Circulating tumour DNA for precision medicine in Non-small cell lung cancer

by

Sebastian Hollizeck

ORCID: 0000-0002-9504-3497

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

in the
Department of Mad Science
Melbourne School of Awesome
THE UNIVERSITY OF MELBOURNE

September 28, 2021

THE UNIVERSITY OF MELBOURNE

Abstract

Department of Mad Science
Melbourne School of Awesome

Doctor of Philosophy

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

The Thesis Abstract is written here (and usually kept to just this page). The page is kept centered vertically so can expand into the blank space above the title too...

Declaration of Authorship

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

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

Signed: _____

Date: _____

Preface

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

Solomon B.J[1]., Tan L.[1], Lin J.J.[1], Wong S.Q.[1], **Hollizeck S.**[1], 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

Hollizeck S., Wong S.Q., Solomon B., Chandrananda D.[2], Dawson S-J.[2] "Custom workflows to improve joint variant calling from multiple related tumour samples: FreeBayesSomatic and Strelka2Pass" *Bioinformatics.* 2021. DOI: 10.1093/bioinformatics/btab606

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

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

Chapter 2: Joint somatic variant calling is an original work describing two workflows for the joint analysis of multiple related tumour samples and has been published in *Bioinformatics* as "Custom workflows to improve joint variant calling from multiple related tumour samples: FreeBayesSomatic and Strelka2Pass" on 21^st September 2021. In addition to the published analysis, I have added longitudinal analysis and its evaluation.

Contributions for this chapter:

- I conceptualised the work
- I implemented the workflows and containerised all required tools
- I performed the data simulation
- I performed the analysis presented in the publication
- I wrote the draft of the manuscript and performed revisions
- D.C. and S-J.D. provided advice in planning and writing the manuscript
- D.C. provided guidance for method development
- S-J.D. provided guidance for method evaluation
- S.W. performed the targeted amplicon validation
- S.W. and B.S. read the draft version and provided feedback
- B.S. provided clinical expertise for human data

Chapter 3:

summary plus contributions

Chapter 4:

Chapter 5:

Funding:

All necessary funding goes here

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

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

Acknowledgements

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

Contents

List of Figures

List of Tables

Abbreviations

DNA	**D**eoxyribo**n**ucleic **A**cid
RNA	**R**ibo**n**ucleic **A**cid
cfDNA	**c**ell **f**ree **DNA**
ctDNA	**c**irculating **t**umour **DNA**
bp	**b**ase **p**air
ChIP	**Ch**romatin **I**mmuno**P**recipitation
WGS	**W**hole **G**enome **S**equencing
WES	**W**hole **E**xome **S**equencing
SCLC	**s**mall **c**ell **l**ung **c**ancer
NSCLC	**n**on-**s**mall **c**ell **l**ung **c**ancer
RAID	**R**edundant **A**rray of **I**ndependent **D**isks
SNP	**S**ingle **N**ucleotide **P**olymorphism
InDel	**In**sertion or **Del**etion
SV	**S**tructural **V**ariant
PON	**P**anel **O**f **N**ormals
GATK	**G**enome **A**nalysis **T**ool**K**it

Constants

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

Symbols

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

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

— Lewis Carroll, *Alice in Wonderland*

1

Introduction

This first introduction chapter contains all the necessary background information as well as an overview for the work discussed in this thesis. It summarised basic biological properties of DNA and cell biology as well as the respective technologies to read, analyse and measure these biological concepts and then how to evaluate the output of these methods. Section 1.1 delineates the role DNA plays for the cell and then section 1.2 shows how these standards are changed in the tumour and cell free context. Section 1.3 introduces the current technologies used to measure and detect DNA and its variations. With section 1.4 covering the computational analysis methods to read out changes in the DNA. Then section 1.5 relates how these changes lead to cancer and what we can learn from them. The introduction concludes with section 1.6 as an overview over the thesis aims and my work in addressing them in the following chapters.

1.1 DNA as a information storage unit

It is a widely accepted fact, that Deoxyribonucleic acid (DNA) serves as the long term information storage molecule of our cells. This information is protected and allows correction of simple errors through its double helix structure [1, 2]. The nucleotides, which consist of a deoxyribose sugar (hence the name), a phosphate group and the nitrogenous base, are joined together by phosphate groups. Even though there are six common naturally occurring nitrogenous bases: Adenine (A), Thymine (T), Guanine (G), Cytosine (C), Uracil (U) and nicotinamide, only the first four are used to encode the genetic information into DNA. Each of the strands mirrors the other, so that an adenine

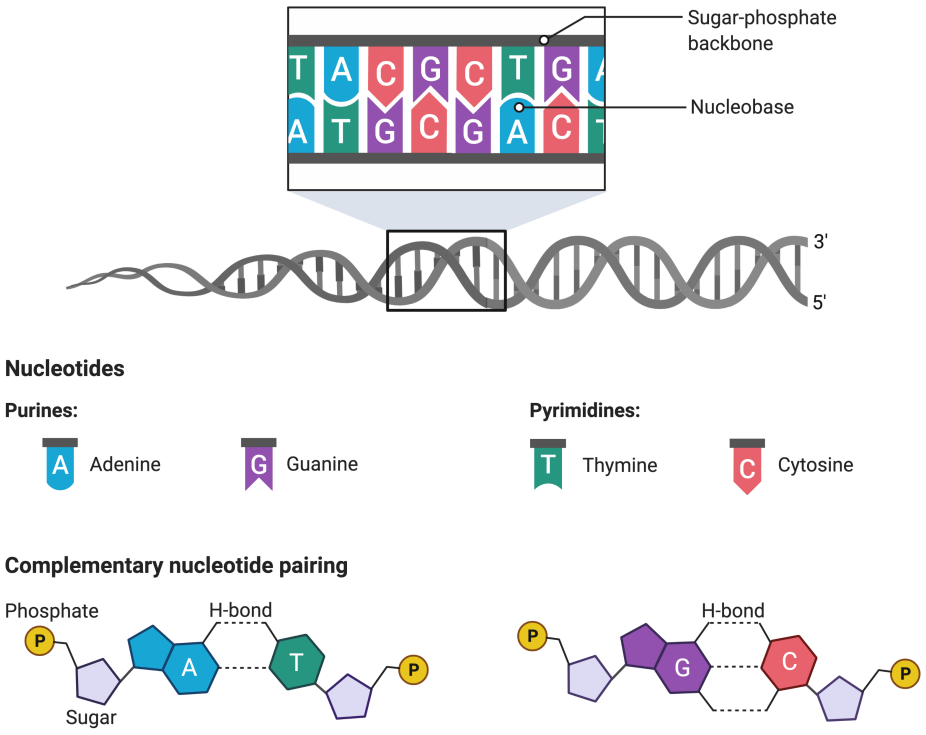

FIGURE 1.1: Overview of DNA structure and the nucleobases, which form DNA strands. Nucleotides are split into Purines and Pyrimidines by the structure of the nitrogen ring; complementary pairing of bases is shown as shapes of the bases as well as with 2D structures; Hyrdogen (H) bonds are shown as dotted lines; Phosphates are shown as P; 3' and 5' ends are defined by the internal number of the carbon atom of the sugar which is exposed

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

The entirety of the DNA encoding the organism is commonly called "the genome" with all genes, which consist of introns and exons are called exome. Unicellular organisms usually only have a very small amount of introns, which to current knowledge only

provide limited information and are only responsible for structure. In vertebrates introns as well as intergeneic DNA (the DNA between genes) contribute most of the DNA in the genome. For example in humans, only 1% of the genetic material is considered to be exonic, whereas introns contribute ≈ 24% and the rest is intergeneic (≈ 75%)[4].

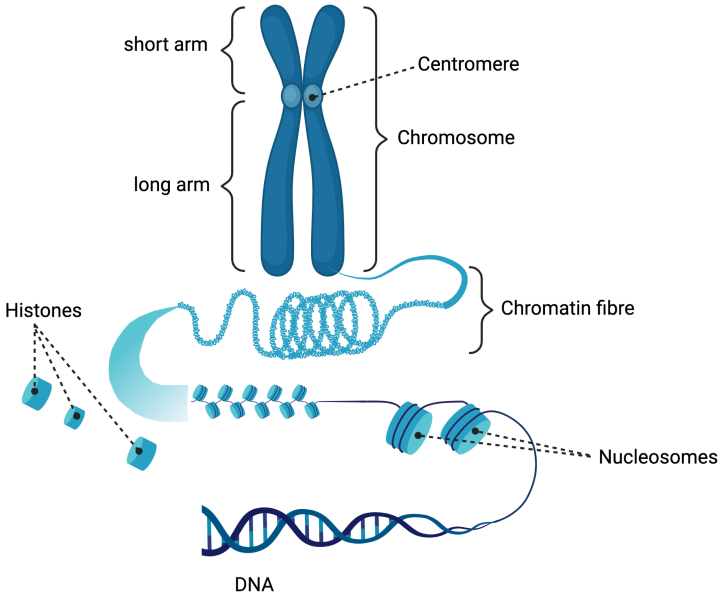

FIGURE 1.2: Structural overview of the metaphase condensed chromosome: DNA is first wrapped around Histones to form nucleosome, which then associate with each other to form the chromatin fiber, which in the metaphase of the cell cycle is condensed even more into the X-shaped chromosome

The DNA in eukaryotes however is not free floating around in the nucleus of a cell, but rather in most eukrayotic organisms, it is highly condensed and structured, first wrapped around nucleosomes like thread on a spool, then organised around histones, into either open (accessible) or closed chromatin, which then can be even further condensed into chromosomes, which have a X-like shape, with two shorter and two longer arms (Figure 1.2). This allows some of the DNA to be accessible where the use of other areas can be restricted[6]. Through this restriction, the availability of certain genes, which are the sections of the DNA, which encode for short term storage molecules like RNA. This restriction plays an important role in cell fate and cell viability. Ultimately all information stored to create a new highly complex organism is stored in just the DNA of one cell. Whichever parts are used and how they are used decides the function and the identity of the cell[7].

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

Similar to the already discussed RAID-like setup of the DNA in two strands, another concept of data security, a spatial different storage is also implemented. Most eukaryotic organisms have at least two of each chromosome (diploid) with some species reaching up to septaploid[8]. However, this concept is not the only reason for the ploidy of somatic cells. For sexually reproducing organisms, at least a diploid set of chromosomes is necessary to enable information to be joined from both parents. Germline cells (sperm and egg) are generally monoploid, such that the resulting cell will be diploid, but the ploidy of the somatic cells is not as uniform within a species, where it can vary between organisms based on gender or rank [9]. In most organisms, a change in ploidy is fatal [10] and only partial ploidy changes like extra copies of chromosome 17 [11], chromosome 18 [12] and chromosome 21 [13] are tolerated. These syndroms can occur when the is an uneven split of chromosomes during cell division. The additional advantage, apart from sexual reproduction, is that a second almost identical copy of a chromosome allows repair of DNA, even when both strands are damaged, for example in a double strand break. In this case, the information from the sister chromosome will be used, by first cutting the double strand break ends to have overhang (resection). This overhang will then merge with the sister chromosome's mirrored strand. In this state, the two chromosomes are fused together in a Holliday junction, which allows the missing part from the resection and the double strand break to be synthesised [14]. During this process, which is part of the homology directed repair (HDR) machinery, the sister chromosomes exchange parts of their DNA, when resolving the Holliday junction. As these stretches of DNA do not need to be 100% identical, this plays and important role in evolution and diversity [15, 16].

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

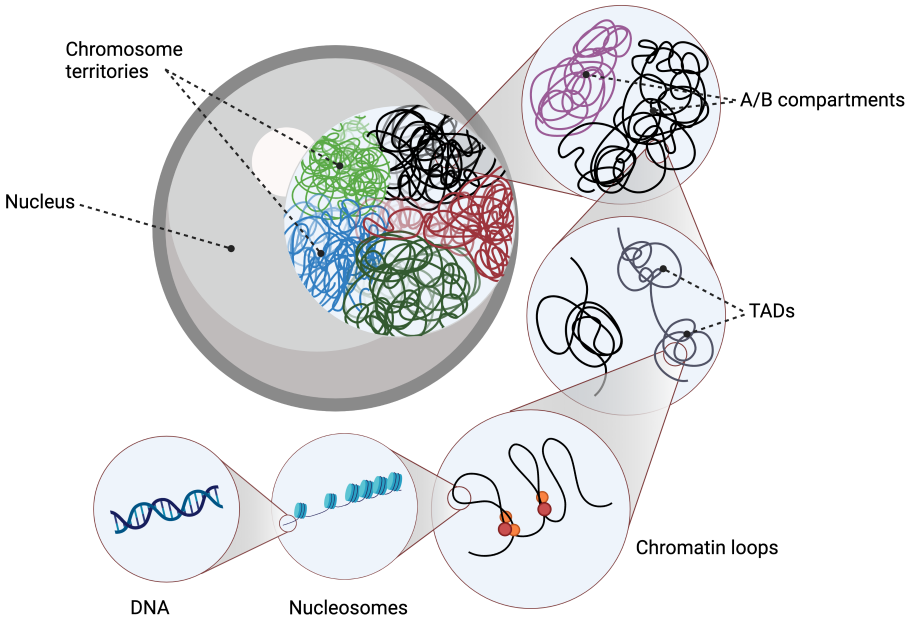

FIGURE 1.3: Individual chromosomes occupy a subspace in the nucleus called chromosome territories. Chromosome territories can be further partitioned to distinct A and B compartments, which are enriched for active and repressed chromatin, respectively. Genomic regions within topologically associating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

1.1.2 Phantastical mutations and where to find them

However even though the DNA is highly stable and error correction methods are constantly working to not introduce any changes in the DNA, the source of evolution and adaptation of species is sourced in a steady mutation rate [18, 19]. These changes in normal tissue are mostly irrelevant to the organism as a whole and will not be passed on to the next generation. These changes are known as somatic mutations. This type of mutation accumulates in a cell linearly over the course of the lifespan of the cell and is not bound to just cell divisions[20, 21]. In contrast, if one of those mutations occurs in the germline cell, eg. sperm or egg producing cells, these mutations will be propagated to all offspring and be present in all cells of that organism and in term all its offspring. These mutations are called germline mutations. These mutations are also called germline variants, as they establish in the population and represent a variation of the organism. Mutations can also be classified depending on either their size ranging from single nucleotide polymorphisms (SNPs) over small insertions or deletions (InDels) to large structural changes, like the deletion of parts of or even a whole chromosome

arm. like previously described with ploidy changes, usually smaller changes have less impact on the overall fitness of the organism, however even SNPs can lead to changes which are not compatible with life[22, 23].

1.2 Cell free DNA is more than just bits and pieces

When a cell from a multicellular organism dies, through which ever method, there will be many different enzymes involved, which clear the debris and recycle material. This means that proteases digest proteins into amino acids, which will later be used for either building new proteins or possibly even digested further for energy production. The same happens with the DNA in the cell. However as discussed in the previous section 1.1 the DNA is wrapped around histones and organised in structures called nuclesomes. These protect the DNA from being cut by DNAases by hindering the access to the DNA, similar to how they stopped the access for transcription into RNA. This then in turn leads to the DNA being cut into pieces mainly in the length of 167 base pairs (bp). These DNA fragments, which are called cell free DNA (cfDNA), can then be detected in bodily fluids, like blood or even stool. By analysing these fragments, non invasive tests for prenatal care have been possible, as the DNA of the foetus is detectable in the mothers blood [24, 25]. Similar to the process, a cancer also sheds DNA, titled circulating tumour DNA (ctDNA), when its cells die, either through intervention of the immune system or through other forcefull processes. These ctDNA fragments can also be analysed and molecular properties measured, without even knowing the exact location of the tumour. As a blood test can be routinely performed in the clinic or even a general practitioner, the monitoring of cancer progression is significantly easier and safer than through other measures. Of course it is, similar to the prenatal test, only a proxy for the cells which are still alive, as these have not shed their DNA. Additionally the amount of shedded DNA is highly variable between tumours, with a general higher amount for later stages, so that sometimes there is almost no ctDNA present, even though the cancer is fairly advanced [26, 27].

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

As we know the building blocks, that make DNA as well as the process and the enzymes responsible, we can synthesise DNA in vitro. By chemically modifying the nucleotides supplied to the synthesis process, the sequence of the copied strand can be analysed. The first method to make use of this used the lambda phage to fuse known ends for the primers needed for the reaction to the piece of DNA and supplied labelled nucleotides [28].This method was then superseded by "Sanger sequencing" after Frederick Sanger who with colleagues published this method in 1977, by adding **di**deoxynucleotides in a low concentration, the polymerase chain reaction would terminate trying to integrate these nucleotides and by labelling them radioactively or flourecently, a gel can be used to determine the sequence of a piece of DNA[29, 30], which made the method better suited for larger scale projects.

However this method has multiple issues for modern research questions. Mostly, that it is fairly labour and time consuming to analyse multiple pieces of DNA at the same time and it is very challenging to sequence all the DNA of an organism. The human genome project, which was started in 1990 used machines which automated the Sanger sequencing procedure and it still took hundreds of researchers 13 years to complete the DNA sequence of just one human [31, 4]. Even though this was a very long project, it laid the ground work for the usage of the current sequencing technologies.

1.3.1 Library preparation - what we learned from using phages

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

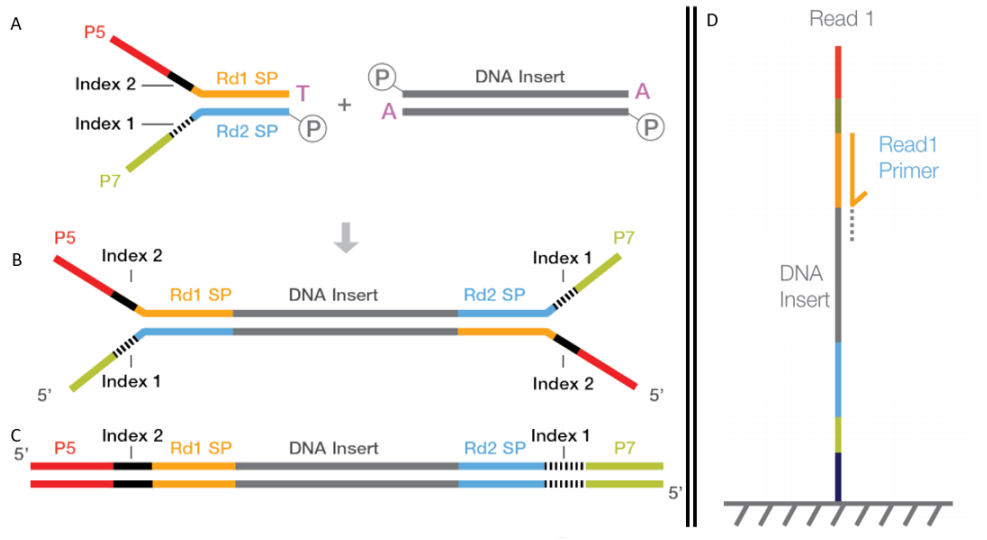

FIGURE 1.4: Adapter ligation during library preparation. The adapters are added to the DNA insert during library preparation. A. The DNA insert is prepared by adding an A-tail and phosphorylation. B. The adapter complex which includes the P5/P7 flow cell binding adapter is added to the DNA insert. C. The DNA insert is ready for sequencing. D. The DNA insert binds to the flow cell for sequencing. Primers bind to the DNA insert to generate reads;
Figure adapted from "How short inserts affect sequencing performance"[32]

complex is ligated. This enabled the DNA to bind to the flow cell which is covered with the reverse complement of the adapter (Figure 1.4).

1.3.2 Next generation sequencing

Next generation sequencing (NGS)is the coined term for basically any standard high-throughput sequencing performed, which includes exome, genome, transcriptome, protein-dna interactions (ChIP) and other epigenome studies. The term NGS is still widely used, even though it has been more than 10 years since the first NGS approach was commercially available. While in the beginning of next generation sequencing there were multiple approaches, the current lion share (80% of sequencing data) of protocols use the Illumina short read sequencing by synthesis approach (Figure 1.5)[33, 34], which is based on the concept of alternating integration of florescently labelled nucleotides and imaging with a microscope (Figure 1.5) as well as multiplexing, where a DNA fragment is ligated to an index, which allows the sequencing of multiple samples at once [35, 36] as it is shown in Figure 1.4. This method allows highly accurate determination of the sequence of a DNA fragment and depending on the flow cell and sequencing machine allows to sequence a whole genome in just 24h.

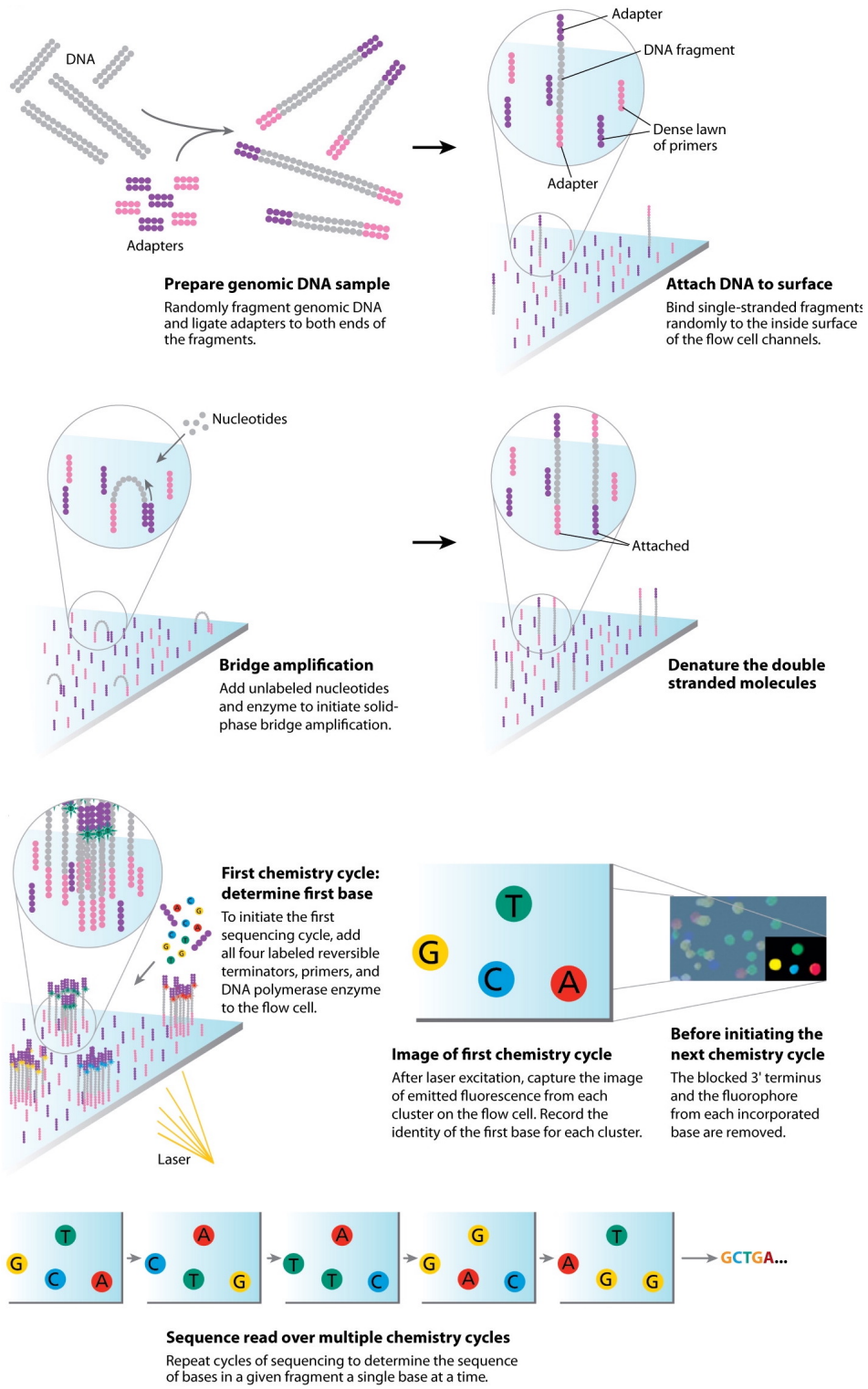

FIGURE 1.5: The Illumina sequencing-by-synthesis approach. Cluster strands created by bridge amplification are primed and all four fluorescently labeled, 3'-OH blocked nucleotides are added to the flow cell with DNA polymerase. The cluster strands are extended by one nucleotide. Following the incorporation step, the unused nucleotides and DNA polymerase molecules are washed away, a scan buffer is added to the flow cell, and the optics system scans each lane of the flow cell by imaging units called tiles. Once imaging is completed, chemicals that effect cleavage of the fluorescent labels and the 3'-OH blocking groups are added to the flow cell, which prepares the cluster strands for another round of fluorescent nucleotide incorporation; Figure adapted from Mardis[33]

1.3.3 Long read sequencing - the "third" generation sequencing

By now, multiple methods which broke free of the size limitations of NGS exist, which are commonly referred to as long read sequencing. Most of the current methods trade the very high accuracy of the second generation NGS methods for the capability of sequencing of sequencing huge continous strands of DNA (current record 2.3 Million bp [37]) with normal library preparation ranging between 10-30 Kbp. These methods are expected to revolutionise our understanding of the highly repetitive elements that exist in the genome, such as the centromeres of chromosomes. Methods such as the direct molecule sequencing approach by Oxford Nanopore are even able to distinguish post transcriptional modifications on RNA[38]. So far, these methods however are still very expensive and as this work is dealing with ctDNA, which is highly fragmented, these methods offer only limited advantages over the short read sequencing, while being much more expensive.

1.4 DNA analysis- what to do with the sequence

The types of analysis that can be done with the output from the sequencing machine stretches far, however, all methods need to first infer the location in the genome, the sequenced piece of DNA originated from. As the current methods randomly fragment the DNA (subsection 1.3.1), the genomic location information is completely lost. This process is referred to as mapping.

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

In this process, the fragments of DNA, which were sequenced, are assigned a genomic coordinate on the reference genome. This is only possible, due to the fact, that we have a resolved genome sequences (section 1.3) for a high number of species. The location a sequenced piece of DNA fits to the reference genome might be unique, but it could also fit to multiple locations, due to highly repetitive regions or due to the existence of pseudo genes with almost 100% identify. In addition to this, the reference genome might not accurately reflect the genome of the organism that has been sequenced. Each mapping position is therefore assigned a quality score, which reflects how likely it is the actual

position of the sequence. As Illumina sequencers have the ability to sequence both ends of the DNA fragment, the position of the ends (read 1 and read 2) to each other can also be used to infer the quality, as they should be within a reasonable distance to each other (Figure 1.4)

As this process is time consuming and the exact location of the fragment might not be as important, there exists a subset of tools called pseudo-mapper, which are based on k-mers, which are predefined DNA sequences of length k, which help to identify certain regions of interest. These tools are especially common for RNAseq, where the exact location of a read doesnt matter, only that the read is within a gene [39, 40], but also for methods that estimate similarity between sequences (DNA, RNA or protein) [41, 42].

For this work however, the exact position of reads is important, so only real mapping methods like BWA [43] or Bowtie 2 [44], which are optimised for short reads from Illumina systems, provide the necessary functions.

add things about alternative contigs and reference genome?

1.4.2 Variant calling - spot the difference

As intra-species genetic variation is intended for adaptation and evolution, there will be places where the DNA sequence of the subject will differ from the sequence of the reference (see subsection 1.1.2). These variants give insight into medical background as well as treatment options for patients and can even be used to guide family planning. Depending on the type of variation that is of interest, a different set of computational methods are needed, as germline and somatic variants have different properties.

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

The most common source of DNA used for germline variant analysis is the mono nuclear layer from the blood of the subject, but really almost any cell can be used for this process, as all cells in the organism will share all germline variants (subsection 1.1.2). The only important input on top of the DNA sequence from the sequencer are the reference genome of the organism as all variant nomenclature is based on the reference and the ploidy of the organism (subsection 1.1.1). The ploidy is important to infer,

at which ranges of allele frequency a variant can biologically occur. For example in a human diploid genome, germline variants can occur either in one or both chromosomes, which mean we assume reads should show an allele frequency of around 50% and 100%, where the hexaploid commercial wheat [45] allele frequency for variants would be 16%, 33%, 50%, 66%, 0.83% and 100%. Due to the random sampling and possible sequencing errors, however the observed allele frequencies will differ. Most state of the art germline variant calling method will also use haplotype reconstructions through de-Bruijn graphs, which features a remapping of reads in relation to each other [46, 47, 48, 49, 50]. These

1.4.4 Somatic variant calling - life is ever changing

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

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

The ideal scenario for tumour somatic variant calling is when a biopsy of the tumour as well as a normal sample of the patient is available. In most clinical cases, this will be the diagnostic biopsy as well as the mono nuclear layer just like for germline calling (subsection 1.4.3). These two samples are then analysed together and only changes that are only in the somatic tumour sample and not in the normal sample are reported. Even though this concept sounds simple, there are some pitfalls[51]. First of all, there might be some tumour contamination in the normal sample, which needs to be adjusted for [48, 52]. Second, there might be normal "contamination" in the tumour sample, this means that not all cells in the tumour sample are actually tumour. This means that

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

change this to not have subsubsections

1.5 Lungcancer

With around 1.6 million deaths world-wide each year, lung cancer is the number one cause of cancer death [53]. 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 [54]. 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 [55]. This group usually shows *EGFR* (epidermal growth factor receptor) driven tumours. EGFR is a transmembrane receptor tyrosine kinase, which is usually only expressed in epithelial, mesenchymal, and neurogenic tissue, but its overexpression in other tissues is a hallmark of many human malignancies, not just NSCLC.

Possibly change this to cancer in general

1.6 Overview

add short description of each chapter

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

— Lucius Annaeus Seneca, *Epistulae morales ad Lucilium*

2

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

2.1 Introduction

When I started exploring the somatic variant calling methods in the beginning of my PhD in 2018, I was surprised about the stark difference between germline and somatic variant calling methods. Where all "modern" germline variant callers have the built-in capability to joint call multiple samples, for example from family trios, virtually no somatic variant caller had this function.

The joint analysis of smaller cohorts improves the performance of germline variant calling methods significantly, by allowing to assess technical artifacts, which might be unique for the individual sequencing machine or the researcher handling the DNA [56, 57]. As certain parts of the genome are more problematic to sequence (section 1.3) and map (subsection 1.4.1), a "control" sample can help to distinguish if a certain observed change occurs commonly, is a technical issue or in fact a real change.

For somatic variant calling, this concept has been adjusted in the genome analysis toolkit (GATK) [58] to allow the use of panel of normals (PON), which contains frequently seen changes in healthy ("normal") individuals analysed with the same sequencing technology [59], but this is a post processing step of the analysis rather than a more intricate model like it is for the germline equivalent. Mutect2, which is the most recent somatic variant calling algorithm provided by the Broad institute, however also provides a multi-sample mode, for which all tumour samples need to be from the same patient, either longitudinal

or spatial different [60]. This mode is hidden quite well and all tutorials published by the developers state that "there is currently no way to perform joint calling for somatic variant discovery" [51], so while all methods in the GATK are considered a beta feature, this seems, that development is not a priority.

There are only two methods currently, which have documented and published capabilities to jointly analyse tumour samples from the same patient to call somatic variants. The first one is a specialised method built on a joint bayesian model for SNVs to occur in multiple samples called multiSNV [61]. However it has multiple shortcomings, which make it not useable for our data. First, as the name suggests, the method can only jointly evaluate SNVs and completely ignores indels and structural variants, which would be acceptable for the superior performance shown. However, multiSNV was optimised only for WES and not for the very deep WGS that is now available. This means exceptionally high runtimes. Even with custom parallelisation that was attempted in this work, the predicted runtime for just one multi sample patient would have been longer than 3 years. This again shows, that while multiSNV was a great step forward at the time, there is a real need for new methods to stem the tide of sequencing data available at low cost.

multiSNV has been the only software available for multi sample analysis, but only recently, during this work, superFreq [62] was published. It combines all standard analysis steps for tumour analysis, like variant calling or clonal deconvolution, into one program and is even able to jointly analyse samples. However similar to multiSNV, its focus when optimising and developing was on WES and RNAseq data, so when applied to our data, we could not find a server with enough memory to execute the workflow.

This then prompted us to investigate possible workflows to enable the analysis of high depth WGS, which we estimate to become more and more normal, with the ever dropping prices of sequencing. The following sections will first show the publication and then discusses additional analysis done after the the publication of the manuscript (section 2.4) and the impact of the joint analysis on downstream methods (section 2.3).

2.2 Publication

The publication about joint somatic variant calling can be found at
https://doi.org/10.1093/bioinformatics/btab606

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

The ability to find additional shared variants has significant impact on our understanding of cancer evolution and the timing of initiation and metastatic seeding. Recent work has shown, that similar to the well known genetic heterogeneity, there is heterogeneity when it comes to metastatic seeding. While traditionally it was thought that tumours only metastesised after they reached a certain size, to escape the restrictions of the niche, like reduced nutrition, recent publications showed, there is also very early metastatic seeding [63]. But all those methods are ultimately based on the somatic variants found in the data, so if we improve on the input of the downstream analysis methods, we can expect a clearer and possibly more granular result.

In this section I will highlight for a few examples on how big the effect can be for methods like phylogenetic reconstruction and clonal decomposition.

2.3.1 Phylogenetic reconstruction

show before vs after for phylogenetic reconstuction

2.3.2 Clonal deconvolution

show before vs after for clonal deconvolution

2.4 Longitudinal analysis - something for the ages

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

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

— Lucius Annaeus Seneca, *De Consolatione ad Marciam*

3

CASCADE - Late stage lung cancer in the spotlight

3.1 Introduction

3.2 Publication

This chapter includes the data analysis for two two publications. The first publication features the resistance mechanism of small cell transformation (https://doi.org/10.1016/j.ccell.2019.08.008[64]) and the second shows the discovery of resistance to a targeted RET-fusion driven cancer (https://doi.org/10.1016/j.jtho.2020.01.006[65])

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

3.3 Cohort analysis

3.4 Mitochondrial phylogenetic reconstruction - the power house of the phylogenies

3.5 Outlook

"Many a mickle makes a muckle."

— proverb

4

MisMatchFinder - hope springs eternal

4.1 Introduction

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

— Epictetus, *The Enchiridion*

5

Conclusion

Bibliography

[1] 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.

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

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

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

[5] Christopher P. Austin. *Chromatin*. URL: https://genome.gov/genetics-glossary/Chromatin (visited on 09/16/2021).

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

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

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

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

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

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

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

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

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

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

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

[17] Ezgi Süheyla Doğan and Chang Liu. "Three-dimensional chromatin packing and positioning of plant genomes". In: *Nature Plants* 4.8 (July 2018), pp. 521–529. DOI: 10.1038/s41477-018-0199-5.

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

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

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

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

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

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

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

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

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

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

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

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

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

[31] and Eric S. Lander et al. “Initial sequencing and analysis of the human genome”. In: *Nature* 409.6822 (Feb. 2001), pp. 860–921. DOI: 10.1038/35057062.

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

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

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

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

[36] G. Church and S Kieffer-Higgins. "Multiplex DNA sequencing". In: *Science* 240.4849 (Apr. 1988), pp. 185–188. DOI: 10.1126/science.3353714.

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

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

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

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

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

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

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

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

[45] and 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.

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

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

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

[49] David Benjamin et al. "Calling Somatic SNVs and Indels with Mutect2". In: (Dec. 2019). DOI: 10.1101/861054.

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

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

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

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

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

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

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

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

[58] Geraldine van der Auwera Brian O'Connor. *Genomics in the Cloud*. O'Reilly UK Ltd., May 1, 2020. 467 pp. ISBN: 1491975199. URL: https://www.ebook.de/de/product/29084351/brian_o_connor_geraldine_van_der_auwera_genomics_in_the_cloud.html.

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

[60] GATK Team. *Mutect2 multi-sample*. Sept. 25, 2020. URL: https://gatk.broadinstitute.org/hc/en-us/community/posts/360062528691-mutect2-multi-sample- (visited on 10/23/2020).

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

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

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

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

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