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
Sir Peter MacCallum department of Oncology
Melbourne School of Awesome
THE UNIVERSITY OF MELBOURNE

October 27, 2021

THE UNIVERSITY OF MELBOURNE

Abstract

Sir Peter MacCallum department of Oncology
Melbourne School of Awesome

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, AUTHOR NAME, declare that this thesis titled, ‘THESIS TITLE’ and the work presented in it are my own. I confirm that:

- The thesis comprises only my original work towards the NAME OF AWARD 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

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.

---

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

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.

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
- 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

Chapter 3:

summary plus contributions

Chapter 4:

Chapter 5:

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

Add Katie and Danes paper

Funding:

All necessary funding goes here

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;

---

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

- 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

0.1 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.

0.1.1 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

0.1.2 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].

Contents

Abstract iii

Declaration of Authorship iv

Preface v

Acknowledgements ix

- 0.1 Software and packages ix
  - 0.1.1 R ix
  - 0.1.2 python x

List of Figures xiii

List of Tables xv

Abbreviations xvii

Constants xix

Symbols xxi

1 Introduction 1

- 1.1 DNA 1
  - 1.1.1 Ploidy 4
  - 1.1.2 Mutations 5
- 1.2 cfDNA 6
- 1.3 DNA sequencing 7
  - 1.3.1 Library preparation 7
  - 1.3.2 Next generation sequencing 8
  - 1.3.3 Long read sequencing 10
- 1.4 DNA analysis 10
  - 1.4.1 Mapping 10
  - 1.4.2 Variant calling 11
  - 1.4.3 Germline 11
  - 1.4.4 Somatic 12
- 1.5 Lungcancer 13
- 1.6 Overview 13

2 Joint somatic variant calling 15

2.1 Introduction . . . 15
2.2 Publication . . . 16
2.3 Effects on downstream analysis . . . 17
2.3.1 Polygenetic reconstruction . . . 17
2.3.2 Clonal deconvolution . . . 20
2.4 Longitudinal analysis . . . 20
2.5 Usage . . . 21

3 CASCADE **23**
3.1 Introduction . . . 23
3.2 Publication . . . 23
3.3 Cohort analysis . . . 23
3.4 Mitochondrial phylogenetic reconstruction . . . 23
3.5 Outlook . . . 23

4 Mismatchfinder **25**
4.1 Introduction . . . 25

5 Conclusion **27**

A Strelka2Pass and FreeBayesSomatic publication **29**
A.1 Introduction . . . 30
A.2 Materials and methods . . . 31
A.2.1 FreeBayesSomatic workflow . . . 31
A.2.2 Strelka2Pass workflow . . . 32
A.3 Validation . . . 33
A.3.1 Simulated data . . . 33
A.3.2 Clinical data . . . 35
A.4 Discussion . . . 37

Bibliography **39**

List of Figures

1.1 Overview DNA structure 2
1.2 Overview Chromosome structure 3
1.3 Overview DNA structure 5
1.4 Library preparation for NGS 8
1.5 Sequencing by synthesis (Illumina) 9

2.1 Reconstructed phylogenies of joint samples 18
2.2 Tanglegram of the reconstructed phylogenies 19
2.3 Usage statistics joint workflows 21

A.1 Figure 1 Comparison of joint multi-sample variant calling and single tumour-normal paired calling methods 34

List of Tables

Abbreviations

DNA	**D**eoxyribo**N**ucleic **A**cid
RNA	**R**ibo**N**ucleic **A**cid
cfDNA	**c**ell **f**ree **DNA**
ctDNA	**c**irculating **t**umour **DNA**
bp	**b**ase **p**air
ChIP	**Ch**romatin **I**mmuno**P**recipitation
WGS	**W**hole **G**enome **S**equencing
WES	**W**hole **E**xome **S**equencing
SCLC	**S**mall **C**ell **L**ung **C**ancer
NSCLC	**N**on-**S**mall **C**ell **L**ung **C**ancer
RAID	**R**edundant **A**rray of **I**ndependent **D**isks
SNP	**S**ingle **N**ucleotide **P**olymorphism
InDel	**In**sertion or **Del**etion
SV	**S**tructural **V**ariant
PON	**P**anel **O**f **N**ormals
GATK	**G**enome **A**nalysis **T**ool**K**it
NJ	**N**eighbour **J**oining
UPGMA	**U**nweighted **P**air **G**roup **M**ethod with **A**rithmetic mean
WPGMA	**W**eighted **P**air **G**roup **M**ethod with **A**rithmetic mean
F81	**F**elsenstein 19**81** model
HKY85	**H**asegawa, **K**ishino and **Y**ano 19**85** model
HPC	**H**igh **P**erformance **C**omputing

Sort alphabetically

Constants

Speed of Light $c = 2.997\ 924\ 58 \times 10^{8}\ \mathrm{ms}^{-\mathrm{S}}$ (exact)

Symbols

a	distance	m
P	power	W (Js^{-1})
ω	angular frequency	rads^{-1}

"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 1.5 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

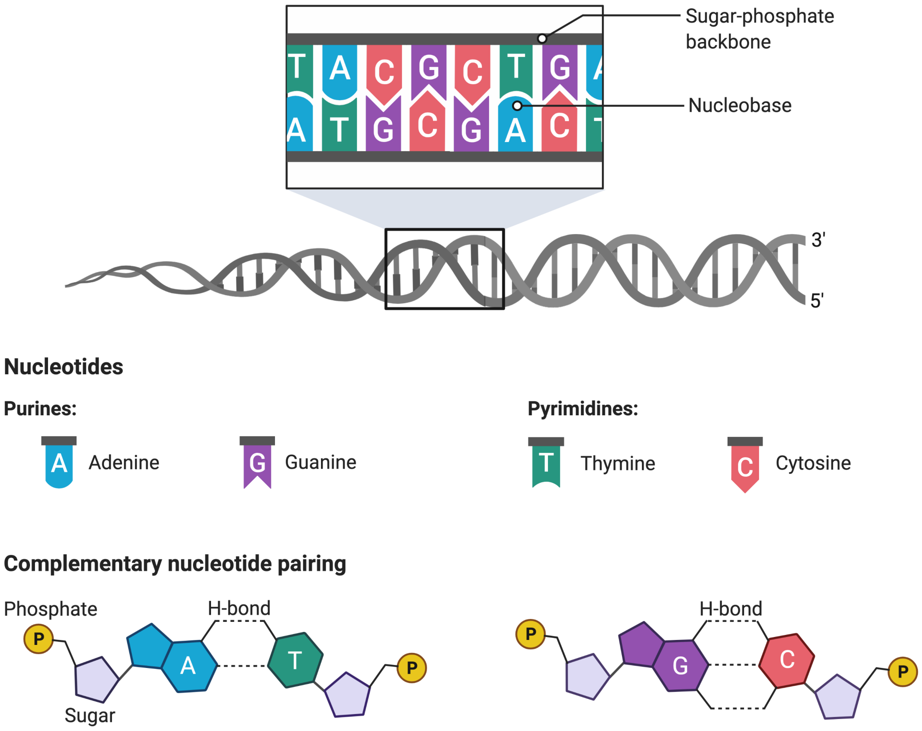

Sugar-phosphate backbone
Nucleobase
T A C G C T G A
A T G C G A C T
3'
5'
Nucleotides
Purines:
A Adenine
G Guanine
Pyrimidines:
T Thymine
C Cytosine
Complementary nucleotide pairing
Phosphate
H-bond
P
A
T
P
Sugar
P
G
H-bond
C
P

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

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 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 ≈ 24% and the rest is intergeneic (≈ 75%)[64].

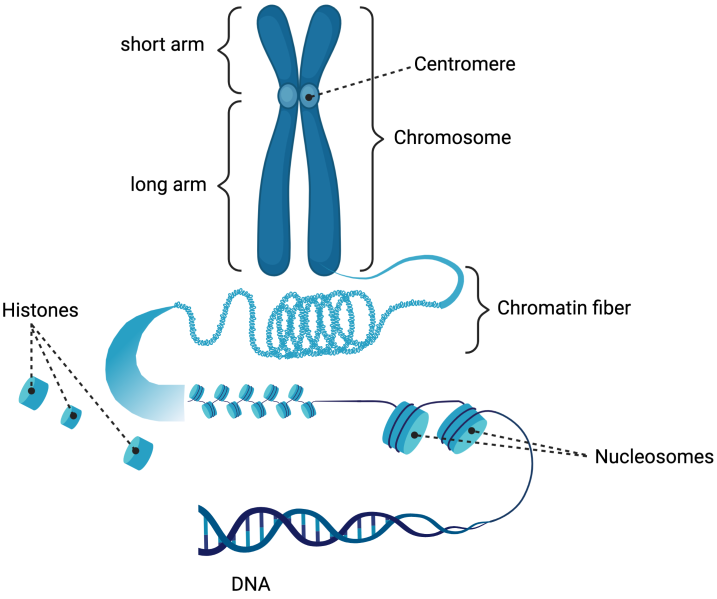


FIGURE 1.2: Structural overview of the metaphase condensed chromosome: DNA is first wrapped around Histones to form nucleosome, which then associate with each other to form the chromatin fiber, which in the metaphase of the cell cycle is condensed even more into the X-shaped chromosome

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 of the 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 - its 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 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 and 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.

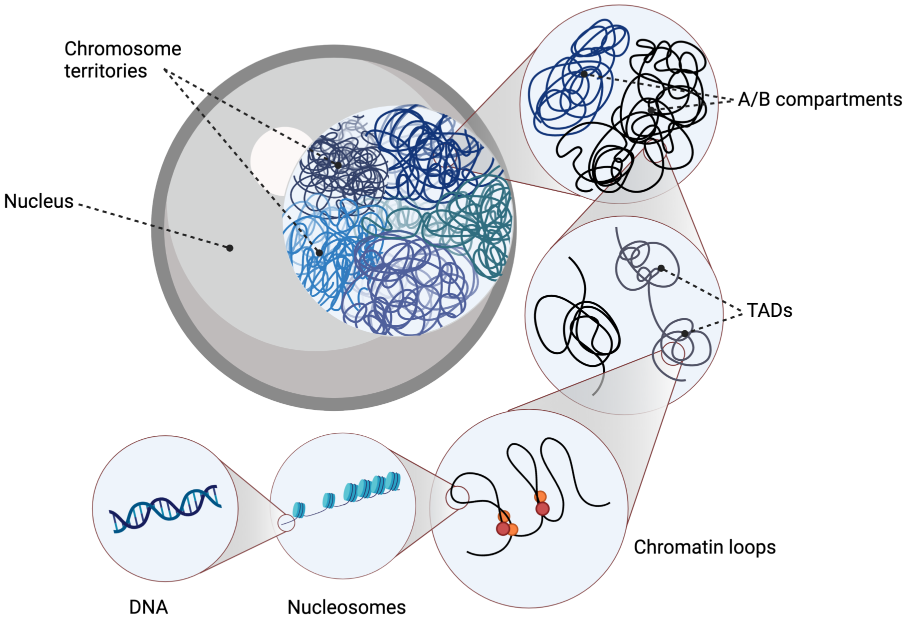


FIGURE 1.3: Individual chromosomes occupy a subspace in the nucleus called chromosome territories. Chromosome territories can be further partitioned to distinct A and B compartments, which are enriched for active and repressed chromatin, respectively. Genomic regions within topologically associating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

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 mutations occurs in the germline cell, eg. 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 many different 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 nucleosomes. 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 mothers 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].

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 flourecently, 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

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 both phosphorelated as well as an A-tail is added, before the adapter

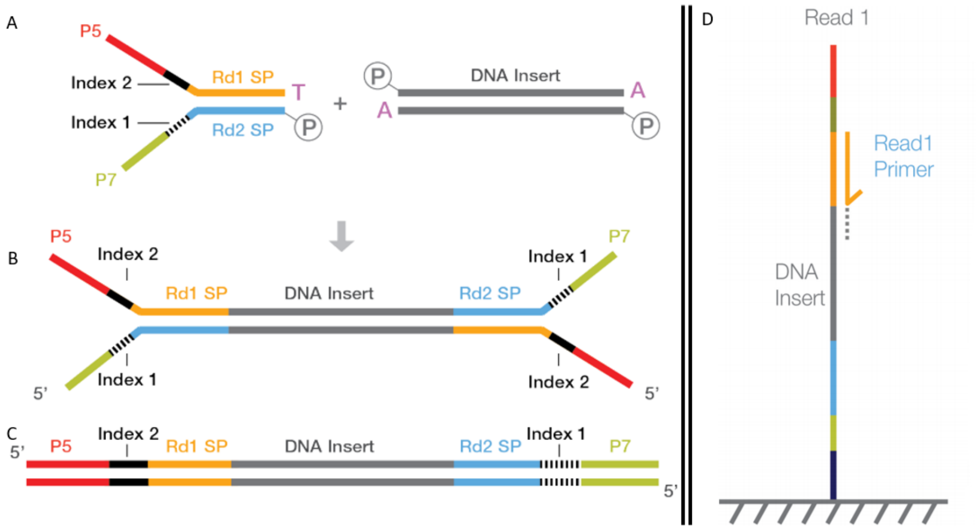
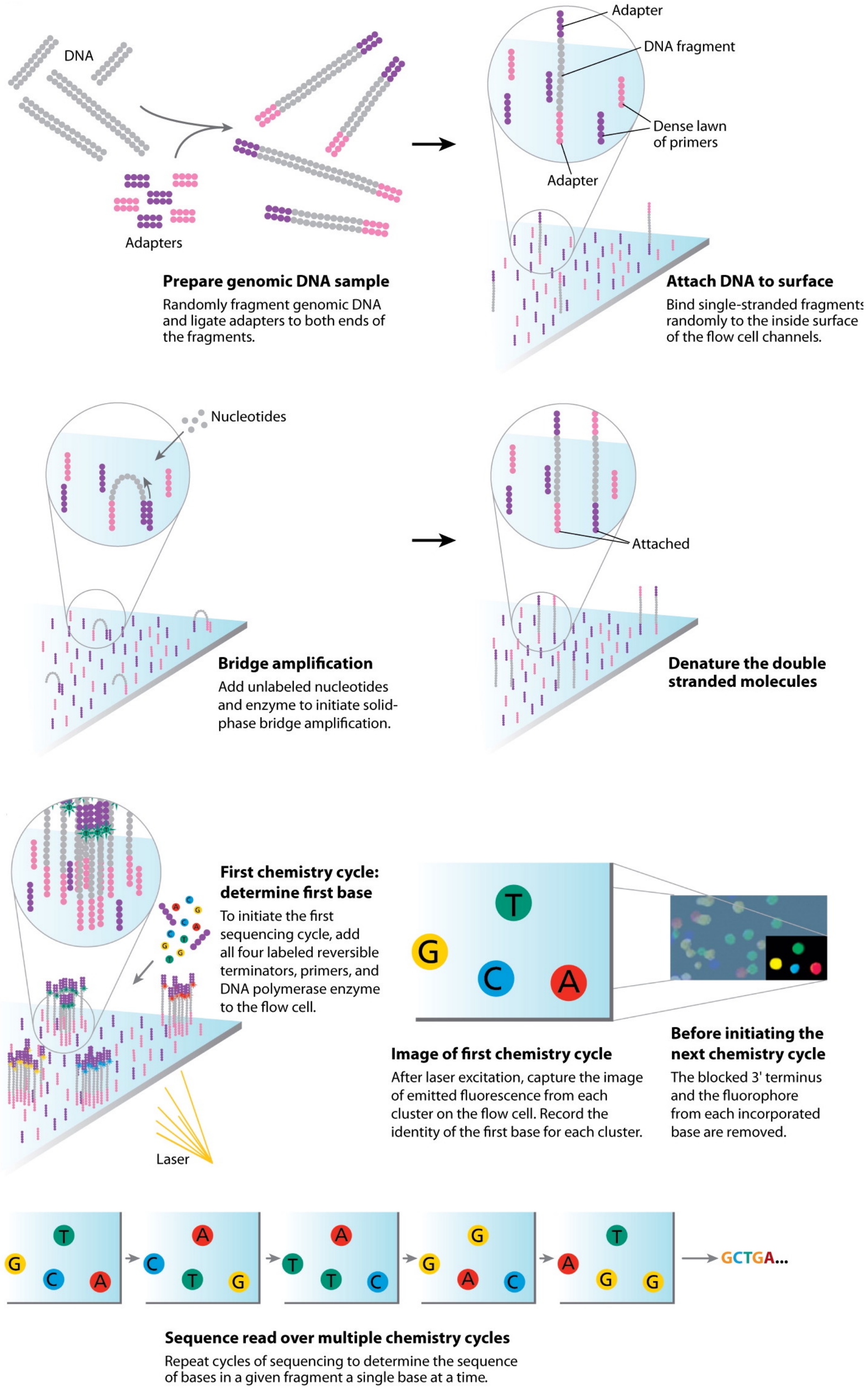


FIGURE 1.4: Adapter ligation during library preparation. The adapters are added to the DNA insert during library preparation. A. The DNA insert is prepared by adding an A-tail and phosphorylation. B. The adapter complex which includes the P5/P7 flow cell binding adapter is added to the DNA insert. C. 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]

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 allows the sequencing of 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.



Prepare genomic DNA sample

Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

Attach DNA to surface

Bind single-stranded fragment randomly to the inside surface of the flow cell channels.

Bridge amplification

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

Denature the double stranded molecules

First chemistry cycle: determine first base

To initiate the first sequencing cycle, add all four labeled reversible terminators, primers, and DNA polymerase enzyme to the flow cell.

Image of first chemistry cycle

After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

Before initiating the next chemistry cycle

The blocked 3' terminus and the fluorophore from each incorporated base are removed.

Sequence read over multiple chemistry cycles

Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at a time.

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]

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 of sequencing huge continous 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 originated from. As the current methods randomly fragment the DNA (subsection 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 doesnt 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 important, 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 subsection 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 (subsection 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 (subsection 1.1.1). The ploidy is important 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, however the observed allele frequencies will differ. 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]. These

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.

1.4.4.1 Tumour-normal variant calling - it's a "simple" substration!?

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 just like for germline calling (subsection 1.4.3). 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 of all, 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

1.4.4.2 Tumour only variant calling - how hard can it be?!

change this to not have subsubsections

1.5 Lungcancer

With around 1.6 million deaths world-wide each year, lung cancer is the number one cause of cancer death [111]. 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 [112]. 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 [113]. 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.

Possibly change this to cancer in general

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 have the built-in capability to joint call multiple samples, for example from family trios, virtually no somatic variant caller had this function.

The joint analysis of smaller cohorts improves the performance of germline variant calling methods significantly, by allowing to assess technical artifacts, which might be unique for the individual sequencing machine or the researcher handling the DNA [114, 115]. As certain parts of the genome are more problematic to sequence (section 1.3) and map (subsection 1.4.1), a “control“ sample can help to distinguish if a certain observed change occurs commonly, is a technical issue or in fact a real change.

For somatic variant calling, this concept has been adjusted in the genome analysis toolkit (GATK) [116] to allow the use of panel of normals (PON), which contains frequently seen changes in healthy (“normal“) individuals analysed with the same sequencing technology [117], but this is a post processing step of the analysis rather than a more intricate model like it is for the germline equivalent. Mutect2, which is the most recent somatic variant calling algorithm provided by the Broad institute, however also provides a multi-sample mode, for which all tumour samples need to be from the same patient, either longitudinal or spatial different [118]. This mode is hidden quite well and all tutorials published 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, this seems, that development is not a priority.

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 to occur in multiple samples called multiSNV [119]. However it has multiple shortcomings, which make it not useable 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 shown. However, multiSNV was optimised only for WES and not for the very deep WGS that is now available. This means exceptionally high runtimes. 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 again 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 at low cost.

multiSNV has been the only software available for multi sample analysis, but only recently, during this work, superFreq [120] was published. It combines all standard analysis steps for tumour analysis, like variant calling or clonal deconvolution, into one program and is even able to jointly analyse samples. However similar to multiSNV, its focus when optimising and developing was on WES and RNAseq data, so when applied to our data, we could not find a server 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 first show the publication and then discusses additional analysis done after the the publication of the manuscript (section 2.4) and the impact of the joint analysis on downstream methods (section 2.3).

2.2 Publication

The publication about joint somatic variant calling can be found at https://doi.org/10.1093/bioinformatics/btab606 and non-journal formated version is also attached as

Appendix A.

2.3 Effects on downstream analysis - not quite the missing link, but close

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 metastatic seeding. While traditionally it was thought that tumours only metastesised after they reached a certain size, to escape the restrictions of the niche, like reduced nutrition, recent publications showed, there is also very early metastatic seeding [121]. But all those methods are ultimately based on the somatic variants found in the data, so if we improve on the input of the downstream analysis methods, we can expect a clearer and possibly more granular result.

In this section I will highlight for a few examples on how big the effect can be for methods, like 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 phylogenetics reconstruction tools, we will not show a comprehensive amount of these tools, but rather focus on the results. For this reason, we chose to use neighbour joining (NJ) [122], because it is fast, readily available in most phylogenetic reconstruction toolkits 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 the real distance, the tree topology will still be reconstructed correctly [123]. Lastly, in contrast to many other methods like UPGMA and WPGMA [124], NJ does not assume an equal mutation rate of each sample, because we know, that the molecular clock hypothesis [125] is not valid for different lineages of cancers [126].

While there are lots of distance measures for DNA sequences, which allow accounting for different probabilities of transitions and transversions as well as uneven base composition, models like F81 [127] or HKY85 [128] are only really designed for germline mutations and are not easily applicable for subclonal somatic mutations, which is why we decided to first transform the variants present in all samples into a binary occurrence vector and then calculating the Hamming distance [129] 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 2.1 shows both the reconstructed phylogenies when using the variants found with the default tumour-normal method and our improved joint method and using the exact same reconstruction method otherwise.

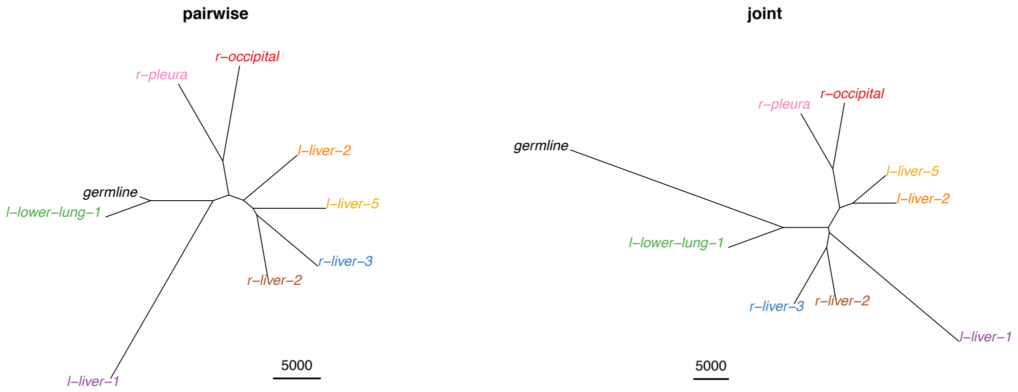


Figure 2.1: Reconstructed phylogenies of a patient with multiple spatially distinct samples; Reconstruction method is neighbour joining for both cases, only the input of the called variants is different. Tip labels describe the location of the sample in the patient. Trees are shown as unrooted with germline as fixated origin point; black line ruler shows the length of an edge with 5000 mutations

There are several obvious changes, first, the longer edge connecting the the germline, which we consider as the state of no somatic variants, to all other samples. This shows that there are many more shared mutations in all samples, than what would have been anticipated with the default method. 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. [130]) this difference of $\approx$ 36000 variants is equivalent to $\approx$ 2000 cell divisions. While the cell doubling rate of lung cancers is highly dependent on the type [131], this difference makes a huge difference when assessing the timing of the tumour initiation and further evolution.

Secondly, there have been topological changes, which generate a bifuricating edge between the warm coloured pleura, occipital, l-liver-2 and l-liver-5 samples, which fits significantly better with the clinical history and PET scans of the patient, as these are the sites of progression, which lead to the death of the patient in contrast to the earlier sites of disease.

Figure 2.2 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. 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.

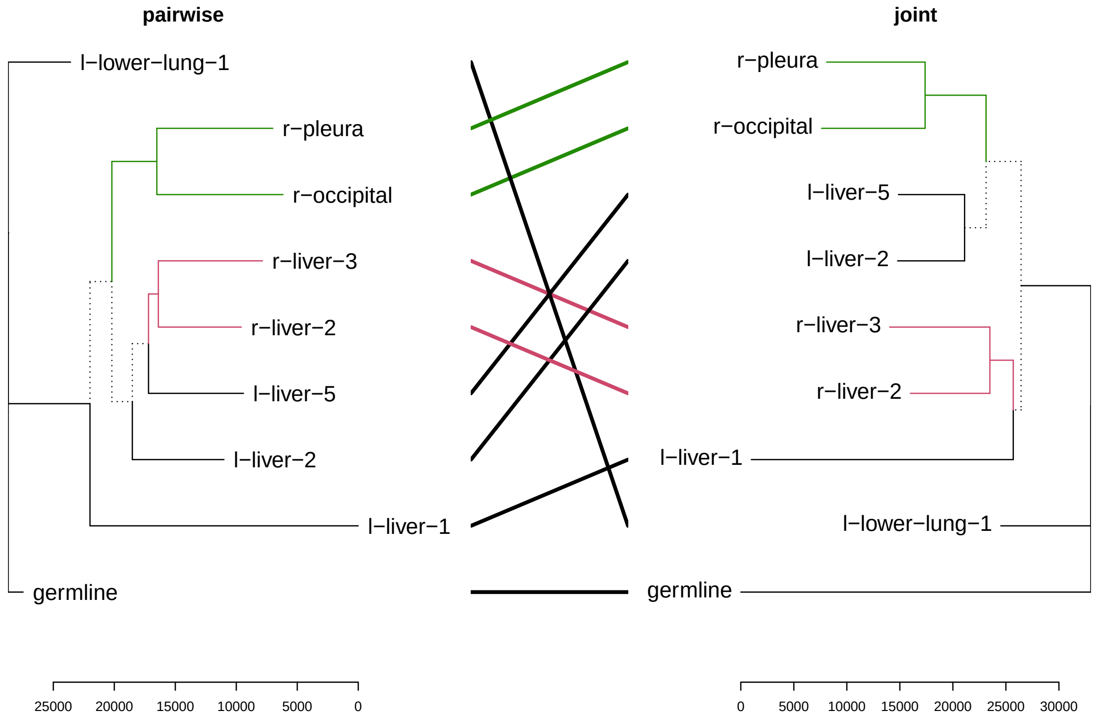

pairwise
l−lower−lung−1
r−pleura
r−occipital
r−liver−3
r−liver−2
l−liver−5
l−liver−2
l−liver−1
germline
25000
20000
15000
10000
5000
0
joint
r−pleura
r−occipital
l−liver−5
l−liver−2
r−liver−3
r−liver−2
l−liver−1
l−lower−lung−1
germline
0
5000
10000
15000
20000
25000
30000

Figure 2.2: Side by side view of the reconstructed trees from Figure 2.1; internal edges, which are distinct between trees are shown as dotted lines; common subtrees are shown in red and green; Tree labels have been sorted to minimise distance between labels; axis show the hamming distance between two nodes. Visualisation generated with dendextend [25]

One example of this is “l-liver-3“ which was a singleton, almost an outlier, in the pairwise reconstructed tree but is clustered together with “r-liver-2“ and “r-liver-3“. Generally, there have been additional changes, which make the topology more granular. Where all of the “l-liver-*“ samples were shown as an almost linear branching off the main stem, they are now further subdivided and each show distinct mutations, like “l-liver-2“ and “l-liver-5“ (Figure 2.2).

show before vs after for phylogenetic reconstruction

2.3.2 Clonal deconvolution

show before vs after for clonal deconvolution

2.4 Longitudinal analysis - something for the ages

Consider the order of longitudinal vs downstream

While the initial motivation for the development of these tools was the analysis of multi-region, so spatial, samples from the same patient coming from the CASCADE rapid autopsy program, a longitudinal application of these methods 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 other datasets, which highlights the flexibility and wide spread use of our new methods.

Specifically, I will show the analysis of longitudinal ctDNA WES of patient “CA-F“ from the manuscript. 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 [132]. In this analysis we aim to improve the quality of detection of low allele frequency somatic variants. This would enable the detection of variants from brains metastasis, as the blood brain barrier decreases the availability of DNA fragments from lesions in the brains in the normal blood stream [133], however it is detectable with sensitive enough methods [134, 135].

select the right dataset to show here (possibly the one from the MisMatchFinder stuff)

2.5 Usage - its not just me that thinks it is good

As published software is amazing, its real value is in the re-usability and portability. 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, many other people even just at the same institute have already shown interest and some research groups are already using the software to analyse their multi-sample data.

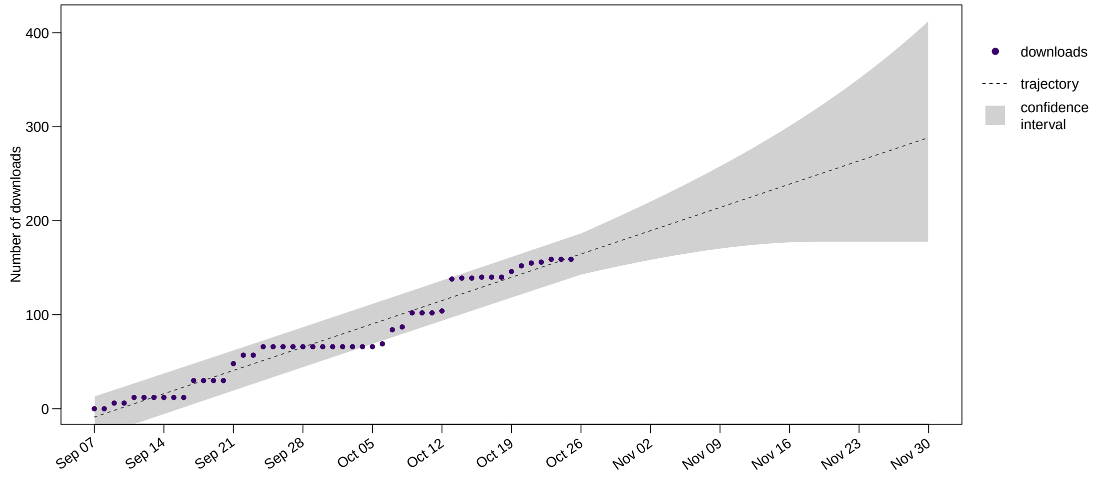


FIGURE 2.3: Cumulative download numbers of the “dawsontoolkit“ docker container since publication of the manuscript; Actual counts are shown as dots, with smoothed trajectory depicted as dotted line with the 95% confidence interval shown as a grey background; confidence interval has been adjusted with exponential decay of prediction accuracy with distance from the last data point; Start date 7th September 2021

To have some proxy of the usage statistics of the workflows, I recorded the download numbers of the “dawsontoolkit“ docker container, which only contains software for re-filtering and joint analysis of the workflows after the publication of the manuscript. 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. However it still shows an interaction and an interest in the methods. The download numbers show a quick increase in numbers in the beginning, just after the publication, followed by a short plateau but then another increase (Figure 2.3). This suggests, that there is a need in the methods, rather than a simple curiosity after publication.

Update the plot at a later timepoint)

> *“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

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[136]) 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[137])

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

“Many a mickle makes a muckle.“

— proverb

4

MisMatchFinder - hope springs eternal

4.1 Introduction

“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

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. 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 [138, 139]. 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 [121, 3, 140].

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 [119], SuperFreq [120], 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 [141].

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 [142] 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 (Eq. 1) and the tumour (Eq. 2) samples genotype likelihoods (GL) with g_0 describing the reference genotype.

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

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

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

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

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

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

A variant is classified as somatic when both somaticLOD as well as somatic VAF pass the criteria somaticLOD (Eq. 3) and somaticVAF (Eq. 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 [143] 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 [144]. 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.

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

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:

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

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 (Eq. 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

A.3.1 Simulated data

We first simulated a phylogeny with somatic and germline variants from ten tumour samples and one normal (Fig. 1A, S1A, B) (Supplementary methods). 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 [145]. The simulated reads were aligned to GRCh38 and both germline and somatic variants from the phylogeny were spiked into the aligned reads using Bamsurgeon [146]. 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

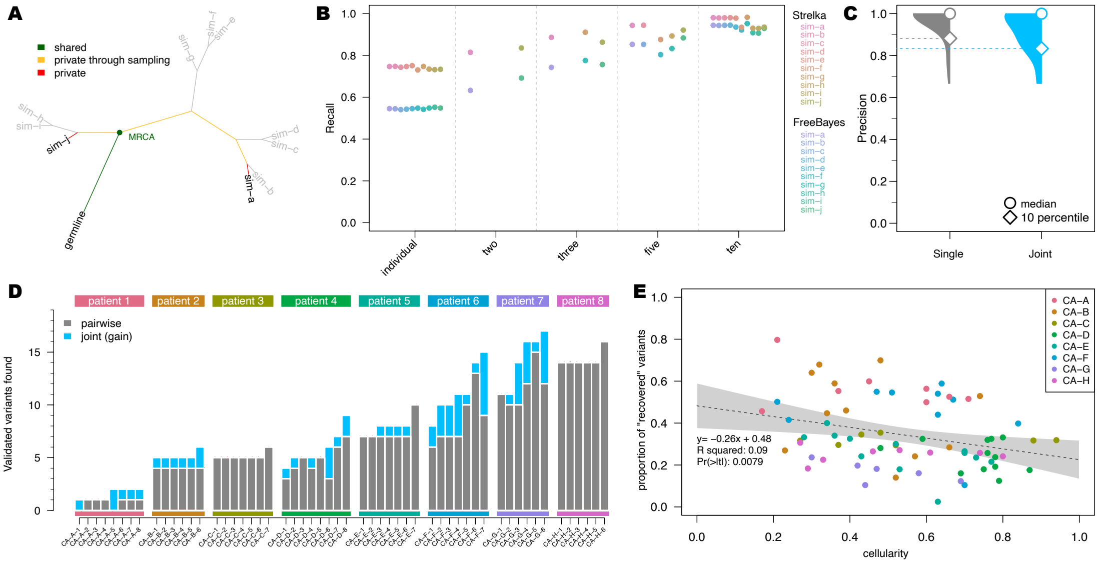


FIGURE A.1: Comparison of joint multi-sample variant calling and single tumour-normal paired calling methods; A) Simulated phylogeny highlighting two samples with high evolutionary distance (sim-a and sim-j) where MRCA denotes the most recent common ancestor. B) Recall estimates of FreeBayes and Strelka2, 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).

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 (Fig. S2). Mutect2 however, had virtually no change in precision, but the recall actually decreased from $\approx 75\%$ to $\approx 41\%$ when analysing the samples jointly (Fig. S2B). 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 (Fig. S3). This improvement in performance for Strelka2 is only achieved after the refiltering step and not just a result of the second pass (Fig. S4) (Supplementary Methods).

The performance of joint variant calling in Mutect2 was inferior compared to all other methods (Fig. S2A, 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 (Fig. S5A, B). This result was unexpected as the simulated variants were evenly distributed along the genome and the corresponding allele frequencies were sampled randomly (Fig. S1D).

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 (Fig. 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 (Fig. S6).

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 S1 and Supplementary methods) [137, 147]. 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 (Fig. S7) 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 (Fig. 1D, S8) (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 (Fig. 1E, S9).

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 (Fig. S8E), similar to its decreased performance in the simulated data (Fig. S2). This was due to true variants being removed by the internal filters of the tool (Fig. S5C, 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 (Fig. S8, S10) but longer run times (Table S2).

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 [148] 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 (Fig. S11).

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.

think about if its smart to put the supp data in here as well

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* 409.6822 (Feb. 2001), 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] 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.

[112] 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.

[113] 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.

[114] 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.

[115] 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.

[116] 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/.

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

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

[119] 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.

[120] 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.

[121] 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.

[122] 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.

[123] 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.

[124] 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.

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

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

[127] 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.

[128] 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.

[129] 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.

[130] 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.

[131] 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.

[132] 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.

[133] "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.

[134] Sang Eun Yoon et al. "Clinical Application of Circulating Tumor DNA in Plasma of Patients with Primary Central Nervous System Lymphoma". In: *Blood* 134.Supplement_1 (Nov. 2019), pp. 1491–1491. DOI: 10.1182/blood-2019-130075.

[135] Chunhua Ma et al. "Detection of circulating tumor DNA from non-small cell lung cancer brain metastasis in cerebrospinal fluid samples". In: *Thoracic Cancer* 11.3 (Mar. 2020), pp. 588–593. DOI: 10.1111/1759-7714.13300.

[136] 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.

[137] 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.

[138] 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.

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

[140] 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.

[141] 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.

[142] 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.

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

[144] 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.

[145] 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.

[146] 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.

[147] 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.

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