

# Development of new methods for accurate estimation of tumour heterogeneity

by

Sebastian Hollizeck

ORCID: 0000-0002-9504-3497

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

in the

The Sir Peter MacCallum Department of Oncology  
Faculty of Medicine, Dentistry, and Health Sciences

**THE UNIVERSITY OF MELBOURNE**

04/01/2023

THE UNIVERSITY OF MELBOURNE

## *Abstract*

The Sir Peter MacCallum Department of Oncology  
Faculty of Medicine, Dentistry, and Health Sciences

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 can be ethically challenging [3]. With circulating tumour DNA (ctDNA) as a proxy it is possible to analyse a snapshot of the unified heterogeneity, but there is still an unmet need for new analysis methods to optimize the analysis of these very valuable data and drive new treatment targets [4].

In this work we will develop new methods to study genetic heterogeneity from next generation sequencing (NGS) of tumour tissue as well as ctDNA to elucidate the role of tumour heterogeneity on treatment resistance.

# Declaration of Authorship

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

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

Signed:

---

Date:

---

# Preface

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

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

## Chapter 1:

The 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 tumour heterogeneity.

## Chapter 2:

The chapter “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<sup>st</sup> September 2021. In addition to the published analysis, I have added longitudinal analysis and its evaluation as well as the impact of this new method on other downstream analysis, like phylogenetic reconstruction and clonal deconvolution.

Contributions for this chapter:

- I conceptualised the work
- I implemented the workflows and containerised all required tools
- I performed the data simulation

---

<sup>1</sup>These authors contributed equally and are considered shared last.

- I performed the analysis presented in the publication
- I wrote the draft of the manuscript and performed revisions
- Dineika Chandrananda (D.C.) and Sarah-Jane Dawson (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
- Stephen Wong (S.W.) performed the targeted amplicon validation
- S.W. and Ben Solomon (B.S.) read the draft manuscript and provided feedback
- B.S. provided clinical expertise for human data

### **Chapter 3:**

In this chapter the analysis of five lung cancer patients is described with regards to intra- and inter-patient tumour heterogeneity. A special focus was put on sample and clonal relationships using the variant calling methods from Chapter 2. We also developed a phylogenetic reconstruction method based on mitochondrial variants, which allowed a different avenue for insight into metastatic seeding and timing. Parts of this analysis has been published[5, 6], but the publications are not included in this thesis, as I did not contribute more than 50% of the work of the articles.

Contributions for this chapter:

- I analysed all data
- I generated all visualisations
- I wrote the draft
- I implemented the mitochondrial phylogeny reconstruction method
- S-J.D., D.C., and Mark Dawson (M.D.) conceptualised the mitochondrial phylogeny reconstruction
- S-J.D., D.C. and B.S. read the draft manuscript and provided feedback
- Lavinia Tan (L.T.) and B.S. provided clinical expertise for human data

**Chapter 4:**

The MisMatchFinder analysis method described in this chapter is an original work using a read-centric variant calling approach to detect tumour somatic mutational signatures from low coverage sequencing data. The work describes individual design decisions and shows the performance of the method in multiple datasets. This work is unpublished and has not yet been submitted for publication.

Contributions for this chapter:

- I conceptualised the work with input from S.W.
- I analysed all data
- I generated all visualisations
- I implemented the method.
- I wrote the draft
- D.C. provided guidance and support for method development
- S-J.D. and D.C. read the draft and provided feedback
- Lavinia Tan (L.T.) and B.S. provided clinical expertise for human data

**Chapter 5:**

The conclusion chapter places the thesis in the wider field of existing methods to assess tumour heterogeneity and outlines future directions of the field.

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

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

---

<sup>2</sup>These authors contributed equally and are considered shared first.

# *Acknowledgements*

I first want to acknowledge the Wurundjeri people of the Kulin nation, the traditional custodians of the land on which my work was conducted. I want to pay respects to their elders: past, present and emerging and all other elders that might happen to read this work. In my time in Australia I was lucky to get a glimpse of this mysterious and special continent and country through their eyes and I am grateful for their ongoing work in keeping the legends and teachings alive.

## **People**

First of all I want to thank Imran House, who got me excited for academical science again. When I finished my Master's degree I was convinced that I would neither need nor want to get a PhD, but after working with him and experiencing his excitement about cutting edge research and his attitude towards academia I opened up to the idea of a PhD. And then he even helped me get a position and shoed me around this beautiful country and it's many loveable customs.

When I said Imran helped me get a PhD position I mean he put me in contact with the lovely Sarah Ftouni, who then helped me every step of the way as if I was her own brother, even though I know how much of a pain I can be. She helped me through all this time with even the dumbest questions and requests with barely any complaints. A genuine treasure and great lab manager.

Caroline Owen was invaluable for me during the application phase, she just took over and just paved the way. Without her I would probably still sit in Germany and trying to figure out the intricacies of both the VISA and stipend applications. When she passed the hat over to Erika Cretney, Erika and her team did everything in their power to help me through the issues of an ever shrinking staff at the university and COVID-19 regulations. So while Caroline made me coming to Australia possible, I want to thank the whole research education team for their continuous efforts.

Over the last years, Sarah-Jane has always shown me that my interests and wishes are important to her. She supported and challenged me to grow both as a person as well as a scientist. Her ability to conceptualise and prioritise were a great inspiration.

With Dineika Chandrananda I struck a real jackpot as a supervisor. She helped me orient myself and offered invaluable advice and patience when discussing methodological ideas. She always had an open ear for my issues and often urged me to take a break and come



back refreshed. I am very thankful to have had a supervisor that cared at least as much about my work as my mental health.

Even though my last supervisor Ben Solomon is a very busy person, he always had time to meet and discuss patient data or research questions. He was always advocating for me and my professional development. I even had the chance to meet Ben's lovely family, which again showed that it was not just about my work. He is a great scientist and doctor, working hard to make the world a better place.

I also want to thank Mark Dawson, who even though he was not officially part of the work in this thesis, he did support and offer advice. Exactly like Sarah-Jane he believed in me and my skills. He was always interested in my progress and valued my opinion.

Lastly I want to acknowledge, that my family supported me however they could during these last years. Even though they were on the other side of the world, I always felt close and connected with them. They enabled me to make my own choices over the years and while it must have been very hard for them to see me move away, they always encouraged me. Thank you so much!

## Software and packages

This section is dedicated to all the software that usually gets un-cited because they are “standard” or backbone.

Many figures in the introductory Chapter 1 were created with the help of BioRender.com

Most analysis in a prototype state was done on a linux cluster running Centos 7 [7] with Bash [8] and due to the high amount of data, parallel [9] was used of the multi-cpu architecture of HPCs.

## R

In depth data analysis and visualisation was done with R [10] with the help of packages listed below.

Most of the parallelisation in R was performed with BiocParallel [11], which is available through BiocManager [12].

Colour scheme selection and manipulation was performed with colorspace [13, 14].

Copy number analysis was performed with sequenza [15], FACETS [16, 17] and PURPLE [18]. Some analysis was also directly performed with copynumber [19, 20].

Variant effect prediction was performed with VEP [21].

Table manipulation was performed with `data.table` [22].

Violin plots were generated with `vioplot` [23].

Heatmaps and UpSet plots were generated with `ComplexHeatmap` [24]

Phylogenetic analysis was performed with both `ape` [25] and `phangorn` [26] followed by `dendextend` [27].

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` [28].

Additional libraries, which were used for a multitude of things are listed in no particular order below: `Rsamtools` [29], `GenomicRanges` [30], `optparse` [31], `VariantAnnotation` [32], `MultiAssayExperiment` [33], `circize` [34], `BioQC` [35], `Biostrings` [36], `deconstructSigs` [37], `BSgenome` [38], `QDNaseq` [39], `RColorBrewer` [40], `pheatmap` [41], `ensemblVEP` [42], `stringdist` [43], `Rsubread` [44], `svglite` [45], `grImport` [46], `XML` [47], `kableExtra` [48], `lsa` [49], `irlba` [50], `ggplot2` [51]

## **python**

Analysis for Chapter 4 was mostly done through `python` [52] with the help of many different packages, which are listed here in no particular order: `numpy` [53], `ncls` [54], `pysam` [55–57], `zarr` [58], `pandas` [59, 60], `quadprog` [61] as well as `scipy` [62].

## **latex**

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

# Contents

<b>Abstract</b>	<b>i</b>
<b>Declaration of Authorship</b>	<b>ii</b>
<b>Preface</b>	<b>iii</b>
<b>Acknowledgements</b>	<b>vii</b>
<b>List of Figures</b>	<b>xiii</b>
<b>List of Tables</b>	<b>xvii</b>
<b>List of Equations</b>	<b>xviii</b>
<b>List of Listings</b>	<b>xviii</b>
<b>Abbreviations</b>	<b>xx</b>
<b>1 Introduction</b>	<b>1</b>
1.1 DNA . . . . .	1
1.1.1 Ploidy . . . . .	4
1.1.2 Mutations . . . . .	5
1.2 cfDNA . . . . .	6
1.3 DNA sequencing . . . . .	8
1.3.1 Library preparation . . . . .	9
1.3.2 Next generation sequencing . . . . .	10
1.3.3 Long read sequencing . . . . .	10
1.4 DNA analysis . . . . .	12
1.4.1 Mapping . . . . .	12
1.4.2 Variant calling . . . . .	13
1.4.3 Germline . . . . .	13
1.4.4 Somatic . . . . .	13
1.5 Cancer . . . . .	15
1.6 Thesis overview and aims . . . . .	24

<b>2</b>	<b>Joint somatic variant calling</b>	<b>25</b>
2.1	Introduction . . . . .	25
2.2	Publication . . . . .	26
2.2.1	Summary . . . . .	27
2.2.2	FreeBayesSomatic workflow . . . . .	27
2.2.3	Strelka2Pass workflow . . . . .	28
2.2.4	Validation . . . . .	29
2.3	Effects on downstream analysis . . . . .	40
2.3.1	Phylogenetic reconstruction . . . . .	40
2.4	Longitudinal analysis . . . . .	43
2.4.1	Clonal deconvolution . . . . .	45
2.4.2	Longitudinal enriched phylogeny . . . . .	47
2.5	Usage . . . . .	48
<b>3</b>	<b>CASCADE</b>	<b>50</b>
3.1	Introduction . . . . .	50
3.1.1	Lung cancer . . . . .	51
3.2	Publications . . . . .	52
3.3	Patient level analysis . . . . .	52
3.3.1	Analysis workflow . . . . .	53
3.3.2	Patient CA-A . . . . .	57
3.3.3	Patient CA-I . . . . .	66
3.3.4	Patient CA-J . . . . .	74
3.3.5	Patient CA-K . . . . .	82
3.3.6	Patient CA-L . . . . .	91
3.4	Mitochondrial phylogenetic reconstruction . . . . .	99
3.4.1	Method . . . . .	100
3.4.2	Results . . . . .	101
3.4.3	Summary . . . . .	105
3.5	Cohort level analysis . . . . .	105
3.6	Outlook . . . . .	107
<b>4</b>	<b>Mismatchfinder</b>	<b>109</b>
4.1	Introduction . . . . .	109
4.1.1	Mutational signature analysis . . . . .	109
4.1.2	Restrictions and pitfalls of standard signature analysis . . . . .	110
4.1.3	Overview . . . . .	111
4.2	Methods . . . . .	111
4.2.1	Mathematical concept . . . . .	111
4.2.2	Data preprocessing . . . . .	112
4.2.3	Mismatch detection . . . . .	113
4.2.4	Filtering steps . . . . .	113
4.2.5	Consensus reads . . . . .	114
4.2.6	Germline filtering . . . . .	115
4.2.7	Count normalisation . . . . .	115
4.2.8	Signature deconvolution . . . . .	116
4.2.9	Signature detection . . . . .	118
4.2.10	Tumour detection . . . . .	120
4.3	Results . . . . .	120
4.3.1	Simulated Data - the validation promised land . . . . .	120

4.3.2	Real world data - analysis of patient data . . . . .	127
4.3.3	Tumour detection analysis . . . . .	138
4.4	Summary . . . . .	141
<b>5</b>	<b>Conclusion</b>	<b>142</b>
<b>Appendices</b>		<b>164</b>
<b>A</b>	<b>Joint somatic variant calling - publication</b>	<b>165</b>
A.1	Introduction . . . . .	166
A.2	Materials and methods . . . . .	167
A.2.1	FreeBayesSomatic workflow . . . . .	167
A.2.2	Strelka2Pass workflow . . . . .	168
A.3	Validation . . . . .	169
A.3.1	Simulated data . . . . .	170
A.3.2	Clinical data . . . . .	171
A.4	Discussion . . . . .	173
A.5	Supplementary methods . . . . .	177
A.5.1	Alignment of clinical data . . . . .	177
A.5.2	Validation of clinical data . . . . .	177
A.5.3	Purity estimation with sequenza . . . . .	178
A.5.4	Performance of individual steps in Strelka2Pass . . . . .	178
A.5.5	Ensemble workflows – user suggestions . . . . .	178
<b>B</b>	<b>Joint somatic variant calling - supplementary data</b>	<b>180</b>
<b>C</b>	<b>CASCADE - supplementary data</b>	<b>181</b>
C.1	supplementary methods . . . . .	181
C.2	CASCADE - supplementary figures . . . . .	181
C.2.1	Patient CA-A . . . . .	181
C.2.2	Patient CA-I . . . . .	190
C.2.3	Patient CA-J . . . . .	196
C.2.4	Patient CA-K . . . . .	200
C.2.5	Patient CA-L . . . . .	205
<b>D</b>	<b>MisMatchFinder - supplementary data</b>	<b>208</b>
D.1	ROI bed files generation . . . . .	208
D.2	Oligo-nucleotide context normalisation . . . . .	208
D.3	Germline filtering with zarr . . . . .	208
D.3.1	Zarr conversion with scikit-allele . . . . .	211
D.3.2	MisMatchFinder filtering - the zarr API . . . . .	212
D.3.3	Data simulation . . . . .	212
D.3.4	Signature simulation - we can spike this punch . . . . .	212
D.3.5	Blacklist generation from healthy samples . . . . .	213
D.3.6	Patient data subsampling . . . . .	213

# List of Figures

1.1	Overview DNA structure . . . . .	2
1.2	Overview Chromosome structure . . . . .	3
1.3	Overview DNA structure . . . . .	4
1.4	Origins of cell-free and circulating tumour DNA schematic; Figure adapted from Racheljunewong - Own work, CC BY-SA 4.0, <a href="https://commons.wikimedia.org/w/index.php?curid=56676758">https://commons.wikimedia.org/w/index.php?curid=56676758</a> . . . . .	7
1.5	Fragment size distribution of ctDNA . . . . .	8
1.6	Library preparation for NGS . . . . .	9
1.7	Sequencing by synthesis (Illumina) . . . . .	11
1.8	Drawing of central nervous system metastasis . . . . .	19
1.9	Original hallmarks of cancer . . . . .	22
1.10	Newest hallmarks of cancer . . . . .	23
1.11	Intra patient heterogeneities in cancer . . . . .	24
2.1	Comparison of joint multi-sample and single tumour-normal paired variant calling methods . . . . .	30
2.2	Characteristics of simulated data . . . . .	31
2.3	Performance of workflows using simulated data . . . . .	32
2.4	Variant allele frequencies (VAF) of variants detected by joint sample analysis . . . . .	33
2.5	Performance of individual steps in the Strelka2pass workflow using the simulated data . . . . .	34
2.6	Summary of variant filters assigned by Mutect2 . . . . .	35
2.7	Assessing the performance of different workflows using tumour samples with different evolutionary relationships in the simulated data . . . . .	36
2.8	Correlation of variant allele frequencies in validation . . . . .	37
2.9	Performance of the different workflows using clinical samples from eight cancer patients . . . . .	38
2.10	Correlation between cellularity and proportion of variants found only with joint calling using FreeBayesSomatic . . . . .	39
2.11	Improvement in recall using FreeBayesSomatic and Strelka2pass over Mutect2 in the clinical samples. . . . .	39
2.12	Reconstructed phylogenies of joint samples . . . . .	41
2.13	Tanglegram of the reconstructed phylogenies . . . . .	42
2.14	Timeline from diagnosis till death for patient CA-F . . . . .	44
2.15	Improved somatic variant calling in longitudinal data . . . . .	45
2.16	Longitudinal data informs diagnostic variant calling . . . . .	46
2.17	Reconstructed clonal trees for joint and pairwise variant calling . . . . .	47
2.18	Reconstructed phylogeny with longitudinal ctDNA samples . . . . .	48
2.19	Usage statistics joint workflows . . . . .	49
3.1	Timeline of patient CA-A from diagnosis until death . . . . .	57
3.2	Allelic frequencies of driver and emerging resistance mutations . . . . .	58
3.3	PET scans of patient CA-A before and during Selpercatinib treatment . . . . .	58

3.4	Schematic of tumour lesions in patient CA-A . . . . .	59
3.5	Phylogeny of autopsy samples from patient CA-A . . . . .	60
3.6	Heatmap of driver gene variants in patient CA-A . . . . .	61
3.7	Circos plot of patient CA-A sample 11 . . . . .	62
3.8	Circos plot of patient CA-A sample 11 without allele frequency filter . . .	63
3.9	Cancer cell fraction of mutation clusters of clonal tree for patient CA-A .	65
3.10	Timeline of patient CA-I from diagnosis until death . . . . .	66
3.11	Schematic of tumour lesions in patient CA-I . . . . .	67
3.12	Phylogeny of autopsy samples from patient CA-I . . . . .	68
3.13	Heatmap of driver gene variants in patient CA-I . . . . .	69
3.14	Circos plot of patient CA-I sample dx . . . . .	71
3.15	Circos plot of patient CA-I sample 557 . . . . .	72
3.16	Clonal evolutionary tree CA-I . . . . .	73
3.17	Cancer cell fraction of mutation clusters of the clonal tree for patient CA-I	73
3.18	Timeline of patient CA-J from diagnosis until death . . . . .	74
3.19	Blood plasma analysis of patient CA-J . . . . .	75
3.20	Schematic of analysed tumour lesions in patient CA-J . . . . .	75
3.21	Phylogeny of autopsy samples from patient CA-J . . . . .	77
3.22	Heatmap of driver gene variants in patient CA-J . . . . .	78
3.23	Circos plot of patient CA-J sample 2 . . . . .	79
3.24	Circos plot of patient CA-J sample 20 . . . . .	80
3.25	Cancer cell fractions of individual mutations for patient CA-J . . . . .	81
3.26	Timeline of patient CA-K from diagnosis until death . . . . .	82
3.27	Schematic of analysed tumour lesions in patient CA-K . . . . .	84
3.28	Phylogeny of autopsy samples from patient CA-K . . . . .	85
3.29	Heatmap of driver gene variants in patient CA-K . . . . .	86
3.30	Circos plot of patient CA-K sample 1 . . . . .	88
3.31	Circos plot of patient CA-K sample 8 . . . . .	89
3.32	Clonal evolutionary tree CA-K . . . . .	90
3.33	Cancer cell fraction of mutation clusters of clonal tree for patient CA-K .	90
3.34	Timeline of patient CA-L from diagnosis until death . . . . .	91
3.35	Schematic of analysed tumour lesions in patient CA-L . . . . .	92
3.36	Phylogeny of autopsy samples from patient CA-L . . . . .	93
3.37	Heatmap of driver gene variants in patient CA-L . . . . .	94
3.38	Circos plot of patient CA-L sample P.1 . . . . .	95
3.39	Circos plot of patient CA-L sample P.2 . . . . .	96
3.40	Clonal evolutionary tree CA-L . . . . .	97
3.41	Cancer cell fraction of mutation clusters of clonal tree for patient CA-L .	98
3.42	Average coverage of mitochondrial DNA of CASCADE patients . . . . .	100
3.43	Mitochondrial and somatic phylogenetic reconstruction of CA-A . . . . .	102
3.44	Mitochondrial and somatic phylogenetic reconstruction of CA-I . . . . .	102
3.45	Mitochondrial and somatic phylogenetic reconstruction of CA-J . . . . .	103
3.46	Mitochondrial and somatic phylogenetic reconstruction of CA-K . . . . .	104
3.47	Mitochondrial and somatic phylogenetic reconstruction of CA-L . . . . .	104
3.48	Percentage of LOH per chromosome in CASCADE patients . . . . .	106
4.1	Trinculeotide count contributions for single base substitution (SBS) sig- nature 7a . . . . .	110
4.2	Schematic of consensus computation method for overlapping reads . . . .	115
4.3	Distance of deconvolution methods from truth . . . . .	117

4.4	Mismatchrate of different filtering methods . . . . .	121
4.5	Signature analysis of spike-in somatic variants . . . . .	122
4.6	Signature weight differences for different deconvolution methods . . . . .	123
4.7	Signature weights differences from normal for SBS7a spike-in . . . . .	124
4.8	Signature weights differences from normal for SBS3 spike-in . . . . .	125
4.9	Signature analysis without germline variant filtering . . . . .	125
4.10	Percent increase of mismatches in analysis with and without germline filter	126
4.11	Signature weights of the normal sample with and without germline filter .	127
4.12	PCA of tri-nucleotide mismatch counts of real world data (PC1 and PC2)	129
4.13	Mismatchrates of healthy samples by age . . . . .	131
4.14	Blacklisted mismatches from healthy individuals . . . . .	131
4.15	Somatic variants found in germline sites . . . . .	134
4.16	Signature weights for the WGS of two BRCA1 mutation positive breast cancer patients . . . . .	135
4.17	Signature weights for subsampled BRCA1 positive patients . . . . .	136
4.18	Signature weights for the WES of two melanoma patients . . . . .	138
4.19	Signature weights of lcWGS of two melanoma samples . . . . .	138
4.20	SBS3 signature weight distribution in healthy and breast cancer samples .	141

## Appendices

164

A.1	Comparison of joint multi-sample variant calling and single tumour-normal paired calling methods . . . . .	169
A.2	Performance of ensemble variant calling strategies . . . . .	175
B.1	Schematic of analysed tumour lesions in patient CA-F . . . . .	180
C.1	Circos plot of patient CA-A sample 26 . . . . .	183
C.2	Circos plot of patient CA-A sample 31 . . . . .	184
C.3	Circos plot of patient CA-A sample 41 . . . . .	185
C.4	Circos plot of patient CA-A sample 47 . . . . .	186
C.5	Circos plot of patient CA-A sample 55 . . . . .	187
C.6	Circos plot of patient CA-A sample 57 . . . . .	188
C.7	Circos plot of patient CA-A sample 59 . . . . .	189
C.8	Number of somatic variants per sample in patient CA-I . . . . .	190
C.9	Phylogeny of samples from patient CA-I with diagnostic sample . . . . .	190
C.10	Circos plot of patient CA-I sample 559 . . . . .	191
C.11	Circos plot of patient CA-I sample 566 . . . . .	192
C.12	Circos plot of patient CA-I sample 573 . . . . .	193
C.13	Circos plot of patient CA-I sample 579 . . . . .	194
C.14	Circos plot of patient CA-I sample 583 . . . . .	195
C.15	Circos plot of patient CA-J sample 24 . . . . .	196
C.16	Circos plot of patient CA-J sample 28 . . . . .	197
C.17	Circos plot of patient CA-J sample 32 . . . . .	198
C.18	Circos plot of patient CA-J sample 42 . . . . .	199
C.19	Circos plot of patient CA-K sample 4 . . . . .	200
C.20	Circos plot of patient CA-K sample 5 . . . . .	201
C.21	Circos plot of patient CA-K sample 6 . . . . .	202
C.22	Circos plot of patient CA-K sample 9 . . . . .	203



C.23	Circos plot of patient CA-K sample 13 . . . . .	204
C.24	Circos plot of patient CA-L sample 8 . . . . .	205
C.25	Circos plot of patient CA-L sample 17A . . . . .	206
C.26	Circos plot of patient CA-L sample 26 . . . . .	207
D.1	Trinucleotide count contributions for single base substitution (SBS) signature 3 . . . . .	214
D.2	Signature weights differences from normal for SBS7a spike-in . . . . .	214
D.3	Signature weights differences from normal for SBS3 spike-in . . . . .	215
D.4	PCA of tri-nucleotide mismatch counts of real world data (PC2 and PC3) . . . . .	216
D.5	Fitted beta distribution for Signature SBS3 in healthy samples . . . . .	217
D.6	Fitted beta distribution for Signature SBS17a in healthy samples . . . . .	218
D.7	Fitted beta distribution for Signature SBS12 in healthy samples . . . . .	219
D.8	Fitted beta distribution for Signature SBS46 in healthy samples . . . . .	220
D.9	Signature detection of variants categorised by presence in gnomAD . . . . .	221

# List of Tables

3.1	Autopsy samples sequenced for patient CA-A . . . . .	59
3.2	Copy number analysis results for patient CA-A . . . . .	64
3.3	Autopsy samples sequenced for patient CA-I . . . . .	66
3.4	Copy number analysis results for patient CA-I . . . . .	70
3.5	Autopsy samples sequenced for patient CA-J . . . . .	76
3.6	Copy number analysis results for patient CA-J . . . . .	81
3.7	Somatic variants found in plasma with AVENIO sequencing for patient CA-K . . . . .	83
3.8	Autopsy samples sequenced for patient CA-K . . . . .	83
3.9	Copy number analysis results for patient CA-K . . . . .	87
3.10	Autopsy samples sequenced for patient CA-L . . . . .	92
3.11	Copy number analysis results for patient CA-L . . . . .	97
4.1	Germline variants retained after germline filtering . . . . .	133
4.2	Confusion matrix for MisMatchFinder leave one out validation on mela- noma training set . . . . .	139
4.3	Confusion matrix for ichorCNA leave one out validation on melanoma training set . . . . .	139
4.4	Confusion matrix for MisMatchFinder leave one out validation on breast cancer training set . . . . .	140
4.5	Confusion matrix for ichorCNA leave one out validation on breast cancer training set . . . . .	140
A.1	Sample name mapping . . . . .	176
A.2	Runtime of different workflows on simulated data . . . . .	177
C.1	Lung cancer genes for CASCADE analysis . . . . .	182
D.1	Dinucleotide counts of GRCh38 . . . . .	209
D.2	Trinucleotide counts of GRCh38 . . . . .	210

## List of Equations

2.1	FreeBayesSomatic: $\text{LOD}_{normal}$ . . . . .	27
2.2	FreeBayesSomatic: $\text{LOD}_{tumour}$ . . . . .	27
2.3	FreeBayesSomatic: somaticLOD definition . . . . .	27
2.4	FreeBayesSomatic: $\text{VAF}_{tumour}$ . . . . .	28
2.5	FreeBayesSomatic: somaticVAF definition . . . . .	28
2.6	Strelka2Pass: pairwise error probability . . . . .	29
2.7	Strelka2Pass: joint error probability . . . . .	29
2.8	Strelka2Pass: joint SomEVS . . . . .	29
3.1	Inclusion criteria for cluster of PhylogicNDT analysis . . . . .	56
3.2	Selective pressure with effective population size . . . . .	99
3.3	Mitochondrial variants based distance function of two samples . . . . .	101
4.1	MisMatchFinder: number of mismatches . . . . .	111
4.2	MisMatchFinder: sequencing error . . . . .	111
4.3	MisMatchFinder: germline variants . . . . .	112
4.4	MisMatchFinder: number of mismatches with distributions . . . . .	112
4.5	MisMatchFinder: number of mismatches correlation with somatic variants	112
4.6	MisMatchFinder: optimisation for signature weights . . . . .	117
4.7	MisMatchFinder: optimisation function restrictions . . . . .	117
4.8	MisMatchFinder: quadratic programming formula . . . . .	117
4.9	Beta distribution probability density function . . . . .	118
4.10	Beta distribution cumulative density function . . . . .	119
4.11	Inverse beta distribution cumulative density function . . . . .	119
4.12	MisMatchFinder: CDF-score calculation per signature and patient . . .	119

# Listings

B.1	parse strelka VCF . . . . .	180
B.2	annotate variants with copy number calls . . . . .	180
B.3	convert to maf format . . . . .	180
C.1	Preprocessing of mitochondrial reads and variants for analysis in R . . . .	181
D.1	scikit-allel conversion vcf_to_zarr . . . . .	211
D.2	field options for reduced memory . . . . .	211
D.3	spike-in variant selection . . . . .	212
D.4	bamsurgeon spike-in . . . . .	213
D.5	Blacklist postprocessing . . . . .	213

## Abbreviations

<b>APOBEC</b>	<b>A</b> Po <b>l</b> ipoprotein <b>B</b> mRNA <b>E</b> ding enzyme, <b>C</b> atalytic polypeptide-like
<b>BAM</b>	<b>B</b> inary <b>A</b> lignment <b>M</b> ap
<b>bp</b>	<b>b</b> ase <b>p</b> air
<b>BQ</b>	<b>B</b> ase <b>Q</b> uality
<b>CASCADE</b>	<b>C</b> Ancer ti <b>S</b> sue <b>C</b> ollection <b>A</b> fter <b>D</b> Eath
<b>CCF</b>	<b>C</b> ancer <b>C</b> ell <b>F</b> raction
<b>cfDNA</b>	<b>c</b> ell <b>f</b> ree <b>D</b> N <b>A</b>
<b>ChIP</b>	<b>C</b> hromatin <b>I</b> mmuno <b>P</b> recipitation
<b>CT</b>	<b>C</b> omputed <b>T</b> omography
<b>ctDNA</b>	<b>c</b> irculating tumour <b>D</b> N <b>A</b>
<b>DBS</b>	<b>D</b> ouble <b>B</b> ase <b>S</b> ubstitution
<b>DNA</b>	<b>D</b> eoxyribo <b>N</b> ucleic <b>A</b> cid
<b>F81</b>	<b>F</b> elsenstein <b>1981</b> model
<b>GATK</b>	<b>G</b> enome <b>A</b> nalysis <b>T</b> ool <b>K</b> it
<b>HGSOC</b>	<b>H</b> igh <b>G</b> rade <b>S</b> erous <b>O</b> varian <b>C</b> arcinoma
<b>HKY85</b>	<b>H</b> asegawa, <b>K</b> ishino and <b>Y</b> ano <b>1985</b> model
<b>HPC</b>	<b>H</b> igh <b>P</b> erformance <b>C</b> omputing
<b>H&amp;E</b>	<b>H</b> ematoxylin and <b>E</b> osin
<b>ILM</b>	<b>I</b> terative <b>L</b> inear <b>M</b> odels
<b>InDel</b>	<b>I</b> nsertion or <b>D</b> eletion
<b>LN</b>	<b>L</b> ymph <b>N</b> ode
<b>MQ</b>	<b>M</b> apping <b>Q</b> uality
<b>MRCA</b>	<b>M</b> ost <b>R</b> ecent <b>C</b> ommon <b>A</b> ncessor
<b>NGS</b>	<b>N</b> ext <b>G</b> eneration <b>S</b> equencing
<b>NJ</b>	<b>N</b> eighbour <b>J</b> oining
<b>NSCLC</b>	<b>N</b> on- <b>S</b> mall <b>C</b> ell <b>L</b> ung <b>C</b> ancer

<b>PCA</b>	<b>P</b> rincipal <b>C</b> omponent <b>A</b> analysis
<b>PCx</b>	<b>P</b> rincipal <b>C</b> omponent number x
<b>PET</b>	<b>P</b> ositron <b>E</b> mission <b>T</b> omography
<b>PON</b>	<b>P</b> anel <b>O</b> f <b>N</b> ormals
<b>QP</b>	<b>Q</b> uadratic <b>P</b> rogramming
<b>RAID</b>	<b>R</b> edundant <b>A</b> rray of <b>I</b> ndependent <b>D</b> isks
<b>RNA</b>	<b>R</b> ibo <b>N</b> ucleic <b>A</b> cid
<b>ROI</b>	<b>R</b> egion <b>O</b> f <b>I</b> nterest
<b>RPRS</b>	<b>R</b> ead <b>P</b> osition <b>R</b> ank <b>S</b> um
<b>SBS</b>	<b>S</b> ingle <b>B</b> ase <b>S</b> ubstitution
<b>SCLC</b>	<b>S</b> mall <b>C</b> ell <b>L</b> ung <b>C</b> ancer
<b>SCT</b>	<b>S</b> mall <b>C</b> ell <b>T</b> ransformation
<b>SNP</b>	<b>S</b> ingle <b>N</b> ucleotide <b>P</b> olymorphism
<b>SNV</b>	<b>S</b> ingle <b>N</b> ucleotide <b>V</b> ariant
<b>SV</b>	<b>S</b> tructural <b>V</b> ariant
<b>TAS</b>	<b>T</b> argeted <b>A</b> mplicon <b>S</b> equencing
<b>TKI</b>	<b>T</b> yrosine <b>K</b> inase <b>I</b> nhibitor
<b>TNBC</b>	<b>T</b> riple <b>N</b> egative <b>B</b> reast <b>C</b> ancer
<b>UPGMA</b>	<b>U</b> nweighted <b>P</b> air <b>G</b> roup <b>M</b> ethod with <b>A</b> rithmetic mean
<b>UV</b>	<b>U</b> ltra <b>V</b> iolet light
<b>VCF</b>	<b>V</b> ariant <b>C</b> all <b>F</b> ormat
<b>VEP</b>	<b>V</b> ariant <b>E</b> ffect <b>P</b> redictor
<b>WES</b>	<b>W</b> hole <b>E</b> xome <b>S</b> equencing
<b>WGD</b>	<b>W</b> hole <b>G</b> enome <b>D</b> oubling
<b>WGS</b>	<b>W</b> hole <b>G</b> enome <b>S</b> equencing
<b>WPGMA</b>	<b>W</b> eighted <b>P</b> air <b>G</b> roup <b>M</b> ethod with <b>A</b> rithmetic mean

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

— Lewis Carroll, *Alice in Wonderland*

# 1

## Introduction

This first introduction chapter contains all the necessary background information <sup>c1</sup>and <sup>c1</sup>as well as an overview of the work discussed in this thesis. It summarises <sup>c2</sup>the basic biological <sup>c2</sup>Text added. 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 <sup>c3</sup>-free context. Section 1.3 <sup>c3</sup>Text added. introduces the current technologies used to measure and detect DNA and its variations. Section 1.4 covers the computational methods developed to read out changes in the DNA. Then section 1.5 <sup>c4</sup> relates how these changes lead to cancer and what we can <sup>c4</sup>fixed reference (R3) learn from them. The introduction concludes with section 1.6 as an overview of the aims of the thesis and my work in addressing them in the following chapters.

### 1.1 DNA as an information storage unit

It is a widely accepted fact <sup>c5</sup> that Deoxyribonucleic acid (DNA) serves as the long <sup>c6</sup>-term <sup>c5</sup> information storage molecule of our cells. This information is protected and allows <sup>c7</sup>the <sup>c6</sup>Text added. correction of simple errors through its double helix structure [63, 64]. The nucleotides, <sup>c7</sup>Text added. 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 <sup>c8</sup> <sub>2</sub> are used to encode <sup>c8</sup>Text added. the genetic information into DNA. Each <sup>c9</sup> strand <sup>c10</sup> mirrors the other, so <sup>c11</sup>an adenine <sup>c9</sup> of the <sup>c10</sup> s <sup>c11</sup> that will be paired up with a thymine forming two hydrogen bonds. Similarly, cytosine



Figures/intro/DNAstructure.pdf

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; Hydrogen (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 pair with guanine forming an even stronger connection with three hydrogen bonds.

While other pairings which do not follow those rules are chemically possible, they are

<sup>c12</sup>primarily observed in ribonucleic acid (RNA) [65]. These <sup>c13</sup>stringent bonding rules

enable the DNA to be similar to a hard drive with backup on a computer. <sup>c14</sup>Further-

more, 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<sup>c15</sup> by proofreading the newly synthesised strand with the template.

In order <sup>c16</sup> to distinguish the two strands, they were assigned the names 3' and 5' <sup>c17</sup>re-

lated to the numbering of the <sup>c18</sup>exposed carbon atom in the sugar<sup>c19</sup> (Figure 1.1).

The <sup>c1</sup>entire DNA encoding the organism is commonly called “the genome“, with all

genes, <sup>c2</sup>consisting of introns and exons. All exonic regions aggregated are termed “the

exome“<sup>c3</sup>. Unicellular organisms usually have a <sup>c4</sup> small number of introns, which to

current knowledge, only provide limited information and are only responsible for <sup>c5</sup>the

secondary and tertiary structure. In vertebrates, introns, as well as intergen<sup>c6</sup>ic DNA

<sup>c12</sup> mostly  
<sup>c13</sup> very  
strict  
<sup>c14</sup> And

<sup>c15</sup> ;  
<sup>c16</sup> to be  
able  
<sup>c17</sup> depend-  
ing on

<sup>c18</sup> Text  
added.  
<sup>c19</sup> , which  
is exposed

<sup>c1</sup> entirety  
of the  
<sup>c2</sup> which  
consist  
<sup>c3</sup> split  
sentence for  
clarity (R3)

2

<sup>c4</sup> very  
<sup>c5</sup> Text  
added.  
<sup>c6</sup> e



(the DNA between genes) contribute most of the DNA in the genome. For example, in humans, only 1% of the genetic material is considered to be exonic, whereas introns contribute  $\approx 24\%$ , and the rest is intergenic ( $\approx 75\%$ ) [66].

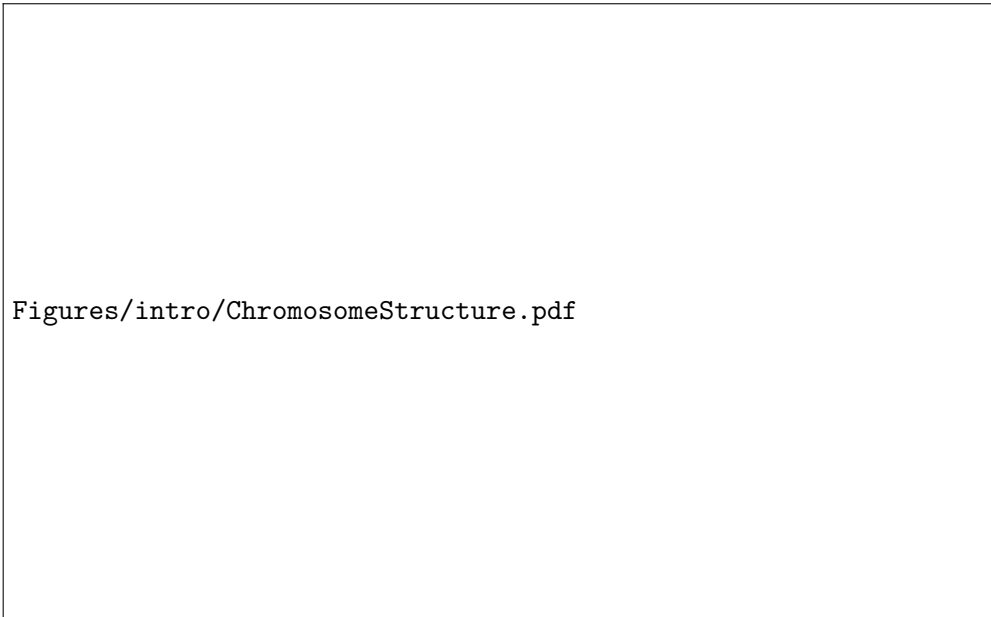


FIGURE 1.2: Structural overview of the metaphase condensed chromosome: DNA is first wrapped around Histones to form nucleosome, which then associate with each other to form the chromatin fiber, which in the metaphase of the cell cycle is condensed even more into the X-shaped chromosome

The DNA in eukaryotes<sup>c1</sup>, however is not free<sup>c2</sup>-floating around in the nucleus of a cell, but rather in most eukaryotic 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<sup>c3</sup> X-like shape, with two shorter and two longer arms (Figure 1.2). This <sup>c4</sup>configuration allows some DNA to be accessible, whereas other areas can be restricted [67]. Through this restriction, the availability of <sup>c5</sup>specific genes, which are the sections of the DNA, which encode for short-term storage molecules like RNA, can be controlled. This restriction plays an <sup>c6</sup>essential 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 <sup>c7</sup>cell's function and <sup>c8</sup> identity <sup>c9</sup> [68].

Even though the X-like structure is the most commonly used and known <sup>c10</sup>, the DNAs 3D structure is usually very different and only takes this shape for a very short time in the cell cycle. Most of the time, the chromosomes are unravelled into something



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

resembling a ball of yarn, where the “open” chromatin regions are on the outside, and the “closed” regions are “hidden” on 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 chromosomes,<sup>c11</sup> called a chiasma.

<sup>c11</sup> ~~which is~~

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

Similar to the already discussed organisation of DNA in two strands, another concept of data security<sup>c1</sup> involving ploidy (the number of complete chromosome sets in a cell)<sup>c1</sup> is also implemented. Most eukaryotic organisms have at least two of each chromosome (diploid), with some species reaching up to septaploid [69]. 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,<sup>c2</sup>so the<sup>c2</sup> ~~such that~~ resulting cell will be diploid. However, the ploidy of the somatic cells is not as uniform within a species, where it can vary between organisms based on gender or rank [70].

In most organisms, a change in ploidy is fatal [71], and only partial ploidy changes are tolerated. In humans, the only chromosomal scale aberrations compatible with life are extra copies of chromosome 17 [72], chromosome 18 [73] and chromosome 21 [74] are tolerated. These syndromes can occur when there is an uneven split of chromosomes during cell division.

However, this concept is not the only reason for the ploidy of somatic cells. The additional advantage, apart from sexual reproduction, is that a second almost identical copy of a chromosome allows repair of DNA, even when both strands are damaged, for example, in a double-strand break. In this case, the information from the sister chromosome will be used by first cutting the double-strand break ends to have an overhang (resection). This overhang will then merge with the sister chromosome's mirrored strand. In this state, the two chromosomes are fused in a Holliday junction, which allows the missing part from the resection and the double-strand break to be repaired [75]. 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 <sup>c1</sup>a vital role in evolution and diversity [76, 77]. <sup>c1</sup> ~~an~~ important

### 1.1.2 “Fantastical” mutations and where to find them

Even though the DNA is highly stable, and error correction methods are constantly working not to introduce any changes in the DNA, the source of evolution and adaptation of species relates to a steady mutation rate [78, 79]. These changes in normal tissue, known as somatic mutations, are mostly irrelevant to the organism <sup>c2</sup> and will not be passed on to the next generation. Somatic mutations accumulate linearly <sup>c3</sup>for the <sup>c4</sup>cell's lifespan <sup>c5</sup> and are not bound to just cell divisions [80–82]. In contrast, if one of those mutations occurs in the germline cell, e.g. sperm or egg-producing cells, these mutations will be propagated to all offspring and be present in all cells of that organism and, in turn, 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 their size, ranging from single nucleotide polymorphisms (SNPs) to small insertions or deletions (InDels) and large structural changes, like the deletion of parts of or even a whole chromosome arm. As previously described, smaller changes usually have less

<sup>c2</sup> ~~as a~~ whole  
<sup>c3</sup> ~~over the~~ course of  
<sup>c4</sup> ~~Text~~ added.  
<sup>c5</sup> ~~of the cell~~

impact on the overall fitness of the organism. However, even SNPs can lead to changes which are not compatible with life [83, 84].

## 1.2 Cell free DNA is more than just bits and pieces

When a cell from a multicellular organism dies, through whichever method, there will be numerous enzymes involved, which clear the debris and recycle material. This <sup>c1</sup>digestion cascade means that proteases <sup>c2</sup>break down 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 when it is released following cell death. 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 <sup>c3</sup>reducing accessibility, similar to how they stop <sup>c4</sup> the access for transcription into RNA. This nucleosomal pattern then in turn, leads to the DNA being cut in the linker regions between nucleosomes into fragments, mainly in the length of 167 base pairs (bp).

These DNA fragments, <sup>c5</sup> called cell free DNA (cfDNA), can then be readily detected in bodily fluids, like blood, urine or even stool. By analysing these fragments, non-invasive tests for prenatal care have been made possible, as the DNA of the developing foetus can be detected in the mother's blood [85, 86].

Similar to this process, a cancer also sheds <sup>c6</sup> “circulating tumour DNA“ (ctDNA), when its cells die, either through <sup>c7</sup>immune system intervention, cancer therapies or other processes. These ctDNA fragments can similarly be analysed and molecular properties measured without even knowing the <sup>c8</sup>tumour's exact location <sup>c9</sup>. As a blood test can be routinely performed in the clinic, <sup>c10</sup> monitoring <sup>c11</sup> cancer progression is significantly easier and safer <sup>c12</sup>with ctDNA than other measures. Of course, it is similar to the prenatal test, acting as a proxy for the cells <sup>c13</sup> still alive, which have not yet shed their DNA. Additionally, the amount of shed <sup>c14</sup> DNA is highly variable between tumours, with a general <sup>c15</sup>ly higher amount seen in later stages when <sup>c16</sup>the tumour burden is high [87, 88].

<sup>c17</sup>The higher tumour burden leads to a higher number of tumour cells turning over and releasing their DNA. As tumour cells are generally resistant to apoptosis (Section 1.5<sup>c18</sup>),

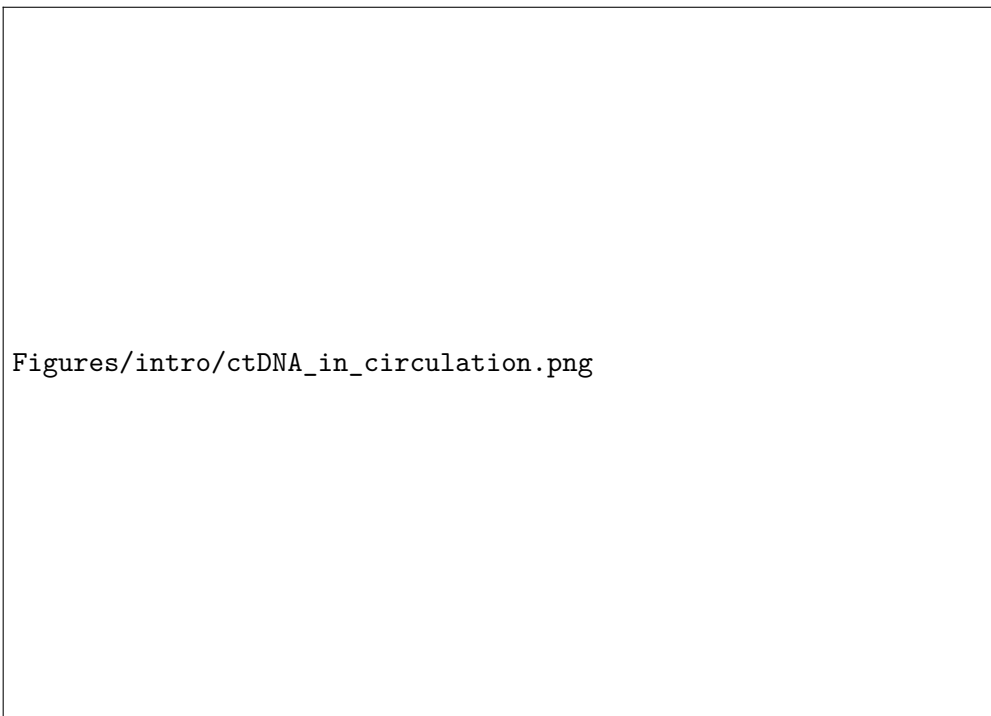
most tumour cells release their DNA in a less organised manner like necroptosis [89]<sup>c19</sup>, ferroptosis [90]<sup>c20</sup>, and autophagy [91]<sup>c21</sup>. Finally some DNA is secreted by cells, both cancer and normal, through exosomes [92]<sup>c22</sup>. The relative contributions of each of these processes to the total amount of ctDNA is unclear and it is likely, that the composition varies between cancer types and patients.

<sup>c19</sup> Text added.

<sup>c20</sup> Text added.

<sup>c21</sup> Text added.

<sup>c22</sup> Text added.



Figures/intro/ctDNA\_in\_circulation.png

FIGURE 1.4: Origins of cell-free and circulating tumour DNA schematic; Figure adapted from Racheljunewong - Own work, CC BY-SA 4.0, <https://commons.wikimedia.org/w/index.php?curid=56676758>

Due to the different biological processes which can lead to the release of DNA from cancer cells, in addition to apoptosis, which is the <sup>c1</sup>primary source of cfDNA from healthy cells, ctDNA has several biological differences from cfDNA. The most prominent features observed are distinct ctDNA fragmentation size patterns, fragment start sites, and fragment ends motives. While the preferred end motive and start site are heavily correlated, restricting the analysis to the lower tail of the mono- and di-nucleosomal peak of the fragment size distribution (74-155bp and 240-325bp) allowed a ctDNA tumour signal enrichment of at least 28%. This <sup>c2</sup> enrichment and the high periodicity of the distribution showed a high dependence on nucleosomal placement. All ctDNA features combined were shown <sup>c3</sup> to predict the presence or absence of tumour DNA in samples regardless of tumour type (Figure 1.5, [93, 94]).

<sup>c1</sup> main

<sup>c2</sup> is

<sup>c3</sup> to be able

<sup>c4</sup>Using the ctDNA contained in a blood sample as a non-invasive method of detection

<sup>c4</sup> Text added.

Figures/intro/fragmentSizeDist.pdf

FIGURE 1.5: Fragment size distribution of 5 high grade serous ovarian carcinoma (HGSOC) patients and panel of healthy controls. The vertical dashed lines are placed on the fragment sizes between 52 and 172 bp where 10 bp periodicity is observed. The vertical lines at 240 and 324 bp show the range at which a shift in the di-nucleosomal peak occurs between HGSOC patients and healthy controls. The inset plot enlarges the density profile in the di-nucleosomal fragment length range. Figure adopted from Markus et al. [93]

and monitoring of a patients cancer status has been used in many recent clinical trials, either to determine tissue of origin in cancers of unknown primary [95]<sup>c5</sup>, predict clinical outcomes [96]<sup>c6</sup> and MRD detection. While ctDNA should theoretically be able to be used as a screening tool for early detections of cancers and risk groups, the general consensus is, that current methods are not sensitive enough for a population wide screening effort [97–99]<sup>c7</sup>.

<sup>c5</sup> Text added.  
<sup>c6</sup> Text added.

<sup>c7</sup> Text added.

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

As we know the building blocks that make DNA <sup>c1</sup>and the processes 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 <sup>c2</sup> used the lambda phage to fuse known ends for the primers needed for the reaction to the piece of DNA and supplied labelled nucleotides [100]. This method was then superseded by "Sanger sequencing" after Frederick Sanger, who, with colleagues, published this method in 1977. Through adding dideoxynucleotides in a low concentration, the polymerase chain reaction would terminate trying to integrate these nucleotides, and by labelling them radioactively or fluorescently, a gel could then be

<sup>c1</sup> ~~as well as~~

<sup>c2</sup> ~~to make use of this~~

used to determine the sequence of the piece of DNA [101, 102], which made the method better suited for large-scale projects.

However, this method had multiple issues for modern research questions. Mostly, <sup>c1</sup> ~~that~~ it was fairly labour-intensive and time-consuming to analyse multiple pieces of DNA <sup>c2</sup> simultaneously, making it very challenging to sequence all the DNA of an organism. <sup>c2</sup> ~~at the same time~~ The human genome project, which was started in 1990, used machines that automated the Sanger sequencing procedure, and it still took hundreds of researchers 13 years to complete the DNA sequence of just one human [66, 103]. Even though this was a very long project, it laid the groundwork for <sup>c3</sup> using the current sequencing technologies. <sup>c3</sup> ~~the usage of~~

### 1.3.1 Library preparation - what we learned from using phages

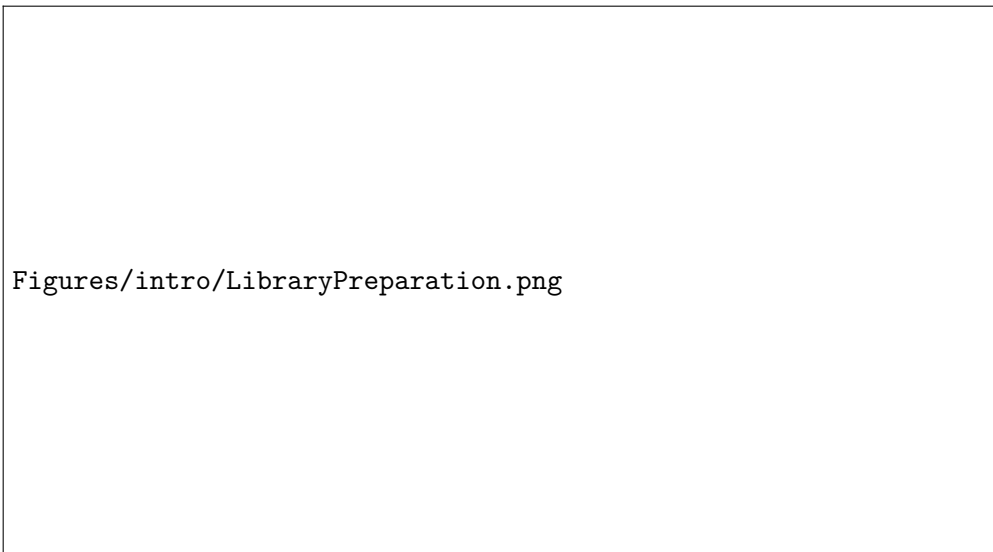


FIGURE 1.6: Adapter ligation during library preparation. The adapters are added to the DNA insert during library preparation. A. The DNA insert is prepared by adding an A-tail and phosphorylation. B. The adapter complex which includes the P5/P7 flow cell binding adapter is added to the DNA insert. C. 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" [104]

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 <sup>c4</sup> the sequencing of ctDNA, <sup>c4</sup> ~~Text added.~~

as discussed in Section 1.2, as the DNA is unbound and already digested into short fragments. Once the DNA is ready, it is phosphorylated, and an A-tail is added, before the adapter complex is ligated. This DNA-tail enables the DNA to bind to the flow cell which is covered with the reverse complement of the adapter (Figure 1.6).

### 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-<sup>c1</sup>DNA interactions (ChIP) and other epigenome studies. The term NGS is still widely used, even though it has been more than <sup>c2</sup>ten years since the first NGS approach was commercially available. While at the beginning of next-generation sequencing, there were multiple approaches, the current lion's share (80% of sequencing data) of protocols use the Illumina short read sequencing by synthesis approach (Figure 1.7) [105, 106], which is based on the concept of alternating integration of fluorescently labelled nucleotides and imaging with a microscope (Figure 1.7), as well as multiplexing, where a DNA fragment is ligated to an index, which allows sequencing of multiple samples at once [107, 108], as it is shown in Figure 1.6. This method enables highly accurate determination of the sequence of a DNA fragment and, depending on the flow cell and sequencing machine, allows one to sequence a whole genome in just 24h.

### 1.3.3 Long read sequencing - the "third" generation sequencing

By now, multiple methods <sup>c3</sup>that broke free of the size limitations of NGS exist, <sup>c4</sup> commonly referred to as long read sequencing. Most <sup>c5</sup> current methods trade the very high accuracy of the second-generation NGS methods for the capability of sequencing huge continuous strands of DNA (current record 2.3 Million bp [109]) with more typical library preparation ranging between 10-30 Kbp. These methods are expected to revolutionise our understanding of the highly repetitive elements <sup>c6</sup> in the genome, such as the centromeres of chromosomes. Methods such as the direct molecule sequencing approach by Oxford Nanopore <sup>c7</sup>can even distinguish post-transcriptional modifications on RNA [110]. However, for ctDNA, which is highly fragmented, these methods offer only limited advantages over <sup>c8</sup> short read sequencing.





Figures/intro/SequencingBySynthesis.jpg

FIGURE 1.7: 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 [105]

## 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 stretch<sup>c9</sup> far. However, all methods first need to infer the location in the genome the sequenced piece of DNA originated from. <sup>c10</sup>Unfortunately, as the current methods randomly fragment the DNA (Section 1.3.1), the genomic location information is completely lost. This process is referred to as mapping. <sup>c9</sup> es <sup>c10</sup> Text added.

### 1.4.1 Mapping - inferring genomic location of a read

In this process, the fragments of DNA, which were sequenced, are assigned a genomic coordinate on the reference genome. This <sup>c1</sup>alignment is only possible because we have resolved genome sequences (Section 1.3) for a high number of species. The location a sequenced piece of DNA matches to the reference genome might be unique, but it could also align to multiple locations, due to highly repetitive regions or <sup>c2</sup> the existence of pseudogenes with almost 100% identity concordance. In addition <sup>c3</sup>, the reference genome might not accurately reflect the genome of the organism that has been sequenced. Each mapping position is <sup>c4</sup> assigned a quality score, which reflects how likely it is to be the actual position of the sequence. As Illumina sequencers <sup>c5</sup>can sequence both ends of the DNA fragment, the position of the ends (read 1 and read 2) relative to each other can also be used to infer the quality, as they should be within a reasonable distance to each other (Figure 1.6). <sup>c1</sup> Text added. <sup>c2</sup> due to <sup>c3</sup> to this <sup>c4</sup> therefore <sup>c5</sup> have the ability to

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-mappers, which are based on  $k$ -mers, <sup>c6</sup> predefined DNA sequences of length  $k$ , which help to identify <sup>c7</sup>specific regions of interest. These tools are <sup>c8</sup>widespread for RNAseq, where the exact location of a read does not matter, only that the read is within a gene [111, 112], but also for methods that estimate <sup>c9</sup>the similarity between sequences (DNA, RNA or protein) [113, 114]. <sup>c6</sup> which are <sup>c7</sup> certain <sup>c8</sup> especially common <sup>c9</sup> Text added.

For this work, however, the exact position of reads is crucial, so only precise mapping methods like BWA [115] or Bowtie 2 [116], which are optimised for short reads from Illumina systems, provide the necessary functions.

### 1.4.2 Variant calling - “Spotting the difference”

As intra-species genetic variation is intended for adaptation and evolution, there will be places where the DNA sequence of the subject will differ from the sequence of the reference (see Section 1.1.2). In the context of cancer these variants can give insights into the development and progression of cancer, and treatment options for patients and can even be used to guide family planning. Depending on the type of variation <sup>c10</sup> of <sup>c10</sup> ~~that is~~ 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 mononuclear layer from the <sup>c1</sup>subject’s blood, but <sup>c2</sup> almost any cell can be used for this process, <sup>c1</sup> ~~blood of the subject~~ as all cells in the organism will share all germline variants (Section 1.1.2). The only <sup>c2</sup> ~~really~~ important input <sup>c3</sup>in addition to the DNA sequence from the sequencer is the <sup>c4</sup>organism’s reference genome <sup>c5</sup>, as all variant nomenclature is based on the reference and the ploidy <sup>c6</sup> (Section 1.1.1). The ploidy is key to understanding the ranges of allele frequencies at which 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%, whereas the hexaploid commercial wheat [117] allele frequency for variants would be 16%, 33%, 50%, 66%, 0.83% and 100%. Due to the random sampling and possible sequencing errors, the observed allele frequencies may differ from the theoretical values. Most state-of-the-art germline variant calling methods will also use haplotype reconstructions through de-Bruijn graphs, which feature a remapping of reads in relation to each other [118–122] where the original mapping location assigned by the aligner (Section 1.4.1) is only used as a guideline. This <sup>c7</sup>realigning allows one to resolve even complex haplotypes of the sample by not restricting the method to the linear setup of the reference genome. <sup>c3</sup> ~~on top of~~ <sup>c4</sup> ~~Text added.~~ <sup>c5</sup> ~~of the organism~~ <sup>c6</sup> ~~of the organism~~ <sup>c7</sup> ~~Text added.~~

### 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 <sup>c8</sup>same change in <sup>c8</sup> ~~Text added.~~ nucleotide. The allele frequency is instead a measure of the subclonal size. A *subclone* is

defined as a set of cells <sup>c9</sup> derived from the cell, <sup>c10</sup>that originally acquired the somatic mutation. Depending on the selective advantage, random drift and also the time point when the variant was introduced, these clones can be very variable in size, and therefore, their contribution within the DNA sequenced could vary. <sup>c11</sup>Furthermore, as not all cells have the variants, the selection of the tissue for library preparation is <sup>c12</sup>essential, unlike for germline calling. In the <sup>c13</sup>cancer setting, somatic variant calling is important in understanding the genetic landscape of cancer samples, where the main question is which changes are present in the tumour and which lead to disease.

<sup>c9</sup> ~~which~~  
<sup>c10</sup> ~~were~~  
<sup>c10</sup> ~~which~~  
<sup>c11</sup> ~~Text~~  
<sup>c11</sup> ~~added.~~  
<sup>c12</sup> ~~very~~  
<sup>c12</sup> ~~important~~  
<sup>c13</sup> ~~setting~~  
<sup>c13</sup> ~~of cancer~~

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 and the mononuclear layer from blood<sup>c1</sup> so <sup>c2</sup>that true somatic variants can be characterised (Section 1.4.3). These two samples are then analysed together, and only changes <sup>c3</sup> in the somatic tumour sample and not in the normal sample are reported. Even though this concept sounds simple, there are some pitfalls [123]. First, there might be some tumour “contamination” in the normal sample, which needs to be adjusted for [120, 124]. Second, there might be normal “contamination” in the tumour sample, which means that not all cells in the tumour sample are actually tumour. This contamination means that the signal of the tumour specific changes may be reduced and harder to find.

<sup>c1</sup> ~~,~~  
<sup>c2</sup> ~~Text~~  
<sup>c2</sup> ~~added.~~  
<sup>c3</sup> ~~that are~~  
<sup>c3</sup> ~~only~~

All of these issues are amplified in the scenario when there is no “normal” sample available, either because the patient did not consent due to other medical issues or because for diagnostic tests, there is often no need for a germline sample. In this case, there is the option for “tumour only” variant calling, which utilises a database of germline variants in the population, to distinguish between somatic and germline variants, as the variant calling is very similar to just germline variant calling (Section 1.4.3) without the restriction of the ploidy. However, even with an extensive database like gnomAD [125] it is unlikely to be able to remove all germline variants from the analysis and as there is no direct comparison, the precision of the “tumour only” method is significantly lower [126].

<sup>c4</sup>Recent studies have shown that the amount of somatic variants in germline cells have been underestimated. The previous assumption of somatic mutations accumulating in cells mainly due to cell division has been shown to be wrong. Abascal et al. [127] <sup>c5</sup>showed

<sup>c4</sup> ~~Text~~  
<sup>c4</sup> ~~added.~~  
<sup>c5</sup> ~~Text~~  
<sup>c5</sup> ~~added.~~

that even non dividing cells like smooth muscle cells acquire somatic mutation at a stable rate over time, where others like sperm are more restricted and show a lower somatic mutation rate. And even that the overall somatic mutation rate of a species is directly tied to its lifespan [82]. <sup>c6</sup>These studies required a significantly different set-up for detection of somatic variants, because there is no “normal“ sample, and instead either single cell colony or highly specific bulk library preparation has to be performed for within sample somatic variant calling. These recent scientific advances highlighted the need for new methods for somatic variant calling even in germline cells. <sup>c6</sup> Text added.

## 1.5 Cancer

For a long time in human history, the origin of cancer as a disease was a mystery, and <sup>c1</sup>many theories, starting in ancient Egypt, were developed. These theories ranged from a curse to chemical imbalance over parasites to trauma. In this section, I will outline the history of cancer as a disease and its treatments starting with ancient times and leading up to the current times. While the first steps are very wide because <sup>c2</sup> biology itself was not understood, it is quite curious how often people with more knowledge came to worse conclusions and theories than were already known thousands of years ago. <sup>c1</sup> a multitude of <sup>c2</sup> the

Around 3000BC, the Egyptians describe the bulging tumour of the breast as an incurable disease [128]. Even then, they already had some ointments, which were used, including resection, cauterisation and salting of the affected areas, all of which were still used until the 19<sup>th</sup> century [129]. This papyrus document is considered the oldest evidence of cancer in humanity. <sup>c3</sup> up

When the ancient Greeks laid the foundation for modern medicine with Hippocrates, the first hypothesis about natural causes of cancers was formulated, and the terms “cancer“ and “carcinoma“ were coined. The abundance and accumulation of “black bile“ in the body <sup>c4</sup>were thought to <sup>c5</sup> cause <sup>c6</sup> the cancers. However, the treatment was still the same as before, with resection and lotions [130]. <sup>c4</sup> was <sup>c5</sup> be the <sup>c6</sup> of

Following Hippocrates, the Roman physician Celsus progressed the understanding of cancers significantly by describing metastatic relapse of treated breast cancer in neighbouring armpits and even the spread to distant organs. <sup>c7</sup>Additionally, he also was <sup>c7</sup> Text added.

aware that the outcome of patients was better, if the tumours were removed early and aggressively [131].

With the destruction of the western Roman Empire, the Middle East became well known for its <sup>c1</sup>substantial advances towards modern medicine.<sup>c2</sup> The court physician to the Emperor of Constantinople, Aetius <sup>c3</sup>succeeded with the first total mastectomy and <sup>c4</sup>was generally <sup>c5</sup> an advocate of the total excision of tumours [132].

<sup>c1</sup> strong  
<sup>c2</sup> split sentence  
<sup>c3</sup> had success  
<sup>c4</sup> Text added.  
<sup>c5</sup> was

Sadly, while both the understanding of cancer and the treatment were steadily improving, the Pope prohibited bloodshed as well as surgeries, and this led to a slow-down of advances, especially because autopsies were also forbidden a hundred years later in 1305. However, there were still illegal experiments conducted and the general classification which is still used currently, was first proposed by Henri de Mondeville, who started classifying tumours by their anatomical site [133].

After the end of the “dark ages“, the wide availability of older medical works from both the Greeks and Romans due to the book print invention led to the re-emergence of the use of chemical ointments and lotions on cancer lesions. Paracelsus laid the groundwork for <sup>c6</sup> modern chemotherapy by promoting the usage of chemicals, which he <sup>c7</sup> warned were poisonous in the wrong concentration, for the treatment of cancer [134].

<sup>c6</sup> the  
<sup>c7</sup> himself

When the <sup>c8</sup>the church no longer banned the dissection of corpses, more <sup>c9</sup> cases of “hidden“ causes of death were found post-mortem, which were often cancers on internal organs, like the brain. The detection of malignant versus benign tumours was also <sup>c10</sup> a major breakthrough. This led to the understanding that benign tumours might turn malignant after some time, and many physicians suggested <sup>c11</sup>the removal of the benign growths [135].

<sup>c8</sup> dissection of corpses was no longer banned by the church  
<sup>c9</sup> and more  
<sup>c10</sup> Text added.  
<sup>c11</sup> Text added.

Due to the apparent genetic disposition of cancer, especially breast cancer, two independent physicians (Zacutus Lusitani and Nicholas Tulp) came to the conclusion, that cancer is contagious and proposed <sup>c12</sup>the isolation of patients [136, 137], which shows that while the treatment of cancer was improving steadily, the origin of the disease was still a mystery. It took until 1700 when Deshaies-Gendron described cancer as a transformation of a normal body part, which continues to grow without control and while he was aware of metastatic disease, he suggested no treatment, as he did not believe cancer to be curable with drugs [138].

<sup>c12</sup> Text added.

However, it took almost 150 years after the theory of cancer being contagious for Nooth [139]<sup>c13</sup> to conduct experiments attempting to infect himself with cancer pieces resected from another person, which proved that cancers generally are not infectious [139].

<sup>c13</sup> change citation style

Another ground-breaking work published in the same year was the collection of almost three thousand autopsy reports and their clinical history, which contained <sup>c1</sup>many detailed cancer cases, including: brain, head and neck, lung, breast, esophagus, stomach, colon, liver, pancreas, kidney, uterus, cervix, bladder, and prostate. Many of the terms used by Theophilus Boneti to describe the cases are still in use, and the work itself was the first step toward tumour pathology [140]

<sup>c1</sup> a-number of

With the invention and, consequently common use of the microscope in pathology, more and more causes of death were identified as caused by cancer. An example is the connection of <sup>c2</sup>chronic cough to lung cancer and swollen joints with sarcoma [141].

<sup>c2</sup> a-

<sup>c3</sup>Moreover, after more and more autopsies of cancer patients, surgeons like Heister [142] found that breast cancer resection needs to include the breast, the axillary lymph node and the pectoralis major muscle, which got to be known as the Halsted radical mastectomy and was the standard of care for a long time.

<sup>c3</sup> Text added.

While the treatment of cancers (mostly surgical) was getting <sup>c4</sup>increasingly advanced, the origin and cause of cancer in patients was still very much debated. As the aetiology of cancer is complex, as we now know, it is maybe not surprising that it took longer, but by the middle of the 18<sup>th</sup>-century, chronic inflammation as a cause of cancer initiation was hypothesised [143].

<sup>c4</sup> more-and more

The next big step was taken when in 1838, the concept of cells as fundamental building blocks of organisms was published. In the following years, many cancers were dissected and microscopically analysed. This <sup>c5</sup>increase in analysis revealed that tumour cells look vastly different from normal cells and it was thought that their morphological features could serve to identify their fate and became known for defining the cellular origin of benign and malignant tumours. <sup>c6</sup>Furthermore while Müller described the tumours as a collection of abnormal cells with stroma, he thought cancer to arise from newly generated cells from a diseased organ and thought the underlying cause to be “amorphous embryonal blastema” [144].

<sup>c5</sup> Text added.

<sup>c6</sup> And

With this foundation, over the next hundred years, <sup>c7</sup>many advances were made in the

<sup>c7</sup> lots-of

morphology of different tumours, and many previously undetected ones, like leukemia, were found and extensively characterised. However, even then, <sup>c8</sup>some researchers <sup>c9</sup> <sup>c8</sup> ~~there~~ <sup>c9</sup> ~~were~~ <sup>c9</sup> ~~who~~ understood that the heterogeneity of cancers is so vast that while he was convinced that the microscope would be a mandatory instrument to diagnose cancers, more effort to collect and study specimens was necessary to have a complete picture [145].

As many shared the view of Bennett, the second half of the 19<sup>th</sup> century was a rich source of surgical pathology and the oncology literature in general. Most outstanding was Rudolf Virchow's "Die krankhaften Geschwülste" [146], which was the first landmark book on the classification of cancers and is still a well of knowledge. From his work, the terms "hyperplasia" and "metaplasia" were derived as pre-cancerous states of cells. He also was one of the first to hypothesise the presence of growth-stimulating substances around cancers, which lead to their uncontrolled growth. Virchow was the first <sup>c1</sup>to <sup>c1</sup> ~~to~~ <sup>c1</sup> ~~Text added.~~ oppose the "amorphous embryonal blastema" theory and instead was convinced that tumour cells were just abnormally changed cells, which he called "chronic irritation theory", and believed that metastases were seeded by the original lesion (like in this melanoma Figure 1.8). He also had <sup>c2</sup>a major scientific impact in <sup>c3</sup>several other fields <sup>c2</sup> ~~Text added.~~ like Parasitology, Forensic and Anthropology<sup>c3</sup>. <sup>c3</sup> ~~a number of~~

While the search for possible cancer-causing substances started to gain more and more interest, only one real cause was thought to be found in the ore of the central European mountains, where miners would have a higher prevalence of lung cancers. Nevertheless, this was later found to be caused by radioactive material and not by the inhaled dust of minerals as expected. Similar<sup>c4</sup>ly, many parasites and bacteria were found as potential <sup>c4</sup> ~~Text added.~~ causes of cancer, but none of the findings could produce proof.

While all these steps were getting closer together in time up until the beginning of the 20<sup>th</sup> century, they were still fairly minor in contrast to the high speed of discoveries that the last hundred years brought with <sup>c5</sup>them. While technically, Willhelm Röntgen <sup>c5</sup> ~~it~~ discovered the X-rays just before the change of the century [147], its impact on the human body and cancer was only clear a few years into the last century [148, 149]. However, similar to how X-rays can cause cancers, researchers also <sup>c6</sup>quickly found <sup>c7</sup> that <sup>c8</sup>they <sup>c6</sup> ~~Text added.~~ can also treat cancer; thus the field of radiotherapy was created. <sup>c9</sup>Radiotherapy then <sup>c7</sup> ~~quickly~~ <sup>c8</sup> ~~it~~ <sup>c9</sup> ~~This~~

---

<sup>c3</sup> Maybe surprising to hear, that he was strongly opposed to Darwin's theory of evolution. In his own words: "The intermediate form is unimaginable save in a dream... We cannot teach or consent that it is an achievement that man descended from the ape or other animal."





FIGURE 1.8: Drawing of central nervous system metastasis from page 121, Volume 2 of “Die krankhaften Geschwülste“ Virchow [146]; translated original caption: Fig. 128: Multiple melanoma of the Pia mater basilaris, most pronounced around the Medulla oblongata, the Pons, the Fossa Sylvii, Fissura longit (sample No. 256a from 1858); Fig. 129: Lower end of spinal cord of Fig. 128 with multiple melanoma of the soft skin with node like growths at the nerve roots (sample No. 256b from 1858)

was the first <sup>c10</sup>significant change in cancer treatment for around five thousand years, <sup>c10</sup> ~~major~~ which also could treat inoperable cancers.

The <sup>c1</sup>following invention <sup>c2</sup> I want to highlight within the vast amount of advances <sup>c1</sup> ~~next~~ made in the advent of the 20<sup>th</sup> century, is the cutting needle aspiration syringe, which <sup>c2</sup> ~~that~~ allowed a non-traumatic biopsy of internal organs for microscopy study. This biopsy method made it possible <sup>c3</sup>not to <sup>c4</sup> have exploratory surgery and <sup>c5</sup> allowed improved <sup>c3</sup> *Text added.* diagnosis and planning of necessary surgery. <sup>c4</sup> ~~not~~ <sup>c5</sup> ~~instead~~

The next major step in the treatment of cancers comes in the form of chemotherapy when Ehrlich [150] published his work “Beiträge zur experimentellen Pathologie und Chemotherapy“, where he injected animals with different toxins in order to destroy cancer cells. Although, it still took another 30 years until after the second world war

when it was discovered, that a chemical designed for warfare also had a potent anti-tumour effect.

In the meantime, the first long-term tissue cultures of animal cancer cell lines were established and further insights like the Warburg effect [151] <sup>c1</sup>were found, which showed <sup>c1</sup> *Text added.* that cancer cells use glucose at a higher rate than healthy cells. This effect ultimately led to the discovery of the positron emission tomography (PET) scan, which allowed a significantly more granular diagnosis and localisation of cancerous lesions than before.

With the success of growing human cell lines in vitro, the USA embarked on a massive experiment to test any potential source of chemical carcinogenesis. However, at the same time, multiple viruses were identified to cause cancers in the 1950s, when electron microscopy was invented [152].

Only a few more years later, the biggest advance in the understanding of biology was made when the structure of DNA was discovered [63] (Section 1.1), <sup>c2</sup>which subsequently <sup>c2</sup> *and* lead to numerous new experiments and breakthroughs. <sup>c3</sup>For example, when studying <sup>c3</sup> *Text added.* how viruses <sup>c4</sup>can reverse transcribe their RNA and insert a new gene into a healthy <sup>c4</sup> *are able to* cell, which then transformed into a cancer cell, the term “oncogene“ was coined [153–155] and the foundation for the understanding of how genes influence the emergence of cancers was laid. This knowledge also lead to the understanding that heritable changes in the genome could predispose a person to cancer, which was previously hypothesised [156]. And while the discovery of DNA was a substantial boost for the understanding of cancer, the diagnostic capabilities increased at a similar speed, with urine tests for biomarkers of certain cancers as well as antigen detection.

<sup>c5</sup>Lastly this is when we arrive at the “current“ times when a few years ago, next- <sup>c5</sup> *And* generation sequencing (NGS) (Section 1.3) was introduced and sped up data generation to improve our understanding of the genetic landscapes of different cancers and the subsequent development of genomic tests based on the molecular profile of certain cancers. The completion of the “Human Genome Project“ (Section 1.3) allowed the accurate resolution of common mutations in cancers, and consortiums coordinated large efforts to build databases of all somatic variations like TCGA, ICGC and PCAWG [157].

<sup>c6</sup>These efforts then allowed the field to classify cancers depending on their tumour mutational burden (TMB), as not all cancer types are mutationally driven. While a typical <sup>c6</sup> *Text added.*

akute myloid leukemia (AML) patient might only present with less than five somatic variants, a melanoma patient's cancer can show thousands of somatic mutations. In addition to somatic variations in cancer, there are many germline variants which cause a predisposition to cancer, like *BRCA1/2*[158]<sup>c7</sup> or *TP53*[156]<sup>c8</sup> and while some cancers have hundreds of thousands of somatic variants, they will also have millions of germline variants[159]<sup>c9</sup>. The combinatory space spanned by the somatic and germline variants of the cancer means that no cancer patients is like another and is known as inter-patient heterogeneity.

<sup>c7</sup> Text added.

<sup>c8</sup> Text added.

<sup>c9</sup> Text added.

However, with the genetic classification and investigation of cancers, researchers observed genetic heterogeneity in addition to the already well-established inter-patient heterogeneity[160]. Not all cancer cells in a patient share the same genetic setup, and this allows the disease to adjust and adapt to treatments and ultimately causes resistance [161]. <sup>c1</sup>While in some patients the genetic heterogeneity is already present before treatment and selective pressure on the subclonal structure of the disease changes the composition[162]<sup>c2</sup>, in others the resistance to treatment is likely non-genetic, but over time the cancer acquires a genetic resistance mechanism[163]<sup>c3</sup>.

<sup>c1</sup> Text added.

<sup>c2</sup> Text added.

<sup>c3</sup> Text added.

Genetic pathology nevertheless enabled the identification of cancer vulnerabilities that then allowed the application of highly specific drugs, like tyrosine kinase inhibitors (TKI)s, which are tailored to target a specific alteration in the genome of a cancer cell, and genetically engineered antibodies <sup>c4</sup>that can <sup>c5</sup>hone in on the cancer.

<sup>c4</sup> which

<sup>c5</sup> be

While the therapeutic world is quickly evolving, many of the questions from previous times are still the same. We still <sup>c6</sup>do not know how and when the heterogeneity in cancers occurs. We <sup>c7</sup>,however, know it is a major source of resistance to treatment. In <sup>c8</sup>most cancer types, we <sup>c9</sup>still do not have an answer to the “cell of origin“ question that has been asked for so long.


<sup>c6</sup> don't

<sup>c7</sup> just

<sup>c8</sup> the majority of

<sup>c9</sup> also

So instead of trying to answer these questions directly, there has been an effort to define fundamental features of malignancies very similar to the early pathology descriptions. The original characteristics comprise 1. Sustaining proliferative signalling 2. Evading growth suppressors 3. Activating invasion and metastasis 4. Enabling replicative immortality 5. Inducing angiogenesis 6. Resisting cell death (Figure 1.9).



Figures/intro/oldHallmarksCancer.jpg

FIGURE 1.9: Acquired capabilities of cancer; Functional capabilities acquired by most cancers during their development; Figure adapted from Hanahan and Weinberg [164]

These hallmarks were for a while considered the core of tumour development, and the authors themselves hypothesised, that these core mechanics allow us to condense the complexity that cancer displays, both in the clinic and in labs, with a small set of rules, which all cancers have to obey [164]. In their exact words: “We foresee cancer research developing into a logical science, where the complexities of the disease, described in the laboratory and clinic, will become understandable in terms of a few underlying principles“

However, with 11 years of additional research into the topic, more hallmarks have been found, and <sup>c1</sup>[the authors revised the original list](#) to contain additional characteristics, namely 1. Avoiding immune destruction 2. Tumour-promoting inflammation 3. Genome instability and mutation 4. Deregulating cellular energetics [165]. And even then, a few years later, even more hallmarks, e.g. metabolic rewiring, are now considered a part of the characteristics of cancer [166].

<sup>c1</sup> the  
original list  
was revised  
by the  
authors

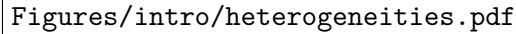
And even during the time of my PhD, further research revealed additional hallmarks, which got characterised by Hanahan [167]. The newest version adds another two characteristics and hallmarks, specifically: 1. unlocking phenotypic plasticity 2. nonmutational epigenetic reprogramming 3. polymorphic microbiomes 4. senescent cells (see Figure 1.10).

The evolution of these hallmarks shows that even though much time and effort was invested in cancer research for multiple centuries, there is still no unifying definition and treatment for cancer. <sup>c2</sup>Additionally, the vast heterogeneity not only between cancer <sup>c2</sup> *Text added.* types but also between patients makes it very hard to study. More so, even within <sup>c3</sup>a <sup>c3</sup> *Text added.* patient there is <sup>c4</sup>a <sup>c4</sup> *Text added.* third type of heterogeneity, which is the <sup>c5</sup>leading <sup>c5</sup> *main* cause of treatment resistance and relapse [1]. Spatial heterogeneity refers to the non-uniform distribution of cancer subclones within different sites of disease. Even when the primary <sup>c6</sup>disease <sup>c6</sup> *site of disease* site is known, metastatic sites may show a very different genetic landscape. In contrast, temporal heterogeneity describes the changes <sup>c7</sup>in <sup>c7</sup> *of* a lesion over time. These changes can be caused by cancer progression or treatment (Figure 1.11). These two concepts, however, are not mutually exclusive, but can occur at the same time.

And while we know that this diversity exists and efforts have been made to measure and classify them [169], there is still a lack of methods to directly analyse this heterogeneity

Figures/intro/NewestHallmarksOfCancer.jpg

FIGURE 1.10: Emerging hallmarks and enabling characteristics of cancer; updated version of the hallmarks figure (Figure 1.9, [164]); Figure adapted from Hanahan [167]; Left, the Hallmarks of Cancer currently embody eight hallmark capabilities and two enabling characteristics. In addition to the six acquired capabilities – Hallmarks of Cancer – proposed in 2000 (Figure 1.9), the two provisional “emerging hallmarks” introduced in 2011 [165] –cellular energetics (now described more broadly as “reprogramming cellular metabolism”) and “avoiding immune destruction” – have been sufficiently validated to be considered part of the core set. Given the growing appreciation that tumors can become sufficiently vascularized either by switching on angiogenesis or by co-opting normal tissue vessels [168], this hallmark is also more broadly defined as the capability to induce or otherwise access, principally by invasion and metastasis, vasculature that supports tumor growth. The 2011 sequel further incorporated “tumor-promoting inflammation” as a second enabling characteristic, complementing overarching “genome instability and mutation,” which together were fundamentally involved in activating the eight hallmark (functional) capabilities necessary for tumor growth and progression. Right, this review incorporates additional proposed emerging hallmarks and enabling characteristics involving “unlocking phenotypic plasticity,” “nonmutational epigenetic reprogramming,” “polymorphic microbiomes,” and “senescent cells.”



Figures/intro/heterogeneities.pdf

FIGURE 1.11: Types of intra patient heterogeneity in cancer: Spatial uneven distribution of subclones in different sites (left); Temporal heterogeneity with regards to genetic landscape within one site, either through progression of tumour or adaptation to treatment; Cells with rough membrane depict cancer cells, smooth membrane are normal cell. Colours show subclonal differences in the cancer. Figure adapted from Dagogo-Jack and Shaw [1]

and utilise this information to guide clinical approaches directly.

## 1.6 Thesis overview and aims

While genomic tumour heterogeneity is now a well accepted concept, there remains a need for computational methods to assess and visualise this heterogeneity. This work aims to contribute to this unmet need by developing three different custom made tools to infer and monitor genomic tumour heterogeneity. I have completed the work in the following three parts:

1. Development of two joint somatic variant calling workflows and the impact of these on downstream analysis (Chapter 2).
2. Analysis of five rapid autopsy probands with state of the art methods to investigate multi-regional tumour heterogeneity and development of a mitochondrial based phylogeny reconstruction method (Chapter 3).
3. Development of a read-centric method to detect somatic mutational signatures from low coverage whole genome sequencing (Chapter 4).

*“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*

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

## 2.1 Introduction

In 2018, at the start of the work presented in this thesis, we observed a difference in methodology between germline and somatic variant calling methods. Where all "modern" germline variant callers, like Strelka2 [120], HaplotypeCaller [170], DRAGEN [171] and DeepVariant [172], have the built-in capability to jointly call multiple related samples, for example from family trios, virtually no somatic variant caller had this functionality.

The joint analysis of smaller cohorts improves the performance of germline variant calling methods significantly, by allowing to assess technical artefacts, which might be unique for the individual sequencing machine or the researcher handling the DNA [173, 174]. Additionally, as certain parts of the genome are more problematic to sequence (Section 1.3) and map (Section 1.4.1), a “control” sample can help to distinguish if a certain observed change occurring frequently is a technical issue or in fact a real change.

For somatic variant calling, this concept has been adopted on in the genome analysis toolkit (GATK) [175] to allow the use of a panel of normals (PON), which contain frequently seen changes in healthy (“normal”) individuals analysed with the same sequencing technology [176]. Although, in contrast to the more intricate model for the germline equivalent, this is a post processing step of the analysis. Mutect2, which is the most recent somatic variant calling algorithm provided by the Broad institute, also provides a multi-sample mode, for which all tumour samples need to be from the same patient, either related longitudinally or spatially [177]. However, this mode is not very well publicised and all tutorials released by the developers state that “there is currently no way to perform joint calling for somatic variant discovery” [123]. So while all methods in the GATK are considered a beta feature, the multi sample mode needs to be used with care.

There are only two methods currently, which have documented and published capabilities to jointly analyse tumour samples from the same patient to call somatic variants. The

first one is a specialised method built on a joint bayesian model for single nucleotide variants (SNV)s that occur in multiple samples called multiSNV [178]. However, it has multiple shortcomings, which make it not usable for our data. First, as the name suggests, the method can only jointly evaluate SNVs and completely ignores INDELs and structural variants, which would be acceptable for the superior performance it provides. However, multiSNV was optimised only for WES and not for the very deep WGS that is now widely available and form a major part of this thesis. This mismatch of input types means exceptionally high runtimes on WGS data. Even with custom parallelisation that was attempted in this work, the predicted runtime for just one multi sample patient would have been longer than 3 years. This shows, that while multiSNV was a great step forward at the time, there is a real need for new methods to stem the tide of sequencing data available due to the ever decreasing sequencing cost.

multiSNV had been the only software available for multi sample analysis for almost five years, but during this work, superFreq [179] was also published. It combines all standard analysis steps for tumour analysis, like quality assessment, variant calling, copy number analysis and clonal deconvolution, into one program and is even able to jointly analyse samples. However, similar to multiSNV, its focus during optimisation and development was on WES and RNAseq data, so when applied to WGS data, we could not find a server node with enough memory to execute the workflow.

This then prompted us to develop a novel workflow to enable the analysis of high depth WGS, which we estimate to become more and more routine, with the ever dropping prices of sequencing. The following sections will show the development and validation of the joint variant calling methods as described in Hollizeck et al. [180] (Section 2.2), and additional analysis on the impact of the joint variant calling on downstream multi-regional tumour analysis (Section 2.3), longitudinal sample analysis (Section 2.4) and clonal deconvolution (Section 2.4.1). The final section provides current information on the usage of the methods by others in the research community (Section 2.5).

## 2.2 Publication

The full publication related to the joint somatic variant calling can be found at <https://doi.org/10.1093/bioinformatics/btab606> and a formatted version is also attached as Appendix A with all supplementary methods.



References to supplementary data will be prefixed with the letter A for tables and methods taken from the publication, supplementary figures from the publication are integrated into the main text and not shown in the appendix. Prefix B was used for additional data generated for this work.

### 2.2.1 Summary

To enable highly sensitive, fast and accurate variant detection from multiple related tumour samples, we have developed joint variant calling extensions to two widely used single-sample algorithms, FreeBayes [118] and Strelka2 [120]. Using both simulated and clinical sequencing data, we show that these workflows are highly accurate and can detect variants at much lower variant allele frequencies than other commonly used methods.

### 2.2.2 FreeBayesSomatic workflow

The original FreeBayes algorithm can jointly evaluate multiple samples, but routinely it does not perform somatic variant calling on tumour-normal pairs. We introduce FreeBayesSomatic which allows concurrent analysis of multiple tumour samples by adapting concepts from SpeedSeq [181] which differentiates the likelihood of a variant between tumour and normal samples instead of imposing an absolute filter for all variants called in the normal. Hence, for each genotype (GT) at SNV sites, FreeBayesSomatic first calculates the difference in likelihoods (LOD) between the normal (Equation 2.1) and the tumour (Equation 2.2) samples genotype likelihoods (GL) with  $g_0$  describing the reference genotype.

$$\text{LOD}_{\text{normal}} = \max_{g_i \in \text{GT}} (\text{GL}(g_0) - \text{GL}(g_i)) \quad (2.1)$$

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

$$\text{somaticLOD} := (\text{LOD}_{\text{normal}} \geq 3.5 \wedge \text{LOD}_{\text{tumour}} \geq 3.5) \quad (2.3)$$

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

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

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

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

The thresholds chosen for both LOD (3.5) and VAF (0.001 and 2.7) calculations were previously fitted by the blue-collar bioinformatics workflow for the ‘‘DREAM synthetic 3’’ dataset using the SpeedSeq likelihood difference approach [182] and were selected to identify high confidence variants.

### 2.2.3 Strelka2Pass workflow

In contrast to FreeBayes, whilst Strelka2 has a multiple-sample mode for germline analysis and tumour-normal pair somatic variant calling capabilities, it cannot jointly analyse multiple related tumour samples. We enable this feature by adapting a two-pass strategy previously used for RNA-seq data [183]. 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.

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

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

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

And therefore:

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

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

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

## 2.2.4 Validation

While the development of new methods can challenge previous assumptions, all methods need be validated against the current gold standard methods in the field, with data which allows objective evaluation. For germline variant calling, there have been multiple community led challenges and specifically designed test datasets, but there is currently no somatic variant calling equivalent. This issue is even more pronounced for our method, as

Appendices/Variantcalling/Figure\_1.pdf

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

we do not only need a tumour-normal pair, but we need the multiple tumour samples in the dataset to be related. To allow a fair comparison, we first generated a fully synthetic dataset, where every variant is known and fully defined (Section 2.2.4.1) to allow a general performance assessment of the methods. Then to ensure that these metrics also hold true in real world data, we then analysed a previously published dataset which had orthogonal validation of a subset of SNVs through targeted amplicon sequencing (TAS) in Section 2.2.4.2.

### 2.2.4.1 Simulated data

We first simulated a phylogeny with somatic and germline variants from ten tumour samples and one normal (Figure 2.1A and Figure 2.2A, B). Germline variants were simulated at a uniform allele frequency of 0.5. Somatic VAFs were sampled from a custom distribution, modelled to favour low allele frequency variants to closely represent real world data (min VAF: 0.001; max VAF: 1; Fig. S1C, D). Paired-end sequencing reads with realistic error profiles were simulated for WGS data at 160X average coverage using

Appendices/Variantcalling/supp/S1.pdf

FIGURE 2.2: Characteristics of simulated data: A) Simulated phylogeny of samples B) Number of simulated germline and somatic variants per sample C) Variant allele frequency distribution of simulated variants per sample D) Distance to nearest variant in each sample.

the ART-MountRainier software [184]. The simulated reads were aligned to GRCh38 and both germline and somatic variants from the phylogeny were spiked into the aligned reads using Bamsurgeon [185]. We compared the workflows for FreeBayes and Strelka2 with and without our extensions for joint variant calling on the simulated datasets. The performance of Mutect2 joint variant calling was also assessed using its proposed best practice workflow. As both Mutect2 and FreeBayes do not return a verdict for each individual sample, we needed to assign each sample in the multi-sample VCF its own FILTER value. We called a somatic variant as present in a sample, if there were at least two reads supporting it for this sample and the overall FILTER showed a “PASS“, which was the same cut-off used in the refiltering step in the Strelka2-pass workflow.

While the precision of each method without our extensions was greater than 99.8%, they all missed at least 25% of all variants in the samples (i.e. recall  $\leq 75\%$ ). In contrast, the recall of the modified workflows increased to  $\approx 95\%$  with only a minute decrease in the

Appendices/Variantcalling/supp/S2.pdf

FIGURE 2.3: Performance of workflows using simulated data: A) Precision and B) Recall of Mutect2, FreeBayes and Strelka2, run in single tumour-normal paired and joint calling configurations.

precision for both FreeBayes and Strelka2 (Figure 2.3). Mutect2 had virtually no change in precision, but the recall actually decreased from  $\approx 75\%$  to  $\approx 41\%$  when analysing the samples jointly (Figure 2.3B). Additionally, with our modified workflows, true positive variants were called with VAFs as low as 0.008 (median detected VAF  $\geq 0.14$  for joint sample analysis and  $\geq 0.21$  for single tumour-normal pair analysis), enabling improved distinction between true variants and technical errors (Figure 2.4). This improvement in performance for Strelka2 is only achieved after the refiltering step and not just a result of the second pass (Figure 2.5, Section A.5.4).

The performance of joint variant calling in Mutect2 was inferior compared to all other methods (Figure 2.3A, B). This was primarily due to the "clustered\_events" filter in Mutect2, which excluded the majority of false negative variants, with negligible contribution to the exclusion of true negative variants (Figure 2.6A, B). This result was unexpected as the simulated variants were evenly distributed along the genome and the corresponding allele frequencies were sampled randomly (Figure 2.2D).

Since the extent of the improvement in our joint calling workflows is bound by the number of shared variants between samples, we sub-sampled the simulated dataset, to show the effect of incomplete sampling on our methods, which is more likely in clinical settings. Furthermore, the evolutionary distance between the related samples in addition to the number of samples, has a major impact on the number of shared variants, as only variants acquired between the germline and the most recent common ancestor (MRCA), will benefit from the joint analysis. Therefore, we selected three sample subsets which included two, three and five samples with high evolutionary distance to show the minimum expected improvement (Figure 2.1A, B). There was a clear linear improvement for both FreeBayesSomatic and Strelka2Pass when increasing the number of samples,



FIGURE 2.4: Variant allele frequencies (VAF) of variants detected by joint sample analysis; A) VAF distribution of true positive variants additionally detected by Strelka2pass B) and FreeBayesSomatic C) VAF distribution of false positive variants additionally detected by FreeBayesSomatic D) and Strelka2pass E) VAF distribution of false negatives not called by FreeBayesSomatic F) and Strelka2pass.

Appendices/Variantcalling/supp/S4.pdf

FIGURE 2.5: Performance of individual steps in the Strelka2pass workflow using the simulated data: A) Precision and B) Recall of tumour-normal paired analysis, two-pass step without refiltering (supplying variants from all tumour-normal pairs for evaluation) and two-pass step with refiltering (the final workflow)

even if they had a distant evolutionary relationship. In contrast, when using only two samples with a small evolutionary distance, the increase in performance was almost as large as when jointly analysing all 10 available samples. This shows that samples with a high number of shared variants will perform better in joint calling workflows (Figure 2.7).



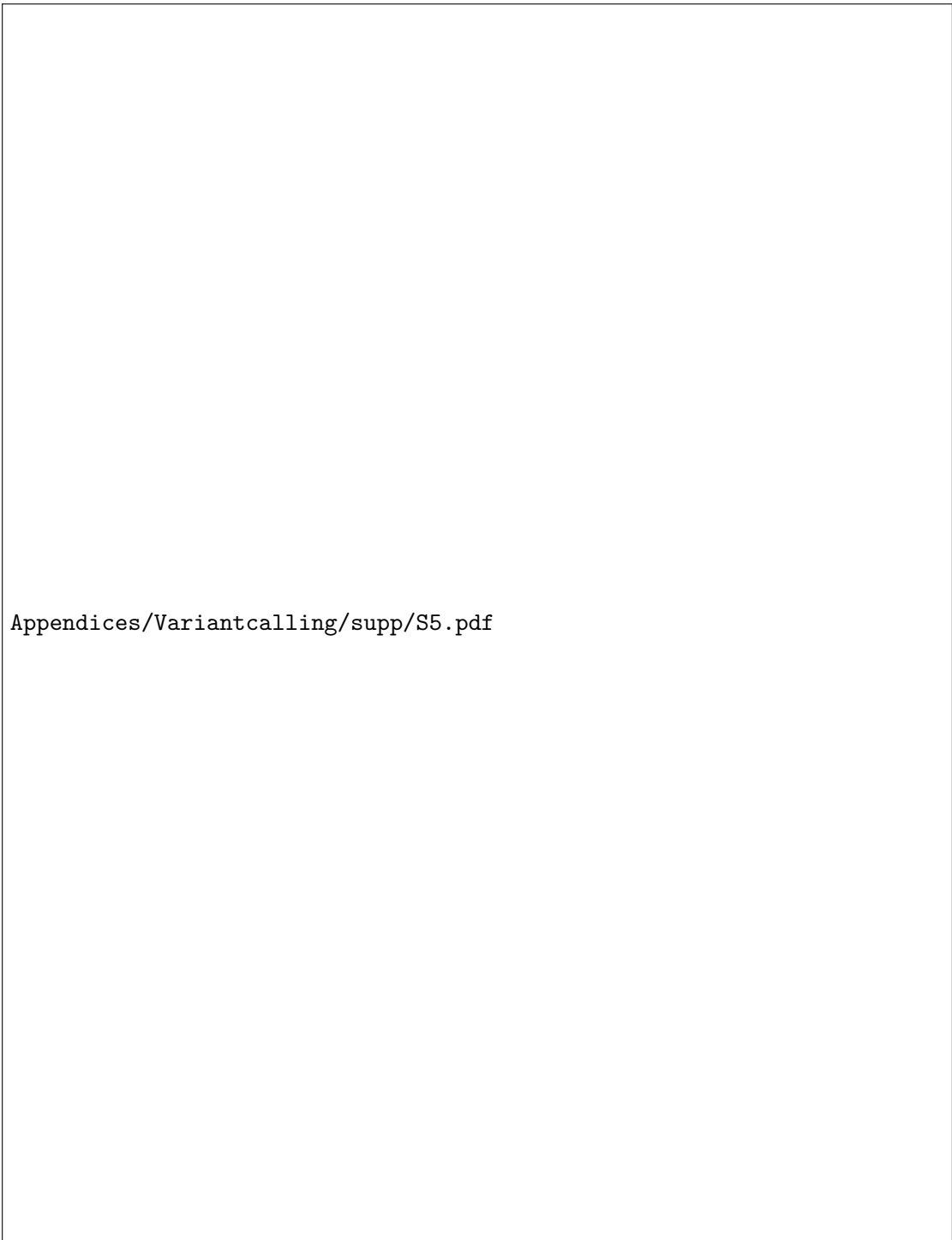
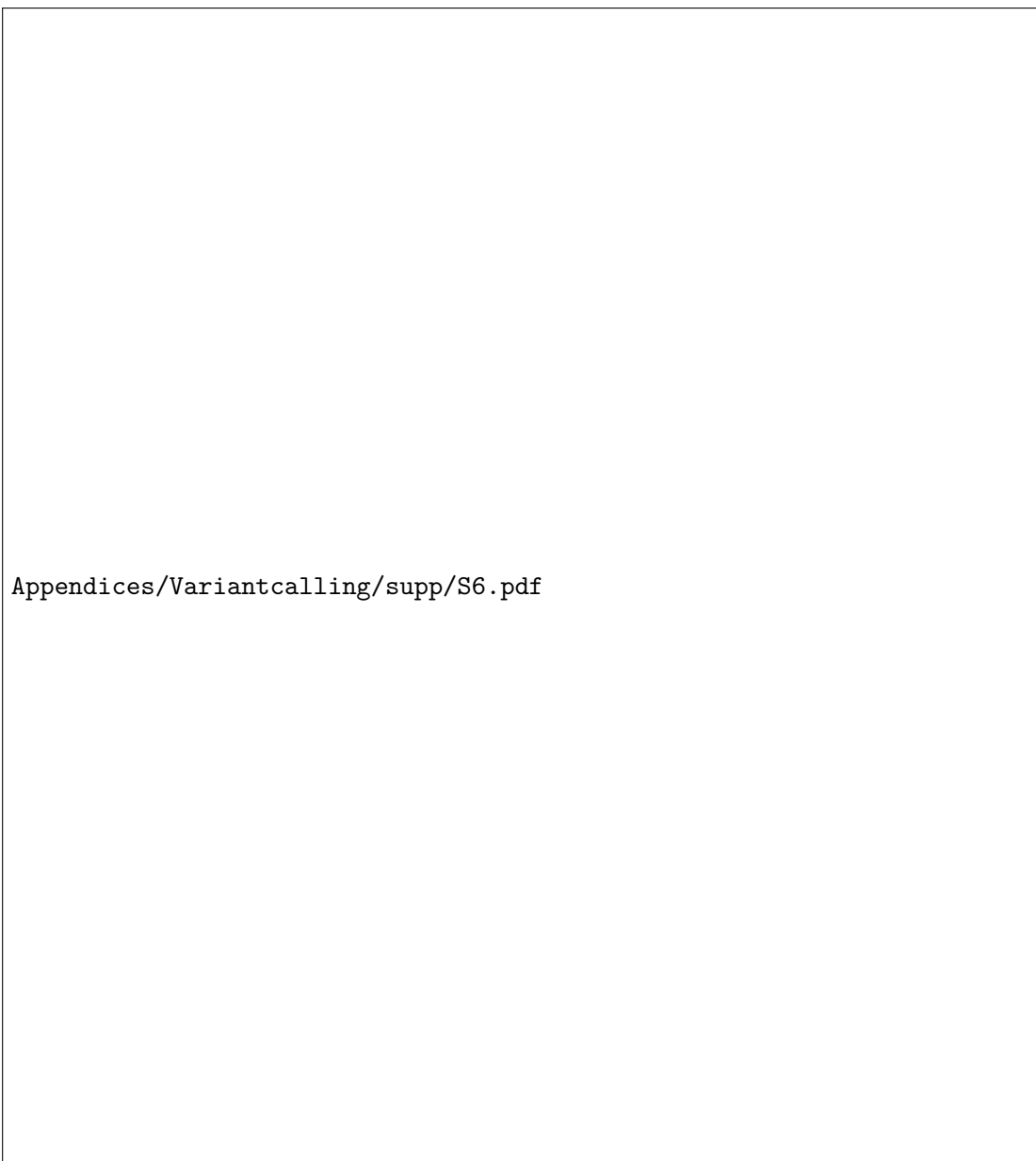


FIGURE 2.6: Summary of variant filters assigned by Mutect2; The counts for each filter type are denoted by black boxplots with white circles depicting the median values. The fitted distribution of variant counts outlines each boxplot; A) Counts of filter assignments for false negative variants and B) true negative variants called by Mutect2 C) Filter assignment for all variants reported for sequenced patient data sequenced with WGS or D) WES.



Appendices/Variantcalling/supp/S6.pdf

FIGURE 2.7: Assessing the performance of different workflows using tumour samples with different evolutionary relationships in the simulated data; A) Simulated phylogeny highlighting two samples with high evolutionary distance (sim-a and sim-j) where MRCA denotes the most recent common ancestor. B) Simulated phylogeny highlighting two samples with low evolutionary distance (sim-a and sim-b).C) Precision and E) Recall of FreeBayes and Strelka, run in individual tumour-normal paired and joint calling configurations using two (sim-a and sim-j), three (sim-a, sim-g and sim-j), five (sim-a, sim-c, sim-f, sim-h and sim-j) and all ten tumour samples D) Precision and F) Recall estimates for FreeBayes and Strelka run in individual tumour-normal paired and joint calling configurations. Comparing the performance of these workflows when using two evolutionary distant samples (sim-a and sim-j), two evolutionary close samples (sim-a and sim-b) and all ten tumour samples.

### 2.2.4.2 Clinical data

To validate the performance of our new workflows, we then analysed WGS and whole-exome sequencing (WES) data of multi-region tumour samples from eight patients, with late stage melanoma, who had multiple tumour samples (average 7 samples per patient; total number of samples 55) collected following enrollment in a rapid autopsy program (CASCADE) conducted at the Peter MacCallum Cancer Centre (Table A.1 and Section A.5.2) [6, 186]. The published studies had multiple somatic variants from the clinical samples orthogonally validated through targeted amplicon sequencing (TAS). We used these TAS-validated variants as the gold standard to evaluate the performance of different workflows, acknowledging that the technical biases inherent to TAS data are different to those present in WGS and WES (Figure 2.8) and that there would be sampling biases depending on different tumour cells analysed in each data type.



FIGURE 2.8: Correlation of variant allele frequencies (VAF) from WES and WGS data against targeted amplicon sequencing VAF values with fitted violin plots of each individual distribution. Grey background shows 95% confidence interval for the fit of the linear model (dotted line).

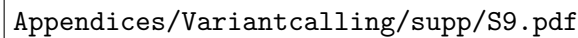
In concordance with the results of the simulated data, our improved workflows found additional variants in all but one patient (Figure 2.1D, Figure 2.9) (total additional variants Strelka2Pass: 64; FreeBayesSomatic: 85) with only a slight drop in precision for FreeBayesSomatic (mean: 0.94 vs. 0.88) and Strelka2Pass (mean: 0.97 vs. 0.92). Since the panel of variants validated by TAS was limited (7108 bp for patients CA-B through -H), this increase in detected variants suggests that a high number of shared variants in

Appendices/Variantcalling/supp/S8.pdf

FIGURE 2.9: Performance of the different workflows using clinical samples from eight cancer patients: A) Number of variants called by Strelka2 run in the tumour-normal paired (grey) and joint calling configurations, which have been validated by targeted amplicon sequencing (TAS). The same for C) FreeBayes and E) Mutect2 workflows. Precision of tumour-normal paired and joint analysis of TAS validated clinical data for B) Strelka2, D) FreeBayes and F) Mutect2; Table A.1 provides the sample naming map to the original publications.

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



Appendices/Variantcalling/supp/S9.pdf

FIGURE 2.10: Correlation between cellularity and proportion of variants found only with joint calling using FreeBayesSomatic. Grey background shows 95% confidence interval for fit of linear model (dotted line)

purity, joint analysis can have a major impact on improving performance (Figure 2.1E, Figure 2.10).



Appendices/Variantcalling/supp/S10.pdf

FIGURE 2.11: Improvement in recall using FreeBayesSomatic and Strelka2pass over Mutect2 in the clinical samples.

Mutect2 in contrast, did not show significant improvement in any sample in its joint calling configuration, but showed inferior performance compared to the tumour-normal pairwise approach in two samples (Figure 2.9E), similar to its decreased performance in the simulated data (Figure 2.3). This was due to true variants being removed by the internal filters of the tool (Figure 2.6C, D). This is in stark contrast to our novel workflows, where the joint analysis preserves all called sites from the pairwise method and finds additional variants. Overall, Mutect2 found less validated variants in all patients than both Strelka2Pass (mean: 2.2) and FreeBayesSomatic (mean: 2.5) with comparable levels of precision (Figures 2.9 and 2.11) but longer run times (Table A.2).

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

## 2.3 Effects of additional somatic variants on downstream analysis

The ability to find additional shared variants has significant impact on our understanding of cancer evolution and the timing of initiation and metastatic seeding. Recent work has shown, that similar to the well known genetic heterogeneity, there is heterogeneity when it comes to the timing of metastatic seeding. While traditionally it was thought that tumours only metastasise after they reach a certain size, to escape the restrictions of the niche, like reduced nutrition, recent publications have shown, that often there is also very early metastatic seeding [187]. For all methods analysing this heterogeneity, evolutionary timing and history are fully reliant on the somatic variants found in the data. Therefore, if we improve the input provided to these analysis methods, we can expect a clearer and possibly more granular result.

The following section will quantify the effect of additionally found variants on phylogenetic reconstruction and clonal decomposition, which use somatic variants as input.

### 2.3.1 Phylogenetic reconstruction

As this work is not about the advantages and shortcomings of different phylogenetic reconstruction tools, we have not performed a comprehensive comparison of tools, but rather focused on the results of using additional variants discovered using our novel joint somatic variant calling workflow. For this reason, we chose to use neighbour joining (NJ) [188], because it is fast, readily available in most phylogenetic reconstruction tool kits and if the input distance is correct, the output will be correct. Furthermore, even if the distance is not 100% correct, if the distance is “nearly additive“ and the input distances are not far off from the real distance, the tree topology will still be reconstructed correctly [189]. Lastly, in contrast to many other methods like UPGMA and WPGMA [190], NJ

does not assume an equal mutation rate of each sample, because we know, that the molecular clock hypothesis [191] is not valid for different lineages of cancers [192].

The only thing that NJ requires as an input is a distance matrix of all samples, so the next step was the selection of the right distance metric. While there are many distance measures for DNA sequences, which allow accounting for different probabilities of transitions and transversions as well as uneven base composition, models like F81 [193] or HKY85 [194] are only really designed for germline mutations and are not easily applicable for subclonal somatic mutations. For this reason, we decided to first transform the variants present in all samples into a binary occurrence vector and then calculate the Hamming distance [195] 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 2.12 shows both the reconstructed phylogenies of the autopsy samples of the previously described late stage melanoma patient “CA-F” from the manuscript (Appendix A, Table A.1), using the variants found with the default tumour-normal method on the left and our improved joint method on the right. The exact same reconstruction methodology was otherwise used.

Figures/jointVariantCalling/phyloCA9.pdf

FIGURE 2.12: Reconstructed phylogenies of a patient with multiple spatially distinct samples; Neighbour joining on Hamming distance on variant occurrence vector. Tip labels describe the location of the sample in the patient. Trees are shown as unrooted with germline as fixated origin point; black line ruler shows the length of an edge with 2000 mutations; LN = lymph node

There are several obvious differences, first in the longer edge connecting the germline to all other samples, which we consider as the state of no somatic variants. This shows

that there are many more shared mutations in all samples, than what would have been anticipated with the default method, which corresponds to an overestimation of the heterogeneity of the samples. As the accumulation of somatic variants is still used as a proxy for timing and cell divisions, when assuming a high mutation rate for lung cancer ( $5.3 \cdot 10^{-8}$  from Werner et al. [196]) this difference of  $\approx 36000$  variants is equivalent to  $\approx 2000$  cell divisions. While the cell doubling rate of cancers is highly dependent on the type [197], this change makes a substantial difference when assessing the timing of the tumour initiation and further evolution.

Secondly, there have been topological changes, which generate a longer bifurcating edge between the olive coloured “right liver lobe” and the “right parietal lobe” lesions showing a bottle neck in cancer evolution, which corresponded with the clinical history, where the patient lived with stable disease for almost ten years, before rapid disease progression just prior to death. The extreme distance of the primary/diagnostic sample from the rest of the samples could be in part due to a difference in sequencing quality, as this was an FFPE tumour biopsy. However, as this feature is preserved between both the joint and the pairwise analysis, it does not appear to be an effect of our new method.

Figures/jointVariantCalling/tanglePhyloCA9.pdf

FIGURE 2.13: Side by side view of the reconstructed trees from Figure 2.12; internal edges, which are distinct between trees are shown as dotted lines; common subtree is shown in red Tree labels have been sorted to minimise distance between labels; LN = lymph node ; Visualisation generated with dendextend [27]

Figure 2.13 shows a topology focused view of the two phylogenetic trees, which highlights the differences between the two reconstructions [198]. The common subtrees are coloured



the same on both sides and connecting lines show identical labels. This format shows that while the trees look quite similar at first glance, they show vastly different topologies.

One example of this is the “small bowel“ tumour sample which was originally connected to the red common subtree, but is now much closer to the “right cerebral lobe“ lesion, forming a parallel clade with the “right liver lobe“ lesion. In general, whereas the pairwise tree shows a very linear topology, with minimal branching and no disjunct subclades, these features are clearly present in the joint reconstruction. (Figure 2.13).

## 2.4 Longitudinal analysis

In addition to the analysis of spatial heterogeneity through multi-regional tumour sampling, we were also interested in the use of our novel joint variant calling workflow for the analysis of temporal heterogeneity through longitudinal samples collected from the same patient over time. Examples of longitudinal analysis could include the joint analysis of diagnostic and relapse tumour tissue samples, or even the repeated serial testing of ctDNA. In the following section, we present work using the published workflows on a longitudinal dataset, which highlights the flexibility and widespread usability of the new methods. Similar to the spatially related samples, the joint analysis can improve the performance, which then in turn enables improved detection of lower allele frequency variants, particularly in the setting of low tumour burden as is commonly seen with ctDNA analysis.

In addition to their autopsy which resected nine distinct tumour sites (Figure B.1), Patient “CA-F“ also had three longitudinal blood samples taken, from which ctDNA was extracted and WES performed. These blood samples were taken as non-invasive surveillance seven, five and two months before the death of the patient (Figure 2.14).

To show that even in longitudinal data, the joint analysis can boost the signal, we jointly variant called variants in the diagnostic biopsy sample with the three ctDNA samples and compared them with the results from the pairwise analysis. On average, we found 2905 additional variants in each of the ctDNA samples, which is more than doubles the average number of variants found with the pairwise analysis (2414). Out of those, we found 534 ( $\approx 20\%$ ) variants in the ctDNA samples, which were found as

Figures/jointVariantCalling/CA-F\_timeline.pdf

FIGURE 2.14: Timeline from diagnosis till death for patient CA-F: 1.9mm melanoma removed after diagnosis 25/07/2003 but with negative sentinel lymph node biopsy; 28/06/2012: PET scan and subsequent liver biopsy confirm relapse with wide spread metastases; trametinib treatment from Oct. 2012 till Jan. 2013 with minor response; blood plasma samples during treatment (1 and 2) as well as after progression (3); death and rapid autopsy of nine metastatic sites (13/04/2013, Figure B.1); Tumour fraction in plasma samples was estimated via digital droplet PCR quantification of the original driver mutation (BRAF:K601E)

a high confidence variant in the diagnostic sample, indicating that these findings were high quality calls.

As in the spatially analysis, in longitudinal data lower tumour purity samples benefit more from the joint analysis. We see that time point 2 (T2) had the highest number of recovered variants (377) which were found as high confidence variants in both other time points (Figure 2.15 A vs. B vs. C) and T2 also has the lowest ctDNA fraction recorded (T1: 60%; T2: 20%; T3: 60%). A total of 106 variants were not found in the ctDNA samples with the pairwise analysis, even though they were high confidence variants in the primary sample (Figure 2.15F). These variants usually showed a lower depth of coverage (dp) in the ctDNA samples, which is likely the explanation as to why they were below the limit of detection of the pairwise analysis.

Finally, we can also find 398 additional variants in the primary sample. 396 of these variants were discarded in the pairwise analysis due to low evidence in the tissue sample, but could be found with a high confidence in the longitudinal data. The last two variants were could be found in the joint analysis, as all 4 samples showed evidence for this variants just below the limit of detection (Figure 2.16).

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

Figures/jointVariantCalling/longitudinalCA9ctDNAVafs.pdf

FIGURE 2.15: Improved somatic variant calling in longitudinal data: Variant allele frequency (VAF) of variants found additionally through joint variant calling which were found as high confidence variants in the primary sample; Variants with less than 0.1 VAF in the primary are coloured grey; “T1 recovered“ shows variants, which were high confidence in all ctDNA samples but T1 and were only found through joint calling there; Axis label show the date of blood collection

### 2.4.1 Clonal deconvolution

One of the most important pieces of information that can be derived from multiple related samples from the same patient is the clonal deconvolution, where subclonal reoccurring patterns of mutations (clones) are resolved both spatially and longitudinally. These reoccurring clones can be linked to either parallel evolution through positive selection pressure, like a targeted drug, or due to the process of developing metastases where a part of the cancer disseminates and grows at a different site. In contrast to the lack of options for joint somatic variant calling, there is a plethora of algorithms and methods available for clonal deconvolution. Since 2015 PhyloWGS [199], Canopy [200], CLOE [201], CloneFinder [202], MACHINA [203] and MOBSTER [204] were published,

Figures/jointVariantCalling/longitudinalCA9primaryVafs.pdf

FIGURE 2.16: Longitudinal data informs diagnostic variant calling: VAFs of variants additionally found through joint calling in the primary samples; A) “Primary recovered by PASS” shows variants which were high confidence in at least one ctDNA sample but not found in the primary; B) “Primary recovered joint” shows variant which were low confidence in all samples in the pairwise analysis individually but were called as high confidence in the joint analysis; Axis label show the date of blood collection

to name a few. Underlying all of these models is the ability to cluster variants with similar variant allele frequencies together, to reduce the combinatorial space and enhance the confidence in the signal [205]. Due to the high number of tools, it is very challenging to select the right tool, especially since all of them have advantages and disadvantages [206]. In this work we decided to use PhylogicNDT [207] as it has been shown to work well on clinical samples [208] and does not have the restriction for the input to be from copy number neutral areas which many of the other tools have.

Both the variants found with the default pairwise as well as with the new joint workflows were annotated with their local allele specific copy number to form a MAF like file format which is required by PhylogicNDT. While PhylogicNDT allows the user to supply the cancer cell fraction for every variant, the program can also estimate them from the supplied allelic counts and the copy number. Local copynumber calls were derived from copy number segment calls made by Sequenza by intersecting the chromosomal location of each variant with the copy number segment containing the variants location. This requires multiple steps and the source code is shown in Listing B.1 (parsing VCF), Listing B.2 and Listing B.3 (convert to MAF format). Variants which couldnt be annotated with copy number information, because their genomic location did not overlap with any called copy number segment, were discarded for this analysis.

Figure 2.17 shows the highest parsimony clonal tree reconstructed by PhylogicNDT for the pairwise as well as the joint variant calling for patient CA-F. As the copynumber calling information is the same for both inputs, the only difference is in the called variants. While there was no subclonal structure detected at all for the pairwise analysis,

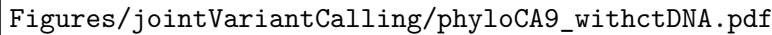
Figures/jointVariantCalling/clonalDeconv.pdf

FIGURE 2.17: Reconstructed clonal trees from PhylogicNDT; Blue circle at top depicts the germline/normal state. The coloured edges with the same coloured circle represents a distinct subclone of the parent from which the edge emerges; The number in braces next to the edge is the number of mutations which define this subclone with an added gene symbol added, if there was a cancer driver gene mutation. The left part shows the result when using the default pairwise method of Strelka2 and the right side uses the results from the Strelka2Pass workflow

there is a highly complex branching structure detected using the jointly called variants with multiple subclones originating from the ancestral clone. As this is a clinical sample, we cannot be certain that the more branched model is the actual truth, but it is biologically more logical that a late stage cancer has developed several subclones, rather than it being a very homogeneous disease at all of the 10 sites at autopsy with no evolution over ten years of disease [208]. It is of particular interest, that the *CDC27* gene was mutated at different time points in different clones (clone 8 vs. clone 4), which is a clear indicator of convergent evolution, which would definitely be missed without the joint analysis.

### 2.4.2 Longitudinal enriched phylogeny

Of course it is finally also possible to build a phylogeny which incorporates data from both the spatial tumour tissue and longitudinal ctDNA analysis. However, as the ctDNA can provide a holistic view of all cancer metastases (Section 1.2) the interpretation needs to accommodate for that.



Figures/jointVariantCalling/phyloCA9\_withctDNA.pdf

FIGURE 2.18: Reconstructed phylogeny with longitudinal ctDNA samples: Tree from Figure 2.12 with three additional ctDNA samples from different time points approximately one year prior to death. The ruler shows the equivalent of 2000 mutations; LN = lymph node

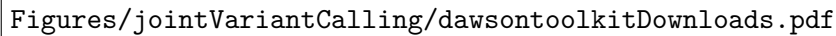
The addition of the ctDNA samples led to a further bipartition edge, which separates the “right liver lobe“, “small bowel“ and “right cerebral lobe“ lesions from the rest of the tree (Figure 2.18). This was already inferable from the topology of the previous tree in Figure 2.13 “joint“, but is even more pronounced with the inclusion of the ctDNA samples.

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

## 2.5 Usage statistics and uptake

Ultimately when choosing research software, publication and citations are not a good metric to evaluate the quality of a method [209]. Many published software packages are not maintained or not even functional even though they are published. While we developed these joint somatic variant calling workflows to deal with a challenge we faced, the interest of others has been continuously expressed by both members of the bioinformatics community in the short period of time since publication.

To have some proxy of the usage statistics of the workflows, we recorded the download numbers of the “dawsontoolkit“ docker container after the publication of the manuscript.



Figures/jointVariantCalling/dawsontoolkitDownloads.pdf

FIGURE 2.19: Cumulative download numbers of the “dawsontoolkit“ docker container since publication of the manuscript; Actual counts are shown as dots, with smoothed trajectory depicted as dotted line with the 95% confidence interval shown as a grey background; confidence interval has been adjusted with exponential decay of prediction accuracy with distance from the last data point; Start date 7<sup>th</sup> September 2021 (publication of method); End point prediction 31<sup>th</sup> December 2022 (End of current year); Numbers were recorded daily from the DockerHub API

The container only consists of software for refiltering and joint analysis of the workflows. Obviously, this is an imperfect measurement, as people can reuse a downloaded container as often as they want, which would not appear in the count and similarly, just because the container was downloaded, the analysis might not have been used. Nevertheless, it still shows an interaction and an interest in the methods. The download numbers show a sustained and stable increase in downloads (Figure 2.19). This suggests, that there is a need in the methods, rather than a simple curiosity after publication, which hopefully will facilitate a higher quality analysis of future projects and therefore lead to a better understanding of cancer evolution and heterogeneity.

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

— Lucius Annaeus Seneca, *De Consolatione ad Marciam*

## CASCADE - Late stage lung cancer in the spotlight

### 3.1 Introduction

As tumour heterogeneity is seen as one of the major causes of resistance to treatment and ultimately relapse, much cell line based research has been conducted to solve tissue of origin and evolutionary trajectories via bulk and single cell sequencing paired with cellular barcoding [210, 211]. However, while cell line models are a great resource for high throughput methods and allow easier reproducibility of results, they are no real substitute for primary patient cells. With the increased availability of patient samples through bio-banking efforts like the UK BioBank [212] and the Victorian Cancer BioBank [213], both patient derived xenografts (PDX) and organoids have gained more and more traction [214] as specialised models to grow primary patient cells in an environment which closely resembles the body of the patient. While this method is superior in many aspects, there are some significant drawbacks. The culturing of the cells requires more effort and is not as easily scalable. These methods also require fresh patient samples, which are not always readily available.

While it is fairly easy to collect diagnostic specimens from tumour biopsies for storage and research, late stage tumour biopsies are rare. Due to the deteriorating health status of the patient biopsies can be dangerous and often an unnecessary burden for the patient. However, these samples are especially critical when answering the question of how the cancer was able to evade treatment and lead to death, as it may reveal an unappreciated insight into spatial and temporal heterogeneity.

To try to combat this issue the cancer tissue collection after death (CASCADE) program was initiated. It recruits cancer patients close to the end of life and enrolls them in a rapid autopsy program. These autopsies are carried out at any time of the day to minimise the impact on the sample to allow high quality assessment including DNA and RNA sequencing of the frozen cells [215]. While the program collects cancer patients unconditionally of their type of disease, the analysis of this thesis is restricted to five lung



cancer patients available at the time of this work. Currently there are no extra lung cancer patients enrolled, but recruitment is still ongoing. Four of these five patients had an Epidermal growth factor receptor (*EGFR*) based cancer and one had a RET Proto-Oncogene (*RET*) fusion with *KIF5B*. Each of those patients had on average 30 specimens resected and put into a bio bank. We then continued to sequence, on average, eight of these samples with either whole genome sequencing (WGS) or whole exome sequencing (WES) to deeply analyse and classify the underlying resistance and driver mechanisms of each patient and their heterogeneity.

### 3.1.1 Lung cancer

With around 1.6 million deaths world-wide each year, lung cancer is the number one cause of cancer death [216]. 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 [217]. 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 [218]. 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.

Even with those strict classifications in place, it is widely accepted, that cancer is a heterogeneous disease, which needs to be accounted for when developing treatments [219]. The ongoing research of lung cancer has led to a shift from cytotoxic chemotherapy to a more personalized approach by accounting for the genetic background of each patient's disease [220]. But not only the inter-patient heterogeneity needs to be taken into account, but also the heterogeneity between different sites of the disease in the same patient [3, 221]. This makes the choice of treatment for one single patient more and more difficult, as some sites might respond to treatment, where others might not. This means, in order to design the perfect treatment regime for a patient, a deep understanding of the overall complexity of the disease is needed. By studying a diverse

background of driver mechanisms of lung cancers and their respective treatment and resistance modes, a general insight in the biologic background is possible. Analysing not only one, but several metastases of the same patients paints a much clearer picture of disease progression and the process behind the resistance to treatment that ultimately led to death.

## 3.2 Publications

This chapter includes and reproduces data analysis that contributed to two publications, however as they were not sole first author publications, they are only mentioned here instead of included in full. The first publication features the resistance mechanism of small cell transformation seen in patient CA-L (Section 3.3.6) (*‘An Evolutionarily Conserved Function of Polycomb Silences the MHC Class I Antigen Presentation Pathway and Enables Immune Evasion in Cancer’* Burr et al. [5]) and the second shows the discovery of emerging novel resistance mutations in a RET-fusion driven NSCLC in patient CA-A (Section 3.3.2) (*‘RET Solvent Front Mutations Mediate Acquired Resistance to Selective RET Inhibition in RET-Driven Malignancies’* Solomon et al. [6]).

## 3.3 Patient level analysis

This section outlines the analyses performed for each patient and highlights work specifically done for certain patients due to their unique clinical features. However, most of the analysis was streamlined with the same workflow applied to each patient. The following sections expand on the individual steps.

1. **Quality control:** Each sample of a patient is checked for kinship and sequencing quality
2. **Read mapping**
3. **Joint somatic variant calling:** SNPs, InDels and SVs are called jointly
4. **Copy number calling**
5. **Variant effect annotation:** short and structural variants are annotated with possible biological effects

## 6. Phylogenetic reconstruction

## 7. Clonal deconvolution

### 3.3.1 Analysis workflow

This section summarises the primary analysis performed for each patient in detail. Specific analysis are discussed in the individual patient sections.

#### 3.3.1.1 Quality control

When multiple samples per patient are available, the possibility of sample mix-ups is higher than when just dealing with a tumour normal pair, so in addition to the standard read depth, sequencing quality and reads-on-target analysis that is routinely performed after sequencing, we performed an additional step of kinship detection. We use concepts commonly employed in germline cohort analysis, like child and parents (trio) or even large databases (gnomAD). As most germline variants are due to mendelian inheritance, we can use the percentage of shared homo- and heterozygous germline variants to estimate the relatedness of two samples. For our analysis we used NGSCheckMate [222] and all the results shown in later sections are based on it, however we also used Somalier [223] on two patient samples with surprising kinship results and Somalier confirmed the result.

While this analysis is very useful to detect samples which do not belong to a patient, either through mislabelling or similar, it does not protect from mix-ups within a patient's samples. However, only orthogonal validation will be able to discern these errors.

Other quality controls were performed with fastQC [224] for read integrity and ‘*CollectWgsMetrics*’ from Picard [225] for WGS samples and ‘*samtools flagstat*’ [57] for on-target estimation for WES samples.

#### 3.3.1.2 Read mapping

For highest mapping performance, reads were aligned alternative contig aware with BWA [115] (v0.7.17) to GRCh38 (*GCA\_000001405.15*) with alternative contigs but no decoy regions. Initial mapping was post-processed with ‘*bwa-postalt.js*’ from bwa-kit to adjust the mapping assignment and quality mapping both to alternative and canonical

contigs. Finally reads were duplicate marked with ‘*MarkDuplicates*’ from the Picard-toolkit.

### 3.3.1.3 Joint somatic variant calling

For short variants (SNPs and InDels), the workflows presented in Chapter 2 were used and while the Strelka2Pass workflow generates structural variant calls, they are not jointly called over all samples. Instead for the structural variants (SVs) we used GRIDSS2 [226], which has a calling model for multiple related tumour samples and as GRIDSS2 is also a prerequisite for copy number calling with PURPLE (Section 3.3.1.4) using the same structural variants allows a higher conformity of analysis.

### 3.3.1.4 Copy number analysis

After somatic variant calling, copy number analysis is critical when dissecting the resistance and driver alterations of a tumour sample. While lung cancers are known for their high mutational burden [227], often genetic amplifications can be found as driver or resistance mechanism. One of the more common resistance mechanisms is a high *EGFR* or *MET* amplification which significantly affect transcription [228]. And while copy number alterations are often shared between metastases [229], the same heterogeneity that can be found in variant calling analysis also affects copy number analysis. Many modern copy number calling methods will use the B-allele frequency, the allele frequency of a heterozygous germline variant, to gain allele specific copy number calls [15, 18, 230]. Although each of those methods will only use the input of one tumour and one germline sample. As described in Chapter 2 we can actually improve the performance by analysing all tumour samples jointly. So far only HATCHet [231] has a joint copy number calling method, but requires significant time investment for installation and subjective manual parameter optimisation on a per patient basis. In contrast both sequenza and PURPLE have very easy installation and usage procedures. To ensure low subjectivity and high reproducibility of our result, we chose to not use HATCHet, and instead use the clinically used and approved PURPLE workflow for all WGS samples and sequenza for WES seeing its successful use in similar situations [3, 186] and because PURPLE is not suitable for WES data. In spite of the potentially higher accuracy of HATCHet, virtually no downstream analysis was equipped to utilise multi clone resolution copy number calls.

### 3.3.1.5 Variant effect annotation

For small variants (SNPs and InDels) “Variant Effect Predictor“ (VEP) version 92 [21] was used to assign possible effects. As a variant can affect multiple genes due to overlapping gene boundaries, effects within a curated list of lung cancer related genes (Table C.1) were assigned an impact in line with the VEP provided impact values of ‘*LOW*’, ‘*Moderate*’ and ‘*HIGH*’. To only have one effect per variant, only the variant with the highest impact was returned. In cases of multiple transcripts being affected with the same impact level, the putative canonical transcript result is used.

For structural variants, the effect annotation depends on the type of the structural variants. For amplifications and deletions, the genes within the variant are compiled and returned as a list. The effect of inversions and similar structural changes are assumed to be fusion based, so the breakpoint is annotated with the gene hit by both breakpoints and a potential fusion gene is returned.

### 3.3.1.6 Phylogenetic reconstruction

Variants called in any sample were transformed into a binary presence/absence vector with a pure absence vector as the germline native state. The vectors were then concatenated into a string representation, and for each pair the Hamming distance were computed [195]. The distance matrix was used as input for the neighbour joining algorithm and visualised with ape [25].

### 3.3.1.7 Clonal deconvolution

Clonal deconvolution for each patient was done with PhylogicNDT Cancer cell fractions (CCF) were left to PhylogicNDT with the option ‘*-maf\_input\_type calc\_ccf*’ by supplying the allele specific local copy number call for each variant the same way as shown in Section 2.4.1 with the copy number calls from Section 3.3.1.4. If no copy number was reported for a variant, it was removed from the analysis.

For the clustering of variants all variants with no known protein changing function were included, by removing all variants with the VEP consequence “intergenic variant“, “intron variant“, “upstream gene variant“, or “downstream gene variant“. While these

add explanation of the neutral selection mutation requirement from page 99

variants certainly might distinguish clones within the sample, they could only arise through random genetic drift and did not relate to resistance mechanisms.

As PhylogicNDT can only visualise 58 distinct clusters, due to the lacking number of distinct colours, we restricted the analysis after clustering all mutations. Clusters with a variant in one of the 319 driver genes suggested by PhylogicNDT were always retained. All other clusters were automatically removed, if the number of variants  $n$  supporting the cluster was smaller than 10, or the CCF value in each sample was too homogeneous, or the confidence interval  $CI$  of the CCF value was too high.

The homogeneity of the CCF value of a cluster  $c$  was assessed by calculating the z-score of each sample  $s$  CCF in respect to all other samples CCF. If one samples z-score indicted a fold change of less than 1.5, the cluster was removed (Equation 3.1).

$$Inclusion(c) = \begin{cases} \text{FALSE,} & \text{for} & n_c(vars) < 10 \\ \text{FALSE,} & \text{for} & \forall s : |z\text{-score}(CCF_s)| < 1.5 \\ \text{FALSE,} & \text{for} & \forall s : CI(CCF_s) < 0.1 \\ \text{TRUE,} & \text{else} \end{cases} \quad (3.1)$$

### 3.3.2 Patient CA-A

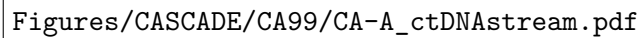
This patient was a 61 year old male with a metastatic *RET-KIF5B* fusion positive NSCLC. After failure of Carboplatin, Pemetrexed, and Pembrolizumab followed by the multi kinase inhibitor Lenvatinib, he then received compassionate access to the RET tyrosine kinase inhibitor Selpercatinib (Figure 3.1). He experienced almost immediate improvement following Selpercatinib with decreased levels of carcinoembryonic antigen and almost 100% reduction of *RET* fusion positive ctDNA after one month (Figure 3.2). Similar to the ctDNA analysis, Positron emission tomography (PET) and computed tomography (CT) imaging revealed significantly reduced tracer uptake in multiple sites and partial response to treatment (Figure 3.3).

Figures/CASCADE/CA99/CA-A\_timeline.pdf

FIGURE 3.1: Timeline of patient CA-A from diagnosis until death: Diagnostic biopsy detected *KIF5B-RET* positive lung adenocarcinoma; SRS: stereotactic radiosurgery; WRBT: whole brain radiation therapy; a total of six blood samples were taken just before and during the selpercatinib treatment.

Serial sampling of the plasma of the patient and analysis with the commercial Guardant360 assay [232] revealed a previously undetected RET G810S resistance mutation after three months of treatment. While at this point the driver mutation allele frequency was still dropping in the plasma, by month four the abundance of RET G810S had increased and was accompanied by additional mutations in the same site (RET G810R, C, and V). In addition there was an increase of fusion positive ctDNA this suggesting the development of acquired resistance to Selpercatinib. While the patient initially was responsive to the treatment, repeat PET scans showed progressive disease after six months, which ultimately led to the death of the patient (Figure 3.3).

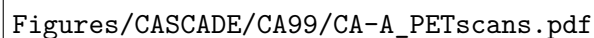
At autopsy, 24 tumour tissue biopsies and a post mortem blood sample were collected and eight of them were selected for WGS at 130x coverage (Figure 3.4, Table 3.1) and analysed with the standard workflow (Section 3.3.1).



Figures/CASCADE/CA99/CA-A\_ctDNASTream.pdf

FIGURE 3.2: Allelic frequencies of driver and emerging resistance mutations during Selpercatinib treatment (11 months after diagnosis); *KIF5B-RET* fusion is the initiating driver with RET G810R/S/C/V the emerging resistance SNPs

Somatic variant calling revealed substantial spatial heterogeneity, where both the occipital lobe and the right pleura sample only contained RET G810S, the right liver lobe harboured predominantly RET G810R with either G810S and G810C as minor clones and lastly, the left liver samples showed almost an even mix between G810C and G810S clones but no G810R presence. The emergence of these mutations in multiple different sites at different allele frequencies, especially in already established sites in the liver, suggests that these mutations are the result of parallel evolution under positive selection



Figures/CASCADE/CA99/CA-A\_PETscans.pdf

FIGURE 3.3: PET scans of patient CA-A before and during Selpercatinib treatment



Figures/CASCADE/CA99/CA-A\_schematic\_CA99\_organColours.pdf

FIGURE 3.4: Schematic of tumour lesions in patient CA-A: Primary diagnostic sample shown in red; All 24 autopsy samples were coloured by organ they were collected from: Brain (7), left lung (2), right lung (1), liver (9), T8 bone (1), ascitic fluid (1), adrenal gland (2), kidney (1); Additionally to the post mortem blood sample, six serial blood samples were taken (Figure 3.1)

TABLE 3.1: Autopsy samples sequenced for patient CA-A: Sample number is the internal sample collection during CASCADE autopsy, the organ of the sample, the fraction of tumour cells from H& E stain and the pathology of the tumour sample.

Sample number	Organ	H&E	Type
11	right occipital lobe	0.7	lung adenocarcinoma
26	right liver lobe	0.6	
31	left lower lung	0.2	
41	left liver lobe	0.2	
47	left liver lobe	0.5	
55	left liver lobe	0.4	
57	right liver lobe	0.6	
59	right pleura	0.7	

through therapy, rather than seeding from one resistant clone. Apart from the mutations changing RET G810 no other variants affecting *RET* or any other lung cancer genes were found in multiple samples. The occipital lobe sample also contained a BRCA1 V939A mutation and one left liver sample (47) showed a synonymous KIT S967%3D mutation. Additionally, no other variant found in non cancer related genes allowed the same explanation of resistance (Figure 3.6).

Phylogeny based on the short variants showed a clear clustering of the right (26 and 57) and left liver samples (41, 47, and 55) with the occipital lobe (11) and pleural sample (59) sharing the most mutations as a hint towards the longest evolutionary trajectory and final progression. There was also a bifurcating line separating several progression (11, 41,47,55; top right) and stable disease sites. The much lower and higher number of both samples 31 and 41 may have been contributed to by the low tumour purity of those samples (Tables 3.1 and 3.2, Figure 3.5).

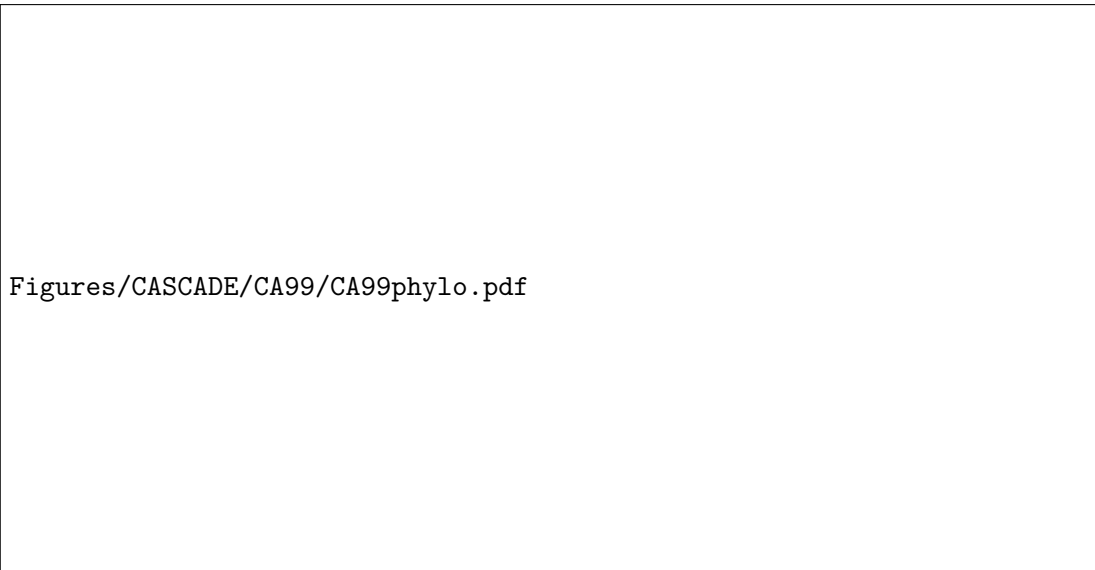


FIGURE 3.5: Phylogeny of autopsy samples from patient CA-A; reconstructed with all somatic SNVs and InDels. Ruler symbolises 4000 variants difference

The structural variant calling with GRIDSS2 showed consistent presence of the *KIF5B-RET* fusion at high allele frequency (min: 0.27 max: 0.535), consistent with a cancer cell fraction of 1 when correcting for local copy number changes (min: 2 max: 3; Figure 3.7), in all but sample 31, which might have been due to the low purity of the sample (Table 3.1). While there was a high number of structural variants present in each sample, consistent with the genomic instability commonly seen in late stages of cancer [208] almost of these rearrangements were sub-clonal and therefore not the main cause of resistance or cancer initiation and rather the result of progressive tumour evolution. To allow a more focused look at structural events and their effect, we restricted the visualisation to events with an allele frequency of 0.2 or higher (Figures 3.7 and 3.8).

While this change also has a minute effect on the PURPLE copy number calls, which are informed by structural variants, these structural variants will also be sub-clonal and therefore removal will result in a cleaner clonal copy number profile.

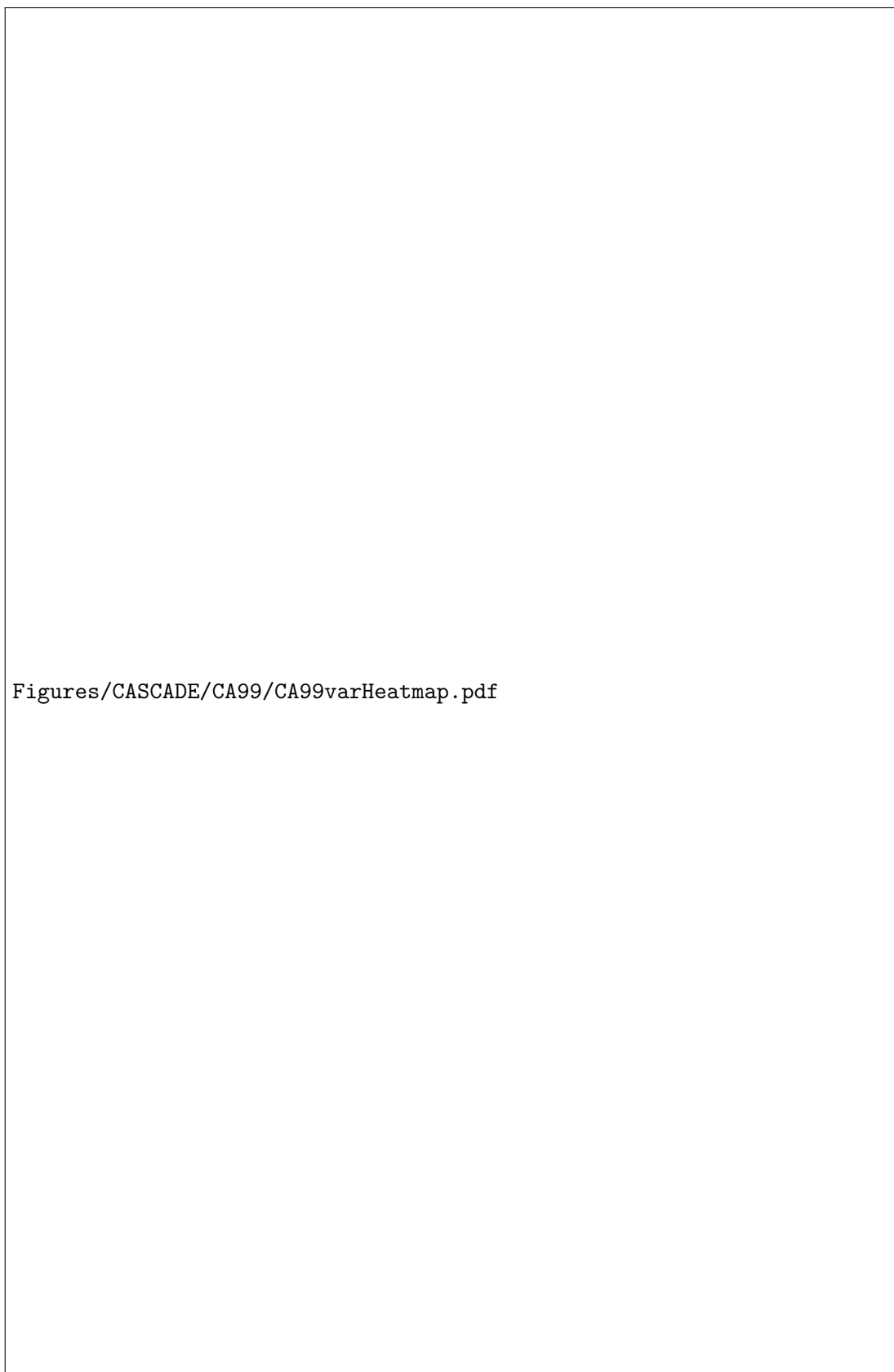


FIGURE 3.6: Heatmap of driver gene variants in patient CA-A: Protein altering mutations are highlighted with their HGVS notation; non protein altering mutations are grouped as “other”.

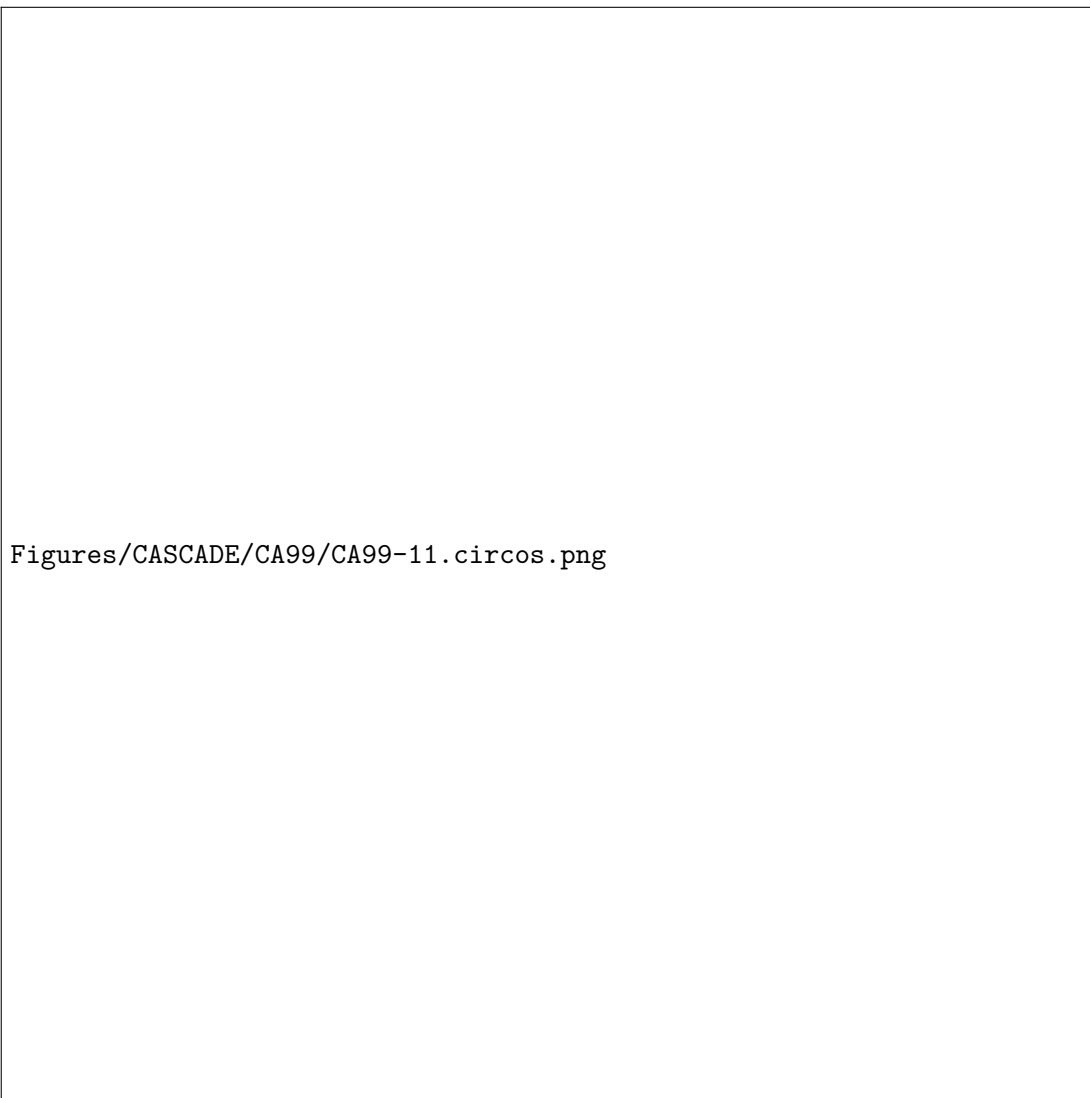


FIGURE 3.7: Circos plot of patient CA-A with somatic structural variants with allele frequency  $> 0.2$ : outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.



FIGURE 3.8: Circos plot of patient CA-A with all somatic structural variants: outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.

Figures 3.7, C.1, C.3, C.4, C.5, C.6 and C.7 show very similar copy number profiles of all tumour samples of patient CA-A, with loss of heterozygosity on almost all chromosomes apart from chromosome 5 and 9. Only Figure C.1 showed a less granular mix of gain and loss of the minor allele, which was most likely due to the lower tumour purity of the sample. However, all samples exhibited high copy number gain levels consistent with whole genome duplication (Tables 3.1 and 3.2). All samples showed a copy number gain in chromosome 7 at the *EGFR* locus leading to *EGFR* amplification (min: 3.9 max: 9.9), which is a known resistance mechanism to Levantinib [234] and was therefore most likely a result of previous treatment (Figure 3.1). Just as expected from the results of the Guardant360 ctDNA results leading up to the death of the patient, there was no *MET* amplification present in the patient at any site.

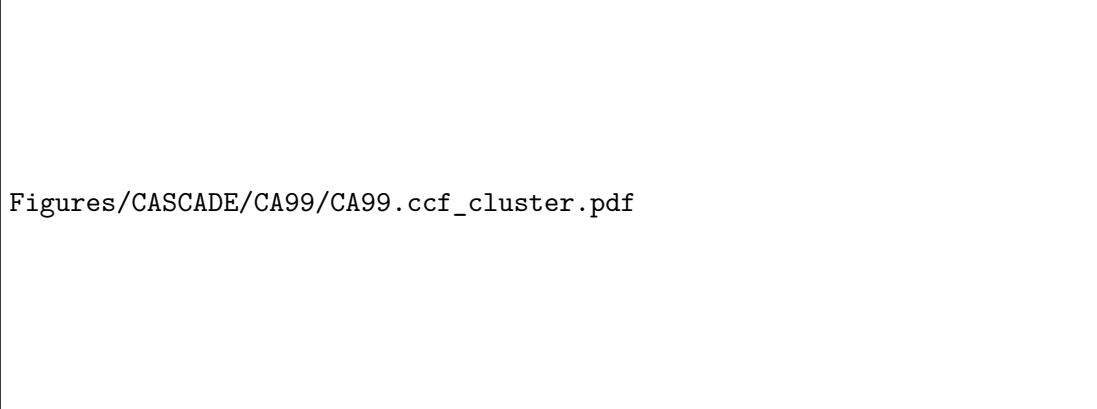
With the absence of any other plausible mechanism, of SNV, SV or CNV nature, the only possible explanation of resistance to Selpercatinib in the patient was the solvent front mutations RET G810C, G810R, G810S, and G810V as described in Solomon et al. [6]. However, the autopsy revealed substantial spatial heterogeneity of the emerged resistance mechanisms within the patient which was previously unappreciated.

TABLE 3.2: Copy number analysis results for patient CA-A: results are taken from the best fit result of PURPLE; WG: whole genome

Sample number	purity	ploidy	polyclonal %	WG duplication
11	0.73	2.86	29.75	True
26	0.61	3.00	26.57	
31	0.21	5.40	49.86	
41	0.28	4.30	42.43	
47	0.46	2.86	29.72	
55	0.38	2.86	38.41	
57	0.61	3.00	28.55	
59	0.69	2.60	37.11	

With the high percentage of likely sub-clonality of samples (Table 3.2) we used PhylogenicNDT to infer clonal hierarchy and complexity. Initial analysis confirmed 141 clone like clusters, which far exceeded the maximum amount PhylogenicNDT was able to visualise. In Figure 3.9 the clonal abundance per sample for the three clones 64, 70 and 139 was shown, which corresponded to RET G810C, RET G810R, and RET G810S respectively, exactly as seen in the longitudinal data (Figure 3.2). The clonal abundance in each sample was highly variable and followed an exclusion pattern, where if one resistant clone was present another was not. Only sample 31 and 55 showed signs of

mixed populations. We attributed these patterns to parallel evolution due to selection pressure of the drug on already established lesions rather than metastatic seeding.



Figures/CASCADE/CA99/CA99.ccf\_cluster.pdf

FIGURE 3.9: Cancer cell fraction of mutation clusters for patient CA-A; transparent polygons show the 95% confidence intervals. Clusters were generated with PhylogiNdt

We attempted to build a clonal tree from the clustered mutations with PhylogiNdt with the restricted clusters using Equation 3.1, however no output was generated after 480 CPU hours. In contrast all other patients required less than 100 CPU hours. We assumed the high subclonality of the data made it impossible to converge to a solution. This again highlights the necessity of computational methods to close the gap between current algorithms and the available datasets.

### 3.3.3 Patient CA-I

This 56 year old female never smoker presented with an *EGFR* exon 19 deletion positive NSCLC in stage IV with metastatic involvement. After an initial good response to Gefitinib treatment, the patient showed progressive nodal disease and was treated with Carboplatin/Gemcitabine chemotherapy with mixed response. After the change to the tyrosine kinase inhibitor Afatinib small intra-cranial metastases were detected and biopsy of a parasternal mass revealed small cell transformation in addition to the *EGFR* T790M resistance mutation. Subsequent treatment changes to Carboplatin/Etoposide as well as CAV (cyclophosphamide, doxorubicin, vincristine) and finally Nivolumab were not successful and the patient died 40 months after diagnosis (Figure 3.10).

Figures/CASCADE/CA51/CA-I\_timeline.pdf

FIGURE 3.10: Timeline of patient CA-I from diagnosis until death: Diagnostic biopsy detected *EGFR* exon 19 deletion lung adenocarcinoma; Second biopsy after 24 months revealed additional *EGFR* T790M mutation and small cell transformation

At autopsy six lesions and one blood sample were collected and biobanked (Figure 3.11). After quality assessment by pathology with H&E stain, all autopsy samples and the initial diagnostic biopsy were sequenced with WES (Table 3.3) and analysed with the standard workflow (Section 3.3.1). The secondary small cell confirmation biopsy at 29 months did not contain enough tissue for sequencing.

TABLE 3.3: Autopsy samples sequenced for patient CA-I: Sample number is the internal sample collection during CASCADE autopsy, the organ of the sample, the fraction of tumour cells from H&E stain and the pathology of the tumour sample. Dx: diagnostic sample

Sample number	Organ	H&E	Type
Dx	right VATS	-	adenocarcinoma
557	parasternal mass	0.9	small cell
559	left diaphragm	0.9	
566	right liver lobe	0.6	
573	right hilar lymph node	0.9	
579	left lung lobe	0.8	
583	left pleura	0.9	



Figures/CASCADE/CA51/CA-I\_schematic\_CA51\_organColours.pdf

FIGURE 3.11: Schematic of tumour lesions in patient CA-I: Primary diagnostic sample shown in red; All six autopsy samples were coloured by organ they were collected from: Parasternal (1), left lung (2), right lung (1), diaphragm (1), liver (1); Additionally a post mortem blood sample was taken

Somatic variant calling showed very little genetic heterogeneity. The original *EGFR* exon 19 deletion was present in all sequenced samples from the diagnostic sample to the 40 months later autopsy samples. No other protein altering somatic mutations were detected at a purity corrected allele frequency  $\geq 0.33$ . While the diagnostic sample presented with a *TERT* H687Q mutation at 25% VAF and sufficient local copy number amplification for 100% cancer cell fraction, none of the autopsy samples showed any support for this variant. Generally, the number of somatic protein altering SNVs and InDels (min: 988 max: 1236 mean: 1090.5) was very close to the average lung adenocarcinoma with an observed tumour mutational burden of 18.75 [233]. In contrast, the diagnostic sample showed a much higher number of mutations, which we attribute to the formalin-fixed paraffin-embedded (FFPE) preservation, which is known to cause DNA damage [235]. This DNA damage could have led to a higher rate of called somatic variants (Figure C.8). In order to appreciate the relationship of the autopsy samples, we removed the diagnostic sample from the phylogenetic analysis. The full phylogeny can be seen in Figure C.9. The phylogeny of SNVs and InDels shows no internal hierarchical structure,

but rather shows that both the stem of tumour initiation and additional private mutations accumulated during the disease progression were approximately equal in number, with the stem being slightly longer. Only a very limited number of somatic variants were shared between cancer samples, which were not part of the initial stem (Figure 3.12). This was consistent with the clinical history, where there was never any clinical remission to treatment, but only mixed response in a subset of already established lesions.

Figures/CASCADE/CA51/CA51phyloAutopsy.pdf

FIGURE 3.12: Phylogeny of autopsy samples from patient CA-I; reconstructed with all somatic SNVs and InDels. Ruler symbolises 2000 variants difference. The phylogeny with diagnostic sample can be found in Figure C.9

In keeping with the small cell transformation, we could highlight an intronic *TP53* mutations, which was present at almost 100% VAF in all autopsy samples. However, the same variant was also found in the diagnostic sample, which was reportedly adenocarcinoma. The variant was therefore not sufficient to drive the transformation, but could have been a predisposition for the patient (Figure 3.13).

When comparing the diagnostic sample with the samples at autopsy the most striking difference was the lower overall copy number gain. The cause of this difference was the amplified minor allele in the diagnostic sample, which was almost completely absent in all autopsy samples. However, this amplification in the diagnostic sample could be rooted in the FFPE DNA damage combined with the lower purity of the sample, which in turn was used as a sign of amplification of the minor allele after purity correction through sequenza. The consistent feature between all samples, diagnostic and autopsy, was the loss of heterozygosity on chromosomes 2, 4, 10 through 13, and 19, with consistent copy number gains at the end of chromosome 1, 3 and 7 and all of chromosome 5 and 6. The high amplification of chromosome 7 is consistent with the origin of the *EGFR* driven primary tumour and the copy number loss of the start of chromosome 17 for all autopsy

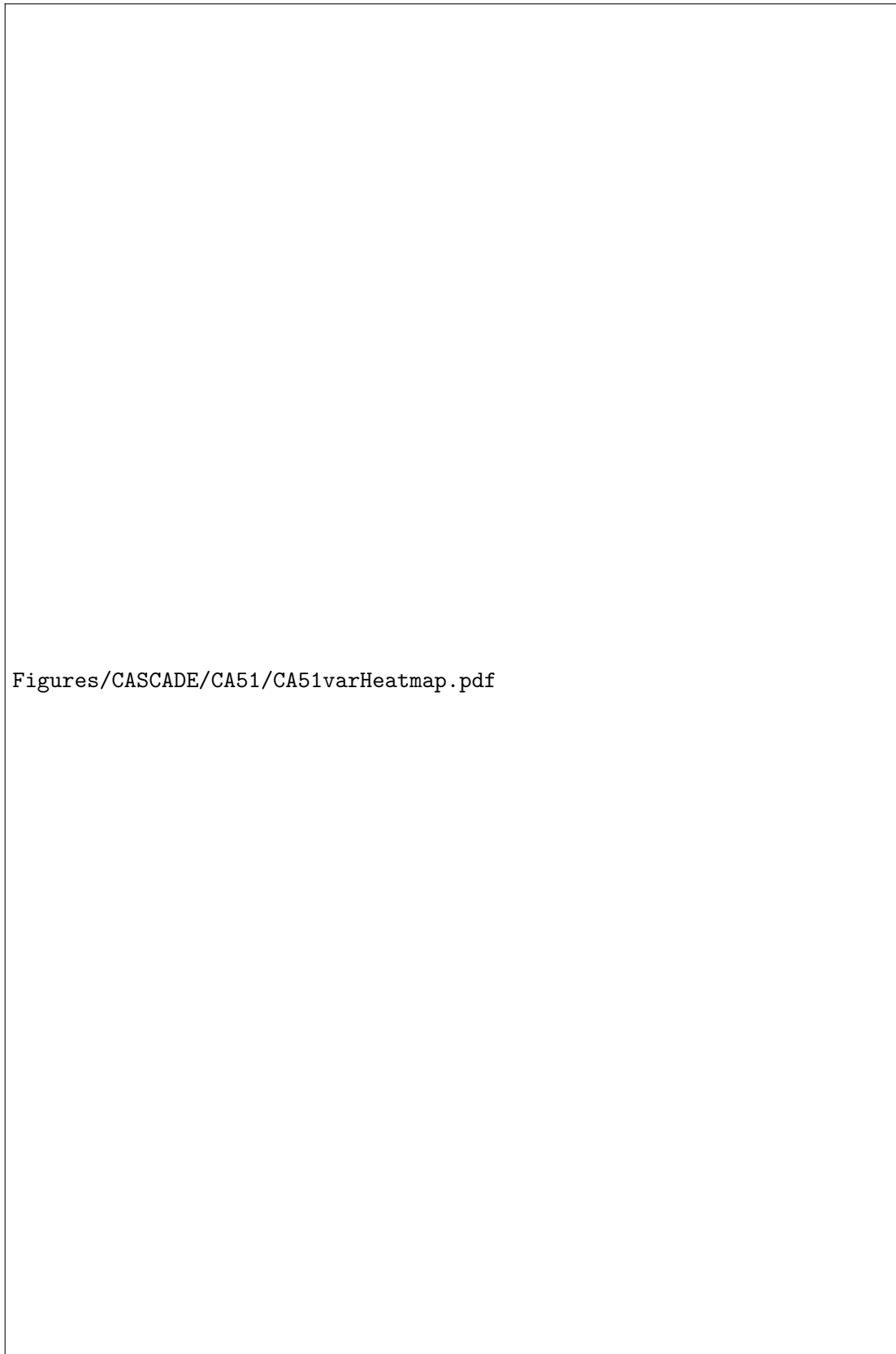


FIGURE 3.13: Heatmap of driver gene variants in patient CA-I: Protein altering mutations are highlighted with their HGVS notation; non protein altering mutations are grouped as “other”.

samples, but only a loss of heterozygosity in the primary sample is consistent with the genetic prerequisites for small cell transformation. (Figure 3.14 vs. Figures 3.15, C.10, C.11, C.12, C.13 and C.14 and Table 3.4).

TABLE 3.4: Copy number analysis results for patient CA-I: results are taken from the best fit result of sequenza

Sample number	purity	ploidy	WG duplication
Dx	0.29	5.5	True
557	0.93	2.6	False
559	0.96	2.6	False
566	0.69	2.7	False
573	0.94	2.6	False
579	0.95	2.6	False
583	0.95	2.6	False

Clonal deconvolution of somatic variants with PhylogicNDT revealed a linear, most likely longitudinal, development of clones adjusting to the changing treatment, with the initial clone 1 containing the exon 19 deletion and individual subclones branching off. While a HLA-A disrupting variant in cluster 13 was observed at high frequency in the diagnostic sample, it was out-competed by other clones at autopsy, and seen as transient cluster 4. Due to the small cell transformation, which correlates with down regulation of major histocompatibility complex (MHC) components, a direct disruption of *HLA-A* was likely not necessary anymore. Some clones also were only observed in specific sites, like cluster 15 or not found at certain sites (cluster 5) which pointed to high heterogeneity of disease at autopsy (Figures 3.16 and 3.17).

These results agree with the phylogenetic reconstruction which showed initial shared somatic evolution of all sites, with strong individual evolution and limited substructure.



FIGURE 3.14: Circos plot of patient CA-I sample dx with somatic structural variants: outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.

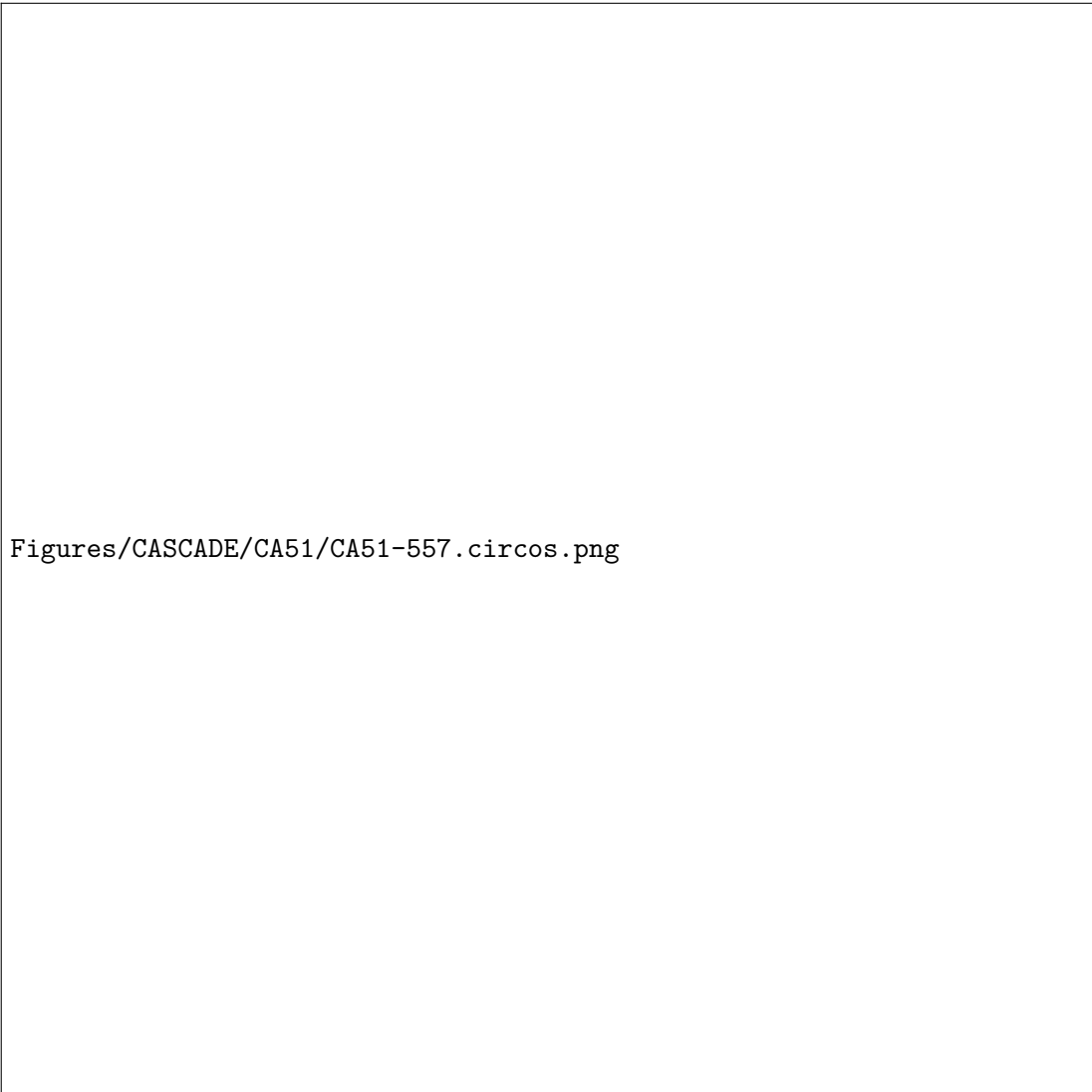
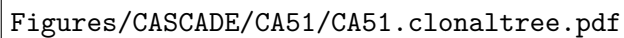
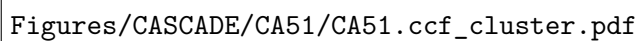


FIGURE 3.15: Circos plot of patient CA-I sample 557 with somatic structural variants: outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.



Figures/CASCADE/CA51/CA51.clonaltree.pdf

FIGURE 3.16: Clonal evolutionary tree of patient CA-I; Highest support tree for clustered ccf clones generated with PhylogicNDT; Support for clone is shown in parenthesis; Major driver alterations of clones were annotated; Clusters with less than 5 supporting variants were discarded; Cluster with 2000 supporting variants only present in sample Dx was discarded as FFPE artefact



Figures/CASCADE/CA51/CA51.ccf\_cluster.pdf

FIGURE 3.17: Cancer cell fraction of mutation clusters of clonal tree for patient CA-I; transparent polygons show the 95% confidence intervals. Clusters and cluster colours are taken from Figure 3.16

### 3.3.4 Patient CA-J

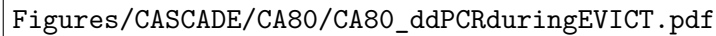
Patient CA-J was a 65 year old female never smoker, who presented with moderately differentiated lung adenocarcinoma (Stage IIIB). Molecular pathology revealed metastatic EGFR L858R positive disease. Initial treatment with both Carboplatin/Paclitaxel and radiotherapy was halted after detection of metastatic disease with bone, left adrenal gland and bilateral lung lesions and she was changed to the tyrosine kinase inhibitor Erlotinib. Progressive pulmonary disease and subsequent left lung core biopsy showed an additional BRAF V600E mutation. Treatment was adjusted to Carboplatin and Pemetrexed, however further disease progression was evident. The patient was finally enrolled in the EVICT trial involving treatment with the BRAF inhibitor Vemurafenib in combination with Erlotinib, which led to stable bone metastasis after one month, but ultimately led to progression of both pulmonary and bone metastases. The patient died 29 months after initial diagnosis (Figure 3.18).

Figures/CASCADE/CA80/CA-J\_timeline.pdf

FIGURE 3.18: Timeline of patient CA-J from diagnosis until death: Diagnostic biopsy detected EGFR L858R positive stage IIIB lung adenocarcinoma; Second diagnosis after 3 months revealed additional brain, bone and lung metastasis with a reclassification to stage IV; Biopsy at the end of erlotinib treatment revealed additional BRAF V600E mutation; one blood sample was taken during the second round of chemotherapy and three more during the time the patient was enrolled in the EVICT trial

Serial plasma sampling Just before and during the enrollment in the EVICT trial allowed us to monitor the genomic landscape of the disease during treatment via specially designed ddPCR analysis. After the second round of chemotherapy a  $\approx 60\%$  variant allele fraction of EGFR L858R was found, suggesting a high ctDNA fraction. The after initial partial response to the change to Vemurafenib and Erlotinib in the EVICT trial, accompanied with a substantial drop in detectable EGFR L858R, the patient relapsed. This progression could also be observed in the steady increase in the TP53 “stop gained“ and BRAF V600E mutation, which were detectable at higher levels than before the EVICT trial (Figure 3.19).

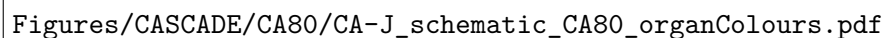




Figures/CASCADE/CA80/CA80\_ddPCRduringEVICT.pdf

FIGURE 3.19: Blood plasma analysis of patient CA-J: Three putative driver mutations were analysed at four time points during treatment. progression and partial response were assigned by clinicians based independent of ctDNA analysis; Y-axis was broken from 25-60 for visibility

At autopsy 18 sites of disease were resected and biobanked. A representative six samples from different organs and sites were selected and WGS was performed after H& E staining confirmed high enough tumour purity (Table 3.5, Figure 3.20). All WGS samples were analysed with the standard analysis workflow (Section 3.3.1).



Figures/CASCADE/CA80/CA-J\_schematic\_CA80\_organColours.pdf

FIGURE 3.20: Schematic of analysed tumour lesions in patient CA-J: Primary diagnostic sample shown in red; All 18 autopsy samples were coloured by organ they were collected from: skull (1), left lung(5), right lung (6), diaphragm (2), liver(2), adrenal gland (2); Additionally to the post mortem blood sample, four serial blood samples were taken (Figure 3.18)

TABLE 3.5: Autopsy samples sequenced for patient CA-J: Sample number is the internal sample collection during CASCADE autopsy, the organ of the sample, the fraction of tumour cells from H& E stain and the pathology of the tumour sample. Dx: diagnostic sample

Sample number	Organ	H&E	Type
Dx	left lung core	-	adenocarcinoma
2	adrenal gland	0.5	
20	right lower lung	0.7	
24	left upper lung	0.9	
28	left middle lung	0.5	
32	right upper lung	0.5	
42	base of skull	0.4	

Somatic variant calling revealed heterogeneity of resistance and driver mutations. While the initial driver variant EGFR L858R was present in all autopsy samples, sample 2 presented with an allele frequency of 0.13 with clonal presence in all other sample. The secondary BRAF V600E mutation, which was detected in the progression biopsy 22 months after diagnosis, was not present in sample 2 and only at low allele frequency (0.13) in sample 28. Only sample 32 showed the *BRAF* mutation at 100% VAF. While the absence in sample 2 could be explained by the overall low tumour purity of the sample, both 24 and 42 had a higher than 50% estimated tumour purity (Table 3.6) and showed a lower VAF, therefore suggesting a lesser involvement of the mutation in resistance.

Additionally to the *BRAF* mutation, some sites (20, 24, 32, and 42) developed a “stop gained” mutation in *TP53* (TP53 G38Ter) at 100% VAF. While both the *BRAF* and *TP53* mutations were present at similar allele fractions in the diagnostic sample (Dx) suggesting a clonal structure, the *TP53* variant was more prevalent at autopsy in multiple samples. So while *BRAF* and *TP53* mutations were correlated in the diagnostic sample, the *TP53* “stop gained” developed independently. Sample 28 in spite of showing traces of BRAF V600E did not develop a *TP53* mutation and sample 2, which did not contain a *BRAF* change instead exhibited two different additional putative driver events (FLT4 V1097M and KEAP1 Q282H) which were not observed in any other sample. Finally, Sample 32 also contained a subclonal KRAS L5Q mutation in addition to both *BRAF* and *TP53* mutations (Figure 3.22).

The emergence of the *TP53* mutations was related to expansion of samples 20, 24, 32, and 42 differentiating them from the adrenal gland (2) and the original site of disease.

This very early split and seeding and the low abundance of the putative *EGFR* driver mutation suggests at very different disease trajectories (Figure 3.21).

Figures/CASCADE/CA80/CA80phylo.pdf

FIGURE 3.21: Phylogeny of autopsy samples from patient CA-J; reconstructed with all somatic SNVs and InDels. Ruler symbolises 4000 variants difference.

Similar to the short variants, there were some structural variants present in all samples. The inversions on chromosome 12 as well as the co-located break and fusion with the start of chromosome 5 could be observed in all samples, even those with very low tumour purity, but the inversions and fusions of chromosome 7, 8, 9, and 11 could only be seen in the higher purity samples 20, 24, 32, and 42 and most were subclonal, as they only had a median allele frequency of 14.3% (min: 10.3%, max: 99.7%) in all samples. While multiple samples exhibited gene fusions with lung cancer driver and resistance genes like *BRAF*, *FGFR1* and *GNAS* these fusions were only present at subclonal levels  $\leq 10\%$ .

All samples apart from sample 2 showed whole genome duplication and high polyclonality, which suggests that in addition to the heterogeneity observed through short and structural variants, there was an additional level of heterogeneity of copy number alterations. The lower purity of sample 2 might have been a confounding factor, however both sample 28 and 32 showed lower purities, but a much higher polyclonality and genome duplication. While sample 2 had several minor focal amplifications in *PMS2*, *STK11* and *GADD45B*, no major copy number amplification was found. All other samples showed amplifications in *KRAS* (min: 2.9, max: 5.0, median: 4.6), *CDK4* (min: 3.2, max: 24.4, median: 21.7) and *BRAF* (min: 2.1, max: 6.0, median: 3.9) in addition to the highly amplified *EGFR* (min: 10.6, max: 266.7, median: 197.4) and *MET* (min: 4.2, max: 6.3, median: 4.6) locus. Both *EGFR* and *MET* copy number gain most likely were the resistance mechanism to the initial treatment with the tyrosine kinase inhibitor Erlotinib. (Figures 3.23, 3.24, C.15, C.16, C.17 and C.18 , Table 3.6).

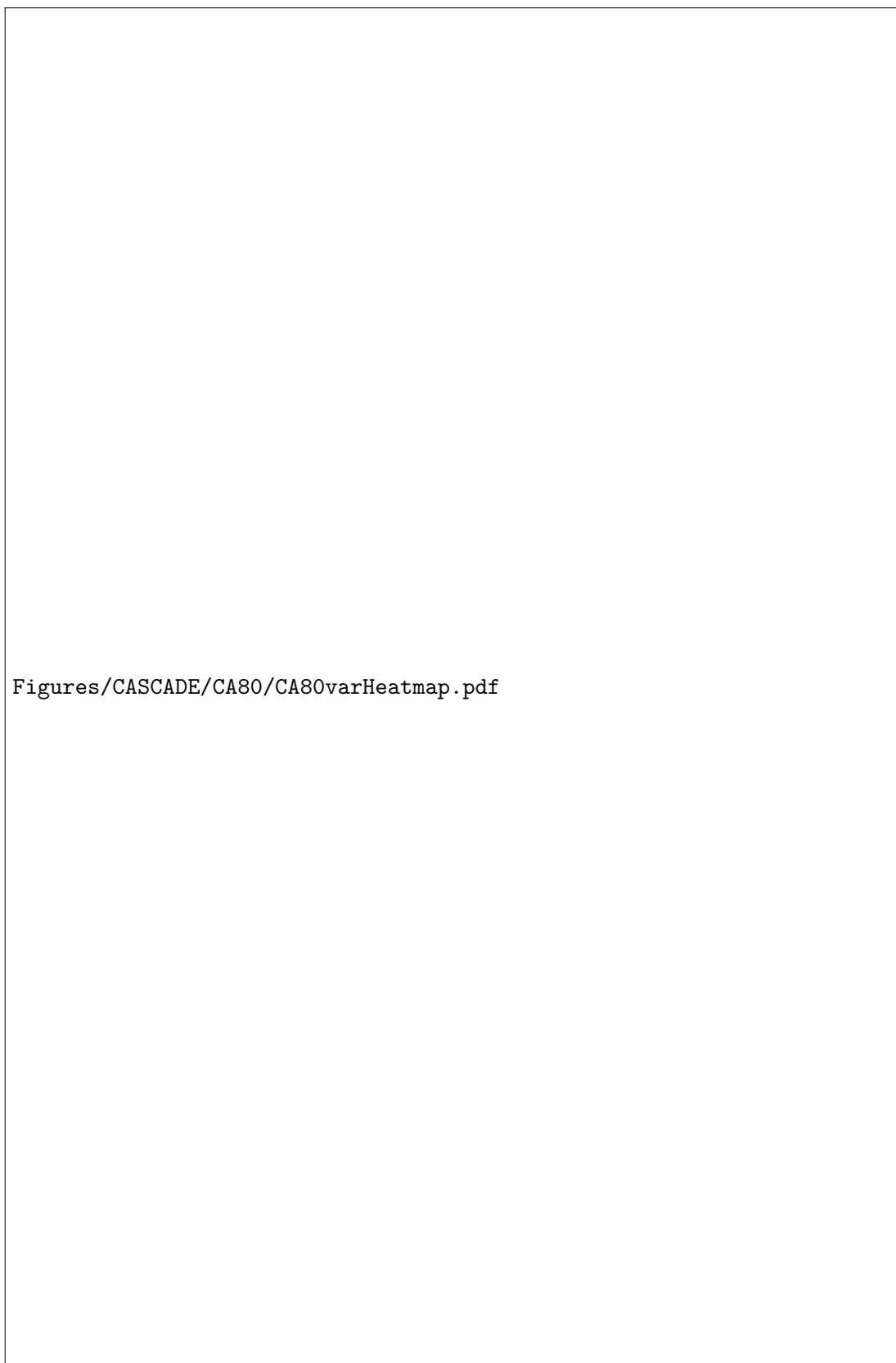


FIGURE 3.22: Heatmap of driver gene variants in patient CA-J: Protein altering mutations are highlighted with their HGVS notation; non protein altering mutations are grouped as “other”.



FIGURE 3.23: Circos plot of patient CA-J sample 2 with somatic structural variants with allele frequency  $\geq 0.1$ : outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.



FIGURE 3.24: Circos plot of patient CA-J sample 20 with somatic structural variants with allele frequency  $\geq 0.1$ : outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.

TABLE 3.6: Copy number analysis results for patient CA-J: results are taken from the best fit result of PURPLE; WG: whole genome

Sample number	purity	ploidy	polyclonal %	WG duplication
2	0.16	2.18	22.53	False
20	0.39	4.80	43.36	True
24	0.73	3.70	30.32	
28	0.18	3.90	42.28	
32	0.25	4.75	48.45	
42	0.52	3.35	41.63	

Both the somatic variants as well as copy number analysis showed clear signs of a *BRAF* driven tumour with both *BRAF* mutations and amplifications as well as amplification of *CDK4*. However the patient also displayed potential alternative methods of resistance, like *KEAP1* and *FLT4* mutations which could only be appreciated by analysing multiple sites of the cancer.

Figures/CASCADE/CA80/CA80.ccf\_mutations.pdf

FIGURE 3.25: Cancer cell fractions of individual mutations of cluster 2 containing TP53 mutations for patient CA-J : each dot represented a distinct variant. Variants were connected with a dotted line to the same variant in other samples. Mutations were clustered with PhylogicNDT;

PhylogicNDT analysis revealed a high degree of subclonality, consistent with the results from PURPLE. However, the low tumour purity of samples 2 and 28 lead to unrealistic clustering of variants in these samples. While the TP53 mutation was not found in either sample 2 or sample 28 (Figure 3.22), the cluster containing this mutation was assigned a 100% cancer cell fraction overall. As the individual mutations do show multiple substructures in this cluster, for example connecting 2 and 20 at low ccf as well as high, and parameter tuning did not lead to a more granular representation, we considered the results to be low quality and not interpretable (Table 3.6, Figure 3.25).

### 3.3.5 Patient CA-K

This 69 year old was a male patient who presented with multifocal lung adenocarcinoma without distant metastasis. The most PET avid location (left upper lung) was designated the primary site over the two other less avid locations (right upper lung and left lower lung). Initial treatment with the tyrosine kinase inhibitor Gefitinib was stopped when the dominant lung lesion and a hilar node lesion showed signs of progression and he was changed to Afatinib. A CT scan after 50 months showed a mild increase at the primary site, stable disease in satellite nodules and no new metastatic sites. After short treatment with Erlotinib, molecular pathology of a biopsy revealed the acquired EGFR T790M resistance mutation. Enrolment in the CLOVIS trial involving treatment with the EGFR tyrosine kinase inhibitor Rociletinib and chemotherapy was ceased due to disease progression. Biopsy 2 confirmed the T790M mutation and therapy with Osimertinib was started, which led to slowly progressive disease. Due to new intracranial disease, and increase in lung and renal disease treatment was switched to the PD-1 inhibitor Nivolumab, but no remission was achieved and the patient died after 103 months (Figure 3.26).

Figures/CASCADE/CA82/CA-K\_timeline.pdf

FIGURE 3.26: Timeline of patient CA-K from diagnosis until death: Diagnostic biopsy detected EGFR L858R positive lung adenocarcinoma; Biopsy 1 after 66 months showed additional EGFR T790M mutation; Biopsy 2 showed no additional variants; one blood sample was taken towards the end of Osimertinib treatment and one second one during PD-1 checkpoint blockade treatment. E: Erlotinib; R: Rociletinib; P: PD-1 inhibitor

Analysis of the plasma sample collected five months prior to the death of the patient with the AVENIO commercial kit revealed an *AKT*, a *KRAS*, and several *EGFR* resistance mutations with high confidence at very low allele frequency. Out of the reported low confidence somatic variants, several more known *EGFR* resistance alterations could be validated in the autopsy samples. Due to the overall low VAF of found variants, we assumed a low ctDNA fraction of the sample. Nevertheless, the multiple *EGFR* mutations suggested a high polyclonality of the disease and a purely genomically *EGFR*



driven cancer (Table 3.7). While these results suggest that a high degree of the final heterogeneity was already present, the longitudinal distance of the plasma sample to the autopsy samples suggests subsequent evolution, which could also be seen absence of the *APC* mutation in the plasma (Figure 3.29).

TABLE 3.7: Somatic variants found in plasma with AVENIO sequencing for patient CA-K: all high confidence variants are shown; low confidence variants also seen at autopsy in any sample were also selected

Gene	Change	VAF (%)	High confidence	Found at autopsy	
AKT1	Glu17Lys	0.88	True	True	
KRAS	Gly12Val	0.18		False	
EGFR	Asp761Tyr	0.05		True	True
	Thr790Met	0.71			
	Cys797Ser	0.26			
	Leu858Arg	1.03			
	Leu718Gln	0.69			
	Ser720Thr	0.67	False		

At autopsy 18 sites were resected and biobanked and seven high quality representative samples from different organs were selected for WGS (Figure 3.27, Table 3.8) and analysed with the standard workflow (Section 3.3.1).

TABLE 3.8: Autopsy samples sequenced for patient CA-K: Sample number is the internal sample collection during CASCADE autopsy, the organ of the sample, the fraction of tumour cells from H& E stain and the pathology of the tumour sample. Dx: diagnostic sample

Sample number	Organ	H&E	Type
Dx	left lung core	0.8	adenocarcinoma
1	right kidney	0.7	
4	right upper lung	0.8	
5	right lower lung	0.7	
6	right middle lung	0.7	
8	left lower lung	0.9	
9	left upper lung	0.5	
13	left brain	0.5	

Joint somatic variant calling on all autopsy samples revealed significant genetic heterogeneity of resistance mechanisms present at each site. While the initial EGFR L858R mutation was still present in all sequenced lesions, the left lower lung sample (8) was the only one with a homozygous variant. All other samples presented with 50% VAF for the activating mutation. Biopsy 1, 66 months after diagnosis, showed the additional EGFR T790M mutation, but at autopsy sample 1 (adrenal gland) did not show evidence of the mutation at all and samples 8, 9, and 13 (left lung and brain) exhibited the variant

Figures/CASCADE/CA82/CA-K\_schematic\_CA82\_organColours.pdf

FIGURE 3.27: Schematic of analysed tumour lesions in patient CA-K: Primary diagnostic sample shown in red; All 17 autopsy samples were coloured by organ they were collected from: Brain (6), left lung (4), right lung (4), kidney (3); Additionally to the post mortem blood sample, two serial blood samples were taken (Figure 3.26)

at subclonal frequencies ( $< 30\%$  VAF). Either the mutation was already subclonal at biopsy, or the resistance was outcompeted by a different clone due to the Osimertinib treatment which targets T790M. The adrenal gland lesion, which did not contain the T790M mutation, instead presented with two other clonal EGFR mutations (S720T and L718Q) which are both known resistant mechanisms to Osimertinib [236, 237].

Additionally the mutations AKT1 G17L and APC E190Ter bifurcated the the autopsy samples into two groups, because the variants were mostly mutually exclusive, where only samples 8 and 9 showed both the stop gained *APC* mutation at 100% VAF with very low VAF of the *AKT1* mutation. Furthermore, several *EGFR* mutations also showed different spatial clonal abundance. Multiple samples exhibited different C797 substitutions both of which are known resistance mechanisms to Osimertinib [238, 239]: Sample 4 contained the EGFR C797G mutation at 7% VAF whereas samples 6, 8, and 9 had a C797S mutation at 1%, 14%, and 22% VAF respectively. However, while 8 shared the sample protein change (C797S) the genomic change was different to both sample 6 and 9. Only sample 4 contained EGFR L792H as a subclonal mutation at 11% VAF

and both sample 1 and 6 contained a subclonal EGFR S720T mutation, another known resistance mechanisms [236, 240]. Additionally to the adrenal sample both lower and middle lower lung samples contained EGFR L718Q. Lastly, all but sample 8 contain the *SMAD4* frameshift mutation at a median cancer cell fraction of 73% (min: 16%, max: 100%) (Figure 3.29).

These genomic changes grouped the samples according to their anatomical location, with right kidney (1) and left brain (13) as outliers, but left (8 and 9) and right lung (4, 5, and 6) samples clustering together (Figure 3.28).

Figures/CASCADE/CA82/CA82phylo.pdf

FIGURE 3.28: Phylogeny of autopsy samples from patient CA-K; reconstructed with all somatic SNVs and InDels. Ruler symbolises 4000 variants difference

Similar to the short variants, structural variants and copy number changes also showed a difference between samples 8 and 9 compared to the rest. While all samples showed inversions on chromosome 6, 8 and 9 with fusions between chromosome 1 and 18, chromosome 6 and 8 and chromosome 8 and 9, samples 8 and 9 also showed a fusion of chromosome 3 with 18 and additional inversions on chromosome 18. The additional inversions and haploinsufficiency directly affect *SMAD4*. These structural changes complemented the “missing” *SMAD4* frameshift mutation in these samples and suggested a key role of *SMAD4* in the resistance to treatment.

Samples 8 and 9 were the only samples in the patient with significantly amplified copy numbers resulting in a whole genome duplication, however they still exhibited the same pattern of loss of heterozygosity. In all samples we observed a copy number gain in the q arm of chromosome 1 amplifying both *NTKR1* and *DDR2* with a loss of heterozygosity for *NRAS* and *MTOR* on the p arm of the same chromosome. The loss of heterozygosity presented in all samples on chromosome 3 reduced the representation of

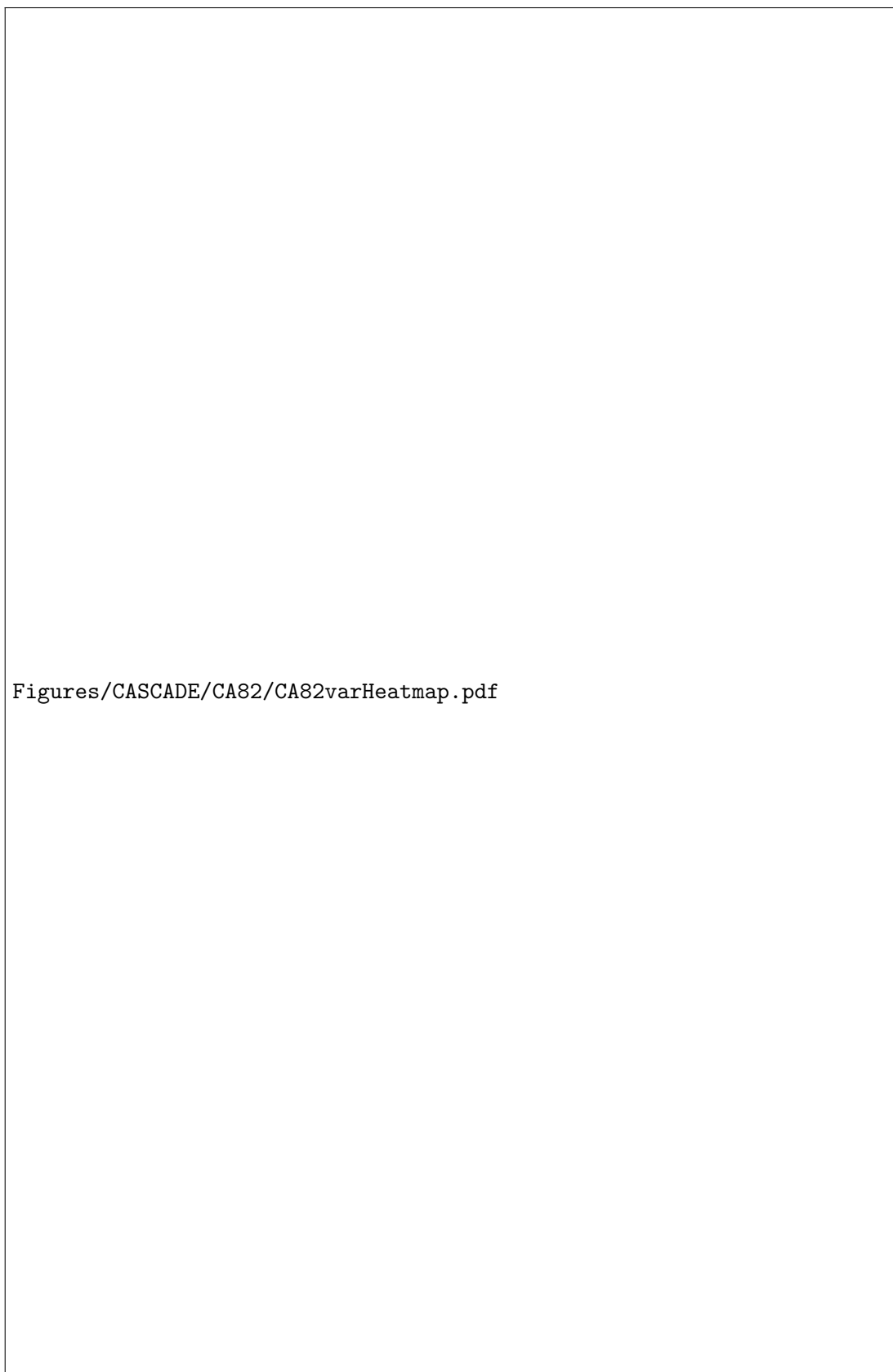


FIGURE 3.29: Heatmap of driver gene variants in patient CA-K: Protein altering mutations are highlighted with their HGVS notation; non protein altering mutations are grouped as “other”.

*TM4SF1*, *PIK3CA*, and *USP13*, however in both sample 8 and 9 the other chromosome was amplified leading to a copy number neutral area. The loss of heterozygosity on chromosome 5 leads to a haploinsufficiency for *APC*, *PIK3R1*, *CSF1R*, *PDGFRB*, and *FLT4* for all samples, which combined with the stop mutation in samples 8 and 9 is suggestive of a common resistance pathway. The loss of heterozygosity on chromosome 8 affected *TUSC3* and *FGFR1*. No lung cancer driver genes were affected by the loss of heterozygosity on chromosome 15. The seemingly heterozygous loss of chromosome X resulted in the loss of *ARAF* and *AR* as a consequence, but the loss must have been subclonal given the sex of the patient was male. Lastly, the additional copy number loss only present in both samples 8 and 9 affected *JAK2*, *CD274*, and *PDCD1LG2* and these samples also showed *EGFR* amplification in contrast to all other samples (Figures 3.30, C.19, C.20, C.21, 3.31, C.22 and C.23, Table 3.9).

TABLE 3.9: Copy number analysis results for patient CA-K: results are taken from the best fit result of PURPLE; WG: whole genome

Sample number	purity	ploidy	polyclonal %	WG duplication
1	0.78	1.84	7.62	False
4	0.48	1.84	4.80	False
5	0.58	1.88	1.02	False
6	0.79	1.86	6.93	False
8	0.30	3.40	6.44	True
9	0.69	3.45	8.32	True
13	0.47	1.90	0.05	False

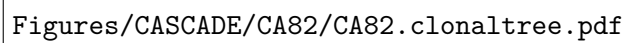
Similar to the phylogeny, the clonal deconvolution with PhylogicNDT revealed a higher complexity of disease after a bottle neck, where the right side lung samples (samples 4, 5, and 6) all presented with a very high prevalence of cluster 11 and followed by clonal diversification after the CNBD1 stop-gained mutation. In contrast samples 1 and 13 (kidney and brain) showed an additional EGFR S720W mutation separating the distant sites from the original lung disease. Finally, the original site of disease in the left lung lobe displayed less diversification, however the early split of cluster 1 into 10 and 3 splitting left lung and all other sites suggests early metastatic seeding (Figures 3.32 and 3.33).



FIGURE 3.30: Circos plot of patient CA-K sample 1: outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.

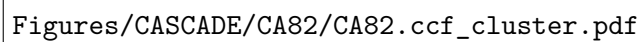


FIGURE 3.31: Circos plot of patient CA-K sample 8: outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.



Figures/CASCADE/CA82/CA82.clonaltree.pdf

FIGURE 3.32: Clonal evolutionary tree of patient CA-K; Highest support tree for clustered ccf clones generated with PhylogicNDT; Support for clone is shown in parenthesis; Major driver alterations of clones were annotated; Clusters with less than 10 supporting variants were discarded



Figures/CASCADE/CA82/CA82.ccf\_cluster.pdf

FIGURE 3.33: Cancer cell fraction of mutation clusters of clonal tree for patient CA-K; transparent polygons show the 95% confidence intervals. Clusters and cluster colours are taken from Figure 3.32



### 3.3.6 Patient CA-L

This 68 year old female ex-smoker presented with *EGFR* mutant NSCLC, however after 12 months of the treatment with the EGFR inhibitor Erlotinib a transformation to small cell lung cancer (SCLC) was detected. While previously it was thought that the different subsets of lung cancers are distinct, more and more evidence is found showing neuroendocrine transformation as a resistance mechanism to targeted therapies not only in lung but also in prostate cancers [241, 242]. The treatment was altered to chemotherapy and then PD-1 inhibition, however due to the loss of MHC-I antigen presentation of small cell lung cancer, the tumour failed to respond [5] and the patient died after 29 months (Figure 3.34).

Figures/CASCADE/CA86/CA-L\_timeline.pdf

FIGURE 3.34: Timeline of patient CA-L from diagnosis until death: Diagnostic biopsy detected EGFR exon 19 deletion positive lung adenocarcinoma; Biopsy after 15 months Erlotinib treatment showed signs of small cell transformation; blood samples were taken at the end of the first Erlotinib treatment, during the chemotherapy treatment and 28 and 29 months after the initial diagnosis.

During autopsy 25 lesions were resected and biobanked and representative samples, both adeno- and small cell carcinoma according to histology, were selected for WES (Figure 3.35, Table 3.10) and analysed with the standard workflow (Section 3.3.1). For granular analysis of the transition from adeno- to small cell carcinoma, the progression sample after 15 months (P) was dissected to the individual types based on histology staining.

Even though not all samples had changed from adeno- to small cell carcinoma, all samples showed the *TP53* “stop gained” mutation at 100% VAF, showing that the *TP53* mutation was not sufficient for the histological transformation [243]. Unsurprisingly, the samples that remained adeno (17A and 26) show a higher dependency on *EGFR* which led to higher clonal abundance of the initial *EGFR* exon 19 deletion and a subsequently

Figures/CASCADE/CA86/CA-L\_schematic\_CA86\_organColours.pdf

FIGURE 3.35: Schematic of analysed tumour lesions in patient CA-L: Primary diagnostic sample shown in red; Samples are coloured by organ they were collected from: left lung (6), right lung (9), sternum (1), diaphragm (2), liver (3), adrenal gland (1) kidney (2), anal (1); Additionally to the post mortem blood sample, five serial blood samples were taken (Figure 3.34)

TABLE 3.10: Autopsy samples sequenced for patient CA-L: Sample number is the internal sample collection during CASCADE autopsy, the organ of the sample, the fraction of tumour cells from H& E stain and the pathology of the tumour sample.

P.1/2: micro-dissected progression biopsy (80% small cell, 20% adeno)

Sample number	Organ	H&E	Type
P.1	right lung core	>0.9	small cell
P.2	right lung core		adenocarcinoma
8	right upper lung	0.9	small cell
17A	left lower lung	-	poorly differentiated adeno
26	right kidney	1	adenocarcinoma

higher VAF of EGFR T790M, while the small cell transformed lung sample (8) acquired a secondary TP53 M40I mutation. No additional variants were close to clonal representation (Figure 3.37).

Surprisingly even though the samples were taken at different times (one at progression and one at autopsy) the small cell transformed samples P.1 and 8 were evolutionarily more closely related and shared more variants, than to the sample P.2 which was taken at the same time. In contrast, the two adenocarcinoma samples taken at autopsy were

clustered together. Additionally, the split of sites of small cell transformation and adenocarcinoma already had happened before the progression sample and the small cell transformed samples appeared to be evolutionary different from samples 17A and 26. In general, the phylogeny suggested the presence of at least 3 distinct trajectories: one giving rise to the adeno sample at progression (P.2), one resulting in the small cell sample P.1 and its longitudinal successor sample 8 and lastly the two adenocarcinoma samples 17A and 26 (Figure 3.36).

Figures/CASCADE/CA86/CA86phylo.pdf

FIGURE 3.36: Phylogeny of samples from patient CA-L; reconstructed with all somatic SNVs and InDels. Ruler symbolises 2000 variants difference.

Copy number analysis with sequenza revealed a high prevalence of loss of heterozygosity in all samples, but both sample P.2 and 8 showed almost no copy number gains on chromosome 9 and 10 with sample 8 even extending through to chromosome 12. In general small cell transformed samples showed a higher level of copy number gain than the original adenocarcinoma. The difference in copy number in the two spatially intertwined types of cancers can only be attributed to the small cell transformation. Additionally to the increased overall ploidy of the small cell sample P.1 over P.2 (Table 3.11), P.2 also lost chromosome X completely (Figure 3.38 vs. Figure 3.39). Interestingly, the small cell samples still had the same high amplification level of EGFR seen in the adenocarcinoma samples (min: 6 max: 13) suggesting the transformation retained EGFR based signalling. While commonly small cell transformation is associated with RB1 loss, the locus was amplified in all samples with a loss of heterozygosity. As this patient's sequencing was restricted to exonic regions, we could not rule out a regulatory defect. Interestingly, while the small cell transformed part of the progression sample (P.1) showed a heterozygous loss of chromosome X, the adenocarcinoma part (P.1) showed an almost

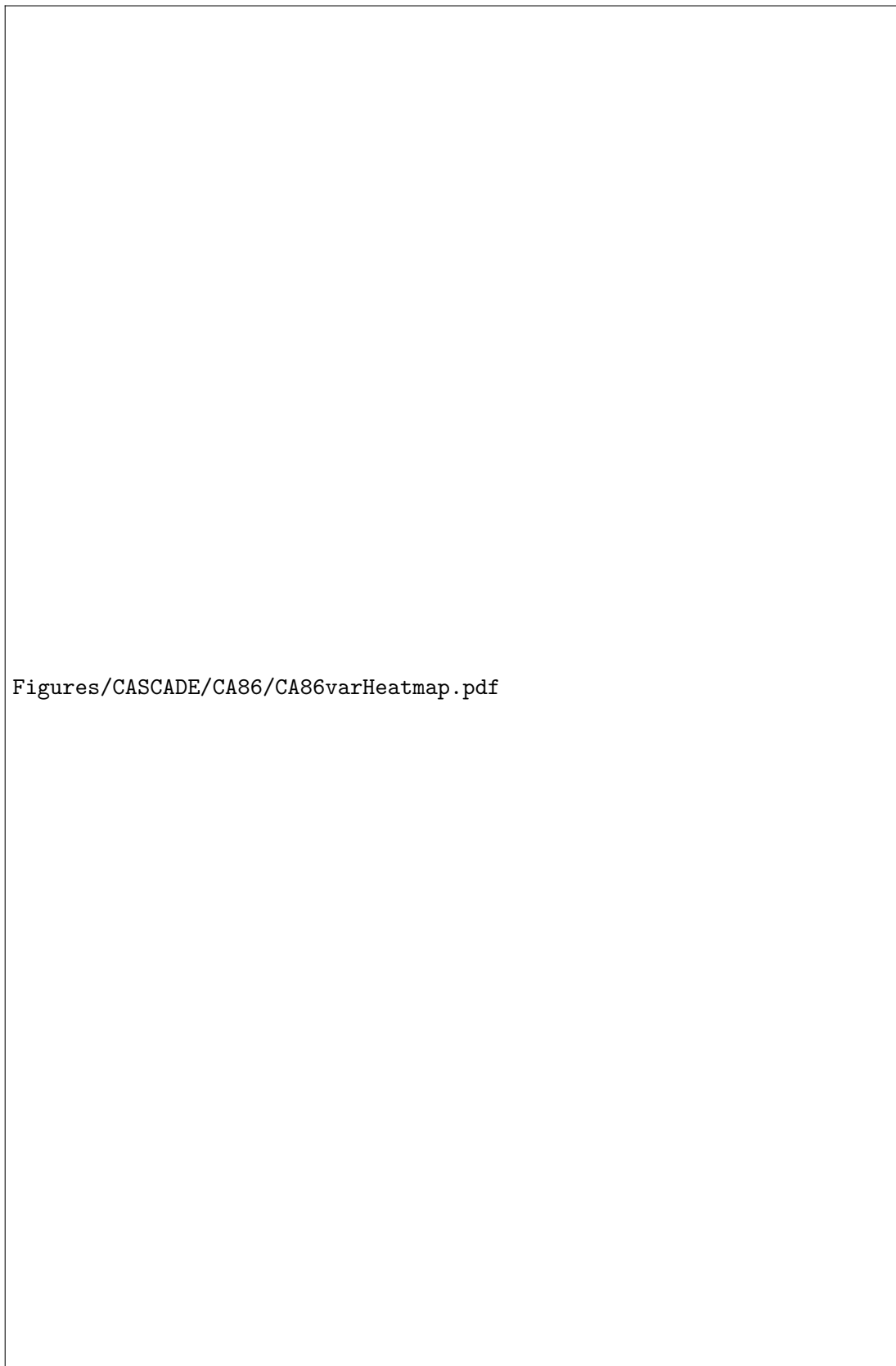


FIGURE 3.37: Heatmap of driver gene variants in patient CA-L: Protein altering mutations are highlighted with their HGVS notation; non protein altering mutations are grouped as “other”.

complete loss of chromosome X, which could not be observed in any of the autopsy samples, which instead showed and amplification. This indicated, that the small cell transformation happened at multiple sites instead of being spread after the transformation (Figures 3.38, 3.39, C.24, C.25 and C.26), .

Figures/CASCADE/CA86/CA86-17B037524-1-S.circos.png

FIGURE 3.38: Circos plot of patient CA-L sample P.1: outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.

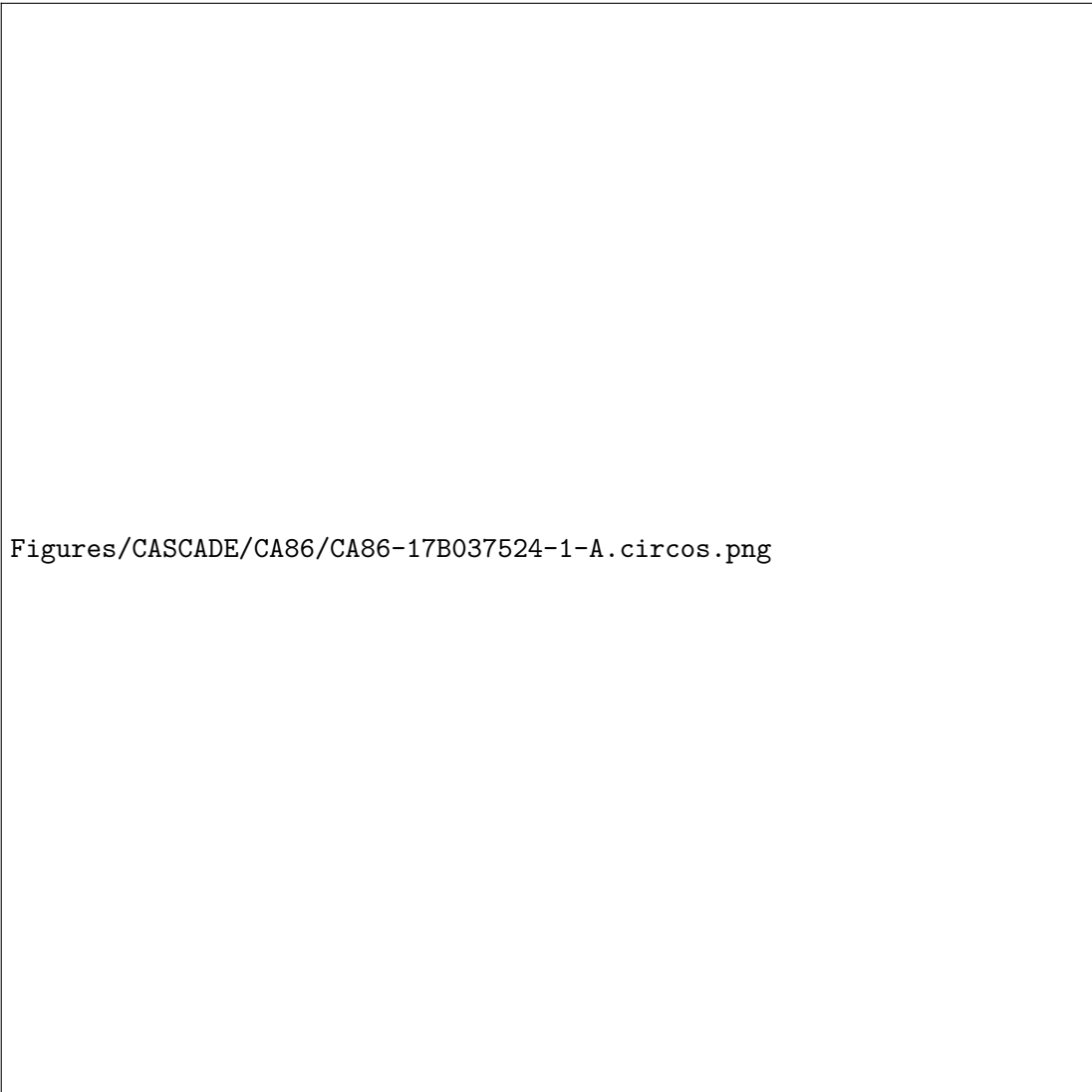


FIGURE 3.39: Circos plot of patient CA-L sample P.2: outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.

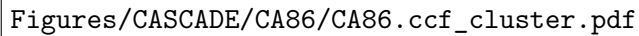
TABLE 3.11: Copy number analysis results for patient CA-L: results are taken from the best fit result of sequenza

Sample number	purity	ploidy	WG duplication
P.1	0.86	3.3	True
P.2	0.27	2.1	False
8	0.96	3.1	True
17A	0.18	4.2	True
26	0.28	3.7	True

Figures/CASCADE/CA86/CA86.clonaltree.pdf

FIGURE 3.40: Clonal evolutionary tree of patient CA-L; Highest support tree for clustered ccf clones generated with PhylogicNDT; Support for clone is shown in parenthesis; Major driver alterations of cluster were annotated; Clusters with less than 5 supporting variants were discarded.

Clonal deconvolution with PhylogicNDT showed a split into clusters 6 and 2 from the initial clone with both the initiating *EGFR* deletion and the *TP53* “stop gained” mutation. While cluster 6 is present in all samples, the small cell transformed samples P.1 and 8 show a much lower cancer cell fraction, suggesting a correlation. Cluster 2 was then split again in to cluster 4 and a successive evolution of cluster 7 to cluster 30. While cluster 4 shows a similar pattern to cluster 6, the high variability of confidence intervals made assessment challenging. Cluster 4 and 6 could possibly be resolved into the same clone with deeper sequencing. Finally cluster 5 and 13, the progeny of cluster 30 allowed the perfect split of samples into small cell transformed and adenocarcinoma



Figures/CASCADE/CA86/CA86.ccf\_cluster.pdf

FIGURE 3.41: Cancer cell fraction of mutation clusters of clonal tree for patient CA-L; transparent polygons show the 95% confidence intervals. Clusters and cluster colours are taken from Figure 3.40

samples. Samples P.2, 17A, and 26, where the presence of cluster 13 could be observed were adenocarcinoma. However, the presence of cluster 5, which was only present in the small cell samples at autopsy suggested an incomplete micro-dissection of the progression sample (Figures 3.40 and 3.41). Surprisingly, there was no known genetic determinant found, which explained the split into EGFR T790M positive and small cell transformed lung cancer. In contrast, it is likely that the priming for transformation or remaining adenocarcinoma was epigenetic [210, 244].



### 3.4 Mitochondrial phylogenetic reconstruction - the power house of the phylogenies

While phylogenetic reconstruction is a well established method for genetic variants from canonical chromosomes to study metastatic progression and timing of evolutionary divergence [187, 199, 245], there are multiple issues. In Section 2.3.1 and Section 2.4.1 we showed how important the proper variant calling method is to accurately recover phylogenies and clonal patterns. In addition, using somatic variants to reconstruct phylogenies is a flawed concept to begin with.

Most models studying genetic variation assume neutral evolution of the DNA loci [246, 247], but cancers almost exclusively exhibit positive selection [248]. And while passenger mutations might not directly affect fitness of the cell, they only exist due to the link to the driver mutation and therefore provide little to no additional information gain in addition to the driver. Furthermore, while in small populations genetic drift as a stochastic process overpowers selective processes (fitness coefficient  $s$ ) and can therefore be assumed to be neutral, in larger populations  $N_e$  (effective population size) where Equation 3.2 does not hold true, mutations are under selective pressure [249].

$$N_e \cdot s \ll 1 \quad (3.2)$$

In summary, we can assume that with cancer cell growth, positive selection through treatment and tumour micro environmental niches, almost all assumptions of the coalescent theory [250] are not applicable for tumour samples and therefore methods using somatic variants and their respective results need to be selected and evaluated carefully.

To tackle this issue, and assist with the interpretation of phylogenetic reconstruction results, we adjusted a method used in single cell sequencing to track clonal expansion with mitochondrial somatic mutations [251] to be usable for standard bulk sequencing. Mitochondrial variants are an ideal source of clonality information, because the mutation rate is significantly higher than nuclear DNA, due to the missing proof reading and repair mechanisms, which allows very granular separation in a shorter time period. Additionally, while there are several diseases caused by defects in mitochondria such as Kearns-Sayre syndrome [252], MERRF [253] and MELAS [254], they usually follow a mendelian inheritance pattern and are hereditary and not somatically acquired. In the

cancer context, somatic mutations in mitochondrial DNA are assumed to be approximately neutral with a possible selection pressure towards healthy ageing and negative selection in cancer [255, 256].

### 3.4.1 Method

First a pileup of all mitochondrial positions was performed. Before the pileup we preselected reads which uniquely mapped to the mitochondrial genome and only retained high mapping quality reads. Then the nucleotide counts in each position were transformed into a MultiEssayExperiment [33] for final analysis in R. The preprocessing code can be found in Listing C.1.

The final MultiEssayExperiment is then read into R and quality metrics applied to exclude samples with not enough coverage on the mitochondrial contig. Our analysed WGS samples showed an extensive coverage of mitochondrial DNA, however WES library preparation procedures might restrict coverage. Patient CA-I had a coverage of more than 100x for all but the germline sample which only had an overall coverage of 17x. Similarly, patient CA-L showed lower depth for the germline sample (127x) but a generally high coverage for all tumour samples (mean: 543x, min: 138x). All other Patients (CA-A/J/K) where samples were sequenced with WGS exhibited a coverage of more than 200 even for low performing samples with a median depth of 67 916, 45 603, and 49 726 per sample (Figure 3.42).

Figures/CASCADE/mito/mtCoverage.pdf

FIGURE 3.42: Average coverage of mitochondrial DNA of CASCADE patients: Orange squares show germline sample for each patient; black points show tumour samples; horizontal red dotted line shows quality cut off suggested by Ludwig et al. [251]

This proved, that even without specifically enriching for mitochondrial DNA, most samples will contained enough tumour reads for this analysis.

To ensure optimal results, we excluded all samples with an average coverage of less than 50x. This means we removed the germline sample for patient CA-I, however as we expected the germline sample to be the ancestral state for all samples, this has virtually no effect on the reconstruction procedure. Additionally, we were more interested in the relationships between the tumour sample, which was still accessible even with the removed germline sample.

In contrast to the simple Hamming distance used for the presence-absence vector representation of canonical somatic variants (Section 3.3.1.6), for mitochondrial variants we employed an allele frequency (*vaf*) based distance (Equation 3.3) of two samples  $s_i$  and  $s_j$ . The difference in read support was normalised with the product of the total allelic depth *cov* and summed up at all sites of variation  $v$ .

$$mitoDist(s_i, s_j) = \sum_{v \in Variants} \left| \frac{vaf_{s_i}(v) \cdot cov_{s_i}(v) - vaf_{s_j}(v) \cdot cov_{s_j}(v)}{cov_{s_i}(v) \cdot cov_{s_j}(v)} \right| \quad (3.3)$$

This distance was only calculated for variant sites where both samples had at least a coverage of 100x to have a representative sampling of the allelic prevalence in each sample, as a human cell usually has more than 100 mitochondria [257].

### 3.4.2 Results

While the mitochondrial variants analysis only used a fraction of the size of the genomic DNA loci and therefore most likely violates the infinite sites assumption [258], it was still able to generate an orthogonal view of the heterogeneity and trajectory of the multi-regional samples in each patient.

#### 3.4.2.1 Patient CA-A

While the separation of progression (11, 47, 55, and 59) and stable (26, 31, 41, and 57) disease sites was already visible in the somatic phylogeny, the bottle neck of treatment

and new metastasis is more obvious in the mitochondrial phylogeny. However the individual resolution of splits appeared to be lower for the mitochondrial reconstruction, as seen in Figure 3.43.

Figures/CASCADE/mito/CA99SomVsMitoPhylo.pdf

FIGURE 3.43: Mitochondrial and somatic phylogenetic reconstruction of CA-A: Somatic variants based reconstruction (A) and mitochondrial variants based reconstruction (B)

#### 3.4.2.2 Patient CA-I

Neither the somatic variants nor the mitochondrial variants resolved the evolutionary trajectory in a granular fashion. The slightly longer stem of shared variants in the mitochondrial phylogeny was most likely due to the low coverage of the germline sample. Similar to all other patients, the substructure of the samples was changed. While using the somatic variants showed sample 566 as the closest to the germline sample, the mitochondrial variant phylogeny instead indicated sample 559 as the closest (Figures 3.42 and 3.44).

Figures/CASCADE/mito/CA51SomVsMitoPhylo.pdf

FIGURE 3.44: Mitochondrial and somatic phylogenetic reconstruction of CA-I: Somatic variants based reconstruction (A) and mitochondrial variants based reconstruction (B)

### 3.4.2.3 Patient CA-J

The mitochondrial reconstruction presented sample 2 as a member of the more genetically complex samples 20, 24, 32, and 42 in spite of the missing *TP53* mutation. Sample 28 on the other hand, which showed almost no evolutionary distance to the normal sample in the somatic analysis presented as a substantial outlier. This shows that the *TP53* mutation of samples 20, 24, 32, and 42 was likely acquired after the seeding of the distant sites like sample 2 in the adrenal gland. The difference in distance to the normal sample for sample 28 was likely due to a “cold” primary site of disease with little cell proliferation, which however still accumulated mitochondrial mutations [127] (Table 3.6, Figure 3.45).

Figures/CASCADE/mito/CA80SomVsMitoPhylo.pdf

FIGURE 3.45: Mitochondrial and somatic phylogenetic reconstruction of CA-J: Somatic variants based reconstruction (A) and mitochondrial variants based reconstruction (B)

### 3.4.2.4 Patient CA-K

In contrast to the somatic variant phylogeny, which showed an outgroup of samples 8 and 9, with a second cluster of samples 4, 5, and 6, the mitochondrial data supported a different split into two groups. These groups almost perfectly bifurcated the samples into those derived from the left and right sided disease sites, with sample 6 being the only sample from the right side clustered with the left lung and brain samples 8, 9, and 13. These data suggested that while only samples 8 and 9 showed a whole genome duplication and the *APC* “stop gained” mutation, they were more closely related to the other samples than assumed from the somatic variant analysis and probably were seeded by the same cells (Table 3.9, Figures 3.29 and 3.46).

Figures/CASCADE/mito/CA82SomVsMitoPhylo.pdf

FIGURE 3.46: Mitochondrial and somatic phylogenetic reconstruction of CA-K: Somatic variants based reconstruction (A) and mitochondrial variants based reconstruction (B)

### 3.4.2.5 Patient CA-L

While the somatic variants linked the small cell carcinoma samples P.1 and 8 together, the mitochondrial analysis showed that the closest relative to P.1 was P.2. As both of the progression samples were taken 14 months ahead of the death of the patient, this agreed with the clinical history of the samples better. Additionally, instead of grouping the adenocarcinoma sample 17A and 26 together, the mitochondrial phylogeny suggested, that while they share a common resistance mechanism (EGFR T790M), it might have been acquired in parallel instead of being seeded from the same lesion, as all samples other than the P.1/2 samples were not grouped together. Lastly, the closeness of sample 8 and the germline sample possibly indicated the presence of small cell disease already “before” the progression samples were collected. However, the FFPE conservation of the P samples could have altered the molecular clock and influenced the branching site on the tree (Figures 3.37 and 3.47).

Figures/CASCADE/mito/CA86SomVsMitoPhylo.pdf

FIGURE 3.47: Mitochondrial and somatic phylogenetic reconstruction of CA-L: Somatic variants based reconstruction (A) and mitochondrial variants based reconstruction (B)

### 3.4.3 Summary

With the analysis of the mitochondrial history of samples, we could shed some light on the timing of lesions and the development of resistance mechanisms, which is not heavily influenced by the treatment and its selection pressure. While the infinite sites hypothesis does not hold true for mitochondrial DNA, due to the limited sites and reduced repair mechanisms, the selection pressure of treatment and their resistance mechanisms parallel evolution bias in the analysis of multiple related tumour samples also violates multiple assumptions for phylogenetic reconstruction when using somatic variants.

Neither options come without pitfalls and caveats, but this method could offer an alternate view on the history and seeding time of lesions and their kinship using data that was previously discarded but was abundantly available. This approach is also available at virtually no additional cost, as mitochondrial variants can be readily detected from standard WGS and WES. Ideally both somatic variants and mitochondrial variants would be integrated into a holistic approach, however due to the substantial difference in scale between nuclear DNA and mitochondrial DNA, the process is intricate and outside the scope of this work.

## 3.5 Cohort level analysis

Even though there were only five lung cancer patients who have participated in the CASCADE program to date, each of the patients had a high number of samples analysed, revealing the complexity of intra-patient heterogeneity in NSCLC (Section 3.3). Despite this heterogeneity, there were several parallels between the patients showing similarities in disease trajectories. The process of small cell transformation (SCT) is still significantly under explored and understood, due to the rarity of the transformation as well as the lower overall survival in comparison to other resistance mechanisms. In the following section, the patients who developed small cell carcinoma transformation were compared and contrasted with the adenocarcinoma cases to further explore this mechanism of resistance.

The generally accepted hallmarks of SCT, apart from the histological changes such as high *MKI67* expression and down regulation of major histocompatibility complex I and II, are a much higher mortality, a high prevalence of *FHIT* and *MAD1L1* deletions or

loss, and *TP53* and *RB1* mutations [259, 260]. However, while in patient CA-L all samples showed a *TP53* “stop gained” mutation, patient CA-I’s transformation did not occur in the setting of *TP53* loss. Additionally, neither of the patients presented with a *RB1*, *FHIT*, or *MAD1L1* loss (Figures 3.13 and 3.37).

In agreement with recent literature showing whole genome doubling (WGD) for SCLC [261], we observed chromosomal arm amplification in patient CA-I and full WGD for patient CA-L. However, all NSCLCs patient also showed at least one sample with complete WGD, casting doubt on WGD being a distinguishing feature of SCT (Tables 3.2, 3.4, 3.6, 3.9 and 3.11). Additionally, the overall loss of heterozygosity could be seen in both NSCLC and SCLC and this seems to be a general feature of late stage lung cancers rather than NSCLC (Figure 3.48) suggesting that copy number alterations are not the main drivers of SCT.

Figures/CASCADE/LOH\_perChrom.pdf

FIGURE 3.48: Percentage of LOH per chromosome in CASCADE patients: Per chromosome violin plots are shown in grey with median as white dot; individual percentages per sample are displayed grouped with colour per patient; LOH was called with a major copy number  $<0.6$  and a major copy number  $>0.6$

The most prominent difference of NSCLC and SCLC in our patients was the reconstructed phylogeny. While the NSCLC showed substructure and meaningful evolutionary splits, the SCLC patients phylogenies showed a distinct “star shaped” pattern. Each sample branch was very close to the others and with substantial amounts of private variants in each sample (Figures 3.44 and 3.47 vs. Figures 3.43, 3.45 and 3.46). Even though the shared variants in CA-L seemed to provide the ability to transform, they do



not necessitate the transformation, as both samples CA-L 17A and 26 remained NSCLC with virtually no known genetic determinant of status. This in turn suggested either a currently undetected genetic determinant or potential epigenetic regulation to explain the SCT (Figure 3.37).

In contrast to CA-L, who presented with both NSCLC and SCLC, CA-I's samples were completely transformed. The biopsied adenocarcinoma, which already had an MHC-I disrupting mutation was completely out-competed by a secondary clone, which did not present with any additional genetic driver alterations. Similar to patient CA-L this suggested, that the genetic prerequisites for SCT were already present in the clonal population, but not sufficient to drive transformation (Figures 3.16 and 3.17).

Gene fusions or regulatory genetic variants leading to aberrant transcription could have been the cause for this phenomenon observed in both SCT cases. These could be detected in RNA sequencing of the biobanked fresh autopsy samples to exclude genetic causes which were not picked up with the performed WES, or detect transcription alterations. However, this analysis is outside the scope of this work.

## 3.6 Outlook

In this chapter we described both the high inter and intra patient heterogeneity of late stage lung cancer patients and showed that mitochondrial variant based phylogeny could help to resolve the sample relationships in the context of selection pressure through treatment. Additionally we uncovered a different disease trajectory for cases showing SCT where the cancers appeared genetically primed for SCT but genetic primed, but genetic alterations alone were not sufficient to drive transformation. This suggests that genetic analysis alone will not allow the prediction or early detection of SCT as a resistance mechanism. This uncoupling of genetic evolution and disease histology was also reported recently in a study focusing on multi-region analysis of treatment naïve SCLC cases [261].

While there exist several multi-region lung cancer studies up to date, their focus was on early stage and treatment naïve disease [3, 262]. While these studies showed ubiquitous intra-tumour heterogeneity and copy number alterations in early stage disease, there is an unmet need for the assessment of late stage lung cancers.

With this work we took a first step towards understanding and measuring the heterogeneity of tumour samples and treatment resistance mechanisms in late stage NSCLC, but many unanswered questions remain. The epigenetic marks and their inheritance patterns in cancer are a massive unexplored field which increases the heterogeneity of cancer even more [263]. Additionally, we could only hypothesise and reconstruct the longitudinal trajectory of the disease from autopsy samples. The next step to validate and further explore these findings would be to analyse temporally spaced samples from the same disease.

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

— Navali, Hatungo of the Karui

# MisMatchFinder - detection of mutational signatures from low coverage WGS

## 4.1 Introduction


Early researchers and physicians realised that cancers can have different morphologies and clinical progression depending on the primary occurrence of the tumour (Section 1.5), with the extensive sequencing of cancer specimens over the last two decades, the mutational signatures of cancers came into focus. These signatures are specific and characteristic combinations of mutations, which stem from distinct biological processes. These processes include exposure to DNA damaging agents like chemotherapy treatment, tobacco and UV radiation, and biological intrinsic pathway errors in DNA-replication or -repair. As each of those processes has a more or less distinct profile of mutations [264, 265] the analysis and deconvolution of the signatures contributing to a patients mutational landscape can help to diagnose and treat a patient. While many signatures occur at a background level and are related to “normal” cellular processes like ageing [233], others can point to defective mismatch repair or gain of function mutations in specific pathways, which can lead to new avenues of therapy for a patient [266].

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

### 4.1.1 Mutational signature analysis

Traditionally cancer mutational signature analysis entails a somatic variant calling process (Section 1.4.2) followed by a counting and deconstruction step, which assigns weights to the individual signatures. These signatures are a precompiled list of mutation count relations (Figure 4.1). While individual SNPs already contains valuable information, there is an improvement in signature granularity when also counting the base up and downstream of the nucleotide change. This expands the feature space of counts from the six base classes of SNPs (C>A, C>T, C>G, T>C, T>A, and T>G) to 96 unique

trinucleotide contexts [233]. While there technically are six more base changes and several more trinucleotide contexts combinatorially possible, they can be collapsed into the afore mentioned 96 by using the reverse complement of the change.



Figures/MisMatchFinder/SBS7aSignature.pdf

FIGURE 4.1: Trinucleotide count contributions for SBS signature 7a (UV exposure); values taken from Alexandrov et al. [227]

Additionally to the single base substitution (SBS) there exists doublet base substitution signatures (DBS) and InDel signatures for somatic mutations of cancers [227], which are all based on the same principle and enable a higher precision for stratification of similar cancer subtypes and DNA damaging agents.

#### 4.1.2 Restrictions and pitfalls of standard signature analysis

Especially for cancer samples, the focus when analysing mutational signatures, is on somatic variants of the sample. This requires deep sequencing of the tumour sample with at least WES or WGS. For optimal results, a matched germline sample for tumour-normal variant calling is required (Section 1.4.4). This means the cost of the assay is surprisingly high for the diffuse result of signature contributions of the variants in the sample. This is especially relevant when it comes to clinical diagnostic tools, where every biopsy of the patient is precious and not easily obtained. Additionally, the matched germline sample might not be available. For an analysis, which is based on the averaged and aggregated somatic variants, to require a high quality input could be seen as counter-intuitive. Especially, as the current gold standard analysis will report signatures, even if there are virtually no variants reported. We therefore developed a method which can be adapted for low coverage whole genome sequencing and requires no prior knowledge of the cancer or a germline sample.

### 4.1.3 Overview

This chapter describes a newly developed method, which allowed the detection of somatic signatures from low coverage WGS of cfDNA. This method, with further optimisation and validation, has the potential to provide a novel approach for non-invasive monitoring of patients and screening of at-risk individuals in a clinical setting.

## 4.2 Methods

With the change from a variant focused approach to a read based method, our method MisMatchFinder analysed “mismatches” of a read from the reference genome, rather than a genomic locus. This had the advantage of not requiring a matched normal and could theoretically be used for virtually any sequencing data source like TAS, WES, WGS or even RNA sequencing. However, it also meant, that the error suppression methods, which are usually used by variant calling methods like read position ranks sum (RPRS) or strand bias were not applicable. To solve the problem of high background noise we developed multiple filtering steps, which are presented in the following sections.

### 4.2.1 Mathematical concept

A mismatch in this work was considered as any position in an aligned read, which did not show the same base as the reference at the aligned position. The mismatch inherited all the metrics of the read such as mapping quality, base quality and read position.

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

$$n(mismatches) = n(somatic\ var.) + n(germline\ var.) + n(seq.\ error) \quad (4.1)$$

With the sequencing error being a function of the sequencing machine and chemistry, the error rate was assumed to be stable, almost constant, when using the same type of sequencing machine and chemistry [173, 174]. We therefore reduced Equation 4.1 to

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

Secondly, the number of germline variants was assumed to be approximately the same between two people [267], which again simplified Equation 4.2:

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

Of course, Equation 4.3 was a crude estimate and instead the constants exhibited variability and were not real constants. To better approximate the inherent variableness of sequencing error and number of germline variants, we instead used Gaussian distributions

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

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

$$n(mismatches) \sim n(som. var.) \quad (4.5)$$

With the help of Equation 4.5 we were theoretically able to approximate tumour mutational burden and signatures from individual reads with MisMatchFinder. The method therefore was independent of read depth and required no matched normal sample for somatic variant calling.

#### 4.2.2 Data preprocessing

As MisMatchFinder employs multiple internal measures to filter and process sequencing data, the steps used for preprocessing were minimal: The reads were aligned to the GRCh38 reference genome (Section 1.4.1). For optimal mapping and additional noise

reduction, paired end sequencing of at least 75 bp is required to ensure a few bases overlap on the standard fragment length of less than 155bp of ctDNA (Section 1.2).

All datasets in this work were sequenced with 100bp paired end, aligned with BWA to GRCh38 and optical duplicates were marked with Picard unless further specified.

### 4.2.3 Mismatch detection

In contrast to conventional variant calling approaches, which find regions of interest through pileups (position wise) and then realign reads in the surrounding area, to accurately estimate the most likely event that led to the observed haplotype (Section 1.4.2), with MisMatchFinder we evaluated every individual read pair as a separate entity to fully span the heterogeneity of all cells and their genetic background. The “MD“- and “CIGAR“-tag of sequencing reads from the preprocessed BAM file were used to reconstruct the sequence of the read and the positions, where the read showed a different base than the reference. These potential mismatch sites were then filtered in multiple steps to reduce the impact of both germline variants as well as sequencing errors, to ensure the assumptions of Section 4.2.1 held true.

### 4.2.4 Filtering steps

Apart from the standard filters, which most variant callers employ, like mapping quality (MQ) and base quality (BQ), which were used to ignore reads as well as read positions respectively, the method also internally filters out common sequencing errors next to homopolymer regions [268]. While we set default cutoffs, for optimal performance on our data (MQ=20, BQ=55, homopolyLength=5), the program allows the user to adjust them to their liking. This is also possible for both the region of interest (ROI) bed-file which was used to restrict the analysis to only highly mappable regions of the genome (Section D.1), as well as for multiple other parameters. These include minimum average base quality, minimum and maximum number of mismatches per read and/or fragment, and the minimum and maximum length of a fragment [269]. If any of these values were not within the specified range, the read was discarded from the analysis. No read flagged as secondary or duplicate was included in the analysis.

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

When paired end sequencing of ctDNA is analysed, the fraction of fragments where reads overlap is higher, than with “normal” tissue based sequencing, due to the shorter fragment length of ctDNA (Section 1.2). This allowed a fragment internal consensus calculation, by adjusting for differences between forward and reverse reads. In many variant calling methods, these differences are used by measuring the “strand bias” [270–272] or “strand balance probability” [118] by looking at a specific locus and evaluating the discrepancy of all forward and all reverse reads. As our method examined each read/fragment independently, the bias could not be calculated, however in the overlapping region of both reads, a consensus was generated. If both reads agreed on the mismatch, the BQ of both reads were summed to emphasise the increased evidence for this variant. In contrast, if they disagreed the base of the higher quality was used and its quality was decreased by half of the BQ of the lower quality base (Figure 4.2 bottom). To increase the stringency of the method, the user can also enable the ‘*-strictOverlap*’ option, which will only consider a mismatch, if both reads agree with each other. As we were only interested in mismatches from the reference, all positions where both agree with the reference were irrelevant for the analysis and were discarded (Figure 4.2 top). For the most stringent analysis, MisMatchFinder can additionally be configured to only use mismatches in the overlap part of a fragment (‘*-onlyOverlap*’), which significantly reduced the number of sequencing errors which were retained in the final analysis (Section 4.3.1.1).

This option however also reduced the available data by restricting the analysis to areas where reads were overlapping. Due to the fragment size distribution of ctDNA a paired end sequencing with 100bp read length will statistically in most cases lead to an overlap of at least 45 nucleotides (Section 1.2) and with 150bp read length most ctDNA fragments will be almost entirely covered by both reads. However due to soft-clipping and incomplete alignment, this number was lower in reality. In our tests, the restriction led to approximately 18 nucleotides (min: 14bp max: 25bp std.dev.: 1.45) being retained in the analysis out of 100. Nevertheless, this meant that with a read depth of 8-10x  $\approx 80\%$  of the genome was covered by the overlap of at least one read pair.



Figures/MisMatchFinder/ConsensusMethodMisMatchFinder.pdf

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

#### 4.2.6 Germline filtering - exclusion of normal variation

To further ensure the Section 4.2.1 assumption that the germline is a very small constant, we needed to remove as many mismatches as possible, which originated from germline variants. For this purpose, we built a custom zarr [58] based storage system from the gnomAD database (v.3.1) [125] using scikit-alel [273]. An in-depth explanation of the generation as well as a script for an end user can be found in Section D.3.

It allowed precise filtering of known germline variant sites from the analysis. The method enables the specification of an allele frequency cut off, but as a default all sites, which were detected in any sample in gnomAD were filtered. We even included sites with low quality variants in the database, as these were sites of sequencing or mapping complications, which most likely interfered with our analysis as well.

#### 4.2.7 Count normalisation

After the filtering steps, all remaining mismatches were aggregated to oligo-nucleotide counts. With this step we also classified directly neighbouring mismatches as DBS, which

were counted as separate entities. SBS and DBS both can be used to identify underlying biological mutational processes, but they have very different signatures associated with them [227].

The counts formed this way were influenced by the background frequency of their reference oligo-nucleotides in the analysed genomic region. As the frequencies of di- and tri-nucleotides are not uniformly distributed in the genome, the chance for a mismatch found in an “AAA” reference context is almost seven times higher than a mismatch in “CGC” (Table D.2, Table D.1). To reduce this bias towards high frequency oligo-nucleotides, we implemented a count normalisation step.

First the di- and tri-nucleotides in the analysed regions were counted using the reference without any black-listed and/or only in white-listed regions. These counts were then either (i) used to directly weight the observed mismatch counts, which led to a more uniform distribution of mismatches, or (ii) by building a fraction of observed oligo-nucleotides and the total counts in the genome (Table D.2, Table D.1), the weighting achieved an approximation of how the counts would have been distributed if the whole genome was analysed. These two options are available with ‘*-normaliseCounts*’ for the approximation to full genome. By also adding ‘*-flatNormalisation*’ only the observed counts are used for normalisation.

All analysis presented in this work was not normalised, as the white listed regions used showed no significant difference in oligo-nucleotide abundance.

#### 4.2.8 Signature deconvolution - finding the original signal

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

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

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

Equation 4.6 can then be written as

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

with the same restrictions as shown in Equation 4.7. These equations and the idea to solve them with quadratic programming (QP) have been taken from Lynch [274], and the iterative linear models (ILM) solving approach was adapted from deconstructSigs [37]. Both methods were reimplemented in python in MisMatchFinder, using the quadprog package [61] for QP and a translation of the R code of deconstructSigs for ILM.

MisMatchFinder allows the use of either QP or ILM, as they in many cases produce very similar results [274]. However, the default method is QP, even though ILM is the more interpretable and more parsimonious method, because of the increased number of signatures, in the latest work by Alexandrov et al. [227], ILM did not resolve the correct signatures if the signal was not strong enough. QP on the other hand showed more stable solutions (Figure 4.3).

Figures/MisMatchFinder/lowInputSignalDeconv.pdf

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

The combinatorial problem in ILM, already shown by Lynch [274], was exacerbated with “wide” signatures like SBS3 (Figure 4.1) and low signature contribution weights. As we were interested in analysing low tumour purity samples with low somatic signature signals from ctDNA, ILM was less sensitive in our test. Especially for SBS3, the contribution of the signature was only assigned by ILM with sufficient signal (15 and 20 mutations per megabase respectively for SBS7a and SBS3) where QP allowed for a more linear increase in signal, even at lower levels. On the other hand, ILM will assign an overall higher proportional weight than QP once the signal reaches a certain threshold (Figure 4.6). ILM was therefore better suited for high confidence signal, but less effective for the more subtle differences we expected from ctDNA.

The deconvolution method could be an area for further optimisation by creating a custom deconvolution system adjusted for ctDNA detection and the signatures present.

For the rest of this work, unless specified differently, the results shown were generated using the QP deconstruction method.

#### 4.2.9 Signature detection

Signature deconvolution with QP resulted in non-negative signature weights for almost all of the signatures when using MisMatchFinder derived counts, however a positive signature did not necessarily indicate the activity of this process in a tumour capacity due to the normal somatic mutation background and germline residual signal (Section 4.3.2.3.1). To enable calling of significantly active signatures in samples, we developed a z-score like system, which used the distribution of each signature weight in the healthy population as a background.

As the weight values after deconvolution were between 0 and 1 inclusive, with a high enrichment for 0 and 0-adjacent weights, we chose the beta distribution with probability density function (PDF, Equation 4.9) with shape parameters  $\alpha$  and  $\beta$  and normalisation constant B to ensure the cumulative density function (CDF) sums up to 1.

$$f(x; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad (4.9)$$

To enable a z-score like estimation, we calculated the  $\lambda$ -quantile of the cumulative density function of the beta distribution (Equation 4.10) for each patient  $p$  samples signature

weight  $w_s(p)$  of signature  $s$  with healthy sample fitted shape parameters  $\alpha_s$  and  $\beta_s$  per signature by solving for  $\lambda$  resulting in the inverse beta cumulative density function (Equation 4.11).

$$F(x; \alpha, \beta) = \frac{\int_0^x t^{\alpha-1} (1-t)^{\beta-1} dt}{B(\alpha, \beta)} \quad (4.10)$$

$$x = F^{-1}(\lambda; \alpha, \beta) = \{x : F(x; \alpha, \beta) = \lambda\} \quad (4.11)$$

This allowed us to estimate how many healthy samples would have a signature weight less than the patient sample for the respective signature  $s$  (Equation 4.12). While we did not see significant down regulation of signatures from the background, the method could support both over- and under-representation analysis. Ten percent of all samples were removed from both tails of the distribution of each signature before fitting the shape parameters to minimise the impact of outliers and instead fit the core distribution.

$$\forall s \in \text{Signatures}, \forall p \in \text{Patients} : \text{CDF-score}(s, p) := \frac{w_s(p)}{F^{-1}(\lambda; \alpha_s, \beta_s)} \quad (4.12)$$

We used  $\lambda = 0.99$  to calculate CDF-scores for all signatures and assumed a CDF-score of more than 2 to be significantly different.

This allowed the prioritization of significantly changed signatures in the tumour samples with regards to our healthy background depending on the desired application. A higher CDF-score cut off could increase specificity at the cost of sensitivity.

While most signature weights could be estimated very well with the moment matching estimation of `fitdistrplus` [275] some signatures did not show an ideal fit. However, in these cases, the fit resulted in a wider distribution with higher theoretical quantiles than empirically observed, which reduced the predictive power of the signature. These signatures also showed a double peak distribution indicative of a population sub structure in the healthy population (Figures D.5 and D.6 vs. Figures D.7 and D.8).

#### 4.2.10 Tumour detection

With the calling of “active” signatures in Section 4.2.9 we were subsequently interested in the classification of samples into “healthy” and perturbed (cancer) signature states. The classification as perturbed would allow a short-listing of samples for a more in-depth analysis with orthogonal validation.

For this purpose we trained a glmnet classifier with the assigned signature weights vector of each sample with an  $\alpha$  value of 0.7 due to the high correlation of the data.

The only other method which enables the classification of lcWGS of cfDNA into healthy and tumour (perturbed) samples was ichorCNA, which uses the copy number state of the sample to infer tumour purity [276]. ichorCNA is currently considered the gold standard and we used their default method as a reference to compare our results to. Any sample with a tumour purity of  $\geq 3\%$  was considered a positive and everything else a negative sample, as suggested by the ichorCNA developers.

The clinical status of disease was used as the ground truth for each sample. Every sample which was taken during a time, when the patient had active disease was assigned the tumour class and every healthy individual the healthy class.

### 4.3 Results

This section presents the results of applying MisMatchFinder to multiple distinct datasets with different configuration. Section 4.3.1 is the evaluation of the method on simulated data, which allowed accurate and definitive insight into the sensitivity of MisMatchFinder and served as proof of concept. Then, Section 4.3.2 summarises the results from multiple real world datasets, demonstrating the methods performance on real world data and the associated clinical insights.

#### 4.3.1 Simulated Data - the validation promised land

Just like in Chapter 2, the novelty of the approach led to the issue of an absence of a gold standard dataset, with which to evaluate the performance of our new method. While there are low coverage WGS datasets of cancer patients, none of them had validated

signatures associated with them. So we simulated data, to allow both optimisation of parameters of our method and granular detection of technical artefacts.

#### 4.3.1.1 Sequencing errors filtering

To judge the ability of our approach to filter out sequencing errors, we first simulated “clean” sequencing reads with neither germline or somatic variants with the ART simulation suite [184]. As current estimates of Illumina sequencing error rates were in the range of 1 in 666 to 1 in 1149 [174] which was significantly higher than even the highest tumour mutational burdens of cancers (melanoma: 1 in 5k; tobacco smoking lung cancer: 1 in 100k) it was very important to be able to eliminate as much of the background errors as possible.

Figures/MisMatchFinder/mismatchrateCleanSequencing.pdf

FIGURE 4.4: Mismatchrate of different filtering methods on sequencing data simulated with ART[184] for both 10x and 3x coverage; Mismatches correspond to simulated sequencing errors; all: no filters, overlap: only use the overlapping parts of paired end reads with consensus building (Section 4.2.5), strict OL: overlap but reads *must* agree, high BQ strict OL: strict OL with high BQ in both variants; A) Absolute counts B) counts from A normalised by the number of analysed bases all: all aligned bases, other: number of bases in read overlap

By only using high base quality mismatches, where both reads agree on the mismatch 99.98% of all sequencing errors could be eliminated and only 1 mismatch in 10M bases would be wrongly counted as a variant (Figure 4.4). This false discovery rate was multiple orders of magnitude lower than without consensus computation and the remaining error rate was lower than most tumour mutational burden estimates [227, 277]. We therefore considered our assumption of constant and small sequencing error contribution to be correct (Section 4.2.1).

#### 4.3.1.2 Spike-in signature detection

With the technical errors eliminated in simulated data, we used a similar method in real world data. However, to also establish a baseline for the detection limit and sensitivity of the method, we decided to first use a hybrid approach. We spiked somatic variants into genuine cfDNA low coverage WGS sequencing data of a healthy control, reducing the amount of unknown variability from other published datasets.

While it would have been possible to simulate the variants completely de novo, without any prior knowledge, we know that somatic mutations follow a certain pattern and there are mutational hotspots [81, 278], so we decided to instead use the COSMIC database [279, 280] as the catalogue to select mutations from. This allowed us to randomly draw mutations, which occurred in a specific cancer subtype. By using COSMIC variants our simulations were less synthetic. The in-depth protocol is shown in Section D.3.4. The downside of this method is that the spike-ins were not predominantly introduced on shorter fragments, as would be the case with real ctDNA.


The following section discusses the results for the simulation of the very distinct SBS7a UV signature (see Figure 4.1) which is predominantly present in Melanoma and secondly the much flatter and more uniform SBS3 (Figure D.1), which is a sign of defective homologous recombination in breast and other cancers.

Figures/MisMatchFinder/spikeInSanityCheck.pdf

FIGURE 4.5: Signature analysis results of spiked-in somatic variants; signatures with a weight less than 1% were collated into “unknown”; The original spike-in signature was coloured in green (SBS3) and purple (SBS7a), unrelated signatures are coloured white and signatures corresponding to sequencing artefacts are coloured in lightgrey; r0.1 corresponds to approximately 0.1 variants per mega base; Weights were generated with deconstructSigs [37]



The spike-in was done at multiple different ratios, to simulate varying tumour purities and tumour mutational burden (TMB). Figure 4.5 shows the signature analysis result of the lowest spike-in ratio “r0.1” which corresponded to 0.1 somatic variants per mega base and resulted in approximately 300 variants for the whole genome. As the spike-in process had to satisfy certain quality measures, not all candidate variants could be used. As such, the final simulated BAM contained 264 additional variants for the SBS3 simulation and 287 for the SBS7a equivalent. The variants corresponded to 304 and 364 “tumour” reads respectively within the  $\approx 261$  million reads of the simulated BAM. With increasing ratio, the spike-in signatures showed decreasing weights for other signatures, which likely got introduced due to the incomplete spike-in process (Section D.3.4).



Figures/MisMatchFinder/deconstructionMethodsDifferences.pdf

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

#### 4.3.1.2.1 Melanoma - UV exposure (SBS7a)

With melanoma, the previously reported normal TMB ranges from 0.1 to 100 mutations per mega base [227]. Melanoma is usually seen as a cancer with very high mutational load, which made it the ideal target for our new mutation based tool. With only the strict overlap (Section 4.2.5) and the germline (Section 4.3.1.3) filtering enabled, we could see that already from r5, which represented 16899 mutated reads (of 260 Mio.),

we could detect the UV signature SBS7a. While this signal would likely be too low to trust in a clinical setting, with r10, the signature weight was already 2% and well established. Additionally, the method was very specific on this dataset. Only SBS7a showed an increase with higher spike-in rate, with minor contributions from other other C>T heavy signatures like SBS2 and SBS30 (Figures 4.6 and 4.7), which partly already stemmed from the spike-in process, which was slightly enriched for SBS2 (Figure 4.5B “unknown”). All other signatures, which were present in the normal sample showed a decrease. This decrease was to accommodate an additional signature, as all signature weights need to sum up to 1.

Figures/MisMatchFinder/SBS7SpikeInSignatureDifferencesFocussed.pdf

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

#### 4.3.1.2.2 Defective homologous recombination-based DNA damage repair (SBS3)

Just as with the SBS7a signatures, even for the much more diffuse signature SBS3, MisMatchFinder specifically revealed the spike-in signature and did not assign the additional mismatches to any other signature. There was a small increase in SBS4 for the very low mutation rate simulations, where no SBS3 was detected. Unsurprisingly, the detection limit for SBS3 was slightly higher than for SBS7a (5 vs. 15 mutations per mega base), because of its more uniform profile. Exactly as with SBS7a, all other signatures showed a slight decrease, to accommodate the additional signature weight (Figures 4.6 and 4.8). While the detection threshold was slightly higher than the currently assumed median TMB in breast cancer, especially triple negative breast cancer (TNBC) has shown a higher TMB, which was at comparable levels to the limit of detection we saw in this simulated dataset [281].

Figures/MisMatchFinder/SBS3SpikeInSignatureDifferencesFocussed.pdf

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

#### 4.3.1.3 Germline filtering

In order to ensure our assumptions also hold true in real patient data, we evaluated the effect of removing germline variants from the analysis. We utilised the same simulated samples from Section 4.3.1.2, where the reads were original cfDNA sequencing reads from a healthy person. These reads had a known natural background germline variant profile as any arbitrary sample would.

Figures/MisMatchFinder/noGermlineFilterAnalysis.pdf

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

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

This extreme change was caused by the much higher number of mismatches which were used in the analysis ( $\approx 1.8$  Mio without germline filter and  $\approx 130$ K with germline filter). This increase in mismatches in the analysis diluted the spike-in variants. Figure 4.10 showed that without the germline filter the additional found mismatches never exceeded 5% of all mismatches, which seemed to be the detection threshold for SBS7a. With germline filtering this threshold corresponded perfectly with the increase of SBS7a weight in those samples (Figure 4.6).

Figures/MisMatchFinder/spikeInPercentage.pdf

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

While we had already established that the spike-in variants could not be detected when retaining germline variant sites, the computed signature weights in the normal sample were vastly different as well. Without the germline filter, the most prevalent signatures were SBS1 and SBS5 which are thought to be molecular clock like signatures, related to the age of the individual [227]. In the germline filtered analysis the most prevalent signatures were SBS4 (tobacco smoking), SBS12 (unknown) and SBS46 (sequencing artefact). In general it seemed like the germline filter removed predominantly SBS1 and SBS5, while most other signatures remained the same (Figure 4.11).

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

This convinced us that germline filtering, additionally to the consensus overlap analysis, was fundamentally important for the method to recover signal. In the following sections, unless further specified, the germline filter was enabled for all analysis.

Figures/MisMatchFinder/noGermlineFilterSignaturesPieChart.pdf

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

### 4.3.2 Real world data - analysis of patient data

While simulated data is perfect to ensure the method performs as expected in edge cases and to estimate detection limits, only real world data allows the final examination to understand if the model used for analysis can mirror biological concepts. To show our new method is usable for a variety of datasets, we used a mixture of different cancer types with different library preparation. In Section 4.3.2.2 we focused on the analysis of 60 healthy individuals. We generated a background noise model and excluded aging as one of the sources of variation from our data. Then we analysed two metastatic breast cancer patients with BRCA1 positive disease, comparing matched tumour-normal sequencing with MisMatchFinder. This dataset allowed us to evaluate how efficient germline filtering was (Section 4.3.2.3.1) and how accurate and sensitive our method was when compared to the current gold standard of tumour-normal tissue analysis (Section 4.3.2.3.2). A second dataset containing samples of two cfDNA time points and corresponding tissue from a patient with metastatic melanoma allowed us to validate performance of MisMatchFinder in a different cancer context (Section 4.3.2.4). Lastly we analysed a dataset of 79 tumour only cfDNA samples both melanoma (40) and breast cancer (39) patients to show the potential clinical application of MisMatchFinder (Section 4.3.3). The mean tumour purity assigned to each sample by ichorCNA was 16.2% (min: 0.3% max: 71.5% sd: 17.0%) displaying a clinically expected range of high and low ctDNA fractions. The mean age of patients was 53.4 (min: 34 max: 74) for the breast cancer patients and 60.2 (min: 34 max: 81) for the melanoma patients which is very closely age matched to our healthy samples (Section 4.3.2.2).

In the following section, we showed that MisMatchFinder exhibits barely any technical bias.

#### 4.3.2.1 Bias detection

A dataset with healthy samples is key to detect biases, because any variability that cannot be accounted to either age or gender is unwanted and will affect the cancer samples in the same way. We expected an increased mismatch rate in the older individuals due to the accumulation of somatic mutations due to “clock like signatures” [127]. In contrast, tumour samples should be biased based on tumour purity, as higher amount of tumour reads would result in a higher amount of mismatches from somatic mutations. To verify that our assumptions were correct, we performed a principle component analysis (PCA) of the raw tri-nucleotide mismatch count numbers, of all 79 tumour only and 60 healthy samples, which MisMatchFinder can report alongside the weights of signatures.

Neither age, nor sex of the sample seemed to have any influence on the mismatches of the sample Figure 4.12A, B. In contrast, there appeared to be a batch effect with regards to the used flowcell on the sequencer and library preparation (Figure 4.12C, D). As these two are strongly intertwined, it was not possible to differentiate the two effects, however the flowcell bias was evident across multiple library preparations and therefore more likely to be the source of the bias. Flowcell 1 contained samples from both ‘g’ and ‘h’ and flowcell 3 both ‘a’ and ‘d’, suggesting that the flowcell had more influence than the preparation. This is consistent with recent literature, which suggests, that there are more and less error prone flow cells [174].

While there was a slight bias towards higher PC1 values for higher DNA input samples, which might be due to a higher library complexity when sequencing, it was minor and the same bias exists for every other method using de-duplicated sequencing data as it might have an effect on the non-redundant data available (Figure 4.12E). Similarly, the very low coverage samples were at the very left of PC1, but there was a substantial spread along the axis for the higher coverage samples as well (Figure 4.12F).

As expected from the model and the biology, there was a trend of separation for both tumour type and tumour purity, with the higher purity samples oriented toward the top right and the healthy and lower purity samples at the left (Figure 4.12G, H).

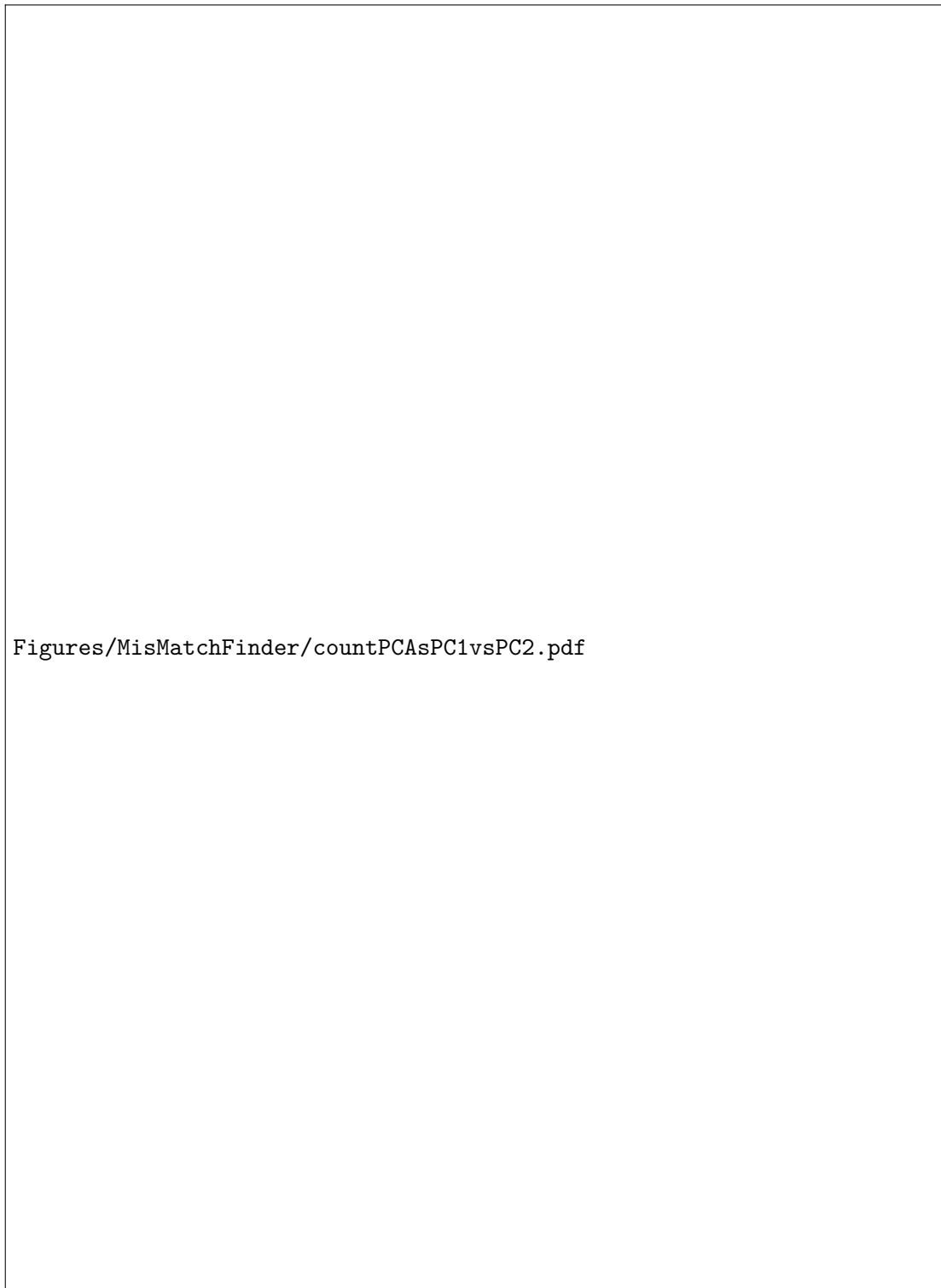


FIGURE 4.12: PCA (PC1 and PC2) of tri-nucleotide mismatch counts of healthy donor and tumour samples (melanoma and metastatic breast cancer) of varying purity; PCA was conducted on scaled and centered data

These tendencies can also be observed when looking at PC2 and PC3 (Figure D.4). PC3 still accounts for  $\approx 9\%$  of the observed variability and PC6 is the first component with an variance value smaller than 1 suggesting that only the first five components should be retained for further analysis (explaining a cumulative 97.3% of the variance). However, as we use the counts for signature deconstruction the PCA analysis only served as a quality control.

#### 4.3.2.2 Healthy cohort

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

With recent literature indicating a linear relationship of aging and somatic mutations [82, 127, 282], we were interested if the MisMatchFinder method could identify the accumulation of somatic mutations with age. While the samples between 30 and 59 showed the expected linear increase in mean mismatch rate, both the 20-29 and 60-69 year old samples showed the inverse trend, where young individuals exhibited a higher mismatch rate than older. This discrepancy could have been rooted in the germline filtering step as shown in Section 4.3.2.3.1, or it may have been a sign of loss of heterogeneity in the hematopoiesis as shown in aging mice [283]. Lastly, with only 40 healthy samples, we might not have had enough representation in each age group to detect subtle changes. Whatever the source, MisMatchFinder was not able to infer the age of a sample with the default settings in this dataset.

##### 4.3.2.2.1 Black list generation

With the strong effect filtering of both technical errors (Section 4.3.1.1) and germline variants (Section 4.3.1.3) as background noise had on our method, we hypothesised that a blacklist of mismatches found in our healthy individuals would help us further cut down on unwanted background signal and refine the somatic mismatch calls. We therefore ran MisMatchFinder with significantly relaxed quality cut-offs to capture as much variation as possible. This included a reduction in mapping quality and base quality as well as



Figures/MisMatchFinder/MisMatchRateByAge.pdf

FIGURE 4.13: Mismatchrates of healthy samples by age: distribution of sample is shown as violin plots on the left hand of each group, with a boxplot and the mean indicated as a black box and a white dot respectively; Observed values were displayed as black dots on the right hand of each group

not restricting the analysis to the highly accurate overlap part of the paired end reads. However we still restricted the analysis to the same highly mappable areas of the genome the same as for the tumour analysis as well as filtering already known germline variants for a better estimation of the impact of this filter step.

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

Figures/MisMatchFinder/mmfbBlackListStats.pdf

FIGURE 4.14: Blacklisted mismatches from healthy individuals: A) Histogram of frequency of mismatches found by at least n samples (multiplicity) B) Percentage of mismatches in healthy samples found in blacklist of mismatches found in at least three healthy samples; violin plot with median as white dot (left) and real values (right)

While there was a substantial number of shared variants in the reduced parameter analysis out of the  $\approx 15$  million mismatches, only 3413 were in more than 50% of all samples (Figure 4.14A), and the blacklist filtering did not lead to a performance boost. When we used the default settings for the healthy samples and used the generated blacklist as a filter on average only 2% of mismatches were removed by the blacklist

assignment (min: 1%, max: 3%) when at least three samples showed the mismatch (Figure 4.14B). Utilising the ten sample blacklist, the mean filtering amount dropped to 0.1% of mismatches of the sample (min: 0.04% max: 0.3%). We therefore concluded that the blacklist filtering step has no additional benefit and it was not included in the default analysis of MisMatchFinder.

#### 4.3.2.3 *BRCA1* mutation positive breast cancer patient samples

The first dataset of cancer patients involved two *BRCA1* mutation positive females. The data contained matched tumour, germline and cfDNA sequencing with high depth ( $\approx 80x$ ) WGS for all samples. With the matched normal, we used the current standard protocol of somatic mutational pattern analysis (Section 4.1.1) and compared it with our new method (Section 4.3.2.3.2).

As the sequencing data had a much higher depth than what is used in standard clinical practice for plasma sequencing, we down-sampled the data to 8x coverage for MismatchFinder analysis, bringing it in line with the simulated data. By using several different seeds for the sampling, we generated pseudo technical replicates of the sequencing (Section D.3.6), which in turn gave an approximation of the stability of the results of MisMatchFinder.

##### 4.3.2.3.1 Germline artifacts

As discussed above, the germline filter step is vital to boost the signal of somatic variants (Section 4.2.6). We were interested in how many germline variants were not filtered out with our filtering step and were still contributing to background noise. The high depth matched healthy WGS samples of the breast cancer patients was used for this analysis. We called germline variants on the matched normal using Strelka2 and compared the called variants with the sites reported by MisMatchFinder as somatic (retained after germline filtering) on the sub-sampled data. All variants with any quality filter assigned by Strelka2 were considered for this analysis, such that possible clonal hematopoiesis (CH) variants were still considered. Table 4.1 showed that on average 2100 germline variants were not filtered out per run. However, this only equated to 0.9% for patient 1 (MBCB196) and 1.5% for patient 2 (MBCB298) of all sites found to be mutated. While the exact numbers of any arbitrary sample will depend on the strictness of the parameters

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

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

of the analysis as well as the mutation rate of the sample, with default parameters a similar result should be expected with other samples, suggesting the germline filter was removing virtually all mismatches caused by germline variants.

The germline variant removal was therefore very effective filtering the 3.75 (MBCB196) and 3.76 (MBCB298) million germline SNPs called by Strelka2 to less than 0.05% of the original number. As the genetic background in gnomAD is not balanced and shows a lack of non-european ancestry data [159], this filtering could become less effective when analysing samples from indigenous or otherwise genetically less characterised patients, as their germline variants might not be comprehensively represented.

Nevertheless, this result combined with the effective filtering of technical errors (Section 4.3.1.1) suggested, that nearly all of the remaining sites were somatic mutations of either the healthy tissue or the cancer cells.

Additionally, we were interested to see if the filtering did remove true somatic signal, by removing somatic variants which mirrored a known germline variant. We therefore

annotated high confidence variants called with Strelka2 in both the patient tumour tissue and cfDNA samples with a germline flag when that variant was previously found in gnomAD. Surprisingly we found that  $\approx 50\%$  of somatic variants found in any sample were also found in gnomAD. To ensure these samples were not outliers, we also analysed the tissue samples from Chapter 3 and saw that all samples had half of the somatic variants called already present in gnomAD (Figure 4.15A). While gnomAD is known to contain potential somatic contamination, these overlaps were most likely due to error prone sites in the genome and the higher rate of deaminations and transitions over transversions [284]. To ensure our filtering did not skew the results we performed signature deconvolution on the somatic variants of a sample, which were not found in gnomAD and variants which were found in gnomAD. While there were minor signatures, which were clinically relevant (SBS13 and SBS87) in the “germline” partition, the majority of signatures were identified using variants not contained in gnomAD. Specifically SBS13’s contribution was larger in the non-germline selection even though it was found in both analyses (Figure 4.15B vs C, Figure D.9).

Figures/MisMatchFinder/somaticVarsInGermlineSites.pdf

FIGURE 4.15: Analysis of somatic variants with respect to germline database gnomAD: A) percentage of somatic variants called with Strelka2 found also in gnomAD; Violin plot with median as white dot (left) and values (right) B) signature analysis of variants not found in gnomAD for sample MBCB196 cfDNA c) signature analysis of variants also found in gnomAD for sample MBCB196 cfDNA; Colours show cancer associated signatures: blue (APOBEC), red (UV exposure), orange (tobacco), purple (chemotherapy), light grey (sequencing artefacts), dark grey (everything below 1% weight)

In summary, the germline filtering was a very powerful step, which boosted the performance of our signature detection method significantly with minimal signal deterioration. The high overlap of germline database variants and somatic variants poses a challenge for perfect concordance between our method and the current gold standard and would be an ideal area for further optimisation especially if ageing signatures like SBS1 and SBS5, which are caused by deamination, are of interest. These variants will be predominantly

removed and were the most likely reason for the absence of ageing signatures seen on healthy individuals (Section 4.3.2.1). A potential solution would be to incorporate a machine learning model to distinguish germline from somatic variants [285, 286].

#### 4.3.2.3.2 Matched tissue and cfDNA WGS samples

Figures/MisMatchFinder/mbcbWGSsignatures.pdf

FIGURE 4.16: Signature weights for the WGS of two BRCA1 mutation positive breast cancer patients: Colours show cancer associated signatures: blue (APOBEC), red (UV exposure), orange (tobacco), purple (chemotherapy), light grey (sequencing artefacts), dark grey (everything below 1% weight)

The matched tissue-normal WGS for the two metastatic breast cancer patients allowed a concordance analysis with a known true signature. We called variants with Strelka2 and performed default signature deconvolution with sigminer to obtain weights using default parameters for the GRCh38 genome build and QP deconstruction method [287]. Due to their known germline BRCA1 mutations we expected a contribution of SBS3. Both the tissue and the ctDNA samples of patient MBCB196 show a greater than 3% SBS3 as expected, however neither samples for MBCB298 showed SBS3 (Figure 4.16).

When applying the default parameter MisMatchFinder analysis to the downsampled WGS, we observed consistent results with any seed, suggesting that the signal was present in all reads and subsampling did not skew the results. Secondly, the expected SBS3 signature was found at  $\approx 18\%$  (min: 14% max: 21%) in the tissue and  $\approx 24\%$  (min: 20% max: 26%) in the cfDNA, recapitulating the results from the standard analysis (Section 4.2.9). The chemotherapy related signatures SBS25, SBS31, and SBS87 seen in both MBCB196 and MBCB298, were not reported by MisMatchFinder. This could have been caused by the germline filtering step, however these signatures were not clinically

actionable and expected after chemotherapy treatment so their exclusion was not of major concern.

Surprisingly, even though the APOBEC signatures SBS2 and SBS13 were observed in the normal signature deconvolution of patient 1 and 2, MisMatchFinder only found very low levels of SBS13 contribution in the cfDNA sample (mean: 0.4% (min: 0.1% max: 0.7%)) and no contribution in the tissue samples. However, MisMatchFinder was able to detect a high prevalence of SBS2 in both tissue (mean: 12% min: 11% max: 12%) and cfDNA (mean: 69% min: 65% max 74%) samples of MBCB298, indicating that MisMatchFinder was capable of detecting APOBEC signatures.

Importantly, even though the WGS analysis with sigminer did not result in a SBS3 positive deconvolution, MisMatchFinder was able to recover the expected signature in the tissue samples (mean: 28%, min: 27%, max: 29%). As the tissue sample in this case was a FFPE sample, with a much higher number of somatic variants called and very low tumour purity ( $\leq 13\%$ ) the standard analysis could have been overwhelmed with FFPE artefacts, which were successfully filtered out using MisMatchFinder. Interestingly, the cfDNA sample did not show any SBS3, but extensively high APOBEC activity. The absence of SBS3 could potentially be explained by reversion mutations in the cancer reenabling BRCA1/2 [288], while the high APOBEC activity has been shown in some

Figures/MisMatchFinder/brca1BarPlots.pdf

FIGURE 4.17: Signature weights for subsampled BRCA1 positive patients: each quadrant represents one high depth WGS sample downsampled to 8x with different seed (x-axis) signature weights per downsampling were shown in the columns in each quadrant. Colours represent clinically relevant signatures: blue (APOBEC), green (HRD), red (UV radiation); light grey and white show sequencing artefact and signatures of unknown significance respectively. Only signatures considered to be active (Section 4.2.9) were displayed

breast cancer types with Kataegis [227, 289]. Both the tissue samples of MBCB196 and the cfDNA sample of MBCB298 showed activity of UV signatures SBS38 and SBS7b respectively. While it is unlikely that these were caused by UV exposure, the APOBEC deamination signature has similarities to UV exposure. We therefore thought the UV signatures may have represented a leaky APOBEC signature rather than genuine UV exposure.

While MisMatchFinder did not detect the expected SBS3 signature in all samples, we could show a high concordance with the current gold standard analysis and MisMatchFinder revealed clinically relevant signatures SBS2 and SBS3.

#### 4.3.2.4 Melanoma patient samples

For melanoma real world data, we analysed samples from a single melanoma patient containing tumour-normal matched tissue WES with two additional longitudinal time points of cfDNA sequenced both through WES and lcWGS. This allowed us to compare the current standard signature analysis for both time points with our novel low coverage method that MisMatchFinder was developed for.

With the help of the WES of both the tissue and the cfDNA samples, we established the gold standard to compare the MisMatchFinder result to. Strelka2 was used to call somatic variants and the high confidence somatic SNPs were used to generate signature weights with sigminer.

As expected from a skin melanoma, the tissue as well as both WES cfDNA samples revealed a high contribution of SBS7a and SBS7b, with approximately 50% of all somatic variants associated with these two signatures in the tissue and time point 1 and 27% in time point 2. While several other signatures were found to be present in all samples, none of them had clinical relevance (Figure 4.18).

This high exposure of SBS7a and SBS7b was highly concordant with our analysis of the lcWGS sample of time point 1, where more than 50% of all somatic mismatches signatures, which were called active (Section 4.2.9), were attributed to SBS7a and SBS7b. While the proportion of SBS7a to SBS7b was different between the WES result and the lcWGS, the weights for those signatures were very similar. Time point 2 on the other hand showed less agreement between WES (sigminer) and lcWGS (MisMatchFinder).

Figures/MisMatchFinder/melanomaWESsignatures.pdf

FIGURE 4.18: Signature weights for the WES of two melanoma patients; First column shows the results for the tissue baseline and middle and right column show the cfDNA; Colours show cancer associated signatures: blue (APOBEC), red (UV exposure), orange (tobacco), purple (chemotherapy), light grey (sequencing artefacts), dark grey (everything below 1% weight)

While SBS7b was still detected at similar levels between both methods, MisMatchFinder failed to detect SBS7a, which was the highest contributing signature at time point 2 using the somatic variants from WES. Additionally both lcWGS samples show a high prevalence of SBS3, which was not detected in the WES. While SBS3 is usually accredited to *BRCA1/2* positive breast cancers, there have been reports of homology repair deficiency in melanomas. These mismatches might have been caused by subclonal variants, which are below detection threshold for conventional variant calling (Figures 4.18 and 4.19).

Figures/MisMatchFinder/melanomaMMFsignatures.pdf

FIGURE 4.19: Signature weights of lcWGS of two melanoma samples Colours show cancer associated signatures: blue (APOBEC), red (UV exposure), orange (tobacco), purple (chemotherapy), light grey (sequencing artefacts), dark grey (everything below 1% weight)

### 4.3.3 Tumour detection analysis

The comparison of our glmnet based analysis of MisMatchFinder signatures showed similar results to ichorCNA, which is the most commonly used method to call tumour presence (i.e. ctDNA detection) from low coverage sequencing of cfDNA. The two methods classified 21 and 23 true positives and 55 and 60 true negatives which resulted in a



sensitivity of 0.525 for MisMatchFinder and 0.575 for ichorCNA and a specificity of 0.91 for MisMatchFinder and 1 for ichorCNA on the melanoma samples (Tables 4.2 and 4.3).

TABLE 4.2: Confusion matrix for MisMatchFinder leave one out validation on melanoma training set

		True class	
		positive	negative
Prediction	positive	21	5
	negative	19	55

TABLE 4.3: Confusion matrix for ichorCNA leave one out validation on melanoma trainings set

		True class	
		positive	negative
Prediction	positive	23	0
	negative	17	60

While MisMatchFinder was inferior to ichorCNA, there were 5 melanoma samples, which showed very low tumour purity with ichorCNA (mean: 0.6% min: 0.4% max: 1.1%) but MisMatchFinder correctly identified them as cancerous samples. As ichorCNA is purely based on copy number alterations, these samples most likely only had mutationally driven disease, which is why MisMatchFinder could detect altered mutational signatures. Interestingly, only one of the patients exhibited melanoma associated signature SBS7c, while all others had a very high contribution of SBS4 and SBS29, which are thought to play a role in tobacco related cancers.

This suggested to us, that while MisMatchFinder is able to detect the right signature if it is present (Section 4.3.1.2.1) the detection of ctDNA in a sample is not purely dependant on the presence of a single dominant signature, and rather is influenced by the combination of signatures detected in the sample. We therefore restricted the detection of the melanoma samples to only melanoma associated signatures SBS7a through SBS7d and SBS38, to validate our hypothesis. With only the melanoma related signatures, only 2 samples were classified as tumour positive, which showed, that the interplay of signatures contains additional data and potentially more signatures than previously thought are related to cancer.

TABLE 4.4: Confusion matrix for MisMatchFinder leave one out validation on breast cancer training set

		True class	
		positive	negative
Prediction	positive	1	0
	negative	38	60

Finally, when comparing our method to ichorCNA on the breast cancer samples, ichorCNA's performance was superior to the melanoma samples Table 4.5, but MisMatchFinder performed considerably worse Table 4.4. The increased sensitivity of ichorCNA (0.74 vs 0.575) on the breast cancer samples was expected, as breast cancers are known to accumulate specific copy number alterations during their evolution [290] and these changes have been shown to be able to stratify and diagnose breast cancer samples [291, 292]. MisMatchFinder on the other hand requires a mutationally driven cancer signal, which is considerably weaker in breast cancers [227].

TABLE 4.5: Confusion matrix for ichorCNA leave one out validation on breast cancer trainings set

		True class	
		positive	negative
Prediction	positive	29	0
	negative	10	60

However, when using information gain feature selection on the signature weights reported for the breast cancer samples by MisMatchFinder, only signatures SBS3 and SBS5 showed a reduced entropy. We therefore used only these signatures in the training and subsequently boosted the sensitivity from 2% to 18%. Finally, by just using a cut-off approach for weights in SBS3, we achieved the highest sensitivity (0.36%) with a minute drop in specificity (0.95%). We used twice the standard deviation added to the mean of SBS3 weights observed in healthy samples as a cut-off (Figure 4.20).

With samples presenting with a SBS3 signature weight above 0.003 considered positive, we found 12 out of 39 (30%) breast cancer samples and only misclassified 3 healthy samples as cancer positive. With only BRCA1/2 mutation positive cancers having a specific signature we only expected to be able to detect a subset of breast cancer patients. With homologous recombination deficiency (SBS3) being reported in up to to 40% of

Figures/MisMatchFinder/SBS3Distributions.pdf

FIGURE 4.20: SBS3 signature weight distribution in healthy and breast cancer samples: MisMatchFinder reported signature weights are shown as both a violinplot with boxplot in black and mean shown as white point on the left and the real measurements on the right; red line depicts classification cut-off twice at 0.003

breast cancers in previous studies [293], our lower sensitivity for breast cancer could be caused by the lack of mutational signatures in the other 60% of cases.

## 4.4 Summary

In this chapter we presented a new method to detect signatures from low coverage whole genome sequencing data as a complementary method to ichorCNA to detect and classify patient samples. Simulated data showed a high specificity of our method for deriving signatures. Real patient data confirmed that while there is a discrepancy between the current standard method of signature detection and MisMatchFinder, the most prevalent clinically relevant signatures were detected. Furthermore, MisMatchFinder was able to detect tumour presence in samples without additional input or need for matched normals.

As the concept has so far only been explored for SNVs, the method can be extended to also analyse DBS and even InDel signatures, which could potentially have an impact on the accuracy and sensitivity of the method. And with more and more insight into the biological features of ctDNA additional enrichment steps, like fragment size selection, could be included to boost the signal of low tumour purity samples [93, 94].

With some improvements and optimisations, MisMatchFinder could be seamlessly integrated into the ctDNA analysis workflows involving low coverage WGS to give insight into the mutational processes and resistance mechanisms [294, 295].

*“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*

## Conclusion

This thesis explored different computational methods to assess the genetic heterogeneity of multiple patient samples using DNA sequencing. With sequencing costs now trending towards \$100 per genome, many clinical studies are now accumulating data at an unprecedented rate [296], but computational methods have not kept up with the development. With multiple spatial and longitudinal samples sequenced, the known concept of spatial and tissue heterogeneity could be assessed and insights into disease trajectory and evolution generated. In the past, the accurate molecular diagnosis of a patient’s cancer led to the discovery of targeted therapies and a massive improvement in cancer care. So it is likely, that similar advances can be made with the accurate analysis of the diversity and history of cancer clones within a patient. While single cell sequencing has already highlighted and described new cell states and types, the methods are far from being able to be used in a clinical situation. Additionally, the amount of data already generated for collaborative efforts like TCGA [157] or PCAWG are cause enough to optimise and develop methods to facilitate further research in the cancer space.

The contributions of this work include the development of multiple proof of principle methods, which show that the analysis of bulk sequencing requires further research and has unrealised potential for both diagnostic and research questions.

This thesis presents three distinct but related projects, which explore the analysis of tumour heterogeneity at different levels and depth, with a focus on method development.

In Chapter 2 we presented the work we conducted to improve the detection of somatic variants at very low allele frequencies. When multiple samples, separated in time or space, of the same patient were available, we were able to improve the detection threshold of variants substantially. These low abundance variants are invaluable in a clinical setting, where they can indicate an arising resistance mechanism, or relapse of disease. With the improved sensitivity of our method, treatment of patients can be adjusted earlier and more accurately. With the increase in multi-region analysis in cancer research, several research projects are already using the joint somatic variant calling approach

developed. In the future these concepts could be adjusted for the use in single cell sequencing and to employ evolutionary modelling to allow usage of priors for the spatial or temporal distance of samples which are analysed jointly.

Chapter 3 illustrates the in-depth analysis of resistance mechanisms and evolutionary history of five lung cancer patients enrolled in the CASCADE autopsy program. Using the joint somatic variant calling from Chapter 2 as a basis, we explored various ways to describe and visualise the disease trajectory of each patient from diagnosis till death. Additionally to the variants, we used copy number analysis and structural variants to contrast and compare each sample within a patient to generate phylogenies to visualise the evolutionary distances and to generate a pseudo time scale for the timing of mutations. In order to further clarify the grouping and distances of samples, we implemented a distance measure based on mitochondrial variants. With this method we could assess the effect of selection pressure on established variants and their timing in the nuclear variant derived phylogenies. This work has already lead to two publications, describing two resistance mechanisms: a novel resistance mechanism to a RET-fusion targeted treatment in patient CA-A [6], and the mechanism of small cell transformation in patient CA-L [5]. Across the cohort with unsurprisingly found few unifying features apart from loss of heterozygosity and large scale genomic amplification. Clear genomic determinants of treatment resistance were identified for the three NSCLC cases which did not show phenotypic switching. The diversity of these genomic mechanisms were profoundly highlighting the true extend of inter patient heterogeneity. In contrast to the NSCLC cases, the two cases with small cell transformation showed distinct evolutionary trajectories, with similarity in their phylogenies, both nuclear and mitochondrial, suggesting initial shared evolution with high private mutations and no clear genetic discriminant for the small cell transformation. Additionally, the small cell transformations did not exhibit the previously hypothesised genomic hallmarks of TP53 and RB1 loss. These findings moved to focus on the importance of characterising non-genomic evolution in parallel to genomic changes in the study of treatment resistance. To fully explore the heterogeneity of disease in these patients and generate a complete landscape, further RNA and epigenetic profiling will be necessary to shed light on the mechanism leading to small cell phenotypic switching.

With Chapter 4 we explored the avenue of measuring tumour evolution over time through the use of ctDNA, as it is often not feasible to serially biopsy a patient during treatment.

We tailored a method to be fully tumour agnostic which can be easily be applied to the clinical setting by using low coverage whole genome sequencing, which has already been used for diagnostics in some cancers. The method uses highly specialised filtering steps to eliminate the background noise from the “normal contamination“ and sequencing errors in these data. We showed that the method can accurately detect specific cancer related signatures at low tumour purity and tumour burden in simulated and patient data for melanoma and breast cancer. MisMatchFinder was able to detect signatures, which were not causing clonal fixation, but rather were present in all clones at different levels. This same effect was also observed when we built a machine learning model to predict the presence or absence of cancer in a plasma sample. Tumour samples were accurately classified without their commonly associated mutational signatures and instead showed substantially perturbed mutational patterns. The predictive ability of MisMatchFinder was already comparable to the current gold standard ichorCNA, and there are multiple possible ways to improve MisMatchFinder further. The restriction of the analysis to fragments of a size which enriches for tumour signal showed substantial performance improvements in our data and recent literature suggests a similar effect for fragment start sites and sequence. These additional filters should increase sensitivity and help reduce the amount of false positive classifications in our predictions. Additionally, MisMatchFinder and predictor could be extended to use double base substitutions and InDel signatures, which are known to be very specific for certain cancer types. Lastly, the classifier could be retrained to classify cancer type in addition to cancer presence with additional data from public datasets.

With the introduction of further sequencing platforms and technologies apart from Illumina [297–299], the ability to generate high quality multi-region/-sample sequencing datasets is more accessible than ever before. These new datasets will require ongoing development of computational tools to analyse the amount of data, learning from the methods already deployed in other fields like population genetics. Ever growing cancer cohort studies will require substantial optimisation and development of methods as can be seen with the rapid development of single cell analysis.

These methods will also require investment of time and money to generate gold standard datasets to further the improvement of algorithms with objective performance measures. As discussed in this thesis, while simulated datasets are ideal to test certain aspects of methods, approaches like “Genome in a Bottle“ are necessary for somatic variant calling

methods to continue improving in the same way that germline variant calling field has evolved [300].

While somatic variant calling in cancer often has the flair of being “solved” every improvement in the field leads to further understanding of concepts and biological processes. In particular, the field of multi-sample variant calling analysis is still very young and requires new guidelines and best practices to evolve, as it challenges current knowledge and hypothesis. The contribution of this work to the field was focused on methods to maximise insight into cancer biology in a multitude of cases. We showed that jointly calling somatic variants is superior and able to uncover previously unappreciated genetic variation. Secondly, we contributed to the biomedical field by detecting and describing novel modes of resistance and lastly, we developed a proof of concept approach to detect somatic mutational patterns from plasma samples. Overall, we believe the work done during this PhD has significantly contributed to the knowledge and capabilities in the field of genetic cancer heterogeneity.

# Bibliography

- [1] I. Dagogo-Jack and A. T. Shaw. ‘Tumour heterogeneity and resistance to cancer therapies’. In: *Nat Rev Clin Oncol* 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] T. L. Leong et al. ‘Deep multi-region whole-genome sequencing reveals heterogeneity and gene-by-environment interactions in treatment-naïve, metastatic lung cancer’. In: *Oncogene* 38.10 (Oct. 2018), pp. 1661–1675. DOI: 10.1038/s41388-018-0536-1.
- [4] T. Yan et al. ‘Multi-region sequencing unveils novel actionable targets and spatial heterogeneity in esophageal squamous cell carcinoma’. In: *Nat Commun* 10.1 (Apr. 2019). DOI: 10.1038/s41467-019-09255-1.
- [5] M. 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.
- [6] B. J. Solomon et al. ‘RET Solvent Front Mutations Mediate Acquired Resistance to Selective RET Inhibition in RET-Driven Malignancies’. In: *J Thorac Oncol* 15.4 (Apr. 2020), pp. 541–549. DOI: 10.1016/j.jtho.2020.01.006.
- [7] The Centos Project. *Centos 7*. July 2014. URL: <https://www.centos.org/> (visited on 26/10/2021).
- [8] 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>.
- [9] O. Tange et al. ‘GNU Parallel - The Command-Line Power Tool’. In: *login: The USENIX Magazine* 36.1 (Feb. 2011), pp. 42–47.
- [10] 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/>.
- [11] M. Morgan et al. *BiocParallel: Bioconductor facilities for parallel evaluation*. R package version 1.24.1. 2020. URL: <https://github.com/Bioconductor/BiocParallel>.
- [12] M. Morgan. *BiocManager: Access the Bioconductor Project Package Repository*. R package version 1.30.10. 2019. URL: <https://CRAN.R-project.org/package=BiocManager>.
- [13] A. Zeileis, K. Hornik and P. 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.
- [14] A. Zeileis et al. ‘colorspace: A Toolbox for Manipulating and Assessing Colors and Palettes’. In: *J Stat Softw* 96.1 (2020). DOI: 10.18637/jss.v096.i01.
- [15] F. Favero et al. ‘Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data’. In: *Ann. Oncol.* 26.1 (Jan. 2015), pp. 64–70. DOI: 10.1093/annonc/mdu479.
- [16] R. Shen and V. 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.
- [17] V. E. Seshan and R. 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 29/09/2021).
- [18] D. 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.
- [19] G. 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.



- 
- [20] G. Nilsen, K. Liestol and O. C. Lingjaerde. *copynumber: Segmentation of single- and multi-track copy number data by penalized least squares regression*. R package version 1.29. 2021.
  - [21] W. McLaren et al. ‘The Ensembl Variant Effect Predictor’. In: *Genome Biology* 17.1 (June 2016). DOI: 10.1186/s13059-016-0974-4.
  - [22] M. Dowle and A. Srinivasan. *data.table: Extension of ‘data.frame’*. R package version 1.14.0. 2021. URL: <https://CRAN.R-project.org/package=data.table>.
  - [23] D. Adler and S. T. Kelly. *vioplot: violin plot*. R package version 0.3.5. 2020. URL: <https://github.com/TomKellyGenetics/vioplot>.
  - [24] Z. Gu, R. Eils and M. Schlesner. ‘Complex heatmaps reveal patterns and correlations in multidimensional genomic data’. In: *Bioinformatics* 32 (May 2016), pp. 2847–2849. DOI: 10.1093/bioinformatics/btw313.
  - [25] E. Paradis and K. Schliep. ‘ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R’. In: *Bioinformatics* 35.3 (July 2018). Ed. by R. Schwartz, pp. 526–528. DOI: 10.1093/bioinformatics/bty633.
  - [26] K. Schliep et al. ‘Intertwining phylogenetic trees and networks’. In: *Methods Ecol Evol* 8.10 (Apr. 2017). Ed. by R. Fitzjohn, pp. 1212–1220. DOI: 10.1111/2041-210x.12760.
  - [27] T. Galili. ‘dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering’. In: *Bioinformatics* 31.22 (July 2015), pp. 3718–3720. DOI: 10.1093/bioinformatics/btv428.
  - [28] J. 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>.
  - [29] M. 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>.
  - [30] M. Lawrence et al. ‘Software for Computing and Annotating Genomic Ranges’. In: *PLoS Computational Biology* 9.8 (8 Aug. 2013). Ed. by A. Prlic, e1003118. DOI: 10.1371/journal.pcbi.1003118.
  - [31] T. L. Davis. *optparse: Command Line Option Parser*. R package version 1.6.6. 2020. URL: <https://CRAN.R-project.org/package=optparse>.
  - [32] V. 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.
  - [33] M. 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.
  - [34] Z. 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.
  - [35] J. D. 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/>.
  - [36] H. Pagès et al. *Biostrings: Efficient manipulation of biological strings*. R package version 2.58.0. 2020. URL: <https://bioconductor.org/packages/Biostrings>.
  - [37] R. Rosenthal. *deconstructSigs: Identifies Signatures Present in a Tumor Sample*. R package version 1.8.0. 2016. URL: <https://CRAN.R-project.org/package=deconstructSigs>.
  - [38] H. 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>.

- 
- [39] I. 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.
- [40] E. Neuwirth. *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-2. 2014. URL: <https://CRAN.R-project.org/package=RColorBrewer>.
- [41] R. Kolde. *pheatmap: Pretty Heatmaps*. R package version 1.0.12. 2019. URL: <https://CRAN.R-project.org/package=pheatmap>.
- [42] V. Obenchain and L. Shepherd. *ensemblVEP: R Interface to Ensembl Variant Effect Predictor*. R package version 1.32.0. 2020.
- [43] M. 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>.
- [44] Y. Liao, G. K. Smyth and W. Shi. ‘The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads’. In: *Nucleic Acids Res.* 47.8 (8 Feb. 2019), e47–e47. DOI: 10.1093/nar/gkz114.
- [45] H. Wickham et al. *svglite: An ‘SVG’ Graphics Device*. R package version 2.0.0. 2021. URL: <https://CRAN.R-project.org/package=svglite>.
- [46] P. Murrell. ‘Importing Vector Graphics: The grImport Package for R’. In: *J Stat Softw* 30.4 (2009), pp. 1–37. DOI: 10.18637/jss.v030.i04. URL: <http://www.jstatsoft.org/v30/i04/>.
- [47] D. 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>.
- [48] H. 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>.
- [49] F. Wild. *lsa: Latent Semantic Analysis*. R package version 0.73.2. 2020. URL: <https://CRAN.R-project.org/package=lsa>.
- [50] J. Baglama, L. 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>.
- [51] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 8th June 2016. 260 pp. ISBN: 978-3-319-24277-4. URL: <https://ggplot2.tidyverse.org>.
- [52] G. VanRossum. *The Python language reference*. Hampton, NHRedwood City, Calif: Python Software FoundationSoHo Books, 2010. ISBN: 9781441412690.
- [53] C. R. Harris et al. ‘Array programming with NumPy’. In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2.
- [54] E. B. Stovner and P. Sætrum. ‘PyRanges: efficient comparison of genomic intervals in Python’. In: *Bioinformatics* (Aug. 2019). Ed. by J. Hancock. DOI: 10.1093/bioinformatics/btz615.
- [55] A. Heger, K. Jacobs et al. *pysam: htlib interface for python*. 25th Oct. 2021. URL: <https://github.com/pysam-developers/pysam> (visited on 26/10/2021).
- [56] J. 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.
- [57] P. Danecek et al. ‘Twelve years of SAMtools and BCFtools’. In: *GigaScience* 10.2 (Jan. 2021). DOI: 10.1093/gigascience/giab008.
- [58] A. Miles et al. *zarr-developers/zarr-python: v2.10.2*. 2021. DOI: 10.5281/ZENODO.5579625.

- 
- [59] W. 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.
  - [60] J. Reback et al. *pandas-dev/pandas: Pandas 1.3.4*. 2021. DOI: 10.5281/ZENODO.5574486.
  - [61] R. T. McGibbon et al. *quadprog: Quadratic Programming Solver (Python) version 0.1.10*. Oct. 2021. URL: <http://github.com/quadprog/quadprog> (visited on 26/10/2021).
  - [62] P. Virtanen et al. ‘SciPy 1.0: fundamental algorithms for scientific computing in Python’. In: *Nat. Methods* 17.3 (Feb. 2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.
  - [63] 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.
  - [64] F. Liang et al. ‘Homology-directed repair is a major double-strand break repair pathway in mammalian cells’. In: *Proc. Natl. Acad. Sci.* 95.9 (Apr. 1998), pp. 5172–5177. DOI: 10.1073/pnas.95.9.5172.
  - [65] R. 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.
  - [66] J. C. Venter et al. ‘The Sequence of the Human Genome’. In: *Science* 291.5507 (Feb. 2001), pp. 1304–1351. DOI: 10.1126/science.1058040.
  - [67] C. M. Hammond et al. ‘Histone chaperone networks shaping chromatin function’. In: *Nat Rev Mol Cell Bio* 18.3 (Jan. 2017), pp. 141–158. DOI: 10.1038/nrm.2016.159.
  - [68] B. Bonev and G. Cavalli. ‘Organization and function of the 3D genome’. In: *Nat Rev Genet* 17.11 (Oct. 2016), pp. 661–678. DOI: 10.1038/nrg.2016.112.
  - [69] T. Tateoka. ‘A contribution to the taxonomy of the *Agrostis mertensii*-*flaccida* complex (Poaceae) in Japan’. In: *The Botanical Magazine Tokyo* 88.2 (June 1975), pp. 65–87. DOI: 10.1007/bf02491243.
  - [70] R. Trivers and H. Hare. ‘Haplodiploidy and the evolution of the social insect’. In: *Science* 191.4224 (Jan. 1976), pp. 249–263. DOI: 10.1126/science.1108197.
  - [71] S. P. Otto. ‘The Evolutionary Consequences of Polyploidy’. In: *Cell* 131.3 (Nov. 2007), pp. 452–462. DOI: 10.1016/j.cell.2007.10.022.
  - [72] M. 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.
  - [73] A. Cereda and J. 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.
  - [74] M. 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.
  - [75] D. M. J. Lilley. ‘Structures of helical junctions in nucleic acids’. In: *Q. Rev. Biophys.* 33.2 (May 2000), pp. 109–159. DOI: 10.1017/s0033583500003590.
  - [76] W. P. Hanage, C. Fraser and B. G. Spratt. ‘The impact of homologous recombination on the generation of diversity in bacteria’. In: *J. Theor. Biol.* 239.2 (Mar. 2006), pp. 210–219. DOI: 10.1016/j.jtbi.2005.08.035.
  - [77] Y. Kong et al. ‘Homologous Recombination Drives Both Sequence Diversity and Gene Content Variation in *Neisseria meningitidis*’. In: *Genome Biol Evol* 5.9 (July 2013), pp. 1611–1627. DOI: 10.1093/gbe/evt116.
  - [78] C. 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.

- 
- [79] K. Sprouffske et al. ‘High mutation rates limit evolutionary adaptation in *Escherichia coli*’. In: *PLOS Genetics* 14.4 (Apr. 2018). Ed. by I. Matic, e1007324. DOI: 10.1371/journal.pgen.1007324.
  - [80] L. B. Alexandrov et al. ‘Clock-like mutational processes in human somatic cells’. In: *Nat. Genet.* 47.12 (Nov. 2015), pp. 1402–1407. DOI: 10.1038/ng.3441.
  - [81] L. Moore et al. ‘The mutational landscape of human somatic and germline cells’. In: *Nature* (Aug. 2021). DOI: 10.1038/s41586-021-03822-7.
  - [82] A. Cagan et al. ‘Somatic mutation rates scale with lifespan across mammals’. In: *Nature* 604.7906 (Apr. 2022), pp. 517–524. DOI: 10.1038/s41586-022-04618-z.
  - [83] H. E. Shamseldin et al. ‘Identification of embryonic lethal genes in humans by autozygosity mapping and exome sequencing in consanguineous families’. In: *Genome Biol.* 16.1 (June 2015). DOI: 10.1186/s13059-015-0681-6.
  - [84] L. Frey et al. ‘Mammalian VPS45 orchestrates trafficking through the endosomal system’. In: *Blood* 137.14 (Apr. 2021), pp. 1932–1944. DOI: 10.1182/blood.202006871.
  - [85] S. 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: *Prenat. Diagn.* 32.13 (Nov. 2012), pp. 1225–1232. DOI: 10.1002/pd.4002.
  - [86] K. 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.
  - [87] F. Diehl et al. ‘Circulating mutant DNA to assess tumor dynamics’. In: *Nat. Med.* 14.9 (July 2008), pp. 985–990. DOI: 10.1038/nm.1789.
  - [88] H. Schwarzenbach, D. S. B. Hoon and K. Pantel. ‘Cell-free nucleic acids as biomarkers in cancer patients’. In: *Nat Rev Cancer* 11.6 (May 2011), pp. 426–437. DOI: 10.1038/nrc3066.
  - [89] Y. Gong et al. ‘The role of necroptosis in cancer biology and therapy’. In: *Molecular Cancer* 18.1 (May 2019). DOI: 10.1186/s12943-019-1029-8.
  - [90] L. Zhao et al. ‘Ferroptosis in cancer and cancer immunotherapy’. In: *Cancer Communications* 42.2 (Feb. 2022), pp. 88–116. DOI: 10.1002/cac2.12250.
  - [91] C. Yun and S. Lee. ‘The Roles of Autophagy in Cancer’. In: *International Journal of Molecular Sciences* 19.11 (Nov. 2018), p. 3466. DOI: 10.3390/ijms19113466.
  - [92] O. E. Andreeva et al. ‘Secretion of Mutant DNA and mRNA by the Exosomes of Breast Cancer Cells’. In: *Molecules* 26.9 (Apr. 2021), p. 2499. DOI: 10.3390/molecules26092499.
  - [93] H. Markus et al. ‘Refined characterization of circulating tumor DNA through biological feature integration’. In: *Scientific Reports* 12.1 (Feb. 2022). DOI: 10.1038/s41598-022-05606-z.
  - [94] F. Mouliere et al. ‘Enhanced detection of circulating tumor DNA by fragment size analysis’. In: *Science Translational Medicine* 10.466 (Nov. 2018), eaat4921. DOI: 10.1126/scitranslmed.aat4921.
  - [95] A. Posner et al. ‘A comparison of DNA sequencing and gene expression profiling to assist tissue of origin diagnosis in cancer of unknown primary’. In: *The Journal of Pathology* 259.1 (Nov. 2022), pp. 81–92. DOI: 10.1002/path.6022.
  - [96] L. Tan et al. ‘Prediction and monitoring of relapse in stage III melanoma using circulating tumor DNA’. In: *Ann. Oncol.* 30.5 (May 2019), pp. 804–814. DOI: 10.1093/annonc/mdz048.
  - [97] A. Campos-Carrillo et al. ‘Circulating tumor DNA as an early cancer detection tool’. In: *Pharmacology & Therapeutics* 207 (Mar. 2020), p. 107458. DOI: 10.1016/j.pharmthera.2019.107458.

- 
- [98] O. D. Pons-Belda, A. Fernandez-Uriarte and E. P. Diamandis. ‘Can Circulating Tumor DNA Support a Successful Screening Test for Early Cancer Detection? The Grail Paradigm’. In: *Diagnostics* 11.12 (Nov. 2021), p. 2171. DOI: 10.3390/diagnostics11122171.
- [99] M. J. Duffy, E. P. Diamandis and J. Crown. ‘Circulating tumor DNA (ctDNA) as a pan-cancer screening test: is it finally on the horizon?’ In: *Clinical Chemistry and Laboratory Medicine (CCLM)* 59.8 (Apr. 2021), pp. 1353–1361. DOI: 10.1515/cclm-2021-0171.
- [100] 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: *Proc. Natl. Acad. Sci.* 71.6 (June 1974), pp. 2510–2514. DOI: 10.1073/pnas.71.6.2510.
- [101] F. Sanger and A. Coulson. ‘A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase’. In: *J. Mol. Biol.* 94.3 (May 1975), pp. 441–448. DOI: 10.1016/0022-2836(75)90213-2.
- [102] F. Sanger, S. Nicklen and A. R. Coulson. ‘DNA sequencing with chain-terminating inhibitors’. In: *Proc. Natl. Acad. Sci.* 74.12 (Dec. 1977), pp. 5463–5467. DOI: 10.1073/pnas.74.12.5463.
- [103] E. S. Lander et al. ‘Initial sequencing and analysis of the human genome’. In: *Nature* 409 (15th Feb. 2001): *The human genome*, pp. 860–921. DOI: 10.1038/35057062.
- [104] I. 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 08/09/2021).
- [105] E. R. Mardis. ‘Next-Generation DNA Sequencing Methods’. In: *Annu. Rev. Genomics Hum. Genet.* 9.1 (Sept. 2008), pp. 387–402. DOI: 10.1146/annurev.genom.9.081307.164359.
- [106] J. 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.
- [107] G. M. Church and W. Gilbert. ‘Genomic sequencing.’ In: *Proc. Natl. Acad. Sci.* 81.7 (Apr. 1984), pp. 1991–1995. DOI: 10.1073/pnas.81.7.1991.
- [108] G. Church and S. Kieffer-Higgins. ‘Multiplex DNA sequencing’. In: *Science* 240 (Apr. 1988), pp. 185–188. DOI: 10.1126/science.3353714.
- [109] A. Payne et al. ‘BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files’. In: *Bioinformatics* 35.13 (Nov. 2018). Ed. by I. Birol, pp. 2193–2198. DOI: 10.1093/bioinformatics/bty841.
- [110] P. N. Pratanwanich et al. ‘Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore’. In: *Nat. Biotechnol.* (July 2021). DOI: 10.1038/s41587-021-00949-w.
- [111] N. L. Bray et al. ‘Near-optimal probabilistic RNA-seq quantification’. In: *Nat. Biotechnol.* 34.5 (Apr. 2016), pp. 525–527. DOI: 10.1038/nbt.3519.
- [112] R. Patro et al. ‘Salmon provides fast and bias-aware quantification of transcript expression’. In: *Nat. Methods* 14.4 (Mar. 2017), pp. 417–419. DOI: 10.1038/nmeth.4197.
- [113] B. D. Ondov et al. ‘Mash: fast genome and metagenome distance estimation using MinHash’. In: *Genome Biol.* 17.1 (June 2016). DOI: 10.1186/s13059-016-0997-x.
- [114] B. B. Luczak, B. T. James and H. Z. Girgis. ‘A survey and evaluations of histogram-based statistics in alignment-free sequence comparison’. In: *Briefings in Bioinformatics* 20 (Dec. 2017), pp. 1222–1237. DOI: 10.1093/bib/bbx161.
- [115] H. Li. ‘Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM’. In: *arXiv* (16th Mar. 2013). arXiv: 1303.3997 [q-bio.GN].
- [116] B. Langmead et al. ‘Scaling read aligners to hundreds of threads on general-purpose processors’. In: *Bioinformatics* 35.3 (July 2018). Ed. by J. Hancock, pp. 421–432. DOI: 10.1093/bioinformatics/bty648.

- 
- [117] K. 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.
  - [118] E. Garrison and G. Marth. ‘Haplotype-based variant detection from short-read sequencing’. In: *arXiv preprint arXiv:1207.3907 [q-bio.GN]* (17th July 2012). arXiv: <http://arxiv.org/abs/1207.3907v2> [q-bio.GN].
  - [119] Z. Lai et al. ‘VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research’. In: *Nucleic Acids Res.* 44.11 (Apr. 2016), e108–e108. DOI: 10.1093/nar/gkw227.
  - [120] S. Kim et al. ‘Strelka2: fast and accurate calling of germline and somatic variants’. In: *Nat. Methods* 15.8 (July 2018), pp. 591–594. DOI: 10.1038/s41592-018-0051-x.
  - [121] D. Benjamin et al. ‘Calling Somatic SNVs and Indels with Mutect2’. In: *bioRxiv* (Dec. 2019). DOI: 10.1101/861054.
  - [122] D. P. Cooke, D. C. Wedge and G. Lunter. ‘A unified haplotype-based method for accurate and comprehensive variant calling’. In: *Nat. Biotechnol.* 39.7 (Mar. 2021), pp. 885–892. DOI: 10.1038/s41587-021-00861-3.
  - [123] GATK Team. *Somatic calling is NOT simply a difference between two callsets*. 15th Sept. 2021. URL: <https://gatk.broadinstitute.org/hc/en-us/articles/360035890491> (visited on 23/09/2021).
  - [124] A. Taylor-Weiner et al. ‘DeTiN: overcoming tumor-in-normal contamination’. In: *Nat. Methods* 15.7 (June 2018), pp. 531–534. DOI: 10.1038/s41592-018-0036-9.
  - [125] K. J. Karczewski et al. ‘The mutational constraint spectrum quantified from variation in 141,456 humans’. In: *Nature* 581.7809 (May 2020), pp. 434–443. DOI: 10.1038/s41586-020-2308-7.
  - [126] A. Karimnezhad et al. ‘Accuracy and reproducibility of somatic point mutation calling in clinical-type targeted sequencing data’. In: *BMC Med Genomics* 13.1 (Oct. 2020). DOI: 10.1186/s12920-020-00803-z.
  - [127] F. Abascal et al. ‘Somatic mutation landscapes at single-molecule resolution’. In: *Nature* 593.7859 (Apr. 2021), pp. 405–410. DOI: 10.1038/s41586-021-03477-4.
  - [128] J. H. Breasted. ‘The Edwin Smith Surgical Papyrus: published in facsimile and hieroglyphic transliteration with translation and commentary in two volumes’. In: (1930).
  - [129] S. I. Hajdu. ‘Greco-Roman thought about cancer’. In: *Cancer* 100.10 (2004), pp. 2048–2051. DOI: 10.1002/cncr.20198.
  - [130] J. Chadwick and W. N. Mann. *The medical works of Hippocrates*. Oxford Blackwell Scientific Publications, 1950, pp. 124–147.
  - [131] Celsus. ‘De Medicina’. In: *Br. J. Surg.* 26.103 (Jan. 1939). With an english translation by W. G. Spencer, M.S. (Lond.), F.R.C.S. (Eng.), pp. 658–659. DOI: 10.1002/bjs.18002610338.
  - [132] E. G. Browne. *Arabian medicine. the FitzPatrick lectures Delivered at the College of Physicians in November 1919 and November 1920*. Cambridge Library Collection - History of Medicine. Cambridge: Cambridge University Press, 2011. 154 pp. ISBN: 9780511709296. DOI: 10.1017/cbo9780511709296.
  - [133] J. E. Pilcher. ‘Guy de Chauliac and Henri de Mondeville,-A Surgical Retrospect.’ In: *Annals of surgery* 21 (1 Jan. 1895), pp. 84–102. ISSN: 0003-4932.
  - [134] Paracelsus (Bombastus von Hohenheim TPA). ‘De Grandibus, de Compositionibus, et Dosibus Receptorum ac Naturalium (Libri Septem)’. In: (1562).
  - [135] M. A. 1. Severino. *De recondita abscessuum natura libri VII*. Apud Octavium Beltranum, 1632.
  - [136] Z. Lusitani. ‘Praxis Medical Admiranda’. In: *Lugduni: J. Hugvetan* (1649).
  - [137] N. Tulp. *Observationes Medicae*. Amstelredami, 1652.

- 
- [138] C. Deshaies-Gendron. *Recherches sur la nature et la guerison des cancers. Enquiries into the nature, knowledge, and cure of cancers. By Mr. Deshaies Gendron, ... Done out of French*. Print edition. Gale ECCO, 1701. 146 pp. ISBN: 978-1385259078.
  - [139] J. Nooth. *Observations on the treatment of scirrhus tumours, and cancers of the breast*. G. and J. Robinson, 1804, p. 101.
  - [140] S. I. Hajdu. 'The First Printed Case Reports of Cancer. The First Printed Case Reports of Cancer'. In: *Cancer* (2010). DOI: 10.1002/cncr.25000.
  - [141] M. Etmüller. *ETMULLERUS ABRIDG'D : or, a compleat system of the theory and practice of physic being a ... description of all diseases incident to men, women*. Gale ECCO, 2018. ISBN: 1385770074.
  - [142] L. Heister. *Kleine Chirurgie, oder, Wund-Artzney: in welcher ein kurzer doch deutlicher Unterricht und Begriff dieser Wissenschaftt gegeben ... werden*. German. Nürnberg: Joh. Adam Stein und Gabriel Nicolaus Raspe, 1747.
  - [143] S. I. Hajdu. 'A note from history: Landmarks in history of cancer, part 2'. In: *Cancer* 117.12 (Dec. 2010), pp. 2811–2820. DOI: 10.1002/cncr.25825.
  - [144] J. Müller. *Über den feinern Bau und die Formen der krankhaften Geschwülste : in zwei Lieferungen*. German. Reimer, 1838. URL: <https://echo.mpiwg-berlin.mpg.de/ECHDocuView?url=/permanent/library/84U9MMK0/pageimg&start=61&mode=imagepath&pn=66> (visited on 07/12/2021).
  - [145] J. Bennett. *On Cancerous and Cancroid Growths*. Sutherland and Knox, 1849, pp. vi–viii. URL: [https://books.google.com.au/books?id=VQg%5C\\_AAAAcAAJ](https://books.google.com.au/books?id=VQg%5C_AAAAcAAJ).
  - [146] R. I. C. Virchow. *Die Krankhaften Geschwülste. dreissig Vorlesungen, gehalten während des Wintersemesters 1862-1863 an der Universität zu Berlin*. German. 3 vols. Berlin: Verlag von August Hirschwald, 1863. URL: <http://resource.nlm.nih.gov/62231840R> (visited on 17/08/2021).
  - [147] W. C. Röntgen. 'Ueber eine neue Art von Strahlen'. In: *Ann. Phys.* 300.1 (1898), pp. 12–17. DOI: 10.1002/andp.18983000103.
  - [148] E. Friebe. 'Demonstration eines cancroids des rechten handrucksens, das sich nach langdauerer einwirkung von roentgenstrahlen entwickelt hatte'. In: *Forsch Roentgenstr* 6 (1902), pp. 106–111.
  - [149] W. Scholtz. 'Ueber den Einfluss der Röntgenstrahlen auf die Haut in gesundem und krankem Zustande'. In: *Archiv für Dermatologie und Syphilis* 59.3 (1902), pp. 421–446.
  - [150] P. Ehrlich. *Beiträge zur experimentellen Pathologie und Chemotherapie*. German. Leipzig: Akademische Verlagsgesellschaft, 1909, pp. 167–194. 247 pp. URL: <https://id.lib.harvard.edu/curiosity/contagion/36-990061083080203941> (visited on 07/12/2021).
  - [151] O. Warburg. 'Photochemie der Eisencarbonylverbindungen und das absolute Absorptionsspektrum des Atmungsferments'. In: *Naturwissenschaften* 16 (1928), pp. 856–861.
  - [152] A. Claude, K. R. Porter and E. G. Pickels. 'Electron Microscope Study of Chicken Tumor Cells'. In: *Cancer Res.* 7.7 (1947), pp. 421–430. ISSN: 0008-5472. eprint: <https://cancerres.aacrjournals.org/content/7/7/421.full.pdf>. URL: <https://cancerres.aacrjournals.org/content/7/7/421>.
  - [153] R. J. Huebner and G. J. Todaro. 'Oncogenes of RNA tumor viruses as determinants of cancer'. In: *Proceedings of the National Academy of Sciences* 64.3 (Nov. 1969), pp. 1087–1094. DOI: 10.1073/pnas.64.3.1087.
  - [154] D. Baltimore. 'Viral RNA-dependent DNA polymerase: RNA-dependent DNA polymerase in virions of RNA tumour viruses'. In: *Nature* 226.5252 (June 1970), pp. 1209–1211. DOI: 10.1038/2261209a0.

- 
- [155] H. M. Temin, S. Mizutami et al. ‘RNA-dependent DNA polymerase in virions of Rous sarcoma virus.’ In: *Cold Spring Harb. Symp. Quant. Biol.* 226.0 (Jan. 1970), pp. 1211–1213. DOI: 10.1101/sqb.1970.035.01.100.
  - [156] F. P. Li and J. F. Fraumeni Jr. ‘Rhabdomyosarcoma in children: epidemiologic study and identification of a familial cancer syndrome’. In: *J. Natl. Cancer Inst.* 43.6 (1969), pp. 1365–1373.
  - [157] The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. ‘Pan-cancer analysis of whole genomes’. In: *Nature* 578.7793 (Feb. 2020), pp. 82–93. DOI: 10.1038/s41586-020-1969-6.
  - [158] J. Mersch et al. ‘Cancers associated with *BRCA1* and *BRCA2* mutations other than breast and ovarian’. In: *Cancer* 121.2 (Sept. 2014), pp. 269–275. DOI: 10.1002/cncr.29041.
  - [159] G. Tiao and J. Goodrich. *gnomAD v3.1. New Content, Methods, Annotations, and Data Availability*. 29th Oct. 2020. URL: <https://gnomad.broadinstitute.org/news/2020-10-gnomad-v3-1-new-content-methods-annotations-and-data-availability/> (visited on 04/03/2023).
  - [160] C. Swanton. ‘Intratumor Heterogeneity: Evolution through Space and Time’. In: *Cancer Research* 72.19 (Sept. 2012), pp. 4875–4882. DOI: 10.1158/0008-5472.can-12-2217.
  - [161] R. A. Burrell and C. Swanton. ‘Tumour heterogeneity and the evolution of polyclonal drug resistance’. In: *Molecular Oncology* 8.6 (July 2014), pp. 1095–1111. DOI: 10.1016/j.molonc.2014.06.005.
  - [162] J. N. Jakobsen et al. ‘Concomitant driver mutations in advanced *EGFR*-mutated non-small-cell lung cancer and their impact on erlotinib treatment’. In: *Oncotarget* 9.40 (May 2018), pp. 26195–26208. DOI: 10.18632/oncotarget.25490.
  - [163] G. R. Oxnard et al. ‘Assessment of Resistance Mechanisms and Clinical Implications in Patients With *EGFR* T790M-Positive Lung Cancer and Acquired Resistance to Osimertinib’. In: *JAMA Oncology* 4.11 (Nov. 2018), pp. 1527–1534. DOI: 10.1001/jamaoncol.2018.2969.
  - [164] D. Hanahan and R. A. Weinberg. ‘The Hallmarks of Cancer’. In: *Cell* 100.1 (Jan. 2000), pp. 57–70. DOI: 10.1016/s0092-8674(00)81683-9.
  - [165] D. Hanahan and R. A. Weinberg. ‘Hallmarks of Cancer: The Next Generation’. In: *Cell* 144.5 (Mar. 2011), pp. 646–674. DOI: 10.1016/j.cell.2011.02.013.
  - [166] Y. A. Fouad and C. Aanei. ‘Revisiting the hallmarks of cancer.’ In: *American journal of cancer research* 7 (5 2017), pp. 1016–1036. ISSN: 2156-6976. epublish.
  - [167] D. Hanahan. ‘Hallmarks of Cancer: New Dimensions’. In: *Cancer Discov* 12.1 (Jan. 2022), pp. 31–46. DOI: 10.1158/2159-8290.cd-21-1059.
  - [168] E. A. Kuczynski et al. ‘Vessel co-option in cancer.’ In: *Nat Rev Clin Oncol* 16.8 (Feb. 2019), pp. 469–493. DOI: 10.1038/s41571-019-0181-9.
  - [169] J. Noorbakhsh et al. ‘Distribution-based measures of tumor heterogeneity are sensitive to mutation calling and lack strong clinical predictive power’. In: *Sci. Rep.* 8.1 (July 2018). DOI: 10.1038/s41598-018-29154-7.
  - [170] R. Poplin et al. ‘Scaling accurate genetic variant discovery to tens of thousands of samples’. In: *bioRxiv* (Nov. 2017). DOI: 10.1101/201178.
  - [171] N. A. Miller et al. ‘A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases’. In: *Genome Med.* 7.1 (Sept. 2015). DOI: 10.1186/s13073-015-0221-8.
  - [172] R. Poplin et al. ‘A universal SNP and small-indel variant caller using deep neural networks’. In: *Nat. Biotechnol.* 36.10 (Sept. 2018), pp. 983–987. DOI: 10.1038/nbt.4235.



- 
- [173] M. Schirmer et al. ‘Illumina error profiles: resolving fine-scale variation in meta-genomic sequencing data’. In: *BMC Bioinformatics* 17.1 (Mar. 2016). DOI: 10.1186/s12859-016-0976-y.
  - [174] N. Stoler and A. Nekrutenko. ‘Sequencing error profiles of Illumina sequencing instruments’. In: *NAR Genomics and Bioinformatics* 3.1 (Jan. 2021). DOI: 10.1093/nargab/lqab019.
  - [175] G. v. d. A. Brian O’Connor. *Genomics in the Cloud*. O’Reilly UK Ltd., 1st May 2020. 467 pp. ISBN: 1491975199. URL: <https://www.oreilly.com/library/view/genomics-in-the/9781491975183/>.
  - [176] GATK Team. *Panel of Normals (PON)*. 23rd July 2021. URL: <https://gatk.broadinstitute.org/hc/en-us/articles/360035890631> (visited on 23/09/2021).
  - [177] GATK Team. *Mutect2 multi-sample*. 25th Sept. 2020. URL: <https://gatk.broadinstitute.org/hc/en-us/community/posts/360062528691> (visited on 23/10/2020).
  - [178] M. Josephidou, A. G. Lynch and S. 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.
  - [179] C. Flensburg et al. ‘SuperFreq: Integrated mutation detection and clonal tracking in cancer’. In: *PLOS Comput. Biol.* 16.2 (Feb. 2020). Ed. by F. Markowetz, e1007603. DOI: 10.1371/journal.pcbi.1007603.
  - [180] S. Hollizeck et al. ‘Custom workflows to improve joint variant calling from multiple related tumour samples: FreeBayesSomatic and Strelka2Pass’. In: *Bioinformatics* (Sept. 2021). Ed. by C. Alkan. DOI: 10.1093/bioinformatics/btab606.
  - [181] C. Chiang et al. ‘SpeedSeq: ultra-fast personal genome analysis and interpretation’. In: *Nat. Methods* 12.10 (Aug. 2015), pp. 966–968. DOI: 10.1038/nmeth.3505.
  - [182] B. Chapman et al. *bcio/bcio-nextgen: v1.2.4*. 2021. DOI: 10.5281/ZENODO.3564938.
  - [183] B. 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.
  - [184] W. Huang et al. ‘ART: a next-generation sequencing read simulator’. In: *Bioinformatics* 28.4 (Dec. 2011), pp. 593–594. DOI: 10.1093/bioinformatics/btr708.
  - [185] A. D. Ewing et al. ‘Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection’. In: *Nature Methods* 12 (May 2015), pp. 623–630. DOI: 10.1038/nmeth.3407.
  - [186] I. A. Vergara et al. ‘Evolution of late-stage metastatic melanoma is dominated by aneuploidy and whole genome doubling’. In: *Nat Commun* 12.1 (Mar. 2021). DOI: 10.1038/s41467-021-21576-8.
  - [187] Z. 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.
  - [188] N. Saitou and M. Nei. ‘The neighbor-joining method: a new method for reconstructing phylogenetic trees.’ In: *Mol. Biol. Evol.* (July 1987). DOI: 10.1093/oxfordjournals.molbev.a040454.
  - [189] R. Mihaescu, D. Levy and L. Pachter. ‘Why Neighbor-Joining Works’. In: *Algorithmica* 54.1 (Dec. 2007), pp. 1–24. DOI: 10.1007/s00453-007-9116-4.
  - [190] R. R. Sokal and C. D. Michener. *A Statistical Method for Evaluating Systematic Relationships*. Vol. 38.2. University of Kansas science bulletin 22. University of Kansas, 1958. 30 pp.
  - [191] E. Zuckerkandl and L. Pauling. ‘Molecular Disease, Evolution, and Genic Heterogeneity’. In: *Horizons in biochemistry* (1962), pp. 189–225.

- 
- [192] D. Shibata. ‘Mutation and epigenetic molecular clocks in cancer’. In: *Carcinogenesis* 32.2 (Nov. 2010), pp. 123–128. DOI: 10.1093/carcin/bgq239.
  - [193] J. Felsenstein. ‘Evolutionary trees from DNA sequences: A maximum likelihood approach’. In: *J. Mol. Evol.* 17.6 (Nov. 1981), pp. 368–376. DOI: 10.1007/bf01734359.
  - [194] M. Hasegawa, H. Kishino and T.-a. Yano. ‘Dating of the human-ape splitting by a molecular clock of mitochondrial DNA’. In: *J. Mol. Evol.* 22.2 (Oct. 1985), pp. 160–174. DOI: 10.1007/bf02101694.
  - [195] 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.
  - [196] B. Werner et al. ‘Measuring single cell divisions in human tissues from multi-region sequencing data’. In: *Nat Commun* 11.1 (Feb. 2020). DOI: 10.1038/s41467-020-14844-6.
  - [197] 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.
  - [198] D. M. de Vienne. ‘Tanglegrams Are Misleading for Visual Evaluation of Tree Congruence’. In: *Molecular Biology and Evolution* 36.1 (Oct. 2018). Ed. by J. Townsend, pp. 174–176. DOI: 10.1093/molbev/msy196.
  - [199] A. G. Deshwar et al. ‘PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors’. In: *Genome Biol.* 16.1 (Feb. 2015). DOI: 10.1186/s13059-015-0602-8.
  - [200] Y. Jiang et al. ‘Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing’. In: *Proceedings of the National Academy of Sciences* 113.37 (Aug. 2016), E5528–E5537. DOI: 10.1073/pnas.1522203113.
  - [201] F. Marass et al. ‘A phylogenetic latent feature model for clonal deconvolution’. In: *Ann. Appl. Stat.* 10.4 (Dec. 2016). DOI: 10.1214/16-aos986.
  - [202] S. Miura et al. ‘Predicting clone genotypes from tumor bulk sequencing of multiple samples’. In: *Bioinformatics* (June 2018). Ed. by J. Hancock. DOI: 10.1093/bioinformatics/bty469.
  - [203] M. El-Kebir, G. Satas and B. J. Raphael. ‘Inferring parsimonious migration histories for metastatic cancers’. In: *Nature Genetics* 50.5 (Apr. 2018), pp. 718–726. DOI: 10.1038/s41588-018-0106-z.
  - [204] G. Caravagna et al. ‘The MOBSTER R package for tumour subclonal deconvolution from bulk DNA whole-genome sequencing data’. In: *BMC Bioinf.* 21.1 (Nov. 2020). DOI: 10.1186/s12859-020-03863-1.
  - [205] M. Tarabichi et al. ‘A practical guide to cancer subclonal reconstruction from DNA sequencing.’ In: *Nature methods* 18.2 (2 Feb. 2021), pp. 144–155. ISSN: 1548-7105. DOI: 10.1038/s41592-020-01013-2. ppublish.
  - [206] S. Miura et al. ‘Power and pitfalls of computational methods for inferring clone phylogenies and mutation orders from bulk sequencing data’. In: *Sci. Rep.* 10.1 (Feb. 2020). DOI: 10.1038/s41598-020-59006-2.
  - [207] I. Leshchiner et al. ‘Comprehensive analysis of tumour initiation, spatial and temporal progression under multiple lines of treatment’. In: *bioRxiv* (Dec. 2018). DOI: 10.1101/508127.
  - [208] M. Gerstung et al. ‘The evolutionary history of 2,658 cancers’. In: *Nature* 578.7793 (Feb. 2020), pp. 122–128. DOI: 10.1038/s41586-019-1907-7.
  - [209] P. P. Gardner et al. ‘Sustained software development, not number of citations or journal choice, is indicative of accurate bioinformatic software’. In: *Genome Biol.* 23.1 (Feb. 2022). DOI: 10.1186/s13059-022-02625-x.

- 
- [210] K. A. Fennell et al. ‘Non-genetic determinants of malignant clonal fitness at single-cell resolution’. In: *Nature* 601.7891 (Dec. 2021), pp. 125–131. DOI: 10.1038/s41586-021-04206-7.
  - [211] L. Penter, S. H. Gohil and C. J. Wu. ‘Natural Barcodes for Longitudinal Single Cell Tracking of Leukemic and Immune Cell Dynamics’. In: *Front. Immunol.* 12 (Jan. 2022). DOI: 10.3389/fimmu.2021.788891.
  - [212] C. Sudlow et al. ‘UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age’. In: *PLOS Medicine* 12.3 (Mar. 2015), e1001779. DOI: 10.1371/journal.pmed.1001779.
  - [213] Cancer Council Victoria. *Victorian Cancer Biobank. Integrity. Excellence. Innovation. Empowering Patients.* 2006. URL: [viccancerbiobank.org.au](http://viccancerbiobank.org.au) (visited on 03/03/2022).
  - [214] G. J. Yoshida. ‘Applications of patient-derived tumor xenograft models and tumor organoids’. In: *Journal of Hematology & Oncology* 13.1 (Jan. 2020). DOI: 10.1186/s13045-019-0829-z.
  - [215] K. Alsop et al. ‘A community-based model of rapid autopsy in end-stage cancer patients’. In: *Nat. Biotechnol.* 34.10 (Sept. 2016), pp. 1010–1014. DOI: 10.1038/nbt.3674.
  - [216] R. L. Siegel, K. D. Miller and A. Jemal. ‘Cancer statistics, 2018’. In: *CA: A Cancer Journal for Clinicians* 68.1 (Jan. 2018), pp. 7–30. DOI: 10.3322/caac.21442.
  - [217] J. R. Molina et al. ‘Non-Small Cell Lung Cancer: Epidemiology, Risk Factors, Treatment, and Survivorship’. In: *Mayo Clin. Proc.* 83.5 (May 2008), pp. 584–594. DOI: 10.4065/83.5.584.
  - [218] S. Sun, J. H. Schiller and A. F. Gazdar. ‘Lung cancer in never smokers — a different disease’. In: *Nat Rev Cancer* 7.10 (Oct. 2007), pp. 778–790. DOI: 10.1038/nrc2190.
  - [219] K. Suda et al. ‘Innate Genetic Evolution of Lung Cancers and Spatial Heterogeneity: Analysis of Treatment-Naïve Lesions’. In: *J Thorac Oncol* 13 (Oct. 2018), pp. 1496–1507. DOI: 10.1016/j.jtho.2018.05.039.
  - [220] N. I. Lindeman et al. ‘Updated Molecular Testing Guideline for the Selection of Lung Cancer Patients for Treatment With Targeted Tyrosine Kinase Inhibitors: Guideline From the College of American Pathologists, the International Association for the Study of Lung Cancer, and the Association for Molecular Pathology’. In: *Archives of Pathology & Laboratory Medicine* 142.3 (Jan. 2018), pp. 321–346. DOI: 10.5858/arpa.2017-0388-cp.
  - [221] P. Savas et al. ‘The Subclonal Architecture of Metastatic Breast Cancer: Results from a Prospective Community-Based Rapid Autopsy Program “CASCADE”’. In: *PLOS Medicine* 13.12 (Dec. 2016). Ed. by M. Ladanyi, e1002204. DOI: 10.1371/journal.pmed.1002204.
  - [222] S. Lee et al. ‘NGSCheckMate: software for validating sample identity in next-generation sequencing studies within and across data types’. In: *Nucleic Acids Res.* 45.11 (Mar. 2017), e103–e103. DOI: 10.1093/nar/gkx193.
  - [223] B. S. Pedersen et al. ‘Somali: rapid relatedness estimation for cancer and germline studies using efficient genome sketches’. In: *Genome Med.* 12.1 (July 2020). DOI: 10.1186/s13073-020-00761-2.
  - [224] S. Andrews. *FastQC: A Quality Control Tool for High Throughput Sequence Data.* 2010. URL: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
  - [225] Broad Institute. *Picard toolkit.* <http://broadinstitute.github.io/picard/>. Comp. software. Version 2.23.8. 2019.
  - [226] D. L. Cameron et al. ‘GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing’. In: *Genome Biol.* 22.1 (July 2021). DOI: 10.1186/s13059-021-02423-x.

- 
- [227] L. B. Alexandrov et al. ‘The repertoire of mutational signatures in human cancer’. In: *Nature* 578.7793 (Feb. 2020), pp. 94–101. DOI: 10.1038/s41586-020-1943-3.
  - [228] M. M. Bjaanaes et al. ‘Whole genome copy number analyses reveal a highly aberrant genome in TP53 mutant lung adenocarcinoma tumors’. In: *BMC Cancer* 21.1 (Oct. 2021). DOI: 10.1186/s12885-021-08811-7.
  - [229] X. Ni et al. ‘Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients’. In: *Proc. Natl. Acad. Sci.* 110.52 (Dec. 2013), pp. 21083–21088. DOI: 10.1073/pnas.1320659110.
  - [230] E. Talevich et al. ‘CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing’. In: *PLOS Computational Biology* 12.4 (Apr. 2016), e1004873. DOI: 10.1371/journal.pcbi.1004873.
  - [231] S. Zaccaria and B. J. Raphael. ‘Accurate quantification of copy-number aberrations and whole-genome duplications in multi-sample tumor sequencing data’. In: *Nat Commun* 11.1 (Sept. 2020). DOI: 10.1038/s41467-020-17967-y.
  - [232] A. Talasz et al. *Use of the GUARDANT360 noninvasive tumor sequencing assay on 300 patients across colorectal, melanoma, lung, breast, and prostate cancers and its clinical utility*. 2014.
  - [233] L. B. Alexandrov et al. ‘Signatures of mutational processes in human cancer’. In: *Nature* 500.7463 (Aug. 2013), pp. 415–421. DOI: 10.1038/nature12477.
  - [234] H. Jin et al. ‘EGFR activation limits the response of liver cancer to lenvatinib’. In: *Nature* 595.7869 (July 2021), pp. 730–734. DOI: 10.1038/s41586-021-03741-7.
  - [235] H. Do and A. Dobrovic. ‘Sequence Artifacts in DNA from Formalin-Fixed Tissues: Causes and Strategies for Minimization’. In: *Clin. Chem.* 61.1 (Jan. 2015), pp. 64–71. DOI: 10.1373/clinchem.2014.223040.
  - [236] F. M. Johnson et al. ‘Phase II Study of Dasatinib in Patients With Advanced Non–Small-Cell Lung Cancer’. In: *J. Clin. Oncol.* 28.30 (Oct. 2010), pp. 4609–4615. DOI: 10.1200/jco.2010.30.5474.
  - [237] M. Bersanelli et al. ‘L718Q Mutation as New Mechanism of Acquired Resistance to AZD9291 in EGFR -Mutated NSCLC’. In: *J Thorac Oncol* 11.10 (Oct. 2016), e121–e123. DOI: 10.1016/j.jtho.2016.05.019.
  - [238] S. Wang et al. ‘EGFR C797S mutation mediates resistance to third-generation inhibitors in T790M-positive non-small cell lung cancer’. In: *Journal of Hematology & Oncology* 9.1 (July 2016). DOI: 10.1186/s13045-016-0290-1.
  - [239] A. Leonetti et al. ‘Resistance mechanisms to osimertinib in EGFR-mutated non-small cell lung cancer’. In: *Br. J. Cancer* 121.9 (Sept. 2019), pp. 725–737. DOI: 10.1038/s41416-019-0573-8.
  - [240] Q. Zhang et al. ‘EGFR L792H and G796R: Two Novel Mutations Mediating Resistance to the Third-Generation EGFR Tyrosine Kinase Inhibitor Osimertinib’. In: *J Thorac Oncol* 13.9 (Sept. 2018), pp. 1415–1421. DOI: 10.1016/j.jtho.2018.05.024.
  - [241] M. G. Oser et al. ‘Transformation from non-small-cell lung cancer to small-cell lung cancer: molecular drivers and cells of origin’. In: *The Lancet Oncology* 16.4 (Apr. 2015), e165–e172. DOI: 10.1016/s1470-2045(14)71180-5.
  - [242] R. Aggarwal et al. ‘Clinical and Genomic Characterization of Treatment-Emergent Small-Cell Neuroendocrine Prostate Cancer: A Multi-institutional Prospective Study’. In: *J. Clin. Oncol.* 36.24 (Aug. 2018), pp. 2492–2503. DOI: 10.1200/jco.2017.77.6880.
  - [243] M. Offin et al. ‘Concurrent RB1 and TP53 Alterations Define a Subset of EGFR-Mutant Lung Cancers at risk for Histologic Transformation and Inferior Clinical Outcomes’. In: *J Thorac Oncol* 14.10 (Oct. 2019), pp. 1784–1793. DOI: 10.1016/j.jtho.2019.06.002.
  - [244] M. L. Suva, N. Riggi and B. E. Bernstein. ‘Epigenetic Reprogramming in Cancer’. In: *Science* 339.6127 (Mar. 2013), pp. 1567–1570. DOI: 10.1126/science.1230184.

- 
- [245] D. Brown et al. ‘Phylogenetic analysis of metastatic progression in breast cancer using somatic mutations and copy number aberrations’. In: *Nat Commun* 8.1 (Apr. 2017). DOI: 10.1038/ncomms14944.
  - [246] M. Kimura. ‘Evolutionary Rate at the Molecular Level’. In: *Nature* 217.5129 (Feb. 1968), pp. 624–626. DOI: 10.1038/217624a0.
  - [247] M. Lynch. ‘Phylogenetic Hypotheses Under the Assumption of Neutral Quantitative-Genetic Variation’. In: *Evolution* 43.1 (Jan. 1989), p. 1. DOI: 10.2307/2409160.
  - [248] V. L. Cannataro and J. P. Townsend. ‘Neutral Theory and the Somatic Evolution of Cancer’. In: *Mol. Biol. Evol.* 35.6 (Apr. 2018). Ed. by S. Kumar, pp. 1308–1315. DOI: 10.1093/molbev/msy079.
  - [249] A. Eyre-Walker and P. D. Keightley. ‘The distribution of fitness effects of new mutations’. In: *Nat Rev Genet* 8.8 (Aug. 2007), pp. 610–618. DOI: 10.1038/nrg2146.
  - [250] J. F. C. Kingman. ‘On the genealogy of large populations’. In: *J Appl Probab* 19.A (1982), pp. 27–43. DOI: 10.2307/3213548.
  - [251] L. S. Ludwig et al. ‘Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics’. In: *Cell* 176.6 (Mar. 2019), 1325–1339.e22. DOI: 10.1016/j.cell.2019.01.022.
  - [252] J. N. Harvey and D. Barnett. ‘Endocrine dysfunction in Kearns-Sayre syndrome’. In: *Clin. Endocrinol. (Oxf.)* 37.1 (July 1992), pp. 97–104. DOI: 10.1111/j.1365-2265.1992.tb02289.x.
  - [253] M. P. Adam et al. ‘GeneReviews’. In: (1993). ppublish.
  - [254] M. Hirano et al. ‘MELAS: An original case and clinical criteria for diagnosis’. In: *Neuromuscul. Disord.* 2.2 (Jan. 1992), pp. 125–135. DOI: 10.1016/0960-8966(92)90045-8.
  - [255] A. Rodell et al. ‘Natural selection of mitochondria during somatic lifetime promotes healthy aging’. In: *Front. Neuroenerg.* 5 (2013). DOI: 10.3389/fnene.2013.00007.
  - [256] Y. Yuan et al. ‘Comprehensive molecular characterization of mitochondrial genomes in human cancers’. In: *Nat. Genet.* 52.3 (Feb. 2020), pp. 342–352. DOI: 10.1038/s41588-019-0557-x.
  - [257] L. W. Cole. ‘The Evolution of Per-cell Organelle Number’. In: *Front. Cell Dev. Biol.* 4 (Aug. 2016). DOI: 10.3389/fcell.2016.00085.
  - [258] M. Kimura. ‘THE NUMBER OF HETEROZYGOUS NUCLEOTIDE SITES MAINTAINED IN A FINITE POPULATION DUE TO STEADY FLUX OF MUTATIONS’. In: *Genetics* 61.4 (Apr. 1969), pp. 893–903. DOI: 10.1093/genetics/61.4.893.
  - [259] J. P. van Meerbeeck, D. A. Fennell and D. K. De Ruyscher. ‘Small-cell lung cancer’. In: *The Lancet* 378.9804 (Nov. 2011), pp. 1741–1755. DOI: 10.1016/s0140-6736(11)60165-7.
  - [260] M. G. Raso, N. Bota-Rabassedas and I. I. Wistuba. ‘Pathology and Classification of SCLC’. In: *Cancers* 13.4 (Feb. 2021), p. 820. DOI: 10.3390/cancers13040820.
  - [261] H. Zhou et al. ‘Multi-region exome sequencing reveals the intratumoral heterogeneity of surgically resected small cell lung cancer’. In: *Nature Communications* 12.1 (Sept. 2021). DOI: 10.1038/s41467-021-25787-x.
  - [262] M. Jamal-Hanjani et al. ‘Tracking the Evolution of Non-Small-Cell Lung Cancer’. In: *N. Engl. J. Med.* 376.22 (June 2017), pp. 2109–2121. DOI: 10.1056/nejmoa1616288.
  - [263] H. Easwaran, H.-C. Tsai and S. B. Baylin. ‘Cancer Epigenetics: Tumor Heterogeneity, Plasticity of Stem-like States, and Drug Resistance’. In: *Molecular Cell* 54.5 (June 2014), pp. 716–727. DOI: 10.1016/j.molcel.2014.05.015.

- 
- [264] M. Hollstein et al. ‘p53 Mutations in Human Cancers’. In: *Science* 253.5015 (July 1991), pp. 49–53. DOI: 10.1126/science.1905840.
  - [265] J. E. Kucab et al. ‘A Compendium of Mutational Signatures of Environmental Agents’. In: *Cell* 177.4 (May 2019), 821–836.e16. DOI: 10.1016/j.cell.2019.03.001.
  - [266] N. J. O’Neil, M. L. Bailey and P. Hieter. ‘Synthetic lethality and cancer’. In: *Nat Rev Genet* 18.10 (June 2017), pp. 613–623. DOI: 10.1038/nrg.2017.47.
  - [267] A. Auton et al. ‘A global reference for human genetic variation’. In: *Nature* 526.7571 (Sept. 2015), pp. 68–74. DOI: 10.1038/nature15393.
  - [268] M. Heydari et al. ‘Illumina error correction near highly repetitive DNA regions improves de novo genome assembly’. In: *BMC Bioinformatics* 20.1 (June 2019). DOI: 10.1186/s12859-019-2906-2.
  - [269] I. Hudecova et al. ‘Characteristics, origin, and potential for cancer diagnostics of ultrashort plasma cell-free DNA’. In: *Genome Res.* (Dec. 2021), gr.275691.121. DOI: 10.1101/gr.275691.121.
  - [270] Y. Guo et al. ‘The effect of strand bias in Illumina short-read sequencing data’. In: *BMC Genomics* 13.1 (Nov. 2012). DOI: 10.1186/1471-2164-13-666.
  - [271] C. T. Saunders et al. ‘Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs’. In: *Bioinformatics* 28 (May 2012), pp. 1811–1817. DOI: 10.1093/bioinformatics/bts271.
  - [272] GATK Team. *StrandBiasBySample*. 18th Sept. 2019. URL: <https://gatk.broadinstitute.org/hc/en-us/articles/360040096492> (visited on 21/12/2021).
  - [273] A. M. pyup.io bot; Murillo R.; Peter Ralph; Nick Harding; Rahul Pisupati; Summer Rae; Tim Millar. *scikit-allel: A Python package for exploring and analysing genetic variation data*. 14th June 2021. DOI: 10.5281/zenodo.4759368.
  - [274] A. G. Lynch. ‘Decomposition of mutational context signatures using quadratic programming methods’. In: *F1000Research* 5 (June 2016), p. 1253. DOI: 10.12688/f1000research.8918.1.
  - [275] M. L. Delignette-Muller and C. Dutang. ‘fitdistrplus: An R Package for Fitting Distributions’. In: *J Stat Softw* 64.4 (2015), pp. 1–34. URL: <https://www.jstatsoft.org/v64/i04/>.
  - [276] V. A. Adalsteinsson et al. ‘Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors’. In: *Nat Commun* 8.1 (Nov. 2017). DOI: 10.1038/s41467-017-00965-y.
  - [277] M. S. Lawrence et al. ‘Mutational heterogeneity in cancer and the search for new cancer-associated genes’. In: *Nature* 499.7457 (June 2013), pp. 214–218. DOI: 10.1038/nature12213.
  - [278] T. Chen et al. ‘Hotspot mutations delineating diverse mutational signatures and biological utilities across cancer types’. In: *BMC Genomics* 17.S2 (June 2016). DOI: 10.1186/s12864-016-2727-x.
  - [279] J. G. Tate et al. ‘COSMIC: the Catalogue Of Somatic Mutations In Cancer’. In: *Nucleic Acids Res.* 47.D1 (Oct. 2018), pp. D941–D947. DOI: 10.1093/nar/gky1015.
  - [280] Wellcome Sanger Institute. *COSMIC database v95*. 14th Nov. 2021. URL: <https://cancer.sanger.ac.uk/>.
  - [281] R. Barroso-Sousa et al. ‘Prevalence and mutational determinants of high tumor mutation burden in breast cancer’. In: *Ann. Oncol.* 31.3 (Mar. 2020), pp. 387–394. DOI: 10.1016/j.annonc.2019.11.010.
  - [282] I. Martincorena et al. ‘Somatic mutant clones colonize the human esophagus with age’. In: *Science* 362.6417 (Nov. 2018), pp. 911–917. DOI: 10.1126/science.aau3879.

- 
- [283] M. Ganuza et al. ‘The global clonal complexity of the murine blood system declines throughout life and after serial transplantation’. In: *Blood* 133.18 (May 2019), pp. 1927–1942. DOI: 10.1182/blood-2018-09-873059.
  - [284] W. Meyerson et al. ‘Origins and characterization of variants shared between databases of somatic and germline human mutations’. In: *BMC Bioinformatics* 21.1 (June 2020). DOI: 10.1186/s12859-020-3508-8.
  - [285] J.-F. Spinella et al. ‘SNooPer: a machine learning-based method for somatic variant identification from low-pass next-generation sequencing’. In: *BMC Genomics* 17.1 (Nov. 2016). DOI: 10.1186/s12864-016-3281-2.
  - [286] S. M. E. Sahraeian et al. ‘Achieving robust somatic mutation detection with deep learning models derived from reference data sets of a cancer sample’. In: *Genome Biology* 23.1 (Jan. 2022). DOI: 10.1186/s13059-021-02592-9.
  - [287] S. Wang et al. ‘Copy number signature analysis tool and its application in prostate cancer reveals distinct mutational processes and clinical outcomes’. In: *PLOS Genetics* 17.5 (May 2021). Ed. by D. A. Gordenin, e1009557. DOI: 10.1371/journal.pgen.1009557.
  - [288] K. K. Lin et al. ‘BRCA Reversion Mutations in Circulating Tumor DNA Predict Primary and Acquired Resistance to the PARP Inhibitor Rucaparib in High-Grade Ovarian Carcinoma’. In: *Cancer Discovery* 9.2 (Nov. 2018), pp. 210–219. DOI: 10.1158/2159-8290.cd-18-0715.
  - [289] S. Rebhandl et al. ‘AID/APOBEC deaminases and cancer’. In: *Oncoscience* 2.4 (Apr. 2015), pp. 320–333. DOI: 10.18632/oncoscience.155.
  - [290] S.-J. Dawson et al. ‘Analysis of Circulating Tumor DNA to Monitor Metastatic Breast Cancer’. In: *N. Engl. J. Med.* 368.13 (Mar. 2013), pp. 1199–1209. DOI: 10.1056/nejmoa1213261.
  - [291] H. G. Russnes et al. ‘Genomic Architecture Characterizes Tumor Progression Paths and Fate in Breast Cancer Patients’. In: *Science Translational Medicine* 2.38 (June 2010). DOI: 10.1126/scitranslmed.3000611.
  - [292] C. Curtis et al. ‘The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups’. In: *Nature* 486.7403 (Apr. 2012), pp. 346–352. DOI: 10.1038/nature10983.
  - [293] S. Akashi-Tanaka et al. ‘BRCAness Predicts Resistance to Taxane-Containing Regimens in Triple Negative Breast Cancer During Neoadjuvant Chemotherapy’. In: *Clinical Breast Cancer* 15.1 (Feb. 2015), pp. 80–85. DOI: 10.1016/j.clbc.2014.08.003.
  - [294] J. R. Homburger et al. ‘Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores’. In: *Genome Med.* 11.1 (Nov. 2019). DOI: 10.1186/s13073-019-0682-2.
  - [295] M. Chen et al. ‘Applying low coverage whole genome sequencing to detect malignant ovarian mass’. In: *J Transl Med* 19.1 (Aug. 2021). DOI: 10.1186/s12967-021-03046-3.
  - [296] Z. D. Stephens et al. ‘Big Data: Astronomical or Genomical?’ In: *PLOS Biology* 13.7 (July 2015), e1002195. DOI: 10.1371/journal.pbio.1002195.
  - [297] Singular Genomics. *Singular Genomics Launches the G4 Sequencing Platform*. 16th Dec. 2021. URL: <https://investor.singulargenomics.com/news-releases/news-release-details/singular-genomics-launches-g4-sequencing-platform> (visited on 05/06/2022).
  - [298] Ultima Genomics. *Ultima Genomics Delivers the \$100 Genome*. 31st May 2022. URL: <https://www.ultimagenomics.com/blog/ultima-genomics-delivers-usd100-genome> (visited on 14/06/2022).
  - [299] Element Biosciences Inc. *Element Launches the AVITI™ System to Democratize Access to Genomics*. 14th Mar. 2022. URL: <https://www.elementbiosciences.com/news/element-launches-the-aviti-system-to-democratize-access-to-genomics> (visited on 03/06/2022).

- 
- [300] J. E. Valle-Inclan et al. ‘A multi-platform reference for somatic structural variation detection’. In: *Cell Genomics* (June 2022). DOI: 10.1016/j.xgen.2022.100139.
  - [301] M. A. DePristo et al. ‘A framework for variation discovery and genotyping using next-generation DNA sequencing data’. In: *Nat. Genet.* 43.5 (Apr. 2011), pp. 491–498. DOI: 10.1038/ng.806.
  - [302] B. Ç. Toptaş et al. ‘Comparing complex variants in family trios’. In: *Bioinformatics* (June 2018). Ed. by O. Stegle. DOI: 10.1093/bioinformatics/bty443.
  - [303] D. Wang et al. ‘Multiregion Sequencing Reveals the Genetic Heterogeneity and Evolutionary History of Osteosarcoma and Matched Pulmonary Metastases’. In: *Cancer Res.* 79.1 (Nov. 2018), pp. 7–20. DOI: 10.1158/0008-5472.can-18-1086.
  - [304] Z. Chen et al. ‘Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency’. In: *Sci. Rep.* 10.1 (Feb. 2020). DOI: 10.1038/s41598-020-60559-5.
  - [305] R. Lupat et al. *Janis: A Python framework for Portable Pipelines*. en. 2021. DOI: 10.5281/ZENODO.4427231.
  - [306] H. Li and R. Durbin. ‘Fast and accurate short read alignment with Burrows-Wheeler transform’. In: *Bioinformatics* 25.14 (May 2009), pp. 1754–1760. DOI: 10.1093/bioinformatics/btp324.
  - [307] K. D. Doig et al. ‘Canary: an atomic pipeline for clinical amplicon assays’. In: *BMC Bioinf.* 18.1 (Dec. 2017). DOI: 10.1186/s12859-017-1950-z.
  - [308] Roche Sequencing Solutions, Inc. *AVENIO ctDNA and Tumour Tissue Expanded Panel*. Sept. 2018. URL: <https://sequencing.roche.com/content/dam/rochesequence/worldwide/resources/brochure-avenio-ctdna-tumor-tissue-expanded-kit-gene-list-SEQ100327.pdf> (visited on 10/03/2022).
  - [309] H. Greulich. ‘The Genomics of Lung Adenocarcinoma: Opportunities for Targeted Therapies’. In: *Genes & Cancer* 1.12 (Dec. 2010), pp. 1200–1210. DOI: 10.1177/1947601911407324.
  - [310] A. El-Telbany and P. C. Ma. ‘Cancer Genes in Lung Cancer: Racial Disparities: Are There Any?’ In: *Genes & Cancer* 3.7-8 (July 2012), pp. 467–480. DOI: 10.1177/1947601912465177.
  - [311] The Cancer Genome Atlas Research Network. ‘Comprehensive molecular profiling of lung adenocarcinoma’. In: *Nature* 511.7511 (July 2014), pp. 543–550. DOI: 10.1038/nature13385.
  - [312] National Comprehensive Cancer Network. *NCCN Guidelines. Non-Small Cell Lung Cancer*. Version 2.2022. 3rd July 2022. URL: [https://www.nccn.org/professionals/physician\\_gls/pdf/nscl.pdf](https://www.nccn.org/professionals/physician_gls/pdf/nscl.pdf) (visited on 15/03/2022).
  - [313] Z. Fan et al. ‘The risk variant rs884225 within EGFR impairs miR-103a-3p’s anti-tumourigenic function in non-small cell lung cancer’. In: *Oncogene* 38.13 (Nov. 2018), pp. 2291–2304. DOI: 10.1038/s41388-018-0576-6.
  - [314] J. Jiang et al. ‘MKRN2 inhibits migration and invasion of non-small-cell lung cancer by negatively regulating the PI3K/Akt pathway’. In: *Journal of Experimental & Clinical Cancer Research* 37.1 (Aug. 2018). DOI: 10.1186/s13046-018-0855-7.
  - [315] J. Hötzel et al. ‘Protein expression of close homologue of L1 (CHL1) is a marker for overall survival in non-small cell lung cancer (NSCLC)’. In: *J. Cancer Res. Clin. Oncol.* 145.9 (Aug. 2019), pp. 2285–2292. DOI: 10.1007/s00432-019-02989-x.
  - [316] H. Do et al. ‘TFAP2C increases cell proliferation by downregulating GADD45B and PMAIP1 in non-small cell lung cancer cells’. In: *Biol. Res.* 52.1 (July 2019). DOI: 10.1186/s40659-019-0244-5.
  - [317] Z.-J. Cheng et al. ‘THZ1 suppresses human non-small-cell lung cancer cells in vitro through interference with cancer metabolism’. In: *Acta Pharmacol. Sin.* 40.6 (Nov. 2018), pp. 814–822. DOI: 10.1038/s41401-018-0187-3.



- 
- [318] L. Ye et al. ‘Transmembrane-4 L-six family member-1 (TM4SF1) promotes non-small cell lung cancer proliferation, invasion and chemo-resistance through regulating the DDR1/Akt/ ERK-mTOR axis’. In: *Respir. Res.* 20 (May 2019). DOI: 10.1186/s12931-019-1071-5.
- [319] Y.-G. Yang et al. ‘Interferon-induced transmembrane protein 1-mediated EGFR/-SOX2 signaling axis is essential for progression of non-small cell lung cancer’. In: *Int. J. Cancer* (Dec. 2018). DOI: 10.1002/ijc.31926.
- [320] S. Feng et al. ‘TUSC3 accelerates cancer growth and induces epithelial-mesenchymal transition by upregulating claudin-1 in non-small-cell lung cancer cells’. In: *Exp. Cell Res.* 373.1-2 (Dec. 2018), pp. 44–56. DOI: 10.1016/j.yexcr.2018.08.012.
- [321] Y. Wu et al. ‘Amplification of USP13 drives non-small cell lung cancer progression mediated by AKT/MAPK signaling’. In: *Biomedicine & Pharmacotherapy* 114 (June 2019), p. 108831. DOI: 10.1016/j.biopha.2019.108831.
- [322] S. Sun et al. ‘KDM4A promotes the growth of non-small cell lung cancer by mediating the expression of Myc via DLX5 through the Wnt/ $\beta$ -catenin signaling pathway’. In: *Life Sci.* 262 (Dec. 2020), p. 118508. DOI: 10.1016/j.lfs.2020.118508.
- [323] T. Derrien et al. ‘Fast Computation and Applications of Genome Mappability’. In: *PLoS ONE* 7.1 (Jan. 2012). Ed. by C. A. Ouzounis, e30377. DOI: 10.1371/journal.pone.0030377.
- [324] V. A. Schneider et al. ‘Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly’. In: *Genome Res.* 27 (Apr. 2017), pp. 849–864. DOI: 10.1101/gr.213611.116.
- [325] D. M. Church et al. ‘Modernizing Reference Genome Assemblies’. In: *PLoS Biology* 9.7 (July 2011), e1001091. DOI: 10.1371/journal.pbio.1001091.
- [326] H. Li. ‘Tabix: fast retrieval of sequence features from generic TAB-delimited files’. In: *Bioinformatics* 27.5 (Jan. 2011), pp. 718–719. DOI: 10.1093/bioinformatics/btq671.

# *Appendices*

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

contains:

- published manuscripts
- supplementary method
- supplementary figures



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

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

---

**Hollizeck S.<sup>1,2</sup>, Wong S.Q.<sup>1,2</sup>, Solomon B.<sup>1,2</sup>, Chandrananda D.<sup>1,2,\*</sup>, and Dawson S-J.<sup>1,2,3,\*</sup>**

<sup>1</sup> Peter MacCallum Cancer Centre, Melbourne 3000, Victoria, Australia

<sup>2</sup> Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne 3000, Victoria, Australia

<sup>3</sup> 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/bi/tbucket/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 [301, 302]. 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 [3, 187, 303].

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 [178], SuperFreq [179], and Mutect2 [121], 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 [304].

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 [118] and Strelka2 [120]. 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 [181] which differentiates the likelihood of a variant between tumour and normal samples instead of imposing an absolute filter for all variants called in the normal. Hence, for each genotype (GT) at SNV sites, FreeBayesSomatic first calculates the difference in likelihoods (LOD) between the normal (Equation A.1) and the tumour (Equation A.2) samples genotype likelihoods (GL) with  $g_0$  describing the reference genotype.

$$\text{LOD}_{\text{normal}} = \max_{g_i \in \text{GT}} (\text{GL}(g_0) - \text{GL}(g_i)) \quad (\text{A.1})$$

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

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

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

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

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

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

The thresholds chosen for both LOD and VAF calculations were previously fitted by the blue-collar bioinformatics workflow for the DREAM synthetic 3 dataset using the SpeedSeq likelihood difference approach [182] 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 [183]. First, somatic variants are called from each tumour-normal pair. All detected variants across the cohort are then used as input for the second pass of the analysis where we re-iterate through each tumour-normal pair but assess allelic information for all input genomic sites.

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

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

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:

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

And therefore:

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

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

error rate of 1%. These "recovered" variants need to pass a set of additional quality metrics related to depth of coverage, mapping quality and read position rank sum score.

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

### A.3 Validation

Appendices/Variantcalling/Figure\_1.pdf

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

### A.3.1 Simulated data

We first simulated a phylogeny with somatic and germline variants from ten tumour samples and one normal (Figure A.1A, Figure 2.2A, B) (Section A.5). Germline variants were simulated at a uniform allele frequency of 0.5. Somatic VAFs were sampled from a custom distribution, modelled to favour low allele frequency variants to closely represent real world data (min VAF: 0.001; max VAF: 1; Figure 2.2C, D). Paired-end sequencing reads with realistic error profiles were simulated for WGS data at 160X average coverage using the ART-MountRainier software [184]. The simulated reads were aligned to GRCh38 and both germline and somatic variants from the phylogeny were spiked into the aligned reads using Bamsurgeon [185]. We compared the workflows for FreeBayes and Strelka2 with and without our extensions for joint variant calling on the simulated datasets. The performance of Mutect2 joint variant calling was also assessed using its proposed best practice workflow. As both Mutect2 and FreeBayes do not return a verdict for each individual sample, we needed to assign each sample in the multi-sample VCF its own FILTER value. We called a somatic variant as present in a sample, if there were at least two reads supporting it for this sample and the overall FILTER showed a "PASS", which was the same cut-off used in the refiltering step in the Strelka2-pass workflow.

While the precision of each method without our extensions was greater than 99.8%, they all missed at least 25% of all variants in the samples (i.e recall  $\leq 75\%$ ). In contrast, the recall of the modified workflows increased to  $\approx 95\%$  with only a minute decrease in the precision for both FreeBayes and Strelka2 (Figure 2.3). Mutect2 however, had virtually no change in precision, but the recall actually decreased from  $\approx 75\%$  to  $\approx 41\%$  when analysing the samples jointly (Figure 2.3B). Additionally, with our modified workflows, true positive variants were called with VAFs as low as 0.008 (median detected VAF  $\geq 0.14$  for joint sample analysis and  $\geq 0.21$  for single tumour-normal pair analysis), enabling improved distinction between true variants and technical errors (Figure 2.4). This improvement in performance for Strelka2 is only achieved after the refiltering step and not just a result of the second pass (Figure 2.5) (Section A.5.4).

The performance of joint variant calling in Mutect2 was inferior compared to all other methods (Figure 2.3A, B). This was primarily due to the "clustered\_events" filter in



Mutect2, which excluded the majority of false negative variants, with negligible contribution to the exclusion of true negative variants (Figure 2.6A, B). This result was unexpected as the simulated variants were evenly distributed along the genome and the corresponding allele frequencies were sampled randomly (Figure 2.2D).

Since the extent of the improvement in our joint calling workflows is bound by the number of shared variants between samples, we sub-sampled the simulated dataset, to show the effect of incomplete sampling on our methods, which is more likely in clinical settings. Furthermore, the evolutionary distance between the related samples in addition to the number of samples, has a major impact on the number of shared variants, as only variants acquired between the germline and the most recent common ancestor (MRCA), will benefit from the joint analysis. Therefore, we selected three sample subsets which included two, three and five samples with high evolutionary distance to show the minimum expected improvement (Figure A.1A, B). There was a clear linear improvement for both FreeBayesSomatic and Strelka2Pass when increasing the number of samples even if they had a distant evolutionary relationship. In contrast, when using only two samples with a small evolutionary distance, the increase in performance was almost as large as when jointly analysing all 10 available samples. This shows that samples with a high number of shared variants will perform better in joint calling workflows (Figure 2.7).

### A.3.2 Clinical data

To validate the performance of our new workflows, we then analysed WGS and whole-exome sequencing (WES) data of multi-region tumour samples from eight patients, with multiple tumour sites (average 7 samples per patient; total number of samples 55), enrolled in a rapid autopsy program conducted at the Peter MacCallum Cancer Centre (Table A.1 and Section A.5) [6, 186]. The published studies had multiple somatic variants from the clinical samples orthogonally validated through targeted amplicon sequencing (TAS). We used these TAS-validated variants as the gold standard to evaluate the performance of different workflows, acknowledging that the technical biases inherent to TAS data are different to those present in WGS and WES (Figure 2.8) and that there would be sampling biases depending on different tumour cells analysed in each data type.

In concordance with the results of the simulated data, our improved workflows found additional variants in all but one patient (Figure A.1D, Figure 2.9) (total additional variants Strelka2Pass: 64; FreeBayesSomatic: 85) with only a slight drop in precision for FreeBayesSomatic (mean: 0.94 vs. 0.88) and Strelka2Pass (mean: 0.97 vs. 0.92). Since the panel of variants validated by TAS was limited (7108 bp for patients CA-B through -H), this increase in detected variants suggests that a high number of shared variants in samples are missed with current approaches, which in turn leads to an overestimation of tumour heterogeneity between samples, as these variants are thought to not be present rather than undetected.

Even though the number of shared variants is a major influencing factor when jointly calling variants, low cellularity samples benefit more from the joint calling, as conventional methods cannot reliably distinguish low allele frequency variants from noise. Through a joint analysis approach, the number of recovered variants is higher in low cellularity samples, which indicates, that especially for clinical samples with variable tumour purity, joint analysis can have a major impact on improving performance (Figure A.1E, Figure 2.10).

Mutect2 in contrast, did not show significant improvement in any sample in its joint calling configuration, but showed inferior performance compared to the tumour-normal pairwise approach in two samples (Figure 2.9E), similar to its decreased performance in the simulated data (Figure 2.3). This was due to true variants being removed by the internal filters of the tool (Figure 2.6C, D). This is in stark contrast to our novel workflows, where the joint analysis preserves all called sites from the pairwise method and finds additional variants. Overall, Mutect2 found less validated variants in all patients than both Strelka2Pass (mean: 2.2) and FreeBayesSomatic (mean: 2.5) with comparable levels of precision (Figure 2.9, Figure 2.11) but longer run times (Table A.2).

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

## A.4 Discussion

Here we present an extension to two widely used variant callers, enabling them to analyse multiple related tumour samples and improve the sensitivity of detecting low allele frequency variants. This is highly relevant in clinical settings where low tumour purities in samples is a common occurrence. These workflows are an important step to satisfy the current unmet need for multi-sample tumour variant calling. While we have showcased their improvements in patient sequencing data, additional validation on larger clinical datasets is warranted to ensure the methodology performs robustly in real world settings. Importantly, these workflows are fully containerised and can be run through Janis [305] on almost any high-performance computing environment, as well as cloud services. Each workflow is highly optimised and parallelised to facilitate the analysis of the large amount of data joint variant calling requires. The workflow specification also allows the easy adjustment of parameters to enable customisation for the user's needs and priorities, whereas building an ensemble workflow using multiple callers is up to the discretion of the user (Figure A.2).

## Acknowledgements

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

## Funding

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

## Conflicts of Interest

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

## Data availability

The simulated data and the respective final variant calling files underlying this article are available from Figshare at <https://melbourne.figshare.com>, and can be accessed with <https://doi.org/10.26188/13635186> for the dataset and <https://doi.org/10.26188/13635187> for the called variants.

The biological data underlying this article are available at the European Genome-Phenome Archive (EGA) at <https://ega-archive.org>, and can be accessed with study id EGAS00001004023 and EGAS00001004950.

# Supplementary data

Appendices/Variantcalling/supp/S11.pdf

FIGURE A.2: Performance of ensemble variant calling strategies. A) Precision and B) Recall of variant detection using the joint multi-sample calling of each tool separately and compared to using Majority-vote ensemble calling (variant is called by at least two callers), Freeka2 (variant is called by both FreeBayesSomatic and Strelka2pass) and Superset (variant is called by either FreeBayesSomatic or Strelka2pass) for the simulated dataset D) Number of TAS validated variants found in the clinical samples with Majority-vote and Superset methods and the corresponding D) Precision estimates.

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

SAMPLE NAME	PUBLISHED IN	ORIGINAL NAME	SEQUENCING
CA-A-1	Solomon et al. [6]	Case 1 Left liver 1	WGS
CA-A-2		Case 1 Right occipital	
CA-A-3		Case 1 Right liver 2	
CA-A-4		Case 1 Right pleura	
CA-A-5		Case 1 Left lower lung lobe	
CA-A-6		Case 1 Left liver 5	
CA-A-7		Case 1 Right liver 3	
CA-A-8		Case 1 Left liver 2	
CA-B-1		CAS-B-21-L-LUNG	WES
CA-B-2		CAS-B-22-R-LUNG	
CA-B-3		CAS-B-14B37035-1B	
CA-B-4		CAS-B-Primary-1	
CA-B-5		CAS-B-15B08317-3A	
CA-B-6		CAS-B-14B37035-1C	
CA-C-1		CAS-A-FR07935894	WGS
CA-C-2		CAS-A-FR07935905	
CA-C-3		CAS-A-FR07935906	
CA-C-4		CAS-A-FR07935907	
CA-C-5		CAS-A-FR07935908	
CA-C-6		CAS-A-FR07935916	
CA-C-7		CAS-A-FR07935918	
CA-D-1		CAS-G-91-2	WES
CA-D-2		CAS-G-75	
CA-D-3		CAS-G-74	
CA-D-4		CAS-G-71	
CA-D-5		CAS-G-91	
CA-D-6		CAS-G-76	
CA-D-7		CAS-G-94	
CA-D-8		CAS-G-72	
CA-E-1	Vergara et al. [186]	CAS-D-70	WES
CA-E-2		CAS-D-61-3	
CA-E-3		CAS-D-66	
CA-E-4		CAS-D-68	
CA-E-5		CAS-D-64	
CA-E-6		CAS-D-61-2	
CA-E-7		CAS-D-62	
CA-F-1		CAS-C-41	WES
CA-F-2		CAS-C-40-Fresh	
CA-F-3		CAS-C-37	
CA-F-4		CAS-C-44	
CA-F-5		CAS-C-42-Fresh	
CA-F-6		CAS-C-43-Fresh	
CA-F-7		CAS-C-46-Primary	
CA-G-1		CAS-F-FR07935922	WGS
CA-G-2		CAS-F-FR07935915	
CA-G-3		CAS-F-FR07935913	
CA-G-4		CAS-F-FR07935909	
CA-G-5		CAS-F-FR07935904	
CA-G-6		CAS-F-FR07935903	
CA-H-1		CAS-E-1	WES
CA-H-2		CAS-E-3	
CA-H-3		CAS-E-4	
CA-H-4		CAS-E-10	
CA-H-5		CAS-E-6	
CA-H-6		CAS-E-8	

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

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

## A.5 Supplementary methods

### A.5.1 Alignment of clinical data

Detailed information on processing of the clinical sequencing datasets was published previously [6, 186]. Briefly, reads were aligned to GRCh38 for patient CAS-A and GRCh37 for patients CAS-B through CAS-H using BWA version 0.7.17 [306] allowing the use of alternative contigs. Reads were then marked as duplicates with Picard software (v2.17.3).

### A.5.2 Validation of clinical data

Detailed information on targeted amplicon sequencing of patient samples can be found in the original publications [6, 186]. A SNV called in WES with any workflow was considered a true positive when the adjusted p-value calculated through an exact binomial test was lower than 0.05 on the TAS data. The probability of success for this test was estimated as the number of bases different from the reference divided by the total number of sequenced bases (0.001) and the number of trials was the read depth covering the variant. For indels, a variant was considered to be validated if either of the panel variant callers primal (in house) or canary [307] called the same variant.

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

### A.5.3 Purity estimation with sequenza

For CA-A the sequenza-utils python program was used to generate input files for the sequenza R program on the aligned BAM files [15]. Kmin and gamma were set to 100 and 500 respectively to discourage a highly fragmented result. For CA-B through -H the reported tumour purities were used from the publication [186].

### A.5.4 Performance of individual steps in Strelka2Pass

As each of the three steps potentially has implications for the performance, we assessed the improvement provided by each step in the Strelka2pass workflow. Figure 2.5 shows, that there is no change in either precision or recall just by supplying variants from all tumour-normal pairs for a second round of evaluation. However, there is a >20% improvement in recall when coupling this to the refiltering step that we have built into the workflow.

### A.5.5 Ensemble workflows – user suggestions

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

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



FreeBayesSomatic (Table A.2). The usage of these workflows can be easily integrated into existing workflows and can be customised to the needs of the user.

# B

## Joint somatic variant calling - supplementary data

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

---

LISTING B.1: parse strelka VCF

---

---

LISTING B.2: annotate variants with copy number calls

---

---

LISTING B.3: convert to maf format

---

Figures/jointVariantCalling/CA-F\_schematic\_organColours.pdf

FIGURE B.1: Schematic of analysed tumour lesions in patient CA-F; Primary (diagnostic) skin sample is shown in red; metastatic sites are shown in blue; From top to bottom: right parietal lobe; left temporal lobe; right cerebellum; posterior mediastinal lymph node; left liver lobe; right liver lobe; liver, hepatic vessel; small bowel; Right side depicts three blood draws with plasma sampling (Figure 2.14)



## CASCADE - supplementary data

### C.1 supplementary methods

LISTING C.1: Preprocessing of mitochondrial reads and variants for analysis in R

---

### C.2 CASCADE - supplementary figures

#### C.2.1 Patient CA-A

TABLE C.1: List of lung cancer related genes used for variant effect prioritisation. If no source was listed, the gene is part of the “AVENIO ctDNA and Tumour Tissue extended Panel” [308], the list of commonly mutated genes in lung cancer [309, 310], which were validated through TCGA [311]. Some of the genes are also part of the targets for molecular analysis of the National Comprehensive Cancer Network guidelines for NSCLC [312].

Gene	source	Gene	source
ABL1		KEAP1	
AKT1		KIT	
AKT2		KRAS	
ALK		MAP2K1	
APC		MAP2K2	
AR		MET	
ARAF		miR-103a-3p	Fan et al. [313]
BRAF		MKRN2	Jiang et al. [314]
BRCA1		MLH1	
BRCA2		MSH2	
CCND1		MSH6	
CCND2		MTOR	
CCND3		NF2	
CD274		NFE2L2	
CDK4		NRAS	
CDK6		NTRK1	
CDKN2A		PDCD1LG2	
CHL1	Hötzel et al. [315]	PDGFRA	
CSF1R		PDGFRB	
CTNNB1		PIK3CA	
DDR2		PIK3R1	
DPYD		PMAIP1	Do et al. [316]
EGFR		PMS2	
ERBB2		PTCH1	
ESR1		PTEN	
EZH2		RAF1	
FBXW7		RB1	
FGFR1		RET	
FGFR2		RNF43	
FGFR3		ROS1	
FLT1		SMAD4	
FLT3		SMO	
FLT4		STK11	
GADD45B	Do et al. [316]	TERT	
GATA3		TFAP2C	Do et al. [316]
GNA11		THZ1	Cheng et al. [317]
GNAQ		TM4SF1	Ye et al. [318]
GNAS		TP53	
IDH1		TSC1	
IDH2		TSC2	
IFITM1	Yang et al. [319]	TUSC3	Feng et al. [320]
JAK2		UGT1A1	
JAK3		USP13	Wu et al. [321]
KDM4	Sun et al. [322]	VHL	
KDR			



FIGURE C.1: Circos plot of patient CA-A sample 26 with somatic structural variants with allele frequency  $> 0.2$ : outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.



FIGURE C.2: Circos plot of patient CA-A sample 31 with somatic structural variants with allele frequency  $> 0.2$ : outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.



FIGURE C.3: Circos plot of patient CA-A sample 41 with somatic structural variants with allele frequency  $> 0.2$ : outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.



FIGURE C.4: Circos plot of patient CA-A sample 47 with somatic structural variants with allele frequency  $> 0.2$ : outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.





FIGURE C.5: Circos plot of patient CA-A sample 55 with somatic structural variants with allele frequency  $> 0.2$ : outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.



FIGURE C.6: Circos plot of patient CA-A sample 57 with somatic structural variants with allele frequency  $> 0.2$ : outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.



FIGURE C.7: Circos plot of patient CA-A sample 59 with somatic structural variants with allele frequency  $> 0.2$ : outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.

### C.2.2 Patient CA-I

Figures/CASCADE/CA51/CA51numVars.pdf

FIGURE C.8: Number of high confidence somatic variants per sample in patient CA-I; variants were called with the Strelka2Pass workflow and retracted to *PASS* only

Figures/CASCADE/CA51/CA51phyloWithDx.pdf

FIGURE C.9: Phylogeny of samples from patient CA-I with diagnostic sample

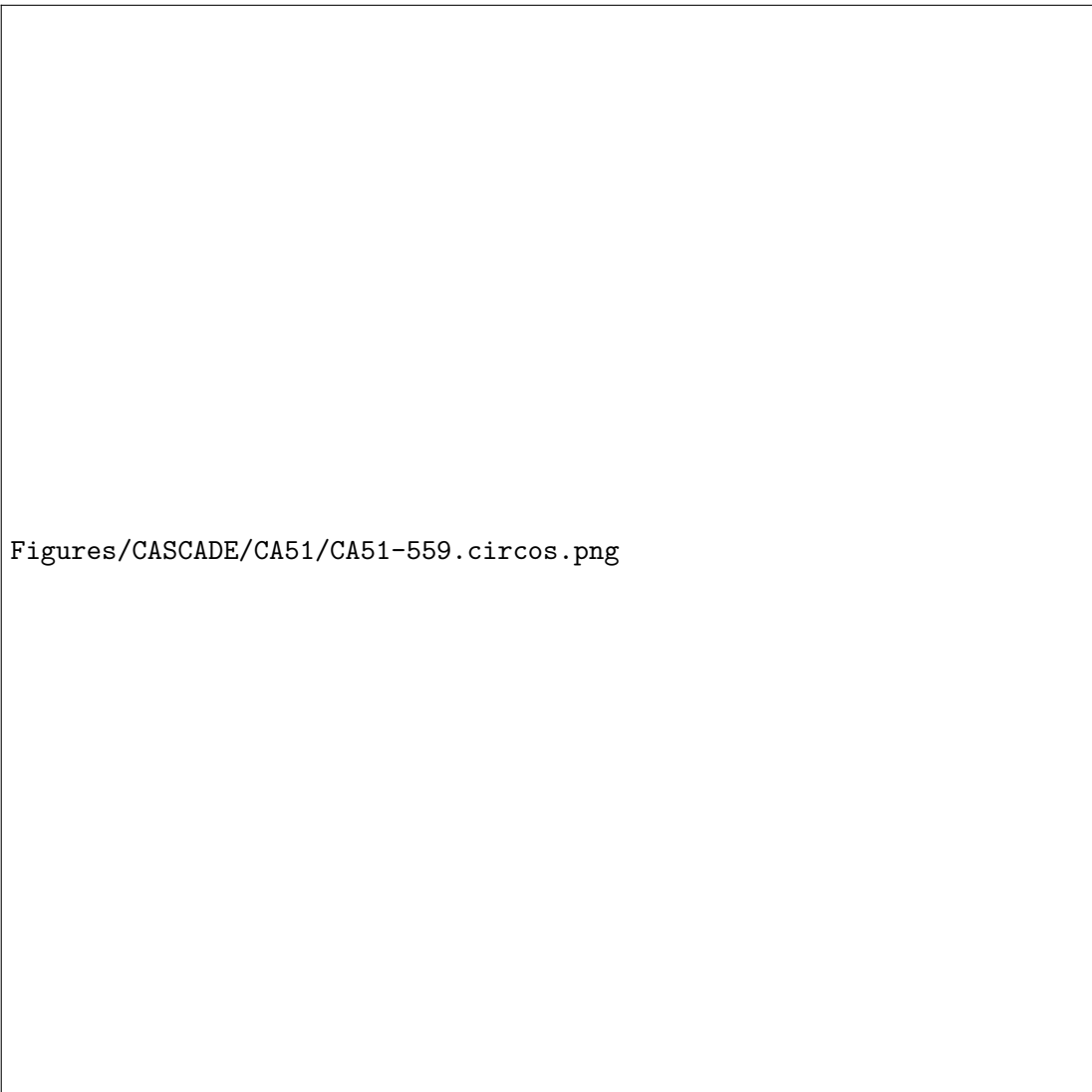


FIGURE C.10: Circos plot of patient CA-I sample 559 with somatic structural variants with allele frequency  $> 0.2$ : outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.



FIGURE C.11: Circos plot of patient CA-I sample 566 with somatic structural variants with allele frequency  $> 0.2$ : outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.

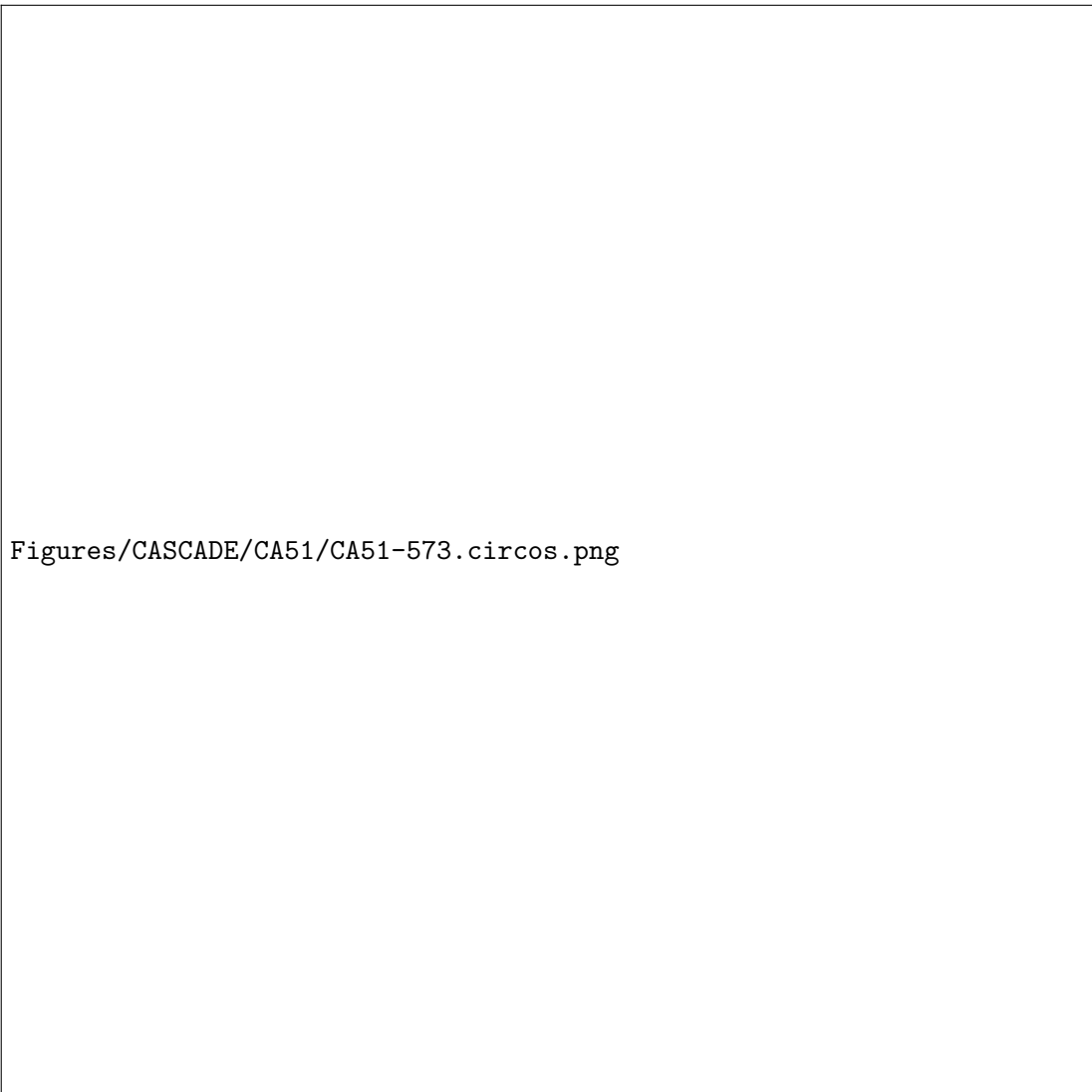


FIGURE C.12: Circos plot of patient CA-I sample 573 with somatic structural variants with allele frequency  $> 0.2$ : outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.



FIGURE C.13: Circos plot of patient CA-I sample 579 with somatic structural variants with allele frequency  $> 0.2$ : outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.



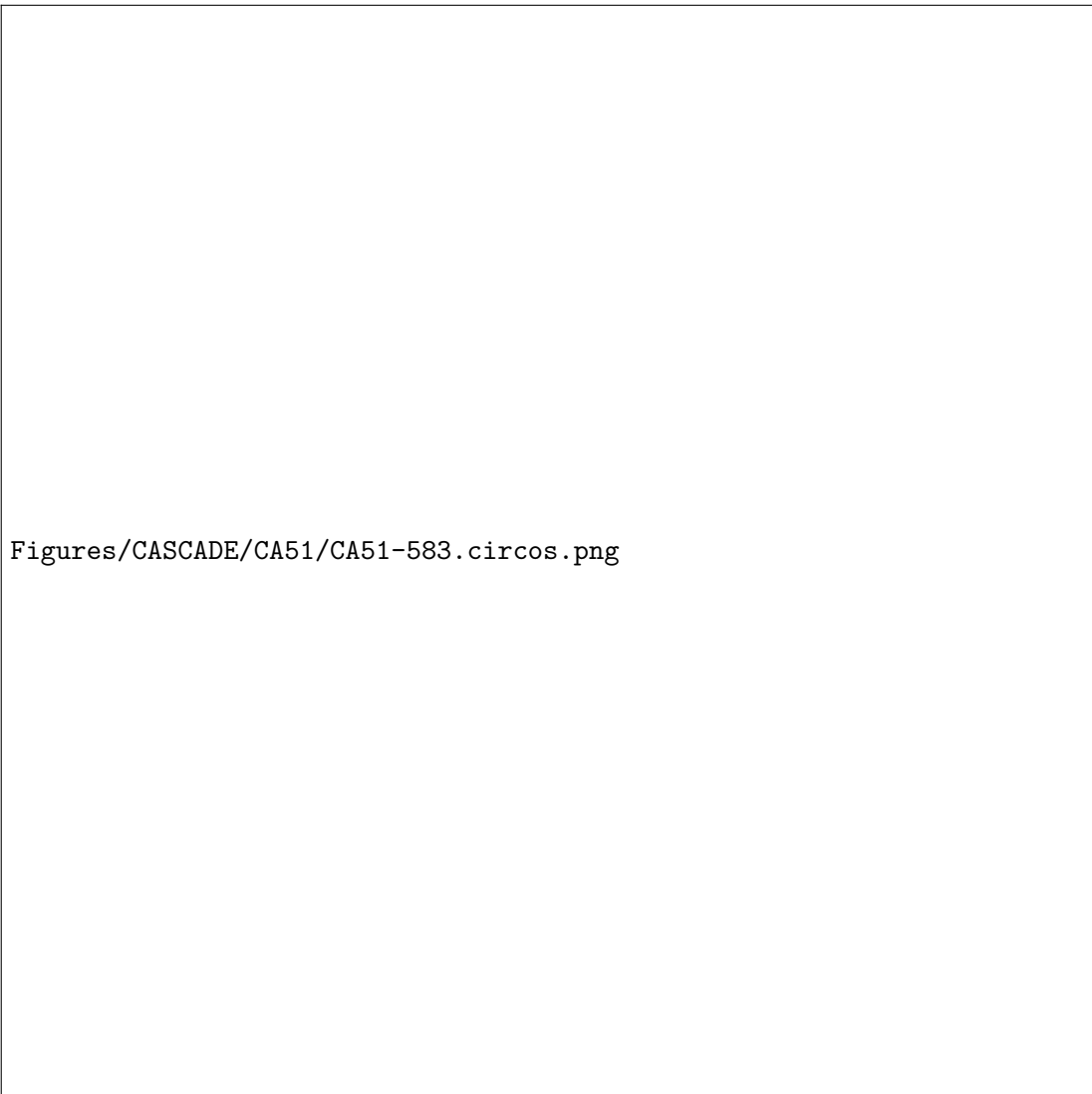


FIGURE C.14: Circos plot of patient CA-I sample 583 with somatic structural variants with allele frequency  $> 0.2$ : outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.

### C.2.3 Patient CA-J

Figures/CASCADE/CA80/CA80-24.circos.png

FIGURE C.15: Circos plot of patient CA-J sample 24 with somatic structural variants with allele frequency  $\geq 0.10$ : outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.



FIGURE C.16: Circos plot of patient CA-J sample 28 with somatic structural variants with allele frequency  $\geq 0.10$ : outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.



FIGURE C.17: Circos plot of patient CA-J sample 32 with somatic structural variants with allele frequency  $\geq 0.10$ : outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.



FIGURE C.18: Circos plot of patient CA-J sample 42 with somatic structural variants with allele frequency  $\geq 0.10$ : outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.

### C.2.4 Patient CA-K

Figures/CASCADE/CA82/CA82-4.circos.png

FIGURE C.19: Circos plot of patient CA-K sample 4: outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.



FIGURE C.20: Circos plot of patient CA-K sample 5: outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.



FIGURE C.21: Circos plot of patient CA-K sample 6: outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.





FIGURE C.22: Circos plot of patient CA-K sample 9: outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.



FIGURE C.23: Circos plot of patient CA-K sample 13: outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.

### C.2.5 Patient CA-L

Figures/CASCADE/CA86/CA86-8.circos.png

FIGURE C.24: Circos plot of patient CA-L sample 8: outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.

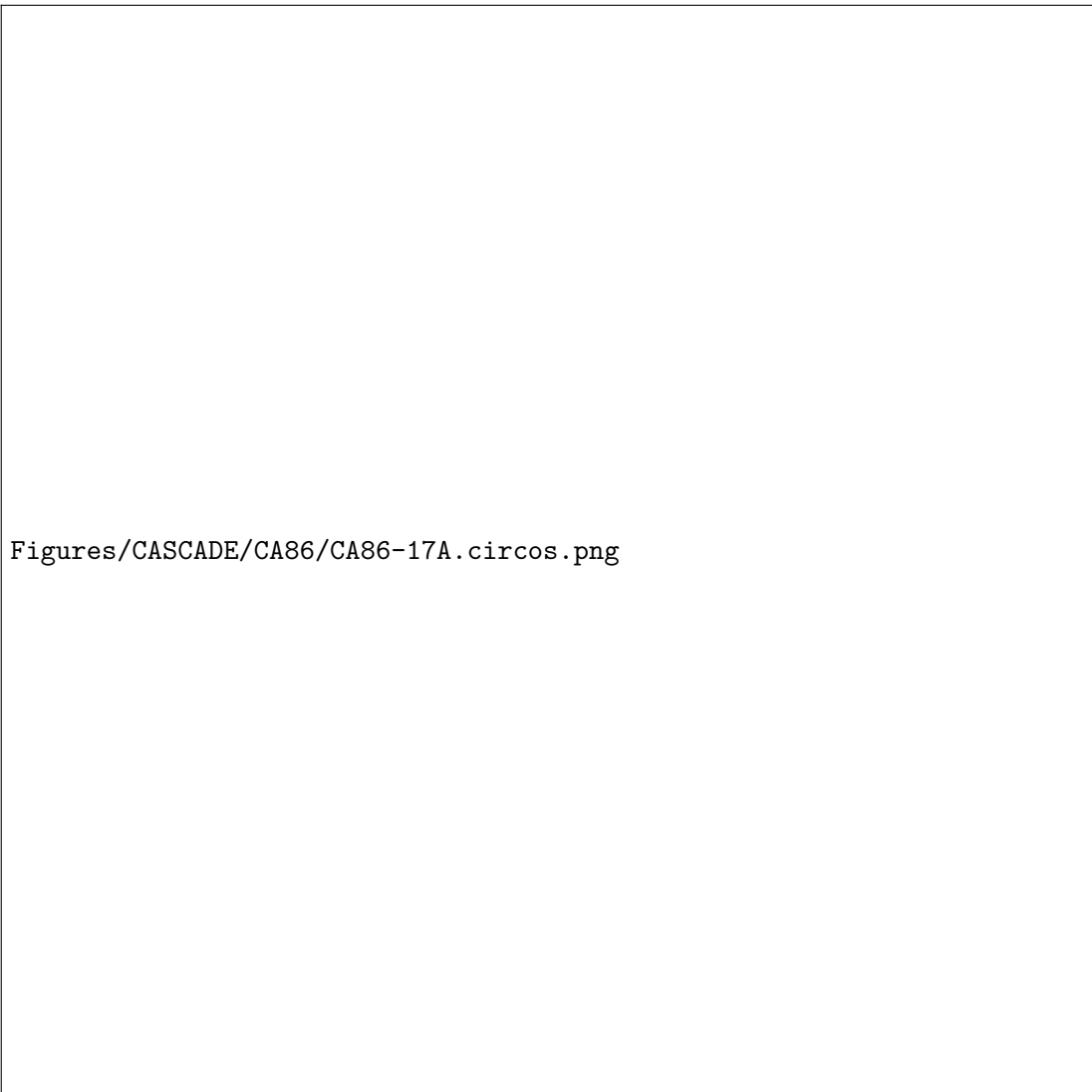


FIGURE C.25: Circos plot of patient CA-L sample 17A: outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.

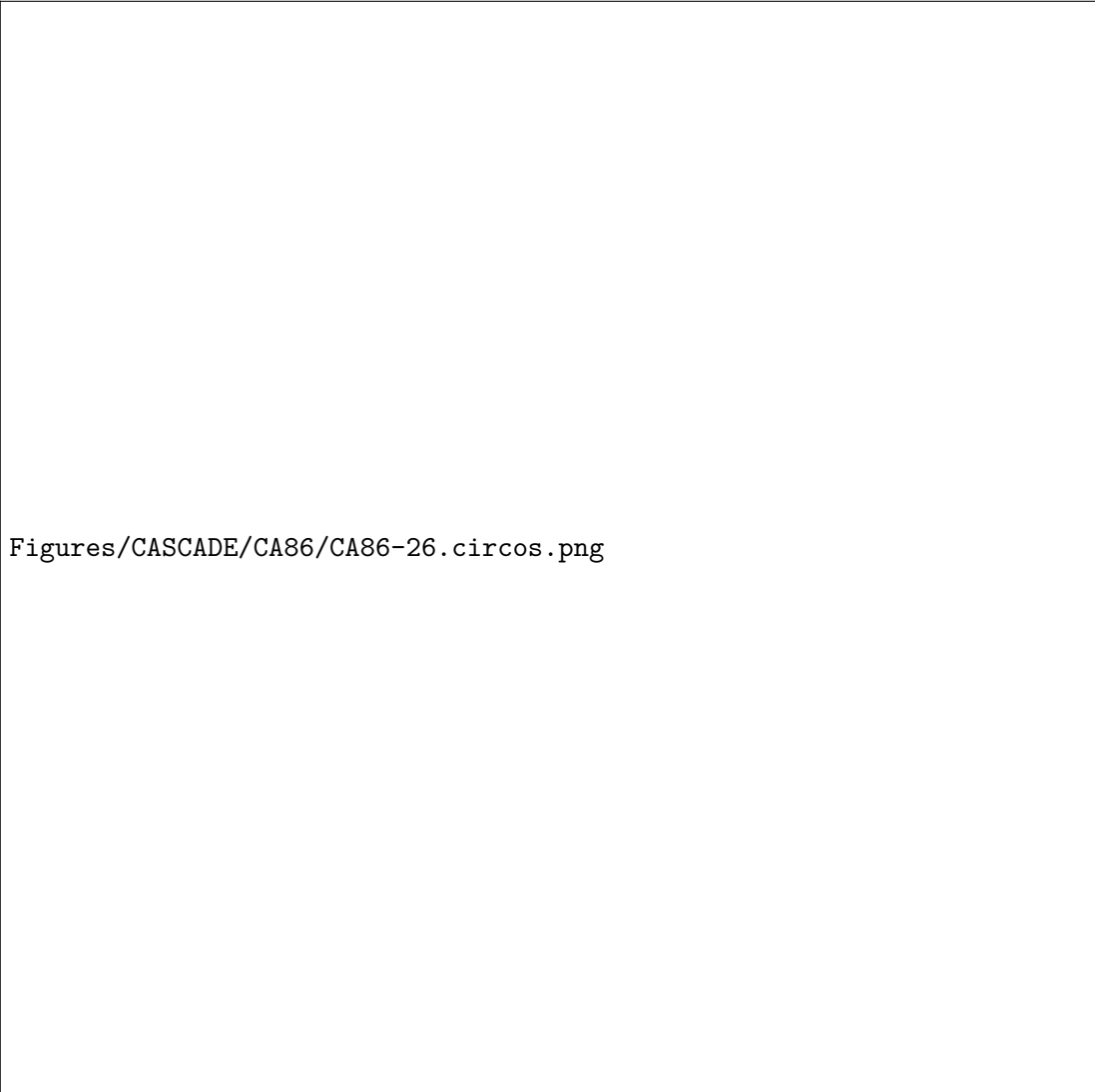


FIGURE C.26: Circos plot of patient CA-L sample 26: outer first ring shows the canonical chromosomes with gaps (centromere, heterochromatin,...) highlighted as darker areas; second ring visualises all somatic SNVs corrected for tumour purity and scaled from 0 to 1, the colour representing the base change of SNV like in Alexandrov et al. [233]; vertical lines directly under the SNVs symbolise InDels, with yellow for insertions and red for deletions; the third ring shows the total copy number alterations, with green showing a copy number gain and red a loss, dots at the outer border show a copy number greater than four; the last ring shows the minor copy number, with blue depicting a gain and orange a loss, this ring allows the detection of copy number neutral changes, like loss of heterozygosity; the center shows all structural variants: translocations in blue, deletions in red, insertions in yellow, tandem duplications in green and inversions in black.



## MisMatchFinder - supplementary methods

### D.1 ROI bed files generation

To ensure optimal mapping rates and no mapping related mismatches, the analysis was restricted to high mappability areas of the genome. These areas were defined as regions, where a k-mer of 100bp had a 85% or higher unique mappability rate. The mappability tracks were first computed with GEM [323] and then collated converted to a bed file with R just like in the best practice instructions of QDNAseq [39] for creating a new bin annotation. This method was only required for GRCh38 [324] as so far, the UCSC mappability data track was only available for GRCh37 [325].

### D.2 Oligo-nucleotide context normalisation

The ROI restriction of the analysis from Section D.1 automatically led to a different tri- and di-nucleotide context frequency in the analysed regions, than the rest of the genome, which was used to generate the original signatures [227]. For this reason, MisMatchFinder analyses the oligo-nucleotide composition of the analysed regions and generates weighted counts by adjusting for the differences.

The baseline frequencies of both di- and tri-nucleotides were generated with the function *oligonucleotideFrequency* from the “Biostrings“ library [36] using the hg38 BSgenome [38]. The raw counts of the di- and tri-nucleotides can be seen in Table D.1 and Table D.2 respectively.

### D.3 Germline filtering with zarr

As shown in Figure 4.4A, the amount of mismatches found in a 10x coverage sample can easily exceed 3 million. In addition to that, the current gnomAD database contains  $\approx 707$  million variants. This means a normal merge for two datasets based on chromosomal position is not feasible for a normal compute resource in a acceptable

TABLE D.1: Dinucleotide counts generated with Biostrings [36] for GRCh38

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

TABLE D.2: Trinucleotide counts generated with Biostrings [36] for GRCh38

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



time frame. To allow an easy query of mismatch positions against the full database, a zarr [58] representation of the gnomAD VCF was generated. However in contrast to the out of the box indexing function shipped with scikit-allel [273] which was used to convert the vcf to zarr, the program uses its own index built with ncls, which is available through PyRanges [54]. The sections below outline first the conversion process with scikit-allel (Section D.3.1) and then details the filtering in the MisMatchFinder program (Section D.3.2)

### D.3.1 Zarr conversion with scikit-allel

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

Importantly, tabix [326] can be used to split the conversion into multiple parts by restricting the process to specific regions.

Listing D.1 shows the code used to convert chromosome ‘chr1’ from the downloaded gnomad vcf

---

LISTING D.1: scikit-allel conversion vcf\_to\_zarr

---

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

---

LISTING D.2: field options for reduced memory

---

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

### D.3.2 MisMatchFinder filtering - the zarr API

### D.3.3 Data simulation

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

### D.3.4 Signature simulation - we can spike this punch

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

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

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

$$n(\text{variants}) = \frac{\text{mutation rate}}{1 \cdot 10^6 \cdot \text{genome length}} \quad (\text{D.1})$$

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

---

LISTING D.3: spike-in variant selection

---

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

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

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

---

LISTING D.4: bamsurgeon spike-in

---

### D.3.5 Blacklist generation from healthy samples

---

LISTING D.5: Blacklist postprocessing

---

### D.3.6 Patient data subsampling

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

# MisMatchFinder - supplementary figures

Figures/MisMatchFinder/SBS3Signature.pdf

FIGURE D.1: Trinculeotide count contributions for SBS signature 3 (Defective homologous recombination-based DNA damage repair); values taken from Alexandrov et al. [227]

Figures/MisMatchFinder/SBS7SpikeInSignatureDifferences.pdf

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

Figures/MisMatchFinder/SBS3SpikeInSignatureDifferences.pdf

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

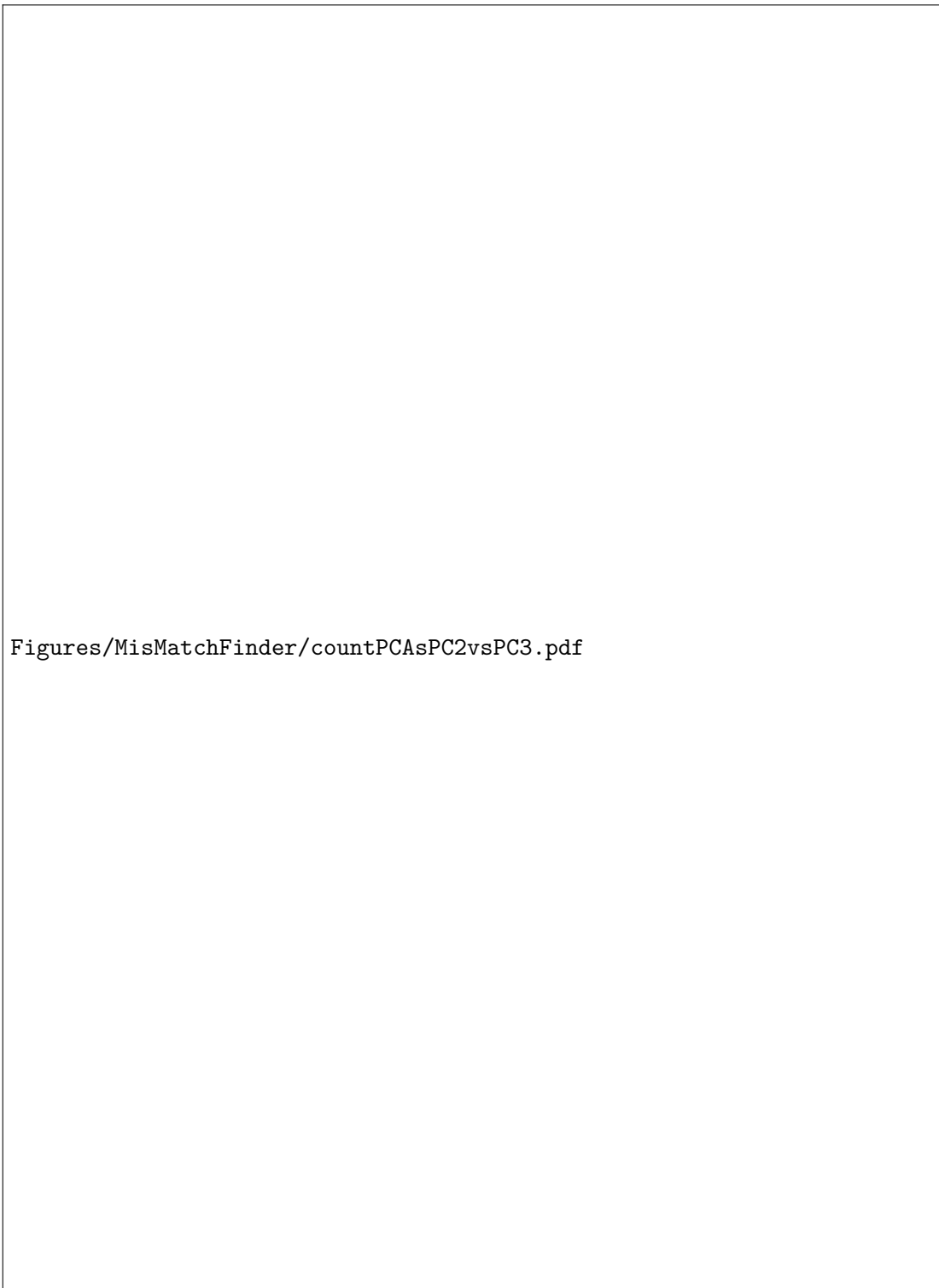


FIGURE D.4: PCA (PC2 and PC3) of tri-nucleotide mismatch counts of healthy donor and tumour samples (melanoma and metastatic breast cancer) of varying purity; PCA was conducted on scaled and centered data

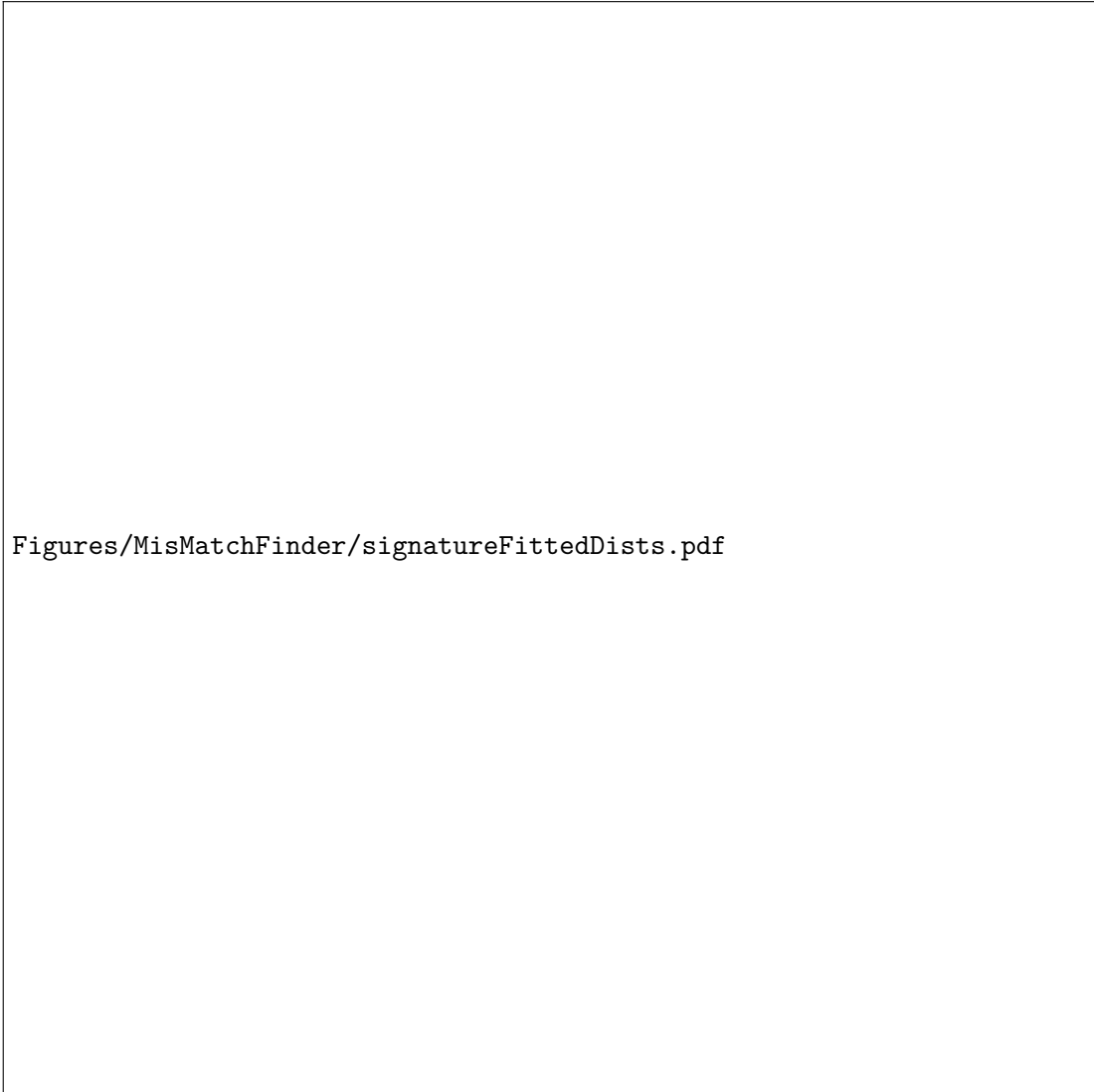


FIGURE D.5: Fitted beta distribution for Signature SBS3 in healthy samples



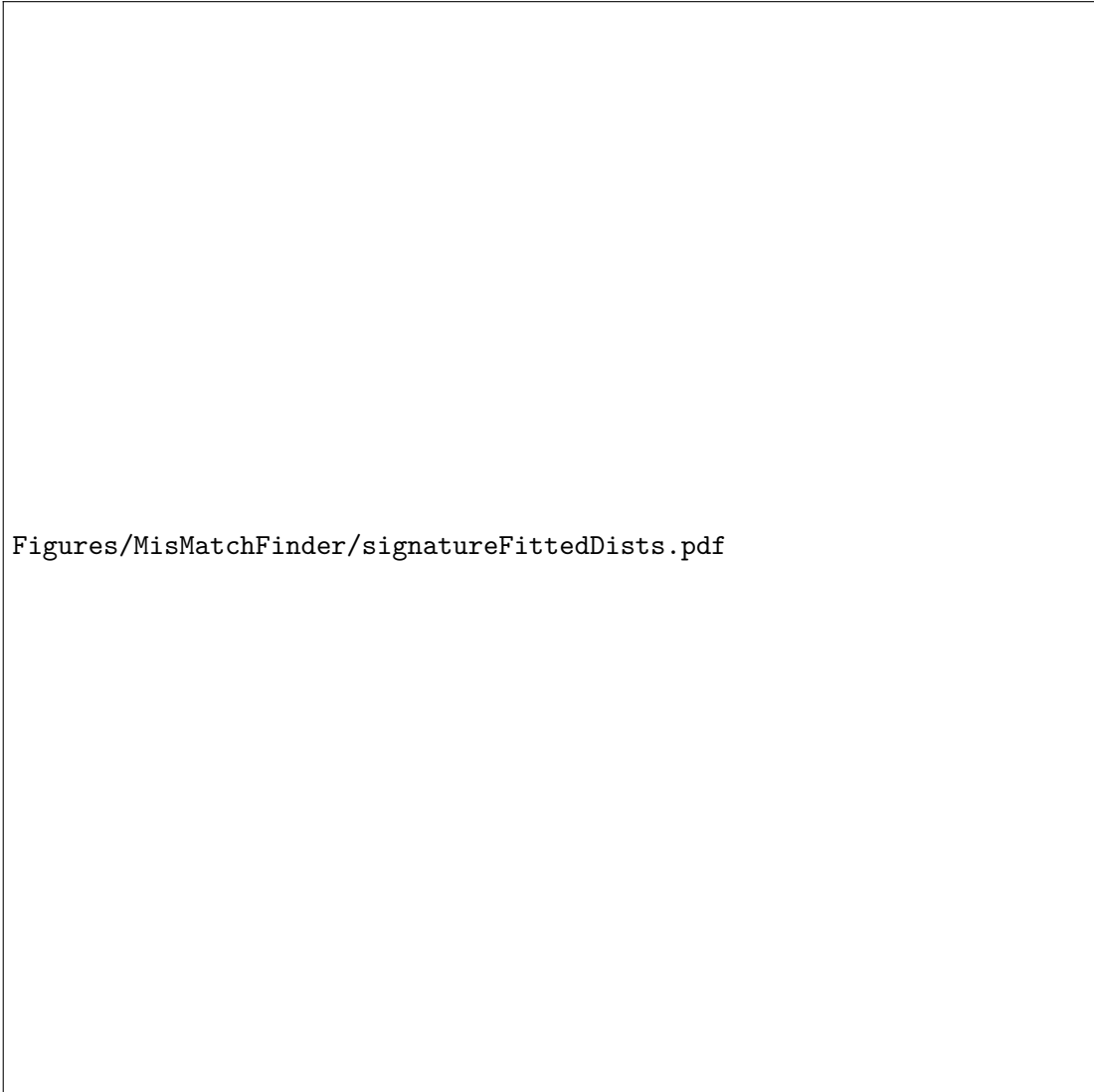
FIGURE D.6: Fitted beta distribution for Signature SBS17a in healthy samples





Figures/MisMatchFinder/signatureFittedDists.pdf

FIGURE D.7: Fitted beta distribution for Signature SBS12 in healthy samples



Figures/MisMatchFinder/signatureFittedDists.pdf

FIGURE D.8: Fitted beta distribution for Signature SBS46 in healthy samples



FIGURE D.9: Signature detection of variants categorised by presence in gnomAD: All pie charts show signature deconvolution results with clinically relevant signatures coloured; blue (APOBEC), red (UV exposure), orange (tobacco), purple (chemotherapy), light grey (sequencing artefacts), dark grey (everything below 1% weight); left column shows results of somatic variants not found in gnomAD, right column shows results for somatic variants found in gnomAD