Development of new methods for accurate estimation of tumour heterogeneity

by

Sebastian Hollizeck

ORCID: 0000-0002-9504-3497

A thesis submitted in total fulfillment for the
degree of Doctor of Philosophy

in the
The Sir Peter MacCallum Department of Oncology
Melbourne Medical School
THE UNIVERSITY OF MELBOURNE

02/03/2022

THE UNIVERSITY OF MELBOURNE

Abstract

The Sir Peter MacCallum Department of Oncology
Melbourne Medical School

Doctor of Philosophy

by Sebastian Hollizeck
ORCID: 0000-0002-9504-3497

Intra-patient tumour heterogeneity is a widely accepted cause of resistance to therapy [1, 2], but the possibility to study this phenomenon is so far underexplored as the acquisition of multi region data sets is costly and ethically challenging [3]. With circulating tumour DNA (ctDNA) as a proxy it is possible to analyze a snapshot of the unified heterogeneity, but there is still an unmet need for new analysis methods to optimize the analysis of these very valuable data and drive new treatment targets [4]. In this work we will develop new methods to study genetic heterogeneity from next generation sequencing (NGS) of tumour tissue as well as ctDNA to elucidate the role of tumour heterogeneity on treatment resistance.

Declaration of Authorship

I, SEBASTIAN HOLLIZECK, declare that this thesis titled, “Development of new methods for accurate estimation of tumour heterogeneity“ and the work presented in it are my own. I confirm that:

- The thesis comprises only my original work towards the DOCTOR OF PHILOSOPHY except where indicated in the preface;
- due acknowledgement has been made in the text to all other material used; and
- the thesis is fewer than the maximum word limit in length, exclusive of tables, maps, bibliographies and appendices as approved by the Research Higher Degrees Committee.

Signed: _____

Date: _____

Preface

This preface includes a summary of all chapters in this work as well as a comprehensive summary of my contributions and everyone else's contribution. This is a thesis *with* publications and each publication included in a chapter is shown here.

Hollizeck S., Wong S.Q., Solomon B., Chandrananda D.[1], Dawson S-J.[1] **"Custom workflows to improve joint variant calling from multiple related tumour samples: FreeBayesSomatic and Strelka2Pass"** *Bioinformatics.* 2021. DOI: 10.1093/bioinformatics/btab606

0.0.0.0.1 Chapter 1: Introduction is an original work providing background and overview relevant to understanding the thesis and its relevance to the field. It includes an introduction to DNA, ctDNA, DNA sequencing, somatic variant calling and lung cancer.

0.0.0.0.2 Chapter 2: Joint somatic variant calling is an original work describing two workflows for the joint analysis of multiple related tumour samples and has been published in *Bioinformatics* as "Custom workflows to improve joint variant calling from multiple related tumour samples: FreeBayesSomatic and Strelka2Pass" on 21^{st} September 2021. In addition to the published analysis, I have added longitudinal analysis and its evaluation as well as the impact of this new method on other downstream analysis, like phylogenetic reconstruction and clonal deconvolution.

Contributions for this chapter:

- I conceptualised the work
- I implemented the workflows and containerised all required tools
- I performed the data simulation
- I performed the analysis presented in the publication
- I wrote the draft of the manuscript and performed revisions
- D.C. and S-J.D. provided advice in planning and writing the manuscript
- D.C. provided guidance for method development

[1]These authors contributed equally and are considered shared last.

- S-J.D. provided guidance for method evaluation
- S.W. performed the targeted amplicon validation
- S.W. and B.S. read the draft manuscript and provided feedback
- B.S. provided clinical expertise for human data

0.0.0.0.3 Chapter 3:

summary plus contributions

0.0.0.0.4 Chapter 4:

0.0.0.0.5 Chapter 5:

0.0.0.0.6 Other publications These publications I have contributed to in my candidature, but they are not presented in this work

Burr M.L., Sparbier C.E., Chan K.L., Chan Y-C., Kersbergen A., Lam E.Y.N., Azidis-Yates E., Vassiliadis D., Bell C.C., Gilan O., Jackson S., Tan L., Wong S.Q., **Hollizeck S.**, Michalak E.M., Siddle H.V., McCabe M.T., Prinjha R.K., Guerra G.R., Solomon B.J., Sandhu S., Dawson S-J., Beavis P.A., Tothill R.W., Cullinane C., Lehner P.J., Sutherland K.D., Dawson M.A. **"An evolutionarily conserved function of polycomb silences the MHC class I antigen presentation pathway and enables immune evasion in cancer"** *Cancer cell.* 2019. DOI: 10.1016/j.ccell.2019.08.008

Solomon B.J.[2]., Tan L.[2], Lin J.J.[2], Wong S.Q.[2], **Hollizeck S.**[2], Ebata K., Tuch B.B., Yoda S., Gainor J.F., Lecia V. Sequist L.V., Oxnard G.R., Gautschi O., Drilon A., Subbiah V., Khoo C., Zhu E.Y., Nguyen M., Henry D., Condroski K.R., Kolakowski G.R., Gomez E., Ballard J., Metcalf A.T., Blake J.F., Dawson S-J., Blosser W., Stancato L.F., Brandhuber B.J., Andrews S., Robinson B.G., Rothenberg S.M **"RET Solvent Front Mutations Mediate Acquired Resistance to Selective RET Inhibition in RET-Driven Malignancies"** *Journal of Thoracic Oncology.* 2020. DOI: 10.1016/j.jtho.2020.01.006

Fennell K.A.[2], Vassiliadis D.[2], Lam E.Y., Martelotto L.G., Balic J.J., **Hollizeck S.**, Weber T.S., Semple T., Wang Q., Miles D.C., MacPherson L., Chan Y-C. Guirguis A.A., Kats L.M., Wong E.S., Dawson S-J., Naik S.H., Dawson M.A. **"Non-genetic determinants of malignant clonal fitness at single cell resolution"** *Nature.* 2021 DOI: 10.1038/s41586-021-04206-7

add more papers if they are published before the end

[2]These authors contributed equally and are considered shared first.

0.0.0.0.7 Funding:

All necessary funding goes here

0.0.0.0.8 Instructions: Where applicable, the following information must be included in a preface:

- a description of work towards the thesis that was carried out in collaboration with others, indicating the nature and proportion of the contribution of others and in general terms the portions of the work which the student claims as original;
- a description of work towards the thesis that has been submitted for other qualifications;
- a description of work towards the thesis that was carried out prior to enrolment in the degree;
- whether any third party editorial assistance was provided in preparation of the thesis and whether the persons providing this assistance are knowledgeable in the academic discipline of the thesis;
- the contributions of all persons involved in any multi-authored publications or articles in preparation included in the thesis;
- the publication status of all chapters presented in article format using the descriptors below;
  - Unpublished material not submitted for publication
  - Submitted for publication to [publication name] on [date]
  - In revision following peer review by [publication name]
  - Accepted for publication by [publication name] on [date]
  - Published by [publication name] on [date]
- an acknowledgement of all sources of funding, including grant identification numbers where applicable and Australian Government Research Training Program Scholarships, including fee offset scholarships.

Acknowledgements

The acknowledgements and the people to thank go here, don't forget to include your project advisor...

Lots of figures in the introductory chapter 1 were reated with the help of BioRender.com

think of where to put the package citations; Probably at the end as appendix

Software and packages

This section is dedicated to all the software that usually gets uncited because they are "standard" or backbone

Most analysis in a prototype state was done on a linux cluster running Centos 7 [5] with Bash [6] and due to the high amount of data, parallel [7] was used of the multi-cpu architecture of HPCs.

R

In depth data analysis and visualisation was done with R [8] with the help of packages listed below.

Most of the parallelisation in R was performed with BiocParallel [9], which is available through BiocManager [10].

Colour shemes and manipulation was performed with colorspace [11, 12].

Copynumber analysis was performed with sequenza [13], FACETS [14, 15] and PURPLE [16]. Some analysis was also directly performed with copynumber [17, 18].

Variant effect prediction was performed with VEP [19].

Table manipulation was performed with data.table [20].

Violin plots were generated with vioplot [21].

Heatmaps and UpSet plots were generated with ComplexHeatmap [22]

Phylogenetic analysis was performed with both ape [23] and phangorn [24] followed by dendextend [25].

Google sheets and its built in scripts were used to collect stats on docker pull requests and the data was then read in R through googlesheets4 [26].

Additional libraries, which were used for a multitude of things are listed in no particular order below: Rsamtools [27], GenomicRanges [28], optparse [29], VariantAnnotation [30], MultiAssayExperiment [31], circlize [32], BioQC [33], Biostrings [34], deconstructSigs [35], BSgenome [36], QDNAseq [37], RColorBrewer [38], pheatmap [39], ensemblVEP [40], stringdist [41], Rsubread [42], svglite [43], grImport [44], XML [45], kableExtra [46], lsa [47], irlba [48], ggplot2 [49]

maybe itemize over just a blob

python

Analysis for chapter 4 was mostly done through python [50] with the help of many different packages, which are listed here in no particular order: numpy [51], ncls [52], pysam [53, 54, 55], zarr [56], pandas [57, 58], quadprog [59] as well as scipy [60].

latex

Of course, finally the typesetting of the thesis itself was done with LaTeX. With these additional packages in no particular order: babel, csquotes, lmodern, CrimsonPro, fontenc, xcolor, hhline, siunitx, biblatex, hyperref, quotchap, todonotes, float, afterpage, multicol, enumitem, array, tocloft, caption, appendix, xurl, graphicx, epstopdf, subfigure, booktabs, rotating and listings. The bese class is 'book' and all packages are available on CTAN and the source code is available at my GitHub repository https://github.com/SebastianHollizeck/PhDThesis.

Contents

List of Figures

List of Tables

List of Equations

Listings

Abbreviations

BAM	**B**inary **A**lignment **M**ap
bp	**b**ase **p**air
BQ	**B**ase **Q**uality
cfDNA	**c**ell **f**ree **DNA**
ChIP	**Ch**romatin **I**mmuno**P**recipitation
ctDNA	**c**irculating **t**umour **DNA**
DBS	**D**ouble **B**ase **S**ubstitution
DNA	**D**eoxyribo**N**ucleic **A**cid
F81	**F**elsenstein 19**81** model
GATK	**G**enome **A**nalysis **T**ool**K**it
HKY85	**H**asegawa, **K**ishino and **Y**ano 19**85** model
HPC	**H**igh **P**erformance **C**omputing
ILM	**I**terative **L**inear **M**odels
InDel	**In**sertion or **Del**etion
MQ	**M**apping **Q**uality
MRCA	**M**ost **R**ecent **C**ommon **A**ncestor
NGS	**N**ext **G**eneration **S**equencing
NJ	**N**eighbour **J**oining
NSCLC	**N**on-**S**mall **C**ell **L**ung **C**ancer
PET	**P**ositron **E**mission **T**omography
PON	**P**anel **O**f **N**ormals
QP	**Q**uadratic **P**rogramming
RAID	**R**edundant **A**rray of **I**ndependent **D**isks
RNA	**R**ibo**N**ucleic **A**cid
ROI	**R**egion **O**f **I**nterest
RPRS	**R**ead **P**osition **R**ank **S**um
SBS	**S**ingle **B**ase **S**ubstitution

SCLC	**S**mall **C**ell **L**ung **C**ancer
SNP	**S**ingle **N**ucleotide **P**olymorphism
SV	**S**tructural **V**ariant
TAS	**T**argeted **A**mplicon **S**equencing
TKI	**T**yrosine **K**inase **I**nhibitor
TNBC	**T**riple **N**egative **B**reast **C**ancer
UPGMA	**U**nweighted **P**air **G**roup **M**ethod with **A**rithmetic mean
UV	**U**ltra **V**iolet light
VCF	**V**ariant **C**all **F**ormat
WES	**W**hole **E**xome **S**equencing
WGS	**W**hole **G**enome **S**equencing
WPGMA	**W**eighted **P**air **G**roup **M**ethod with **A**rithmetic mean

"Begin at the beginning," the King said, very gravely, "and go on till you come to the end: then stop."

— Lewis Carroll, *Alice in Wonderland*

1

Introduction

This first introduction chapter contains all the necessary background information as well as an overview for the work discussed in this thesis. It summarised basic biological properties of DNA and cell biology as well as the respective technologies to read, analyse and measure these biological concepts and then how to evaluate the output of these methods. Section 1.1 delineates the role DNA plays for the cell and then section 1.2 shows how these standards are changed in the tumour and cell free context. Section 1.3 introduces the current technologies used to measure and detect DNA and its variations. With section 1.4 covering the computational analysis methods to read out changes in the DNA. Then section 3.1.1 relates how these changes lead to cancer and what we can learn from them. The introduction concludes with section 1.6 as an overview over the thesis aims and my work in addressing them in the following chapters.

1.1 DNA as a information storage unit

It is a widely accepted fact, that Deoxyribonucleic acid (DNA) serves as the long term information storage molecule of our cells. This information is protected and allows correction of simple errors through its double helix structure [61, 62]. The nucleotides, which consist of a deoxyribose sugar (hence the name), a phosphate group and the nitrogenous base, are joined together by phosphate groups. Even though there are six common naturally occurring nitrogenous bases: Adenine (A), Thymine (T), Guanine (G), Cytosine (C), Uracil (U) and nicotinamide, only the first four are used to encode the genetic information into DNA. Each of the strands mirrors the other, so that an adenine will be paired up with a thymine forming two hydrogen bonds. Similarly, cytosine will pair with guanine forming an even stronger bond with three hydrogen bonds. While other pairings which do

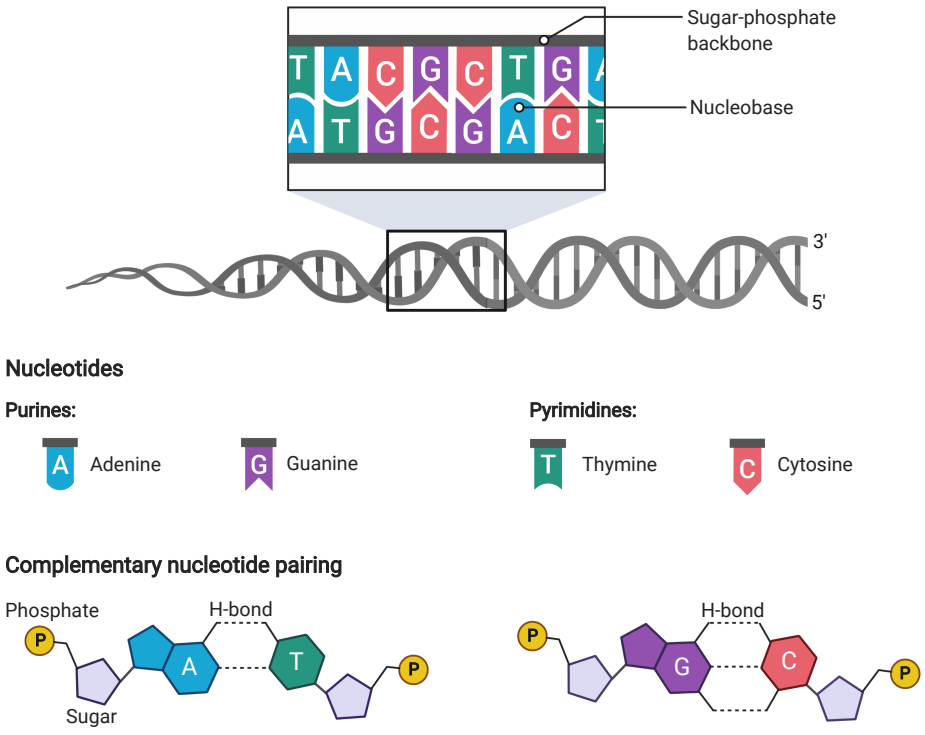

Figure 1.1: Overview of DNA structure and the nucleobases, which form DNA strands. Nucleotides are split into Purines and Pyrimidines by the structure of the nitrogen ring; complementary pairing of bases is shown as shapes of the bases as well as with 2D structures; Hyrdogen (H) bonds are shown as dotted lines; Phosphates are shown as P; 3' and 5' ends are defined by the internal number of the carbon atom of the sugar which is exposed; Adapted from "DNA structure" by BioRender.com (2021) Retrieved from https://app.biorender.com/biorender-templates

not follow those rules are chemically possible, they are mostly observed in ribonucleic acid (RNA) [63]. These very strict bonding rules enable the DNA to be similar to a hard drive with backup on a computer. And as only one strand contains all the information, the DNA polymerase enzyme does only need access to one strand, which allows parallel replication during cell division, but also error corrections, by proof reading the newly synthesised strand with the template. In order to be able to distinguish the two strands, they were assigned the names 3' and 5' depending on the numbering of the carbon atom in the sugar, which is exposed (Figure 1.1).

The entirety of the DNA encoding the organism is commonly called "the genome" with all genes, which consist of introns and exons are called exome. Unicellular organisms usually only have a very small amount of introns, which to current knowledge only provide limited information and are only responsible for structure. In vertebrates introns as well as intergeneic DNA (the DNA between genes) contribute most of the DNA in the genome. For example in humans, only 1% of the genetic

material is considered to be exonic, whereas introns contribute $\approx 24\%$ and the rest is intergeneic ($\approx 75\%$)[64].

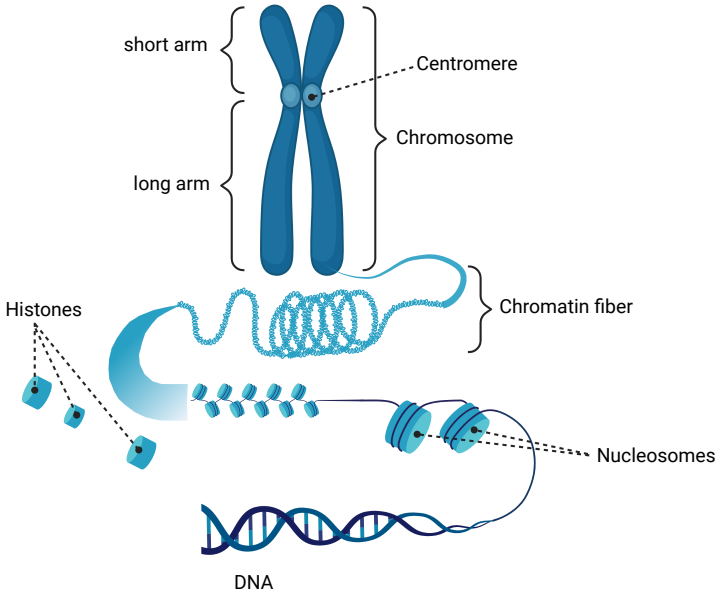

Figure 1.2: Structural overview of the metaphase condensed chromosome: DNA is first wrapped around Histones to form nucleosome, which then associate with each other to form the chromatin fiber, which in the metaphase of the cell cycle is condensed even more into the X-shaped chromosome

The DNA in eukaryotes however is not free floating around in the nucleus of a cell, but rather in most eukrayotic organisms, it is highly condensed and structured, first wrapped around nucleosomes like thread on a spool, then organised around histones, into either open (accessible) or closed chromatin, which then can be even further condensed into chromosomes, which have a X-like shape, with two shorter and two longer arms (Figure 1.2). This allows some DNA to be accessible where the use of other areas can be restricted[65]. Through this restriction, the availability of certain genes, which are the sections of the DNA, which encode for short term storage molecules like RNA. This restriction plays an important role in cell fate and cell viability. Ultimately, all information stored to create a new highly complex organism is stored in just the DNA of one cell. Whichever parts are used and how they are used decides the function and the identity of the cell[66].

1.1.1 Ploidy - it is good to have a backup, if you do it right

Similar to the already discussed RAID-like setup of the DNA in two strands, another concept of data security, a spatial different storage is also implemented. Most eukaryotic organisms have at least

two of each chromosome (diploid) with some species reaching up to septaploid[67]. However, this concept is not the only reason for the ploidy of somatic cells. For sexually reproducing organisms, at least a diploid set of chromosomes is necessary to enable information to be joined from both parents. Germline cells (sperm and egg) are generally monoploid, such that the resulting cell will be diploid, but the ploidy of the somatic cells is not as uniform within a species, where it can vary between organisms based on gender or rank [68]. In most organisms, a change in ploidy is fatal [69] and only partial ploidy changes like extra copies of chromosome 17 [70], chromosome 18 [71] and chromosome 21 [72] are tolerated. These syndroms can occur when the is an uneven split of chromosomes during cell division. The additional advantage, apart from sexual reproduction, is that a second almost identical copy of a chromosome allows repair of DNA, even when both strands are damaged, for example in a double strand break. In this case, the information from the sister chromosome will be used, by first cutting the double strand break ends to have an overhang (resection). This overhang will then merge with the sister chromosome's mirrored strand. In this state, the two chromosomes are fused together in a Holliday junction, which allows the missing part from the resection and the double strand break to be synthesised [73]. During this process, which is part of the homology directed repair (HDR) machinery, the sister chromosomes exchange parts of their DNA, when resolving the Holliday junction. As these stretches of DNA do not need to be 100% identical, this plays an important role in evolution and diversity [74, 75].

Even though this X-like structure is the most commonly used and known structure, the DNAs 3D structure is usually very different and only takes this shape for the very short time of the cell cycle. Most of the time, the chromosomes are unravelled into something resembling a ball of yarn, where the "open" chromatin regions are on the outside and the "closed" regions are "hidden" in the inside and each chromosome establishes its own "territory" inside the nucleus (Figure 1.3). This structure allows another DNA cross over with non-sister chromsomes, which is called a chiasma.

1.1.2 Phantastical mutations and where to find them

However even though the DNA is highly stable, and error correction methods are constantly working to not introduce any changes in the DNA, the source of evolution and adaptation of species is sourced in a steady mutation rate [76, 77]. These changes in normal tissue are mostly irrelevant to the organism as a whole and will not be passed on to the next generation. These changes are known as somatic mutations. This type of mutation accumulates in a cell linearly over the course of the lifespan of the cell and is not bound to just cell divisions[78, 79]. In contrast, if one of those

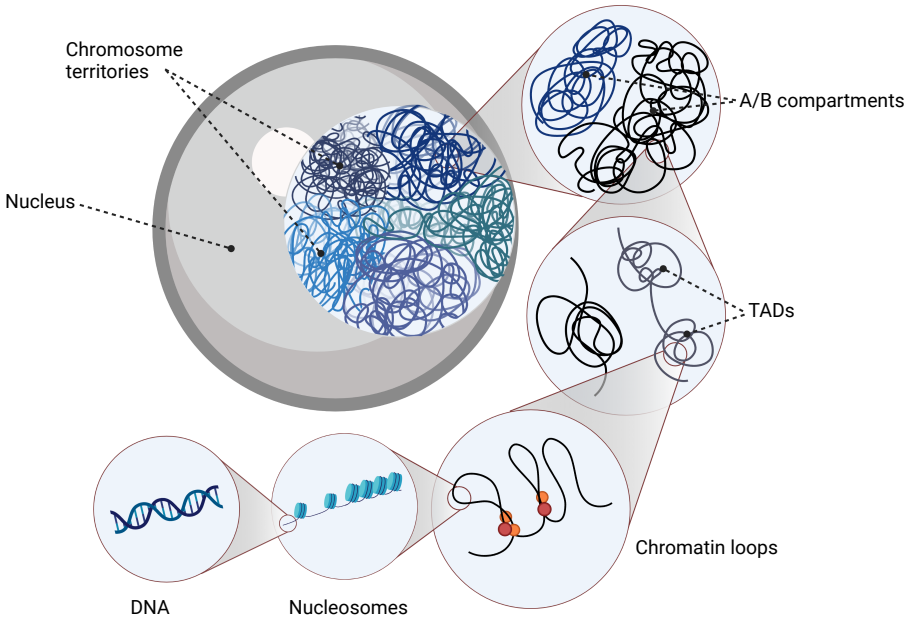

Figure 1.3: Individual chromosomes occupy a subspace in the nucleus called chromosome territories. Chromosome territories can be further partitioned to distinct A and B compartments, which are enriched for active and repressed chromatin, respectively. Genomic regions within topologically associating domains (TADs) display increased interactions, while their interactions with neighbouring regions outside the TADs are rather limited.

mutations occurs in the germline cell, e.g. sperm or egg producing cells, these mutations will be propagated to all offspring and be present in all cells of that organism and in term all its offspring. These mutations are called germline mutations. These mutations are also called germline variants, as they establish in the population and represent a variation of the organism. Mutations can also be classified depending on either their size, ranging from single nucleotide polymorphisms (SNPs) over small insertions or deletions (InDels) to large structural changes, like the deletion of parts of or even a whole chromosome arm. Like previously described with ploidy changes, usually smaller changes have less impact on the overall fitness of the organism, however even SNPs can lead to changes which are not compatible with life[80, 81].

1.2 Cell free DNA is more than just bits and pieces

When a cell from a multicellular organism dies, through which ever method, there will be numerous enzymes involved, which clear the debris and recycle material. This means that proteases digest proteins into amino acids, which will later be used for either building new proteins or possibly

even digested further for energy production. The same happens with the DNA in the cell. However, as discussed in the previous section 1.1 the DNA is wrapped around histones and organised in structures called nuclesomes. These protect the DNA from being cut by DNAases by hindering the access to the DNA, similar to how they stopped the access for transcription into RNA. This then in turn leads to the DNA being cut into pieces mainly in the length of 167 base pairs (bp). These DNA fragments, which are called cell free DNA (cfDNA), can then be detected in bodily fluids, like blood or even stool. By analysing these fragments, non invasive tests for prenatal care have been possible, as the DNA of the foetus is detectable in the mother's blood [82, 83]. Similar to the process, a cancer also sheds DNA, titled circulating tumour DNA (ctDNA), when its cells die, either through intervention of the immune system or through other forcefull processes. These ctDNA fragments can also be analysed and molecular properties measured, without even knowing the exact location of the tumour. As a blood test can be routinely performed in the clinic or even a general practitioner, the monitoring of cancer progression is significantly easier and safer than through other measures. Of course, it is, similar to the prenatal test, only a proxy for the cells which are still alive, as these have not shed their DNA. Additionally, the amount of shedded DNA is highly variable between tumours, with a general higher amount for later stages, so that sometimes there is almost no ctDNA present, even though the cancer is fairly advanced [84, 85].

include the length of ctDNA

1.3 DNA sequencing - when is next generation sequencing the current generation?

As we know the building blocks, that make DNA as well as the process and the enzymes responsible, we can synthesise DNA in vitro. By chemically modifying the nucleotides supplied to the synthesis process, the sequence of the copied strand can be analysed. The first method to make use of this used the lambda phage to fuse known ends for the primers needed for the reaction to the piece of DNA and supplied labelled nucleotides [86]. This method was then superseded by "Sanger sequencing" after Frederick Sanger who with colleagues published this method in 1977, by adding **di**deoxynucleotides in a low concentration, the polymerase chain reaction would terminate trying to integrate these nucleotides and by labelling them radioactively or florescently, a gel can be used to determine the sequence of a piece of DNA [87, 88], which made the method better suited for larger scale projects.

However, this method has multiple issues for modern research questions. Mostly, that it is fairly labour and time consuming to analyse multiple pieces of DNA at the same time, and it is very challenging to sequence all the DNA of an organism. The human genome project, which was started in 1990 used machines which automated the Sanger sequencing procedure and it still took hundreds of researchers 13 years to complete the DNA sequence of just one human [89, 64]. Even though this was a very long project, it laid the ground work for the usage of the current sequencing technologies.

1.3.1 Library preparation - what we learned from using phages

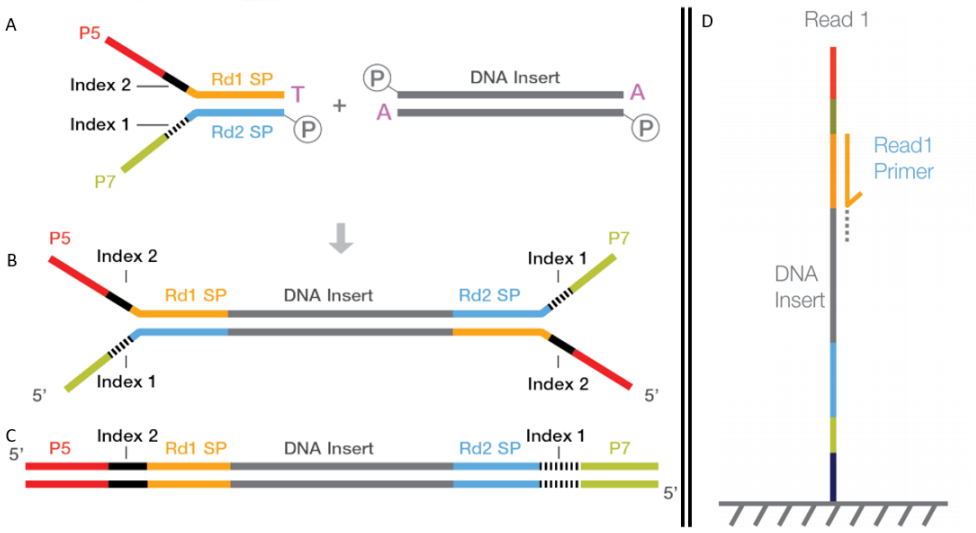

Figure 1.4: Adapter ligation during library preparation. The adapters are added to the DNA insert during library preparation. A. The DNA insert is prepared by adding an A-tail and phosphorylation. B. The adapter complex which includes the P5/P7 flow cell binding adapter is added to the DNA insert. C. The DNA insert is ready for sequencing. D. The DNA insert binds to the flow cell for sequencing. Primers bind to the DNA insert to generate reads;
Figure adapted from "How short inserts affect sequencing performance"[90]

Library preparation is the name of the preprocessing step, which is done before it is sequenced with the current technologies. The first step to sequence DNA is to obtain the DNA, which is done by lysing the cells of interest, which disrupts the cell membrane and therefore spills all its contents. The then spilled DNA is fragmented into smaller pieces, by either restriction enzymes or sonication, to have a size of about between 200-800bp. These steps are not necessary when preparing sequencing of ctDNA, as discussed in Section 1.2, the DNA is unbound and already digested into short fragments. Once the DNA is ready, it is phosphorelated and an A-tail is added, before the adapter complex is ligated. This enabled the DNA to bind to the flow cell which is covered with the reverse complement of the adapter (Figure 1.4).

1.3.2 Next generation sequencing

Next generation sequencing (NGS)is the coined term for basically any standard high-throughput sequencing performed, which includes exome, genome, transcriptome, protein-dna interactions (ChIP) and other epigenome studies. The term NGS is still widely used, even though it has been more than 10 years since the first NGS approach was commercially available. While in the beginning of next generation sequencing there were multiple approaches, the current lion share (80% of sequencing data) of protocols use the Illumina short read sequencing by synthesis approach (Figure 1.5)[91, 92], which is based on the concept of alternating integration of florescently labelled nucleotides and imaging with a microscope (Figure 1.5) as well as multiplexing, where a DNA fragment is ligated to an index, which allow to sequence multiple samples at once [93, 94] as it is shown in Figure 1.4. This method allows highly accurate determination of the sequence of a DNA fragment and depending on the flow cell and sequencing machine allows to sequence a whole genome in just 24h.

1.3.3 Long read sequencing - the "third" generation sequencing

By now, multiple methods which broke free of the size limitations of NGS exist, which are commonly referred to as long read sequencing. Most of the current methods trade the very high accuracy of the second generation NGS methods for the capability of sequencing huge continuous strands of DNA (current record 2.3 Million bp [95]) with normal library preparation ranging between 10-30 Kbp. These methods are expected to revolutionise our understanding of the highly repetitive elements that exist in the genome, such as the centromeres of chromosomes. Methods such as the direct molecule sequencing approach by Oxford Nanopore are even able to distinguish post transcriptional modifications on RNA[96]. So far, these methods however are still very expensive and as this work is dealing with ctDNA, which is highly fragmented, these methods offer only limited advantages over the short read sequencing, while being much more expensive.

1.4 DNA analysis - what to do with the sequence

The types of analysis that can be done with the output from the sequencing machine stretches far, however, all methods need to first infer the location in the genome, the sequenced piece of DNA

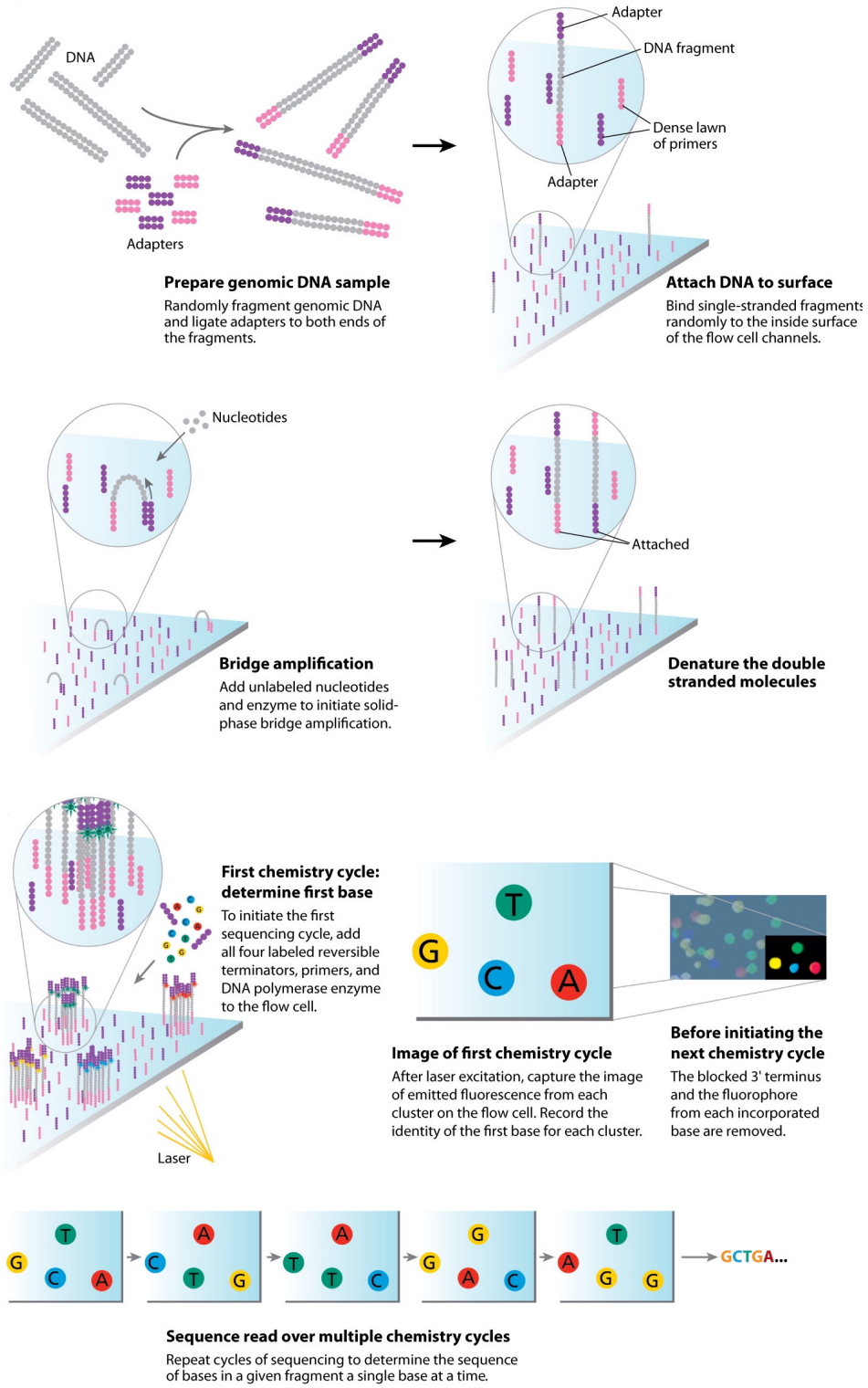

Figure 1.5: The Illumina sequencing-by-synthesis approach. Cluster strands created by bridge amplification are primed and all four fluorescently labeled, 3'-OH blocked nucleotides are added to the flow cell with DNA polymerase. The cluster strands are extended by one nucleotide. Following the incorporation step, the unused nucleotides and DNA polymerase molecules are washed away, a scan buffer is added to the flow cell, and the optics system scans each lane of the flow cell by imaging units called tiles. Once imaging is completed, chemicals that effect cleavage of the fluorescent labels and the 3'-OH blocking groups are added to the flow cell, which prepares the cluster strands for another round of fluorescent nucleotide incorporation; Figure adapted from Mardis[91]

originated from. As the current methods randomly fragment the DNA (Section 1.3.1), the genomic location information is completely lost. This process is referred to as mapping.

1.4.1 Mapping - Ey man, where is my genomic location?

In this process, the fragments of DNA, which were sequenced, are assigned a genomic coordinate on the reference genome. This is only possible, due to the fact, that we have a resolved genome sequences (Section 1.3) for a high number of species. The location a sequenced piece of DNA fits to the reference genome might be unique, but it could also fit to multiple locations, due to highly repetitive regions or due to the existence of pseudo genes with almost 100% identify. In addition to this, the reference genome might not accurately reflect the genome of the organism that has been sequenced. Each mapping position is therefore assigned a quality score, which reflects how likely it is the actual position of the sequence. As Illumina sequencers have the ability to sequence both ends of the DNA fragment, the position of the ends (read 1 and read 2) to each other can also be used to infer the quality, as they should be within a reasonable distance to each other (Figure 1.4)

As this process is time-consuming and the exact location of the fragment might not be as important, there exists a subset of tools called pseudo-mapper, which are based on k-mers, which are predefined DNA sequences of length k, which help to identify certain regions of interest. These tools are especially common for RNAseq, where the exact location of a read does not matter, only that the read is within a gene [97, 98], but also for methods that estimate similarity between sequences (DNA, RNA or protein) [99, 100].

For this work however, the exact position of reads is crucial, so only real mapping methods like BWA [101] or Bowtie 2 [102], which are optimised for short reads from Illumina systems, provide the necessary functions.

add things about alternative contigs and reference genome?

1.4.2 Variant calling - spot the difference

As intra-species genetic variation is intended for adaptation and evolution, there will be places where the DNA sequence of the subject will differ from the sequence of the reference (see Section 1.1.2). These variants give insight into medical background as well as treatment options for patients and can even be used to guide family planning. Depending on the type of variation that is

of interest, a different set of computational methods are needed, as germline and somatic variants have different properties.

1.4.3 Germline variant calling - the cards you have been dealt at birth

The most common source of DNA used for germline variant analysis is the mono nuclear layer from the blood of the subject, but really almost any cell can be used for this process, as all cells in the organism will share all germline variants (Section 1.1.2). The only important input on top of the DNA sequence from the sequencer are the reference genome of the organism, as all variant nomenclature is based on the reference and the ploidy of the organism (Section 1.1.1). The ploidy is key to infer, at which ranges of allele frequency a variant can biologically occur. For example in a human diploid genome, germline variants can occur either in one or both chromosomes, which mean we assume reads should show an allele frequency of around 50% and 100%, where the hexaploid commercial wheat [103] allele frequency for variants would be 16%, 33%, 50%, 66%, 0.83% and 100%. Due to the random sampling and possible sequencing errors, the observed allele frequencies will differ from the theoretical values. Most state of the art germline variant calling method will also use haplotype reconstructions through de-Bruijn graphs, which features a remapping of reads in relation to each other [104, 105, 106, 107, 108] where the original mapping location assigned by the aligner (Section 1.4.1) is only used as a guideline. This allows to resolve even complex haplotypes of the sample by not restricting the method to the linear setup of the reference genome.

1.4.4 Somatic variant calling - life is ever-changing

In contrast to germline variant calling, somatic variant calling methods cannot rely on allele frequency, as not all cells sequenced are expected to have the change in nucleotide. The allele frequency is instead a measure of the sub clonal size. A subclone is here defined as the set of cells, which were derived from the cell, which originally acquired the somatic mutation. Depending on the selective advantage, just random drift and also the time point when the variant was introduced, these clones can be very variable in size and therefore their contribution to the DNA in the sequencing. As not all cells have the variants, the selection of the tissue for library preparation is very important, unlike for germline calling. The main use of somatic variant calling is the genetic diagnosis and research of cancer samples, where the main question is, which changes are present in the tumour, which lead to the disease.

The ideal scenario for tumour somatic variant calling is when a biopsy of the tumour as well as a normal sample of the patient is available. In most clinical cases, this will be the diagnostic biopsy as well as the mono nuclear layer from blood, just like for germline calling (Section 1.4.3). This needs to be adjusted depending on the type of malignancy, because if the tumour is a leukemia, the mono nuclear layer of the blood might contain tumour cells, but for solid tumours, the blood is a routine, minimally invasive option. These two samples are then analysed together and only changes that are only in the somatic tumour sample and not in the normal sample are reported. Even though this concept sounds simple, there are some pitfalls [109]. First, there might be some tumour contamination in the normal sample, which needs to be adjusted for [106, 110]. Second, there might be normal "contamination" in the tumour sample, this means that not all cells in the tumour sample are actually tumour. This means that the signal of the tumour changes is reduced and harder to find.

All of these issues are amplified in the case, when there is no "normal" sample available, either because the patient didnt consent, due to other medical issues, or because for diagnostic tests there usually is no need for a germline sample. In this case, there is the option for "tumour only" variant calling, which requires a database of germline variants in the population, to distinguish between somatic and germline variants, as the variant calling is very similar to just germline variant calling (Section 1.4.3) without the restriction of the ploidy. However, even with an extensive database like gnomAD [111] it is unlikely to be able to remove all germline variants from the analysis and as there is no direct comparison, the precision of the "tumour only" method is significantly lower [112].

1.5 Cancer

For a long time in human history, the origin of cancer as a disease was a mystery and a multitude of theories, starting in ancient Egypt, were developed. These theories ranged from a curse to chemical imbalance over parasites to trauma. In this section I will outline both the history of cancer as a disease and the treatments starting with ancient times leading up right until the current times. While the first steps are very wide, because the biology itself was not understood, it is quite curious how often people with more knowledge came to worse conclusions and theories, than were already known thousands of years ago.

Around 3000BC the Egyptians describe the bulging tumour of the breast as an incurable disease[113], even then they already had some ointments, which were used, including resection, cauterisation

and salting of the affected areas, all of which were still used up until the 19^{th} century [114]. This papyrus document is considered the oldest evidence of cancer in humanity.

When the ancient Greeks laid the foundation for modern medicine with Hippocrates, the first hypothesis about natural causes of cancers was formulated and the terms “cancer“ and “carcinoma“ were coined. The abundance and accumulation of “black bile“ in the body was thought to be the cause of the cancers. However, the treatment was still the same as before, with resection and lotions [115].

Following Hippocrates, the Roman physician Celsus progressed the understanding of cancers significantly, by describing metastatic relapse of treated breast cancer in neighbouring armpits and even the spread to distant organs. He also was aware, that the outcome of patients was better, if the tumours were removed early and aggressively [116].

With the destruction of the western Roman Empire, the Middle East became well known for their strong advances towards modern medicine and the court physician to the Emperor of Constantinople Aetius had success with the first total mastectomy and generally was an advocate of the total excision of tumours [117].

Sadly, while both the understanding of cancer and the treatment were steadily improving, the Pope prohibited bloodshed as well as surgeries and therefore lead to a slow-down of advances, especially because autopsies were also forbidden a hundred years later in 1305. However, there were still illegal experiments conducted and the general classification which is still used up to date was started, by Henri de Mondeville, who started classifying tumours by their anatomical site[118].

After the end of the “dark ages“, the wide availability of older medical works from both Greek and Roman due to the book print invention, led to the re-emergence of the use of chemical ointments and lotions on cancer lesions. With Paracelsus promoting the usage of chemicals, which he himself warns are poisonous in the wrong concentration, for the treatment of cancer, he laid the groundwork for the modern Chemotherapy [119].

As the dissection of corpses was no longer banned by the church, more and more cases of “hidden“ causes of death were found post mortem, which were often cancers on internal organs, like the brain but also the detection of malignant and benign tumours was a major breakthrough. This lead to the understanding, that benign tumours might turn malignant after some time and many physicians suggested removal of the benign growths [120].

Due to genetic disposition of cancer, especially breast cancer, two independent physicians (Zacutus Lusitani and Nicholas Tulp) came to the conclusion, that cancer is contagious and proposed isolation of patients [121, 122] which shows, that while the treatment of cancer was improving steadily, but the origin of the disease was still a mystery. It took until 1700 when Deshaies-Gendron described cancer as a transformation of a normal body part, which continues to grow without control and while he was aware of metastatic disease, he suggested no treatment, as he did not believe cancer to be curable with drugs [123].

Another ground-breaking work published in the same year was the collection of almost three thousand autopsy reports and their clinical history, which contain a number of detailed cancer cases including: brain, head and neck, lung, breast, esophagus, stomach, colon, liver, pancreas, kidney, uterus, cervix, bladder, and prostate. Many of the terms used by Theophilus Boneti to describe the cases are still in use and the work itself was the first step toward tumour pathology [124]

However, it took almost 150 years after the theory of cancer being contagious for Nooth [125] to conduct experiments trying to infect himself with cancer pieces resected from another person, which proved that cancers generally are not infectious.

With the invention and consequently common use of the microscope in the pathology, more and more causes of deaths were identified as caused by cancer. An example is the connection of a chronic cough to lung cancer and swollen joints with sarcoma [126].

After more and more autopsies of cancer patients, surgeons like Heister [127] found that breast cancer resection needs to include the breast, the axillary lymph node and the pectoralis major muscle which got to be known as the Halsted radical mastectomy and was the standard of care for a long time.

While the treatment of cancers (mostly surgical) was getting more and more advanced, but the origin and cause of cancer in patients was still very much debated. As there are a manifold of causes as we now know, it is maybe not surprising that it took longer, but by the middle of the 18^{th} century chronic inflammation as a cause of cancer initiation was hypothesised [128].

The next big step was taken, when in 1838 the concepts of cells as fundamental building blocks of organisms was published. In the following years, many cancers were dissected and microscopically analysed. This revealed that tumour cells look vastly different from normal cells, and it was thought that they morphological features could serve to identify their fate [129] and became known

for defining the cellular origin of benign and malignant tumours. And while he described the tumours as a collection of abnormal cells with stroma, he thought cancer to arise from newly generated cells from a diseased organ and thought the underlying cause to be “amorphous embryonal blastema“.

With this foundation, over the next hundred years, lots of advances were made into the morphology of different tumours and many previously undetected ones, like leukemia, were found and extensively characterised. However, even then, there were researchers, which understood that the heterogeneity of cancers is so vast, that while he was convinced that the microscope will be a mandatory instrument to diagnose cancers, more effort to collect and study specimen is necessary to have a complete picture [130].

As many shared the view of Bennett, the second half of the 19^{th} century was a rich source of surgical pathology and the oncology literature in general. Most outstanding was Rudolf Virchow’s “Die krankhaften Geschwülste“ [131] which is a first landmark book on the classification of cancers, and is still a well of knowledge. From his work, the terms “hyperplasia“ and “metaplasia“ we derived, as pre-cancerous states of cells. He also was one of the first to hypothesis the presence of growth stimulating substances around cancers, which lead to their uncontrolled growth. While he also was the first to again oppose the “amorphous embryonal blastema“ theory and instead was convinced that tumour cells were just abnormally changed cells, which he called “chronic irritation theory“ and had a theory that metastasis were seeded by the original lesion (like in this melanoma Figure 1.6), he also had major scientific impact in a number of other fields like Parasitology, Forensic and Anthropology[1].

While the search of possible cancer causing substances started to get more and more interest, only one real cause was thought to be found in the ore of the central European mountains, where miners would have a higher prevalence of lung cancers. Nevertheless, this was later found to be caused by radio active material and not by the inhaled dust of minerals as expected. Similar, many parasites and bacteria were found as potential causes of cancer, but none of the findings could produce proof.

While all these steps were getting closer together in time up until the beginning of the 20^{th} century, they were still fairly minor in the contrast to the high speed and throughput results that the last hundred years brought with it. While technically Willhelm Röntgen discovered the X-rays just before the change of the century [132], both its impact on the body and cancer were only clear a few

[1] Maybe surprising to hear, that he was strongly opposed to Darwin’s theory of evolution. In his own words: “The intermediate form is unimaginable save in a dream... We cannot teach or consent that it is an achievement that man descended from the ape or other animal.“

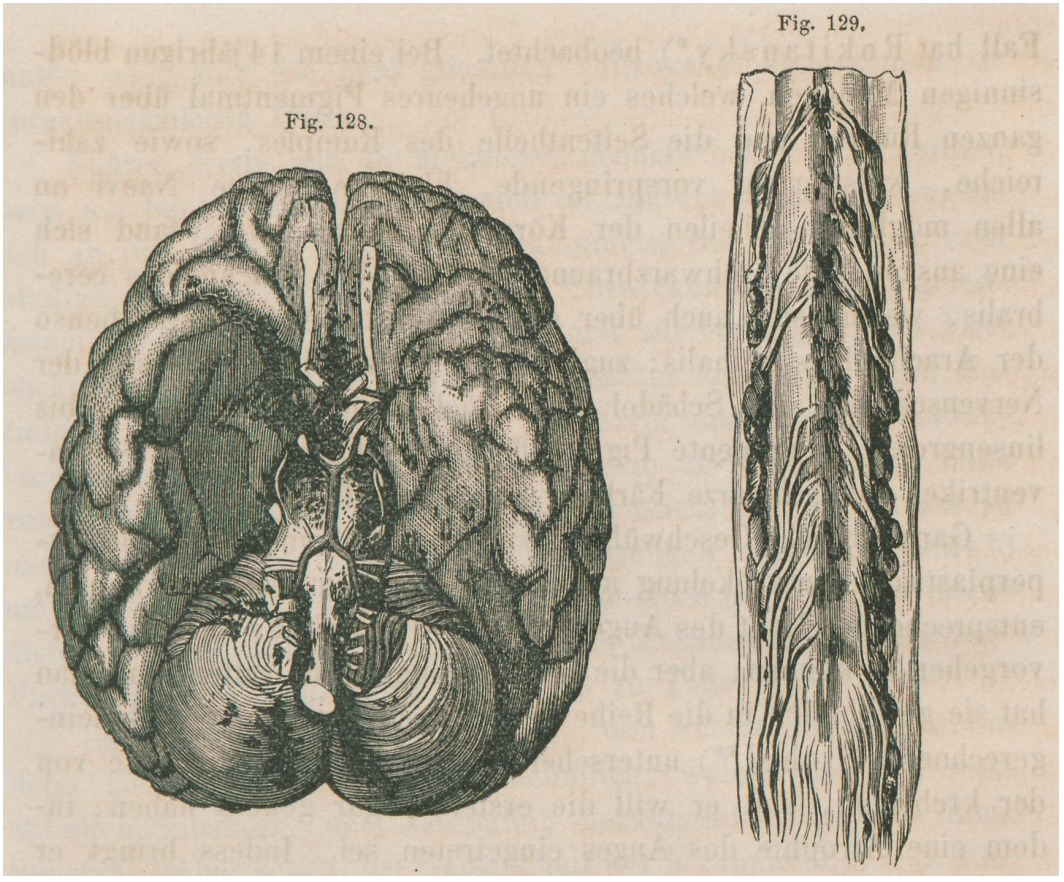

Figure 1.6: Drawing of central nervous system metastasis from page 121, Volume 2 of "Die krankhaften Geschwülste" Virchow [131]; translated original caption: Fig. 128: Multiple melanoma of the Pia mater basilaris, most pronounced around the Medulla oblongata, the Pons, the Fossa Sylvii, Fissura longit (sample No. 256a from 1858); Fig. 129: Lower end of spinal cord of Fig. 128 with multiple melanoma of the soft skin with node like growths at the nerve roots (sample No. 256b from 1858)

years into the last century [133, 134]. However, similar to how X-rays can cause cancers, researcher also found quickly, that it can also treat cancer and thus the field of radiotherapy was created. This then was the first major change in cancer treatment for around five thousand years, which also could treat inoperable cancers.

The next invention, that I want to highlight in the vast amount of advances made in the advent of the 20^{th} century, is the cutting needle aspiration syringe, which allowed a non-traumatic biopsy of internal organs for microscopy study. Which made it possible to not have exploratory surgery and instead allow planning of necessary operations.

The next mayor step in the treatment of cancers comes in the form of chemotherapy, when Ehrlich [135] published his work "Beiträge zur experimentellen Pathologie und Chemotherapy" where he injected animals with different toxins in order to destroy cancer cells. Although, it still took another

30 years till after the second world war when the discovery, that a chemical design for warfare also had potent anti-tumour effect.

In the meantime, the first long term tissue cultures of animal cancer cell lines were established and further insights like the Warburg effect [136] found, which showed, that cancer cells use glucose at a higher rate than healthy cells. This effect ultimately to the discovery of the positron emission tomography (PET) scan, which allowed a significantly more granular diagnosis and localisation of cancerous lesions than before.

With the success of growing human cell lines in vitro, the USA embarked on a massive experiment to test any potential source of chemical carcinogenesis. But at the same time, multiple viruses were identified to cause cancers in the 1950s, when electron microscopy was invented [137].

Only a few more years later, the biggest advance in the understanding of biology was made, when the structure of DNA was discovered [61] (Section 1.1)and subsequently lead to numerous new experiments and breakthroughs. When studying how viruses are able to reverse transcribe their RNA and insert a new gene into a healthy cell, which then transformed into a cancer cell, the term "oncogene" was coined[138, 139, 140] and the foundation for the understanding of how genes influence the emergence of cancers was laid. This also lead to the understanding, that heritable changes in the genome could predispose a person to cancer, which was previously hypothesised [141]. And while the discovery of DNA was a substantial boost for the understanding of cancer, the diagnostic capabilities increased at a similar speed, with urine tests for biomarkers of certain cancers as well as antigen detection.

And this is when we arrive at the "current" times, when a few years ago next generation sequencing (NGS) (Section 1.3) was introduced and sped up data generation on genomic and non-genomic diagnostic tests, from targeted amplicon sequencing to whole genome sequencing. These highly specific tests then allowed the application of highly specific drugs, like tyrosine kinase inhibitors (TKI)s, which are tailored to target a specific alteration in the genome of a cancer cell, and genetically engineered antibodies which can be homed in on the cancer. And while the therapeutic world is quickly evolving, many of the questions from previous times are still the same. We still don't know how and when the heterogeneity in cancers occurs, we just know it is a major source of resistance to treatment. We also still do not have an answer to the "cell of origin" question that has been asked for so long, but we do know that some cancers can de-differentiate and morph between cell types.

So instead of trying to answer these questions directly, there has been an effort to define fundamental features malignancies have to be considered cancers, very similar to the early pathology descriptions. The original characteristics comprise 1. Sustaining proliferative signalling 2. Evading growth suppressors 3. Activating invasion and metastasis 4. Enabling replicative immortality 5. Inducing angiogenesis 6. Resisting cell death (Figure 1.7).

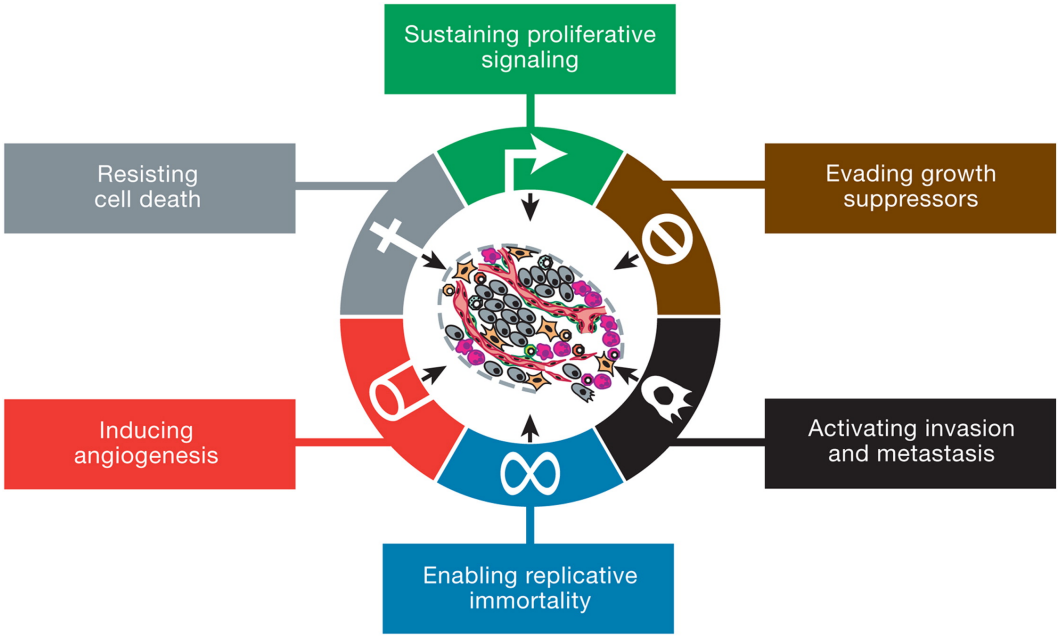

Figure 1.7: Acquired capabilities of cancer; Functional capabilities acquired by most cancers during their development; Figure adapted from Hanahan et al.[142]

These hallmarks were for a while considered the core of tumour development and the authors themselves hypothesised, that these core mechanics allow us to condense the complexity that cancer displays, both in the clinic and in labs, with a small set of rules, which all cancers have to obey [142]. In their exact words: "We foresee cancer research developing into a logical science, where the complexities of the disease, described in the laboratory and clinic, will become understandable in terms of a a few underlying principles"

However, with 11 years of additional research into the topic, more hallmarks have been found, and the original list was revised by the authors to contain additional characteristics, namely 1. Avoiding immune destruction 2. Tumour-promoting inflammation 3. Genome instability and mutation 4. Deregulating cellular energetics [143]. And even then a few years later, even more hallmarks e.g. metabolic rewiring are now considered a part of the characteristics of cancer [144].

And even during the time of my PhD, further research revealed additional hallmarks, which got

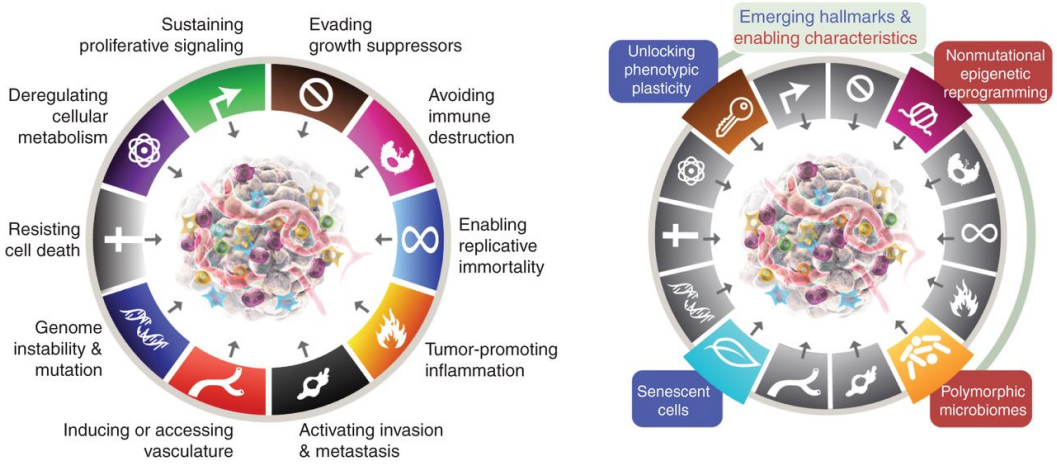

Figure 1.8: Emerging hallmarks and enabling characteristics of cancer; updated version of the hallmarks figure (Figure 1.7,[142]); Figure adapted from Hanahan[145]; Left, the Hallmarks of Cancer currently embody eight hallmark capabilities and two enabling characteristics. In addition to the six acquired capabilities – Hallmarks of Cancer – proposed in 2000 (Figure 1.7), the two provisional "emerging hallmarks" introduced in 2011 ([143]) –cellular energetics (now described more broadly as "reprogramming cellular metabolism") and "avoiding immune destruction" – have been sufficiently validated to be considered part of the core set. Given the growing appreciation that tumors can become sufficiently vascularized either by switching on angiogenesis or by co-opting normal tissue vessels [146], this hallmark is also more broadly defined as the capability to induce or otherwise access, principally by invasion and metastasis, vasculature that supports tumor growth. The 2011 sequel further incorporated "tumor-promoting inflammation" as a second enabling characteristic, complementing overarching "genome instability and mutation," which together were fundamentally involved in activating the eight hallmark (functional) capabilities necessary for tumor growth and progression. Right, this review incorporates additional proposed emerging hallmarks and enabling characteristics involving "unlocking phenotypic plasticity," "nonmutational epigenetic reprogramming," "polymorphic microbiomes," and "senescent cells."

characterised by Hanahan [145]. The newest version adds another two characteristics and hallmarks, specifically: 1. unlocking phenotypic plastisity 2. nonmutational epigenetic reprogramming 3. polymorphic microbiomes 4. senescent cells (see Figure 1.8).

These evolution of these hallmarks shows, why even though lots of time and effort was invested into cancer research for multiple centuries, there still is no unifying definition and treatment for cancer. The vast heterogeneity not only between cancer types, but also between patients makes it very hard to study. But even within patient there is third type of heterogeneity, which is the main cause of treatment resistance and relapse [1]. And while we know, that this diversity exists and efforts have been made to measure and classify them [147], there is still a lack of methods, which deal with the heterogeneity in their models to inform clinical approaches directly.

Write how this shows that the heterogeneity of cancer is the reason we still havent found a unifying description and treatment

1.6 Overview

add short description of each chapter

"It is the main source of our mistakes, when making making decision, that we only look at life piece by piece and not as a whole."

— Lucius Annaeus Seneca, *Epistulae morales ad Lucilium*

2

Joint somatic variant calling - if germline can do it, so can we

2.1 Introduction

When I started exploring the somatic variant calling methods in the beginning of my PhD in 2018, I was surprised about the stark difference between germline and somatic variant calling methods. Where all "modern" germline variant callers, like Strelka2 [106], HaplotypeCaller [148], DRAGEN [149] and DeepVariant [150], have the built-in capability to jointly call multiple related samples, for example from family trios, virtually no somatic variant caller had this functionality.

The joint analysis of smaller cohorts improves the performance of germline variant calling methods significantly, by allowing to assess technical artefacts, which might be unique for the individual sequencing machine or the researcher handling the DNA [151, 152]. Additionally, as certain parts of the genome are more problematic to sequence (Section 1.3) and map (Section 1.4.1), a "control" sample can help to distinguish if a certain observed change occurring frequently is a technical issue or in fact a real change.

For somatic variant calling, this concept has been adopted on in the genome analysis toolkit (GATK) [153] to allow the use of panel of normals (PON), which contains frequently seen changes in healthy ("normal") individuals analysed with the same sequencing technology [154]. Although, in contrast to the more intricate model for the germline equivalent, this is a post processing step of the analysis. Mutect2, which is the most recent somatic variant calling algorithm provided by the Broad institute, also provides a multi-sample mode, for which all tumour samples need to be from the same patient, either related longitudinally or spatially [155]. However, this mode is not very well publicised and

all tutorials released by the developers state that "there is currently no way to perform joint calling for somatic variant discovery" [109]. So while all methods in the GATK are considered a beta feature, the multi sample mode needs to be used with care.

There are only two methods currently, which have documented and published capabilities to jointly analyse tumour samples from the same patient to call somatic variants. The first one is a specialised method built on a joint bayesian model for SNVs that occur in multiple samples called multiSNV [156]. However, it has multiple shortcomings, which make it not usable for our data. First, as the name suggests, the method can only jointly evaluate SNVs and completely ignores INDELs and structural variants, which would be acceptable for the superior performance it provides. However, multiSNV was optimised only for WES and not for the very deep WGS that is now available and part of this thesis. This mismatch of input types means exceptionally high runtimes on our data. Even with custom parallelisation that was attempted in this work, the predicted runtime for just one multi sample patient would have been longer than 3 years. This shows, that while multiSNV was a great step forward at the time, there is a real need for new methods to stem the tide of sequencing data available due to the ever decreasing sequencing cost.

multiSNV has been the only software available for multi sample analysis for almost five years, but during this work, superFreq [157] was published. It combines all standard analysis steps for tumour analysis, like quality assessment, variant calling, copy number analysis and clonal deconvolution, into one program and is even able to jointly analyse samples. However, similar to multiSNV, its focus during optimisation and development was on WES and RNAseq data, so when applied to our data, we could not find a server node with enough memory to execute the workflow.

This then prompted us to investigate possible workflows to enable the analysis of high depth WGS, which we estimate to become more and more normal, with the ever dropping prices of sequencing. The following sections will show the development and validation of the joint variant calling methods as described in Hollizeck et al. [158] (Section 2.2), additional analysis on the impact of the joint variant calling on downstream analysis (Section 2.3), longitudinal analysis (Section 2.4) and clonal deconvolution (Section 2.4.1) and lastly information on the usage of the methods by others in the research community (Section 2.5).

2.2 Publication

The full publication about joint somatic variant calling can be found at https://doi.org/10.1093/bioinformatics/btab606 and non-journal formatted version is also attached as Appendix A with all supplementary methods.

References to supplementary data will be prefixed with the letter A in the text.

2.2.1 Summary

To enable highly sensitive, fast and accurate variant detection from multiple related tumour samples, we have developed joint variant calling extensions to two widely used single-sample algorithms, FreeBayes [104] and Strelka2 [106]. Using both simulated and clinical sequencing data, we show that these workflows are highly accurate and can detect variants at much lower variant allele frequencies than other commonly used methods.

2.2.2 FreeBayesSomatic workflow

The original FreeBayes algorithm can jointly evaluate multiple samples, but routinely it does not perform somatic variant calling on tumour-normal pairs. We introduce FreeBayesSomatic which allows concurrent analysis of multiple tumour samples by adapting concepts from SpeedSeq [159] which differentiates the likelihood of a variant between tumour and normal samples instead of imposing an absolute filter for all variants called in the normal. Hence, for each genotype (GT) at SNV sites, FreeBayesSomatic first calculates the difference in likelihoods (LOD) between the normal (Equation 2.1) and the tumour (Equation 2.2) samples genotype likelihoods (GL) with g_0 describing the reference genotype.

$$\mathrm{LOD}_{\mathrm{normal}} = \max_{g_i \in \mathrm{GT}} \left(\mathrm{GL}(g_0) - \mathrm{GL}(g_i)\right) \tag{2.1}$$

$$\mathrm{LOD}_{\mathrm{tumour}} = \min_{s \in \mathrm{Samples}} \left(\min_{g_i \in \mathrm{GT}} \left(\mathrm{GL}_s(g_i) - \mathrm{GL}_s(g_0)\right) \right) \tag{2.2}$$

$$\mathrm{somaticLOD} := \left(\mathrm{LOD}_{\mathrm{normal}} \geq 3.5 \wedge \mathrm{LOD}_{\mathrm{tumour}} \geq 3.5\right) \tag{2.3}$$

Next, the variant allele frequencies (VAF) in both the tumour and the normal samples are compared at each site.

$$\text{VAF}_{\text{tumour}} = \max_{s \in \text{Samples}} (\text{VAF}_s) \tag{2.4}$$

$$\begin{aligned} \text{somaticVAF} := (&\text{VAF}_{\text{normal}} \leq 0.001 \vee \\ &(\text{VAF}_{\text{tumour}} \geq 2.7 \cdot \text{VAF}_{\text{normal}})) \end{aligned} \tag{2.5}$$

A variant is classified as somatic when both somatic LOD as well as somatic VAF pass the criteria somaticLOD (Equation 2.3) and somaticVAF (Equation 2.5).

The thresholds chosen for both LOD and VAF calculations were previously fitted by the blue-collar bioinformatics workflow for the "DREAM synthetic 3" dataset using the SpeedSeq likelihood difference approach [160] and were selected to identify high confidence variants.

2.2.3 Strelka2Pass workflow

In contrast to FreeBayes, whilst Strelka2 has a multiple-sample mode for germline analysis and tumour-normal pair somatic variant calling capabilities, it cannot jointly analyse multiple related tumour samples. We enable this feature by adapting a two-pass strategy previously used for RNA-seq data [161]. First, somatic variants are called from each tumour-normal pair. All detected variants across the cohort are then used as input for the second pass of the analysis, where we re-iterate through each tumour-normal pair but assess allelic information for all input genomic sites.

The method re-evaluates the likelihood of each variant, by integrating every genotype from each tumour-normal pair. This step can "call" a variant (v) in a sample that initially did not present enough evidence to pass the Strelka2 internal filtering using two conditions: 1) if this variant was called as a proper "PASS" by Strelka2 in any other tumour sample, or 2) if the integrated evidence for this variant across all tumour-normal pairs reached a sufficiently high level. The second condition was based on the somatic evidence score (SomEVS) reported by Strelka2, which is the logarithm of the probability of the variant v being an artefact.

$$p_{error}(v) = 10^{\left(\frac{-\text{SomEVS}(v)}{10}\right)} \tag{2.6}$$

While the germline sample is shared between all processes, we can approximate these individual probabilities as being independent, since one variant calling process is agnostic of the other. Hence, we derive the following:

$$p_{error}(v_{s_1}, v_{s_2}, \ldots, v_{s_n}) = \prod_{s \in \text{Samples}} p_{error}(v_s) \tag{2.7}$$

And therefore:

$$\text{SomEVS}(v_{s_1}, v_{s_2}, \ldots, v_{s_n}) = \sum_{s \in \text{Samples}} \text{SomEVS}(v_s) \tag{2.8}$$

This allows the summation (Equation 2.8) of the SomEVS score across all supporting variants to assign a "PASS" filter, if it reached a joint SomEVS score threshold. This threshold can be set by the user and is 20 by default, which corresponds to an estimated error rate of 1%. These "recovered" variants need to pass a set of additional quality metrics related to depth of coverage, mapping quality and read position rank sum score.

As an additional improvement, we also built multiallelic support into Strelka2 which originally only reports the most prevalent variant at a specific site. Within the two-pass analysis, we reconstruct the available evidence for a multiallelic variant at a called site from the allele-specific read counts and report the minor allele at this site, if there is sufficient support from other samples. This method allows recovery of minor alleles only if another sample has this variant called by Strelka2, as SomEVS scores are not available for minor alleles.

2.2.4 Validation

While the development of new methods can challenge previous assumptions and allow to challenge previous ruled, all methods need be validated against the current gold standard methods in the field with data which allows objective measurements. For germline variant calling, there have

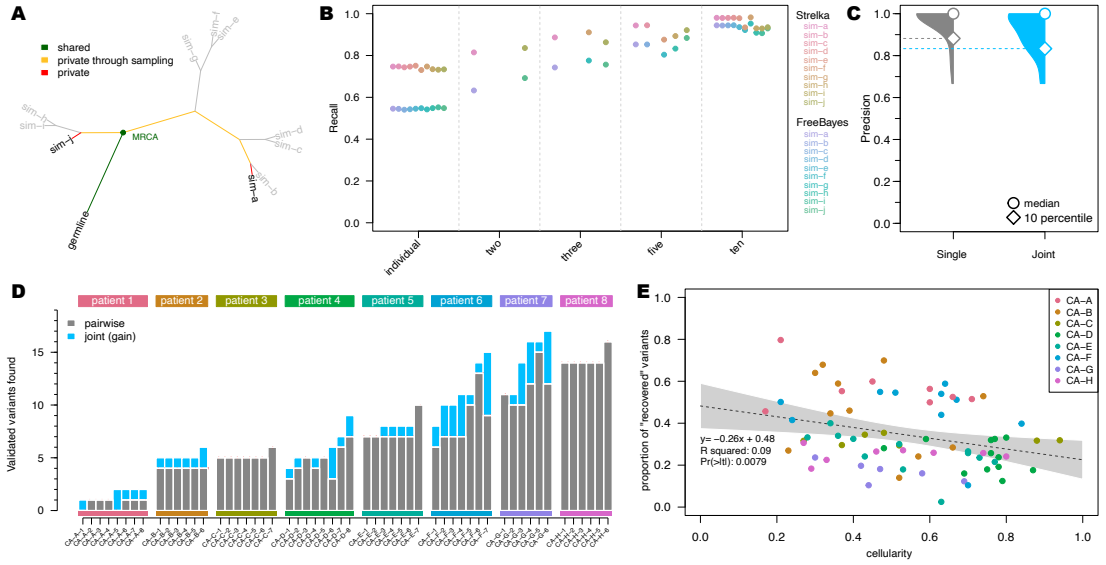

Figure 2.1: Comparison of joint multi-sample variant calling and single tumour-normal paired calling methods; A) Simulated phylogeny highlighting two samples with high evolutionary distance (sim-a and sim-j) where MRCA denotes the most recent common ancestor. B) Recall estimates of FreeBayes and Strelka2, run in individual tumour-normal paired and joint calling configurations using two (sim-a and sim-j), three (sim-a, sim-g and sim-j), five (sim-a, sim-c, sim-f, sim-h and sim-j) and all ten tumour samples. C) Precision of Strelka2 and D) Number of variants called by Strelka2 run in both tumour-normal paired (grey) and added with joint calling configurations (blue), which have been validated by targeted amplicon sequencing (TAS). E) Correlation between cellularity and proportion of variants found only with joint calling using Strelka2Pass for clinical samples; grey area shows the "95%" confidence interval for the linear model fit (dotted line).

been multiple community lead challenges and specifically designed test datasets, but there is currently no somatic variant calling equivalent. This issue is even more pronounced for our method, as we do not only need a tumour-normal pair, but we need the multiple tumour samples in the dataset to be related. To allow a fair comparison, I first generated a fully synthetic dataset, where every variant is known and fully defined (Section 2.2.4.1) to allow a general performance assessment of the methods. Then to ensure that these metrics also hold true in real world data, we then re-analysed previously published datasets which have orthogonal validation in the form of targeted amplicon sequencing (TAS) (Section 2.2.4.2).

2.2.4.1 Simulated data

We first simulated a phylogeny with somatic and germline variants from ten tumour samples and one normal (Figure 2.1A and Figure A.2A, B). Germline variants were simulated at a uniform allele frequency of 0.5. Somatic VAFs were sampled from a custom distribution, modelled to favour low allele frequency variants to closely represent real world data (min VAF: 0.001; max VAF: 1; Fig. S1C, D). Paired-end sequencing reads with realistic error profiles were simulated for WGS data at 160X

average coverage using the ART-MountRainier software [162]. The simulated reads were aligned to GRCh38 and both germline and somatic variants from the phylogeny were spiked into the aligned reads using Bamsurgeon [163]. We compared the workflows for FreeBayes and Strelka2 with and without our extensions for joint variant calling on the simulated datasets. The performance of Mutect2 joint variant calling was also assessed using its proposed best practice workflow. As both Mutect2 and FreeBayes do not return a verdict for each individual sample, we needed to assign each sample in the multi-sample VCF its own FILTER value. We called a somatic variant as present in a sample, if there were at least two reads supporting it for this sample and the overall FILTER showed a "PASS", which was the same cut-off used in the refiltering step in the Strelka2-pass workflow.

While the precision of each method without our extensions was greater than 99.8%, they all missed at least 25% of all variants in the samples (i.e. recall $\leq 75\%$). In contrast, the recall of the modified workflows increased to $\approx 95\%$ with only a minute decrease in the precision for both FreeBayes and Strelka2 (Figure A.3). Mutect2 had virtually no change in precision, but the recall actually decreased from $\approx 75\%$ to $\approx 41\%$ when analysing the samples jointly (Figure A.3B). Additionally, with our modified workflows, true positive variants were called with VAFs as low as 0.008 (median detected VAF ≥ 0.14 for joint sample analysis and ≥ 0.21 for single tumour-normal pair analysis), enabling improved distinction between true variants and technical errors (Figure A.4). This improvement in performance for Strelka2 is only achieved after the refiltering step and not just a result of the second pass (Figure A.5, Section A.5.4).

The performance of joint variant calling in Mutect2 was inferior compared to all other methods (Figure A.3A, B). This was primarily due to the "clustered_events" filter in Mutect2, which excluded the majority of false negative variants, with negligible contribution to the exclusion of true negative variants (Figure A.6A, B). This result was unexpected as the simulated variants were evenly distributed along the genome and the corresponding allele frequencies were sampled randomly (Figure A.2D).

Since the extent of the improvement in our joint calling workflows is bound by the number of shared variants between samples, we sub-sampled the simulated dataset, to show the effect of incomplete sampling on our methods, which is more likely in clinical settings. Furthermore, the evolutionary distance between the related samples in addition to the number of samples, has a major impact on the number of shared variants, as only variants acquired between the germline and the most recent common ancestor (MRCA), will benefit from the joint analysis. Therefore, we selected three sample subsets which included two, three and five samples with high evolutionary distance to show the

minimum expected improvement (Figure 2.1A, B). There was a clear linear improvement for both FreeBayesSomatic and Strelka2Pass when increasing the number of samples, even if they had a distant evolutionary relationship. In contrast, when using only two samples with a small evolutionary distance, the increase in performance was almost as large as when jointly analysing all 10 available samples. This shows that samples with a high number of shared variants will perform better in joint calling workflows (Figure A.7).

2.2.4.2 Clinical data

To validate the performance of our new workflows, we then analysed WGS and whole-exome sequencing (WES) data of multi-region tumour samples from eight patients, with multiple tumour sites (average 7 samples per patient; total number of samples 55), enrolled in a rapid autopsy program conducted at the Peter MacCallum Cancer Centre (Table A.1 and Section A.5.2) [164, 165]. The published studies had multiple somatic variants from the clinical samples orthogonally validated through targeted amplicon sequencing (TAS). We used these TAS-validated variants as the gold standard to evaluate the performance of different workflows, acknowledging that the technical biases inherent to TAS data are different to those present in WGS and WES (Figure A.8) and that there would be sampling biases depending on different tumour cells analysed in each data type.

In concordance with the results of the simulated data, our improved workflows found additional variants in all but one patient (Figure 2.1D, Figure A.9) (total additional variants Strelka2Pass: 64; FreeBayesSomatic: 85) with only a slight drop in precision for FreeBayesSomatic (mean: 0.94 vs. 0.88) and Strelka2Pass (mean: 0.97 vs. 0.92). Since the panel of variants validated by TAS was limited (7108 bp for patients CA-B through -H), this increase in detected variants suggests that a high number of shared variants in samples are missed with current approaches, which in turn leads to an overestimation of tumour heterogeneity between samples, as these variants are thought to not be present rather than undetected.

Even though the number of shared variants is a major influencing factor when jointly calling variants, low cellularity samples benefit more from the joint calling, as conventional methods cannot reliably distinguish low allele frequency variants from noise. Through a joint analysis approach, the number of recovered variants is higher in low cellularity samples, which indicates, that especially for clinical samples with variable tumour purity, joint analysis can have a major impact on improving performance (Figure 2.1E, Figure A.10).

Mutect2 in contrast, did not show significant improvement in any sample in its joint calling configuration, but showed inferior performance compared to the tumour-normal pairwise approach in two samples (Figure A.9E), similar to its decreased performance in the simulated data (Figure A.3). This was due to true variants being removed by the internal filters of the tool (Figure A.6C, D). This is in stark contrast to our novel workflows, where the joint analysis preserves all called sites from the pairwise method and finds additional variants. Overall, Mutect2 found less validated variants in all patients than both Strelka2Pass (mean: 2.2) and FreeBayesSomatic (mean: 2.5) with comparable levels of precision (Figure A.9, Figure A.11) but longer run times (Table A.2).

Our improved workflow also enabled the discovery of multiallelic variants with Strelka2, which led to the discovery of on average 42 additional variants (min: 1; max: 535) in the analysed WES and 987 additional variants in the WGS (min: 81; max 2329). These variants are strong indicators of sub clonal structure and are invaluable for the study of evolutionary trajectories in cancer, as shown in the following sections.

2.3 Effects of calling additional somatic variants on downstream analysis

The ability to find additional shared variants has significant impact on our understanding of cancer evolution and the timing of initiation and metastatic seeding. Recent work has shown, that similar to the well known genetic heterogeneity, there is heterogeneity when it comes to the timing of metastatic seeding. While traditionally it was thought that tumours only metastasise after they reach a certain size, to escape the restrictions of the niche, like reduced nutrition, recent publications showed, there is also very early metastatic seeding [166]. But all methods analysing heterogeneity, evolutionary timing and history are fully reliant on the somatic variants found in the data. Therefore, if we improve the input provided to these analysis methods, we can expect a clearer and possibly more granular result.

In the following sections, I will quantify the effect of using additional variants on phylogenetic reconstruction and clonal decomposition, which use somatic variants as input.

2.3.1 Phylogenetic reconstruction

As this work is not about the advantages and shortcomings of different phylogenetic reconstruction tools, I have not performed a comprehensive comparison of these tools, but rather focused on the results of using additional variants. For this reason, I chose to use neighbour joining (NJ) [167], because it is fast, readily available in most phylogenetic reconstruction tool kits and if the input distance is correct, the output will be correct. And even, if the distance is not 100% correct, if the distance is "nearly additive" and the input distances are not far off from the real distance, the tree topology will still be reconstructed correctly [168]. Lastly, in contrast to many other methods like UPGMA and WPGMA [169], NJ does not assume an equal mutation rate of each sample, because we know, that the molecular clock hypothesis [170] is not valid for different lineages of cancers [171].

The only thing that NJ requires as an input is a distance matrix of all samples, so the next step was the selection of the right distance metric. While there are many distance measures for DNA sequences, which allow accounting for different probabilities of transitions and transversions as well as uneven base composition, models like F81 [172] or HKY85 [173] are only really designed for germline mutations and are not easily applicable for subclonal somatic mutations, which is why I decided to first transform the variants present in all samples into a binary occurrence vector and then calculating the Hamming distance [174] between all samples. This generates a maximum parsimony approach and the branch length of the trees will be directly translatable to the amount of variants which are different between samples.

Figure 2.2 shows both the reconstructed phylogenies of the autopsy samples of the late stage melanoma patient "CA-F" from the manuscript (Appendix A, Table A.1), using the variants found with the default tumour-normal method on the left and our improved joint method on the right. The exact same reconstruction methodology was used otherwise, such that only the different inputs lead to the final difference.

Maybe adjust the font size in the trees to make it more readable

There are several obvious changes, first in the longer edge connecting the germline to all other samples, which we consider as the state of no somatic variants. This shows that there are many more shared mutations in all samples, than what would have been anticipated with the default method, which corresponds to an overestimation of the heterogeneity of the samples. As the accumulation of somatic variants is still used as a proxy for timing and cell divisions, when assuming a high mutation rate for lung cancer ($5.3 \cdot 10^{-8}$ from Werner et al. [175]) this difference of ≈ 36000 variants

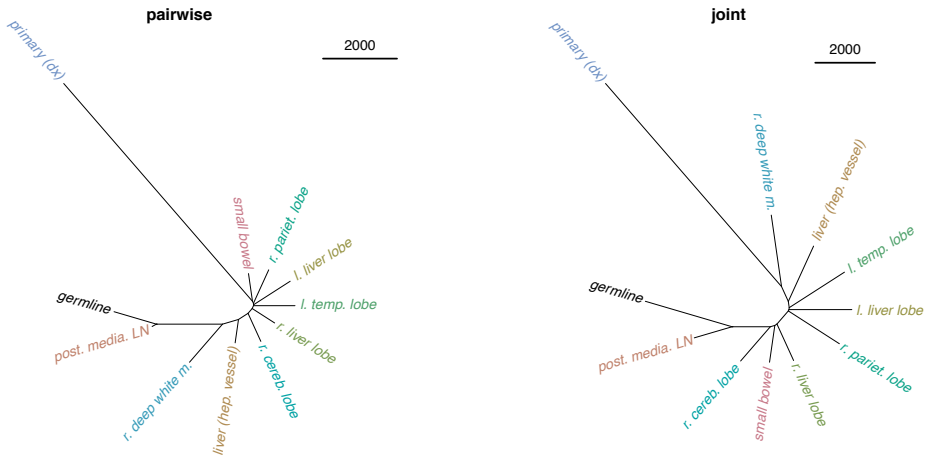

Figure 2.2: Reconstructed phylogenies of a patient with multiple spatially distinct samples; Neighbour joining on Hamming distance on variant occurrence vector. Tip labels describe the location of the sample in the patient. Trees are shown as unrooted with germline as fixated origin point; black line ruler shows the length of an edge with 2000 mutations

is equivalent to $\approx$ 2000 cell divisions. While the cell doubling rate of lung cancers is highly dependent on the type [176], this change makes a substantial difference when assessing the timing of the tumour initiation and further evolution.

Secondly, there have been topological changes, which generate a longer bifuricating edge between the olive coloured "r. liver lobe" and the "r. pariet. lobe" showing a bottle neck in cancer evolution, which fits very well with the clinical history, where the patient lived with stable disease for almost ten years, before progressing and dying. The extreme distance of the primary/diagnostic sample from the rest of the samples could be either a difference in sequencing quality, or due to the exposure to FFPE for the ten years between tumour diagnosis and death. However, as this feature is preserved between both the joint and the pairwise analysis, it does not appear to be an effect of our new method.

maybe increase the line width of the edges

Figure 2.3 shows a topology focused view of the two trees, which highlights the breaks which are needed to morph one tree into the other with dotted edges [177]. The common subtrees are coloured the same on both sides and connecting lines show identical labels. This format shows that while the trees look quite similar at first glance, they show vastly different topologies.

One example of this is "small bowel" which was connected to the red common subtree, but is now much closer to the "r. cereb. lobe" and forms a parallel clade with the "r.liver lobe". In general, where

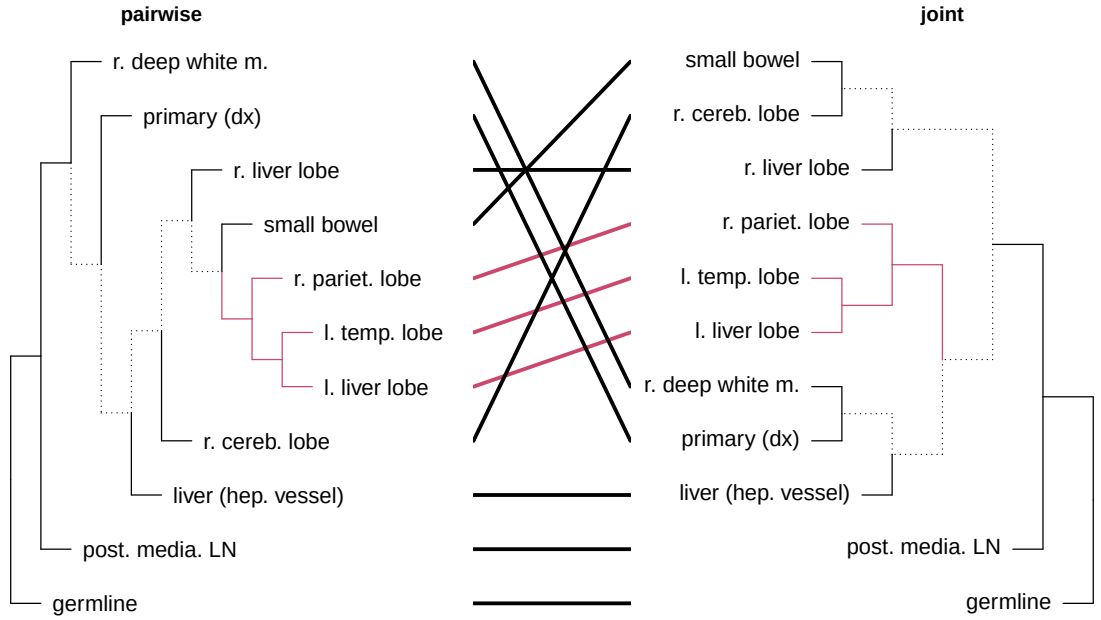

Figure 2.3: Side by side view of the reconstructed trees from Figure 2.2; internal edges, which are distinct between trees are shown as dotted lines; common subtree is shown in red Tree labels have been sorted to minimise distance between labels; Visualisation generated with dendextend [25]

the pairwise tree shows a very linear topology, which leaves only branching out of the main with no disjunct subclades, which are clearly present in the joint reconstruction. (Figure 2.3).

2.4 Longitudinal analysis

The initial motivation for the development of our workflows was the analysis of multi-region, or spatial, samples from the same patient coming from the CASCADE rapid autopsy program. However, we were very interested on applying the methods on longitudinal samples from patients, for example, for the joint analysis of diagnostic and relapse sample, or even the repeated testing of ctDNA are quite worth thinking about. In this part, I will present work using the published workflows on a longitudinal dataset, which highlights the flexibility and widespread usability of the new methods.

In addition to their autopsy which resected nine distinct metastatic sites (Figure A.13), Patient "CA-F" also had three longitudinal blood samples taken, from which ctDNA was extracted and WES performed. These blood samples were taken as non-invasive surveillance seven, five and two months before the death of the patient (Figure 2.4). In a study of late stage melanoma patients, Tan et al.

identified ctDNA sequencing as a way to stratify patients into high and low risk of relapse and therefore inform adjuvant therapy [178]. This makes patient "CA-F" a very good test dataset to showcase the improvement with joint variant calling. Similar to the spatially related samples, the joint analysis can improve the performance, which then in turn enable the detection of lower allele frequency variants, either through lower tumour burden or through the limited availability of DNA fragments from brain lesions due to the blood brain barrier [179].

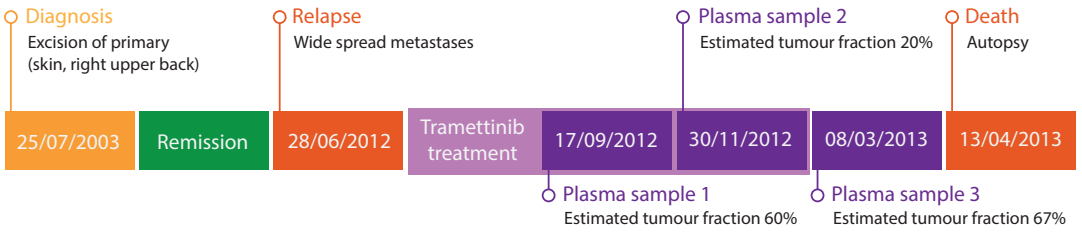

Figure 2.4: Timeline from diagnosis till death for patient CA-F: 1.9mm melanoma removed after diagnosis Friday 25th July, 2003 but with negative sentinel lymph node biopsy; Thursday 28th June, 2012: PET scan and subsequent liver biopsy confirm relapse with wide spread metastases; trametinib treatment from Oct. 2012 till Jan. 2013 with minor response; blood plasma samples during treatment (1 and 2) as well as after progression (3); death and rapid autopsy of nine metastatic sites (Saturday 13th April, 2013, Figure A.13); Tumour fraction in plasma samples was estimated via the original driver mutation (BRAF:K601E)

To show that even in longitudinal data, the joint analysis can boost the signal, we jointly variant called the diagnostic biopsy sample with the three ctDNA samples and compared them with the results from the pairwise analysis. On average, we found 2905 additional variants in each of the ctDNA samples, which is more than doubles the average number of variants found with the pairwise analysis (2414). Out of those, we found 534 variants in the ctDNA samples, which were found as a high confidence variant in the diagnostic sample, indicating that these findings are high quality calls.

Exactly like in the spatially different samples, in longitudinal data lower tumour purity samples benefit more from the joint analysis. We see that time point 2 (T2) has the highest amount of recovered variants (377) which are found as high confidence variants in both other time points (Figure 2.5 A vs. B vs. C) and T2 also has the lowest tumour purity in the cfDNA recorded (T1: 60%; T2: 20%; T3: 60%) however, there are still 106 variants, which were not found in the ctDNA samples at all with the pairwise analysis at all, even though they were high confidence variants in the primary sample (Figure 2.5F). These variants usually show a lower depth of coverage (dp) in the ctDNA samples, which may possibly indicate a problematic region in the genome, but rather than it not being called a variant, it is just a sign of incomplete data, which can be used with our joint approach.

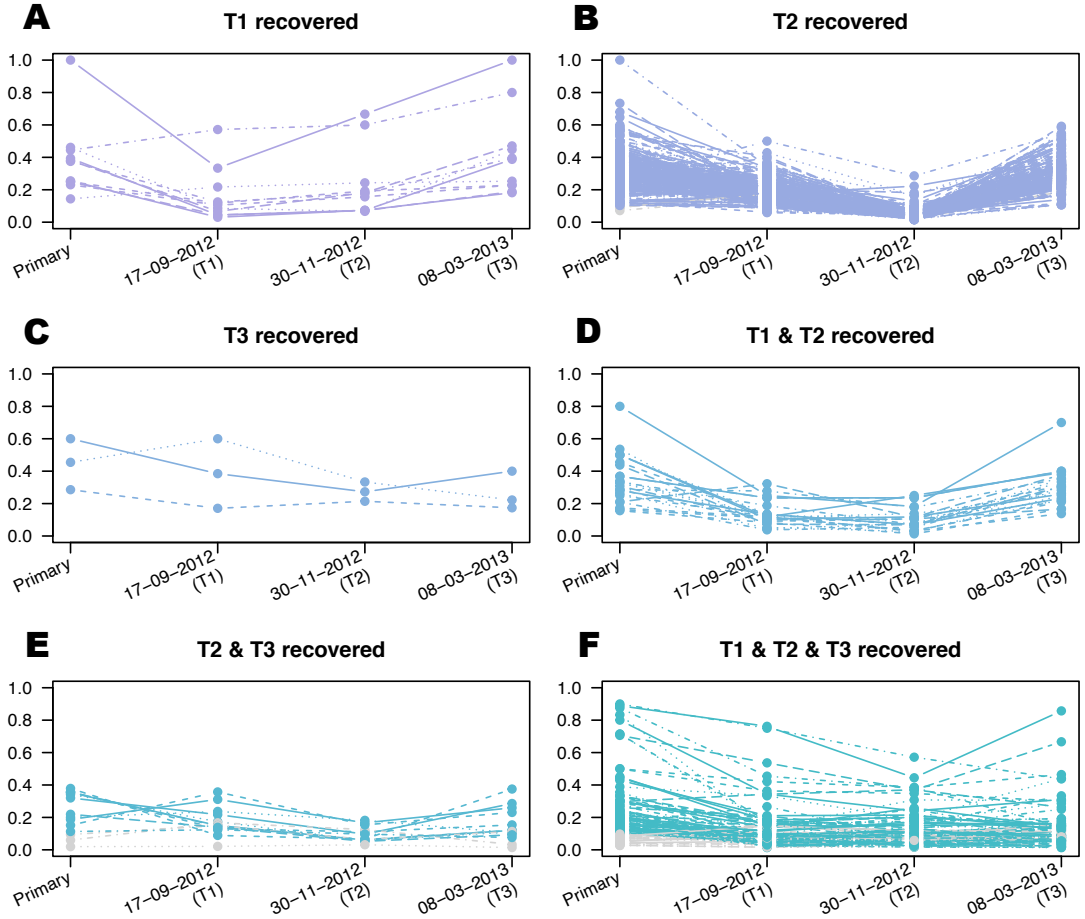

Figure 2.5: Improved somatic variant calling in longitudinal data: Variant allele frequency (VAF) of variants found additionally through joint variant calling which were found as high confidence variants in the primary sample; Variants with less than 0.1 VAF in the primary are coloured grey; "T1 recovered" shows variants, which were high confidence in all ctDNA samples but T1 and were only found through joint calling there; Axis label show the date of blood collection

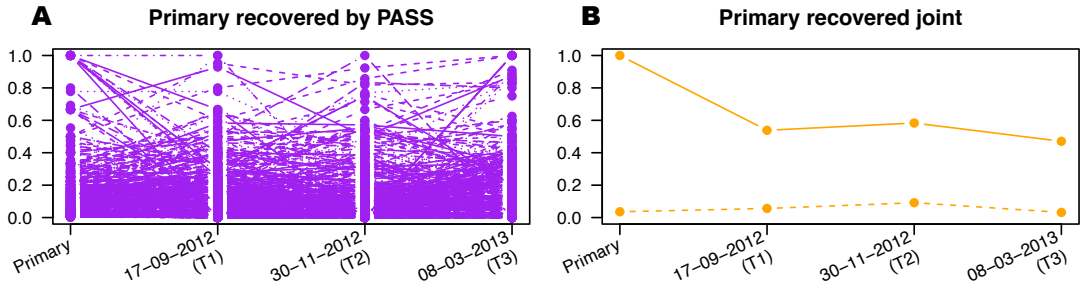

Figure 2.6: Longitudinal data informs diagnostic variant calling: Vafs of variants additionally found through joint calling in the primary samples; Primary recovered by PASS shows variants which were high confidence in at least one ctDNA sample; Primary recovered joint shows variant which were low confidence in all samples in the pairwise analysis; Axis label show the date of blood collection

Finally, we can also find 398 additional variants in the primary sample. 398 were discarded due to missing data in the tissue sequencing, but could be found with a high confidence in the longitudinal

data and two of the variants were included, as all 4 samples had this variant below the detection threshold (Figure 2.6). The missing depth in the primary also leads to the occasional very high allele frequency of the variant, as all available reads show the variant, but their numbers are below the threshold normal variant callers will report variants.

This shows that both spatially and longitudinal related samples should be analysed jointly, as it substantially increases the amount of true variants found, which as shown before, can have a large impact on downstream analysis of the samples.

2.4.1 Clonal deconvolution

On of the most important information derived from multiple related samples from the same patient is the clonal deconvolution, where subclonal reoccurring patterns of mutations (clones) are resolved both spatially and longitudinally. These reoccurring clones can be linked to either parallel evolution through positive selection pressure, like a targeted drug, or to the process of developing metastases where a piece of the cancer "breaks" off and grows at a different site. In contrast to the lack of options for joint somatic variant calling, there is a plethora of algorithms and methods available for clonal deconvolution. Since 2015 PhyloWGS [180], Canopy [181], CLOE [182], CloneFinder [183], MACHINA [184] and MOBSTER [185] were published, to name a few. Underlying all these models is a form of clustering variants with similar variant allele frequency together, to reduce the combinatorial space and enhance the confidence in the signal [186]. Due to the high number of tools, it is very challenging to select the right tool, especially since all of them have advantages and disadvantages [187]. In this work I decided to use PhylogicNDT [188] as it has been shown to work well on clinical samples [189] and does not have the restriction for the input to be from copy number neutral areas which many of the other tools have.

Both the variants found with the default pairwise as well as with the new joint workflows were annotated with their local allele specific copy number to form a MAF like file format which is required by PhylogicNDT. While PhylogicNDT allows the user to supply the cancer cell fraction for every variant, the program can also estimate them from the supplied allelic counts and the copy number. Local copynumber calls were derived from copy number segment calls made by sequenza by intersecting chromosomal location of each variant with the copy number segment containing the variants location. This requires multiple steps and the source code is shown in Listing A.1 (parsing VCF), Listing A.2 and Listing A.3 (convert to MAF format). Variants which couldnt be annotated

with copy number information, because their genomic location did not overlap with any called copy number segment, were discarded for this analysis.

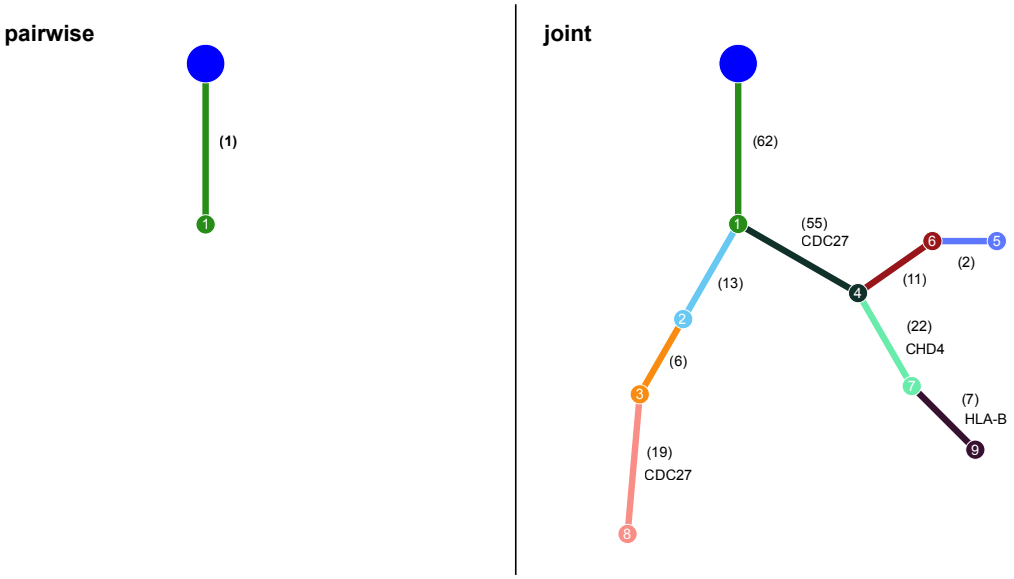

Figure 2.7: Reconstructed clonal trees from PhylogicNDT; Blue circle at top depicts the germline/normal state. The coloured edges with the same coloured circle represents a distinct subclone of the parent from which the edge emerges; The number in braces next to the edge is the number of mutations which define this subclone with an added gene symbol added, if there is a cancer driver gene mutation. The left part shows the result when using the default pairwise method of Strelka2 and the right side uses the results from the Strelka2Pass workflow

Figure 2.7 shows the highest parsimony clonal tree reconstructed by PhylogicNDT for the pairwise as well as the joint variant calling. As the copynumber calling information is the same for both inputs, the only difference is in the called variants. While there was no subclonal structure detected at all for the pairwise analysis, there is a highly variable structure detected using the jointly called variants. As this is a clinical sample, we cannot be certaine that the more branched model is the actual truth, but its biologically more logical that a late stage cancer has developed several subclones, rather than it being a very homogeneous disease at all of the 10 sites at autopsy with no evolution over ten years of disease [189]. It is of particular interest, that the *CDC27* gene got mutated at different time points in different clones (clone 8 vs. clone 4), which a clear sign of parallel evolution, which would definitely be missed without the joint analysis.

2.4.2 Longitudinal enriched phylogeny

Of course it is finally also possible to build a phylogeny with the spatial tissue samples and the longitudinal ctDNA samples. However, as the ctDNA give a holistic view of all cancer metastases

(Section 1.2) the interpretation needs to accommodate for that.

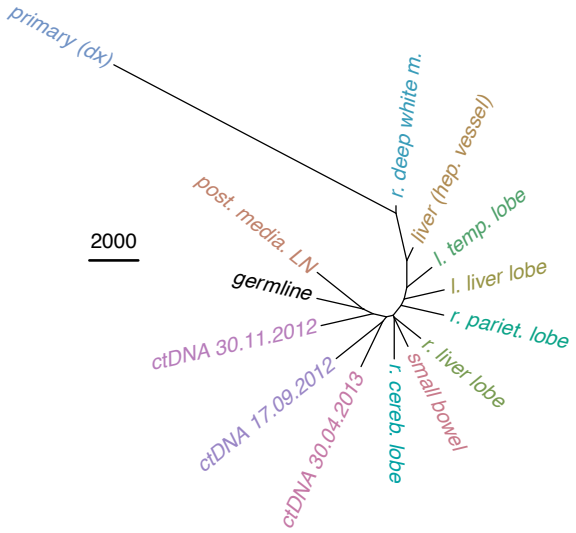

Figure 2.8: Reconstructed phylogeny with longitudinal ctDNA samples: Tree from Figure 2.2 with three additional ctDNA samples from different time points about one year prior to death. The ruler shows the equivalent of 2000 mutations

The maybe most surprising thing is that the more temporally distant ctDNA samples from 17.09.2012 and 30.04.2013 are in a subclade together, away from the "ctDNA 30.11.2012" sample. Secondly, the addition of the ctDNA samples also lead to a further bipartition edge, which separates "r. liver lobe", "small bowel" and "r. cereb. lobe" from the rest of the tree (Figure 2.8). This was already inferable from the topology of the previous tree in Figure 2.3 "joint", but is even more pronounced with the inclusion of the ctDNA samples.

This shows that the addition of more samples helps to refine and improve the trajectory and history of cancer samples and it is vital to do this analysis jointly to generate the optimal result.

2.5 Usage statistics and uptake

Ultimately when choosing research software, publication and citations are not a good metric to evaluate the quality of a method[**Gardner2022**]. Many published software packages are not maintained or not even functional even though they are published. While I developed these joint somatic variant calling workflows to deal with a challenge I faced, the interest of others was continuously expressed by bother members of the bioinformatics community whenever I presented this work.

To have some proxy of the usage statistics of the workflows, I recorded the download numbers of the "dawsontoolkit" docker container after the publication of the manuscript. The container only

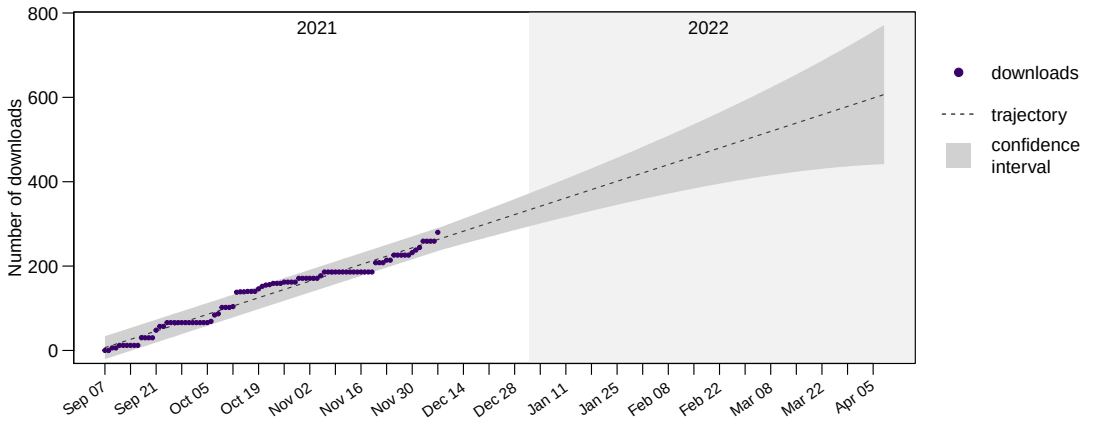

Figure 2.9: Cumulative download numbers of the "dawsontoolkit" docker container since publication of the manuscript; Actual counts are shown as dots, with smoothed trajectory depicted as dotted line with the 95% confidence interval shown as a grey background; confidence interval has been adjusted with exponential decay of prediction accuracy with distance from the last data point; Start date 7^{th} September 2021 (publication of method); End point prediction 8^{th} April 2022 (End of candidature); Numbers were recorded daily from the DockerHub API

consists of software for refiltering and joint analysis of the workflows . Obviously, this is an imperfect measurement, as people can reuse a downloaded container as often as they want, which would not appear in the count and similarly, just because the container was downloaded, the analysis might not have been used. Nevertheless, it still shows an interaction and an interest in the methods. The download numbers show a sustained and stable increase in downloads (Figure 2.9). This suggests, that there is a need in the methods, rather than a simple curiosity after publication, which hopefully will facilitate a higher quality analysis of future projects and therefore lead to a better understanding of cancer evolution and heterogeneity.

"Death is a release from and an end of all pains: beyond it our sufferings cannot extend: it restores us to the peaceful rest in which we lay before we were born"

— Lucius Annaeus Seneca, *De Consolatione ad Marciam*

3

CASCADE - Late stage lung cancer in the spotlight

3.1 Introduction

talk about cascade autopsies

3.1.1 Lungcancer

With around 1.6 million deaths world-wide each year, lung cancer is the number one cause of cancer death [190]. Every year about twelve thousand Australians get diagnosed with lung cancer. These cases can be generally split into two groups: small cell lung cancers (SCLC) and non-small cell lung cancers (NSCLC), which account for around 15% and 85% of cases, respectively. The majority of NSCLC are either lung adenocarcinoma or lung squamous cell carcinoma [191]. Even though smoking is highly associated with lung cancers, there is a big group of never smokers, with a high risk of lung cancers in East Asia, especially women, which is correlated with outside influences like pollution and occupational carcinogens and paired with genetic susceptibility [192]. This group usually shows *EGFR* (epidermal growth factor receptor) driven tumours. EGFR is a transmembrane receptor tyrosine kinase, which is usually only expressed in epithelial, mesenchymal, and neurogenic tissue, but its overexpression in other tissues is a hallmark of many human malignancies, not just NSCLC.

3.2 Publication

This chapter includes the data analysis for two two publications. The first publication features the resistance mechanism of small cell transformation (https://doi.org/10.1016/j.ccell.2019.08.008[193]) and the second shows the discovery of resistance to a targeted RET-fusion driven cancer (https://doi.org/10.1016/j.jtho.2020.01.006[164])

Cant include papers like this, will have to write the chapter as a whole

3.3 Cohort analysis

3.4 Mitochondrial phylogenetic reconstruction - the power house of the phylogenies

3.5 Outlook

"When the sum is already greater than the parts, there is room to make it greater still."

— Navali, Hatungo of the Karui

4

MisMatchFinder - hope springs eternal

4.1 Introduction

While even very early on, researchers realised, that cancers have different morphologies and clinical progression depending on the primary occurrence of the tumour (Section 1.5), with the extensive sequencing of cancer specimen over the last decade, the mutational signatures of cancers came into focus. These signatures are specific and characteristic combinations of mutations, which stem from distinct biological processes. These processes include exposure to DNA damaging agents like Chemotherapy treatment, tobacco and UV, as well as biological intrinsic pathways errors in DNA-replication or -repair. As each of those processes has a more or less distinct profile of mutations [194, 195] the analysis and deconvolution of the signatures involved in a patients mutational landscape can help diagnosing and treating a patient. While many signatures occur at a background level and are related to "normal" cellular processes like ageing [196], others can point to defective mismatch repair or gain of function mutations in specific pathways, which then lead to new avenues of therapy for the patient [197].

Supplementary information and plots for this chapter are attached in the appendix and prepended with B.

4.1.1 Mutational signature analysis

Traditionally the cancer mutational signature analysis entails a somatic variant calling process (Section 1.4.2) followed by a counting and deconstructing step, which assigns weights to the individual signatures. These signatures are precompiled list of mutation count relations (Figure 4.1). While

the individual SNP already contains valuable information, there is an improvement in granularity when also counting the base up and downstream of the change. This expands the feature space of counts from the six base classes of SNPs (C>A, C>T, C>G, T>C, T>A, and T>G)to 96 unique trinucleotide contexts [196]. While there technically are six more base changes and many more trinucleotide contexts combinatorially possible, they can be collapsed into the afore mentioned ones by using the reverse complement.

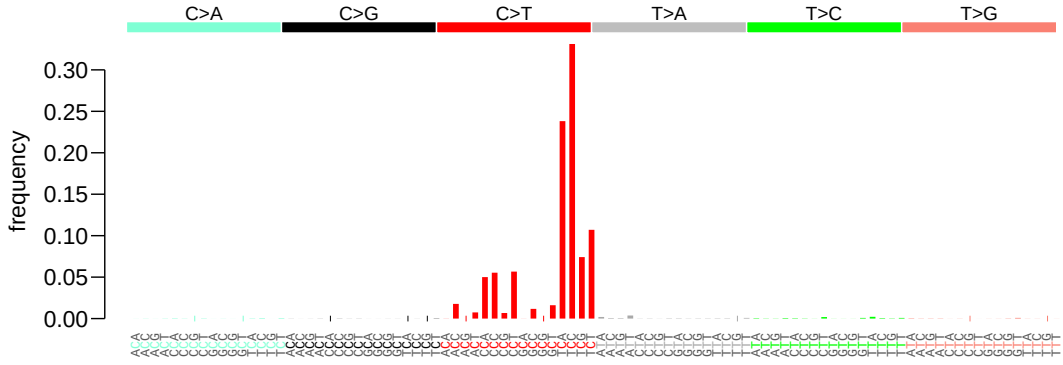

Figure 4.1: Trinculeotide count contributions for SBS signature 7a (UV exposure); values taken from Alexandrov et al. [198]

Additionally to the single base substitution (SBS) there now exist doublet base substitution signatures and InDel signatures for somatic mutations of cancers [198], which are all based on the same principle and enable a higher precision for stratification of similar cancer subtypes and DNA damaging agents.

4.1.2 Restrictions and pitfalls of standard signature analysis

Especially for cancer samples, the focus is usually on somatic variants of the sample. This requires a fairly deep sequencing of the tumour sample with at least WES, better WGS, and for optimal results, a germline sample for tumour-normal variant calling is required as well to not have noise from potentially missed germline variants (Section 1.4.4). This means the cost of the data of the analysis is surprisingly high for a fairly diffuse result of signatures that contribute to the variants found. This is especially relevant when it comes to clinical diagnostic tools, where every biopsy of the patient is valuable and a germline sample might not always be available. For an analysis, which is based on the averaged and aggregated somatic variants to require a high quality input seems counter-intuitive. Especially as the analysis will always report signatures, even if there are virtually

no variants reported, it will still suggest a mixture of potentially clinically relevant signatures, even if the patient is healthy.

4.1.3 Overview

These reasons make the actual analysis, which should come as a first step to inform both clinicians and pathology what to look out for more of a side arm, which is nice to have but not significant or informative. In this chapter I will describe a newly developed method, which allows the detection of somatic signatures from low coverage WGS of cfDNA. This allows the non-invasive monitoring of patients and possibly screening of at-risk individuals with very little cost.

4.2 Methods

With the change from a variant focused approach to a read based method, this new method will call "mismatches" of a read from the reference genome, rather than a variant. This has the advantage of not requiring a matched normal and its use for virtually any sequencing data source, be it TAS, WES, WGS or even nanopore sequencing[1]. However it also means, that the error suppression method, which are usually used by variant calling methods like read position ranks sum (RPRS) or strand bias are not usable, which leads to a higher degree of background noise. In the following sections I will describe how we filter and curate the found mismatches to retain as much signal as possible.

4.2.1 Mathematical concept

With the change from site based method, the concept of a mismatch from the reference needs to be introduced. A mismatch in the following is any position in an aligned read, which does not show the same base as the reference at the aligned position. The mismatch will inherit all the metrics of the read such as mapping quality, base quality and read position.

This then means, there are three sources of mismatches in a read, which are somatic variants, germline variants and sequencing errors (Equation 4.1).

$$n(mismatches) = n(somatic\,var.) + n(germline\,var.) + n(seq.\,error) \tag{4.1}$$

[1] however nanopore is not really usefull due to the short fragments naturally occurring in cfDNA

With the sequencing error being a function of the sequencing machine and chemistry, the error rate should be a stable almost constant, when using the same sequencing machine and chemistry [151, 152]. We can therefore reduce Equation 4.1 to

$$n(mismatches) = n(som.\ var.) + n(germ.\ var.) + c_{seq.\ err.} \tag{4.2}$$

Secondly, the number of germline variants is approximately the same between two people [199], which again simplifies Equation 4.2 by replacing $n(germline\ var.)$.

$$n(mismatches) = n(som.\ var.) + c_{germ.\ var.} + c_{seq.err.} \tag{4.3}$$

Of course, Equation 4.3 is a crude approximation and instead the constants are not real constants, but instead are better approximated with Gaussian distributions which leads to the following equation

$$n(mismatches) = n(som.\ var.) + \mathcal{N}(\mu_{germ.\ var.}, \sigma^2_{germ.\ var.}) + \mathcal{N}(\mu_{seq.\ err.}, \sigma^2_{seq.\ err.}) \tag{4.4}$$

However, both Equation 4.3 and 4.4 allow to make the conclusion, that with small enough values for either $c_{germ.\ var}/c_{seq.\ err.}$ or $\mu_{germ.\ var}/\mu_{seq.\ err.}$ and $\sigma_{germ.\ var}/\sigma_{seq.\ err.}$ respectively, there is a linear correlation between the amount of mismatches on a read and the somatic variants it contains:

$$n(mismatches) \sim n(som.\ var.) \tag{4.5}$$

With the help of Equation 4.5 we can approximate tumour mutational burden and signatures from individual reads. This method is therefore independent from read depth and requires no matched normal sample for somatic variant calling.

4.2.2 Data preprocessing

As this new method has sophisticated internal measures to filter and process sequencing data, the steps for preprocessing are minimal: The reads only need to be aligned to a reference genome (Section 1.4.1). For optimal mapping and additional noise reduction, paired end sequencing of at least 75 bp is suggested. This ensures a few bases overlap on the standard fragment length of less than 150bp of ctDNA (Section 1.2). Another optional suggested step is the duplication marking of the BAM file.

4.2.3 Mismatch detection

In contrast to conventional variant calling approaches, which find regions of interest through pile-ups (position wise) and then realign reads in the surrounding area, to accurately estimate the most likely event that lead to the observed haplotype (Section 1.4.2), with this new method, we take every individual read as a separate entity to fully span the heterogeneity of all cells and their genetic background. A sequencing reads “MD“- and “CIGAR“- tag from the preprocessed BAM file are used to reconstruct the sequence of the read and the positions, where the read shows a different base than the reference. These potential mismatch sites will then filtered in multiple steps to reduce the impact of both germline variants as well as sequencing errors

4.2.4 Filtering steps

Apart from the filters, which most variant callers will employ, like mapping quality (MQ) and base quality (BQ), which are used to ignore reads as well as positions respectively, the method also internally filters out common sequencing errors next to homopolymer regions [200]. While these cutoffs were preselected by me for optimal performance on our data (MQ=20, BQ=55, homopolyLength=5), the program allows the user to adjust them to their liking. This is also possible for both the region of interest (ROI) bed-file which was used to restrict the analysis to only highly mappable regions of the genome (Section B.1), as well as for multiple other parameters which are unique to our method, like minimum average base quality, minimum and maximum number of mismatches per read and/or

fragment, and the minimum and maximum length of a fragment [201]. If any of these values are not within the specified range a read will be discarded in the analysis. This is also the default for reads which have a secondary alignment position or are considered duplicates of any kind.

4.2.5 Consensus reads - what happens when the sequencer isn't sure

When paired end sequencing of ctDNA is analysed, the fraction of fragments where reads overlap is higher, than with "normal" tissue based sequencing, due to the shorter fragment length of ctDNA (Section 1.2). This allows an fragment internal consensus generation, by adjusting for differences between forward and reverse read. In many variant calling methods, these differences are used by measuring the "strand bias" [202, 203, 204] or "strand balance probability" [104] by looking at a specific locus and evaluating the discrepancy of all forward and all reverse reads. As our method evaluates each read/fragment independently, the bias cannot be calculated, however in the overlapping region of both reads, a consensus can be generated. If both reads agree on the mismatch, the BQ of both reads will be added together to emphasise the increased evidence for this variants. However, if they disagree the base of the higher quality will be used and its quality will be decreased by half of the BQ of the lower quality base (Figure 4.2 bottom). To increase the stringency of the method, the user can also enable the '*–strictOverlap*' option, which will only consider a mismatch, if both reads agree with each other and decrease the BQ to zero otherwise. As we are only interested in mismatches from the reference, all positions where both agree with the reference are irrelevant for the analysis and will be discarded (Figure 4.2 top). For the most stringent analysis, MisMatchFinder can additionally be configured to only use mismatches in the overlap part of a fragment ('*–onlyOverlap*'), which significantly reduces the number of sequencing errors which end up in the final analysis (Section 4.3.1.1).

This method however also reduces the available data by restricting the analysis to areas where reads are overlapping. Due to the fragment size distribution of ctDNA a paired end sequencing with 100bp read length will in most cases lead to an overlap of at least 45 nucleotides (Section 1.2) and with 150bp most ctDNA fragments will be almost entirely covered by both reads in theory. However due to soft-clipping and incomplete alignment, this number will be lower in reality. In our tests, the restriction leads to on average 18 nucleotides (min: 14bp max: 25bp std.dev.: 1.45) being retained in the analysis from a read on average for a 100bp read in real world data. This however means that with a read depth of 8-10x ≈80% of the genome will be covered by the overlap of at least one read pair.

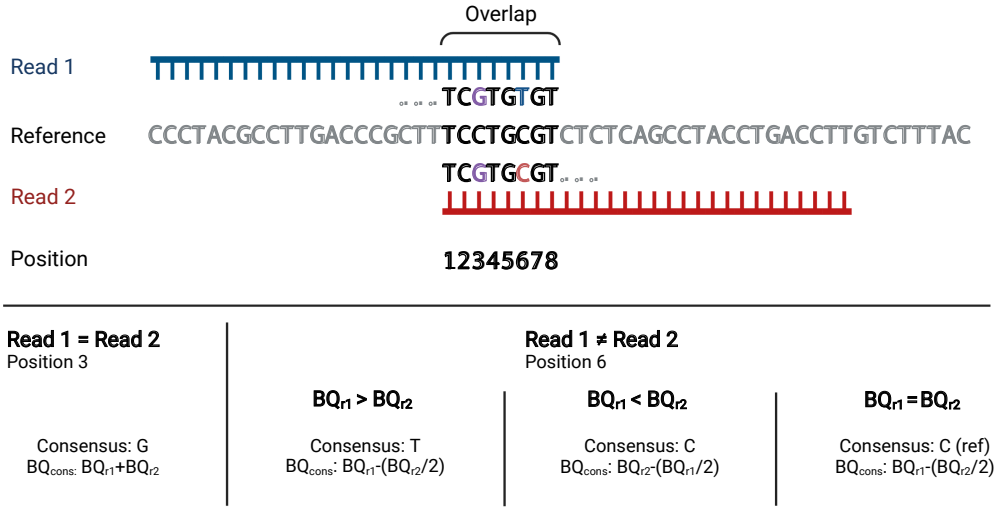

Figure 4.2: Schematic of consensus computation method for overlapping reads in MismatchFinder; Read 1 and Read 2 depict two overlapping paired end reads aligned to the reference sequence; Positions in the overlap are numbered for later referral; Read positions agreeing with the reference are coloured black, positions differing from the reference but agreeing in both reads are coloured purple (position 3) and differences between reads are coloured in the respective read colours (blue and red, position 6); Calculation for the resulting base quality (BQ_{cons} for each possibility is shown as formulas)

4.2.6 Germline filtering - exclusion of normal variation

To further enable the Section 4.2.1 claim that the germline is a very small constant, we need to remove as many mismatches as possible, which stem from germline variants. For this purpose, I built a zarr [56] based storage system from the gnomAD database (v.3.1) [111] using scikit-allel [205]. A in-depth explanation of the generation as well as a script for for an end user can be found in Section B.3.

This then allows very precise filtering of known germline variants sites from the analysis. The method allows the specification of an allele frequency to consider a variant to be filtered, however as baseline, it will filter all sites, which were detected in any sample in gnomAD. This even includes sites with low quality variants, as these are signs for sequencing or mapping complications, which will most likely interfere with our method as well.

4.2.7 Count normalisation - not everyone has the same chances

Finally when having filtered all “noise“ mismatches from the dataset, we can aggregate all mismatches to oligo-nucleotide counts. With this step also comes the classification of directly neighbouring mismatches as DBS, which are counted as separate entities. SBS and DBS both can be used

to identify underlying biological mutational processes, but they have very different signatures associated with them [198]. The counts formed this way are influenced by the background frequency of their reference oligo-nucleotides in the analysed genomic region. As the frequencies of di- and tri-nucleotides are not uniform in the genome, the chance for a mismatch found in an "AAA" reference contex is almost seven times higher than a mismatch in "CGC" (Table B.2, Table B.1). To reduce this bias towards high frequency oligo-nucleotides, I implemented a count normalisation step.

First the di- and tri-nucleotides in the analysed regions are counted using the supplied reference without any black-listed and/or only in white-listed regions. These counts are then either used to directly weight the observed mismatch counts, which leads to a more uniform distribution of mismatches, or by building a fraction of observed oligo-nucleotides and the total counts in the genome (Table B.2, Table B.1), the weighting achieves an approximation of how the counts would be distributed over the whole genome. These two options are available with '*–normaliseCounts*' for the approximation to full genome. By also adding '*–flatNormalisation*' only the observed counts are used for normalisation.

4.2.8 Signature deconvolution - find the original signal

The deconvolution of the involved signatures from known set of signatures is equivalent to finding the minimal distance between m as the observed number of mismatches in each oligo-nucleotide context (a vector of length 96) and $\mathbf{S}w$, where $\mathbf{S}$ is the matrix of oligo-nucleotide defined contributions for each signature, resulting in a matrix of $96 \times k$ with k being the number of known signatures. Lastly, w is the vector of weights of each signature, which we want to estimate.

$$\text{minimise:} \quad (m - \mathbf{S}w)^T(m - \mathbf{S}w) = m^T m - w^T\mathbf{S}^T m - m^T\mathbf{S}w + w^T\mathbf{S}^T\mathbf{S}w \tag{4.6}$$

$$\text{with:} \quad \sum_j w_j = 1 \quad \text{and} \quad \forall j \; w_j \geq 0 \tag{4.7}$$

Equation 4.6 can then be written as

$$\text{minimise:} \quad -m^T\mathbf{S}w + \frac{1}{2}w^T\mathbf{S}^T\mathbf{S}w \tag{4.8}$$

With the same restrictions as shown in Equation 4.7. These equations and the idea to solve it with quadratic programming (QP) have been taken from Lynch [206], the iterative linear models (ILM) solving approach was adapted from deconstructSigs [35]. Both methods are reimplemented in python in MisMatchFinder, using the quadprog package [59] for QP and a translation of the R code of deconstructSigs for ILM.

MisMatchFinder allows the use of either QP or ILM, as they in many cases produce very similar results [206]. The default method is QP however, even though ILM is the more interpretable method and is the more parsimonious method, because with the increased number of signatures, in the latest work by Alexandrov et al. [198], ILM does not lead to the right signatures if the signal is not strong enough but QP seems to be more stable (Figure 4.3).

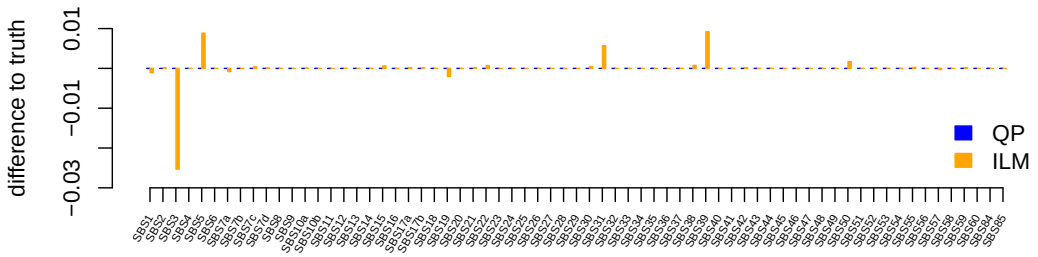

Figure 4.3: Distances of the estimated weights generated with ILM and QP from the true weight used as input; Truth is a synthetic count sample with (SBS1: 0.25; SBS3: 0.05; SBS5: 0.46; SBS7a: 0.1; SBS19: 0.03; SBS21: 0.01; SBS31: 0.08; SBS57: 0.02;)

The combinatorial problem in ILM, already shown by Lynch [206], seems to be especially strong with “wide“ signatures like SBS3 (Figure 4.1) and low signature contribution. That makes it less useful for our approach, as we expect low tumour purity and therefore low somatic signature signals in cfDNA, but even with ILM as deconvolution method impacts the detection of SBS7a less than the detection of SBS3. especially for SBS3 the weight for ILM will only be assigned with sufficient signal (15 and 20 mutations per megabase respectivly for for SBS7a and SBS3) where QP allows a more linear increase in signal, even at lower levels. In contrast ILM will assign more weight overall than QP once the signal reaches a certain threshold (Figure 4.6). This means, that ILM will be better for high powered signal, but less effective for the more subtle differences we expect from ctDNA.

The deconvolution method might be a spot for further optimisation by creating a custom deconvolution system adjusted for ctDNA detection.

For the rest of this chapter, unless specified differently, the results shown will use the QP deconstruction method.

maybe move the simulation description here

4.3 Results

In this section I will present the results when applying MisMatchFinder to various different datasets. First the evaluation of the method on simulated data, which allows accurate and definitive insight into the sensitivity and a proof of concept (Section 4.3.1). Then the analysis of real world applications is displayed in Section 4.3.2 demonstrating that the method does not only work in cleanly simulated data, but find clinically relevant insight in patient samples.

4.3.1 Simulated Data - the validation promised land

Just like in chapter 2, the novelty of the approach leads to the issue of no gold standard dataset, with which to evaluate the performance of a new method. While there are low coverage WGS datasets of cancer patients, none of them have validated signatures associated with them. So again simulated data is the optimal starting point to allow both optimisation of parameters as well as granular detection of artefacts which can originate from any step starting from sequencing over mapping to the signature deconvolution.

4.3.1.1 Sequencing errors - there is always a cleaner data

To judge the ability of our approach to filter out sequencing errors, we first simulated "clean" sequencing reads with neither germline or somatic variants with the ART simulation suite [162]. As current estimates of Illumina sequencing is in the range of 1 in 666 to 1 in 1149 [152] which is significantly higher than even the highest tumour mutational burdens of cancers (melanoma: 1 in 5k; tobacco smoking lung cancer: 1 in 100k) it is very important to be able to eliminate as much of the background noise of sequencing errors as possible.

By only using high base quality mismatches, where both reads agree on the mismatch 99.98% of all sequencing errors can be eliminated and only 1 in 10M bases will be wrongly counted as a variant (Figure 4.4). This false discovery rate is multiple orders of magnitude lower than before and in a similar range to normal mutationally driven cancers tumour mutational burden [198, 207].

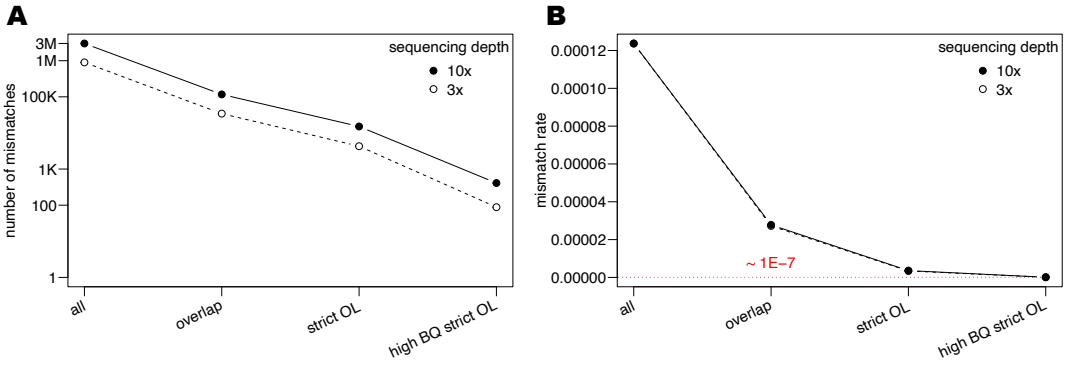

Figure 4.4: Mismatchrate of different filtering methods on sequencing data simulated with ART[162] for both 10x and 3x coverage; Mismatches correspond to simulated sequencing errors; all: no filters, overlap: only use the overlapping parts of paired end reads with consensus building (Section 4.2.5), strict OL: overlap but reads *must* agree, high BQ strict OL: strict OL with high BQ in both variants; A) Absolute counts B) counts from A normalised by the number of analysed bases all: all aligned bases, other: number of bases in read overlap

4.3.1.2 Spike-in signature detection

With the technical error eliminated in simulated data, the question was would our method work in a real world data, however to establish a baseline for detection limit and sensitivity of the method, we decided to first use a hybrid approach, were we spike-in somatic variants into a genuine low coverage WGS sequencing of a healthy control. That reduces the amount of unknown variability from other published datasets.

While it would be possible simulate the variants completely de novo, without any prior knowledge, we know that somatic mutations follow a certain pattern and there a mutational hotspots [208, 79], so we decided to instead use the COSMIC database [209, 210] as the to select mutations from. This allows us to select mutations, which definitely occurred in a specific cancer subtype, which leads to a simulation which closer resembles real data. The in-depth protocol is shown in Section B.3.4. The downside of this method is that the spike-in will not predominantly happen on shorter fragments, as it would be the case with ctDNA.

In the following section I will discuss the results for the simulation of the very prominent SBS7a signature (see Figure 4.1)which is predominantly present in Melanoma (see Section 4.3.1.2.1) and secondly the much flatter and more uniform SBS3 (see Figure B.1), which is a sign of defective homologous recombination in breast cancers (Section 4.3.1.2.2).

The spike-in was done at multiple different ratios, to simulate varying tumour purity and tumour mutational burden (TMB). Figure 4.5 shows the signature analysis result of the lowest spike-in ratio "r0.1" which corresponds to 0.1 somatic variants per mega base and results in approximately 300

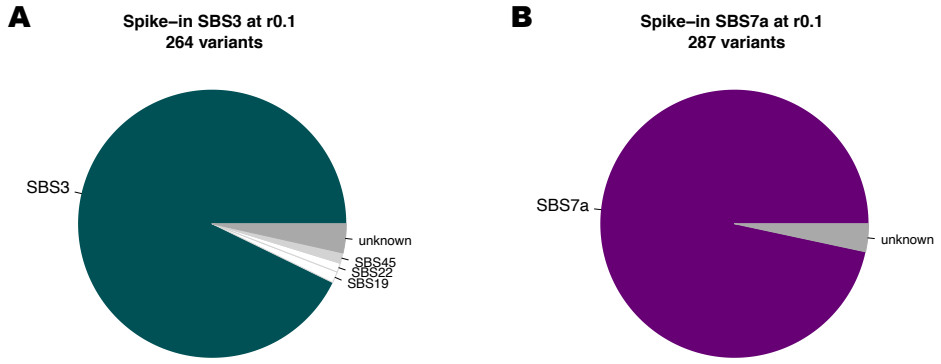

Figure 4.5: Signature analysis results of spiked-in somatic variants; signatures with a weight less than 1% were collated into "unknown"; The original spike-in signature was coloured in green (SBS3) and purple (SBS7a), unrelated signatures are coloured white and signatures corresponding to sequencing artefacts are coloured in lightgrey; r0.1 corresponds to approximately 0.1 variants per mega base; Weights were generated with deconstructSigs [35]

variants for the whole genome. As the spike-in process has to satisfy certain quality measures, not all candidate variants can be used. As such, the final simulated BAM contains 264 additional variants for the SBS3 simulation and 287 for the SBS7a equivalent. That corresponds to 304 and 364 "tumour" reads respectively within the $\approx$ 261 million reads of the simulated BAM. With increasing ratio, the spike-in signatures show decreasing weights for other signatures, which likely got introduced due to the incomplete spike-in (Section B.3.4).

4.3.1.2.1 Melanoma - UV exposure (SBS7a)

With melanoma, TMB ranges from 0.1 to 100 mutations per mega base [198], however Melanoma is usually seen as a very cancer with very high mutational load, which makes it the ideal target for this new mutational based tool. With only the strict overlap (Section 4.2.5) and the germline (Section 4.3.1.3) filtering enabled, we can see that already from r5, which represents 16899 mutated reads (of 260 Mio.), we see an increase in signature SBS7a. While this signal is likely too low to trust this in a real world setting, with r10, the weight is already 2% and well established. Secondly, the method is very specific on this dataset, where only SBS7a shows an increase with higher spike-in, which minor leaks to other heavily $C > T$ signatures like SBS2 and SBS30 (Figure 4.6,Figure 4.7), which partly already stem from the spike-in process, which slightly enriched for SBS2 (Figure 4.5B "unknown"). All other signatures, which are present in the normal show a decrease, which is to be expected, as all signature weights need to sum up to one.

4.3.1.2.2 BRCA1/2 - Defective homologous recombination-based DNA damage repair (SBS3)

Just as with the SBS7a signatures, even for the much less specific signature SBS3, MisMatchFinder

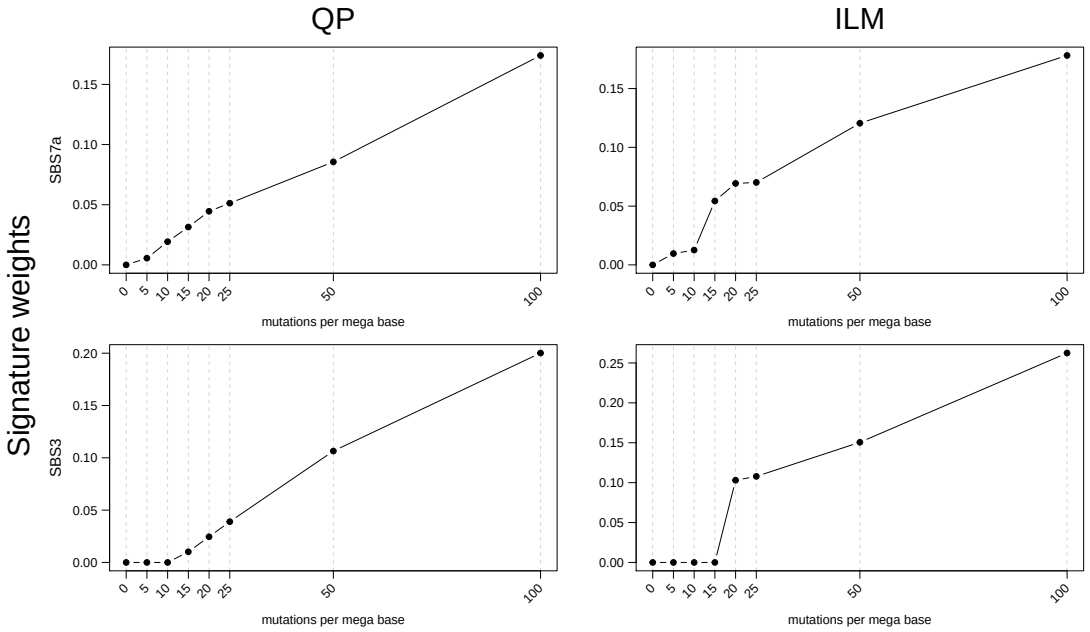

Figure 4.6: Signature weight differences for different deconvolution methods; Methods are the quadratic programming (QP) and iterative linea model (ILM); deconvolution was performed on the same counts generated with MisMatchFinder on 7 simulated dataset with increasing mutational burden from 5 to 100 mutations per mega base spike-in; for 0 mutations per mega base, the normal sample used for the spike-in was used

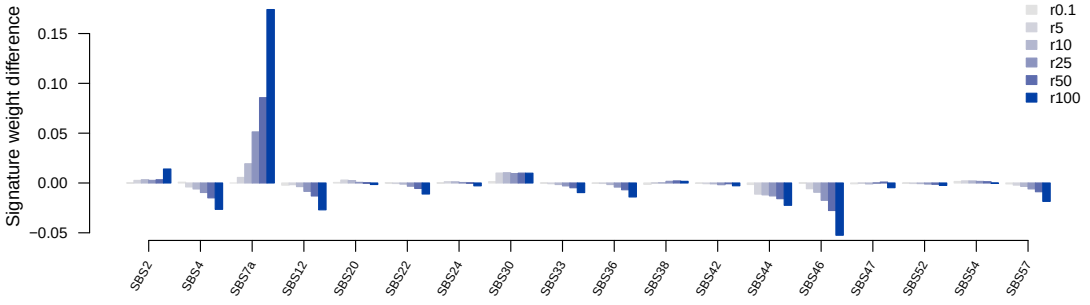

Figure 4.7: Signature weights differences from normal for SBS7a spike-in; Weights were deconstructed with QP method in MisMatchFinder and the weights assigned to the normal sample used for the spike-in were subtracted; Only Signatures with original weight $\geq 1\%$ or a minimum difference of 0.5% are shown. The full weights can be seen in Figure B.2; r0.1 corresponds to 0.1 mutations per mega base (287 variants) and r100 is the equivalent of 100 mutations per mega base (286974 variants)

specifically picks out the spike-in signature and does not assign it to any other signature. There is a small increase in SBS4 for the very low mutation rate simulations, where no SBS3 is detected. Not surprisingly, the detection limit for SBS3 is slightly higher than for SBS7a (5 vs. 15 mutations per mega base), because it is a more uniform signal. Exactly as with SBS7a, all other signatures show a slight decrease, to accommodate the additional signature weight which sums to one (Figure 4.6, Figure 4.8). While this is slightly higher than the currently assumed median TMB in breast cancer, especially triple negative breast cancer (TNBC) has shown a higher TMB, which is at comparable

levels to the limit of detection we see in this simulated dataset [211].

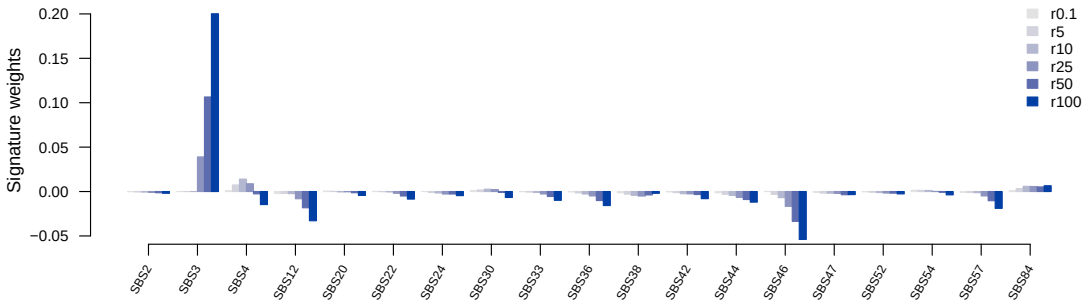

Figure 4.8: Signature weights differences from normal for SBS3 spike-in; Weights were deconstructed with QP method in MisMatchFinder and the weights assigned to the normal sample used for the spike-in were substracted; Only Signatures with original weight $\geq 1\%$ or a minimum difference of 0.5% are shown. The full weights can be seen in Figure B.3; r0.1 corresponds to 0.1 mutations per megabase (264 variants) and r100 is the equivalent of 100 mutations per megabase (285367 variants)

4.3.1.3 Germline filtering

With real patient data, we can evaluate the effect of removing germline variants from the analysis is. For this I used the same simulated samples from above (Section 4.3.1.2), where the reads are original ctDNA sequencing reads from a healthy person. These reads will have a natural background germline variant signature.

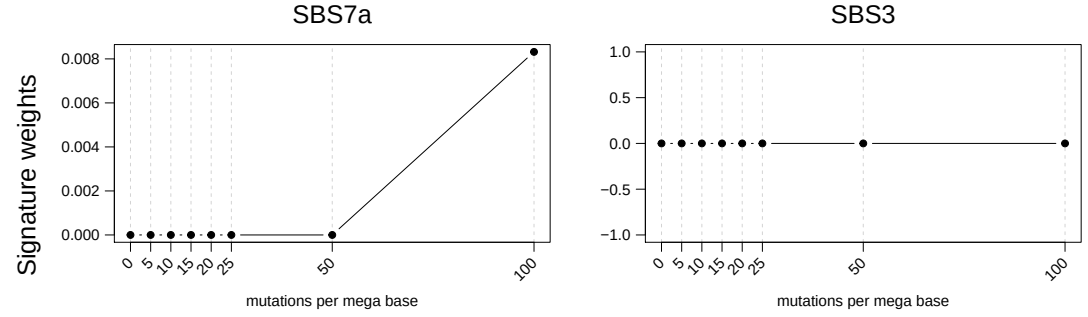

Figure 4.9: Signature analysis without germline variant filtering; Weights were deconstructed with QP method in MisMatchFinder, but in contrast to Figure 4.6, the filter removing all known germline variants was disabled

In stark contrast to the previous analysis (Figure 4.6), when retaining mismatches in known germline variant sites, the sensitivity of the method reduces significantly. Only for the SBS7a spike-in at the very highest mutation frequency (100 mutations per mega base) a signal is detected. This signal is still weaker than what was previously found with just 10 mutations per mega base. Unsurprisingly SBS3 performs worse, just as before, and no signal is detected at all (Figure 4.9).

This extreme change is caused by the much higher number of mismatches which were used in the analysis ($\approx$ 1.8 Mio without germline filter and $\approx$ 130K with germline filter). This increase in mismatches in the analysis dilutes the spike-in variants. Figure 4.10 shows that without the germline filter the additional found mismatches never exceeds 5% which seems to be the detection threshold for SBS7a, as with germline filtering this threshold corresponds nicely with the arise of SBS7a weight in those samples (Figure 4.6).

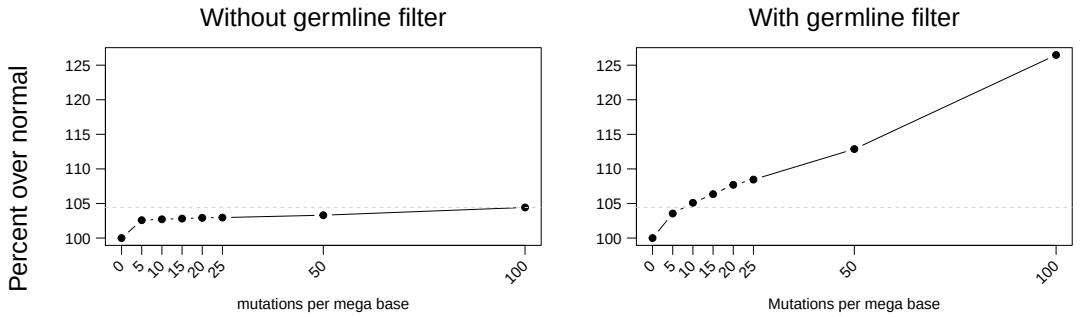

Figure 4.10: Percent increase of mismatches in analysis with and without germline filter; Values are normalised to the number of mismatches found in the normal sample (depicted as 0 mutations per mega base); dotted grey line shows the maximum increase in the left panel (without germline filter)

While we already know that the spike-in variants were not detected when retaining germline variant sites, the signatures detected in the normal sample are vastly different as well. Without the germline filter, the most prevalent signatures are SBS1 and SBS5 which are thought to be molecular clock like signatures, related to the age of the individual [198]. In the germline filtered analysis the most prevalent signatures are SBS4 (tabacco smoking), SBS12 (unkown) and SBS46 (sequencing artefact). In general it seems like the germline filter removes predominantly SBS1 and SBS5, while most other signatures remain the same (Figure 4.11).

As the sample was acquired through a healthy donor blood bank, we have no way to verify if the sample was or is a smoker.

This shows that germline filtering similar to the consensus overlap analysis is fundamentally important for the method to recover signal. In the following sections, unless further specified, the germline filter will be enabled for all analysis.

4.3.2 Real world data - the only things that matters

While simulated data is perfect to ensure the method performs as expected in edge cases and to estimate detection limits, only real world data allows the final examination if the model used for

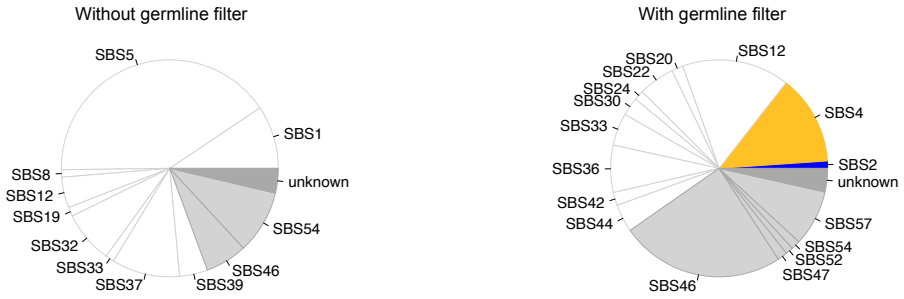

Figure 4.11: Signature weights of the normal sample with and without germline filter; MismatchFinder derived signature weights with and without germline filter; weights below 1% contribution are accumulated in "unknown" (darkgrey), lightgrey signatures show sequencing artefact signatures, yellow shows smoking related signatures and blue depicts APOBEC signatures

analysis can mirror biological concepts. To show our new method is usable for a variety of datasets, we used a mixture of different cancer types from various different sequencers. In Section 4.3.2.1 I present the dataset of healthy individuals to show specificity of our method. Then in Section 4.3.2.2 I present the analysis for metastatic breast cancer patients with matched tumour and germline sequencing. First how efficient germline filtering is in Section 4.3.2.2.1 and then how accurate and sensitive our method is when compared to the current gold standard of tumour-normal tissue analysis (Section 4.3.2.2.2) and then the envisioned approach of tumour only analysis for clinical samples in Section 4.3.2.2.3. These analysis were also performed on their melanoma equivalent in Section 4.3.2.3.1 for the matched analysis and Section 4.3.2.3.2 to show that the detection is not cancer specific.

4.3.2.1 Healthy cohort

We sequenced the 60 healthy samples, from varying age groups (24 yrs.-70 yrs. median: 48.5 yrs.) with 24 males and 36 females, in the exact same way as the tumour only samples for MBCB and melanoma (Section 4.3.2.2.3 and Section 4.3.2.3.2 respectively) to an effective average coverage of 8x WGS, with mixed healthy samples and cancer samples on sequencing flow cells to account for batch effects.

4.3.2.1.1 Bias detection This dataset of healthy samples is ideal to detect biases, because any variability that cannot be accounted to either age or gender is unwanted and will affect the cancer samples in the same way. We expect an increased mismatch rate in the older individuals do to the accumulation of somatic mutations, but most of these mutations should be accounted for by SBS1

and SBS5, as they represent the "clock-like" signatures [212]. Any other signatures would have to be assumed to be either normal variation in the healthy population or technical artefacts.

maybe add more info to describe the healthies

4.3.2.1.2 Black list generation With the strong influence filtering our both technical errors (Section 4.3.1.1) and germline variants (Section 4.3.1.3) as background noise had on our method, we hypothesised that a blacklist of mismatches found in our healthy individuals would help us further cut down on unwanted background signal and refine the somatic mismatch calls. I therefore ran mismatchfinder with significantly relaxed quality cut-offs to capture as much variation as possible. This includes a reduction in mapping quality and base quality as well as not restricting the analysis to the highly accurate overlap part of the paired end reads. However we still restricted the analysis to the same highly mappable areas of the genome the same as for the tumour analysis as well as filtering already known germline variants for a better estimation of the impact of this filter step.

The site files generated through MisMatchFinder were then concatenated and aggregated to multiple blacklists with cut-offs of a variant present in at least 3, 5 or 10 times. The bash code used for the post processing of MisMatchFinder site files can be found in Listing 4.1.

Listing 4.1: Blacklist postprocessing

```
awkCounting=$(cat <<'AWK'
{
    key=$1"\t"$2"\t"$3"\t"$4
    counts[key]++
}
END{
    for(i in counts){
        occ = counts[i]
        if(occ >= 3){
            print i"\t"counts[i] > "healthy_blacklist_sites_m3.tsv"
        }
        if(occ >= 5){
            print i"\t"counts[i] > "healthy_blacklist_sites_m5.tsv"
        }
        if(occ >= 10){
            print i"\t"counts[i] > "healthy_blacklist_sites_m10.tsv"
        }
    }
}
```

```
AWK
)

cat *_sites.tsv | awk "$awkCounting"
```

Add the health cohort in there/ possible age

4.3.2.2 MBCB patient samples

are they actually published?

The first dataset of patients is two previously published BRCA1/2 positive breast cancers. The data contains matched tumour, germline and ctDNA sequencing as high depth WGS for both patients. With the matched normal, we can use the current standard protocol of somatic mutational pattern analysis (Section 4.1.1) and compare it with our new method (Section 4.3.2.2.2).

As the sequencing data of the ctDNA is much higher depth than what is used in standard clinical practice, we down sample the data to 10x coverage, which is also the coverage of the simulated data. By using numerous different seeds for the sampling, we can generate pseudo technical replicates of the sequencing (Section B.3.5), which then in term can give an approximation of the stability of the results for both the tissue and the ctDNA samples .

And secondly we analysed a set of tumour only sequencing samples, how they would be available in clinical testing, e.g. copy number analysis [213, 214]. To show that our method can be applied to supply additional clinically relevant information from already available data (Section 4.3.2.2.3).

add the subsampled data

4.3.2.2.1 Germline artifacts As I have pointed out above, how important the germline filter step is to boost the signal of somatic variants (Section 4.2.6), we were interested how many germline variants were not filtered out with our method. The high depth matched healthy WGS samples of the breast cancer dataset is ideal for this analysis. I called germline variants on the matched normal using Strelka2 and compared the called variants with the sites which MisMatchFinder found to be somatic (retained after germline filtering) on the sub-sampled data. All variants with any quality filter assigned by Strelka2 were considered for this analysis, such that possible clonal hematopoiesis (CH) variants are still considered. Table 4.1 shows that on average 2100 germline variants are not filtered out. However, this only equates to 0.9% for MBCB196 and 1.5% for MBCB298 of all sites

Table 4.1: Germline variants retained after germline filtering with in MisMatchFinder analysis; Default parameters were used when running MisMatchFinder with gnomAD zarr for filtering. seed column shows the seed used to subsample the high depth sequencing BAM, "mismatch sites" column contains number of sites found to be changed, "germline sites" contains the number of sites also found with germline variant calling, fraction shows fraction of column 4 and 3

Patient ID	seed	mismatch sites	germline sites	fraction
MBCB196	1007	216 950	2107	0.0097
	1234	217 145	2073	0.0095
	1337	216 823	2080	0.0096
	1717	217 593	2089	0.0096
	2358	217 317	2097	0.0096
	3311	217 219	2046	0.0094
	5229	216 876	2062	0.0095
	6060	217 388	2080	0.0096
	9876	217 656	2008	0.0092
MBCB298	1756	148 495	2168	0.0146
	3599	149 901	2224	0.0148
	4117	149 382	2277	0.0152
	4306	149 549	2248	0.0150
	4359	149 805	2205	0.0147
	5788	150 103	2241	0.0149
	5887	150 099	2287	0.0152
	8387	149 533	2248	0.0150
	9754	149 547	2229	0.0149

found to be mutated. The exact numbers depend on the strictness of the parameters of the analysis as well as the mutation rate of the sample, so they should only be taken as a guideline.

Nevertheless, this result combined with the effective filtering of technical errors (Section 4.3.1.1) suggests, that the remaining sites are somatic mutations of either the healthy tissue or the cancer.

4.3.2.2.2 Matched WGS samples - when you know what the results should be

4.3.2.2.3 Tumour only WGS samples - the large scale clinical approach

4.3.2.3 Melanoma patients

For melanoma real world data, we analysed multiple datasets, which allowed to highlight and dissect specific aspects of signature detection from lcWGS when dealing with melanoma. First I will show the results on a set of two patients, were WES of tumour tissue, normal tissue and both cfDNA samples is available on top of lcWGS of the cfDNA (Section 4.3.2.3.1) and then I will highlight the

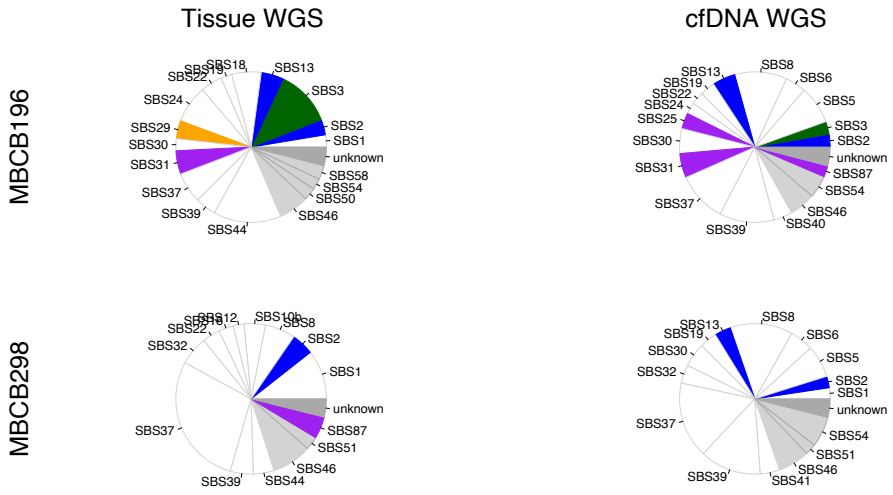

Figure 4.12: Signature weights for the WGS of two MBCB patients; Colours show cancer associated signatures: blue (APOBEC), red (UV exposure), orange (tobacco), purple (chemotherapy), light grey (sequencing artefacts), dark grey (everything below 1% weight)

results of the analysis of a bigger dataset of melanoma patients without matched normals but with varying tumour fraction (Section 4.3.2.3.2).

4.3.2.3.1 Matched WES samples - when you know what the results should be

With the help of the WES of both the tissue and the cfDNA samples, we know what the expected outcome for each of the lcWGS samples is, as the standard procedure for signature detection starts with somatic variants. Strelka2 was used to call somatic variants and the high confidence somatic SNPs were then used in R to generate signature weights with deconstructSigs. As the most recent signatures were generated using whole genome sequencing, the input was normalised with the '*exome2genome*' option when calculating weights.

For Patient 1, both the tissue as well as the cfDNA samples show a very high signature weight related to UV radiation SBS7a and SBS7b (combined $\geq 30\%$ however patient 2 only shows very little SBS7a in only the cfDNA samples.

4.3.2.3.2 Large detection level cohort - how sensitive can we be?

4.3.3 Summary

Dont know if we need this

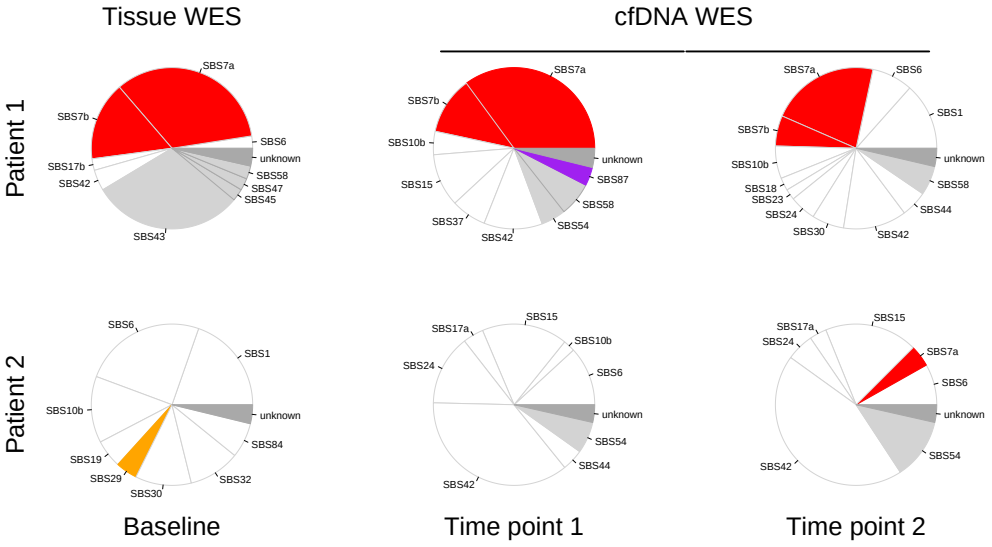

Figure 4.13: Signature weights for the WES of two melanoma patients; First column shows the results for the tissue baseline and middle and right column show the cfDNA; Colours show cancer associated signatures: blue (APOBEC), red (UV exposure), orange (tobacco), purple (chemotherapy), light grey (sequencing artefacts), dark grey (everything below 1% weight)

"As you think, so you become. Our busy minds are forever jumping to conclusions, manufacturing and interpreting signs that aren't there."

— Epictetus, *The Enchiridion*

5

Conclusion

we should have some stuff to write here in the end

Bibliography

[1] Ibiayi Dagogo-Jack and Alice T. Shaw. "Tumour heterogeneity and resistance to cancer therapies". In: *Nature Reviews Clinical Oncology* 15.2 (Nov. 2017), pp. 81–94. doi: 10.1038/nrclinonc.2017.166.

[2] R. Fisher, L. Pusztai, and C. Swanton. "Cancer heterogeneity: implications for targeted therapeutics." In: *British journal of cancer* 108 (3 Feb. 2013), pp. 479–485. issn: 1532-1827. doi: 10.1038/bjc.2012.581. ppublish.

[3] Tracy L. Leong et al. "Deep multi-region whole-genome sequencing reveals heterogeneity and gene-by-environment interactions in treatment-naive, metastatic lung cancer". In: *Oncogene* 38.10 (Oct. 2018), pp. 1661–1675. doi: 10.1038/s41388-018-0536-1.

[4] Ting Yan et al. "Multi-region sequencing unveils novel actionable targets and spatial heterogeneity in esophageal squamous cell carcinoma". In: *Nature Communications* 10.1 (Apr. 2019). doi: 10.1038/s41467-019-09255-1.

[5] The Centos Project. *Centos 7*. July 2014. url: https://www.centos.org/ (visited on 10/26/2021).

[6] Free Software Foundation. *Bash (3.2.48). [Unix shell program]*. Version 5.1.8(1)-release. 2007. url: http://ftp.gnu.org/gnu/bash/bash-3.2.48.tar.gz.

[7] Ole Tange et al. "GNU Parallel - The Command-Line Power Tool". In: *login: The USENIX Magazine* 36.1 (Feb. 2011), pp. 42–47.

[8] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2021. url: https://www.R-project.org/.

[9] Martin Morgan et al. *BiocParallel: Bioconductor facilities for parallel evaluation*. R package version 1.24.1. 2020. url: https://github.com/Bioconductor/BiocParallel.

[10] Martin Morgan. *BiocManager: Access the Bioconductor Project Package Repository*. R package version 1.30.10. 2019. url: https://CRAN.R-project.org/package=BiocManager.

[11] Achim Zeileis, Kurt Hornik, and Paul Murrell. "Escaping RGBland: Selecting colors for statistical graphics". In: *Computational Statistics & Data Analysis* 53.9 (July 2009), pp. 3259–3270. doi: 10.1016/j.csda.2008.11.033.

[12] Achim Zeileis et al. "colorspace: A Toolbox for Manipulating and Assessing Colors and Palettes". In: *Journal of Statistical Software* 96.1 (2020). doi: 10.18637/jss.v096.i01.

[13] F. Favero et al. "Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data". In: *Annals of Oncology* 26.1 (Jan. 2015), pp. 64–70. doi: 10.1093/annonc/mdu479.

[14] Ronglai Shen and Venkatraman E. Seshan. "FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing". In: *Nucleic Acids Research* 44.16 (June 2016), e131–e131. doi: 10.1093/nar/gkw520.

[15] Venkatraman E. Seshan and Ronglai Shen. *facets: Cellular Fraction and Copy Numbers from Tumor Sequencing*. R package version 0.6.0. 2018. url: https://github.com/mskcc/facets (visited on 09/29/2021).

[16] Daniel L. Cameron et al. "GRIDSS, PURPLE, LINX: Unscrambling the tumor genome via integrated analysis of structural variation and copy number". In: *bioRxiv* (Sept. 2019). doi: 10.1101/781013. url: https://doi.org/10.1101/781013.

[17] Gro Nilsen et al. "Copynumber: Efficient algorithms for single- and multi-track copy number segmentation". In: *BMC Genomics* 13.1 (Nov. 2012). doi: 10.1186/1471-2164-13-591.

[18] Gro Nilsen, Knut Liestol, and Ole Christian Lingjaerde. *copynumber: Segmentation of single- and multi-track copy number data by penalized least squares regression*. R package version 1.29.0.9000. 2021.

[19] William McLaren et al. "The Ensembl Variant Effect Predictor". In: *Genome Biology* 17.1 (June 2016). doi: 10.1186/s13059-016-0974-4.

[20] Matt Dowle and Arun Srinivasan. *data.table: Extension of 'data.frame'*. R package version 1.14.0. 2021. url: https://CRAN.R-project.org/package=data.table.

[21] Daniel Adler and S. Thomas Kelly. *vioplot: violin plot*. R package version 0.3.5. 2020. url: https://github.com/TomKellyGenetics/vioplot.

[22] Zuguang Gu, Roland Eils, and Matthias Schlesner. "Complex heatmaps reveal patterns and correlations in multidimensional genomic data". In: *Bioinformatics* 32.18 (May 2016), pp. 2847–2849. doi: 10.1093/bioinformatics/btw313.

[23] Emmanuel Paradis and Klaus Schliep. "ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R". In: *Bioinformatics* 35.3 (July 2018). Ed. by Russell Schwartz, pp. 526–528. doi: 10.1093/bioinformatics/bty633.

[24] Klaus Schliep et al. "Intertwining phylogenetic trees and networks". In: *Methods in Ecology and Evolution* 8.10 (Apr. 2017). Ed. by Richard Fitzjohn, pp. 1212–1220. doi: 10.1111/2041-210X.12760.

[25] Tal Galili. "dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering". In: 31.22 (July 2015), pp. 3718–3720. doi: 10.1093/bioinformatics/btv428.

[26] Jennifer Bryan. *googlesheets4: Access Google Sheets using the Sheets API V4*. R package version 1.0.0. 2021. url: https://CRAN.R-project.org/package=googlesheets4.

[27] Martin Morgan et al. *Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import*. R package version 2.8.0. 2021. url: https://bioconductor.org/packages/Rsamtools.

[28] Michael Lawrence et al. "Software for Computing and Annotating Genomic Ranges". In: *PLoS Computational Biology* 9.8 (8 Aug. 2013). Ed. by Andreas Prlic, e1003118. doi: 10.1371/journal.pcbi.1003118. url: http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1003118.

[29] Trevor L Davis. *optparse: Command Line Option Parser*. R package version 1.6.6. 2020. url: https://CRAN.R-project.org/package=optparse.

[30] Valerie Obenchain et al. "VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants". In: *Bioinformatics* 30.14 (2014), pp. 2076–2078. doi: 10.1093/bioinformatics/btu168.

[31] Marcel Ramos et al. "Software for the integration of multi-omics experiments in Bioconductor". In: *Cancer Research* 77(21); e39-42 (June 2017). doi: 10.1101/144774.

[32] Zuguang Gu et al. "circlize implements and enhances circular visualization in R". In: *Bioinformatics* 30.19 (19 June 2014), pp. 2811–2812. doi: 10.1093/bioinformatics/btu393.

[33] Jitao David Zhang et al. "Detect tissue heterogeneity in gene expression data with BioQC". In: *BMC Genomics* 18.1 (Apr. 2017), p. 277. doi: 10.1186/s12864-017-3661-2. url: http://accio.github.io/BioQC/.

[34] H. Pagès et al. *Biostrings: Efficient manipulation of biological strings*. R package version 2.58.0. 2020. url: https://bioconductor.org/packages/Biostrings.

[35] Rachel Rosenthal. *deconstructSigs: Identifies Signatures Present in a Tumor Sample*. R package version 1.8.0. 2016. url: https://CRAN.R-project.org/package=deconstructSigs.

[36] Hervé Pagès. *BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs*. R package version 1.58.0. 2020. url: https://bioconductor.org/packages/BSgenome.

[37] Ilari Scheinin et al. "DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly". In: *Genome Research* 24.12 (Sept. 2014), pp. 2022–2032. doi: 10.1101/gr.175141.114.

[38] Erich Neuwirth. *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-2. 2014. url: https://CRAN.R-project.org/package=RColorBrewer.

[39] Raivo Kolde. *pheatmap: Pretty Heatmaps*. R package version 1.0.12. 2019. url: https://CRAN.R-project.org/package=pheatmap.

[40] Valerie Obenchain and Lori Shepherd. *ensemblVEP: R Interface to Ensembl Variant Effect Predictor*. R package version 1.32.0. 2020.

[41] M.P.J. van der Loo. "The stringdist Package for Approximate String Matching". In: *The R Journal* 6.1 (1 2014), pp. 111–122. doi: 10.32614/rj-2014-011. url: https://CRAN.R-project.org/package=stringdist.

[42] Yang Liao, Gordon K. Smyth, and Wei Shi. "The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads". In: *Nucleic Acids Research* 47.8 (8 Feb. 2019), e47–e47. doi: 10.1093/nar/gkz114.

[43] Hadley Wickham et al. *svglite: An 'SVG' Graphics Device*. R package version 2.0.0. 2021. url: https://CRAN.R-project.org/package=svglite.

[44] Paul Murrell. "Importing Vector Graphics: The grImport Package for R". In: *Journal of Statistical Software* 30.4 (2009), pp. 1–37. doi: 10.18637/jss.v030.i04. url: http://www.jstatsoft.org/v30/i04/.

[45] Duncan Temple Lang. *XML: Tools for Parsing and Generating XML Within R and S-Plus*. R package version 3.99-0.5. 2020. url: https://CRAN.R-project.org/package=XML.

[46] Hao Zhu. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. R package version 1.3.4. 2021. url: https://CRAN.R-project.org/package=kableExtra.

[47] Fridolin Wild. *lsa: Latent Semantic Analysis*. R package version 0.73.2. 2020. url: https://CRAN.R-project.org/package=lsa.

[48] Jim Baglama, Lothar Reichel, and B. W. Lewis. *irlba: Fast Truncated Singular Value Decomposition and Principal Components Analysis for Large Dense and Sparse Matrices*. R package version 2.3.3. 2019. url: https://CRAN.R-project.org/package=irlba.

[49] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, June 8, 2016. 260 pp. isbn: 978-3-319-24277-4. url: https://ggplot2.tidyverse.org.

[50] Guido VanRossum. *The Python language reference*. Hampton, NHRedwood City, Calif: Python Software FoundationSoHo Books, 2010. isbn: 9781441412690.

[51] Charles R. Harris et al. “Array programming with NumPy”. In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. doi: 10.1038/s41586-020-2649-2.

[52] Endre Bakken Stovner and Pål Sætrom. “PyRanges: efficient comparison of genomic intervals in Python”. In: *Bioinformatics* (Aug. 2019). Ed. by John Hancock. doi: 10.1093/bioinformatics/btz615.

[53] Andreas Heger, Kevin Jacobs, et al. *pysam: htslib interface for python*. Oct. 25, 2021. url: https://github.com/pysam-developers/pysam (visited on 10/26/2021).

[54] James K Bonfield et al. “HTSlib: C library for reading/writing high-throughput sequencing data”. In: *GigaScience* 10.2 (Jan. 2021). doi: 10.1093/gigascience/giab007.

[55] Petr Danecek et al. “Twelve years of SAMtools and BCFtools”. In: *GigaScience* 10.2 (Jan. 2021). doi: 10.1093/gigascience/giab008.

[56] Alistair Miles et al. *zarr-developers/zarr-python: v2.10.2*. 2021. doi: 10.5281/ZENODO.5579625.

[57] Wes McKinney et al. “Data Structures for Statistical Computing in Python”. In: *Proceedings of the 9th Python in Science Conference*. Vol. 445. Austin, TX. SciPy, 2010, pp. 51–56. doi: 10.25080/majora-92bf1922-00a.

[58] Jeff Reback et al. *pandas-dev/pandas: Pandas 1.3.4*. 2021. doi: 10.5281/ZENODO.5574486.

[59] Robert T. McGibbon et al. *quadprog: Quadratic Programming Solver (Python) v0.1.10*. Oct. 1, 2021. url: https://github.com/quadprog/quadprog (visited on 10/26/2021).

[60] Pauli Virtanen et al. “SciPy 1.0: fundamental algorithms for scientific computing in Python”. In: *Nature Methods* 17.3 (Feb. 2020), pp. 261–272. doi: 10.1038/s41592-019-0686-2.

[61] J. D. Watson and F. H. C. Crick. “Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid”. In: *Nature* 171.4356 (Apr. 1953), pp. 737–738. doi: 10.1038/171737a0.

[62] F. Liang et al. “Homology-directed repair is a major double-strand break repair pathway in mammalian cells”. In: *Proceedings of the National Academy of Sciences* 95.9 (Apr. 1998), pp. 5172–5177. doi: 10.1073/pnas.95.9.5172.

[63] Richard R. Sinden. “Introduction to the Structure, Properties, and Reactions of DNA”. In: *DNA Structure and Function*. Elsevier, 1994, pp. 1–57. doi: 10.1016/b978-0-08-057173-7.50006-7.

[64] J. Craig Venter et al. “The Sequence of the Human Genome”. In: *Science* 291.5507 (Feb. 2001), pp. 1304–1351. doi: 10.1126/science.1058040.

[65] Colin M. Hammond et al. “Histone chaperone networks shaping chromatin function”. In: *Nature Reviews Molecular Cell Biology* 18.3 (Jan. 2017), pp. 141–158. doi: 10.1038/nrm.2016.159.

[66] Boyan Bonev and Giacomo Cavalli. “Organization and function of the 3D genome”. In: *Nature Reviews Genetics* 17.11 (Oct. 2016), pp. 661–678. doi: 10.1038/nrg.2016.112.

[67] Tuguo Tateoka. “A contribution to the taxonomy of theAgrostis mertensii-flaccida complex (Poaceae) in Japan”. In: *The Botanical Magazine Tokyo* 88.2 (June 1975), pp. 65–87. doi: 10.1007/bf02491243.

[68] R. Trivers and H Hare. “Haploidploidy and the evolution of the social insect”. In: *Science* 191.4224 (Jan. 1976), pp. 249–263. doi: 10.1126/science.1108197.

[69] Sarah P. Otto. “The Evolutionary Consequences of Polyploidy”. In: *Cell* 131.3 (Nov. 2007), pp. 452–462. doi: 10.1016/j.cell.2007.10.022.

[70] Marvin I. Gottlieb et al. “Trisomy-17 syndrome”. In: *The American Journal of Medicine* 33.5 (Nov. 1962), pp. 763–773. doi: 10.1016/0002-9343(62)90253-x.

[71] Anna Cereda and John C Carey. “Trisomy 18 Syndrome”. In: *Atlas of Genetic Diagnosis and Counseling*. Vol. 7. 1. Humana Press, 2012, pp. 990–996. doi: 10.1186/1750-1172-7-81.

[72] Maj A Hultén et al. “On the origin of trisomy 21 Down syndrome”. In: *Molecular Cytogenetics* 1.1 (2008), p. 21. doi: 10.1186/1755-8166-1-21.

[73] David M. J. Lilley. “Structures of helical junctions in nucleic acids”. In: *Quarterly Reviews of Biophysics* 33.2 (May 2000), pp. 109–159. doi: 10.1017/s0033583500003590.

[74] William P. Hanage, Christophe Fraser, and Brian G. Spratt. “The impact of homologous recombination on the generation of diversity in bacteria”. In: *Journal of Theoretical Biology* 239.2 (Mar. 2006), pp. 210–219. doi: 10.1016/j.jtbi.2005.08.035.

[75] Ying Kong et al. “Homologous Recombination Drives Both Sequence Diversity and Gene Content Variation in Neisseria meningitidis”. In: *Genome Biology and Evolution* 5.9 (July 2013), pp. 1611–1627. doi: 10.1093/gbe/evt116.

[76] Charles Darwin. “On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life”. In: *Evolutionary Writings*. Oxford University Press, May 2010. doi: 10.1093/owc/9780199580149.003.0005.

[77] Kathleen Sprouffske et al. “High mutation rates limit evolutionary adaptation in Escherichia coli”. In: *PLOS Genetics* 14.4 (Apr. 2018). Ed. by Ivan Matic, e1007324. doi: 10.1371/journal.pgen.1007324.

[78] Ludmil B Alexandrov et al. “Clock-like mutational processes in human somatic cells”. In: *Nature Genetics* 47.12 (Nov. 2015), pp. 1402–1407. doi: 10.1038/ng.3441.

[79] Luiza Moore et al. “The mutational landscape of human somatic and germline cells”. In: *Nature* (Aug. 2021). doi: 10.1038/s41586-021-03822-7.

[80] Hanan E. Shamseldin et al. “Identification of embryonic lethal genes in humans by autozygosity mapping and exome sequencing in consanguineous families”. In: *Genome Biology* 16.1 (June 2015). doi: 10.1186/s13059-015-0681-6.

[81] Laura Frey et al. “Mammalian VPS45 orchestrates trafficking through the endosomal system”. In: *Blood* 137.14 (Apr. 2021), pp. 1932–1944. doi: 10.1182/blood.2020006871.

[82] Shan Dan et al. “Clinical application of massively parallel sequencing-based prenatal noninvasive fetal trisomy test for trisomies 21 and 18 in 11 105 pregnancies with mixed risk factors”. In: *Prenatal Diagnosis* 32.13 (Nov. 2012), pp. 1225–1232. doi: 10.1002/pd.4002.

[83] Kypros H. Nicolaides et al. “Noninvasive Prenatal Testing for Fetal Trisomies in a Routinely Screened First-Trimester Population”. In: *Obstetrical & Gynecological Survey* 68.3 (Mar. 2013), pp. 173–175. doi: 10.1097/ogx.0b013e318285bf66.

[84] Frank Diehl et al. “Circulating mutant DNA to assess tumor dynamics”. In: *Nature Medicine* 14.9 (July 2008), pp. 985–990. doi: 10.1038/nm.1789.

[85] Heidi Schwarzenbach, Dave S. B. Hoon, and Klaus Pantel. “Cell-free nucleic acids as biomarkers in cancer patients”. In: *Nature Reviews Cancer* 11.6 (May 2011), pp. 426–437. doi: 10.1038/nrc3066.

[86] R. Padmanabhan, E. Jay, and R. Wu. “Chemical Synthesis of a Primer and Its Use in the Sequence Analysis of the Lysozyme Gene of Bacteriophage T4”. In: *Proceedings of the National Academy of Sciences* 71.6 (June 1974), pp. 2510–2514. doi: 10.1073/pnas.71.6.2510.

[87] F. Sanger and A.R. Coulson. “A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase”. In: *Journal of Molecular Biology* 94.3 (May 1975), pp. 441–448. doi: 10.1016/0022-2836(75)90213-2.

[88] F. Sanger, S. Nicklen, and A. R. Coulson. “DNA sequencing with chain-terminating inhibitors”. In: *Proceedings of the National Academy of Sciences* 74.12 (Dec. 1977), pp. 5463–5467. doi: 10.1073/pnas.74.12.5463.

[89] Eric S. Lander et al. “Initial sequencing and analysis of the human genome”. In: *Nature* 409 (Feb. 15, 2001): *The human genome*, pp. 860–921. doi: 10.1038/35057062.

[90] Inc Illumina. *How short inserts affect sequencing performance*. Sept. 2020. url: https://sapac.support.illumina.com/bulletins/2020/12/how-short-inserts-affect-sequencing-performance.html (visited on 09/08/2021).

[91] Elaine R. Mardis. "Next-Generation DNA Sequencing Methods". In: *Annual Review of Genomics and Human Genetics* 9.1 (Sept. 2008), pp. 387–402. doi: 10.1146/annurev.genom.9.081307.164359.

[92] Jenny Straiton et al. "From Sanger sequencing to genome databases and beyond". In: *BioTechniques* 66.2 (Feb. 2019), pp. 60–63. doi: 10.2144/btn-2019-0011.

[93] G. M. Church and W. Gilbert. "Genomic sequencing." In: *Proceedings of the National Academy of Sciences* 81.7 (Apr. 1984), pp. 1991–1995. doi: 10.1073/pnas.81.7.1991.

[94] G. Church and S Kieffer-Higgins. "Multiplex DNA sequencing". In: *Science* 240 (Apr. 1988), pp. 185–188. doi: 10.1126/science.3353714.

[95] Alexander Payne et al. "BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files". In: *Bioinformatics* 35.13 (Nov. 2018). Ed. by Inanc Birol, pp. 2193–2198. doi: 10.1093/bioinformatics/bty841.

[96] Ploy N. Pratanwanich et al. "Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore". In: *Nature Biotechnology* (July 2021). doi: 10.1038/s41587-021-00949-w.

[97] Nicolas L Bray et al. "Near-optimal probabilistic RNA-seq quantification". In: *Nature Biotechnology* 34.5 (Apr. 2016), pp. 525–527. doi: 10.1038/nbt.3519.

[98] Rob Patro et al. "Salmon provides fast and bias-aware quantification of transcript expression". In: *Nature Methods* 14.4 (Mar. 2017), pp. 417–419. doi: 10.1038/nmeth.4197.

[99] Brian D. Ondov et al. "Mash: fast genome and metagenome distance estimation using MinHash". In: *Genome Biology* 17.1 (June 2016). doi: 10.1186/s13059-016-0997-x.

[100] Brian B Luczak, Benjamin T James, and Hani Z Girgis. "A survey and evaluations of histogram-based statistics in alignment-free sequence comparison". In: *Briefings in Bioinformatics* 20.4 (Dec. 2017), pp. 1222–1237. doi: 10.1093/bib/bbx161.

[101] Heng Li. "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM". In: (Mar. 16, 2013). arXiv: 1303.3997 [q-bio.GN].

[102] Ben Langmead et al. “Scaling read aligners to hundreds of threads on general-purpose processors”. In: *Bioinformatics* 35.3 (July 2018). Ed. by John Hancock, pp. 421–432. doi: 10.1093/bioinformatics/bty648.

[103] Klaus F. X. Mayer et al. “A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome”. In: *Science* 345.6194 (July 2014), pp. 1251788–1251788. doi: 10.1126/science.1251788.

[104] Erik Garrison and Gabor Marth. “Haplotype-based variant detection from short-read sequencing”. In: *arXiv preprint arXiv:1207.3907 [q-bio.GN]* (July 17, 2012). arXiv: http://arxiv.org/abs/1207.3907v2 [q-bio.GN].

[105] Zhongwu Lai et al. “VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research”. In: *Nucleic Acids Research* 44.11 (Apr. 2016), e108–e108. doi: 10.1093/nar/gkw227.

[106] Sangtae Kim et al. “Strelka2: fast and accurate calling of germline and somatic variants”. In: *Nature Methods* 15.8 (July 2018), pp. 591–594. doi: 10.1038/s41592-018-0051-x.

[107] David Benjamin et al. “Calling Somatic SNVs and Indels with Mutect2”. In: *bioRxiv* (Dec. 2019). doi: 10.1101/861054. url: https://doi.org/10.1101/861054.

[108] Daniel P. Cooke, David C. Wedge, and Gerton Lunter. “A unified haplotype-based method for accurate and comprehensive variant calling”. In: *Nature Biotechnology* 39.7 (Mar. 2021), pp. 885–892. doi: 10.1038/s41587-021-00861-3.

[109] GATK Team. *Somatic calling is NOT simply a difference between two callsets*. Sept. 15, 2021. url: https://gatk.broadinstitute.org/hc/en-us/articles/360035890491 (visited on 09/23/2021).

[110] Amaro Taylor-Weiner et al. “DeTiN: overcoming tumor-in-normal contamination”. In: *Nature Methods* 15.7 (June 2018), pp. 531–534. doi: 10.1038/s41592-018-0036-9.

[111] Konrad J. Karczewski et al. “The mutational constraint spectrum quantified from variation in 141,456 humans”. In: *Nature* 581.7809 (May 2020), pp. 434–443. doi: 10.1038/s41586-020-2308-7.

[112] Ali Karimnezhad et al. “Accuracy and reproducibility of somatic point mutation calling in clinical-type targeted sequencing data”. In: *BMC Med Genomics* 13.1 (Oct. 2020). doi: 10.1186/s12920-020-00803-z.

[113] James Henry Breasted. “The Edwin Smith Surgical Papyrus: published in facsimile and hieroglyphic transliteration with translation and commentary in two volumes”. In: (1930).

[114] Steven I Hajdu. “Greco-Roman thought about cancer”. In: *Cancer* 100.10 (2004), pp. 2048–2051. doi: 10.1002/cncr.20198.

[115] John Chadwick and William N Mann. *The medical works of Hippocrates*. Oxford Blackwell Scientific Publications, 1950, pp. 124–147.

[116] Celsus. “De Medicina”. In: *British Journal of Surgery* 26.103 (Jan. 1939). With an english translation by W. G. Spencer, M.S. (Lond.), F.R.C.S. (Eng.), pp. 658–659. doi: 10.1002/bjs.18002610338.

[117] E. G. Browne. *Arabian medicine. the FitzPatrick lectures Delivered at the College of Physicians in November 1919 and November 1920*. Cambridge Library Collection - History of Medicine. Cambridge: Cambridge University Press, 2011. 154 pp. isbn: 9780511709296. doi: 10.1017/cbo9780511709296.

[118] J. E. Pilcher. “Guy de Chauliac and Henri de Mondeville,-A Surgical Retrospect.” In: *Annals of surgery* 21 (1 Jan. 1895), pp. 84–102. issn: 0003-4932.

[119] Paracelsus (Bombastus von Hohenheim TPA). “De Grandibus, de Compositionibus, et Dosibus Receptorum ac Naturalium (Libri Septem”. In: (1562).

[120] Marco Aurelio 1580-1656 Severino. *De recondita abscessuum natura libri VII*. Apud Octavium Beltranum, 1632.

[121] Zacutus Lusitani. “Praxis Medical Admiranda”. In: *Lugduni: J. Hugvetan* (1649).

[122] Nicolai Tulpii. *Observationes Medicae*. Amstelredami, 1652.

[123] Claude Deshaies-Gendron. *Recherches sur la nature et la guerison des cancers. Enquiries into the nature, knowledge, and cure of cancers. By Mr. Deshaies Gendron, ... Done out of French.* Print edition. Gale ECCO, 1701. 146 pp. isbn: 978-1385259078.

[124] Steven I. Hajdu. “The First Printed Case Reports of Cancer. The First Printed Case Reports of Cancer”. In: *Cancer* (2010). doi: 10.1002/cncr.25000.

[125] James Nooth. *Observations on the treatment of scirrhous tumours, and cancers of the breast*. G. and J. Robinson, 1804, p. 101.

[126] Michael Etmüller. *ETMULLERUS ABRIDG'D : or, a compleat system of the theory and practice of physic being a ... description of all diseases incident to men, women*. Gale ECCO, 2018. isbn: 1385770074.

[127] Lorenz Heister. *Kleine Chirurgie, oder, Wund-Artzney: in welcher ein kurzer doch deutlicher Unterricht und Begriff dieser Wissenschafft gegeben ... werden*. German. Nürnberg: Joh. Adam Stein und Gabriel Nicolaus Raspe, 1747.

[128] Steven I. Hajdu. "A note from history: Landmarks in history of cancer, part 2". In: *Cancer* 117.12 (Dec. 2010), pp. 2811–2820. doi: 10.1002/cncr.25825.

[129] Johannes Müller. *Über den feinern Bau und die Formen der krankhaften Geschwülste : in zwei Lieferungen*. German. Reimer, 1838. url: https://echo.mpiwg-berlin.mpg.de/ECHOdocuView?url=/permanent/library/84U9MMK0/pageimg&start=61&mode=imagepath&pn=66 (visited on 12/07/2021).

[130] J.H. Bennett. *On Cancerous and Cancroid Growths*. Sutherland and Knox, 1849, pp. vi–viii. url: https://books.google.com.au/books?id=VQg%5C_AAAAcAAJ.

[131] Rudolf ludwig Carl10.1002/andp.18983000103 Virchow. *Die Krankhaften Geschwülste. dreissig Vorlesungen, gehalten während des Wintersemesters 1862-1863 an der Universität zu Berlin*. German. 3 vols. Berlin: Verlag von August Hirschwald, 1863. url: http://resource.nlm.nih.gov/62231840R (visited on 08/17/2021).

[132] W. C. Röntgen. "Ueber eine neue Art von Strahlen". In: *Annalen der Physik* 300.1 (1898), pp. 12–17. doi: 10.1002/andp.18983000103.

[133] EAFA Frieben. "Demonstration eines cancroids des rechten handruckens, das sich nach langdauerner einwirkung von roentgenstrahlen entwickelt hatte". In: *Forsch Roentgenstr* 6 (1902), pp. 106–111.

[134] W Scholtz. "Ueber den Einfluss der Röntgenstrahlen auf die Haut in gesundem und krankem Zustande". In: *Archiv für Dermatologie und Syphilis* 59.3 (1902), pp. 421–446.

[135] Paul Ehrlich. *Beiträge zur experimentellen Pathologie und Chemotherapie*. German. Germany, Leipzig: Akademische Verlagsgesellschaft, 1909, pp. 167–194. 247 pp. url: https://id.lib.harvard.edu/curiosity/contagion/36-990061083080203941 (visited on 12/07/2021).

[136] Otto Warburg. "Photochemie der Eisencarbonylverbindungen und das absolute Absorptionsspektrum des Atmungsferments". In: *Naturwissenschaften* 16.45-47 (1928), pp. 856–861.

[137] Albert Claude, Keith R. Porter, and Edward G. Pickels. "Electron Microscope Study of Chicken Tumor Cells". In: *Cancer Research* 7.7 (1947), pp. 421–430. issn: 0008-5472. eprint: https://cancerres.aacrjournals.org/content/7/7/421.full.pdf. url: https://cancerres.aacrjournals.org/content/7/7/421.

[138] Robert J Huebner and George J Todaro. “Oncogenes of RNA tumor viruses as determinants of cancer”. In: *Proceedings of the National Academy of Sciences* 64.3 (Nov. 1969), pp. 1087–1094. doi: 10.1073/pnas.64.3.1087.

[139] David Baltimore. “Viral RNA-dependent DNA polymerase: RNA-dependent DNA polymerase in virions of RNA tumour viruses”. In: *Nature* 226.5252 (June 1970), pp. 1209–1211. doi: 10.1038/2261209a0.

[140] Howard M Temin, S Mizutami, et al. “RNA-dependent DNA polymerase in virions of Rous sarcoma virus.” In: *Cold Spring Harbor Symposia on Quantitative Biology* 226.0 (Jan. 1970), pp. 1211–1213. doi: 10.1101/sqb.1970.035.01.100.

[141] Frederick P Li and Joseph F Fraumeni Jr. “Rhabdomyosarcoma in children: epidemiologic study and identification of a familial caneer syndrome”. In: *Journal of the National Cancer Institute* 43.6 (1969), pp. 1365–1373.

[142] Douglas Hanahan and Robert A Weinberg. “The Hallmarks of Cancer”. In: *Cell* 100.1 (Jan. 2000), pp. 57–70. doi: 10.1016/s0092-8674(00)81683-9.

[143] Douglas Hanahan and Robert A. Weinberg. “Hallmarks of Cancer: The Next Generation”. In: *Cell* 144.5 (Mar. 2011), pp. 646–674. doi: 10.1016/j.cell.2011.02.013.

[144] Yousef Ahmed Fouad and Carmen Aanei. “Revisiting the hallmarks of cancer.” In: *American journal of cancer research* 7 (5 2017), pp. 1016–1036. issn: 2156-6976. epublish.

[145] Douglas Hanahan. “Hallmarks of Cancer: New Dimensions”. In: *Cancer Discovery* 12.1 (Jan. 2022), pp. 31–46. doi: 10.1158/2159-8290.cd-21-1059.

[146] Elizabeth A. Kuczynski et al. “Vessel co-option in cancer.” In: *Nature Reviews Clinical Oncology* 16.8 (Feb. 2019), pp. 469–493. doi: 10.1038/s41571-019-0181-9.

[147] Javad Noorbakhsh et al. “Distribution-based measures of tumor heterogeneity are sensitive to mutation calling and lack strong clinical predictive power”. In: *Scientific Reports* 8.1 (July 2018). doi: 10.1038/s41598-018-29154-7.

[148] Ryan Poplin et al. “Scaling accurate genetic variant discovery to tens of thousands of samples”. In: *bioRxiv* (Nov. 2017). doi: 10.1101/201178.

[149] Neil A. Miller et al. “A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases”. In: *Genome Medicine* 7.1 (Sept. 2015). doi: 10.1186/s13073-015-0221-8.

[150] Ryan Poplin et al. “A universal SNP and small-indel variant caller using deep neural networks”. In: *Nature Biotechnology* 36.10 (Sept. 2018), pp. 983–987. doi: 10.1038/nbt.4235.

[151] Melanie Schirmer et al. “Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data”. In: *BMC Bioinformatics* 17.1 (Mar. 2016). doi: 10.1186/s12859-016-0976-y.

[152] Nicholas Stoler and Anton Nekrutenko. “Sequencing error profiles of Illumina sequencing instruments”. In: *NAR Genomics and Bioinformatics* 3.1 (Jan. 2021). doi: 10.1093/nargab/lqab019.

[153] Geraldine van der Auwera Brian O’Connor. *Genomics in the Cloud*. O’Reilly UK Ltd., May 1, 2020. 467 pp. isbn: 1491975199. url: https://www.oreilly.com/library/view/genomics-in-the/9781491975183/.

[154] GATK Team. *Panel of Normals (PON)*. July 23, 2021. url: https://gatk.broadinstitute.org/hc/en-us/articles/360035890631 (visited on 09/23/2021).

[155] GATK Team. *Mutect2 multi-sample*. Sept. 25, 2020. url: https://gatk.broadinstitute.org/hc/en-us/community/posts/360062528691 (visited on 10/23/2020).

[156] Malvina Josephidou, Andy G. Lynch, and Simon Tavaré. “multiSNV: a probabilistic approach for improving detection of somatic point mutations from multiple related tumour samples”. In: *Nucleic Acids Research* 43.9 (Feb. 2015), e61–e61. doi: 10.1093/nar/gkv135.

[157] Christoffer Flensburg et al. “SuperFreq: Integrated mutation detection and clonal tracking in cancer”. In: *PLOS Computational Biology* 16.2 (Feb. 2020). Ed. by Florian Markowetz, e1007603. doi: 10.1371/journal.pcbi.1007603.

[158] S Hollizeck et al. “Custom workflows to improve joint variant calling from multiple related tumour samples: FreeBayesSomatic and Strelka2Pass”. In: *Bioinformatics* (Sept. 2021). Ed. by Can Alkan. doi: 10.1093/bioinformatics/btab606.

[159] Colby Chiang et al. “SpeedSeq: ultra-fast personal genome analysis and interpretation”. In: *Nature Methods* 12.10 (Aug. 2015), pp. 966–968. doi: 10.1038/nmeth.3505.

[160] Brad Chapman et al. *bcbio/bcbio-nextgen: v1.2.4*. 2021. doi: 10.5281/ZENODO.3564938.

[161] Brendan A. Veeneman et al. “Two-pass alignment improves novel splice junction quantification”. In: *Bioinformatics* 32 (Oct. 2015), pp. 43–49. doi: 10.1093/bioinformatics/btv642.

[162] Weichun Huang et al. “ART: a next-generation sequencing read simulator”. In: *Bioinformatics* 28.4 (Dec. 2011), pp. 593–594. doi: 10.1093/bioinformatics/btr708.

[163] Adam D Ewing et al. “Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection”. In: *Nature Methods* 12.7 (May 2015), pp. 623–630. doi: 10.1038/nmeth.3407.

[164] Benjamin J. Solomon et al. “RET Solvent Front Mutations Mediate Acquired Resistance to Selective RET Inhibition in RET-Driven Malignancies”. In: *Journal of Thoracic Oncology* 15.4 (Apr. 2020), pp. 541–549. doi: 10.1016/j.jtho.2020.01.006.

[165] Ismael A. Vergara et al. “Evolution of late-stage metastatic melanoma is dominated by aneuploidy and whole genome doubling”. In: *Nature Communications* 12.1 (Mar. 2021). doi: 10.1038/s41467-021-21576-8.

[166] Zheng Hu et al. “Quantitative evidence for early metastatic seeding in colorectal cancer”. In: *Nature Genetics* 51.7 (June 2019), pp. 1113–1122. doi: 10.1038/s41588-019-0423-x.

[167] N Saitou and M Nei. “The neighbor-joining method: a new method for reconstructing phylogenetic trees.” In: *Molecular Biology and Evolution* (July 1987). doi: 10.1093/oxfordjournals.molbev.a040454.

[168] Radu Mihaescu, Dan Levy, and Lior Pachter. “Why Neighbor-Joining Works”. In: *Algorithmica* 54.1 (Dec. 2007), pp. 1–24. doi: 10.1007/s00453-007-9116-4.

[169] Robert Reuven Sokal and Charles Duncan Michener. *A Statistical Method for Evaluating Systematic Relationships*. Vol. 38.2. University of Kansas science bulletin 22. University of Kansas, 1958. 30 pp.

[170] Emile Zuckerkandl and Linus Pauling. “Molecular Disease, Evolution, and Genic Heterogeneity”. In: *Horizons in biochemistry* (1962), pp. 189–225.

[171] D. Shibata. “Mutation and epigenetic molecular clocks in cancer”. In: *Carcinogenesis* 32.2 (Nov. 2010), pp. 123–128. doi: 10.1093/carcin/bgq239.

[172] Joseph Felsenstein. “Evolutionary trees from DNA sequences: A maximum likelihood approach”. In: *Journal of Molecular Evolution* 17.6 (Nov. 1981), pp. 368–376. doi: 10.1007/bf01734359.

[173] Masami Hasegawa, Hirohisa Kishino, and Taka-aki Yano. “Dating of the human-ape splitting by a molecular clock of mitochondrial DNA”. In: *Journal of Molecular Evolution* 22.2 (Oct. 1985), pp. 160–174. doi: 10.1007/bf02101694.

[174] R. W. Hamming. “Error Detecting and Error Correcting Codes”. In: *Bell System Technical Journal* 29.2 (Apr. 1950), pp. 147–160. doi: 10.1002/j.1538-7305.1950.tb00463.x.

[175] Benjamin Werner et al. “Measuring single cell divisions in human tissues from multi-region sequencing data”. In: *Nature Communications* 11.1 (Feb. 2020). doi: 10.1038/s41467-020-14844-6.

[176] T. Arai et al. “Tumor doubling time and prognosis in lung cancer patients: evaluation from chest films and clinical follow-up study”. In: *Japanese journal of clinical oncology* 24 (4 Aug. 1994), pp. 199–204. issn: 0368-2811. ppublish.

[177] Damien M de Vienne. “Tanglegrams Are Misleading for Visual Evaluation of Tree Congruence”. In: 36.1 (Oct. 2018). Ed. by Jeffrey Townsend, pp. 174–176. doi: 10.1093/molbev/msy196.

[178] L. Tan et al. “Prediction and monitoring of relapse in stage III melanoma using circulating tumor DNA”. In: *Annals of Oncology* 30.5 (May 2019), pp. 804–814. doi: 10.1093/annonc/mdz048.

[179] “ctDNA is a specific and sensitive biomarker in multiple human cancers.” In: *Cancer discovery* 4.4 (4 Apr. 2014), OF8. issn: 2159-8290. doi: 10.1158/2159-8290.CD-RW2014-051. ppublish.

[180] Amit G Deshwar et al. “PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors”. In: *Genome Biology* 16.1 (Feb. 2015). doi: 10.1186/s13059-015-0602-8.

[181] Yuchao Jiang et al. “Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing”. In: 113.37 (Aug. 2016), E5528–E5537. doi: 10.1073/pnas.1522203113.

[182] Francesco Marass et al. “A phylogenetic latent feature model for clonal deconvolution”. In: *The Annals of Applied Statistics* 10.4 (Dec. 2016). doi: 10.1214/16-aoas986.

[183] Sayaka Miura et al. “Predicting clone genotypes from tumor bulk sequencing of multiple samples”. In: (June 2018). Ed. by John Hancock. doi: 10.1093/bioinformatics/bty469.

[184] Mohammed El-Kebir, Gryte Satas, and Benjamin J. Raphael. “Inferring parsimonious migration histories for metastatic cancers”. In: 50.5 (Apr. 2018), pp. 718–726. doi: 10.1038/s41588-018-0106-z.

[185] Giulio Caravagna et al. “The MOBSTER R package for tumour subclonal deconvolution from bulk DNA whole-genome sequencing data”. In: *BMC Bioinformatics* 21.1 (Nov. 2020). doi: 10.1186/s12859-020-03863-1.

[186] Maxime Tarabichi et al. “A practical guide to cancer subclonal reconstruction from DNA sequencing.” In: *Nature methods* 18.2 (2 Feb. 2021), pp. 144–155. issn: 1548-7105. doi: 10.1038/s41592-020-01013-2. ppublish.

[187] Sayaka Miura et al. “Power and pitfalls of computational methods for inferring clone phylogenies and mutation orders from bulk sequencing data”. In: *Scientific Reports* 10.1 (Feb. 2020). doi: 10.1038/s41598-020-59006-2.

[188] Ignaty Leshchiner et al. “Comprehensive analysis of tumour initiation, spatial and temporal progression under multiple lines of treatment”. In: *bioRxiv* (Dec. 2018). doi: 10.1101/508127.

[189] Moritz Gerstung et al. “The evolutionary history of 2,658 cancers”. In: *Nature* 578.7793 (Feb. 2020), pp. 122–128. doi: 10.1038/s41586-019-1907-7.

[190] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal. “Cancer statistics, 2018”. In: *CA: A Cancer Journal for Clinicians* 68.1 (Jan. 2018), pp. 7–30. doi: 10.3322/caac.21442.

[191] Julian R. Molina et al. “Non-Small Cell Lung Cancer: Epidemiology, Risk Factors, Treatment, and Survivorship”. In: *Mayo Clinic Proceedings* 83.5 (May 2008), pp. 584–594. doi: 10.4065/83.5.584.

[192] Sophie Sun, Joan H. Schiller, and Adi F. Gazdar. “Lung cancer in never smokers — a different disease”. In: *Nature Reviews Cancer* 7.10 (Oct. 2007), pp. 778–790. doi: 10.1038/nrc2190.

[193] Marian L. Burr et al. “An Evolutionarily Conserved Function of Polycomb Silences the MHC Class I Antigen Presentation Pathway and Enables Immune Evasion in Cancer”. In: *Cancer Cell* 36.4 (Oct. 2019), 385–401.e8. doi: 10.1016/j.ccell.2019.08.008.

[194] Monica Hollstein et al. “p53 Mutations in Human Cancers”. In: *Science* 253.5015 (July 1991), pp. 49–53. doi: 10.1126/science.1905840.

[195] Jill E. Kucab et al. “A Compendium of Mutational Signatures of Environmental Agents”. In: *Cell* 177.4 (May 2019), 821–836.e16. doi: 10.1016/j.cell.2019.03.001.

[196] Ludmil B. Alexandrov et al. “Signatures of mutational processes in human cancer”. In: *Nature* 500.7463 (Aug. 2013), pp. 415–421. doi: 10.1038/nature12477.

[197] Nigel J. O’Neil, Melanie L. Bailey, and Philip Hieter. “Synthetic lethality and cancer”. In: *Nature Reviews Genetics* 18.10 (June 2017), pp. 613–623. doi: 10.1038/nrg.2017.47.

[198] Ludmil B. Alexandrov et al. “The repertoire of mutational signatures in human cancer”. In: *Nature* 578.7793 (Feb. 2020), pp. 94–101. doi: 10.1038/s41586-020-1943-3.

[199] Adam Auton et al. “A global reference for human genetic variation”. In: *Nature* 526.7571 (Sept. 2015), pp. 68–74. doi: 10.1038/nature15393.

[200] Mahdi Heydari et al. “Illumina error correction near highly repetitive DNA regions improves de novo genome assembly”. In: *BMC Bioinformatics* 20.1 (June 2019). doi: 10.1186/s12859-019-2906-2.

[201] Irena Hudecova et al. “Characteristics, origin, and potential for cancer diagnostics of ultra-short plasma cell-free DNA”. In: *Genome Research* (Dec. 2021), gr.275691.121. doi: 10.1101/gr.275691.121.

[202] Yan Guo et al. “The effect of strand bias in Illumina short-read sequencing data”. In: *BMC Genomics* 13.1 (Nov. 2012). doi: 10.1186/1471-2164-13-666.

[203] Christopher T. Saunders et al. “Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs”. In: *Bioinformatics* 28.14 (May 2012), pp. 1811–1817. doi: 10.1093/bioinformatics/bts271.

[204] GATK Team. *StrandBiasBySample*. Sept. 18, 2019. url: https://gatk.broadinstitute.org/hc/en-us/articles/360040096492 (visited on 12/21/2021).

[205] Alistair Miles; pyup.io bot; Murillo R.; Peter Ralph; Nick Harding; Rahul Pisupati; Summer Rae; Tim Millar. *scikit-allel: A Python package for exploring and analysing genetic variation data*. June 14, 2021. doi: 10.5281/zenodo.4759368.

[206] Andy G. Lynch. “Decomposition of mutational context signatures using quadratic programming methods”. In: *F1000Research* 5 (June 2016), p. 1253. doi: 10.12688/f1000research.8918.1.

[207] Michael S. Lawrence et al. “Mutational heterogeneity in cancer and the search for new cancer-associated genes”. In: *Nature* 499.7457 (June 2013), pp. 214–218. doi: 10.1038/nature12213.

[208] Tenghui Chen et al. “Hotspot mutations delineating diverse mutational signatures and biological utilities across cancer types”. In: *BMC Genomics* 17.S2 (June 2016). doi: 10.1186/s12864-016-2727-x.

[209] John G Tate et al. “COSMIC: the Catalogue Of Somatic Mutations In Cancer”. In: *Nucleic Acids Research* 47.D1 (Oct. 2018), pp. D941–D947. doi: 10.1093/nar/gky1015.

[210] Wellcome Sanger Institute. *COSMIC database v95*. Nov. 14, 2021. url: https://cancer.sanger.ac.uk/.

[211] R. Barroso-Sousa et al. “Prevalence and mutational determinants of high tumor mutation burden in breast cancer”. In: *Annals of Oncology* 31.3 (Mar. 2020), pp. 387–394. doi: 10.1016/j.annonc.2019.11.010.

[212] Federico Abascal et al. “Somatic mutation landscapes at single-molecule resolution”. In: *Nature* 593.7859 (Apr. 2021), pp. 405–410. doi: 10.1038/s41586-021-03477-4.

[213] Julian R. Homburger et al. “Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores”. In: *Genome Medicine* 11.1 (Nov. 2019). doi: 10.1186/s13073-019-0682-2.

[214] Ming Chen et al. “Applying low coverage whole genome sequencing to detect malignant ovarian mass”. In: *Journal of Translational Medicine* 19.1 (Aug. 2021). doi: 10.1186/s12967-021-03046-3.

[215] Mark A DePristo et al. “A framework for variation discovery and genotyping using next-generation DNA sequencing data”. In: *Nature Genetics* 43.5 (Apr. 2011), pp. 491–498. doi: 10.1038/ng.806.

[216] Berke Ç Toptaş et al. “Comparing complex variants in family trios”. In: *Bioinformatics* (June 2018). Ed. by Oliver Stegle. doi: 10.1093/bioinformatics/bty443.

[217] Di Wang et al. “Multiregion Sequencing Reveals the Genetic Heterogeneity and Evolutionary History of Osteosarcoma and Matched Pulmonary Metastases”. In: *Cancer Research* 79.1 (Nov. 2018), pp. 7–20. doi: 10.1158/0008-5472.can-18-1086.

[218] Zixi Chen et al. “Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency”. In: *Scientific Reports* 10.1 (Feb. 2020). doi: 10.1038/s41598-020-60559-5.

[219] Richard Lupat et al. *Janis: A Python framework for Portable Pipelines*. en. 2021. doi: 10.5281/ZENODO.4427231.

[220] H. Li and R. Durbin. “Fast and accurate short read alignment with Burrows-Wheeler transform”. In: *Bioinformatics* 25.14 (May 2009), pp. 1754–1760. doi: 10.1093/bioinformatics/btp324.

[221] Kenneth D. Doig et al. “Canary: an atomic pipeline for clinical amplicon assays”. In: *BMC Bioinformatics* 18.1 (Dec. 2017). doi: 10.1186/s12859-017-1950-z.

[222] Thomas Derrien et al. “Fast Computation and Applications of Genome Mappability”. In: *PLoS ONE* 7.1 (Jan. 2012). Ed. by Christos A. Ouzounis, e30377. doi: 10.1371/journal.pone.0030377.

[223] Valerie A. Schneider et al. “Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly”. In: *Genome Research* 27.5 (Apr. 2017), pp. 849–864. doi: 10.1101/gr.213611.116.

[224] Deanna M. Church et al. “Modernizing Reference Genome Assemblies”. In: *PLoS Biology* 9.7 (July 2011), e1001091. doi: 10.1371/journal.pbio.1001091.

[225] H. Li. “Tabix: fast retrieval of sequence features from generic TAB-delimited files”. In: *Bioinformatics* 27.5 (Jan. 2011), pp. 718–719. doi: 10.1093/bioinformatics/btq671.

Appendices

by Sebastian Hollizeck

ORCID: 0000-0002-9504-3497

contains:

- published manuscripts
- supplementary method
- supplementary figures

A

Custom workflows to improve joint variant calling from multiple related tumour samples: FreeBayesSomatic and Strelka2Pass

This appendix contains the manuscript published at *Bioinformatics* in an non journal style format with the supplementary methods and figures. It can also be found at 10.1093/bioinformatics/btab606/6361543 for a paper style version.

Hollizeck S.[1,2], **Wong S.Q.**[1,2], **Solomon B.**[1,2], **Chandrananda D.**[1,2,*], **and Dawson S-J.**[1,2,3,*]

[1] Peter MacCallum Cancer Centre, Melbourne 3000, Victoria, Australia

[2] Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne 3000, Victoria, Australia

[3] Centre for Cancer Research, University of Melbourne, Melbourne 3000, Victoria, Australia

[*] D.C and S.J.D are co-senior authors and contributed equally to this article

Received on 27-Jan-2021; revised on 13-Jul-2021; accepted on 12-Aug-2021

Abstract

Summary: This work describes two novel workflows for variant calling that extend the widely used algorithms of Strelka2 and FreeBayes to call somatic mutations from multiple related tumour samples and one matched normal sample. We show that these workflows offer higher precision and

recall than their single tumour-normal pair equivalents in both simulated and clinical sequencing data.

Availability and Implementation: Source code freely available at the following link: https://atlassian.petermac.org.au/bitbucket/projects/DAW/repos/multisamplevariantcalling and executable through Janis (https://github.com/PMCC-BioinformaticsCore/janis) under the GPLv3 licence.

Contact: Dineika.Chandrananda@petermac.org, Sarah-Jane.Dawson@petermac.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

A.1 Introduction

Joint variant calling methods are routinely used to call germline variants by leveraging population-wide information across multiple related samples [215, 216]. This concept is also advantageous for somatic variant calling to potentially overcome the challenges of spatial heterogeneity and low tumour purity. However, there is a critical lack of robust algorithms that allow multi-sample somatic calling. Most studies still rely on variant calling of separate tumour-normal pairs, subsequently combining the results across a sample cohort [166, 3, 217].

There are two major pitfalls for combining variants called from individual tumour samples. First, it is very difficult to differentiate between a false negative result due to "missing data" versus the true absence of a variant. Second, there is limited sensitivity for low allele frequency variants thus, decreasing the ability to detect minor clones, particularly in samples with low tumour purity.

Currently, only three algorithms claim to have the functionality to jointly analyse multiple samples: multiSNV [156], SuperFreq [157], and Mutect2 [107], each presenting different limitations. For instance, multiSNV cannot call indels and along with SuperFreq, is not optimised for analysis of deep coverage whole-genome sequencing (WGS) data. Mutect2 has previously been shown to be disadvantageously conservative as well as computationally inefficient [218].

To enable highly sensitive, fast and accurate variant detection from multiple related tumour samples, we have developed joint variant calling extensions to two widely used single-sample algorithms, FreeBayes [104] and Strelka2 [106]. Using both simulated and clinical sequencing data, we show that these workflows are highly accurate and can detect variants at much lower variant allele frequencies than commonly used methods.

A.2 Materials and methods

A.2.1 FreeBayesSomatic workflow

The original FreeBayes algorithm can jointly evaluate multiple samples but routinely it does not perform somatic variant calling on tumour-normal pairs. We introduce FreeBayesSomatic which allows concurrent analysis of multiple tumour samples by adapting concepts from SpeedSeq [159] which differentiates the likelihood of a variant between tumour and normal samples instead of imposing an absolute filter for all variants called in the normal. Hence, for each genotype (GT) at SNV sites, FreeBayesSomatic first calculates the difference in likelihoods (LOD) between the normal (Equation A.1) and the tumour (Equation A.2) samples genotype likelihoods (GL) with g_0 describing the reference genotype.

$$\mathrm{LOD}_{\mathrm{normal}} = \max_{g_i \in \mathrm{GT}} \left(\mathrm{GL}(g_0) - \mathrm{GL}(g_i)\right) \tag{A.1}$$

$$\mathrm{LOD}_{\mathrm{tumour}} = \min_{s \in \mathrm{Samples}} \left(\min_{g_i \in \mathrm{GT}} \left(\mathrm{GL}_s(g_i) - \mathrm{GL}_s(g_0)\right) \right) \tag{A.2}$$

$$\mathrm{somaticLOD} := \left(\mathrm{LOD}_{\mathrm{normal}} \geq 3.5 \wedge \mathrm{LOD}_{\mathrm{tumour}} \geq 3.5\right) \tag{A.3}$$

Next, the variant allele frequencies (VAF) in both the tumour and the normal samples are compared at each site.

$$\mathrm{VAF}_{\mathrm{tumour}} = \max_{s \in \mathrm{Samples}} \left(\mathrm{VAF}_s\right) \tag{A.4}$$

$$\begin{aligned} \mathrm{somaticVAF} := &\left(\mathrm{VAF}_{\mathrm{normal}} \leq 0.001 \vee \right. \\ &\left. \left(\mathrm{VAF}_{\mathrm{tumour}} \geq 2.7 \cdot \mathrm{VAF}_{\mathrm{normal}}\right)\right) \end{aligned} \tag{A.5}$$

A variant is classified as somatic when both somaticLOD as well as somatic VAF pass the criteria somaticLOD (Equation A.3) and somaticVAF (Equation A.5).

The thresholds chosen for both LOD and VAF calculations were previously fitted by the blue-collar bioinformatics workflow for the DREAM synthetic 3 dataset using the SpeedSeq likelihood difference approach [160] and were selected to identify high confidence variants.

A.2.2 Strelka2Pass workflow

In contrast to FreeBayes, whilst Strelka2 has a multiple-sample mode for germline analysis and tumour-normal pair somatic variant calling capabilities, it cannot jointly analyse multiple related tumour samples. We enable this feature by adapting a two-pass strategy previously used for RNA-seq data [161]. First, somatic variants are called from each tumour-normal pair. All detected variants across the cohort are then used as input for the second pass of the analysis where we re-iterate through each tumour-normal pair but assess allelic information for all input genomic sites.

The method re-evaluates the likelihood of each variant, by integrating every genotype from each tumour-normal pair. This step can "call" a variant (v) in a sample that initially did not present enough evidence to pass the Strelka2 internal filtering using two conditions: 1) if this variant was called as a proper "PASS" by Strelka2 in any other tumour sample, or 2) if the integrated evidence for this variant across all tumour-normal pairs reached a sufficiently high level. The second condition was based on the somatic evidence score (SomEVS) reported by Strelka2, which is the logarithm of the probability of the variant v being an artefact.

$$p_{error}(v) = 10^{\left(\frac{-\text{SomEVS}(v)}{10}\right)} \tag{A.6}$$

While the germline sample is shared between all processes, we can approximate these individual probabilities as being independent, since one variant calling process is agnostic of the other. Hence, we derive the following:

$$p_{error}(v_{s_1}, v_{s_2}, \ldots, v_{s_n}) = \prod_{s \in \text{Samples}} p_{error}(v_s) \tag{A.7}$$

And therefore:

$$\text{SomEVS}(v_{s_1}, v_{s_2}, \ldots, v_{s_n}) = \sum_{s \in \text{Samples}} \text{SomEVS}(v_s) \tag{A.8}$$

This allows the summation (Equation A.8) of the SomEVS score across all supporting variants to assign a "PASS" filter, if it reached a joint SomEVS score threshold. This threshold can be set by the user and is 20 by default, which corresponds to an estimated error rate of 1%. These "recovered" variants

need to pass a set of additional quality metrics related to depth of coverage, mapping quality and read position rank sum score.

As an additional improvement, we also built multiallelic support into Strelka2 which originally only reports the most prevalent variant at a specific site. Within the two-pass analysis, we reconstruct the available evidence for a multiallelic variant at a called site from the allele-specific read counts and report the minor allele at this site, if there is sufficient support from other samples. This method allows recovery of minor alleles only if another sample has this variant called by Strelka2, as SomEVS scores are not available for minor alleles.

A.3 Validation

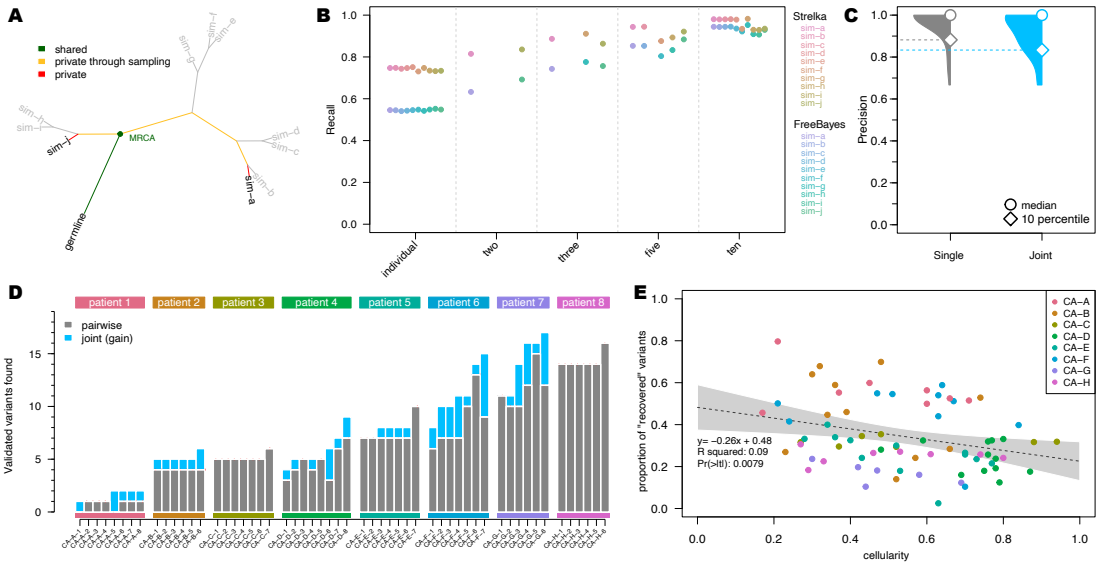

Figure A.1: Comparison of joint multi-sample variant calling and single tumour-normal paired calling methods; A) Simulated phylogeny highlighting two samples with high evolutionary distance (sim-a and sim-j) where MRCA denotes the most recent common ancestor. B) Recall estimates of FreeBayes and Strelka2, run in individual tumour-normal paired and joint calling configurations using two (sim-a and sim-j), three (sim-a, sim-g and sim-j), five (sim-a, sim-c, sim-f, sim-h and sim-j) and all ten tumour samples. C) Precision of Strelka2 and D) Number of variants called by Strelka2 run in both tumour-normal paired (grey) and added with joint calling configurations (blue), which have been validated by targeted amplicon sequencing (TAS). E) Correlation between cellularity and proportion of variants found only with joint calling using Strelka2Pass for clinical samples; grey area shows the "95%" confidence interval for the linear model fit (dotted line).

A.3.1 Simulated data

We first simulated a phylogeny with somatic and germline variants from ten tumour samples and one normal (Figure A.1A, Figure A.2A, B) (Section A.5). Germline variants were simulated at a uniform allele frequency of 0.5. Somatic VAFs were sampled from a custom distribution, modelled to favour low allele frequency variants to closely represent real world data (min VAF: 0.001; max VAF: 1; Figure A.2C, D). Paired-end sequencing reads with realistic error profiles were simulated for WGS data at 160X average coverage using the ART-MountRainier software [162]. The simulated reads were aligned to GRCh38 and both germline and somatic variants from the phylogeny were spiked into the aligned reads using Bamsurgeon [163]. We compared the workflows for FreeBayes and Strelka2 with and without our extensions for joint variant calling on the simulated datasets. The performance of Mutect2 joint variant calling was also assessed using its proposed best practice workflow. As both Mutect2 and FreeBayes do not return a verdict for each individual sample, we needed to assign each sample in the multi-sample VCF its own FILTER value. We called a somatic variant as present in a sample, if there were at least two reads supporting it for this sample and the overall FILTER showed a "PASS", which was the same cut-off used in the refiltering step in the Strelka2-pass workflow.

While the precision of each method without our extensions was greater than 99.8%, they all missed at least 25% of all variants in the samples (i.e recall $\leq 75\%$). In contrast, the recall of the modified workflows increased to $\approx 95\%$ with only a minute decrease in the precision for both FreeBayes and Strelka2 (Figure A.3). Mutect2 however, had virtually no change in precision, but the recall actually decreased from $\approx 75\%$ to $\approx 41\%$ when analysing the samples jointly (Figure A.3B). Additionally, with our modified workflows, true positive variants were called with VAFs as low as 0.008 (median detected VAF ≥ 0.14 for joint sample analysis and ≥ 0.21 for single tumour-normal pair analysis), enabling improved distinction between true variants and technical errors (Figure A.4). This improvement in performance for Strelka2 is only achieved after the refiltering step and not just a result of the second pass (Figure A.5) (Section A.5.4).

The performance of joint variant calling in Mutect2 was inferior compared to all other methods (Figure A.3A, B). This was primarily due to the "clustered_events" filter in Mutect2, which excluded the majority of false negative variants, with negligible contribution to the exclusion of true negative variants (Figure A.6A, B). This result was unexpected as the simulated variants were evenly distributed along the genome and the corresponding allele frequencies were sampled randomly (Figure A.2D).

Since the extent of the improvement in our joint calling workflows is bound by the number of shared variants between samples, we sub-sampled the simulated dataset, to show the effect of incomplete sampling on our methods, which is more likely in clinical settings. Furthermore, the evolutionary distance between the related samples in addition to the number of samples, has a major impact on the number of shared variants, as only variants acquired between the germline and the most recent common ancestor (MRCA), will benefit from the joint analysis. Therefore, we selected three sample subsets which included two, three and five samples with high evolutionary distance to show the minimum expected improvement (Figure A.1A, B). There was a clear linear improvement for both FreeBayesSomatic and Strelka2Pass when increasing the number of samples even if they had a distant evolutionary relationship. In contrast, when using only two samples with a small evolutionary distance, the increase in performance was almost as large as when jointly analysing all 10 available samples. This shows that samples with a high number of shared variants will perform better in joint calling workflows (Figure A.7).

A.3.2 Clinical data

To validate the performance of our new workflows, we then analysed WGS and whole-exome sequencing (WES) data of multi-region tumour samples from eight patients, with multiple tumour sites (average 7 samples per patient; total number of samples 55), enrolled in a rapid autopsy program conducted at the Peter MacCallum Cancer Centre (Table A.1 and Section A.5) [164, 165]. The published studies had multiple somatic variants from the clinical samples orthogonally validated through targeted amplicon sequencing (TAS). We used these TAS-validated variants as the gold standard to evaluate the performance of different workflows, acknowledging that the technical biases inherent to TAS data are different to those present in WGS and WES (Figure A.8) and that there would be sampling biases depending on different tumour cells analysed in each data type.

In concordance with the results of the simulated data, our improved workflows found additional variants in all but one patient (Figure A.1D, Figure A.9) (total additional variants Strelka2Pass: 64; FreeBayesSomatic: 85) with only a slight drop in precision for FreeBayesSomatic (mean: 0.94 vs. 0.88) and Strelka2Pass (mean: 0.97 vs. 0.92). Since the panel of variants validated by TAS was limited (7108 bp for patients CA-B through -H), this increase in detected variants suggests that a high number of shared variants in samples are missed with current approaches, which in turn leads to an overestimation of tumour heterogeneity between samples, as these variants are thought to not be present rather than undetected.

Even though the number of shared variants is a major influencing factor when jointly calling variants, low cellularity samples benefit more from the joint calling, as conventional methods cannot reliably distinguish low allele frequency variants from noise. Through a joint analysis approach, the number of recovered variants is higher in low cellularity samples, which indicates, that especially for clinical samples with variable tumour purity, joint analysis can have a major impact on improving performance (Figure A.1E, Figure A.10).

Mutect2 in contrast, did not show significant improvement in any sample in its joint calling configuration, but showed inferior performance compared to the tumour-normal pairwise approach in two samples (Figure A.9E), similar to its decreased performance in the simulated data (Figure A.3). This was due to true variants being removed by the internal filters of the tool (Figure A.6C, D). This is in stark contrast to our novel workflows, where the joint analysis preserves all called sites from the pairwise method and finds additional variants. Overall, Mutect2 found less validated variants in all patients than both Strelka2Pass (mean: 2.2) and FreeBayesSomatic (mean: 2.5) with comparable levels of precision (Figure A.9, Figure A.11) but longer run times (Table A.2).

Our improved workflow also enabled the discovery of multiallelic variants with Strelka2, which led to the discovery of on average 42 additional variants (min: 1; max: 535) in the analysed WES and 987 additional variants in the WGS (min: 81; max 2329). These variants are strong indicators of sub clonal structure and could be invaluable for the study of evolutionary trajectories in cancer.

A.4 Discussion

Here we present an extension to two widely used variant callers, enabling them to analyse multiple related tumour samples and improve the sensitivity of detecting low allele frequency variants. This is highly relevant in clinical settings where low tumour purities in samples is a common occurrence. These workflows are an important step to satisfy the current unmet need for multi-sample tumour variant calling. While we have showcased their improvements in patient sequencing data, additional validation on larger clinical datasets is warranted to ensure the methodology performs robustly in real world settings. Importantly, these workflows are fully containerised and can be run through Janis [219] on almost any high-performance computing environment, as well as cloud services. Each workflow is highly optimised and parallelised to facilitate the analysis of the large

amount of data joint variant calling requires. The workflow specification also allows the easy adjustment of parameters to enable customisation for the user's needs and priorities, whereas building an ensemble workflow using multiple callers is up to the discretion of the user (Figure A.12).

Acknowledgements

The authors would like to thank all patients who provided tissue samples utilised in this study. The authors acknowledge Dr Lavinia Tan for assistance provided with the collection of patient clinical samples.

Funding

This work was supported by the National Health and Medical Research Council [grant numbers 1196755 to S.J.D, 1158345 to S.J.D and B.J.S, 1194783 to S.Q.W, 1173450 to B.J.S]; and CSL Centenary Fellowship to S.J.D; Victorian Cancer Agency [grant numbers 19008 to D.C, 19002 to S.Q.W]

Conflicts of Interest

S.J.D has been a member of advisory boards for AstraZeneca and Inivata. The S.J.D. lab has received funding from Cancer Therapeutics CRC and Roche-Genentech. B.J.S. has been a member of advisory boards for AstraZeneca, Roche-Genentech, Pfizer, Novartis, Amgen, Bristol Myers Squibb and Merk

Data availability

The simulated data and the respective final variant calling files underlying this article are available from Figshare at https://melbourne.figshare.com, and can be accessed with https://doi.org/10.26188/13635186 for the dataset and https://doi.org/10.26188/13635187 for the called variants.
The biological data underlying this article are available at the European Genome-Phenome Archive (EGA) at https://ega-archive.org, and can be accessed with study id EGAS00001004023 and EGAS00001004950.

Supplementary data

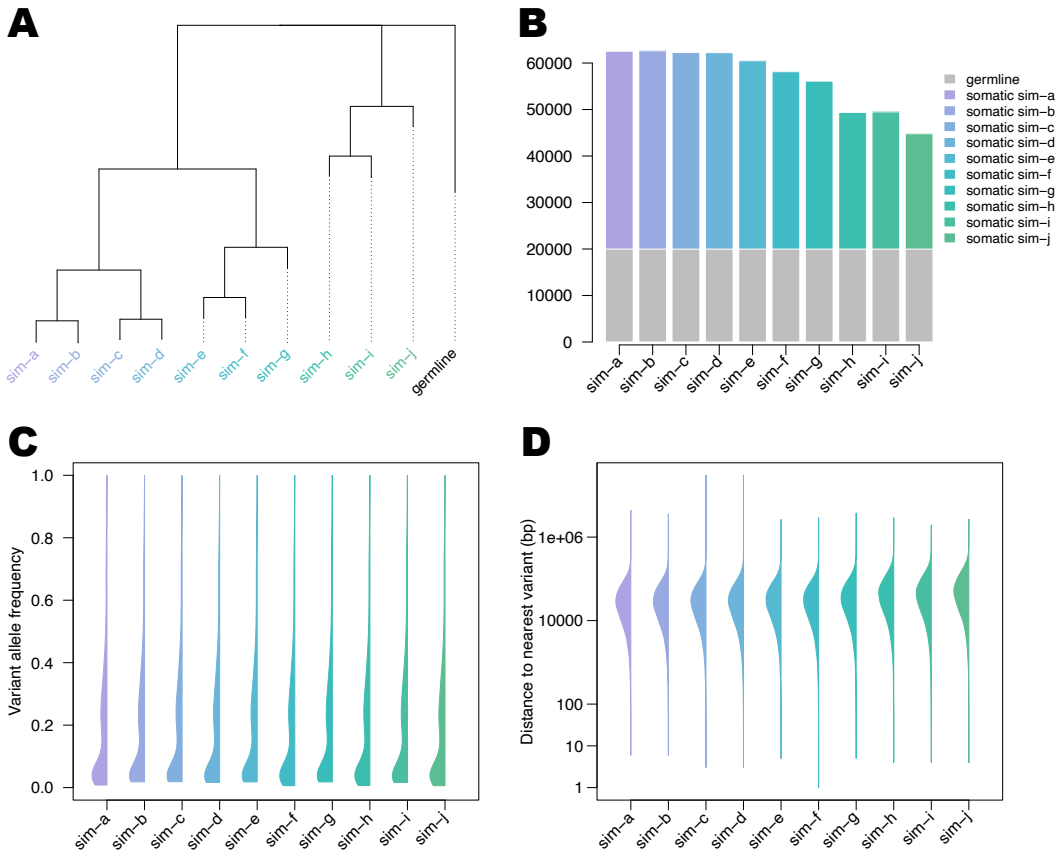

Figure A.2: Characteristics of simulated data: A) Simulated phylogeny of samples B) Number of simulated germline and somatic variants per sample C) Variant allele frequency distribution of simulated variants per sample D) Distance to nearest variant in each sample.

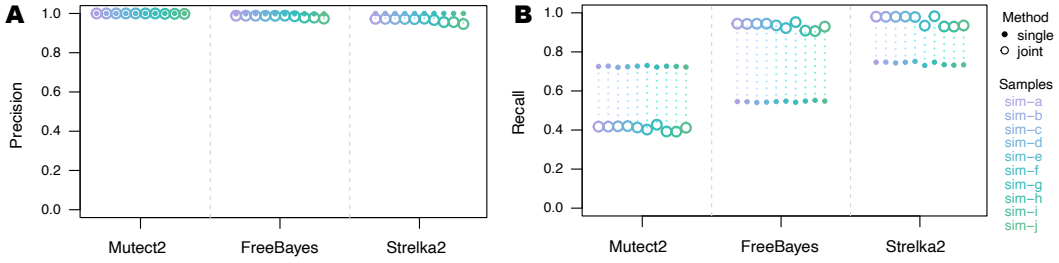

Figure A.3: Performance of workflows using simulated data: A) Precision and B) Recall of Mutect2, FreeBayes and Strelka2, run in single tumour-normal paired and joint calling configurations.

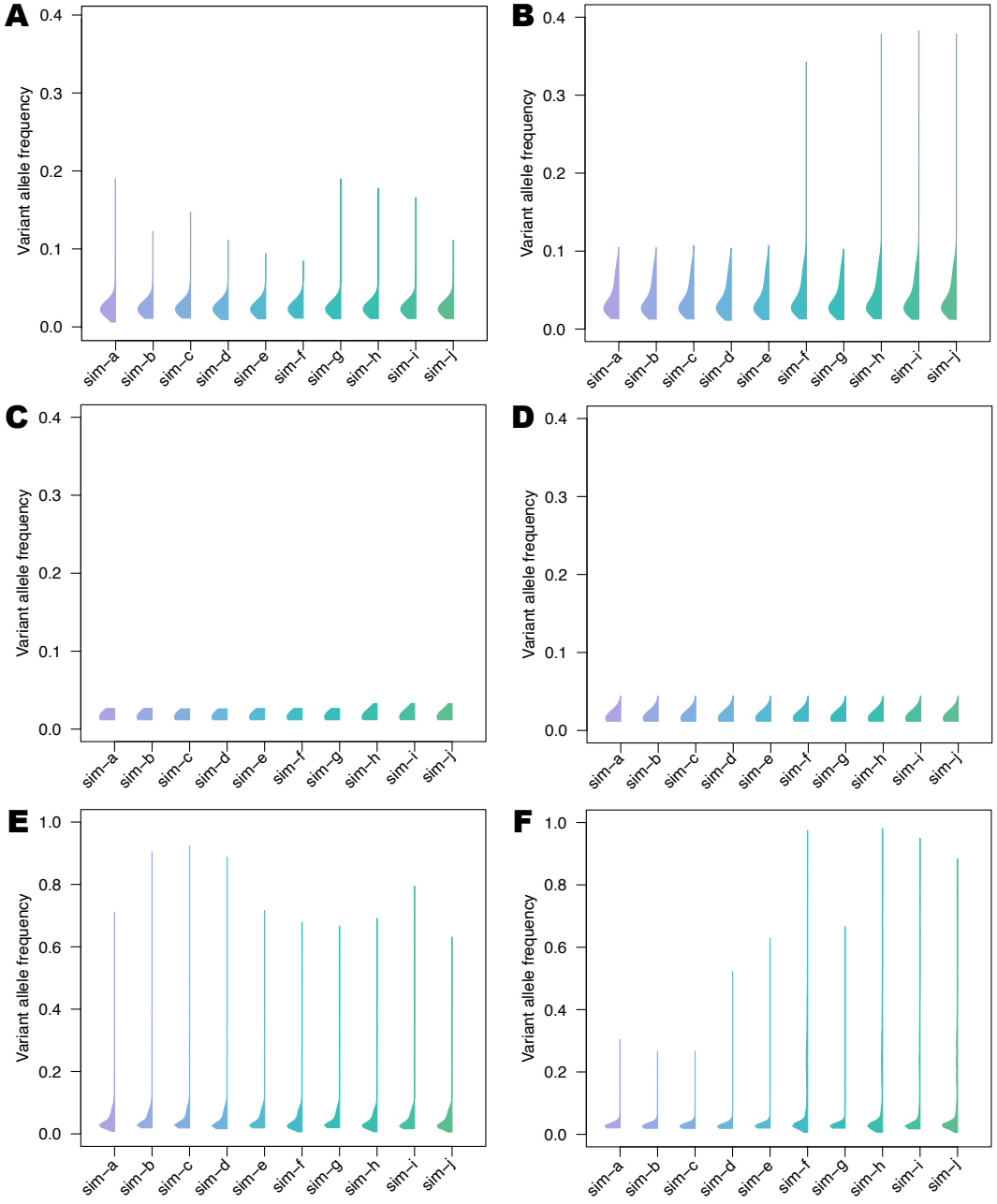

Figure A.4: Variant allele frequencies (VAF) of variants detected by joint sample analysis; A) VAF distribution of true positive variants additionally detected by Strelka2pass B) and FreeBayesSomatic C) VAF distribution of false positive variants additionally detected by FreeBayesSomatic D) and Strelka2pass E) VAF distribution of false negatives not called by FreeBayesSomatic F) and Strelka2pass.

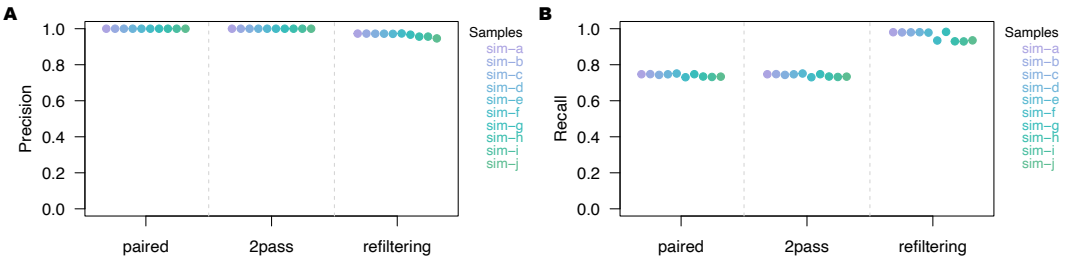

Figure A.5: Performance of individual steps in the Strelka2pass workflow using the simulated data: A) Precision and B) Recall of tumour-normal paired analysis, two-pass step without refiltering (supplying variants from all tumour-normal pairs for evaluation) and two-pass step with refiltering (the final workflow)

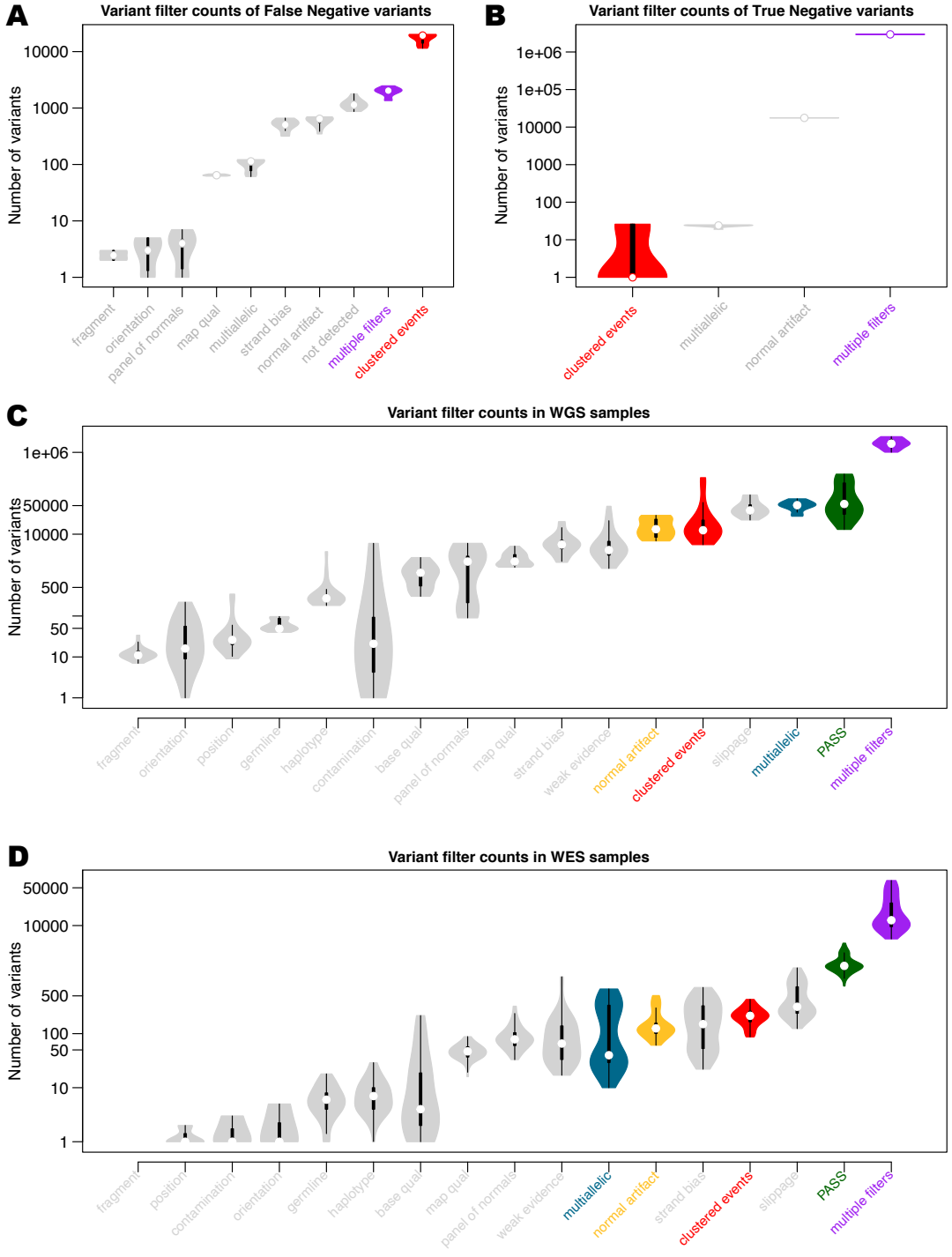

Figure A.6: Summary of variant filters assigned by Mutect2; The counts for each filter type are denoted by black boxplots with white circles depicting the median values. The fitted distribution of variant counts outlines each boxplot; A) Counts of filter assignments for false negative variants and B) true negative variants called by Mutect2 C) Filter assignment for all variants reported for sequenced patient data sequenced with WGS or D) WES.

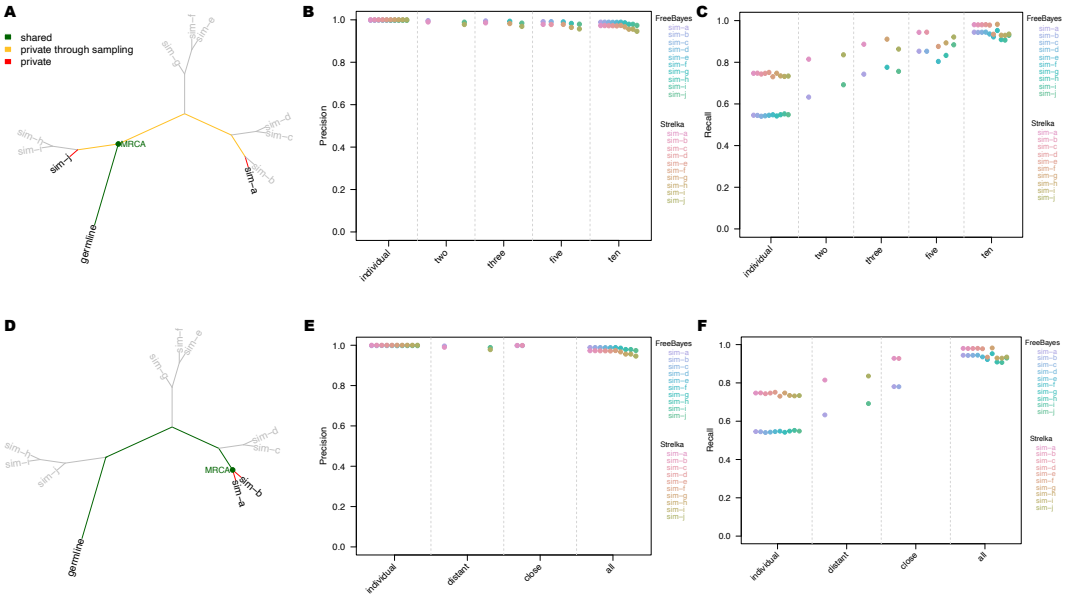

Figure A.7: Assessing the performance of different workflows using tumour samples with different evolutionary relationships in the simulated data; A) Simulated phylogeny highlighting two samples with high evolutionary distance (sim-a and sim-j) where MRCA denotes the most recent common ancestor. B) Precision and C) Recall estimates of FreeBayes and Strelka, run in individual tumour-normal paired and joint calling configurations using two (sim-a and sim-j), three (sim-a, sim-g and sim-j), five (sim-a, sim-c, sim-f, sim-h and sim-j) and all ten tumour samples D) Simulated phylogeny highlighting two samples with low evolutionary distance (sim-a and sim-b). E) Precision and F) Recall estimates for FreeBayes and Strelka run in individual tumour-normal paired and joint calling configurations. The plots compare the performance of these workflows when using two evolutionary distant samples (sim-a and sim-j), two evolutionary close samples (sim-a and sim-b) and all ten tumour samples.

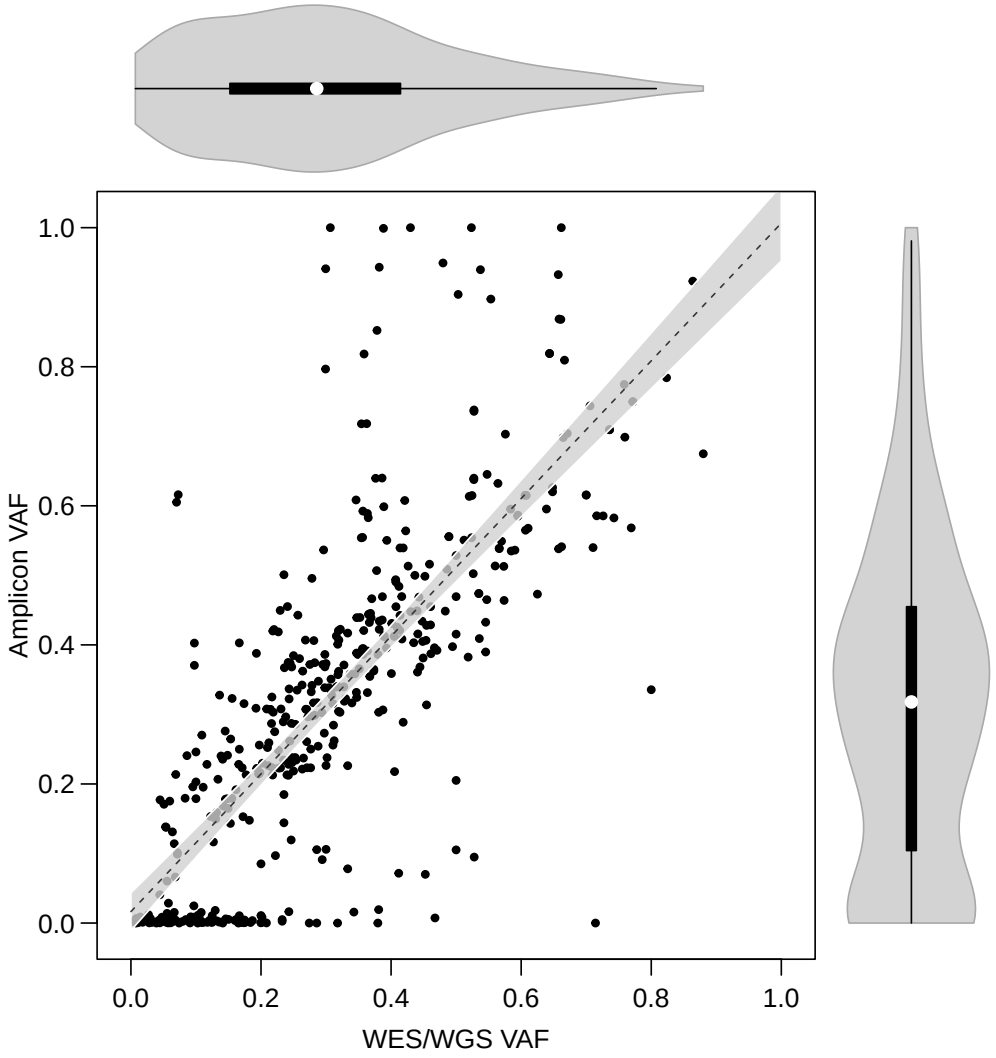

Figure A.8: Correlation of variant allele frequencies (VAF) from WES and WGS data against targeted amplicon sequencing VAF values with fitted violin plots of each individual distribution. Grey background shows 95% confidence interval for the fit of the linear model (dotted line).

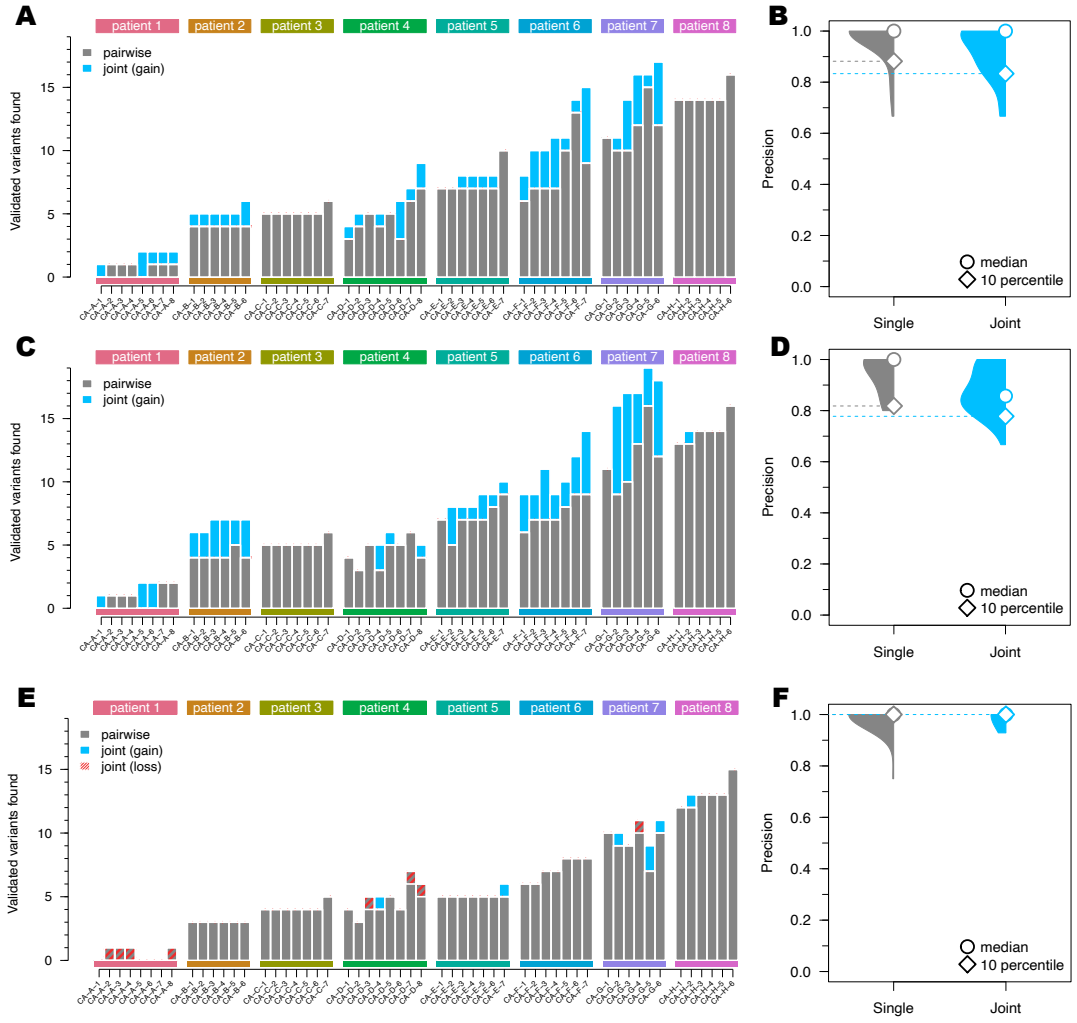

Figure A.9: Performance of the different workflows using clinical samples from eight cancer patients: A) Number of variants called by Strelka2 run in the tumour-normal paired (grey) and joint calling configurations, which have been validated by targeted amplicon sequencing (TAS). The same for C) FreeBayes and E) Mutect2 workflows. Precision of tumour-normal paired and joint analysis of TAS validated clinical data for B) Strelka2, D) FreeBayes and F) Mutect2; Sup. Table 1 provides the sample naming map to the original publications.

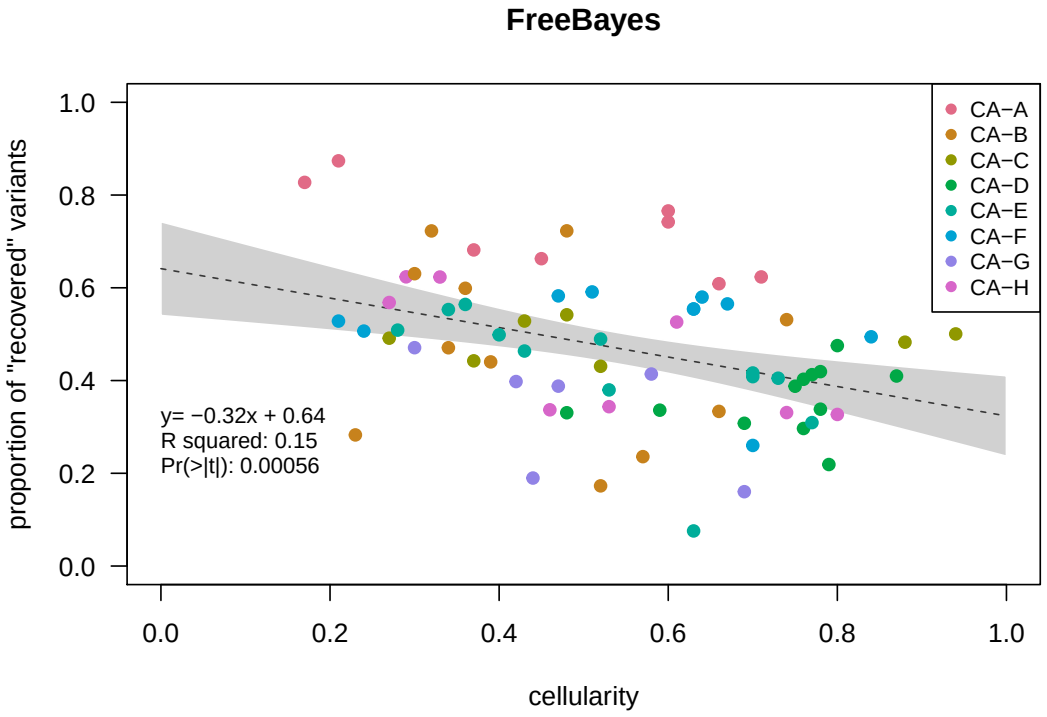

Figure A.10: Correlation between cellularity and proportion of variants found only with joint calling using FreeBayesSomatic. Grey background shows 95% confidence interval for fit of linear model (dotted line)

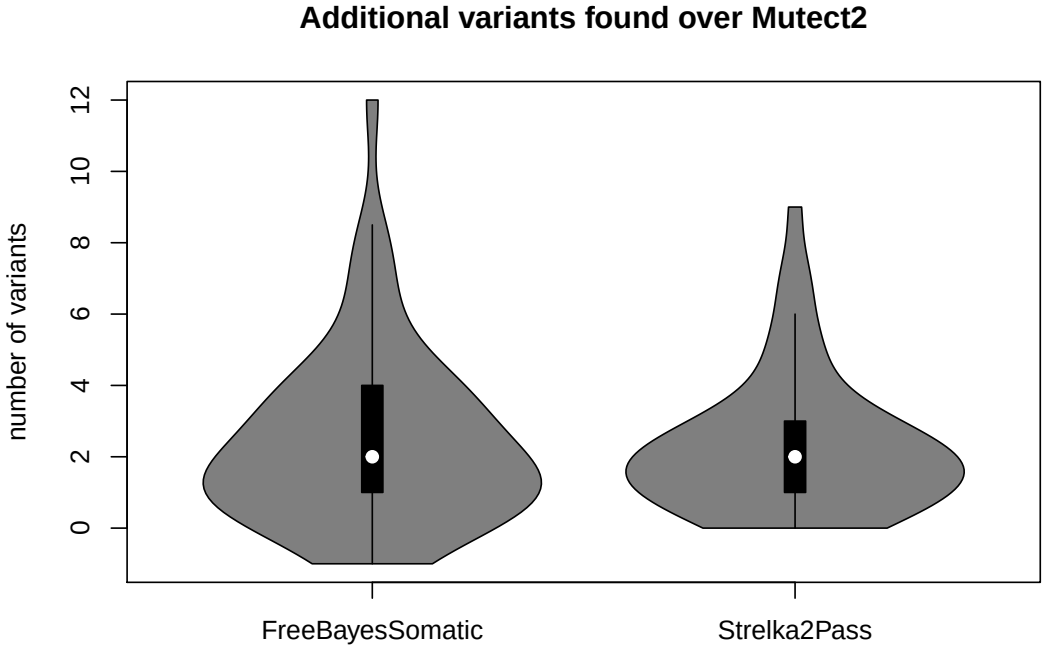

Figure A.11: Improvement in recall using FreeBayesSomatic and Strelka2pass over Mutect2 in the clinical samples.

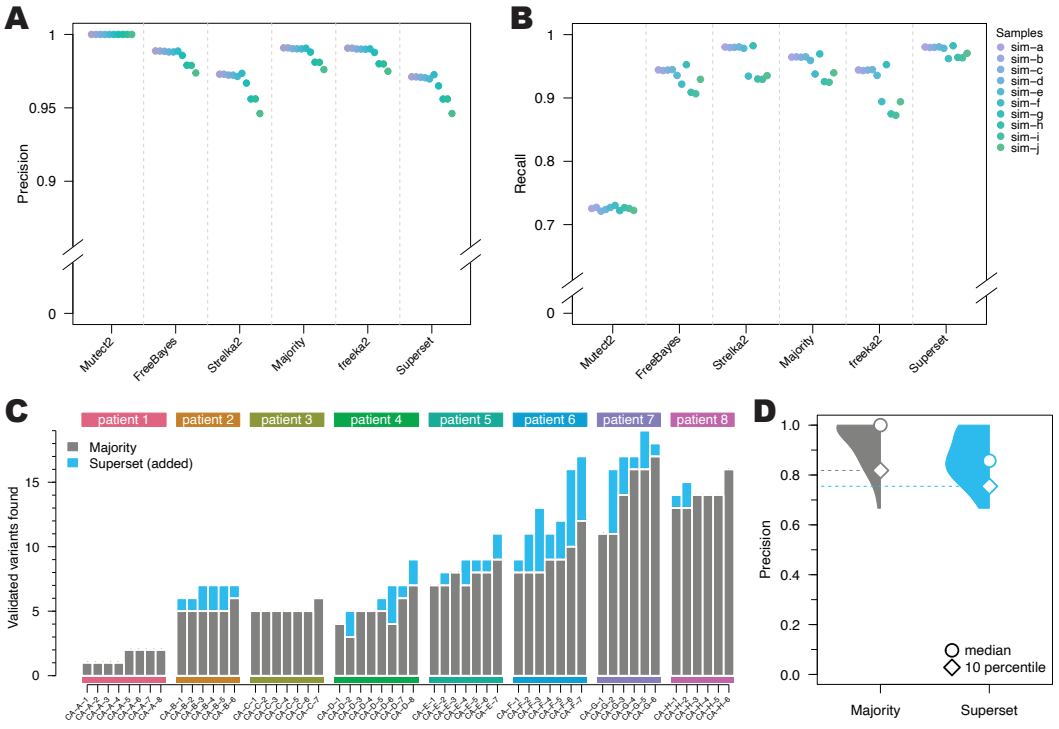

Figure A.12: Performance of ensemble variant calling strategies. A) Precision and B) Recall of variant detection using the joint multi-sample calling of each tool separately and compared to using Majority-vote ensemble calling (variant is called by at least two callers), Freeka2 (variant is called by both FreeBayesSomatic and Strelka2pass) and Superset (variant is called by either FreeBayesSomatic or Strelka2pass) for the simulated dataset D) Number of TAS validated variants found in the clinical samples with Majority-vote and Superset methods and the corresponding D) Precision estimates.

Table A.1: Sample naming map relating to previously published datasets. The first column contains sample names as they appear in this work, and the third column denotes how the samples are referred to in the original studies. Forth column shows the type of sequencing WES: whole-exome sequencing; WGS: whole genome sequencing.

SAMPLE NAME	PUBLISHED STUDY	ORIGINAL NAME	SEQUENCING TYPE
CA-A-1	Solomon et al. [164]	Case 1 Left liver 1	WGS
CA-A-2		Case 1 Right occipital	
CA-A-3		Case 1 Right liver 2	
CA-A-4		Case 1 Right pleura	
CA-A-5		Case 1 Left lower lung lobe	
CA-A-6		Case 1 Left liver 5	
CA-A-7		Case 1 Right liver 3	
CA-A-8		Case 1 Left liver 2	
CA-B-1	Vergara et al. [165]	CAS-B-21-L-LUNG	WES
CA-B-2		CAS-B-22-R-LUNG	
CA-B-3		CAS-B-14B37035-1B	
CA-B-4		CAS-B-Primary-1	
CA-B-5		CAS-B-15B08317-3A	
CA-B-6		CAS-B-14B37035-1C	
CA-C-1		CAS-A-FR07935894	WGS
CA-C-2		CAS-A-FR07935905	
CA-C-3		CAS-A-FR07935906	
CA-C-4		CAS-A-FR07935907	
CA-C-5		CAS-A-FR07935908	
CA-C-6		CAS-A-FR07935916	
CA-C-7		CAS-A-FR07935918	
CA-D-1		CAS-G-91-2	WES
CA-D-2		CAS-G-75	
CA-D-3		CAS-G-74	
CA-D-4		CAS-G-71	
CA-D-5		CAS-G-91	
CA-D-6		CAS-G-76	
CA-D-7		CAS-G-94	
CA-D-8		CAS-G-72	
CA-E-1		CAS-D-70	WES
CA-E-2		CAS-D-61-3	
CA-E-3		CAS-D-66	
CA-E-4		CAS-D-68	
CA-E-5		CAS-D-64	
CA-E-6		CAS-D-61-2	
CA-E-7		CAS-D-62	
CA-F-1		CAS-C-41	WES
CA-F-2		CAS-C-40-Fresh	
CA-F-3		CAS-C-37	
CA-F-4		CAS-C-44	
CA-F-5		CAS-C-42-Fresh	
CA-F-6		CAS-C-43-Fresh	
CA-F-7		CAS-C-46-Primary	
CA-G-1		CAS-F-FR07935922	WGS
CA-G-2		CAS-F-FR07935915	
CA-G-3		CAS-F-FR07935913	
CA-G-4		CAS-F-FR07935909	
CA-G-5		CAS-F-FR07935904	
CA-G-6		CAS-F-FR07935903	
CA-H-1		CAS-E-1	WES
CA-H-2		CAS-E-3	
CA-H-3		CAS-E-4	
CA-H-4		CAS-E-10	
CA-H-5		CAS-E-6	
CA-H-6		CAS-E-8	

Table A.2: Runtime of different workflows on simulated data; The runtimes were generated on the Peter MacCallum Cancer Centre HPC cluster with Intel(R) Xeon(R) CPU E5-2660 v3 @ 2.60GHz. The times are displayed in single CPU runtime, but each workflow is highly parallelised, such that the user runtime is far lower.

	Number of tumour samples used for joint calling			
Method	2	3	5	10
FreeBayesSomatic	562h	811h	1185h	2292h
Strelka2Pass	310h	465h	776h	1552h
Mutect2	-	-	-	28 418h

A.5 Supplementary methods

A.5.1 Alignment of clinical data

Detailed information on processing of the clinical sequencing datasets was published previously [164, 165]. Briefly, reads were aligned to GRCh38 for patient CAS-A and GRCh37 for patients CAS-B through CAS-H using BWA version 0.7.17 [220] allowing the use of alternative contigs. Reads were then marked as duplicates with Picard software (v2.17.3).

A.5.2 Validation of clinical data

Detailed information on targeted amplicon sequencing of patient samples can be found in the original publications [164, 165]. A SNV called in WES with any workflow was considered a true positive when the adjusted p-value calculated through an exact binomial test was lower than 0.05 on the TAS data. The probability of success for this test was estimated as the number of bases different from the reference divided by the total number of sequenced bases (0.001) and the number of trials was the read depth covering the variant. For indels, a variant was considered to be validated if either of the panel variant callers primal (in house) or canary [221] called the same variant.

Only amplicons with an average mapping rate of at least 80% over all samples, as well as an average coverage of more than 300 were considered for further analysis. WES variants were first subsetted to be within the area of the respective amplicons.

A.5.3 Purity estimation with sequenza

For CA-A the sequenza-utils python program was used to generate input files for the sequenza R program on the aligned BAM files [13]. Kmin and gamma were set to 100 and 500 respectively to discourage a highly fragmented result. For CA-B through -H the reported tumour purities were used from the publication [165].

A.5.4 Performance of individual steps in Strelka2Pass

As each of the three steps potentially has implications for the performance, we assessed the improvement provided by each step in the Strelka2pass workflow. Figure A.5 shows, that there is no change in either precision or recall just by supplying variants from all tumour-normal pairs for a second round of evaluation. However, there is a >20% improvement in recall when coupling this to the refiltering step that we have built into the workflow.

A.5.5 Ensemble workflows – user suggestions

An overall workflow can contain any number of additional variant callers, when not restricted to callers with joint analysis capability. Importantly, there is no benefit of jointly analysing samples with Mutect2, and it may decrease the performance in some cases. Each of our presented workflows outperformed Mutect2 on the data shown here, so when assembling an ensemble method, these methods, should have a higher confidence assigned to them in joint analysis cases, than tumour-normal pair approaches.

Depending on the end needs of the user, an ensemble workflow can be optimised towards precision or recall. In Figure A.12 we show the performance changes improvement that can be achieved by combining Mutect2 in tumour-normal paired analysis with the two new workflows FreeBayesSomatic and Strelka2Pass. First, in a "best of three" majority vote, where the variant needs to be called by two out of three variant callers, we enhance the precision of each of the individual tools, with slightly lower recall. On the other hand, with the super set approach, where any variant called in either FreeBayesSomatic or Strelka2Pass is included in the end result, this improves the recall even further, but slightly reduces the precision. This approach has the additional benefit of not needing to run Mutect2 which is an order of magnitude slower in our tests, than Strelka2Pass and FreeBayesSomatic (Table A.2). The usage of these workflows can be easily integrated into existing workflows and can be customised to the needs of the user.

Additional supplementary data

This section contains supplementary data for the joint somatic variant calling chapter (chapter 2) not contains in the published paper but for the work shown in this thesis

Listing A.1: parse strelka VCF

```
#this function will parse a strelkaVcf after the joined refiltering put all the
#relevant info into a data.table
parseStrelkaVcf <- function(vcfFilePath=stop("need file as input"), geneList){
  require(ensemblVEP)
  vcfObj <- VariantAnnotation::readVcf(vcfFilePath)
  vcfObj <- VariantAnnotation::expand(vcfObj)

  res <- data.table()
  #add the fixed columns to the result
  fixed <- .getFixedVcfColumns(vcfObj)
  res[,c("chr","pos","ref","alt","filter"):=fixed]

  #get the vep annotation
  vepcolumns <- .getVepAnnotationColumns(vcfObj, geneList)
  res <- cbind(res,vepcolumns)

  #add additional info, which is interesting but we dont us TUMOR and NORMAL
  # because what if the columns were renamed
  res[, dp:=as.numeric(geno(vcfObj)$DP[,2])]
  if(is.null(geno(vcfObj)$AF)){
    res[ ,freq:=extractStrelkaAD(vcfObj, sample=2)/dp]
    res[ ,nfreq:=extractStrelkaAD(vcfObj, sample=1)/dp]
  }else{
    res[ ,freq:=as.numeric(geno(vcfObj)$AF[,2])]
    res[ ,nfreq:=as.numeric(geno(vcfObj)$AF[,1])]
  }

  return(res)
}
```

Listing A.2: annotate variants with copy number calls

```
for (sample in dnaSamples){

```

```
  varName <- paste0("vars",sample)
  varTmp <- as.data.frame(get(varName))

  segments <- read.table(paste0(pathPrefix,'/dawson_genomics/Projects/
    CASCADE/', .sampleBase, '/analysis/', .sampleBase, "-",  sample, '/CNV
    /', .sampleBase,"-",sample,"_segments.txt"),header=T)
  cnInfo <- apply(varTmp,1,function(x){
    # print(x["chr"])
    cnRow <- segments[(segments$chromosome == x["chr"] & segments$start.
    pos <= as.numeric(x["pos"]) & segments$end.pos >= as.numeric(x["pos"])
    ),]
    cnCols <- cnRow[,c("CNt","A","B")]
    if (nrow(cnCols) == 0){
      #no info means we assume its normal
      cnCols <- data.frame(CNt=2,A=1,B=1)
    }
    return(cnCols)
  })

  cnInfo <- bind_rows(cnInfo)
  varTmp <- cbind(varTmp, cnInfo)

  #dividing by purity and then dividing by CNt tells us how many cells
    have that variant
  varTmp$vaf <- varTmp$freq / purities[counter] #/ varTmp$CNt
  #does not help if we have more than 1 in a percentage
  varTmp$vaf[varTmp$vaf > 1] <- 1
  assign(varName, as.data.table(varTmp))
  counter <- counter +1
}
```

Listing A.3: convert to maf format

```
convertToMaf <- function(varTable, chrs=paste0("chr",c(1:22, "X", "Y"))){

  keep <- varTable[chr %in% chrs,]

  ret <- data.table(Hugo_Symbol=keep$symbol,
                    Chromosome=keep$chr,
                    Start_position=keep$pos,
                    Reference_Allele=keep$ref,
                    Tumor_Seq_Allele2=keep$alt,
```

```
                    t_ref_count=round(keep$dp*(1-keep$freq)),
                    t_alt_count=round(keep$dp*keep$freq),
                    local_cn_a1=keep$A,
                    local_cn_a2=keep$B)

  #if there is no gene annotated, we set it to unknown
  ret[is.na(ret$Hugo_Symbol),"Hugo_Symbol"] <- "Unkown"

  return(ret)

}
```

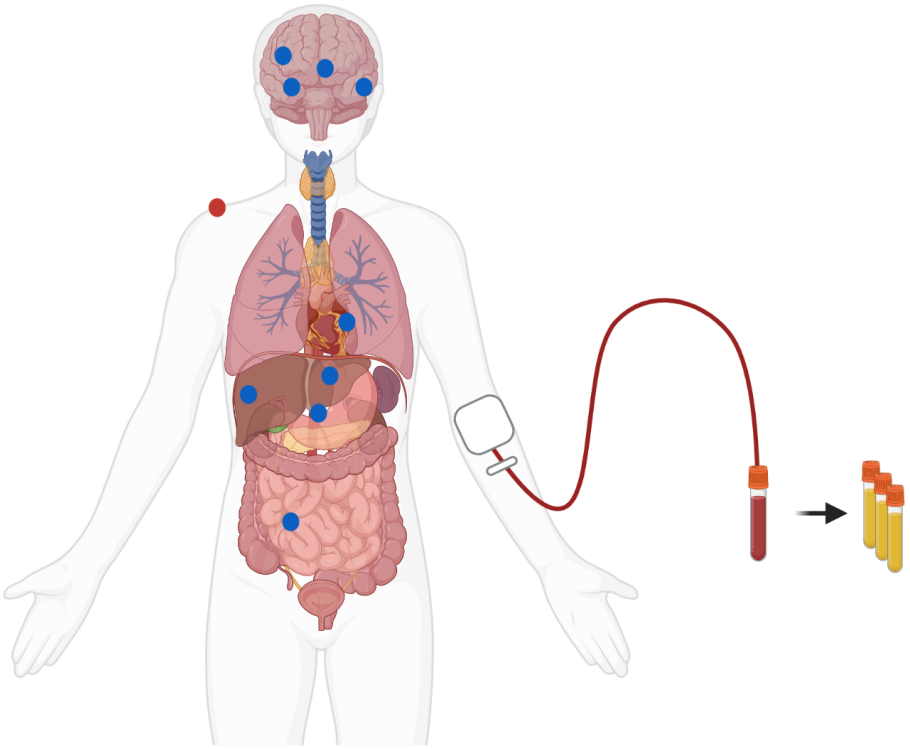

Figure A.13: Schematic of analysed tumour lesions in patient CA-F; Primary (diagnostic) skin sample is shown in red; metastatic sites are shown in blue; From top to bottom: right parietal lobe; left temporal lobe; right cerebellum; posterior mediastinal lymph node; left liver lobe; right liver lobe; liver, hepatic vessel; small bowel; Right side depicts three blood draws with plasma sampling (Figure 2.4)

B

MisMatchFinder - supplementary methods

B.1 ROI bed files generation

To ensure optimal mapping rates and no mapping related mismatches, the analysis is restricted to high mappability areas of the genome. These areas were defined as regions, where a k-mer of 100bp has a 85% or higher unique mappability rate. The mappability tracks were first computed with GEM [222] and then collated converted to a bed file with R just like in the best practice instructions of QDNAseq [37] for creating a new bin annotation. This method is only required for GRCh38 [223] as so far, the UCSC data track is only available for GRCh37 [224].

B.2 Oligo-nucleotide context normalisation

The ROI restriction of the analysis from Section B.1 automatically leads to a different tri- and di-nucleotide context frequency in the analysed regions, than the rest of the genome, which was used to generate the original signatures [198]. For this reason, MisMatchFinder analyses the oligo-nucleotide composition of the analysed regions and generates weighted counts by adjusting for the differences.

The baseline frequencies of both di- and tri-nucleotides were generated with the function *oligonucleotideFrequency* from the "Biostrings" library [34] using the hg38 BSgenome [36]. The raw counts of the di- and tri-nucleotides can be seen in Table B.1 and Table B.2 respectively.

Table B.1: Dinucleotide counts generated with Biostrings [34] for GRCh38

DINUCLEOTIDE	COUNT
AA	287 025 139
AC	148 150 331
AG	205 752 406
AT	226 225 785
CA	212 880 749
CC	151 236 932
CG	29 401 795
CT	205 524 144
GA	175 847 498
GC	124 732 844
GG	152 432 158
GT	148 502 457
TA	191 400 248
TC	174 923 630
TG	213 928 532
TT	289 690 054

Table B.2: Trinucleotide counts generated with Biostrings [34] for GRCh38

TRINUCLEOTIDE	COUNT	TRINUCLEOTIDE	COUNT
AAA	112 465 943	GAA	58 990 420
AAC	43 532 050	GAC	27 737 004
AAG	58 439 928	GAG	49 560 877
AAT	72 587 151	GAT	39 559 024
ACA	59 305 516	GCA	42 481 943
ACC	33 784 390	GCC	34 497 599
ACG	7 584 302	GCG	7 078 395
ACT	47 476 086	GCT	40 674 873
AGA	65 552 680	GGA	46 022 042
AGC	41 073 623	GGC	34 474 720
AGG	51 723 263	GGG	38 148 838
AGT	47 402 783	GGT	33 786 518
ATA	60 308 591	GTA	33 265 786
ATC	39 076 747	GTC	27 466 578
ATG	53 548 035	GTG	44 578 403
ATT	73 292 370	GTT	43 191 653
CAA	55 220 609	TAA	60 348 082
CAC	44 001 434	TAC	32 879 810
CAG	59 791 771	TAG	37 959 659
CAT	53 866 888	TAT	60 212 654
CCA	53 293 160	TCA	57 800 075
CCC	38 036 593	TCC	44 918 305
CCG	8 026 845	TCG	6 712 244
CCT	51 880 303	TCT	65 492 835
CGA	6 511 692	TGA	57 760 931
CGC	7 021 552	TGC	42 162 935
CGG	8 229 568	TGG	54 330 453
CGT	7 638 969	TGT	59 674 158
CTA	37 666 053	TTA	60 159 779
CTC	49 481 013	TTC	58 899 235
CTG	59 039 769	TTG	56 762 262
CTT	59 337 262	TTT	113 868 707

B.3 Germline filtering with zarr

As shown in Figure 4.4A, the amount of mismatches found in a 10x coverage sample can easily exceed 3 million. In addition to that, the current gnomAD database contains $\approx$ 707 million variants. This means a normal merge for two datasets based on chromosomal position is not feasible for a normal compute resource in a acceptable time frame. To allow an easy query of mismatch positions against the full database, a zarr [56]representation of the gnomAD VCF was generated. However in contrast to the out of the box indexing function shipped with skikit-allel [205] which was used to convert the vcf to zarr, the program uses its own index built with ncls, which is available through PyRanges [52]. The sections below outline first the conversion process with scikit-allel (Section B.3.1) and then details the filtering in the MisMatchFinder program (Section B.3.2)

B.3.1 Zarr conversion with scikit-allel

While it is easy to access a zarr archive, both for reading and writing, once it is created, the generation requires time. The time is mostly computational and not so much development, as the scikit-allel package contains the function '$allel.vcf\_to\_zarr$', which allows the direct conversion of VCF to zarr with only a few prerequisites.

Importantly, tabix [225] can be used to split the conversion into multiple parts by restricting the process to specific regions.

Listing B.1 shows the code used to convert chromosome 'chr1 'from the downloaded gnomad vcf

Listing B.1: scikit-allel conversion vcf_to_zarr

```
import scikit-allel as allel

allel.vcf_to_zarr(input="gnomad.genomes.r3.1.2.sites.vcf.bgz", output="/
    out/put/folder/",  group="chr1", region="chr1", fields="*")
```

When MisMatchFinder is installed on your system, the function '$generateZarrStorage$'is a wrapper, which allows the parallel conversion as well to resume a failed or incomplete attempt. It is equivalent to the above code and has only usability and ease of access as priorities. This automated version will convert all fields, which include fields never used in MisMatchFinder to optimise the memory footprint of the zarr representation, the option fields in Listing B.1 can be set to the value shown in Listing B.2.

Listing B.2: field options for reduced memory

```
fields="['variants/CHROM', 'variants/POS', 'variants/REF', 'variants/ALT
    ', 'variants/AF', 'variants/FILTER_PASS']"
```

Which contains only the information used in MisMatchFinder. The same result can be achieved with adding the option '–$mandatoryOnly$'to the supplied wrapper.

actually implement the wrapper i am talking about here

B.3.2 MisMatchFinder filtering - the zarr API

B.3.3 Data simulation

This section contains all the additional information required to replicate the simulation of data used in the MisMatchFinder chapter (chapter 4)

B.3.4 Signature simulation - we can spike this punch

This section describes the signature spike-in simulation. The full code of the variant selection is available in Listing B.3 with the bamsurgeon code shown in Listing B.4.

For the selection of variants to spike-in with bamsurgeon, I use the fully annotated "CosmicMutantExport.tsv" from https://cancer.sanger.ac.uk/cosmic/download, then restrict the list to SNPs. These are loaded into R and annotated with their tri-nucleotide context. Because the signatures are based on the pyrimidine nucleotides, the reverse complement is generated for variants with a purine in the center position.

The sampling amount is calculated by using the intended signatures percentages (e.g. Figure 4.1, Figure B.1) and multiplying with the desired amount of variants, which can be derived from the chosen mutation rate in per million (Equation B.1).

$$n(variants) = \frac{mutation\,rate}{1 \cdot 10^6 \cdot genome\,length} \tag{B.1}$$

For our data, we assume a genome length of $3 \cdot 10^9$ and use four different mutations rates (0.1, 5, 25, 50 and 100). For the final sampling I used "data.table" and finally variants are assign an allele frequency of 0.1.

Listing B.3: spike-in variant selection

```
#get the snps
bed <- data.table(read.table("/data/reference/dawson_labs/COSMIC/v92/CosmicMutantSNPs.bed", sep="\t"))
colnames(bed) <- c("chr", "start", "end", "ref", "alt", "cancer", "status")

#get the surrounding variants
varTriNuc <- GRanges(seqnames=bed$chr, IRanges(start=(bed$start-1), end=(bed$end+1)))

#select the right genome
genome <- BSgenome.Hsapiens.UCSC.hg38::BSgenome.Hsapiens.UCSC.hg38

#get trinucs
seq <- Biostrings::getSeq(genome, varTriNuc)

# if we dont have a C or a T as the ref, we build the reverse complement, because thats
# what the signatures are based on
seq[bed$ref %in% c("G", "A")] <- Biostrings::reverseComplement(seq[bed$ref %in% c("G", "A")])

#put the trinucs with the variants
bed[,tri:= as.character(seq) ]
#add in the trinuc alt, so that we know where things are goings
bed[,triAlt:=ifelse(ref %in% c("C", "T"), alt, as.character(Biostrings::complement(Biostrings::DNAStringSet(bed$alt))))]

#now we build the name of the trinuc change as it is used in COSMIC
bed[,cosmicName:=paste(tri,triAlt)]
bed[,cosmicName:=gsub(pattern="(.)(.)(.) (.)", replacement = "\\1[\\2>\\4]\\3", cosmicName)]

#read in the signatures profile
sig7a <- data.table(read.table("/home/shollizeck/workspace/myDawsonRep/TMB/v3.2_SBS7a_PROFILE.txt", header=T, sep="\t"))

# mutations per megabase
rate <- 100

```

```
nVars <- rate/1E6*3E9

#get the number of each trinucChange we need
numCols <- colnames(sig7a)[-1]
sig7a[,(numCols):=lapply(.SD, function(x) round(x*nVars)), .SDcols=
    numCols]

#merge the two tables together to enable sampling
sampleTableSBS7a <- merge(bed, sig7a, by.x="cosmicName", by.y="X")

selectionSBS7a <- sampleTableSBS7a[,.SD[sample(.N,size=SBS7a_GRCh38)], by
    ="cosmicName"]

#add in the frequency (VAF) of the variants which is a uniform 10%
selectionSBS7a[,vaf:=0.1]
```

With the generated bed, bamsurgeon can be used to create the final mutated BAM. As we are using low coverage WGS as input, some parameters need to be adjusted to allow variants to be generated. Mostly, we need to allow bamsurgeon to even mutate regions with very low coverage ('–$mindepth$ 1'), ignore the pileup of the original ('–$ignorepileup$'), allow a higher coverage difference ('–d 0.7') and lastly allow bamsurgeon to NOT mutate a position ('–$minmutreads$ 0'). To make the data creation reproducible, we also assign a seed of 1234.

After the BAM generation, the actually spiked-in variants are generated sorted and indexed.

Lastly, the bam needs to be postprocessed to be in line with the SAM specifications (Listing B.4).

Listing B.4: bamsurgeon spike-in

```
bamsurgeon addsnv.py -d 0.7 --ignorepileup --mindepth 1 --minmutreads 0 -
    v mutations.bed -r $reference -o mutated.bam --aligner mem --seed 1234
     -f input.bam

bamsurgeon makevcf.py addsnv_logs_mutated.bam | vcfstreamsort -a |
    bcftools view -o variants.vcf.gz -O z && bcftools index -t variants.
    vcf.gz

bamsurgeon  postprocess.py -f ${reference}.fai mutated.bam
```

B.3.5 Patient data subsampling

Subsampling of high depth WGS data was done with samtools (v1.13) supplying random seeds, but stable sampling rates. Sampling rates were selected, such that the output file would have an average coverage of 10x to be comparable with other sequencing data.

Supplementary Figures

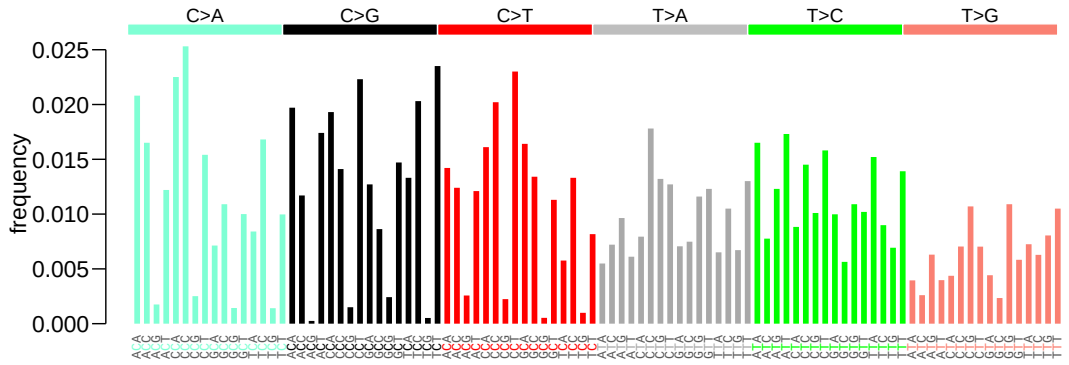

Figure B.1: Trinculeotide count contributions for SBS signature 3 (Defective homologous recombination-based DNA damage repair); values taken from Alexandrov et al. [198]

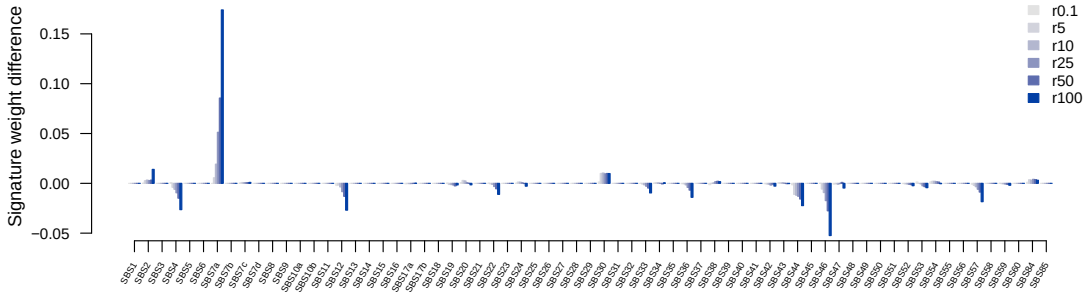

Figure B.2: Signature weights differences from normal for SBS7a spike-in; Weights were deconstructed with QP method in MisMatchFinder and the weights assigned to the normal sample used for the spike-in were substracted; r0.1 corresponds to 0.1 mutations per megabase (287 variants) and r100 is the equivalent of 100 mutations per megabase (286974 variants)

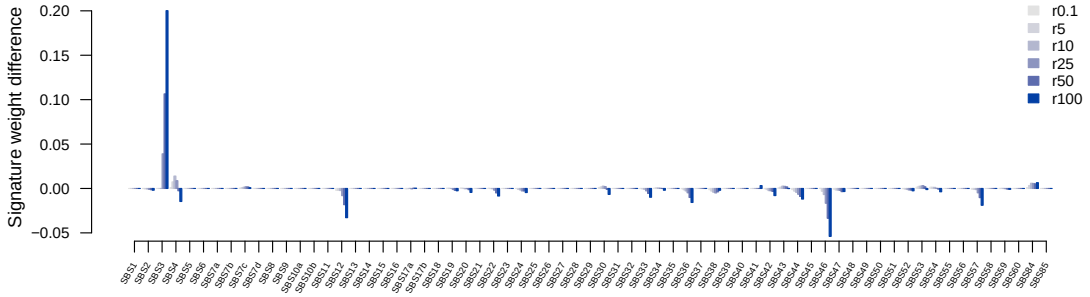

Figure B.3: Signature weights differences from normal for SBS3 spike-in; Weights were deconstructed with QP method in MisMatchFinder and the weights assigned to the normal sample used for the spike-in were substracted; r0.1 corresponds to 0.1 mutations per megabase (264 variants) and r100 is the equivalent of 100 mutations per megabase (285367 variants)