated using these data, containing between-patient standard deviation of allelic fractions for each autosomal genomic locus and also the corresponding aggregate allelic fraction values (**Fig. 3.4**), without maintaining any sample-specific information. This ensured that individuals remained anonymous during the quality control process of examining reads using an IncDB. An aggregate allelic fraction was defined as the mean of all of the allelic fractions across the patient cohort for that specific allele at a specific single-nucleotide genomic position. All patients were noncancer patients, and in addition, patients in data sets 2–5 all had neurological disorders. The main differences between data sets were the use of the reference genome GRCh37 for alignment in data sets 1 and 2, while GRCh38 was used for data sets 3–5, and the use of BWA-MEM (H. Li 2013) for alignment in data set 1, while the Isaac-aligner (Raczy et al. 2013) was used in data sets 2–5. Since clinical sequencing facilities will continually upgrade their sequencing pipelines with new versions of algorithms and genome builds as time goes on, we wanted to make comparisons between different sequencing protocols in different cohorts of patients. In order to verify our findings, we also sequenced two additional cohorts (data sets 4 and 5) using exactly the same protocol as data set 3.

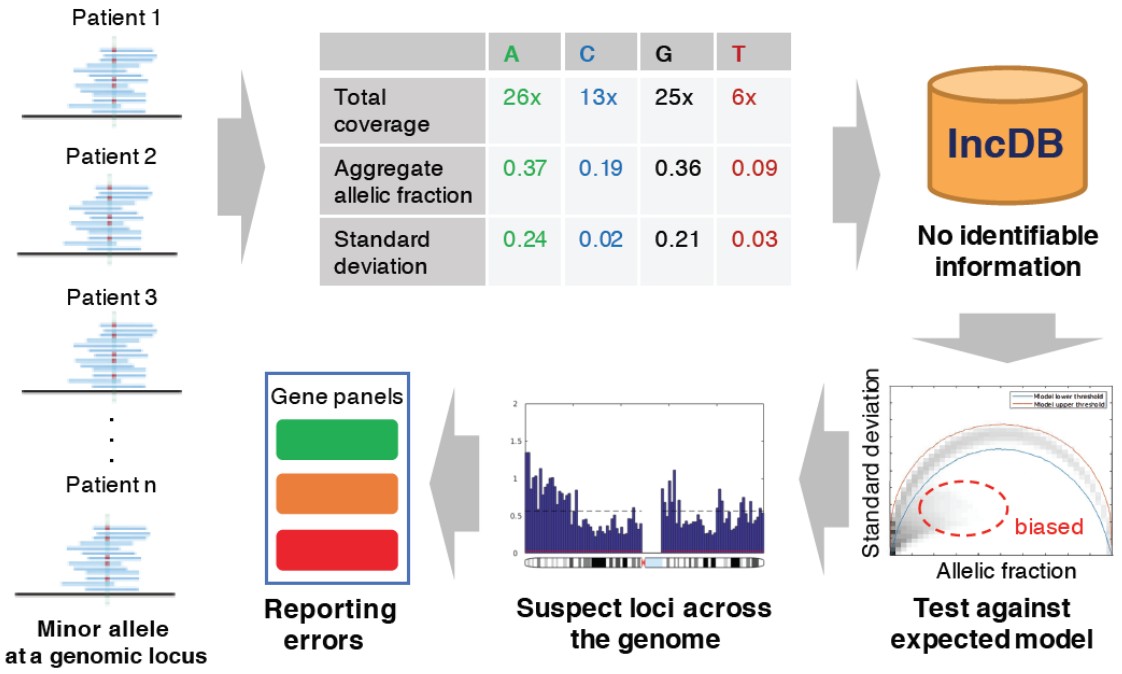


Figure 3.4: Approach for detecting loci with systematic sequence bias (Created by Dennis Wang): Alternative allele fractions are collected from a cohort of individuals at every locus and aggregated into allele-specific summary statistics. These aggregate allelic fractions and standard deviations at each genomic position are stored in the Incremental Database (IncDB), which does not contain any patient-specific information. The 99.9% confidence interval for expected standard deviation at each allelic fraction was generated. Genomic positions where the observed standard deviation was below the confidence interval expected were catalogued as "suspect loci", and mapped to variant calls in clinically relevant genes. Prioritisation of genes for diagnostic and reporting purposes can be adjusted according to the presence of suspect loci.

The observed relationship between standard deviation and the aggregate allelic fraction at each genomic locus was compared to the expected distribution assuming inherited variants in Hardy-Weinberg equilibrium in Figure 3.5A–C. We labeled the positions which fell below the 99.9% confidence interval of the Mendelian model as 'suspect loci'; ~1%–3% of all autosomal loci were suspect loci for at least one allele, which we define as unique suspect loci in this study. The upper and lower 99.9% confidence intervals of the expected Mendelian model are illustrated in Figure 3.6. In all five data sets, suspect loci are mostly a low allelic fraction (up to ~40%) but with much lower standard deviations across samples compared to Mendelian variants with the same allelic fractions. The distribution of allelic fractions at suspect loci are illustrated in Figure 3.7 for data sets 1–3. Unique suspect loci were counted across data sets 1 and 2, and the overlap revealed that most unique suspect loci were not shared between both pipelines (**Fig. 3.5D**). Single nucleotide variants (SNVs) that were called in the gold-standard reference NA12878, an individual sample separate from all of the data sets

used, were analyzed to check if they validated both suspect locus positions and their corresponding suspect alleles in data sets 1 and 2 (**Fig. 3.5E**). There were 147,000 SNVs reported in NA12878 that were also annotated as suspect in both data sets 1 and 2, which corresponded to an overlap proportion roughly twice as high as in **Figure 3.5D**, showing that suspect variant calls were more frequent at suspect loci present in both data sets than in either one data set alone.

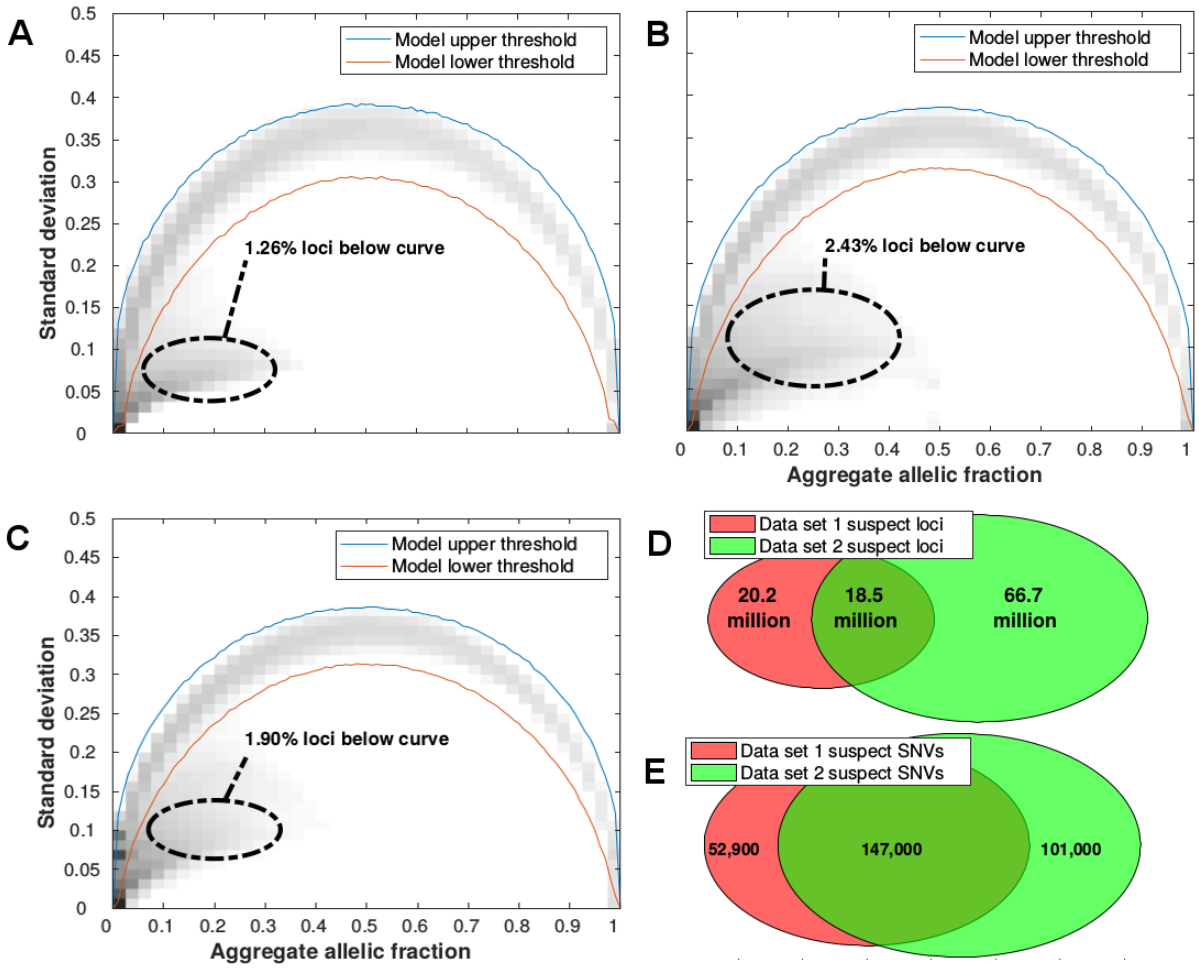


Figure 3.5: Identification of suspect autosomal loci/allele combinations with persistent low allelic fractions across patients. Observed and expected alternative allele fractions were estimated from five different whole genome sequenced cohorts: (**A**) Data set 1, (**B**) Data set 2, (**C**) Data set 3. For all loci in autosomal chromosomes, the standard deviation and aggregate allelic fraction values from the IncDB were plotted against each other in a density plot. The darker regions have the highest concentration of loci. The red lines indicate the upper and lower boundaries of the 99.9% confidence interval for the expected allelic standard deviation (shown in **Fig. 3.6A,B** respectively). Suspect loci for each cohort were defined as the loci with standard deviation below their simulated model lower threshold. A total of 2.8 billion autosomal loci were assessed. (**D**) Venn diagram showing the overlap of all suspect loci between data sets 1 and 2. (**E**) Venn diagram of overlap at suspect loci where SNVs

have been called. A total of 3.44 million autosomal SNVs were not annotated as suspect in either data set 1 or 2.

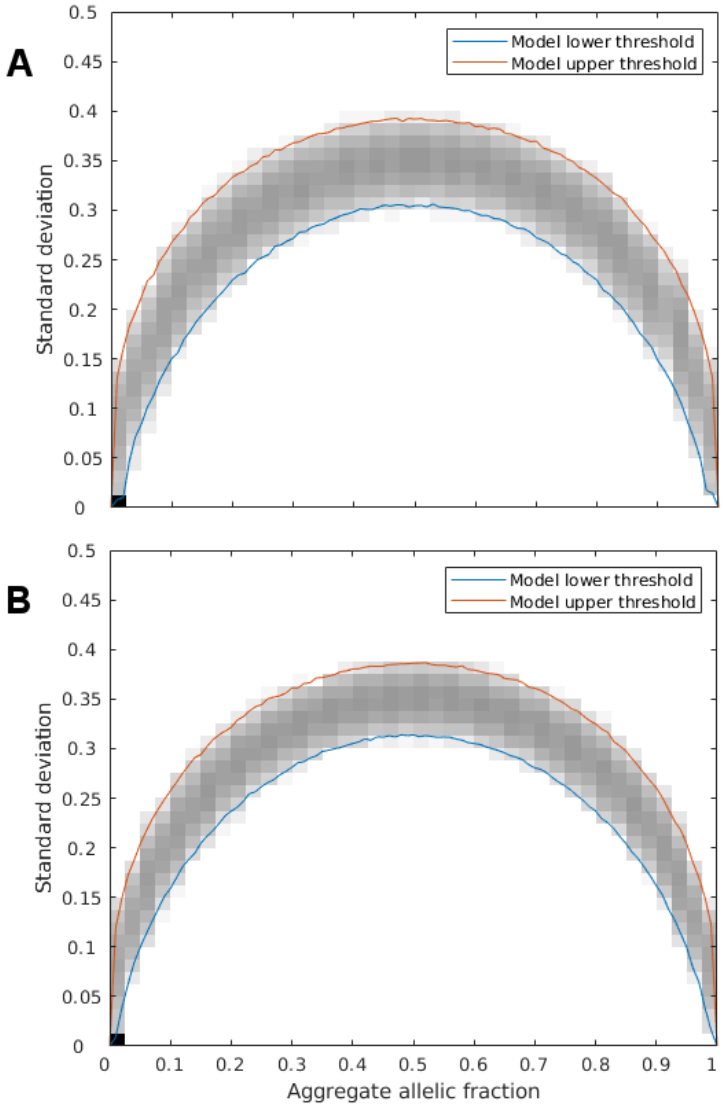


Figure 3.6: Monte Carlo random sampling was used to generate the Hardy-Weinberg expected range of standard deviation at each allelic fraction with 1000 repetitions at each. The standard deviation and aggregate allelic fraction for these simulated values are plotted on the above density plot. The maximum and minimum boundaries of these values are marked in red and blue respectively and represent the upper and lower 99.9% confidence interval for all autosomal chromosomes with a population of 150 patients (**A**), corresponding to data set 1, or 215 patients (**B**), corresponding to data sets 2 and 3.

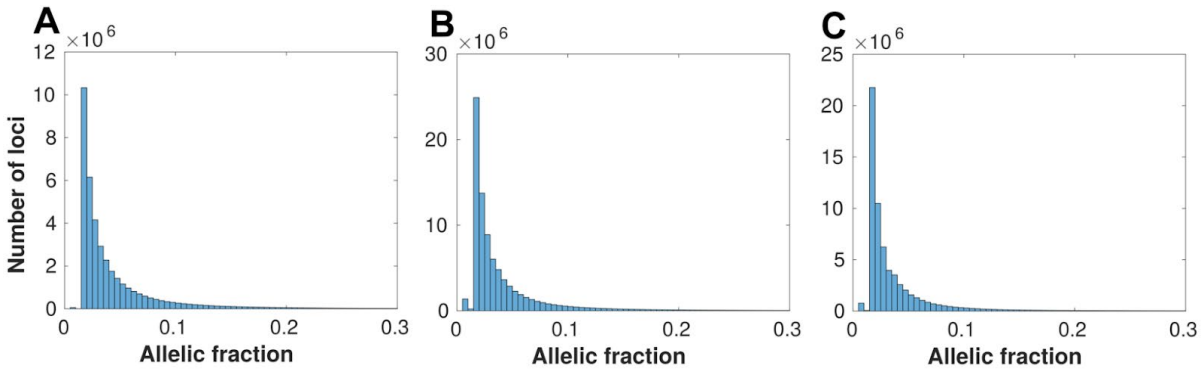


Figure 3.7: Allelic fractions observed at suspect loci in data sets 1-3 respectively (A-C), using the default suspect locus classification parameters for this study (confidence interval=99.9%, assumed error rate=0.01). Each bin has a width of 0.005 allelic fraction and the scale is cut off at allelic fraction=0.3 because the number of suspect loci decreases to near-zero. The allele fractions of the systematic errors were generally very small and heavily right-skewed (median= 0.0282, 0.0261, 0.0241; 95% confidence interval = 0.0154-0.1884, 0.0151-0.2059, 0.0151-0.1770 for data sets 1-3 respectively).

Enrichment of suspect loci within specific genomic regions

Unique suspect loci were found to be present across the entire sequenced autosomal genome and only absent in unsequenced sections, although their prevalence varied across sequenced regions both locally and showing larger trends across chromosomes (**Fig. 3.8**). We examined the distribution of unique suspect loci across different regions of the genome (**Fig. 3.9**) and recorded the regional enrichment of unique suspect loci (**Table 3.5**) using odds ratios (ORs). All odds ratios calculated were highly statistically significant due to the very large number of genomic positions sampled, even when the odds ratios were close to 1. The 95% confidence interval lower and upper bounds were both equal to the reported odds ratios to three significant figures in all cases. The highest/least significant P-value recorded was for the enrichment of suspect loci in the intellectual disability gene panel in data set 2 (OR = 1.01, $P = 8.69 \times 10^{-41}$). All other P-values ranged from 10^{-322} to 10^{-79}.

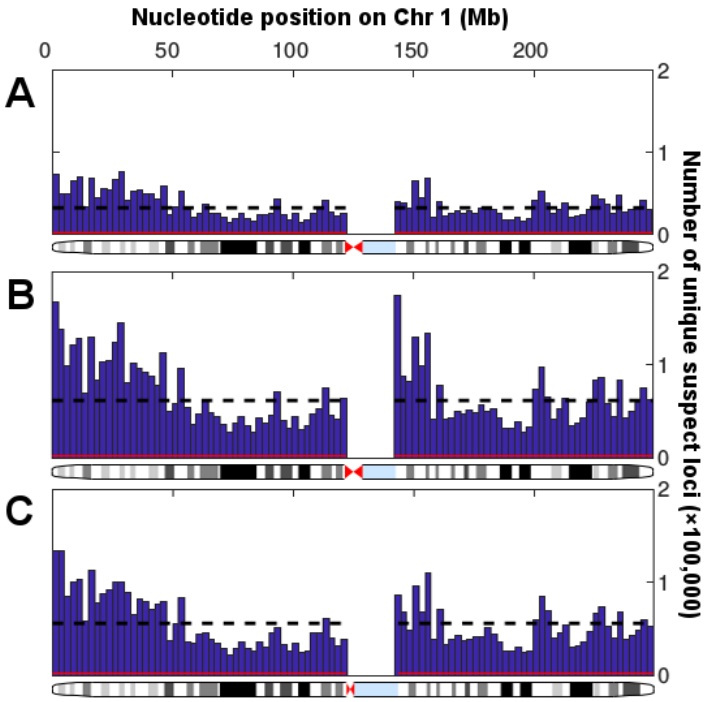


Figure 3.8: Variability in distribution of unique suspect loci in sequenced regions of Chromosome 1. Histograms show the numbers of suspect loci in 100 regular intervals across Chromosome 1, with the number of suspect loci per 2.49 million bp bin on the y axis and the nucleotide position on the x axis. There were no suspect loci at the centromere since this could not be sequenced. The black dotted line shows the mean number of suspect loci per bin, while the red line shows the number of suspect loci in each bin that would be expected by chance (1 per 1000 loci). (**A**) Data set 1, Personalis IncDB and GRCh37. (**B)** Data set 2, 100,000 Genomes Project and GRCh37. (**C**) Data set 3, 100,000 Genomes Project and GRCh38.

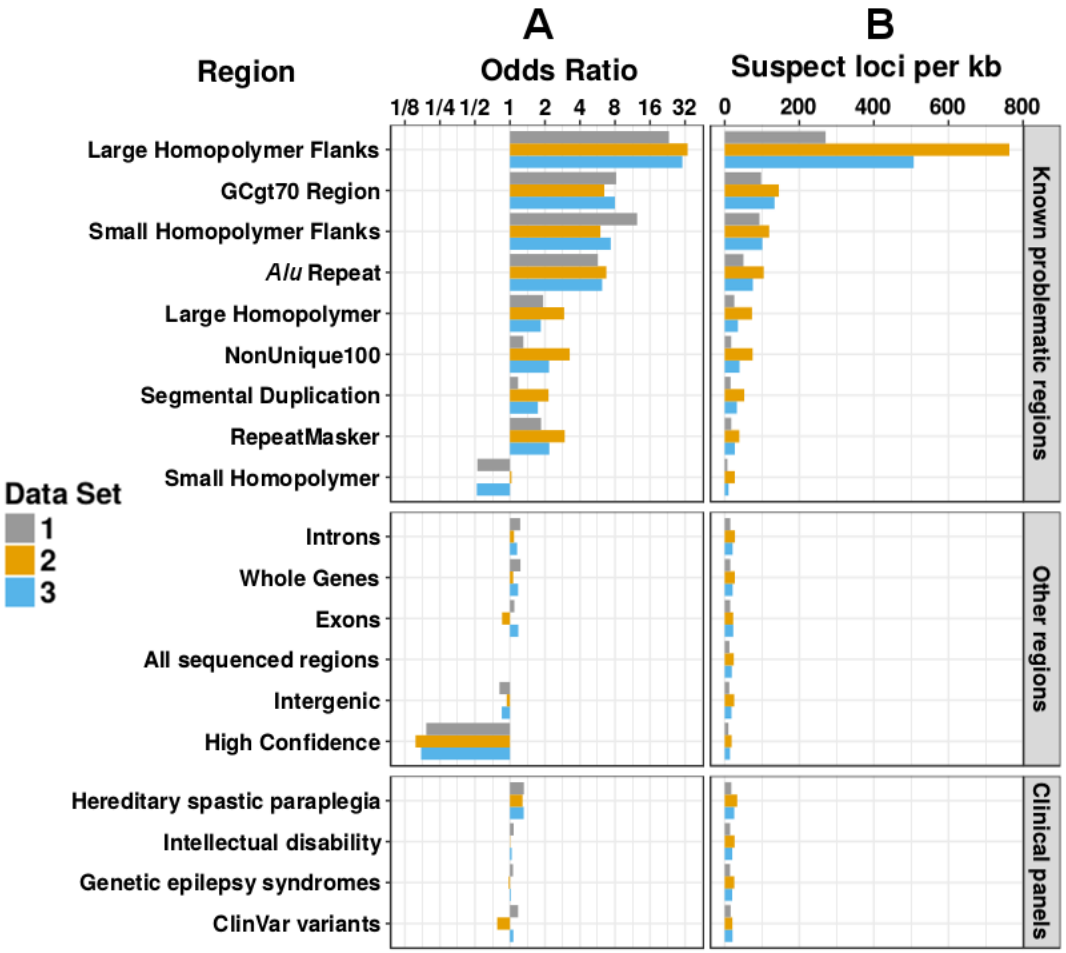


Figure 3.9: Enrichment of unique suspect loci in different types of genomic regions. All odds ratios shown are statistically significant and equivalent to the 95% upper and lower confidence intervals to 3 significant figures. (**A**) Log_2 scale barplot showing the odds ratios for regional enrichment of unique suspect loci across all autosomal chromosomes from three different data sets. All odds ratios were calculated and shown to be significant using Fisher's exact test. Regions were compared to "All sequenced regions" (odds ratio of 1). (**B**) Barplot showing the number of unique suspect loci per kb across all autosomal chromosomes from three different data sets. Although clinical and high confidence regions had a lower rate of suspect loci per kb, over the entire genome they still contained a large number of suspect loci overall in high confidence regions (24.1/45.1/34.6 million in data sets 1, 2 and 3) and clinical regions (2.55/4.64/3.55 million in data sets 1, 2 and 3 within the intellectual disability panel alone).

	Number of suspect loci (x1,000,000)			**Number of non-suspect loci (x1,000,000)**			Odds ratio		
Data set	1	2	3	1	2	3	1	2	3
Inside GC-rich regions	1.59	2.31	2.12	14.7	13.6	13.8	8.19	6.47	7.99
Outside GC-rich regions	34.8	67.8	50.1	2630	2580	2600			
Inside Alu Repeat regions	14.3	29.6	21.3	271	253	262	5.70	6.74	6.21
Outside Alu Repeat regions	22	40.5	30.9	2370	2340	2350			
Inside large homopolymer regions	0.0867	0.238	0.107	3.28	3.02	2.89	1.93	2.92	1.84
Outside large homopolymer regions	36.3	69.8	52.1	2640	2590	2610			
Inside large homopolymer flanks	5.79	11.6	9.02	21.4	15.2	17.8	23.29	33.69	30.37
Outside large homopolymer flanks	30.6	58.4	43.2	2630	2580	2590			
Inside NonUnique100 regions	2.21	8.55	4.54	125	106	110	1.30	3.26	2.17
Outside NonUnique100 regions	34.2	61.5	47.7	2520	2490	2500			

Inside segmental duplication	2.08	6.31	3.93	130	114	117	1.17	2.15	1.74
Outside segmental duplication	34.3	63.8	48.3	2520	2480	2490			
Inside RepeatMasker regions	23.5	51.9	34.8	1320	1270	1250	1.85	2.96	2.18
Outside RepeatMasker regions	12.9	18.2	17.5	1330	1320	1360			
Inside small homopolymer regions	0.0918	0.321	0.0976	12.7	11.5	9.38	0.53	1.03	0.52
Outside small homopolymer regions	36.3	69.8	52.1	2640	2580	2600			
Inside small homopolymer flanks	18	22.3	19.2	193	187	190	12.40	5.99	7.36
Outside small homopolymer flanks	18.4	47.8	33.1	2460	2400	2420			
Inside introns	18	32.7	25.1	1180	1160	1160	1.22	1.09	1.15
Outside introns	18.4	37.4	27.1	1470	1440	1450			
Inside whole genes	19.2	34.5	27	1260	1230	1240	1.23	1.07	1.18
Outside whole genes	17.2	35.6	25.3	1390	1360	1370			

Inside exons	1.2	1.84	1.86	80.5	79.2	79.2	1.09	0.86	1.18
Outside exons	35.2	68.2	50.4	2570	2510	2530			
Inside intergenic regions	17.2	35.6	25.3	1390	1360	1370	0.81	0.94	0.85
Outside intergenic regions	19.2	34.5	27	1260	1230	1240			
Inside high confidence regions	24.1	45.1	34.6	2410	2390	2400	0.19	0.15	0.17
Outside high confidence regions	12.2	25	17.7	234	204	211			
Inside hereditary spastic paraplegia gene panel	0.134	0.253	0.193	7.43	7.29	7.35	1.32	1.28	1.31
Outside hereditary spastic paraplegia gene panel	36.2	69.8	52	2640	2590	2600			
Inside intellectual disability gene panel	2.55	4.64	3.55	174	171	172	1.08	1.01	1.04
Outside intellectual disability gene panel	33.8	65.4	48.7	2470	2420	2440			

Inside genetic epilepsy syndromes gene panel	0.754	1.34	1.04	51.7	50.9	51.2	1.06	0.98	1.02
Outside genetic epilepsy syndromes gene panel	35.6	68.7	51.2	2600	2540	2560			
Inside clinVar short variants	0.00779	0.0101	0.0103	0.483	0.479	0.479	1.18	0.78	1.07
Outside clinVar short variants	36.4	70.1	52.2	2650	2590	2610			

Table 3.3: Fisher Exact contingency table. The full contingency table for all regions across data sets 1-3, showing these statistics for all 18 types of region examined in this study. P-values are not shown since these were extremely low and statistically significant in all cases. The largest p-value was 8.69×10^{-41}, for the enrichment of suspect loci in the intellectual disability gene panel in data set 2, which was still highly statistically significant. This high level of statistical significance was expected due to the very large sample sizes used.

The 100 base-pair flanks of large (≥20 bp) homopolymers were the most heavily enriched for suspect loci by a large amount (OR = 23.29, 33.69, 30.37 for data sets 1, 2, 3). Small (<20 bp) homopolymers' 100 base-pair flanks (OR = 12.40, 5.99, 7.36), GC-rich regions (OR = 8.19, 6.47, 7.99), and *Alu* repeats (OR = 5.70, 6.74, 6.21) were also strongly enriched for suspect loci. Large homopolymers (OR = 1.93, 2.92, 1.84) and the RepeatMasker regions (OR = 1.85, 2.96, 2.18) were mildly enriched for suspect loci in data sets 1–3. Small homopolymers, on the other hand, were depleted or unenriched (OR = 0.525, 1.03, 0.519), and the NIST GIAB high-confidence region was strongly depleted (OR = 0.191, 0.154, 0.172) for suspect loci. The NonUnique100 region showed much greater enrichment of suspect loci in data sets 2 (OR = 3.26) and 3 (OR = 2.17), than in data set 1 (OR = 1.30), and segmental duplications also displayed this (OR = 1.17, 2.15, 1.74).

Systematic biases confirmed in the gold-standard reference sample

Our analysis based on aggregate allele fraction statistics found the presence of suspect loci in all cohorts. In order to confirm that these loci occur independently from the samples examined, specific suspect loci were examined in the reference sample NA12878 cell line, which was sequenced and had variants called by Personalis, Inc. using the same pipeline as data set 1 for comparison but which was not part of any of the data sets used to generate the IncDBs. We divided the SNVs called using this pipeline into groups based on whether they were annotated as suspect or nonsuspect, whether they occurred in the NIST GIAB benchmark regions, and whether they matched the v3.3.2 NIST GIAB benchmark variants at those positions.

Suspect SNVs, which accounted for ~5% of all SNV calls, mostly reported low allele fractions (0.300 median compared to 0.579 in nonsuspect SNVs) in the read pileup of NA12878, with the exception of suspect SNVs within the benchmark region that matched the benchmark variants. These were the least common type of suspect SNV and most closely resembled the allelic fraction distribution of nonsuspect SNVs (1358 SNVs) (**Fig. 3.10A**), suggesting that suspect loci that matched the benchmark variants were likely to be false positives. Suspect SNVs within the benchmark regions usually did not match the benchmark variants within these high-confidence regions (1839 SNVs) (**Fig. 3.10B**). Most suspect SNVs occurred outside of the benchmark regions (23,802 SNVs) (**Fig. 3.10C**).

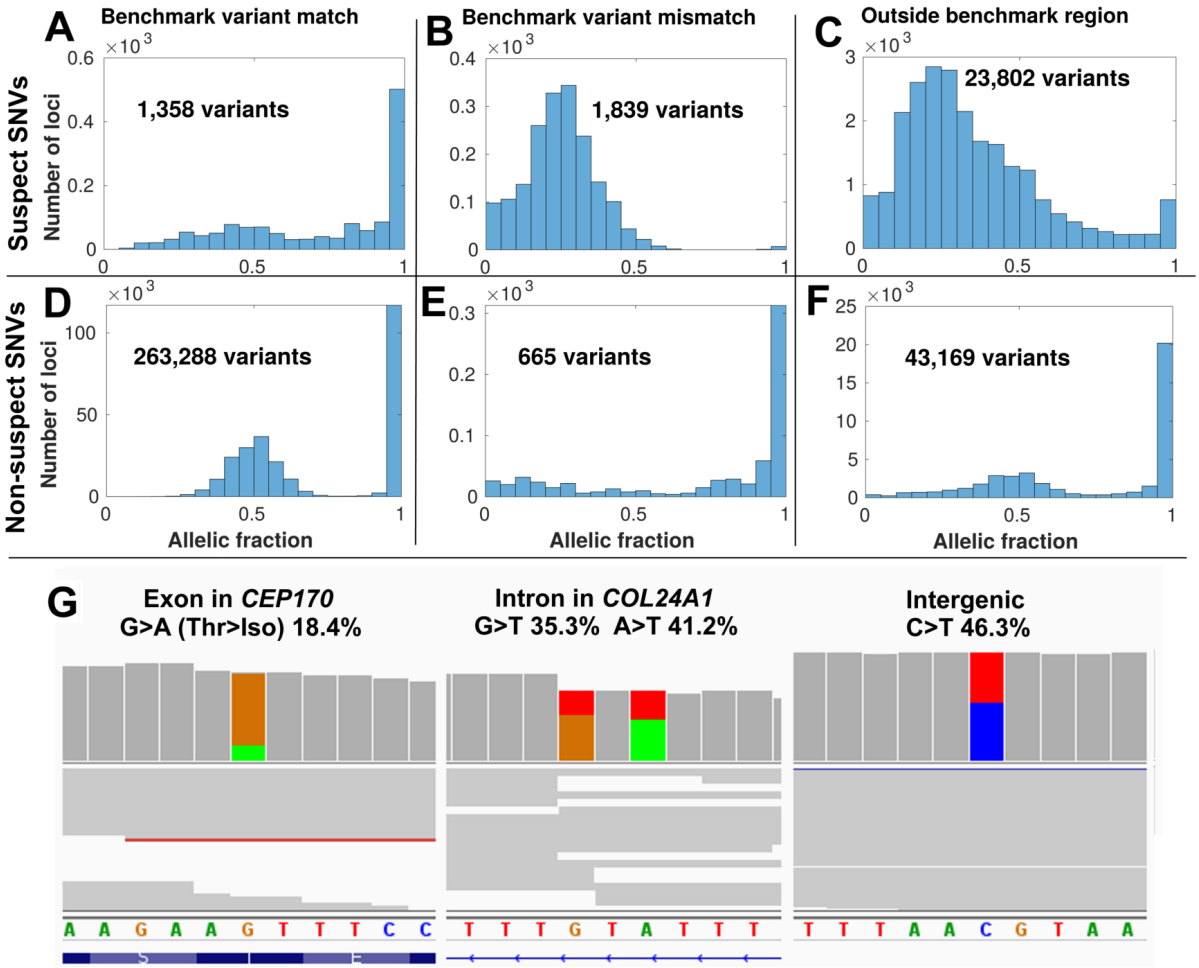


Figure 3.10: Suspect loci in detected variants of a gold-standard genome. Distribution of allelic fractions of SNVs called in Chromosome 1 of NA12878, classified as either suspect SNVs (top row - **A-C**), if they corresponded to suspect alleles in data set 1 (Personalis), or non-suspect SNVs (second row - **D-F**). SNVs were also classified based on whether they matched the NIST GIAB v3.3.2 benchmark variants (left column), didn't match the benchmark variants (middle column) or were outside of the GIAB benchmark region (right column). Low coverage variants (<10 supporting reads) were excluded from this analysis. (**G**) Cropped panels from the Integrative Genomics Viewer (Robinson et al. 2017), highlighting suspect loci from data set 1 in Chromosome 1 which were called as variants separately in NA12878. NA12878 was sequenced with Illumina HiSeq, but not used as part of the patient data set to create the IncDB (Zook et al. 2014, 2019). Reads are shown in grey with coloured bands where non-reference allelic reads were observed (A=Green, C=Blue, G=Brown, T=Red). Suspect SNVs and their respective read proportions in the NA12878 cell line are indicated above - these systematically occur at similar levels across all patients in the IncDBs used to identify them. Left/Middle: Suspect SNVs in exonic and intronic regions of genes in the PanelApp Intellectual Disability panel (Martin et al. 2019). Right: Suspect SNV in an intergenic region.

In contrast, most called SNVs in NA12878 were nonsuspect, were within the GIAB high-confidence benchmark region, and matched the GIAB benchmark variants for those

positions (263,288 SNVs) (**Fig. 3.10D**). These conformed to the expected distribution of allele fractions for heterozygous or homozygous SNVs, with peaks around 50% and 100%. Nonsuspect SNVs that were inside the high-confidence benchmark region but which did not match the benchmark variants were the least common and only showed a peak around 100%, with no peak observed for heterozygous variants and low levels of SNVs at all allelic fractions, suggesting that these were sequenced the worst of the nonsuspect loci (665 SNVs) (**Fig. 3.10E**). Nonsuspect SNVs outside of the benchmark regions were also numerous and had a similar pattern to nonsuspect SNVs which matched benchmark variants, although the peaks were broader (43,169 SNVs) (**Fig. 3.10F**), suggesting that the allelic fractions recorded for these were slightly less accurate.

Suspect loci called as SNVs occurred across all types of genomic regions, including clinically important positions (**Fig. 3.10G**). Altogether, this suggests that the systematic biases at suspect loci can contribute to false positive variant calls in clinically important regions, even within the NIST GIAB high-confidence regions. However, suspect SNVs that match the v3.3.2 NIST GIAB benchmark variants should be treated with caution, since their allelic fraction distributions suggest that these may indeed be true variants, so the NIST GIAB benchmark variants are a necessary and complementary resource to combine with IncDB-based methods for annotating systematic errors accurately.

Discussion

The main aim of this study has been to develop and evaluate a novel statistical method to identify positions of the genome that are prone to systematic bias in genomic sequencing and alignment, using anonymized summary patient data. We developed an approach to quality control sequenced reads in the autosomal genome by cataloguing genomic positions that persistently present a low-fraction alternate allele across patients rather than reflecting true biological variation, which we have labeled as 'suspect loci.' We have explored the extent to which these systematic biases occur across varied genomic regions, including regions known to be difficult to sequence, higher confidence regions, and clinical panels. We have also confirmed the existence of these systematic biases in an independent gold-standard reference genome and the utility of our approach.

Using the GIAB benchmark calls to assess called SNVs removes roughly 95% of suspect SNV calls (**Figure 3.10**). However, it does not remove the remaining suspect SNV calls within the benchmark region and it also has the disadvantage of filtering out all SNV calls outside of the GIAB benchmark region, 64% of which are not suspect, completely preventing SNV calling outside of the benchmark regions. By contrast, the IncDB-based filtering approach enables called SNVs to be assessed throughout the human genome. I was unable to test how the IncDB-based filtering approach compared with simply using a binomial call to assess allelic fraction, due to data access restrictions on extracting binomial call data. However, I would predict that using binomial calls alone would mask low-level systematic biases, since these would effectively be hidden from the diploid genotype after a binomial call, and would therefore only identify very strong systematic biases that affected genotype calls.

For the purposes of this paper, we have defined systematic biases as locus-specific errors in the accurate quantification of an allelic fraction that affect all samples sequenced with the same pipeline to a similar extent. Within this definition, our method should successfully identify almost all single-nucleotide allelic fraction biases >1.5%, but it does not have the sensitivity to detect smaller biases at the confidence level chosen (**Fig. 3.7**). For example, using a confidence interval of 99.9% and a sample size of 150 (reflecting the parameters used to classify suspect loci for data set 1), ~9.6% of sites with a 1% systematic error rate were classified as suspect loci, while ~99.2%/99.8%/100%/100% of sites with a 2%/3%/4%/5% systematic error rate were classified as suspect loci, respectively, when we simulated sites with these systematic error rates, representing a very high resolution. We have not tested this method for identifying systematic biases in the detection of indels or structural variants, but, with modifications, a similar approach can be used to identify those types of systematic error.

We did not investigate systematic biases in insertions and deletions in this study due to technical difficulties in implementing this for human genomes, which are extremely large and would have resulted in an excessively long runtime for the computational code used as a result. This was not ideal because insertions and deletions can be a key source of systematic bias, as highlighted in Chapter IV of this thesis, and we recommend that indels are considered when permitted by the computational resources available. We carried out Monte Carlo simulations to check if this method could be used to identify CNVs, but found that CNVs were not annotated by this method below a population frequency of 10% and could only be detected if intervening SNV alleles had unbalanced aggregate allelic fractions >10%. We expect that CNVs that fulfill these criteria are exceedingly rare, so this is unlikely to cause much of the systematic bias this method identifies. Our method cannot be used for identifying types of systematic bias that are not locus-specific, such as general biases in the detection of certain alleles (e.g., GC bias) (Chen et al. 2013), biases that systematically affect the ends of genomic reads being sequenced (Ma et al. 2019), or other context-specific biases that do not occur in a systematic or locus-specific manner. For example, we would not expect to identify mosaicism unless this was systematic, with a high proportion of individuals having the same alternative genotypes at the same genomic coordinates in the same fraction of cells in their tissue. To evaluate whether age-related mosaicism could have been present in our results, we compared the numbers of suspect loci (54.6 million) and suspect SNVs (43,832) in data set 5, which used a more elderly cohort of patients, with data set 3 (63.2 million suspect loci and 46,287 suspect SNVs) and data set 4 (90.3 million suspect loci and 59,634 suspect SNVs), which both used the same sequencing pipeline. We did not observe an increase in suspect loci in the gene *TP53*, which is thought to be correlated with this (Yizhak et al. 2019), or in general across the genome (**Fig. 3.11**), so we could not find evidence for mosaicism affecting suspect locus annotation. The main aim of our approach is not to identify a specific cause of systematic bias but rather to act as a quality control method, to catalogue where these systematic biases occur so that they can be filtered out of scientific and clinical results where they may lead to inaccurate conclusions. As a result, our study does not seek to identify why the systematic biases we have identified occur, even if some of our results suggest possible causes that could be followed up by future studies.

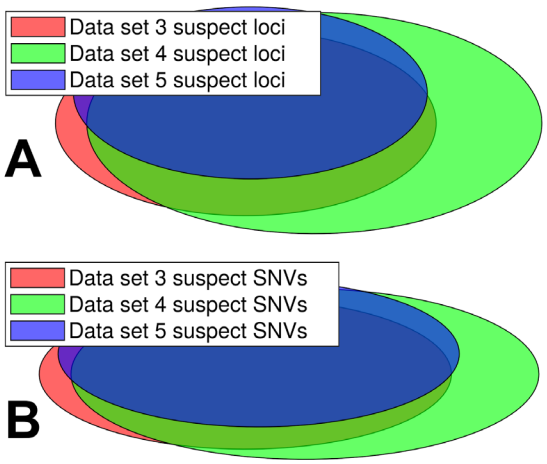


Figure 3.11: Venn diagrams showing the overlap in data sets 3-5 of (**A**) all unique suspect loci and (**B**) called SNVs in NA12878 annotated as suspect, with shaded areas proportional to the numbers and overlap of suspect loci and suspect SNVs in each data set. All 3 of these data sets use exactly the same technical conditions, but share no patients in common. All the patients used have the same distribution of clinical attributes but cohort 5 are on average an older group to explore the possibility of age-related mosaicism impacting suspect loci detected. These 3 IncDBs show considerable suspect loci overlap (mean overlap=65.7%) and suspect SNV overlap (mean overlap=71.9%) between each other, although there are some differences likely due to private mutations. The higher degree of overlap between data sets 3 and 5 than between data sets 3 and 4, as well as the lower numbers of suspect loci and suspect SNVs in data set 5 than data set 4 suggest that age-related mosaicism is unlikely to impact on suspect locus annotation.

To investigate the extent to which differences in mutations between large patient cohorts affected the annotation of suspect loci when technical procedures were held constant, we compared the overlap between suspect loci for data sets 3–5, for which all patients were sequenced using the same protocol and pipeline. These three data sets showed considerable suspect loci overlap (mean overlap = 65.7%) and suspect SNV overlap (mean overlap = 71.9%) in the intersection of all three data sets, although there were some differences likely due to private mutations (**Fig. 3.11**). Compared with the overlap between data sets 1 and 2, for which separate technical procedures were used (mean overlap = 34.8% for suspect loci, 66.4% for suspect SNVs), the overlap was ~30% greater for suspect loci but only ~5% greater for suspect SNVs, suggesting that differences in the technical procedures used played a large role in suspect locus annotation in general, but genetic differences between cohorts were particularly important when identifying systematic biases that could affect variant calls. We would therefore recommend users increase the size of IncDBs they generate as much as is reasonably possible, to maximize the numbers of systematic biases they can detect by this method, at least until overlap between separate IncDBs approaches a maximum percentage.

For loci unaffected by systematic bias, the standard deviation and aggregate allelic fraction stored in the IncDB were expected to relate to each other in accordance with Hardy-Weinberg equilibrium. However, ~1%–3% of autosomal loci had a significantly lower standard deviation than predicted by Hardy-Weinberg equilibrium ($P = 0.0005$), suggesting the presence of systematic bias across numerous genomic loci. At these loci, most individuals appear to present a low-fraction allele. Persistent low-fraction alleles would be inconsistent with our understanding of human genetics and are presumably a technological artifact: a bias or systematic error in the sequencing technology itself or perhaps in the read mapping. The impact of these suspect loci is magnified in the context of studies looking at large numbers of genomic positions, since a small percentage of this would still correspond to a high number of genomic positions affected by systematic bias. It is therefore clear that these systematic biases are of concern and deserve further attention. Previous studies examining systematic sequencing bias have primarily focused on biases in total coverage across loci and have not examined position-dependent systematic biases in the allelic fraction (Cheung et al. 2011; Ross et al. 2013); therefore, no previous estimates for the prevalence of systematic sequencing bias in allelic fractions were available to compare with our estimate of ~1%–3% of all autosomal loci in the human genome.

The Hardy-Weinberg equilibrium model used as the basis for calculating the expected allelic fraction standard deviation relies upon a range of assumptions. These include that organisms are diploid, do not reproduce asexually and have very large populations, which is true of the cohorts observed. However, the Hardy-Weinberg model also assumes that random mating occurs and that no migration occurs into or out of the population, which is not representative of real world cohorts. Taken to an extreme, if an IncDB was generated for a cohort derived from two genetically distinct populations that did not interbreed, then there would be many positions on the genome where a SNP common to one population rarely occurred in another population, leading to a lower-than-modelled number of heterozygous individuals occurring in the IncDB cohort. This would result in a higher allelic fraction standard deviation being observed in the IncDB, making those positions less likely to fall under the minimum standard deviation threshold to be annotated as having systematic bias. It would therefore be more difficult to find systematic bias at genomic positions where genetically-isolated human populations differed, so our estimate of systematic bias would be an underestimate. However, given that more than 99.8% of genomic positions are identical even for genetically distinct humans, this would have very little effect on the overall number of genomic positions displaying systematic bias. In the cohorts used in this study, this scenario was not the case, with participants primarily coming from a white ethnic background and

Hardy-Weinberg equilibrium generally holding true (**Table 3.2**). Another assumption of the Hardy-Weinberg model is that alleles are not under selection. There are likely to be positions in the human genome where this is not the case, resulting in fewer than expected individuals homozygous for a deleterious recessive mutation, with a lower allelic fraction standard deviation in the cohort as a result. This would result in the opposite effect, with such alleles more likely to be incorrectly annotated as showing systematic bias. Limitations due to the genetic makeup of the cohorts used and the presence of alleles under selection could be reduced in future applications by filtering genomic positions out of the results if the ratios of genotypes were observed to differ from the Hardy-Weinberg predictions. Given that these scenarios would only affect very small numbers of genomic positions, we would not expect them to alter the IncDB results significantly.

Another possible limitation to using IncDB-based methods to assess systematic bias in genetically-isolated populations is caused by the phenomenon of reference bias. Reference bias is an effect whereby individuals who are more genetically different from the reference human genome are not sequenced as effectively, as a result of reads that are very different from the reference genome, such as those containing structural variants, not mapping correctly, leading to variants in these reads not being detected. This can alter the detected allelic fraction at these positions in individuals, resulting in them being incorrectly genotyped as homozygous when they are actually heterozygous and causing the observed standard deviation to be higher than it actually is, resulting in an underestimate of systematic bias in these populations. Long read sequencing technologies such as Oxford Nanopore Technologies can be used to improve mapping of structural variants to minimise the incidence of this.

IncDBs could potentially also be used to identify systematic biases in whole exome sequencing, using the same method. However, whole exome sequencing is affected by additional sources of systematic and non-systematic bias that could alter the allelic fraction standard deviation observed significantly, in different directions for different genomic positions. In preliminary tests on whole exome sequencing data, we found that the observed standard deviations were very different to the expected curve, with a much broader arc indicating that our model was a poor fit, so we do not recommend applying IncDBs to whole exome sequencing data.

Our estimate indicated a very large number of genomic sites affected by systematic bias, potentially more than might be expected. Therefore, we analyzed what proportion of these would be filtered out by applying a strongly conservative confidence threshold

(99.999%) in addition to the default threshold used in this study (99.9%) to data sets 1–3 and compared how this affected suspect locus numbers, allelic fractions, and overlap between data sets 1 and 2. This filtered out a very high proportion (99.98%/91.06%/92.68% in data sets 1, 2, and 3, respectively) of the suspect loci that had allelic fractions <0.03, since it was rare for loci exhibiting very small systematic biases to have a standard deviation lower than the 99.999% confidence threshold (**Fig. 3.12**) compared to the 99.9% confidence threshold (**Fig. 3.7**). In contrast, a much lower proportion (25.95%/17.74%/23.40% in data sets 1, 2, and 3, respectively) of the suspect loci with larger systematic biases (allelic fractions > 0.03) were filtered out at the 99.999% confidence threshold. Since it is the larger systematic biases that are most likely to affect variant calling, this shows that the method is robust for identifying most important systematic biases, even at a strongly conservative confidence level. The percentage overlap of suspect loci between data sets 1 and 2 decreases slightly when a more conservative threshold (99.999% confidence interval) is used for defining suspect loci (from 47.9%/21.7% to 41.1%/16.1% of data sets 1 and 2, respectively) (**Fig. 3.13**).

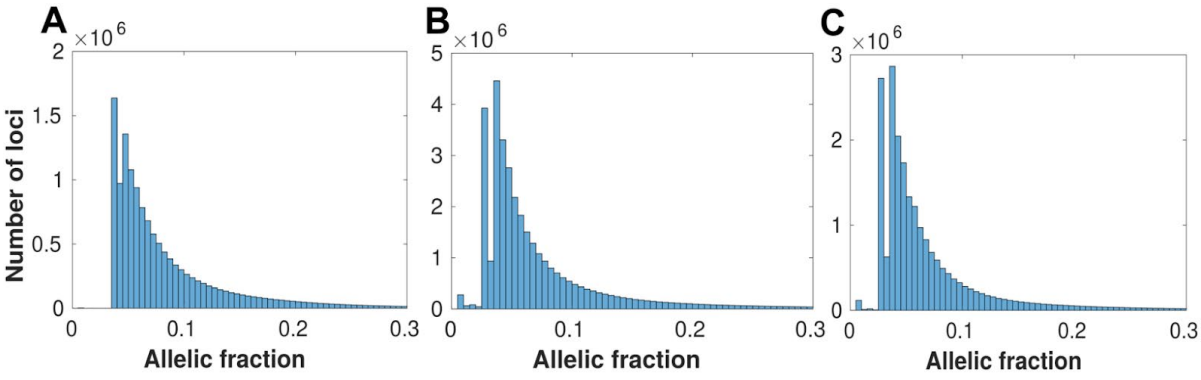


Figure 3.12: Distribution of allelic fractions observed at suspect loci in data sets 1-3 respectively (A-C), using a stricter confidence threshold for suspect locus classification than the default parameters used in this study (confidence interval=99.999%, assumed error rate=0.01). Each bin has a width of 0.005 allelic fraction and the scale is cut off at allelic fraction=0.3 because the number of suspect loci decreases to near-zero. Compared to Fig. 3.7, the number of suspect loci at lower allelic fractions (<3%) drops off sharply since the systematic error rate is unlikely to be significant at such a low effect size, at this confidence level.

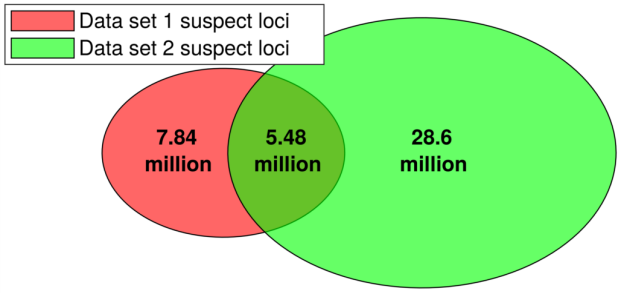


Figure 3.13: Overlap of all unique suspect loci between data sets 1 and 2, using a stricter confidence threshold for suspect locus classification (confidence interval=99.999%, assumed error rate=0.01). Data set 3 used the GRCh38 reference, so it was not included in the Venn diagram. Applying this alternative threshold filtered out a very high proportion (99.98/91.06/92.68% in data sets 1, 2 and 3 respectively) of the suspect loci that had allelic fractions <0.03, since fewer low allelic fraction loci exhibited a standard deviation lower than the 99.999% confidence threshold. In contrast, a much lower proportion (25.95/17.74/23.40% in data sets 1, 2 and 3 respectively) of the suspect loci with larger systematic biases (allelic fractions>0.03) were filtered out at the 99.999% confidence threshold. The percentage overlap of suspect loci between data sets 1 and 2 decreases slightly when the more conservative threshold (99.999% confidence interval) is used for defining suspect loci (from 47.9/21.7% with the default threshold (Fig. 3.5D) to 41.1/16.1% in this figure, in data sets 1/2 respectively).

Suspect loci were widespread across all sequenced chromosomal segments, but there was variability between different types of genomic region. Despite this, there was little or no depletion of suspect loci in some regions expected to have more accurate sequencing, such as exons and the clinical gene panels, suggesting that greater caution in these areas is justified when calling variants. The NIST high-confidence regions displayed the greatest depletion of suspect loci as expected and had the lowest proportion of suspect loci per kb, while homopolymer flanks displayed the greatest enrichment of suspect loci.

The correlation between chromosomal suspect locus density and the proportions of chromosomes within large homopolymer flanks was high (Spearman's rho = 0.955, 0.861, 0.820, $P = 5.11 \times 10^{-12}$, 2.76×10^{-7}, 3.07×10^{-6} for data sets 1, 2, and 3, respectively), accounting for the majority of the variability in the proportion of suspect loci in autosomal chromosomes (**Fig. 3.14**). The flanks of large homopolymers are particularly prone to errors in alignment compared to other types of sequencing error, especially for Illumina sequencing (Laehnemann, Borkhardt, and McHardy 2016) as was done here, suggesting that the misalignment of reads could be a source of systematic biases.

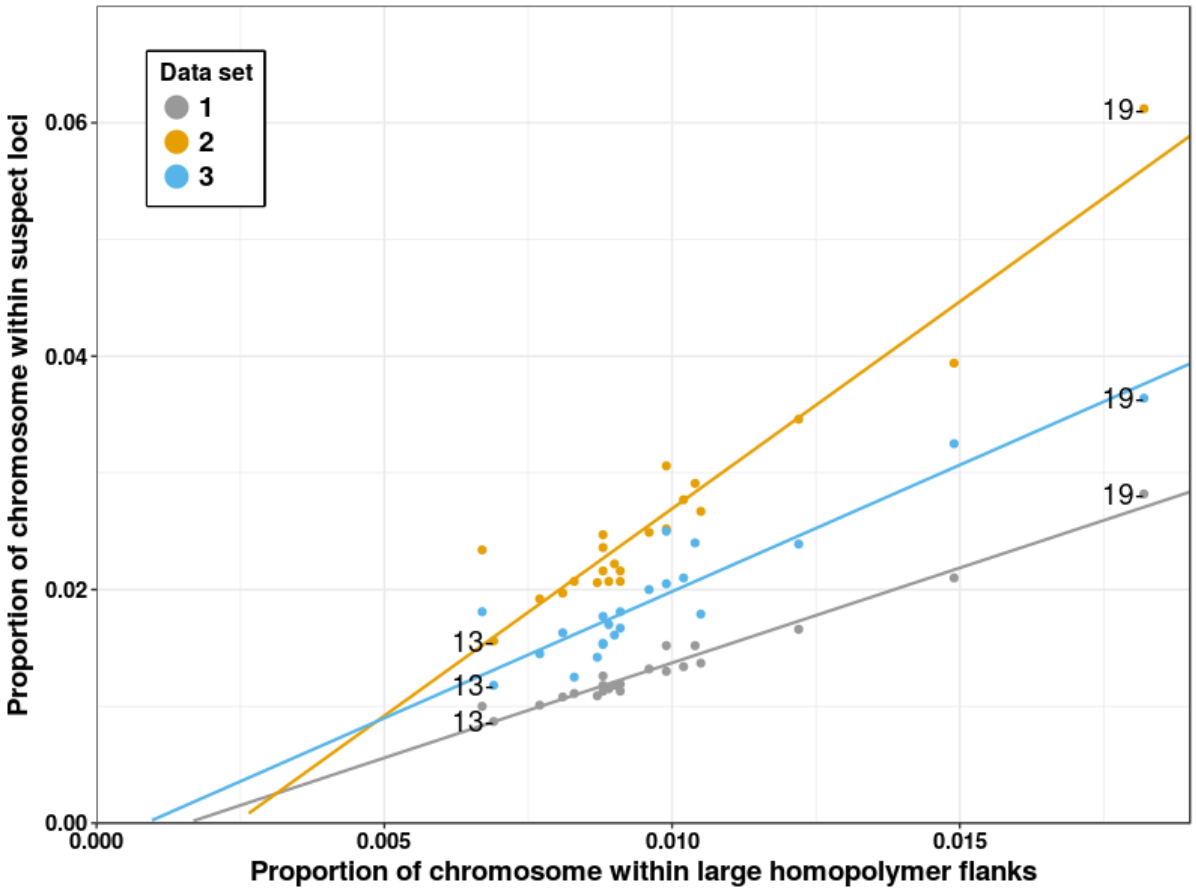


Figure 3.14: Proportion of chromosomal positions with at least one suspect locus vs. proportion of chromosome within large homopolymer flanks, with linear line of best fit. There is a strong linear correlation between the proportion of unique suspect loci in each chromosome and the proportion of loci in large homopolymer flanks, indicating that these regions are a key contributor to systematic sequencing and/or alignment errors. Chromosomes 13 and 19 are labelled as the chromosomes with the lowest and highest levels of suspect loci respectively in all 3 data sets.

One hundred base pairs was chosen as the flank length because it is of the same order of size as the read lengths used for the Illumina sequencing and we would not expect anything outside of 100 bp from a homopolymer to be covered by sequencing reads, so the 100-bp length is the theoretical maximum length at which we might expect effects to occur. Homopolymers cannot affect the sequencing beyond one read length from their edge. In practice, most suspect locus enrichment occurred within the 3-bp flanking regions however, with enrichment odds ratios greatly increasing when the flank size was reduced to 3 bp in both large homopolymers (from 23.3/33.7/30.4 in data sets 1, 2, and 3, respectively, in 100-bp flanks, to 171.3/94.3/97.6 in 3-bp flanks) and small homopolymers (from 12.4/6.0/7.4 in 100-bp flanks in data sets 1, 2, and 3, respectively, to 30.9/12.2/13.3 in 3-bp flanks).

In addition, we found differences in regional systematic biases between data sets. For example, the NonUnique100 region showed much greater enrichment of suspect loci in data sets 2 (OR = 3.26) and 3 (OR = 2.17) than in data set 1 (OR = 1.30). This region is defined as containing all sequenced positions where a single 100-bp read cannot map uniquely, so we would expect a large proportion of systematic biases here to mainly correspond to alignment errors rather than sequencing chemistry. The higher enrichment of suspect loci in NonUnique100 in data sets 2 and 3 could therefore indicate that the different aligners used (Isaac aligner (Raczy et al. 2013) for data sets 2 and 3 rather than BWA-MEM (H. Li 2013) for data set 1) could be the main explanation for the different levels of systematic bias found between the data sets in this region. The Isaac aligner is known to be a faster aligner than BWA-MEM, but previous validation attempts found that it was slightly worse than BWA-MEM in terms of accuracy, especially outside of NA12878 (Mainzer et al., n.d.). The increased frequency of systematic biases we detected with the Isaac aligner in data sets 2 and 3 therefore seems to indicate that BWA-MEM is still superior in terms of accuracy, at least compared with the versions of the Isaac aligner used (SAAC00776.15.01.27 and iSAAC-03.16.02.19, respectively).

We also found variability in the distribution of suspect loci within individual genes tested (**Supplemental Data S1,S2** at https://genome.cshlp.org/content/30/3/415/suppl/DC1). For example, within the clinical gene panels, there were very low proportions of suspect loci (0% out of 2021 total loci in *LIPT2* in data set 2) to very high proportions of suspect loci (36.6% out of 3759 total loci in *NPRL2* in data set 2), suggesting that some genes might be particularly prone to systematic sequencing biases. This is likely because the genes intersect with the problematic regions we examined in different ways. Both of these genes are associated with genetic epilepsy syndromes, but *LIPT2* is not affected by systematic sequencing biases at all, while these heavily affect *NPRL2*. Clinicians focusing on specific genes for diagnostic purposes could use this information to identify how much caution they need when assessing pathogenic variants in those genes (**Table 3.4**). These suspect loci were confirmed in an independent reference sample sequenced using the same pipeline, including within introns, exons, intergenic regions, and NIST GIAB high-confidence regions, and within called variants in clinically relevant regions such as the PanelApp disease panels (Martin et al. 2019). In addition, we demonstrated that our approach could be used in combination with the NIST GIAB benchmarking variants to improve suspect locus annotation by identifying which suspect SNVs were likely to be false positives.

dbSNP ID	Gene Affected	Variant Type	Diagnostic Panel	ClinVar Associated Disease
rs1131691563	*AMPD2*	splice donor variant	Ataxia and cerebellar anomalies	Not provided
rs1553201258	*DARS2*	intron variant	Ataxia and cerebellar anomalies	Cerebral cortical atrophy
rs370132645	*OTOF*	splice donor variant	Auditory Neuropathy Spectrum Disorder	Rare genetic deafness
rs587779190	*MSH2*	stop gained	Adult solid tumours cancer susceptibility	Lynch syndrome
rs63750640	*MSH2*	missense variant	Adult solid tumours cancer susceptibility	Lynch syndrome
rs1114167845	*MSH2*	stop gained	Adult solid tumours cancer susceptibility	Lynch syndrome
rs587779197	*MSH2*	missense variant	Adult solid tumours cancer susceptibility	Lynch syndrome
rs587779195	*MSH2*	splice donor variant	Adult solid tumours cancer susceptibility	Lynch syndrome
rs121918737	*SCN1A*	missense variant	Genetic epilepsy syndromes	Severe myoclonic epilepsy in infancy
rs267607819	*MLH1*	splice acceptor variant	Adult solid tumours cancer susceptibility	Lynch syndrome
rs1559551570	*MLH1*	frameshift	Adult solid tumours cancer susceptibility	Hereditary nonpolyposis colon cancer
rs148891849	*DNAH5*	stop gained	Primary ciliary disorders	Primary ciliary dyskinesia
rs7755898	*CYP21A2*	stop gained	Disorders of sex development	Classic congenital adrenal hyperplasia
rs1554247637	*ARID1B*	frameshift	Coffin-Siris syndrome	Inborn genetic diseases

rs1562671039	*PMS2*	stop gained	Adult solid tumours cancer susceptibility	Not provided
rs193929376	*GCK*	splice donor variant	Diabetes - neonatal onset	Permanent neonatal diabetes mellitus
rs121909112	*HSPB1*	missense variant	Distal myopathies	Charcot-Marie-Tooth disease type 2F
rs111033565	*PRSS1*	missense variant	Pancreatitis	Hereditary pancreatitis
rs794728419	*KCNH2*	splice donor variant	Short QT syndrome	Not provided
rs587777641	*GPIHBP1*	missense variant	Severe hypertriglyceridaemia	Hyperlipoproteinemia, type ID
rs1563963464	*APTX*	splice acceptor variant	Ataxia and cerebellar anomalies	Ataxia-oculomotor apraxia type 1
rs146292819	*ABCA1*	missense variant	Hereditary neuropathy	ABCA1-Related Disorders
rs864321692	*WAC*	stop gained	Intellectual disability	Desanto-shinawi syndrome
rs587782455	*PTEN*	splice acceptor variant	Adult solid tumours for rare disease	PTEN hamartoma tumour syndrome

Table 3.4: Verified pathogenic variants in ClinVar that are included in diagnostic gene panels and at locations of suspect loci found in data sets 3-5.

We further confirmed that suspect variants were filtered out of the Genome Aggregation Database (gnomAD) using their quality control processes (**Fig. 3.15**). The gnomAD database is a large database of all of the variation found across a large ethnically diverse population, taken from 125,748 exomes and 15,708 genomes (Karczewski et al. 2019). Our results revealed that suspect variants were also widespread in the gnomAD database, even after filtering by gnomAD's quality control process, across allelic frequencies. We also evaluated whether suspect loci could be identified simply by using a quality threshold **(Fig. 3.16**). Nonsuspect loci had significantly higher proportions of high quality reads, with >90% of reads having sequencing and mapping quality scores >20 in ~90% of nonsuspect loci. This demonstrated that low-quality reads were more frequent among suspect loci to a

large degree, suggesting that improving sequencing and alignment quality could help with decreasing these systematic biases. However, there was still significant overlap between read quality at suspect and nonsuspect loci. Finally, we analyzed whether read depth could be used to quality-control for suspect loci (**Fig. 3.17**). Kolmogorov–Smirnov tests confirmed that there was a different distribution of read depths between the sets of genomic positions with and without suspect loci ($D = 0.4289, 0.2885, 0.3079$ and $P < 10^{-15}$ for data sets 1, 2, and 3, respectively), but there was not a clear correlation between read depth and likelihood of a position having suspect loci. While the lower quartile and median values for read depth were slightly lower for unique suspect loci, for data sets 2 and 3 the upper quartiles were significantly higher at unique suspect loci. We could only conclude that there was a greater breadth of read depth in positions annotated as unique suspect loci than other positions. A possible explanation for our results could be that most unique suspect loci have slightly lower read depths than nonsuspect loci, but some unique suspect loci have extremely high read depth, perhaps as a result of systematic alignment errors attributing reads from multiple locations in the genome to the same genomic position at these suspect loci. Our analyses therefore suggest that the existing read depth thresholds and quality control procedures commonly used in sequencing would not be sufficient to filter out the systematic biases, and the reported variants in large population studies such as gnomAD may need to be reassessed.

In addition to read depth and quality thresholds, there exist other quality metrics for assessing the validity of variant calls, including filtering out entire genomic regions with poorer sequencing quality, carrying out matched tumour-normal sequence comparisons to distinguish between real low-level tumour variants and noise, and using machine learning-based approaches such as Variant Quality Score Recalibration (VQSR) (Zhang and Ochoa 2020). VQSR learns from a training set of genomic data and compares this with validated variants in order to identify the probability that a positive or negative variant call is correct. VQSR learns general context-based features of true positive variant calls based on many different positions across the genome as a result, whereas our approach carries out separate quantification of a single metric at each individual genomic position. As a result, our approach requires many genomes in order to filter out false-positive variants, but provides an output that is more customised to the exact positions at which variants are called.

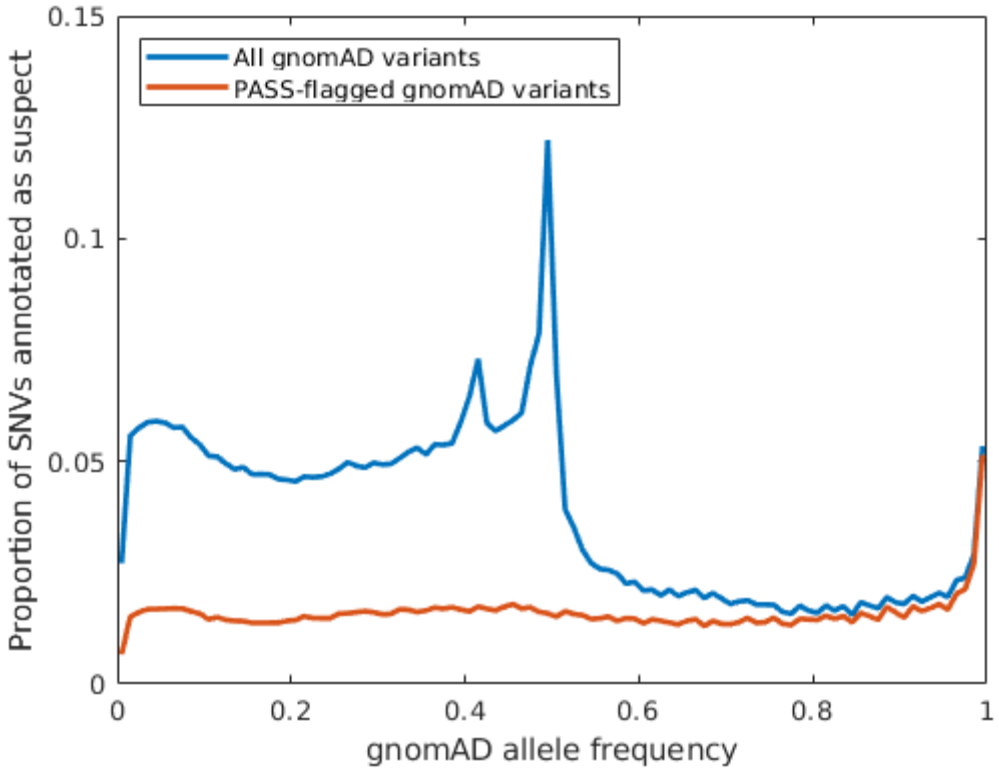


Figure 3.15: Proportion of autosomal gnomAD SNVs annotated as suspect at different gnomAD allele frequencies in data set 1. A large proportion of the low allelic frequency gnomAD variants annotated as suspect did not pass gnomAD's quality control, but ~1.5% of SNVs that passed gnomAD's quality control checks were annotated as suspect across most allele frequencies, suggesting that systematic biases are prevalent in gnomAD's called SNVs, albeit at a low rate.

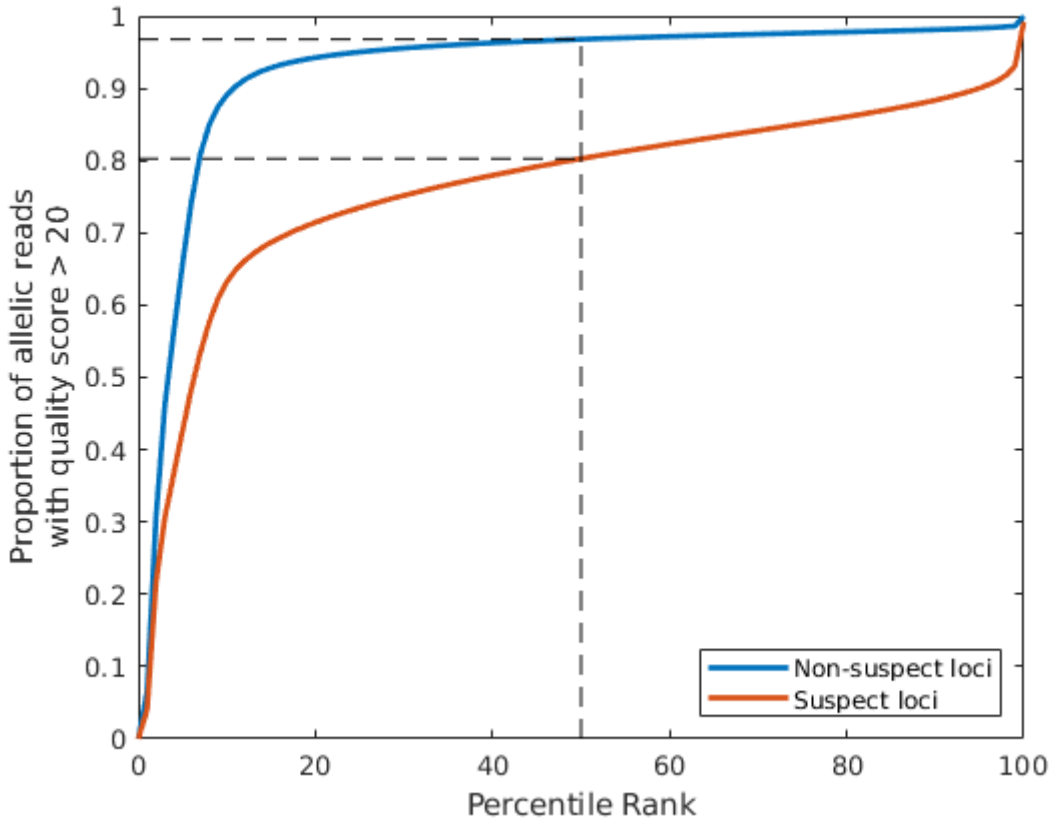


Figure 3.16: The proportion of allelic reads that had quality scores > 20 (for both sequencing and mapping), plotted against the percentile ranks for both non-suspect loci (blue) and suspect loci (red) respectively using this measure, including reads for all locus/allele combinations for chromosome 1 in data set 1 (Personalis Inc.). Non-suspect loci had significantly higher proportions of high quality reads, with >90% of reads having quality scores >20 in ~90% of non-suspect loci/allele combinations vs. ~5% of suspect loci/allele combinations.

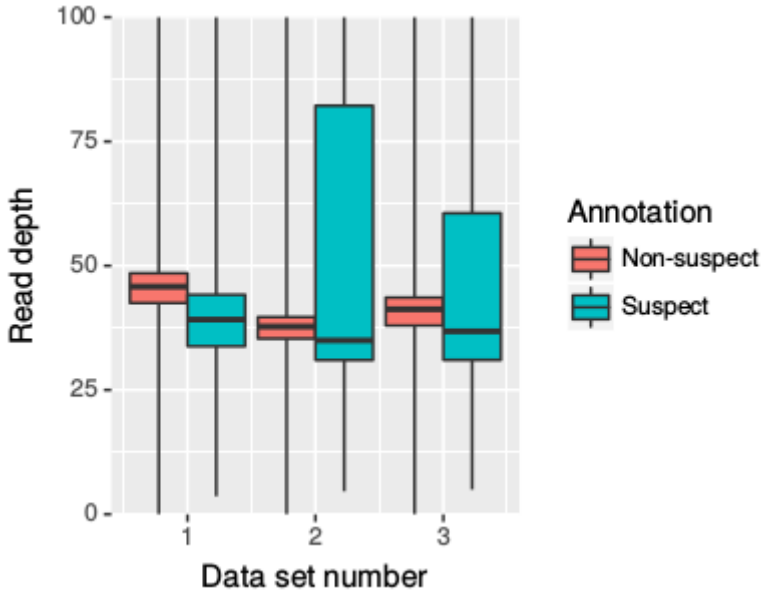


Figure 3.17: Distribution (quartiles) of read depths observed across suspect (cyan) and non-suspect (red) loci positions in data sets 1-3 respectively. The maximum read depth values for the whiskers were 5,304/6,165/6,318 for non-suspect loci and 5,195/9,093/7,764 for suspect loci in data sets 1-3 respectively, but the figure y limit was kept at a read depth of 100 for clarity over the main range of read depths. Two-Sample Kolmogorov-Smirnov tests confirmed that there was a different distribution of read depth between genomic positions with and without suspect loci (D=0.4289, 0.2885, 0.3079 and $p<10^{-15}$ for data sets 1,2,3 respectively), but there was no clear correlation between read depth and likelihood of a position having suspect loci. While the lower quartile and median values for read depth were slightly lower for unique suspect loci, for data sets 2 and 3 the upper quartiles were significantly higher at unique suspect loci.

Conclusion

We have demonstrated the utility of IncDBs to assess the quality of clinical whole genomes of five independent cohorts sequenced by commercial and public healthcare organizations while maintaining patient anonymity. In addition to showing the utility of this approach on whole-genome Illumina sequencing, IncDBs could be applied to data from different types of sequencing platforms in the future, including specific targeted, exome-sequencing, and long-read technologies such as Pacific Biosciences (PacBio) and Oxford Nanopore. Our agile approach for detecting suspect loci could be deployed in various settings where the raw data for individual genomes cannot be accessed—for instance, when patient confidentiality must be maintained. Under those conditions, being able to identify systematic biases would enable improvements to variant calling and has the potential to reduce errors in clinical genomic testing.

References

1000 Genomes Project Consortium, Gonçalo R. Abecasis, David Altshuler, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Richard A. Gibbs, Matt E. Hurles, and Gil A. McVean. 2010. “A Map of Human Genome Variation from Population-Scale Sequencing.” *Nature* 467 (7319): 1061–73.

Benjamini, Yuval, and Terence P. Speed. 2012. “Summarizing and Correcting the GC Content Bias in High-Throughput Sequencing.” *Nucleic Acids Research* 40 (10): e72.

Chen, Yen-Chun, Tsunglin Liu, Chun-Hui Yu, Tzen-Yuh Chiang, and Chi-Chuan Hwang. 2013. “Effects of GC Bias in next-Generation-Sequencing Data on de Novo Genome Assembly.” *PloS One* 8 (4): e62856.

Cheung, Ming-Sin, Thomas A. Down, Isabel Latorre, and Julie Ahringer. 2011. “Systematic Bias in High-Throughput Sequencing Data and Its Correction by BEADS.” *Nucleic Acids Research* 39 (15): e103.

Cunningham, Fiona, Premanand Achuthan, Wasiu Akanni, James Allen, M. Ridwan Amode, Irina M. Armean, Ruth Bennett, et al. 2019. “Ensembl 2019.” *Nucleic Acids Research* 47 (D1): D745–51.

Goldfeder, Rachel L., James R. Priest, Justin M. Zook, Megan E. Grove, Daryl Waggott, Matthew T. Wheeler, Marc Salit, and Euan A. Ashley. 2016. “Medical Implications of Technical Accuracy in Genome Sequencing.” *Genome Medicine* 8 (1): 24.

Hasler, J., and K. Strub. 2007. “Survey and Summary: Alu Elements as Regulators of Gene Expression.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkm044.

Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2019. “Variation across 141,456 Human Exomes and Genomes Reveals the Spectrum of Loss-of-Function Intolerance across Human Protein-Coding Genes.” *bioRxiv*. https://doi.org/10.1101/531210.

Karolchik, Donna, Angela S. Hinrichs, Terrence S. Furey, Krishna M. Roskin, Charles W. Sugnet, David Haussler, and W. James Kent. 2004. “The UCSC Table Browser Data Retrieval Tool.” *Nucleic Acids Research* 32 (Database issue): D493–96.

Kent, W. James, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, and David Haussler. 2002. “The Human Genome Browser at UCSC.” *Genome Research* 12 (6): 996–1006.

King, Daniel A., Alejandro Sifrim, Tomas W. Fitzgerald, Raheleh Rahbari, Emma Hobson, Tessa Homfray, Sahar Mansour, et al. 2017. “Detection of Structural Mosaicism from Targeted and Whole-Genome Sequencing Data.” *Genome Research* 27 (10): 1704–14.

Laehnemann, David, Arndt Borkhardt, and Alice Carolyn McHardy. 2016. “Denoising DNA

Deep Sequencing Data—high-Throughput Sequencing Errors and Their Correction." *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbv029.

Li, H. 2013. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM." *arXiv:1303.3997*. https://arxiv.org/abs/1303.3997v2.

Li, Heng. 2011. "A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data." *Bioinformatics* 27 (21): 2987–93.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79.

Mainzer, Liudmila Sergeevna, Brad A. Chapman, Oliver Hofmann, Gloria Rendon, Zachary D. Stephens, and Victor Jongeneel. n.d. "Validation of Illumina's Isaac Variant Calling Workflow." https://doi.org/10.1101/031021.

Martin, Antonio Rueda, Eleanor Williams, Rebecca E. Foulger, Sarah Leigh, Louise C. Daugherty, Olivia Niblock, Ivone U. S. Leong, et al. 2019. "PanelApp Crowdsources Expert Knowledge to Establish Consensus Diagnostic Gene Panels." *Nature Genetics*, November. https://doi.org/10.1038/s41588-019-0528-2.

Ma, Xiaotu, Ying Shao, Liqing Tian, Diane A. Flasch, Heather L. Mulder, Michael N. Edmonson, Yu Liu, et al. 2019. "Analysis of Error Profiles in Deep next-Generation Sequencing Data." *Genome Biology* 20 (1): 50.

Raczy, Come, Roman Petrovski, Christopher T. Saunders, Ilya Chorny, Semyon Kruglyak, Elliott H. Margulies, Han-Yu Chuang, et al. 2013. "Isaac: Ultra-Fast Whole-Genome Secondary Analysis on Illumina Sequencing Platforms." *Bioinformatics* 29 (16): 2041–43.

Robinson, James T., Helga Thorvaldsdóttir, Aaron M. Wenger, Ahmet Zehir, and Jill P. Mesirov. 2017. "Variant Review with the Integrative Genomics Viewer." *Cancer Research*. https://doi.org/10.1158/0008-5472.can-17-0337.

Ross, Michael G., Carsten Russ, Maura Costello, Andrew Hollinger, Niall J. Lennon, Ryan Hegarty, Chad Nusbaum, and David B. Jaffe. 2013. "Characterizing and Measuring Bias in Sequence Data." *Genome Biology* 14 (5): R51.

Sandmann, Sarah, Aniek O. de Graaf, Mohsen Karimi, Bert A. van der Reijden, Eva Hellström-Lindberg, Joop H. Jansen, and Martin Dugas. 2017. "Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data." *Scientific Reports* 7 (February): 43169.

Smit, AFA, Hubley, R & Green, P. 2013-2015. "RepeatMasker Open-4.0." 2013-2015. http://www.repeatmasker.org.

Vattathil, Selina, and Paul Scheet. 2016. “Extensive Hidden Genomic Mosaicism Revealed in Normal Tissue.” *American Journal of Human Genetics* 98 (3): 571–78.

Wall, Jeffrey D., Ling Fung Tang, Brandon Zerbe, Mark N. Kvale, Pui-Yan Kwok, Catherine Schaefer, and Neil Risch. 2014. “Estimating Genotype Error Rates from High-Coverage next-Generation Sequence Data.” *Genome Research* 24 (11): 1734–39.

Xiao, Chunlin, Justin Zook, Shane Trask, Stephen Sherry, and the Genome-in-a-Bottle Consortium. 2014. “Abstract 5328: GIAB: Genome Reference Material Development Resources for Clinical Sequencing.” *Cancer Research*. https://doi.org/10.1158/1538-7445.am2014-5328.

Yizhak, Keren, François Aguet, Jaegil Kim, Julian M. Hess, Kirsten Kübler, Jonna Grimsby, Ruslana Frazer, et al. 2019. “RNA Sequence Analysis Reveals Macroscopic Somatic Clonal Expansion across Normal Tissues.” *Science* 364 (6444). https://doi.org/10.1126/science.aaw0726.

Zhang, Chuanyi, and Idoia Ochoa. 2020. “VEF: A Variant Filtering Tool Based on Ensemble Methods.” Bioinformatics 36 (8): 2328–36.

Zook, Justin M., Brad Chapman, Jason Wang, David Mittelman, Oliver Hofmann, Winston Hide, and Marc Salit. 2014. “Integrating Human Sequence Data Sets Provides a Resource of Benchmark SNP and Indel Genotype Calls.” *Nature Biotechnology* 32 (3): 246–51.

Zook, Justin M., Jennifer McDaniel, Nathan D. Olson, Justin Wagner, Hemang Parikh, Haynes Heaton, Sean A. Irvine, et al. 2019. “An Open Resource for Accurately Benchmarking Small Variant and Reference Calls.” *Nature Biotechnology* 37 (5): 561–66.

Chapter IV: Quality control of SARS-CoV-2 genomes

Subchapter A: Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus

As part of my PhD thesis, I am including a section adapted from my published paper *"Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus"*, which was published in *Cell*, DOI: 10.1016/j.cell.2020.06.043

This work was done in collaboration with the Los Alamos National Laboratory, who carried out most of the study, including all of the wet lab work and most of the analyses. I have adapted the text from this chapter to only include the work I contributed, which is detailed in the section of the original publication titled "*Sequence quality control*". I have added background to put the results and interpretation into context, explaining their importance. The corresponding figure for this section (S7/4.1) was plotted by Dr Matt Parker, but I include it here because it shows the output of my work. I did not contribute towards other figures in this publication, so these are not shown. Figures are numbered differently for this thesis chapter, than in the published version.

The work documented in this publication was also part of a larger study "*Benchmarking SARS-CoV-2 Oxford Nanopore Sequencing Pipelines and Generation of a Variant Blacklist*" which I was the primary contributor to — I present a preprint manuscript for this larger study later in this chapter, which used many of the same methods used in the *Cell* publication. I have repeated these where necessary in each subchapter so that they can be read as self-contained sections.

Yours sincerely,

Timothy Freeman (PhD candidate)

Introduction

In the previous chapter I developed a novel method (Freeman et al. 2020) for identifying systematic sequencing biases in clinical whole genome sequencing, using a data structure termed *Incremental Databases* (IncDBs), combined with a Monte Carlo model, to identify genomic positions showing consistent detection of an allele at a systematic allelic fraction across all patients sequenced using the same protocol and sequencing pipeline. In this subchapter I aimed to adapt these methods to apply them to a cohort of haploid SARS-CoV-2 genomes, rather than diploid human genomes, to investigate the flexibility and utility of the approach in a very different scenario. No previous methods had been developed for establishing systematic sequencing biases in SARS-CoV-2, and there were no blacklists generated at the time to flag variants arising from systematic bias. I aimed to identify whether there were any systematic biases present in the ARTIC protocol ONT sequencing pipeline being used for SARS-CoV-2 samples, and if so, to establish what could be the cause of these, so that they could be removed to improve sequencing accuracy.

The modifications needed for applying our previous method to this project included the following: The sample sizes used for calculating the IncDB statistics were increased to 884 to reflect the number of SARS-CoV-2 samples present; the Monte Carlo model was modified to simulate a haploid rather than diploid genotype; the assumed error rate for the Monte Carlo model was increased to reflect the higher observed error rate for ONT vs. Illumina sequencing. In addition to these necessary changes, we also wanted to investigate systematic bias in the detection of indels, which was not carried out in the previous chapter due to computational limitations imposed by the large size of the human genome that would have made runtime too long to be practical. We chose to include systematic bias at indels for SARS-CoV-2 due to its much smaller genome removing these computational limitations, so that we could gauge the importance of these.

Methods

SARS-CoV-2 Sample Collection and Processing

Samples from 884 SARS-CoV-2 positive individuals were obtained from either throat or combined nose/throat swabs. Nucleic acids were extracted from 200µl of each sample using the MagnaPure96 extraction platform (Roche Diagnostics Ltd, Burgess Hill, UK). SARS-CoV-2 RNA was detected using primers and probes targeting the E gene and the RdRp genes of SARS-CoV-2 and the human gene RNaseP, to allow normalisation, for routine clinical diagnostic purposes, with thermocycling and fluorescence detection on ABI Thermal Cycler (Applied Biosystems, Foster City, United States) using previously described primer and probe sets (Corman et al. 2020).

Sample Preparation and Sequencing

Nucleic acids from positive cases underwent long-read whole genome sequencing (Oxford Nanopore Technologies (ONT), Oxford, UK) using the ARTIC Network protocol (accessed the 19th of April 2020, https://artic.network/ncov-2019, https://www.protocols.io/view/ncov-2019-sequencing-protocol-bbmuik6w). In most cases 23 isolates and one negative control were barcoded per 9.4.1D nanopore flow cell. Following base calling, data were demultiplexed using ONT Guppy (--require-both-ends).

Base calling

After initial fast base calling and demultiplexing with guppy (v3.2 - GridION --require-both-ends). We employed 4 additional basecallers, using Oxford Nanopore guppy versions v3.3 (hac3), v4.0 (hac4), and guppy v4.0 in combination with research base calling models (rerio); run length encoded (rle), and flipflop, using default parameter settings.

Mapping & Variant Calling

Using ARTIC Network v1.1.3 we processed pass basecalled data using both nanopolish v0.13.2 (Simpson 2018) and medaka v1.0.1 (Oxford Nanopore Technologies 2018) variant calling modes with default parameter settings.

Reads were filtered based on quality and length (400 to 700bp), then mapped to the Wuhan reference genome (MN908947.3) and primer sites trimmed. Reads were then downsampled to 200x coverage in each direction.

Percent identity was calculated by outputting alignments (primer trimmed) from minimap2 in paf format and dividing column 10 by column 11.

Coverage was calculated by counting non N nucleotides in the consensus sequence produced by ARTIC and divided by the total genome size.

Incremental Database Generation and Systematic Bias Detection

An Incremental Database (IncDB) (Freeman et al. 2020) was generated for the cohort of 884 patients, repeated using each of the 4 basecalling models (hac3, hac4, rle, flipflop), with 4 IncDBs in total. IncDB generation was carried out as described in our previous publication (Freeman et al. 2020), but adapted to apply to the smaller, haploid SARS-CoV-2 genome, as performed in previous quality control efforts (Korber et al. 2020), rather than the human genome.

For each basecalling model, non-reference alleles with allelic fraction standard deviation below the lower standard deviation bound and a mean allelic fraction greater than 5% across patients were considered to have significant levels of systematic bias, since there was consistent low-to-mid level support for those reads across most or all BAM files indicating a systematic bias towards their detection even when absent. Individual sample mutations in all VCF files were annotated with the z value of their respective allelic fraction, z = (variant MAF - systematic MAF)/(allelic fraction standard deviation) , if they occurred at a position at which systematic bias was present above 5%.

The read depth values for each allele (A, C, G, T, DEL, INS) at every autosomal genomic locus were calculated from aligned BAMs and divided by the total read depth at the corresponding loci — including reads that supported indels rather than A, C, G or T — to get the allelic coverage fraction, x_p, for each allele at each locus in each patient. Unlike the previous chapter, allelic fraction systematic biases for indels were examined in this study as well as SNVs, since the SARS-CoV-2 genome is comparably very small and the same . Individual IncDBs were created for each data set from the aggregate allelic fraction and standard deviation values for each allele at each locus across the entire cohort, which were calculated from x_p as described below:

Aggregate allelic fraction = $\frac{1}{N}\sum_{p=1}^{N} x_p$

Standard deviation = $\sqrt{\frac{1}{N}\sum_{p=1}^{N}(x_p - \bar{x})^2}$=$\sqrt{\frac{1}{N}\sum_{p=1}^{N}({x_p}^2) - (\frac{1}{N}\sum_{p=1}^{N} x_p)^2}$

N is the number of patients, *p* is the patient identifier, x_p is the allelic coverage fraction for a specific allele in patient *p*, and $\bar{x}$ is the mean of all of the allelic coverage fractions for that same allele across all patients (aggregate allelic fraction). Notice that to compute the aggregate allelic fraction, we do not store each individual genome's x_p values, but the sum of x_p across all genomes. Similarly, we can compute the standard deviation across genomes by storing the sum of x_p , as well as the sum of ${x_p}^2$. This approach not only removes all individual-specific genomic information, but also allows the IncDB to grow indefinitely, as more samples are sequenced and analyzed: They can simply contribute to the running sums of x_p and ${x_p}^2$. Also note that the sums in the equations above do not take up more size on disk as the number of samples increases, so the overall IncDB file size does not increase as new samples are added.

Identifying loci affected by systematic bias (suspect loci)

For each locus in all autosomal chromosomes, the standard deviation and aggregate allelic fraction values were taken from the IncDB and plotted against each other in a density plot using python scripts. The main bow-shaped feature of this plot was the expected result of haploid alleles present at a variety of population frequencies, while observed positions with standard deviations below the 99.9% expected confidence interval for a given nucleotide were defined as suspect loci for that allele. Positions that displayed at least one suspect allele at that position were termed unique suspect loci. The total count of unique suspect loci was therefore lower than the total count of allele-specific suspect loci, since some positions had multiple suspect alleles.

The 99.9% confidence interval was estimated using Monte Carlo sampling as detailed in the pseudocode below. Monte Carlo sampling used three nested loops which respectively simulated the standard deviation at a single genomic position between *n* individual sample allelic fractions (loop 3), 1000 times to calculate the upper and lower 99.9% confidence intervals (loop 2), for each aggregate allelic fraction from 0 to 1 in intervals of 0.01 (loop 1). The standard deviation values were recorded and used to classify suspect loci with *n* = 884. The model assumed an error rate of 0.05, corresponding to an approximation of the observed error rate of ONT WGS in the output sequencing data.

Monte Carlo simulation of standard deviation (pseudocode)

The Monte Carlo model was generated for a haploid cohort of 884 individual samples with a mean error rate of 0.05, reflecting what was observed in the ONT sequencing, in order to calculate the 0.1% lower and upper bounds of the standard deviation confidence interval.

1. For aggregate allelic fractions, AAF, from 0 to 1 in intervals of 0.01 (each representing a simulated single autosomal genomic position with that aggregate allelic fraction across all patients) do {

    2. repeat 1000 times {

        3. repeat for *n* simulated patients {

        Randomly generate haploid genotype for each simulated patient using the binomial distribution (assuming the major and minor allelic fractions sum to 1) at the given AAF value;

        Assuming a sequencing error rate of 0.05, randomly draw *c* reads from the binomial distribution to simulate observed major/minor allelic reads for the simulated biallelic diploid genotype. No positions were modelled as multiallelic;

        Divide by total read depth, *c*, to get the individual allelic fractions for each patient;

        }

    Calculate the standard deviations between the individual allelic fractions for all *n* patients at the simulated genomic position;

    }

Maximum and minimum values of 1000 repetitions mark upper and lower 99.9% confidence intervals for standard deviation at given AAF;

}

Results

We created an IncDB from a cohort of ONT-sequenced Sheffield SARS-CoV-2 sequences, and used this to flag suspect loci. We discovered a systematic sequencing bias that gave rise to what appeared at first to be a mutation of interest at position 943 (24389 A > C and 24390 G > C) in the spike protein that was evident in sequences from Belgium, but was actually a false positive variant. It was frequent enough to be a site of interest, and was tracked since there were concerns that it could be rapidly increasing in population frequency as a variant that increased the ability of SARS-CoV-2 to spread. We interrogated these positions in the raw sequencing data from Sheffield, and although these two variants were not present in the final consensus sequence from any of the Sheffield isolates, the raw, untrimmed bam files showed their presence in only one of the amplicons covering the site (**Figure 4.1**). We noticed that this position was to the left of the 5' primer of amplicon 81 in what we believed to be an adapter sequence. Comparison of the Wuhan reference and the adapter sequence revealed similarity around this position:

Nanopore adapter sequence:

CAGCACCTT

The Wuhan reference sequence:

CAGCAAGTT

We saw a C present at around 50% of called bases at both these positions in the raw sequencing reads. After trimming the nanopore adapter sequences used by the ARTIC pipeline, the two mutations of interest were no longer supported by the remaining sequencing reads, which supported the reference alleles at an allelic fraction of nearly 100%, and did not contribute to the final consensus sequence. Although it is evident in amplicon 81, in this region, there is no evidence for these variants in the data from amplicon 80, which also covers these positions. We include a figure (**Figure 4.1**) to explain our finding.

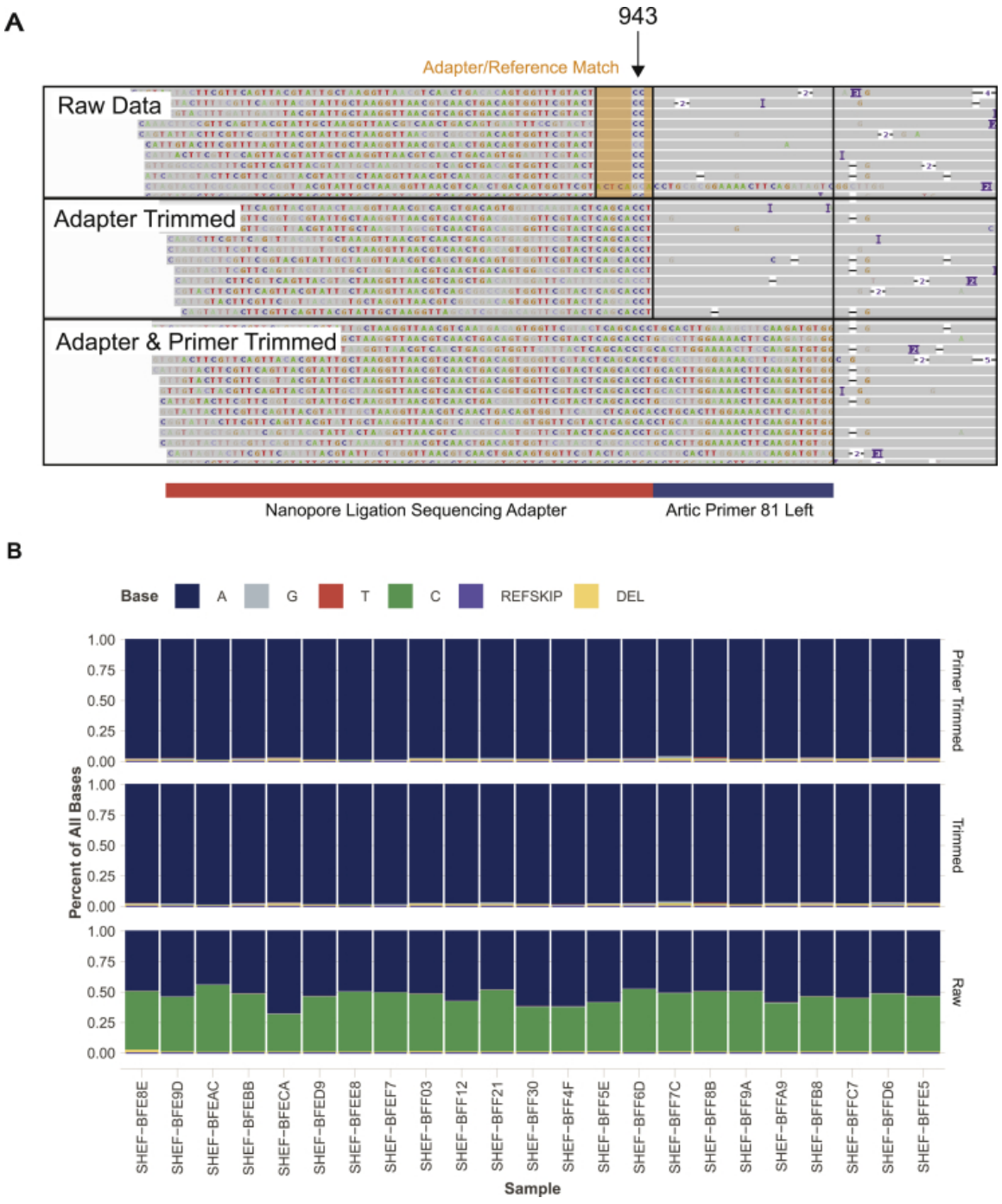


Figure 4.1: Investigation of S943P

A. IGV plots showing bam files from nanopore sequencing data of amplicons produced by the ARTIC network protocol. Raw data from amplicon 81 contains a portion of an adapter sequence which is homologous to the reference genome, apart from the C variants which lead to a S943P mutation call. This region is therefore included in variant calling if location-based trimming is not carried out. Subsequent panels show that this region is soft clipped when trimming adapters and primers and is therefore not available for variant calling. **B.** Base frequencies at position 24,389 in 23 samples from the Sheffield data show that C is present in half of the reads in the raw data, but is absent from trimmed and primer trimmed data.

Discussion

We employed an IncDB-based approach similar to the previous chapter, but with important modifications to the IncDB generation, Monte Carlo model used for suspect locus annotation, and the choice to include indels for systematic bias analysis. This showeds that the methods presented in Chapter III could be applied to a haploid viral genome sequenced using a different sequencing technology. We demonstrated the flexibility of this technique by identifying an error that has arisen due to a combination of improper trimming of adapter and primer regions from raw sequencing reads before downstream analysis, and the coincidental homology between the nanopore adapter sequence and the Wuhan reference genome in this region. This is included here as a cautionary note; resolving rare biological mutations and sequencing error will be an important balance going forward in terms of interpretation of rare mutations (De Maio et al. 2020a). A recurrent amino acid change like L5F could potentially result from a recurrent sequencing or sequence processing error (De Maio et al. 2020a), or alternatively, it may be of particular interest if it is naturally recurring homoplasy. These results demonstrate the value of IncDB-based methods across species and sequencing pipelines. The previous chapter used these to provide sequence quality control for Illumina sequencing, while in this chapter I showed that it was valuable in ONT sequencing also, even in much smaller, haploid genomes. Korber et al. contacted the group in Belgium who had recorded high frequencies of the S943P variant to inform them of our findings. They already suspected a sequencing issue existed at this position, concurred with our interpretation, and contacted GISAID with a request to remove the problematic sequences. As a result of our findings this variant was excluded from the list of mutations of concern that was published in *Cell* for the study, and was also added to a prominent SARS-CoV-2 variant blacklist (De Maio et al. 2020b).

Subchapter B: Benchmarking SARS-CoV-2 Oxford Nanopore Sequencing Pipelines and Generation of a Variant Blacklist

As part of my PhD thesis, I am including a manuscript *"Benchmarking SARS-CoV-2 Oxford Nanopore Sequencing Pipelines and Generation of a Variant Blacklist"*, that I have prepared for submission as a journal article.

I am the primary author of this work, done under the guidance of my PhD supervisor Dr Dennis Wang and my colleague Dr Matt Parker, who is also the corresponding author. Dr Matt Parker implemented the ARTIC pipeline, base calling and variant calling to sequence the samples. I designed and carried out all of the data analyses detailed in this publication, drafted and edited the manuscript, and produced all results, figures and tables.

Yours sincerely,

Timothy Freeman (PhD candidate)

Introduction

Background

SARS-CoV-2 is a betacoronavirus that is the causative agent of the respiratory disease COVID-19 (Coronaviridae Study Group of the International Committee on Taxonomy of Viruses 2020) that was first reported in China in December 2019 (Zhou et al. 2020) and has resulted in an ongoing pandemic. SARS-CoV-2 has a 29.9kb RNA-based genome, which encodes four canonical 3′ structural proteins, termed the spike (S), envelope (E), membrane (M) and nucleocapsid (N) proteins and a 5′ frameshifted polyprotein (ORF1a/ORF1ab) found in all coronaviruses, in addition to other accessory proteins (Ashour et al. 2020). Accurate sequencing of the SARS-CoV-2 genome is vital in order to study the structure and function of SARS-CoV-2 genes to guide the development of vaccines and other treatments (Michel et al. 2020), and to identify and track variants as they emerge globally, so that their effects may be characterised and appropriate public health actions taken in response (Korber et al. 2020). However, while SARS-CoV-2 sequencing accuracy has improved rapidly in the past year, it is

important to regularly assess to what extent apparent viral sequence diversity is caused by sequencing errors as opposed to being real variation (Kubik et al. 2020; Bull et al. 2020).

ONT sequencing of SARS-CoV-2

An abundance of sequencing data has been generated for the SARS-CoV-2 virus, available at large-scale data repositories such as GISAID (Elbe and Buckland-Merrett 2017; Shu and McCauley 2017), which contains over 1,000,000 SARS-CoV-2 viral sequences (accessed 13/03/2021). National and international initiatives have been established to sequence samples accurately in order to trace the epidemiology of the virus and monitor the emergence of new lineages (COVID-19 Genomics UK (COG-UK) 2020; Korber et al. 2020; Tegally et al. 2021). Many of the existing SARS-CoV-2 sequences available have been sequenced using ARTIC network protocols (N. Loman, Rowe, and Rambaut 2020) followed by Oxford Nanopore Technology (ONT) sequencing (Korber et al. 2020; Franco-Muñoz et al. 2020; Kumar et al. 2020). ONT sequencing benefits from realtime, simple and low-cost protocols, providing advantages for sequencing viral genomes accurately and allowing early termination of sequencing when sufficient data has been produced (Pollard et al. 2018; Depledge et al. 2019). Base calling models for ONT sequencing are rapidly evolving (Amarasinghe et al. 2020) and in combination with improvements to the ARTIC/ONT sequencing protocol (Itokawa et al. 2020; Tyson et al. 2020) sequencing quality has increased over the course of the pandemic. Assessing the differences in results across these tools would assist with understanding the effect on sequencing accuracy of the choices made when analysing SARS-CoV-2 sequencing data. However, a systematic comparison of these has not yet been carried out.

Existing quality control approaches

Existing approaches for quality control of sequencing data commonly include setting thresholds on read coverage, quality and mutant allele fraction (MAF) when calling mutations and excluding known sequencing artefacts and mutations that are not concordant across different sequencing methods and analyses. For SARS-CoV-2 specifically, we aimed to compare our results with two established mutation blacklists from the literature (Bull et al. 2020; De Maio et al. 2020b). These blacklists incorporated standard quality control methods, but also some additional features: Bull et al. 2020 blacklisted 15 regions of low sequence complexity (i.e. with homopolymeric or repetitive content) at which mutations showing discordance between ONT and Illumina sequencing were enriched. De Maio et al. 2020 blacklisted mutations exhibiting a range of features, including ambiguous mutations only found in a small number of sequencing centres, mutations removed from GISAID sequences upon

recalling, mutations confirmed as false-positives resulting from specific nanopore adapter sequences or interspecific contamination, and mutations showing a high level of homoplasy — defined as appearing to emerge de novo at high rates inconsistent with standard mutation rates, suggesting that they were sequencing artefacts. De Maio et al. also blacklisted mutations that had near perfect linkage to other proximal blacklisted mutations. Bull et al. and De Maio et al. blacklist mutations at 310 and 477 unique positions respectively, including 6 positions that have mutations in both blacklists.

Systematic sequencing bias in ONT-sequenced SARS-CoV-2 samples

We define systematic sequencing bias as consistent, erroneous detection of reads supporting a specific mutation at a similar allelic fraction across all samples sequenced using the same sequencing technology and protocol. We have previously published a method for identifying systematic sequencing bias (Freeman et al. 2020), and adapted this to the haploid SARS-CoV-2 genome. In this method we catalogue the allelic fraction values of all alleles at all genomic loci across a large sample cohort in a structure called an Incremental Database (IncDB) and identifying where the standard deviation in allelic fraction values is significantly lower than would be predicted by a Monte Carlo model. Using this approach, we are able to confirm false positive mutation calls at genomic loci affected by systematic bias if the sample mutation allelic fraction (MAF) is not significantly higher than the systematic bias allelic fraction. Even if the numbers of mutations affected by systematic sequencing bias are low, assessing this is still valuable since it confirms false positive mutations that are likely to be called in many samples. It also helps with identifying and removing sources of systematic bias in a sequencing pipeline that might not have been identified previously or addressed with other quality control measures, since this method was only published recently. In a recent study we used this approach (Korber et al. 2020) to reveal examples of false positive mutations that were missed by other quality control methods, such as the commonly reported false positive mutation S943P in the SARS-CoV-2 spike protein. This mutation was highlighted by its systematic inclusion at roughly 50% of called bases at the corresponding genomic positions (24389 A>C and 24390 G>C) when carrying out sequencing without first trimming the ONT adapter sequence, which is homologous to this part of the SARS-CoV-2 reference genome. Trimming adapter sequences soft-clips this region in amplicon 81, leading to this mutation no longer being called, while no systematic bias exists among the remaining called bases at the same positions derived from amplicon 80. This shows that the basecalls supporting this mutation did not originate from the SARS-CoV-2 sequence itself.

Study aims

Here we present analysis of 884 SARS-CoV-2 genomes using the ARTIC Network protocol (Tyson et al. 2020) and subsequent ONT sequencing (Amarasinghe et al. 2020; Itokawa et al. 2020), and compare results with four base calling models and two variant calling workflows. We describe a range of novel methods to blacklist SARS-CoV-2 mutations in ONT sequencing data (**Figure 4.2**), including evaluating systematic bias, deletion-prone loci and discrepancies in mutation calling between genomic pipelines. By comparing our blacklisted mutations against these other mutation blacklists we aim to assist the SARS-CoV-2 genomics community in avoiding miscalling mutations, so that they can accurately identify new mutations of concern and their patterns of circulation.

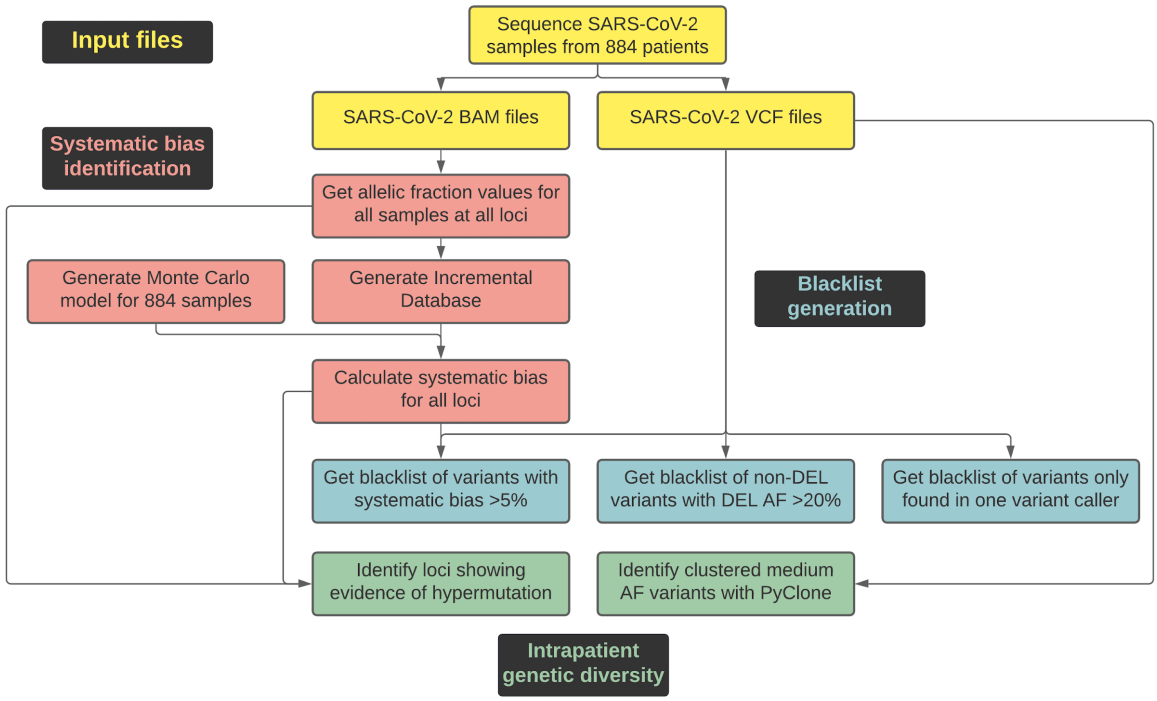


Figure 4.2: Summary of the steps taken to blacklist mutations prone to a range of sequencing errors.

Methods

SARS-CoV-2 Sample Processing

Samples from 884 SARS-CoV-2 positive individuals were obtained from either throat or combined nose/throat swabs. Nucleic acids were extracted from 200µl of each sample using the MagnaPure96 extraction platform (Roche Diagnostics Ltd, Burgess Hill, UK). SARS-CoV-2 RNA was detected using primers and probes targeting the E gene and the RdRp genes of SARS-CoV-2 and the human gene RNaseP, to allow normalisation, for routine clinical diagnostic purposes, with thermocycling and fluorescence detection on ABI Thermal Cycler (Applied Biosystems, Foster City, United States) using previously described primer and probe sets (Corman et al. 2020).

Sample Preparation and Sequencing

Nucleic acids from positive cases underwent long-read whole genome sequencing (Oxford Nanopore Technologies (ONT), Oxford, UK) using the ARTIC Network protocol (accessed the 19th of April, https://artic.network/ncov-2019, https://www.protocols.io/view/ncov-2019-sequencing-protocol-bbmuik6w). In most cases 23 isolates and one negative control were barcoded per 9.4.1D nanopore flow cell. Following base calling, data were demultiplexed using ONT Guppy (--require-both-ends).

Base calling

After initial fast base calling and demultiplexing with guppy (v3.2 - GridION --require-both-ends). We employed 4 additional basecallers (**Table 4.1**), using Oxford Nanopore guppy versions v3.3 (hac3), v4.0 (hac4), and guppy v4.0 in combination with research base calling models (rerio); run length encoded (rle), and flipflop, using default parameter settings.

Guppy base calling model	Key attributes
High accuracy v3.3 (hac3)	Default base calling model, existing version at the start of the SARS-CoV-2 outbreak.
High accuracy v4.0 (hac4)	Default base calling model, most up-to-date version at the time of sequencing.
High accuracy v4.0 + run length encoded model (rle)	Same as hac4, except uses a model in which read sequence is encoded in a homopolymer-compressed form, proposed to give better performance at repetitive regions.
High accuracy v4.0 + flipflop model (flipflop)	Same as hac4, except uses the flipflop base calling model, which is slower due to the more complex neural network involved, but thought to give better read accuracy.

Table 4.1: Comparison of Guppy base calling models' key attributes

Mapping & Variant Calling

Using ARTIC Network v1.1.3 we processed pass basecalled data using both nanopolish v0.13.2 (Simpson 2018) and medaka v1.0.1 (Oxford Nanopore Technologies 2018) variant calling modes with default parameter settings.

Reads were filtered based on quality and length (400 to 700bp), then mapped to the Wuhan reference genome (MN908947.3) and primer sites trimmed. Reads were then downsampled to 200x coverage in each direction.

Percent identity was calculated by outputting alignments (primer trimmed) from minimap2 in paf format and dividing column 10 by column 11.

Coverage was calculated by counting non N nucleotides in the consensus sequence produced by ARTIC and divided by the total genome size.

SARS-CoV-2 viral lineages were determined using Pangolin (https://github.com/cov-lineages/pangolin) , with default settings.

Incremental Database Generation and Systematic Bias Detection

An Incremental Database (IncDB) (Freeman et al. 2020) was generated for the cohort of 884 patients, repeated using each of the 4 basecalling models (hac3, hac4, rle, flipflop), with 4 IncDBs in total. IncDB generation was carried out as described in our previous publication (Freeman et al. 2020), but adapted to apply to the smaller, haploid SARS-CoV-2 genome, as

performed in previous quality control efforts (Korber et al. 2020), rather than the human genome. Standard deviation values for allelic fractions across the cohort were calculated along with the mean allelic fraction values for each allele A, C, G, T, DEL, INS at each position of the SARS-CoV-2 genome. The Monte Carlo model was generated for a haploid cohort of 884 individual samples with a mean error rate of 0.05, reflecting what was observed in the ONT sequencing, in order to calculate the 0.1% lower and upper bounds of the standard deviation confidence interval. For each basecalling models, non-reference alleles with allelic fraction standard deviation below the lower standard deviation bound and a mean allelic fraction greater than 5% across patients were considered to have significant levels of systematic bias, since there was consistent low-to-mid level support for those reads across most or all BAM files indicating a systematic bias towards their detection even when absent. Individual sample mutations in all VCF files were annotated with the z value of their respective allelic fraction, z = (variant MAF - systematic MAF)/(allelic fraction standard deviation) , if they occurred at a position at which systematic bias was present above 5%.

Processing and Annotation of Merged VCF Files

Merged VCF files were generated by the ARTIC pipeline described above for all combinations of the two variant callers, nanopolish and medaka, and the four basecalling models, hac3, hac4, rle and flipflop, for all 884 patient samples. medaka VCF files were processed with VT to decompose multiallelic variants (Tan, Abecasis, and Kang 2015), so that variants would be in the same format for medaka and nanopolish. Allelic depths, fractions and systematic biases were annotated as additional columns for each variant. VCF files were filtered to remove any variants with QUAL<20 or total read depth < 20. Merged VCF files were formed by combining the outputs from two separate VCF files corresponding to the different strand directionality of the adapter sets used for amplification. It was therefore possible for the same variant to be supported in both of the constituent VCF files where there was overlap, resulting in a duplicate in the merged VCF. If the same variant was duplicated within a VCF file for a single patient, only the higher quality variant was kept. All VCF processing and annotation was carried out within the main python script included in the code.

Annotation of Blacklisted Mutation Peaks

The 30,000 bp SARS-CoV-2 genome was divided up into 150 consecutive 200bp bins. Any bins at which blacklisted mutations covered more than 10% of genomic positions across the bin were classified as blacklisted mutation peaks. If consecutive peaks occurred, they were collectively considered as the same peak, resulting in 5 genomic peaks overall with more

than 10% of positions covered by blacklisted mutations: 4200-4600, 5200-5800, 10000-10600, 18000-18400 and 22800-23000.

Blacklist Classifications

Mutations were blacklisted for “caution” or full masking if they fell within one or more of the following five subsets:

Caution:

1) Mutations with systematic bias > 5% for that allele.
2) NonDEL20 mutations: Non-deletion mutations at positions where more than 20% of sequencing reads supported a deletion in the patient sample. These always occurred adjacent to a homopolymer. These were also manually examined in IGV (Robinson et al. 2011) to produce the output in **Figure 4.5**.
3) Mutations which were exclusively detected with one variant caller (either medaka or nanopolish).
4) Mutations which were not called across all four basecalling models used in this study (HAC3, HAC4, RLE, FLIPFLOP), but only a subset of these, where the mutation was called with at least one basecalling model.

Mask (takes precedence over caution if mutation also in caution blacklist):

1) Mutations with systematic bias > 5% for that allele, which did not have a mutation allelic fraction significantly (two standard deviations) higher than the systematic allelic fraction.
2) In cases where separately blacklisted mutations were recommended for masking in the (De Maio et al. 2020b) or (Bull et al. 2020) blacklists, the recommendation was upgraded to “mask” if not already recommended for full masking. These “caution” and “masking” definitions are used by the python command line tool “sarscov2vcfblacklister.py” which we share for other users to annotate and filter their SARS-CoV-2 ONT-sequenced VCF files using our mutation blacklist.

Genomic Region Definitions/Sources

Regions used for the Fisher Exact tests in **Table 4.6** were defined as follows and calculated in the R code or downloaded from the listed source:

Homopolymers: A section of the SARS-CoV-2 genome containing only one type of base, repeated for at least 5bp.

Homopolymer 3bp and 100bp flanks: The 3bp and 100bp flanks respectively on both sides of each homopolymer, as defined above.

GCgt99p: Genomic loci in which the surrounding 50bp flanks on both sides have in total >52 GC bases. This is the 99th percentile for GC content across the SARS-CoV-2 reference genome.

ARTIC primers: These are the SARS-CoV-2 genomic regions to which the ARTIC primers anneal during the ARTIC sequencing pipeline.

PyClone Visualisation of MAF Clusters

VCF files for individual samples were converted into the TSV input format for PyClone within the main python script. The PyClone command “PyClone run_analysis_pipeline” was used for each sample individually, with 10,000 iterations used for the model in each case (--num_iters 10000). The PyClone model was run with the sequencing error rate set to 0.05; All other parameters were left with default values.

Statistical Tests

Fisher Exact tests used to calculate odds ratios for tables 4.5 & 4.6 showing enrichment of blacklisted loci subsets within each other and within genomic regions of interest. Bonferroni correction applied to p-values to determine statistical significance.

Data access

All code used in this study, to repeat these analyses and reproduce the associated tables and figures, as well as a python command line tool named sarscov2vcfblacklister.py, for annotating and filtering blacklisted mutations in ONT-sequenced SARS-CoV-2 VCF files, is available at https://github.com/tmfreeman400/SARS-CoV-2_blacklist_code, along with instructions. Files containing the ARTIC primers, qPCR primers, SARS-CoV-2 reference genome fasta and sgRNAs, which are used as inputs in the main R script, are also provided in this folder. Fasta files for all of the sequencing data are available on GISAID for all 884 sequenced samples (Elbe and Buckland-Merrett 2017; Shu and McCauley 2017) (https://www.gisaid.org/) and

COG-UK FASTQs are available on the European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/view/) with study accession ERP121228.

The reference fasta file for SARS-CoV-2 was downloaded from https://raw.githubusercontent.com/artic-network/artic-ncov2019/master/primer_schemes/nCoV-2019/V3/nCoV-2019.reference.fasta . The Bull et al. blacklist was downloaded from the publication “Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis” (Bull et al. 2020), while the De Maio et al. blacklist (De Maio et al. 2020b) was downloaded on 13th April 2021 from https://raw.githubusercontent.com/W-L/ProblematicSites_SARS-CoV2/master/problematic_sites_sarsCov2.vcf .

Results

Identifying mutations that are not consistently called across variant callers and base calling models

Swab samples from 884 individuals positive for SARS-CoV-2 by RT-PCR were sequenced using the ARTIC Network protocol (Tyson et al. 2020) and subsequent Nanopore sequencing. Base calling was repeated with four different models for comparison. These were Guppy production high accuracy v3.3 (hac3), and v4.0 (hac4), and using experimental base calling models; run length encoded (rle) and flipflop (later forming the basis of hac4). After base calling, two different variant calling workflows were employed — using medaka or nanopolish as the variant caller — giving a total of eight different pipeline combinations. The frequencies with which mutations were called within the cohort (**Figure 4.3A,B**), and the fractions of sequencing reads supporting each mutant allele within each sample were calculated (**Figure 4.3C,D**) for each combination of base calling model and variant calling workflow.

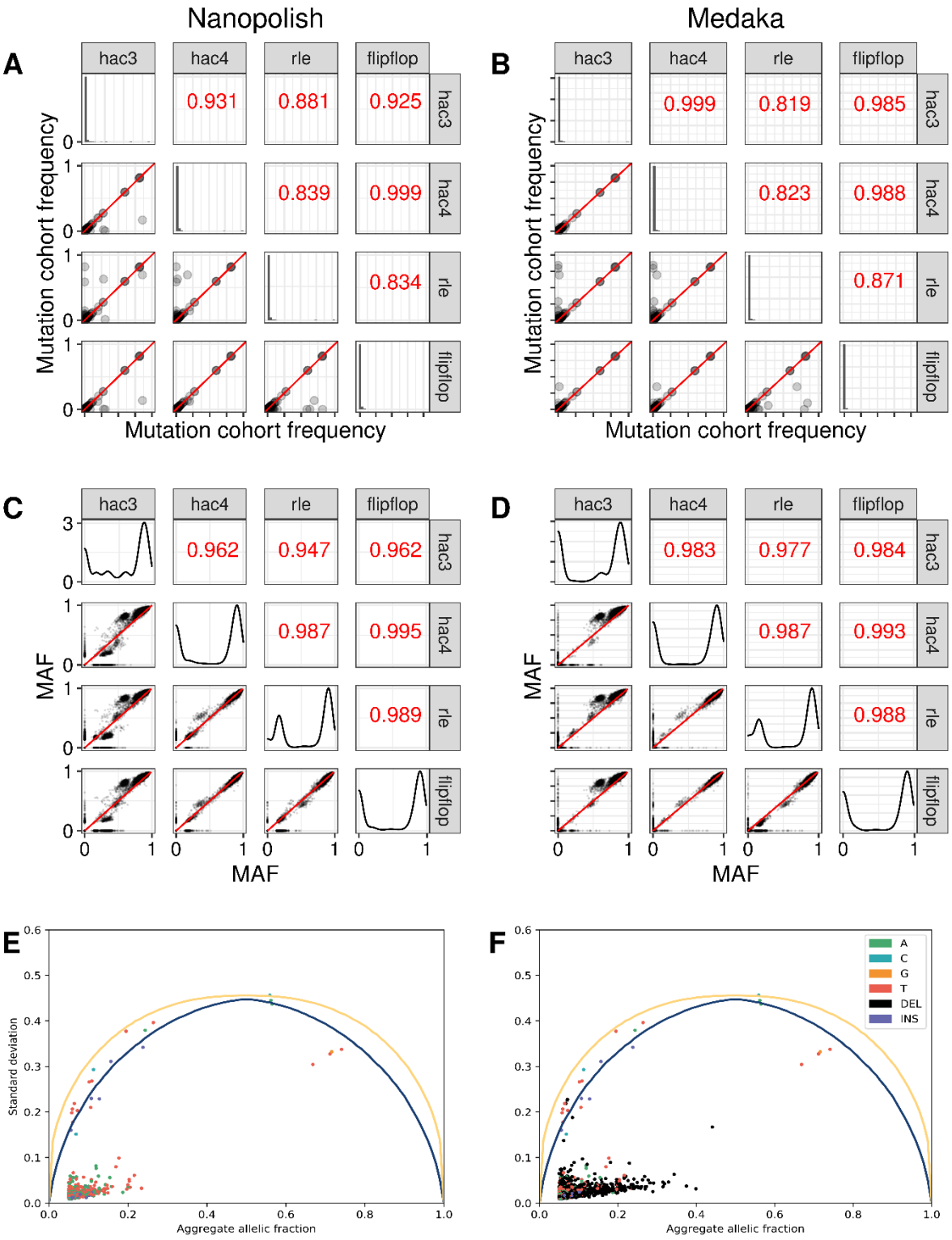


Figure 4.3: Mutation cohort frequency, MAF and systematic bias of all SARS-CoV-2 mutations in 884 patient samples.

A-D: Comparisons of results using four base calling models (hac3, hac4, rle, flipflop). Mutation cohort frequency **(A,B)** and MAF **(C,D)** values displayed in the bottom left grids, with their distributions shown in the diagonal grids. Pearson correlations between base calling models shown in the top right grids. Mutation cohort frequency indicates the fraction of individual samples within the cohort which had that mutation, with one point plotted for each unique mutation detected across the cohort, while

MAF refers to the allelic fraction at which each mutation was detected in an individual sample, with each point corresponding to a mutation in a single sample and not the mean MAF across samples. **E,F:** Standard deviation between the individual allelic fractions for the cohort of 884 ARTIC/hac4 sequenced SARS-CoV-2 genomes versus the aggregate (mean) allelic fraction, for each allele at each genomic locus. The upper and lower 0.1% confidence intervals for expected standard deviation from the Monte Carlo model are plotted as curves over the scatterplot. All points under the lower confidence interval represent alleles that have systematic bias towards their detection at a given allelic fraction regardless of actual mutation presence. Reference alleles and all alleles with an aggregate allelic fraction below 5% of supporting reads for the corresponding allele are not shown since these are numerous and largely unaffected by systematic bias. This explains why no alleles are visible on the plot under 5% AAF or near 100% AAF. Panels **E** and **F** display all alleles excluding/including deletions respectively.

More than 98% of unique SARS-CoV-2 mutations in our cohort were rare (<2.5% cohort frequency) or only occurred in one sample. Seven unique mutations were consistently called across basecallers and variant callers at a high frequency (>30% of samples). These were A23403G, C3037T, C14408T, C241T (~80% cohort frequency), and G28881A, G28882A, G28883C (~60% minimum cohort frequency), corresponding with the cohort frequencies of SARS-CoV-2 lineages descended from B.1 and B.1.1 respectively, in which these mutations are described. This was confirmed using Pangolin to classify these sequences into their respective lineage (**Figure 4.4**). The low number of mutations within the standard deviation bounds, i.e. not displaying systematic bias, was as expected since SARS-CoV-2 has a low mutation rate and very small genome size, and there were few genomic differences between the lineages observed at the time of sequencing. In accordance with this, the cohort frequencies we observed for these mutations had similar values to the general population frequencies reported by Pangolin, and accounted for all of the lineage markers that differed across the lineages. The run-length-encoded base calling model (rle) uses reads stored in a homopolymer-compressed form and showed the lowest Pearson correlation values with the other base calling models for both variant callers (0.8-0.9 with all rle combinations), while all other combinations shared standard read encoding and had higher Pearson correlation values for the mutations called (>0.9).

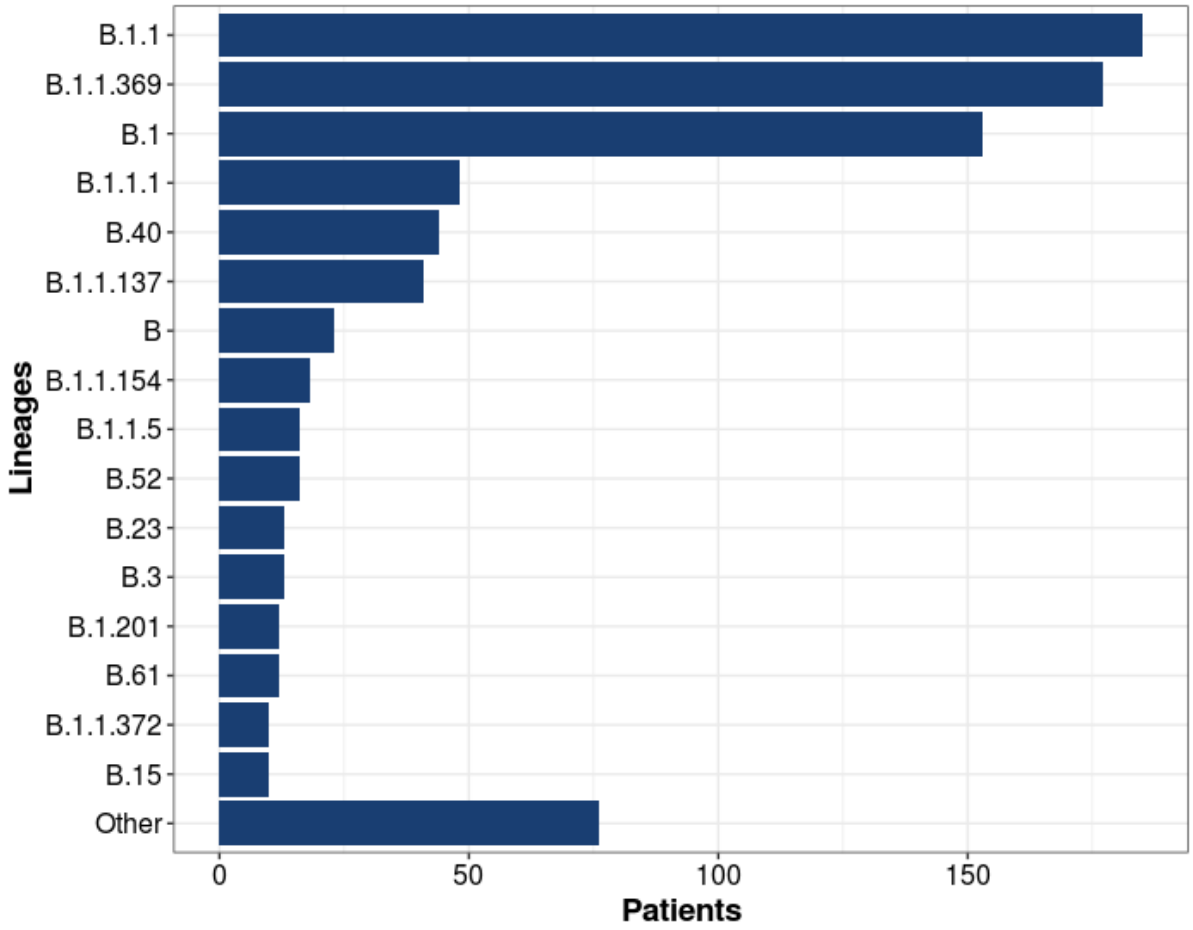


Figure 4.4: Viral lineages of SARS-CoV-2 samples in our cohort, calculated using Pangolin. Lineages with fewer than ten samples included in "Other".

Most mutations were called at high MAFs, with >75% of reads supporting the mutation across all combinations of base calling models and variant calling workflows. A high Pearson correlation was observed between all basecallers (0.987-0.995) except for hac3 (0.947-0.962), which displayed a large cluster of called mutations at 50-60% allelic fraction that were at >75% allelic fraction with the other basecallers (**Figure 4.3C,D**). Most of the mutations in this cluster were either C3037T or T12184C - both of which were adjacent to homopolymers and had higher levels of reads supporting deletions with hac3 base calling compared to the other basecallers, explaining this discrepancy in MAF values. The hac3/nanopolish combination also called a cluster of 710 mutations at an MAF of 25-40% that were called with an MAF of <25% with other basecallers, 557 of which were CT24981C. CT24981C was never called with medaka, but did not display any other blacklist features.

Identifying mutations prone to systematic sequencing bias

Next we sought to identify if any mutations were the result of systematic bias. We have previously applied a method to detect systematic bias in Human Illumina WGS data (Freeman et al. 2020), which we adapted to the haploid SARS-CoV-2 genome in this study, and applied after adapter and primer trimming to remove systematic bias mutations such as S943P in the spike protein that we identified earlier as resulting from adapter sequence homology with the SARS-CoV-2 reference sequence (Korber et al. 2020).

By summarising the 884 BAM files available for each of the four base calling algorithms tested, using our previously published method (Freeman et al. 2020), we were able to catalogue all SARS-CoV-2 alleles affected by systematic bias across each base calling model (systematic biases in hac4 shown in **Figure 4.3E,F** as an example) and determine if any mutations were called at these positions.

The number of SARS-CoV-2 genomic loci which displayed systematic bias of at least 5% allelic fraction was counted for each allele, for each base calling model used (**Table 4.2**), revealing that there was a consistent, systematic presence of reads detected for all alleles at many positions. The hac4 base calling model displayed the fewest loci with systematic bias >5%, for every allele. Systematic detections of deletions around 35-52% allelic fraction were the highest allelic fraction systematic biases observed. Deletions were also the most frequent source of systematic bias, comprising 49.2-61.5% of systematic biases across all basecallers, followed by A/T bias and insertions.

Basecaller/ allele	hac3	hac4	rle	flipflop
A	691 (13.2)	312 (13.2)	514 (15.0)	359 (13.7)
C	285 (5.4)	40 (1.7)	136 (4.0)	47 (1.8)
G	298 (5.7)	42 (1.8)	142 (4.1)	43 (1.6)
T	836 (15.9)	340 (14.4)	597 (17.4)	415 (15.9)
DEL	2579 (49.2)	1454 (61.5)	1495 (43.6)	1556 (59.6)
INS	555 (10.6)	178 (7.5)	548 (16.0)	191 (7.3)

Table 4.2: Systematic bias breakdown by allele and basecaller.
Number of genomic loci with systematic bias >5% for each allele (A, C, G, T, DEL, INS) with each base calling algorithm (hac3, hac4, rle, flipflop). Relative allelic percentages given in brackets. Hac4 exhibited the lowest numbers of loci with systematic bias >5% for every allele.

Across all workflows examined, 19 unique mutations were called at positions where the mutant allele had systematic bias >5%, indicating that caution should be taken when calling these mutations. Of these 19 unique mutations, 5 called mutations (G1265A, G2539A, G10704A, G18002A, G18317A) had allelic fractions that were not significantly higher than the systematic allelic fractions found within the IncDB for the corresponding basecaller, suggesting that the mutation was likely to be a false positive. All of these 5 mutations were called exclusively with the medaka variant caller, and were not called with nanopolish. These mutations are recommended for masking.

Non-deletion mutations found at positions where a high proportion of reads support deletions

Deletions were the most common allele to be incorrectly supported by a consistent fraction of reads at many positions across the cohort, as a result of systematic bias towards deletions detected by the IncDB Monte Carlo algorithm at these positions (**Figure 4.3E,F, Table 4.2**). We therefore sought to examine positions with a high proportion of reads supporting deletions, including cases where this was sporadic rather than a systematic bias, to identify mutations that could be susceptible to sequencing error. We examined all 68 non-deletion mutations with >20% sequencing reads supporting a deletion, which we refer to as “NonDEL20” mutations. 66 of these (97%) were called with medaka, while 55 (81%) were called with nanopolish, likely

reflecting the larger number of mutations in general called by medaka, but still showing that both variant calling workflows called a large proportion of NonDEL20 mutations identified. The majority (68%) of these mutations were called as a T. All NonDEL20 variants, including non-T mutations, occurred adjacent to, or as part of, a 2bp+ homopolymer matching the reference or mutant allele, in 93% of cases. Two examples of these, which were called by both variant callers used, and with all four base calling models, are shown in **Figure 4.5**. This seemed to indicate that ONT sequencing was prone to manifesting mutations at positions adjacent to their respective homopolymers, especially T mutations, this seems to occur in tandem with a high proportion of reads supporting a deletion at these positions. The preference for T NonDEL20 variants could be partially explained by the overrepresentation of T homopolymers within the SARS-CoV-2 genome: 49% of homopolymers of length 5bp or more corresponded to T homopolymers. However, since the preference for T NonDEL20 variants was stronger than this, it was not a sufficiently high percentage to wholly explain this phenomenon. In addition, there was an equal (49%) proportion of A homopolymers of this length, but the proportion of A NonDEL20 variants was much lower by comparison (10%), indicating that the high occurrence of T NonDEL20 variants was mostly due to a preference for bad base calls being T rather than due to a high prevalence of T homopolymers. Similar observations supporting this have previously been made in human (Cornelis et al. 2017) and bacterial (N. J. Loman, Quick, and Simpson 2015) ONT sequencing and are thought to occur due to random changes in the speed at which DNA strands are passed through the nanopore (Szalay and Golovchenko 2015), which make it difficult to accurately measure how long homopolymeric stretches are.

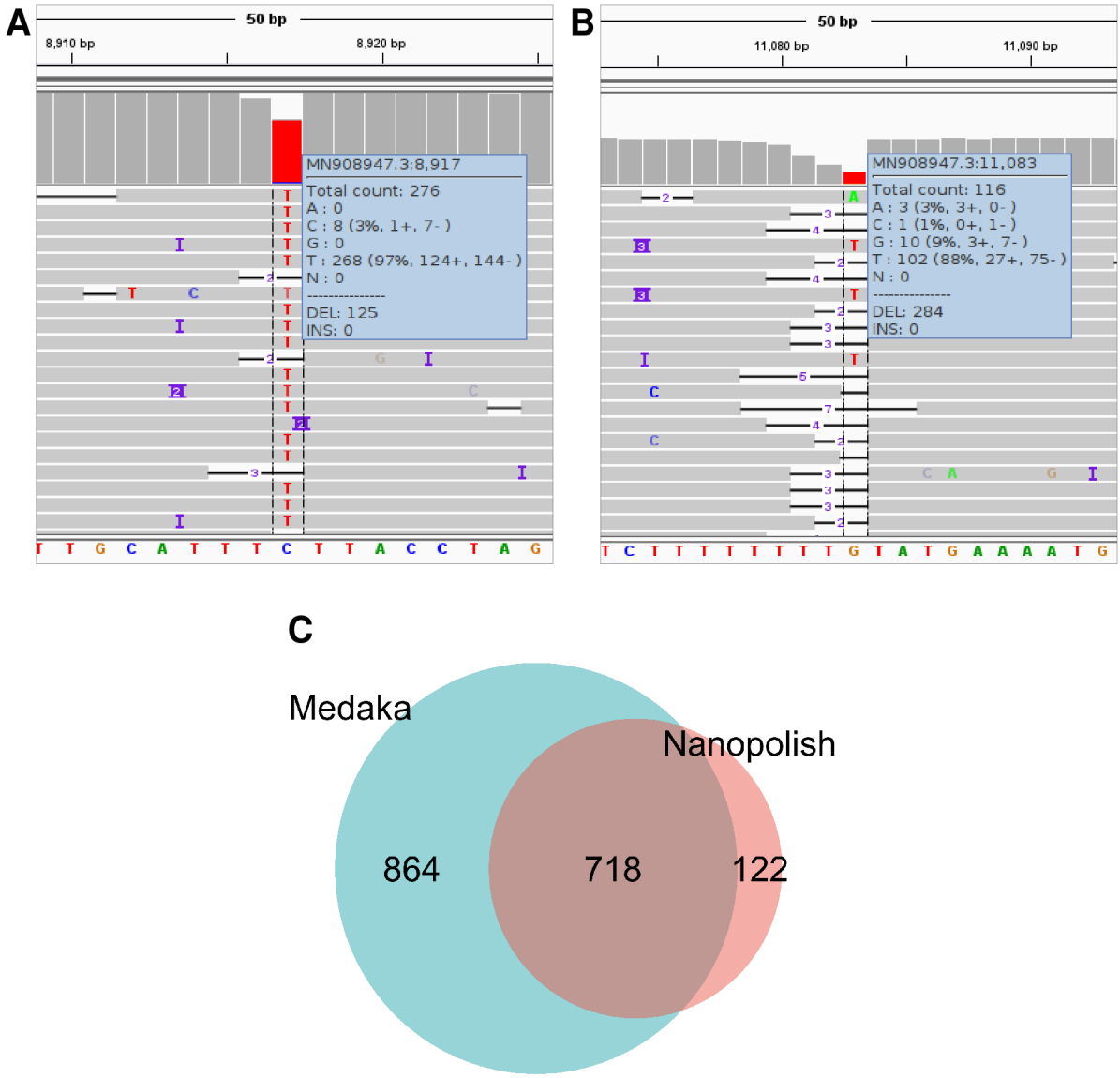


Figure 4.5: Examples of NonDEL20 mutations (A,B) and numbers of mutations called with medaka/nanopolish across the patient cohort (C).

A: Read coverage of position 8,917 of the SARS-CoV-2 genome in the hac4 BAM file for sample SHEF-C6C7B. 125/401 (31%) reads support a deletion at this position, but 97% of remaining reads support a T, with the reference allele being a C.

B: Read coverage of position 11,083 of the SARS-CoV-2 genome in the hac4 BAM file for sample SHEF-D273F. 284/400 (71%) reads support a deletion at this position, but 88% of remaining reads support the presence of a T, with the reference allele being a G.

C: Numbers of unique mutations across all base calling models, with medaka and nanopolish respectively.

Comparing and summarising blacklisted mutations

The blacklists we generated were combined and summarised for comparison with each other and the Bull et al. and De Maio et al. published blacklists (**Figure 4.6**). Due to the large number of our blacklist mutations which were not called across all four basecallers, these were plotted as a histogram above the main bar chart, instead of as individual positions. We also produced a python command line tool, sarscov2vcfblacklister.py, to annotate and filter variants in VCF files using this blacklist summary. There were five genomic peaks where blacklisted mutations in general were detected at a rate of more than 10% of loci, and which contained 559 blacklisted mutations in total. These were 4200-4600, 5200-5800, 10000-10600, 18000-18400 and 22800-23000. The 22800-23000 peak also covers the part of the spike protein (**Figure 4.7**) in which all of the key pandemic mutations of concern are located (**Table 4.3**), associated with increases in viral transmission in the UK, South Africa, Brazil and USA, although none of these mutations were themselves blacklisted. We therefore recommend that identification of new mutations of concern within this region be carried out with caution if they are not found in many patients or sequencing centres, but confirm that none of these three currently identified mutations are prone to sequencing error. In our cohort we blacklisted 211 other spike protein mutations requiring caution when interpreting SARS-CoV-2 sequencing data.

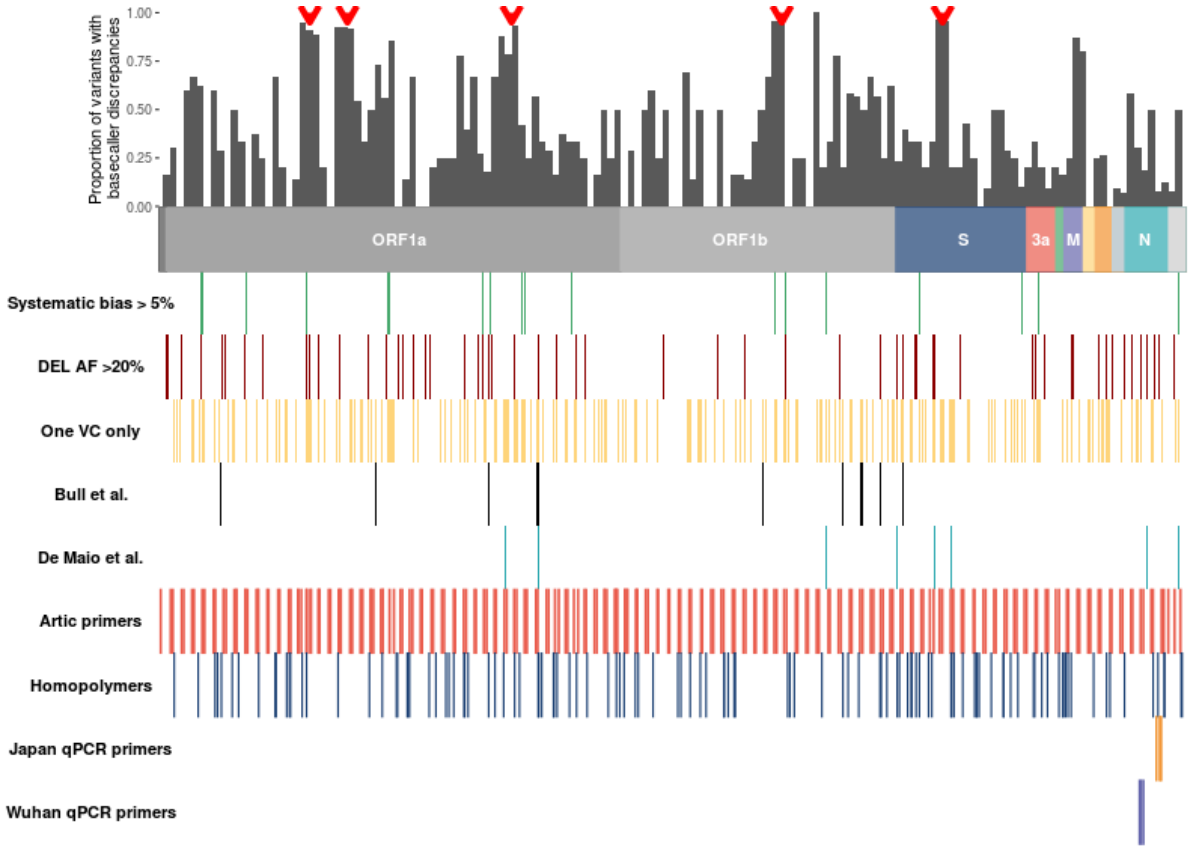


Figure 4.6: SARS-CoV-2 genomic distribution of blacklist features.

Summary of positions at which different blacklisted mutations occur. Top: Bar chart showing the proportion of mutations which are not called across all basecallers in 200bp bins, above coloured panel showing the genomic coordinates of different SARS-CoV-2 ORFs. Peaks where blacklisted mutations made up more than 10% of genomic loci are indicated with red arrows. Middle: Genomic locations of blacklisted mutations plotted using the colour-coded key. Bottom: Genomic coordinates of primers and homopolymers shown for comparison.

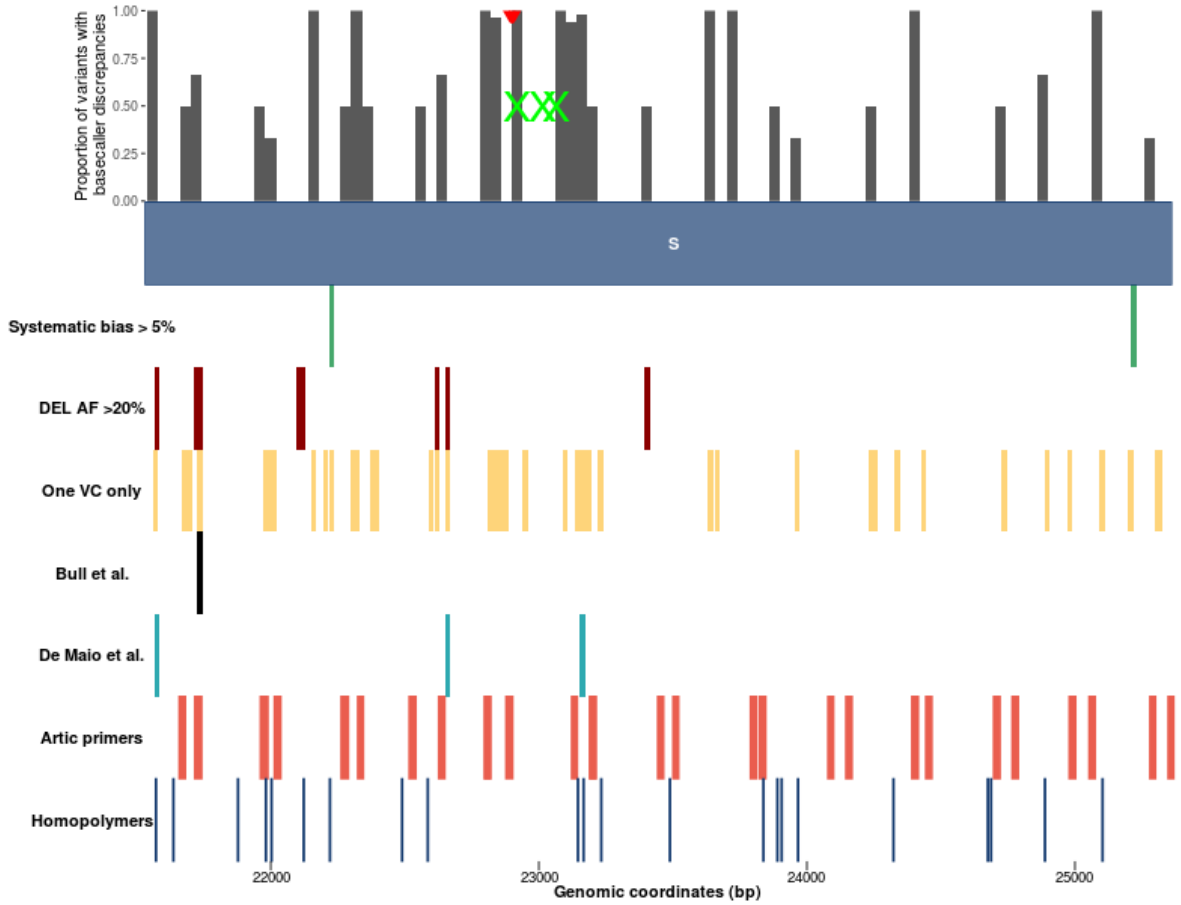


Figure 4.7: SARS-CoV-2 spike protein distribution of blacklist features.
Summary of positions at which different blacklisted mutations occur within the spike protein. Top: Bar chart showing the proportion of mutations which are not called across all basecallers in 40bp bins. Peak where blacklisted mutations made up more than 10% of genomic loci is indicated with a red arrow. Locations of key mutations of concern are plotted with green Xs. Middle: Genomic locations of blacklisted mutations plotted using the colour-coded key. Bottom: Genomic coordinates of primers and homopolymers shown for comparison.

Variant amino acid change (spike protein)	Nucleotide change	Strains with this variant (country most associated with outbreak)
N501Y	A23063T	B.1.1.7 (UK), B.1.351 (South Africa), B.1.1.28 (Brazil), P.1 (Brazil)
E484K	G23012A	B.1.351 (South Africa), B.1.1.28 (Brazil), B.1.525 (Nigeria)
L452R	T22917G	B.1.427/B.1.429 (USA)

Table 4.3: Key mutations of concern in the spike protein of SARS-CoV-2.

Variant calling workflow and base calling model discrepancies were the features accounting for the most blacklist mutations by a large margin (**Table 4.4**). 54.2% of all unique mutations had at least one discrepancy between base calling models, mostly with the medaka variant caller (900 vs. 106 discrepancies). Likewise, most unique mutations were only detected with medaka (50.7%, **Figure 4.5C**).

Blacklist feature	Overall	Nanopolish	Medaka
Systematic bias > 5% allelic fraction	19	14	17
MAF not higher than systematic bias	5	0	5
NonDEL20	68	55	66
Only called with one variant caller	986	122	864
Mutation call not consistent across different basecallers	985	106	900
Mutation call not consistent across different basecallers (excluding HAC3)	698	68	641
In Bull et al. blacklist (mask)	20	12	16
In De Maio et al. blacklist (mask)	6	5	5
In De Maio et al. blacklist (caution)	4	3	4

Table 4.4: Breakdown of blacklisted mutations by feature and variant caller.
Numbers of unique blacklisted mutations in our combined blacklist, that fall within specific subsets. These are defined in more detail in the “Blacklist Classifications” section of the methods.

We compared the overlap between each group of mutations blacklisted for the different reasons established in this study, in order to determine if any of the blacklist features were related. Our analysis (**Table 4.5**) indicated that mutation calls that were discordant between variant calling workflows were significantly (adjusted $P<0.0005$) depleted in both NonDEL20 mutations (OR=0.194) and De Maio et al. blacklisted mutations (OR=0.122), but overlapped with mutations that were discordant between base calling workflows strongly (OR=99). De Maio et al. and Bull et al. have independently flagged some of our blacklisted mutations as likely false positives previously using other methods, such as establishing homoplasy (De Maio et al. 2020a; Crispell, Balaz, and Gordon 2019). We therefore examined overlap between our blacklisted mutations and these smaller blacklists from the literature, but the blacklists from the literature had low numbers of flagged mutations in our cohort (see **Table 4.4**), so their odds ratios, although high, were not statistically significant. For example, NonDEL20

mutations had high overlap with the Bull et al. blacklist (OR=6.31), and mutations with systematic bias > 5% (OR=9.75) and NonDEL20 mutations (OR=5.61) had high overlap with the De Maio blacklist, but these overlaps were not statistically significant.

Blacklist feature	Systematic bias >5%	NonDEL20	Only called in one variant caller	Variant call not consistent across all four basecallers	Bull et al.	De Maio et al.
Systematic bias >5%						
NonDEL20	1.341					
Only called in one variant caller	0.421	0.194***				
Variant call not consistent across all four basecallers	0.527	0.529	99.016***			
Bull et al.	0	6.313	1.093	4.189		
De Maio et al.	9.752	5.606	0.112***	0.413	4.161	

Table 4.5: Fisher Exact Test odds ratios showing overlap between different blacklists across all mutations detected in the patient cohort. Statistical significance is indicated by a */**/*** (Bonferroni-adjusted p=0.05/0.0.005/0.0005 respectively) and significant odds ratios are also highlighted (blue for depletion, orange for enrichment).

We also compared the overlap of each group of mutations blacklisted for the different reasons established in this study with genomic regions known for poor sequencing quality, in order to determine if any of the blacklist features were enriched in these. Key significant (adjusted P<0.0005) results (**Table 4.6**) indicated that mutations with discrepancies between variant callers (OR=3.72) and/or base calling models (OR=5.39) were significantly enriched in the 3bp flanks of homopolymers. The Bull et al. blacklisted mutations were particularly significantly enriched in homopolymers themselves (OR=17.0) and homopolymer 3bp (OR=10.8) and 100bp (OR=2.83) flanks also to a lesser extent, reflecting the known difficulty sequencing in these regions. The De Maio et al. mutations were also significantly enriched in homopolymers, (OR=4.99), albeit to a lesser extent, and were also significantly enriched in the ARTIC primers (OR=1.84).

Blacklist features (columns) / Region features (rows)	Systematic bias >5%	NonDEL20	Only called in one variant caller	Variant call not consistent across all four basecallers	Bull et al.	De Maio et al.	Non-blacklisted variants
Homopolymers	0	0.774	1.072	1.07	16.951***	4.986***	0.495
Homopolymer 3bp flanks	0	3.695	3.722***	5.394***	10.824***	1.121	1.372
Homopolymer 100bp flanks	0.79	1.464	1.298	1.362*	2.827***	1.017	0.981
GCgt99p	0	3.961	0	0.184	0	0.253	1.698
ARTIC primers	0.571	0.899	0.734	0.732	0.759	1.841***	0.941
sgRNAs	0.44	1.645	1.094	1.173	0.692	0.905	1.453

Table 4.6: Fisher Exact Test odds ratios showing enrichment of various blacklisted mutations in a range of notable SARS-CoV-2 genomic regions.

Statistical significance is indicated by a */**/*** (Bonferroni-adjusted p=0.05/0.0.005/0.0005 respectively) and significant odds ratios are also highlighted (orange for enrichment). There was no significant depletion of mutations across any regions.

Differentiating between intra-patient viral genetic diversity and sequencing error

Given the variation in MAFs observed in our cohort, another important question to answer is whether multiple genetically-distinct SARS-CoV-2 viruses could be present within some individuals as a coinfection, or whether certain genomic positions undergo relatively high levels of intra-patient variation across the cohort. Both of these effects would in theory cause intermediate MAFs that could appear similar to sequencing artefacts despite being caused by real biological variation, such as systematic sequencing bias, or deletion bias adjacent to homopolymers (as seen in the NonDEL20 mutation blacklist described in this paper). We therefore established how coinfection and high position-specific intra-patient mutation respectively would differ from these other effects in theory, and identified all cases where there was evidence that they might be occurring.

Usually the MAF of a specific mutation in a haploid genome shows a bimodal distribution across a cohort with two narrow peaks, corresponding to absence (initial peak at very low MAF, e.g. 0-10%) and presence of the mutation (second peak at high MAF, e.g. 90-100%). However, if a specific locus is prone to a certain mutation at high levels, then this can result in that mutation being absent at infection, but becoming present at a noticeable MAF at the point of sequencing. We refer to these as mutation-prone loci (MPLs). The mutations affecting MPLs would be expected to be deleterious since they would otherwise become fixed

in the viral population rapidly. For these mutations, we would expect to see a broader range of MAF values for the MPL ranging from low to high values, depending on the specific locus and how prone it is to mutation. This lack of distinct, separated bimodal peaks would result in a lower standard deviation in allelic fraction across the cohort for that mutation, but the standard deviation would still be significantly higher than it would be for mutations called as a result of systematic bias. Mutation calls resulting from systematic bias in sequencing or alignment can be distinguished from MPL mutation alleles because systematic bias alleles are consistently called at very similar levels across all or most samples, exhibiting a single, very narrow peak at a low-to-mid allelic fraction rather than a broader one.

Candidates for alleles fitting this description have been suggested before (Kuipers et al. 2020). Kuipers et al. have previously generated ranked lists of mutations from two different Illumina-sequenced cohorts based upon their apparent intra-patient genetic diversity, proposing the top ten candidate loci from each as potential loci at which mutations may occur at relatively high rates compared to the rest of the SARS-CoV-2 genome. Two of these 20 genomic loci appeared to fulfill our criteria for classifying an MPL in our own ONT-sequenced cohort (**Figure 4.8A**). Since MPLs are a real biological source of genetic diversity rather than a sequencing artefact, we determined that they should theoretically be present regardless of the sequencing pipeline used, so we did not consider any other loci as MPLs. The MPLs we identified were at positions 6696 (C>T, C>CT) and 15965 (G>T, G>GT) and could be distinguished from systematic bias loci due to their higher standard deviation between patients, since systematic biases result in much more narrow allelic fraction ranges across patient cohorts. C6696T and G15965T would respectively cause the amino acid changes P1326L in nsp3 and C842F in nsp12 and have been detected in GISAID sequences visible on CoV-GLUE (Singer et al. 2020). Insertions at these positions would lead to frameshifts, but have not been detected in any samples. Three MPL mutations were called as mutations in our cohort (G15965GT, C6696CT and C6696T) at the upper ranges of their MAF values (they were not called when the MPL MAF was low), while G15965T was not called as a mutation by either nanopolish or medaka. All other candidates either were not called as mutations in our cohort, or they had a bimodal MAF distribution, or they had a variable MAF distribution that could be explained by the varying prevalence of reads supporting a deletion due to sequencing errors at positions adjacent to a homopolymer (i.e. they were in our NonDEL20 blacklist).

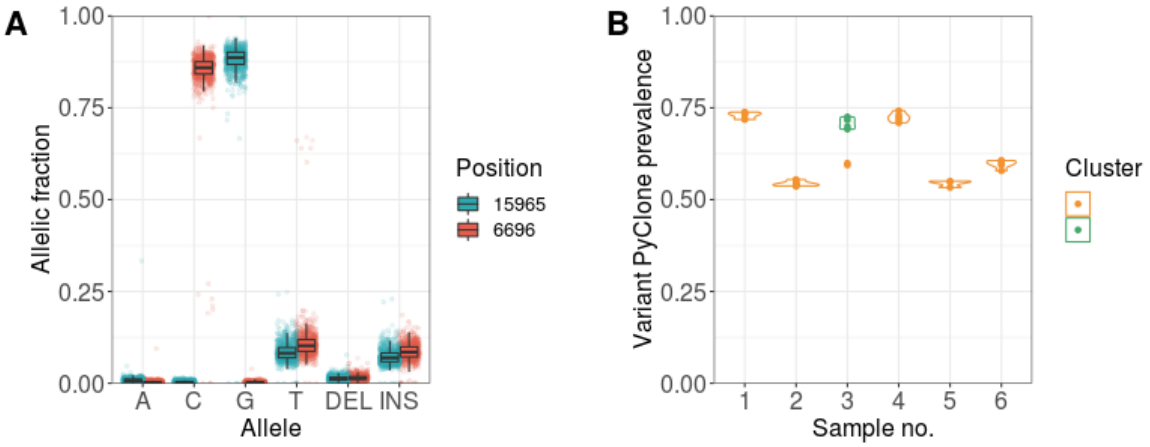


Figure 4.8: Potential mutation-prone loci at two genomic positions (A) and medium MAF clustering in samples (B).

A: HAC4 allelic fractions for all alleles recorded at the two genomic loci showing evidence of high position-specific intra-patient mutation which (Kuipers et al. 2020) separately identified with Illumina. Positions 6696 and 15965 both show a range of low MAF values for two ALT alleles that are never recorded at high allelic fractions in the cohort (6696:C>T, C>CT and 15965:G>T, G>GT), resulting in decreased observed allelic fractions of the reference/major alleles at those positions.

B: Nanopolish/hac4 PyClone intra-patient variant prevalence clusters for mutations between 25-75% MAF across all basecallers, in six patients with 5+ called mutations. Colours indicated separate clusters where this occurred in sample 3. No corresponding cluster below 50% prevalence was observed in any sample, such that a pair of clusters had 100% total prevalence.

In a small number of cases within our cohort we observed mutations called at MAF values closer to 50%, despite that mutant allele having a bimodal MAF distribution across the rest of the cohort reflecting clear presence/absence, in contrast with the MPLs analysed above. This suggested that the mutations could potentially be present in combination with either the reference or another alternative allele as a coinfection.

We observed 61 samples in which there were two or more mutations with an MAF between 25-75% (which we define as a medium MAF value) across all base calling models with either variant caller. All combinations of base calling models and variant callers exhibited mutations with medium MAF values with high correlation (**Figure 4.3C,D**), suggesting that this was not specific to any particular workflow. In total, 150 unique mutations had a medium MAF value within at least one of these samples, and which had a clear bimodal distribution of MAF values indicating presence/absence in all other samples. 79% of these unique mutations were not at NonDEL20 positions, so high levels of genomic reads supporting deletions were not present and did not contribute to the medium MAF of most mutations. The large number and variety of unique mutations with a medium MAF value in at least one sample indicated that

contamination was unlikely to be the cause of these medium MAF values. Contamination would be expected to cause low-level detection of much less variable mutation sets than observed in affected samples if it were occurring.

The six samples with the most possible coinfection mutations had 5-8 mutations called at 25-75% MAF across all basecallers. The HAC4/nanopolish VCF files for each of these six samples were visualised with PyClone (**Figure 4.8B**). PyClone is a program that uses Bayesian clustering to group sets of mutations into putative clonal clusters based on the numbers of reads that do or do not support those mutations. It also estimates the prevalence of clusters with these sets of mutations as a fraction (Roth et al. 2014). Within these samples, the MAF values of most or all of these mutations clustered together, even across mutations that were not close to each other on the SARS-CoV-2 genome, providing further evidence of a possible coinfection. However, opposite clusters of called mutations with MAF values of 100% minus the MAF values of these clusters, representing the alternative variant cluster in each case, were not observed, even though this would be expected to occur if a coinfection with two genetically different viruses were present. This reflected the fact that the alternate alleles to the cluster mutations in these cases likely corresponded to the reference allele, which by default is not shown within VCF files and therefore not input to PyClone. Any additional non-reference alleles at these positions were unlikely to be supported by enough sequencing reads to be called within the VCF file and would also need to be supported by at least 25% of reads to be displayed on the figure, so it was unsurprising that only one cluster at 0.5-0.75 allelic fraction was observed for most patients, with the exception of sample 3. PyClone indicated support for separate clusters at 0.6 and 0.7 for sample 3, colour coded in yellow and green respectively.

Discussion

Basecaller comparisons

Hac3 was responsible for the most discrepancies in variant calling compared to the other base calling models examined (**Figure 4.3**), and also exhibited the most NonDEL20 mutations of all basecallers. 40% of NonDEL20 mutations were exclusive to hac3, while 0-3% were exclusive to the other base calling models, and hac4 had none at all. Finally, hac3 was prone to much higher levels of systematic base calling bias than the other base calling models used in this study (**Table 4.2**), while hac4 performed best in this regard also. Out of all possible alleles, deletions accounted for the highest level of systematic bias, especially with the hac3 model, exhibiting 2579 loci for which at least 5% of reads supported the presence of a deletion due to systematic bias towards deletions at this level. The high prevalence of systematic biases towards detection of deletions may in part be related to sequencing errors in the proximity of homopolymers which result in high levels of deletions being detected there (**Figure 4.5**). In addition to the base calling models examined in this study, new ONT base calling models and algorithms such as Bonito v0.3.8 (Silvestre-Ryan and Holmes 2021), released on 21/04/2021, may further improve basecalling accuracy, but were not tested.

Apart from this study, there are no reports comparing the accuracy of these base calling models and their application to SARS-CoV-2 sequencing. Hac3 may not be in use among many sequencing centres anymore due to the release of hac4, but many early SARS-CoV-2 genomes were sequenced before this: The standalone version of hac4 was released on 18/06/2020, but it was not incorporated into ONT's MinKNOW pipeline until 22/07/2020, at which point 70,658 human SARS-CoV-2 sequences had been submitted to GISAID, many sequenced using hac3, which could benefit from recalling with hac4 if the raw ONT sequencing data is still available. This may have affected the results of publications using hac3-derived data, covering areas such as confirming SARS-CoV-2 infections and identifying virus mutations (Wang et al. 2020), tracing their geographic spread (Walker et al. 2020), and determining chains of infection in a healthcare setting (Meredith et al. 2020), so these recommendations could impact on a wide diversity of research topics. Our findings highlight the importance of retaining raw sequencing data in long-term repositories so that cutting edge base calling models and mutation detection workflows can be reapplied as they are released and updated. We reiterate the importance of sequencing centres updating their workflows from hac3 to hac4 where they have not yet done so.

Variant caller comparison

Nanopolish and medaka mutation calling workflows exhibited differences in called mutations across multiple measures of how prone each was to error, and nanopolish performed better in all of these aspects.

Nanopolish did not call any mutations with MAF matching the systematic bias allelic fraction, while medaka called five. All of these medaka-exclusive false-positive mutations were G>A mutations, at positions 1265, 2539, 10704, 18002 and 18317 respectively, suggesting that a G>A systematic bias was the specific cause. G>A bias is commonly found at higher levels than noise from other allelic substitutions when sequencing, usually due to spontaneous deamination of methylated cytosine residues to uracil (Chen et al. 2014; Ma et al. 2019). This type of G>A bias is systematic, concordant with what we observe, because it would be consistent across all samples, and therefore appears to be the most likely explanation for these errors. This suggests that nanopolish variant calling is generally not affected by systematic bias in sequencing for any of the ONT base calling models, while medaka is, albeit at a low level.

Medaka exhibited higher numbers of blacklisted mutations in every category examined (**Table 4.4**), with a particularly large difference in the number of discordant mutation calls across base calling models (106 with nanopolish vs. 900 with medaka). In addition, the majority (85%) of mutations called with nanopolish were concordant with medaka, while fewer (45%) medaka mutations were concordant with nanopolish (**Figure 4.5C**). The base callers compared are similar, so the large number of differences between their medaka variant calling results, combined with medaka generally calling more variants, suggests that medaka has a higher variant calling sensitivity and calls more edge cases where small differences in the base calling outputs change the variant call outcome, with a higher corresponding rate of false positive calls. Previous benchmarking of medaka against nanopolish appears to show that nanopolish performs better at positions with low sequencing quality, but not high quality (Wick, Judd, and Holt 2019a, [b] 2019), which may be the basis for why nanopolish called fewer ambiguous variants.

Despite these drawbacks, medaka is commonly used as an alternative to nanopolish due to the fact that it does not require raw sequencing data as an input — only FASTQ files are required — as well as its shorter runtimes — medaka is 20-50X faster — and greater accuracy for some genomes (Oxford Nanopore Technologies 2018), albeit not SARS-CoV-2 as shown in this study. Nanopolish also carries out variant calling on FASTQ files that have been output by a separate base calling algorithm, but additionally requires raw signal output

in FAST5 files, which it uses to compute a new consensus sequence against which variant calling is carried out rather than the reference sequence, thus improving its accuracy (Oxford Nanopore Technologies 2021; Loman, Quick, and Simpson 2015). Nanopolish does not carry out its own base calling. For scientists who still wish to use medaka for sequencing despite its lower mutation calling accuracy, we would recommend masking or taking caution with calling any of the mutations described in our blacklist, in order to compensate for these to some extent. A newer version of ARTIC (v1.2.1+) has since been released using an updated version of medaka, which has an option to specify the basecalling model. Using the updated medaka model with this specification may enable improved performance for medaka beyond what was reported in this study.

Summary of blacklists and utility in quality control

We generated novel blacklists of SARS-CoV-2 mutations using three approaches: (1) Identification of discrepancies between base calling models and variant callers, (2) presence of systematic bias, and (3) presence of high levels of reads supporting deletions at non-deletion mutations (NonDEL20). We compared these against two existing blacklists in the literature which used different methods for blacklisting mutations, including phylogenetic and sequencing-based approaches (Bull et al. 2020; De Maio et al. 2020b). Only 2.08/5.16% were independently flagged in the De Maio/Bull et al. blacklists respectively. Discrepancies between basecallers and variant callers were responsible for the majority of blacklisted mutations, while NonDEL20 mutations were uncommon (55 with nanopolish, 66 with medaka) and systematic biases caused five false positive mutation calls with medaka (**Table 4.4**). Mutations that show discrepancies between workflows may be false positives in workflows where they are detected or false negatives in workflows where they are not called. We therefore recommend caution with these rather than outright masking, since masking would remove real mutations that were false negatives in at least one of the other basecallers or variant callers examined.

We have shown that ONT sequencing recorded high proportions of reads supporting a deletion at many positions adjacent to homopolymers, and that this often coincided with the detection of non-deletion mutations which have decreased MAFs as a result and appear suspect. Some examples of these NonDEL20 mutations are commonly called across many sequencing methods, such as G28881A (Korber et al. 2020; Nguyen et al. 2020; Weber, Ramirez, and Doerfler 2020). These are unlikely to be false positives due to the consensus between completely different sequencing methods such as Illumina and ONT. Nevertheless, we recommend caution with calling NonDEL20 mutations due to their significant enrichment in other blacklists (odds ratio (OR)=8.43 in the Bull et al. blacklist, **Table 4.5**) and poorer base

calling accuracy as shown by the prevalence of reads incorrectly supporting a deletion at these positions, likely resulting from their adjacent homopolymers in cases where NonDEL20 mutations occur.

Presence of reads supporting deletions was the most common reason for why called mutations did not have high MAFs, but this was usually not systematic, even though systematic bias towards deletions was the most common source of systematic bias (**Table 4.2**). Both systematic and (in proximity to homopolymers) non-systematic errors in sequencing seem to lead to higher-than-expected detection of reads supporting deletions with ONT technology. A recent SARS-CoV-2 publication (Thielen et al. 2021) has found lower MAF values at mutations adjacent to homopolymers with ONT sequencing — including many of the same mutations we have classed as NonDEL20 — but displayed near-100% MAF values for these same mutations with Illumina sequencing, suggesting that this effect is ONT-specific. Furthermore, Thielen et al. found strong concordance between ONT and Illumina consensus sequencing results at all other positions. This may be related to the underlying way in which nanopore sequencing works, which relies upon measuring an electrical current through the pore that is altered as different bases pass through — when repeated homopolymeric bases pass through there is little detectable change in this electrical current, making it difficult to accurately gauge how long the homopolymer is, resulting in mismapping of reads at homopolymers and erroneously suggesting the presence of deletions adjacent to them (N. J. Loman, Quick, and Simpson 2015; Cornelis et al. 2017), as we observed. Future ONT sequencing work could benefit from stricter filtering of reads supporting deletions at a low-to-medium allelic fraction, although this might result in real low-to-medium MAF deletions being missed if these exist.

As of 12th April 2021, there were five SARS-CoV-2 lineages of concern (https://www.gisaid.org/hcov19-variants/) due to large recent increases in population frequency, as well as possible functional effects of their key SARS-CoV-2 spike protein mutations of concern (N501Y, E484K and L452R, **Table 4.3**) (Tegally et al. 2021; Galloway et al. 2021; Faria et al. 2021; Zhang et al. 2021; Rambaut Group 2021). None of the patients in our cohort had any of these mutations, nor were any of these mutations themselves blacklisted, even though they all had genomic loci within the 22800-23000 peak in blacklisted mutations. In addition, systematic sequencing bias was not found for any of these mutations (although we blacklisted 211 other spike protein mutations), confirming that none of these mutations of concern are prone to sequencing inaccuracy with the ONT ARTIC sequencing pipeline, with important ramifications for sequencing efforts tracking outbreaks associated with these mutations.

Differentiating between intra-patient viral genetic diversity and sequencing error

We ensured that no mutations were being blacklisted as potential sequencing errors when they might be caused by real biological genetic diversity (Simmonds 2020), justifying the quality control processes suggested in this paper. A summary of possible causes of intra-patient genetic diversity, their different effects on MAF, and their frequency within our cohort is shown in **Table 4.7**.

Three called mutations (G15965GT, C6696CT and C6696T) exhibited cross-cohort MAF values with both ONT and Illumina sequencing on separate cohorts that suggested that positions 6696 and 15965 always displayed the reference allele at infection, but were prone to these mutations at high rates leading to intra-patient genetic diversity that was not transmitted. This has been further corroborated by observations in the literature that the allelic fraction values of these mutations correlate with patient age (Kuipers et al. 2020), which would be expected given that older patients have longer infections and accumulate more mutations over time.

There were 61 samples that displayed single clusters of 50-75% MAF mutations, but were missing smaller corresponding 25-50% MAF mutation clusters that would be expected if a coinfection were occurring, although this is likely explained by the corresponding clusters being formed of reference alleles and therefore absent from the VCF. In addition, it is difficult to distinguish between real coinfection mutations and sequencing artefacts below 50% allelic fraction, limiting our ability to confirm whether low-level coinfections were indeed occurring in these samples. Other studies (Lythgoe et al. 2020; Tonkin-Hill et al. 2020) have achieved findings which support a similar rate of occurrence of putative coinfections in different cohorts, but with the same limitations. This would be a promising area for further research with other sequencing technologies to confirm the existence or absence of SARS-CoV-2 coinfections in these circumstances.

Feature	MPL	Coinfection (unconfirmed)	Systematic Bias	NonDEL20 mutation
Source of effect	Real genetic diversity	Real genetic diversity (if coinfection is the cause)	Sequencing artefact consistent across samples	Sequencing artefact adjacent to homopolymer
Allelic fraction distribution across patient cohort	Broad range of MAF values across many patients, since mutations expected to accumulate with time since infection.	Samples where a coinfection exists have medium MAF mutation clusters (25-75%) that are between bimodal peaks covering the majority of mutations, (at ~0/100% MAF).	Similar fraction of reads consistently support mutation in all/most samples (low to medium MAF).	Bimodal, but called mutation MAF is lower than expected due to high number of reads supporting DEL allele.
MAF SD at this position	Medium	Highest	Low	High
Number of called mutations affected in cohort	3 mutations at 2 loci	61/884 samples, with 150 medium MAF unique mutations in total	19 mutations	68 mutations

Table 4.7: Sources of apparent intra-patient genetic diversity

Hypothesised categories of medium allelic fraction mutations detected, distinguishing between real biological diversity (MPLs and coinfection) and sequencing artefacts (systematic bias and NonDEL20 mutations), with defining features and number of potential occurrences in our data set.

Conclusion

This study establishes best-practice guidelines for optimising SARS-CoV-2 sequencing by benchmarking a range of base calling models and variant calling workflows used in combination with the ARTIC sequencing pipeline for ONT, on a cohort of 884 SARS-CoV-2 samples in the UK. We recommend guppy v4.0 (high-accuracy mode) and nanopolish as the best-performing base calling model and variant calling workflow respectively. We reiterate the importance of ONT sequencing centres staying up-to-date with the latest version of the Guppy basecaller and retaining raw sequencing data for recalling as new algorithms and models are released or updated over time. In addition, we share a blacklist of mutations that should be treated with caution or masked during sequencing, listing a range of features for each blacklisted mutation that may affect sequencing accuracy at that position, such as discrepancies between pipelines, susceptibility to systematic sequencing bias, and high levels of reads incorrectly supporting deletions adjacent to homopolymers. If these limitations to sequencing accuracy are not controlled for then erroneous detection of blacklisted mutations may occur, altering conclusions about mutations present in pandemic lineages and their prevalence. We anticipate that this blacklist will provide a valuable quality control resource for the broader scientific community using ONT SARS-CoV-2 sequencing data, enabling false positive mutation calls to be identified and filtered out with ease, and we share a command line tool for users to apply to their own VCFs for this purpose.

This study does not fully establish the mechanistic differences between base calling algorithms and variant callers that cause the differences in sequencing results detailed here. While this is not necessary for the aims of our research, knowledge of the underlying causes of poor sequence accuracy could aid software developers in improving sequencing bioinformatics software to improve sequencing accuracy at the blacklisted genomic positions. In addition, our research only utilised ONT sequencing data. Future studies could build upon this by applying our approaches to other types of sequencing, such as Illumina. This would allow benchmarking against ONT and could identify SARS-CoV-2 sequencing issues that do not apply to ONT sequencing, which are not included in our blacklist.

References

Amarasinghe, Shanika L., Shian Su, Xueyi Dong, Luke Zappia, Matthew E. Ritchie, and Quentin Gouil. 2020. “Opportunities and Challenges in Long-Read Sequencing Data Analysis.” *Genome Biology* 21 (1): 30.

Ashour, Hossam M., Walid F. Elkhatib, Md Masudur Rahman, and Hatem A. Elshabrawy. 2020. “Insights into the Recent 2019 Novel Coronavirus (SARS-CoV-2) in Light of Past Human Coronavirus Outbreaks.” *Pathogens* 9 (3). https://doi.org/10.3390/pathogens9030186.

Bull, Rowena A., Thiruni N. Adikari, James M. Ferguson, Jillian M. Hammond, Igor Stevanovski, Alicia G. Beukers, Zin Naing, et al. 2020. “Analytical Validity of Nanopore Sequencing for Rapid SARS-CoV-2 Genome Analysis.” *Nature Communications* 11 (1): 6272.

Chen, Guoli, Stacy Mosier, Christopher D. Gocke, Ming-Tseh Lin, and James R. Eshleman. 2014. “Cytosine Deamination Is a Major Cause of Baseline Noise in next-Generation Sequencing.” *Molecular Diagnosis & Therapy* 18 (5): 587–93.

Corman, Victor M., Olfert Landt, Marco Kaiser, Richard Molenkamp, Adam Meijer, Daniel Kw Chu, Tobias Bleicker, et al. 2020. “Detection of 2019 Novel Coronavirus (2019-nCoV) by Real-Time RT-PCR.” *Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin* 25 (3). https://doi.org/10.2807/1560-7917.ES.2020.25.3.2000045.

Cornelis, Senne, Yannick Gansemans, Lieselot Deleye, Dieter Deforce, and Filip Van Nieuwerburgh. 2017. “Forensic SNP Genotyping Using Nanopore MinION Sequencing.” *Scientific Reports* 7 (February): 41759.

Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. 2020. “The Species Severe Acute Respiratory Syndrome-Related Coronavirus: Classifying 2019-nCoV and Naming It SARS-CoV-2.” *Nature Microbiology* 5 (4): 536–44.

COVID-19 Genomics UK (COG-UK). 2020. “An Integrated National Scale SARS-CoV-2 Genomic Surveillance Network.” *The Lancet. Microbe* 1 (3): e99–100.

Crispell, Joseph, Daniel Balaz, and Stephen Vincent Gordon. 2019. “HomoplasyFinder: A Simple Tool to Identify Homoplasies on a Phylogeny.” *Microbial Genomics* 5 (1). https://doi.org/10.1099/mgen.0.000245.

De Maio, Nicola, Conor Walker, Rui Borges, Lukas Weilguny, Greg Slodkowicz, and Nick Goldman. 2020a. “Issues with SARS-CoV-2 Sequencing Data.” *Virological.org*. https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473.

———. 2020b. “Masking Strategies for SARS-CoV-2 Alignments.” Virological. July 29, 2020. https://virological.org/t/masking-strategies-for-sars-cov-2-alignments/480.

Depledge, Daniel P., Kalanghad Puthankalam Srinivas, Tomohiko Sadaoka, Devin Bready, Yasuko Mori, Dimitris G. Placantonakis, Ian Mohr, and Angus C. Wilson. 2019. “Direct RNA Sequencing on Nanopore Arrays Redefines the Transcriptional Complexity of a Viral Pathogen.” *Nature Communications* 10 (1): 754.

Elbe, Stefan, and Gemma Buckland-Merrett. 2017. “Data, Disease and Diplomacy: GISAID’s Innovative Contribution to Global Health.” *Global Challenges*. https://doi.org/10.1002/gch2.1018.

Faria, Nuno R., Thomas A. Mellan, Charles Whittaker, Ingra M. Claro, Darlan da S. Candido, Swapnil Mishra, Myuki A. E. Crispim, et al. 2021. “Genomics and Epidemiology of a Novel SARS-CoV-2 Lineage in Manaus, Brazil.” *medRxiv : The Preprint Server for Health Sciences*, March. https://doi.org/10.1101/2021.02.26.21252554.

Franco-Muñoz, Carlos, Diego A. Álvarez-Díaz, Katherine Laiton-Donato, Magdalena Wiesner, Patricia Escandón, José A. Usme-Ciro, Nicolás D. Franco-Sierra, et al. 2020. “Substitutions in Spike and Nucleocapsid Proteins of SARS-CoV-2 Circulating in South America.” *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* 85 (November): 104557.

Freeman, Timothy M., Genomics England Research Consortium, Dennis Wang, and Jason

Harris. 2020. “Genomic Loci Susceptible to Systematic Sequencing Bias in Clinical Whole Genomes.” *Genome Research* 30 (3): 415–26.
Galloway, Summer E., Prabasaj Paul, Duncan R. MacCannell, Michael A. Johansson, John T. Brooks, Adam MacNeil, Rachel B. Slayton, et al. 2021. “Emergence of SARS-CoV-2 B.1.1.7 Lineage - United States, December 29, 2020-January 12, 2021.” *MMWR. Morbidity and Mortality Weekly Report* 70 (3): 95–99.
Itokawa, Kentaro, Tsuyoshi Sekizuka, Masanori Hashino, Rina Tanaka, and Makoto Kuroda. 2020. “Disentangling Primer Interactions Improves SARS-CoV-2 Genome Sequencing by Multiplex Tiling PCR.” *PloS One* 15 (9): e0239403.
Korber, Bette, Will M. Fischer, Sandrasegaram Gnanakaran, Hyejin Yoon, James Theiler, Werner Abfalterer, Nick Hengartner, et al. 2020. “Tracking Changes in SARS-CoV-2 Spike: Evidence That D614G Increases Infectivity of the COVID-19 Virus.” *Cell* 182 (4): 812–27.e19.
Kubik, Slawomir, Ana Claudia Marques, Xiaobin Xing, Janine Silvery, Claire Bertelli, Flavio De Maio, Spyros Pournaras, et al. 2020. “Guidelines for Accurate Genotyping of SARS-CoV-2 Using Amplicon-Based Sequencing of Clinical Samples.” *BioRXiv*, December. https://doi.org/10.1101/2020.12.01.405738.
Kuipers, Jack, Aashil A. Batavia, Kim P. Jablonski, Fritz Bayer, Nico Borgsmüller, Arthur Dondi, Monica-Andreea Drăgan, et al. 2020. “Within-Patient Genetic Diversity of SARS-CoV-2.” *bioRxiv*. https://doi.org/10.1101/2020.10.12.335919.
Kumar, Pramod, Rajesh Pandey, Pooja Sharma, Mahesh S. Dhar, Vivekanand A, Bharathram Uppili, Himanshu Vashisht, et al. 2020. “Integrated Genomic View of SARS-CoV-2 in India.” *Wellcome Open Research* 5 (August): 184.
Loman, Nicholas J., Joshua Quick, and Jared T. Simpson. 2015. “A Complete Bacterial Genome Assembled de Novo Using Only Nanopore Sequencing Data.” *Nature Methods* 12 (8): 733–35.
Loman, Nick, Will Rowe, and Andrew Rambaut. 2020. “nCoV-2019 Novel Coronavirus Bioinformatics Protocol.” https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html.
Lythgoe, Katrina A., Matthew Hall, Luca Ferretti, and Mariateresa de Cesare. 2020. “Within-Host Genomics of SARS-CoV-2.” *bioRxiv*. https://doi.org/10.1101/2020.05.28.118992.
Ma, Xiaotu, Ying Shao, Liqing Tian, Diane A. Flasch, Heather L. Mulder, Michael N. Edmonson, Yu Liu, et al. 2019. “Analysis of Error Profiles in Deep next-Generation Sequencing Data.” *Genome Biology* 20 (1): 50.
Meredith, Luke W., William L. Hamilton, Ben Warne, Charlotte J. Houldcroft, Myra Hosmillo, Aminu S. Jahun, Martin D. Curran, et al. 2020. “Rapid Implementation of SARS-CoV-2 Sequencing to Investigate Cases of Health-Care Associated COVID-19: A Prospective Genomic Surveillance Study.” *The Lancet Infectious Diseases* 20 (11): 1263–71.
Michel, Christian Jean, Claudine Mayer, Olivier Poch, and Julie Dawn Thompson. 2020. “Characterization of Accessory Genes in Coronavirus Genomes.” *Virology Journal* 17 (1): 131.
Nguyen, Tam Thi, Thach Ngoc Pham, Trang Dinh Van, Trang Thu Nguyen, Diep Thi Ngoc Nguyen, Hoa Nguyen Minh Le, John-Sebastian Eden, et al. 2020. “Genetic Diversity of SARS-CoV-2 and Clinical, Epidemiological Characteristics of COVID-19 Patients in Hanoi, Vietnam.” *PloS One* 15 (11): e0242537.
Oxford Nanopore Technologies. 2018. “Medaka.” https://nanoporetech.github.io/medaka/.
Oxford Nanopore Technologies. 2021. “GitHub - Jts/nanopolish: Signal-Level Algorithms for MinION Data.” Accessed October 31, 2021. https://github.com/jts/nanopolish.
Pollard, Martin O., Deepti Gurdasani, Alexander J. Mentzer, Tarryn Porter, and Manjinder S. Sandhu. 2018. “Long Reads: Their Purpose and Place.” *Human Molecular Genetics* 27 (R2): R234–41.
Rambaut Group. 2021. “B.1.525.” PANGO Lineages. February 20, 2021. https://cov-lineages.org/global_report_B.1.525.html.
Shu, Yuelong, and John McCauley. 2017. “GISAID: Global Initiative on Sharing All Influenza Data – from Vision to Reality.” *Eurosurveillance*. https://doi.org/10.2807/1560-

7917.es.2017.22.13.30494.
Silvestre-Ryan, Jordi, and Ian Holmes. 2021. “Pair Consensus Decoding Improves Accuracy of Neural Network Basecallers for Nanopore Sequencing.” *Genome Biology* 22 (1): 1–6.
Simmonds, P. 2020. “Rampant C→U Hypermutation in the Genomes of SARS-CoV-2 and Other Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary Trajectories.” *mSphere*. https://doi.org/10.1128/msphere.00408-20.
Simpson, J. 2018. “Nanopolish: Signal-Level Algorithms for MinION Data.” *Github Available at: Https://github. Com/jts/nanopolish [Accessed January 10, 2019]*.
Szalay, Tamas, and Jene A. Golovchenko. 2015. “De Novo Sequencing and Variant Calling with Nanopores Using PoreSeq.” *Nature Biotechnology* 33 (10): 1087–91.
Tan, Adrian, Gonçalo R. Abecasis, and Hyun Min Kang. 2015. “Unified Representation of Genetic Variants.” *Bioinformatics* 31 (13): 2202–4.
Tegally, Houriiyah, Eduan Wilkinson, Marta Giovanetti, Arash Iranzadeh, Vagner Fonseca, Jennifer Giandhari, Deelan Doolabh, et al. 2021. “Emergence of a SARS-CoV-2 Variant of Concern with Mutations in Spike Glycoprotein.” *Nature*, March. https://doi.org/10.1038/s41586-021-03402-9.
Thielen, Peter M., Shirlee Wohl, Thomas Mehoke, Srividya Ramakrishnan, Melanie Kirsche, Oluwaseun Falade-Nwulia, Nídia S. Trovão, et al. 2021. “Genomic Diversity of SARS-CoV-2 during Early Introduction into the Baltimore–Washington Metropolitan Area.” *JCI Insight*. https://doi.org/10.1172/jci.insight.144350.
Tonkin-Hill, Gerry, Inigo Martincorena, Roberto Amato, and Andrew R. J. Lawson. 2020. “Patterns of within-Host Genetic Diversity in SARS-CoV-2.” *bioRxiv*. https://doi.org/10.1101/2020.12.23.424229.
Tyson, John R., Phillip James, David Stoddart, Natalie Sparks, Arthur Wickenhagen, Grant Hall, Ji Hyun Choi, et al. 2020. “Improvements to the ARTIC Multiplex PCR Method for SARS-CoV-2 Genome Sequencing Using Nanopore.” *bioRxiv : The Preprint Server for Biology*, September. https://doi.org/10.1101/2020.09.04.283077.
Walker, Andreas, Torsten Houwaart, Tobias Wienemann, Malte Kohns Vasconcelos, Daniel Strelow, Tina Senff, Lisanna Hülse, et al. 2020. “Genetic Structure of SARS-CoV-2 Reflects Clonal Superspreading and Multiple Independent Introduction Events, North-Rhine Westphalia, Germany, February and March 2020.” *Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin* 25 (22). https://doi.org/10.2807/1560-7917.ES.2020.25.22.2000746.
Wang, Ming, Aisi Fu, Ben Hu, Yongqing Tong, Ran Liu, Zhen Liu, Jiashuang Gu, et al. 2020. “Nanopore Targeted Sequencing for the Accurate and Comprehensive Detection of SARS-CoV-2 and Other Respiratory Viruses.” *Small*. https://doi.org/10.1002/smll.202002169.
Weber, Stefanie, Christina Ramirez, and Walter Doerfler. 2020. “Signal Hotspot Mutations in SARS-CoV-2 Genomes Evolve as the Virus Spreads and Actively Replicates in Different Parts of the World.” *Virus Research* 289 (November): 198170.
Wick, Ryan R., Louise M. Judd, and Kathryn E. Holt. 2019a. “Performance of Neural Network Basecalling Tools for Oxford Nanopore Sequencing.” *Genome Biology* 20 (1): 129.
———. 2019b. “August 2019 Consensus Accuracy Update.” https://github.com/rrwick/August-2019-consensus-accuracy-update.
Zhang, Wenjuan, Brian D. Davis, Stephanie S. Chen, Jorge M. Sincuir Martinez, Jasmine T. Plummer, and Eric Vail. 2021. “Emergence of a Novel SARS-CoV-2 Variant in Southern California.” *JAMA: The Journal of the American Medical Association*, February. https://doi.org/10.1001/jama.2021.1612.
Zhou, Peng, Xing-Lou Yang, Xian-Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, Hao-Rui Si, et al. 2020. “A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin.” *Nature* 579 (7798): 270–73.

Chapter V: Conclusion

This thesis documents research in three areas that cover the common themes of assessing the accuracy of genomic variant detection and understanding the clinical impact of variants. I explore the utility of somatic tumour variants in predicting survival of non-small cell lung cancer (NSCLC) patients, describe a novel statistical approach to catalogue systematic sequencing bias across the human genome, and show how this method can be adapted and applied to improve quality control of SARS-CoV-2 genomes in the context of a range of different quality control methods.

The burden of somatic mutations and neoantigens has been associated with improved survival in cancer treated with immunotherapies, especially non-small cell lung cancer (NSCLC). However, there was uncertainty about their effect on outcome in early-stage untreated cases. From a small cohort of 36 NSCLC cases, somatic mutations and copy number alterations in 865 genes that contributed to patient overall survival were identified. Simply, the number of altered genes (NAG) among these 865 genes was associated with longer disease-free survival (hazard ratio (HR) = 0.153, $p = 1.48 \times 10^{-4}$). Patients with a high NAG could be further stratified based on the presence of immunogenic mutations, revealing a further subgroup of stage I NSCLC with even better prognosis compared to the low NAG group ($HR = 0.0807$, $p = 7.8 \times 10^{-5}$), and associated with cytotoxic T-cell expression.

Accurate massively parallel sequencing (MPS) of genetic variants is key to many areas of science and medicine, such as cataloguing population genetic variation and diagnosing genetic diseases. Certain genomic positions can be prone to higher rates of systematic sequencing and alignment bias that limit accuracy, resulting in false positive variant calls. Current standard practices to differentiate between loci that can and cannot be sequenced with high confidence utilize consensus between different sequencing methods as a proxy for sequencing confidence. These practices have significant limitations, and alternative methods are required to overcome them. I have developed a novel statistical method based on summarizing sequenced reads from whole-genome clinical samples and cataloguing them in “Incremental Databases” that maintain individual confidentiality. Allele statistics were catalogued for each genomic position that consistently showed systematic biases with the corresponding MPS sequencing pipeline. I found systematic biases present at ~1%–3% of the human autosomal genome across five patient cohorts. I identified which genomic regions were more or less prone to systematic biases, including large homopolymer flanks (odds ratio = 23.29–33.69) and the NIST high confidence genomic regions (odds ratio = 0.154–0.191). I

confirmed my predictions on a gold-standard reference genome and showed that these systematic biases can lead to suspect variant calls within clinical panels. My results recommend increased caution to address systematic biases in whole-genome sequencing and alignment. My study provides the implementation of a simple statistical approach to enhance quality control of clinically sequenced samples by flagging variants at suspect loci for further analysis or exclusion. Additionally, this project revealed synergistic insights that provided mutual benefits to the projects in chapters II and IV. For example, the results of the IncDB-based approach on germline data from chapter III revealed lists of genomic positions at which systematic sequencing and alignment biases were present. Somatic neoantigenic variants can be checked against this germline-derived IncDB in the same way that germline variants are checked against it, to identify whether there were false positive variant calls caused by systematic bias — i.e. by checking whether their allelic fractions were significantly higher than the systematic allelic fraction at those positions — and therefore filter these out. This enables false positive tumour variant calls to be identified even though a similar IncDB could not be produced from tumour sequences due to their non-adherence to Hardy-Weinberg equilibrium. My preliminary analysis of the neoantigenic variants found in chapter II showed that none of these were caused by systematic bias, supporting the accuracy of these variant calls and the results of the survival analysis.

Accurate SARS-CoV-2 genome sequencing is vital for effective monitoring of the global pandemic. It is therefore important to ensure that the accuracy of SARS-CoV-2 genome analysis pipelines is assessed regularly to keep best-practice bioinformatics recommendations up to date. Using sequenced SARS-CoV-2 samples from a cohort of 884 patients under the ARTIC network protocol with subsequent Oxford Nanopore Technology (ONT) sequencing, I examined the differences in consensus sequences produced by base and variant callers. I performed a range of quality control measures to annotate SARS-CoV-2 mutations that exhibited features suggesting a risk of decreased sequencing accuracy with one or more of these base or variant callers. More than 90% of these were not identified in two recent blacklists of ONT-called SARS-CoV-2 mutations in independent cohorts where sequencing was shown to be inaccurate. Guppy v4.0 (in the high-accuracy mode) and nanopolish were the best performers respectively, suggesting that the accuracy of the first wave of SARS-CoV-2 genomic sequences could be improved with more up-to-date base calling. I identified systematic and non-systematic biases in base calling, and differentiated between sequencing artefacts and potentially real intra-patient genetic diversity. I propose the use of a mutation blacklist when considering the accuracy of SARS-CoV-2 variant calls to improve the quality of future SARS-CoV-2 sequences. The application of IncDBs to SARS-CoV-2 sequencing in this project revealed valuable insights about systematic bias that added

to the approach developed in chapter III. In particular, the small size of the SARS-CoV-2 genome allowed systematic bias to be measured for indels as well as SNVs, and this revealed that systematic bias towards deletions was the largest source of systematic bias in SARS-CoV-2 ONT sequencing. This suggests that improving the computational efficiency of the IncDB-based approach in chapter III would be a valuable future direction, since it would allow the detection of systematic deletion bias, which may be an important source of variant call errors in human genome sequencing also.

I have developed flexible bioinformatics methods, which I share so that further advances in genomic sequencing accuracy, quality control and clinical utility can take place. These methods can be easily adapted to different disease areas, organisms and types of sequencing, and hold potential beyond the uses I have established. There remain challenges in the broad fields of quality control and clinical utility of variants that this thesis does not address. The utility of the tools described here depends upon routine sequencing and clinical staff having the confidence and training to utilise genomic information in tailoring patient care, regardless of how developed these tools are. The use of genomic data and bioinformatics is likely to become more widespread as its availability increases, with diagnostic and prognostic measures shifting further away from traditional methods, but it may take longer to integrate novel bioinformatics tools into clinical practice, especially within large healthcare systems. The methods described here will be used and adapted in the coming years to build upon what I have achieved, but may have stronger impacts across scientific and pharmaceutical research than clinical practice as a result. More specifically, gaps remain in the use of mutation and neoantigen-based prognostic tools across many less-studied cancer types and non-cancer diseases influenced by the immune system. Future applications of my research could focus on these in combination with non-genomic prognostic tools. This could help pharmaceutical companies to more effectively target candidate drugs towards the patients who would most benefit from them across disease areas, and help clinicians to choose optimal treatment regimens tailored to individuals.