ctDNA 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 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., Chandranada 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.

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

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

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.

FUNDING:

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

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}

Chapter 1

Introduction

In 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 and analyse and 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 **??** 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 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. The DNA in eukaryotes however is not free floating around in the nucleus of a cell, but rather it is highly organised around histones, which then form something resembling a spool of thread. This allows some of the DNA to be accessible where the use of other areas can be restricted. 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.

1.1.1 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. 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[4, 5]. 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.

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 [6, 7]. 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 still alive. 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 [8, 9].

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

[10].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[11, 12], 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 [13, 14]. Even though this was a very long project, it laid the ground work for the usage of the current sequencing technologies.

1.3.1 Library preparation - what we learned from using phages

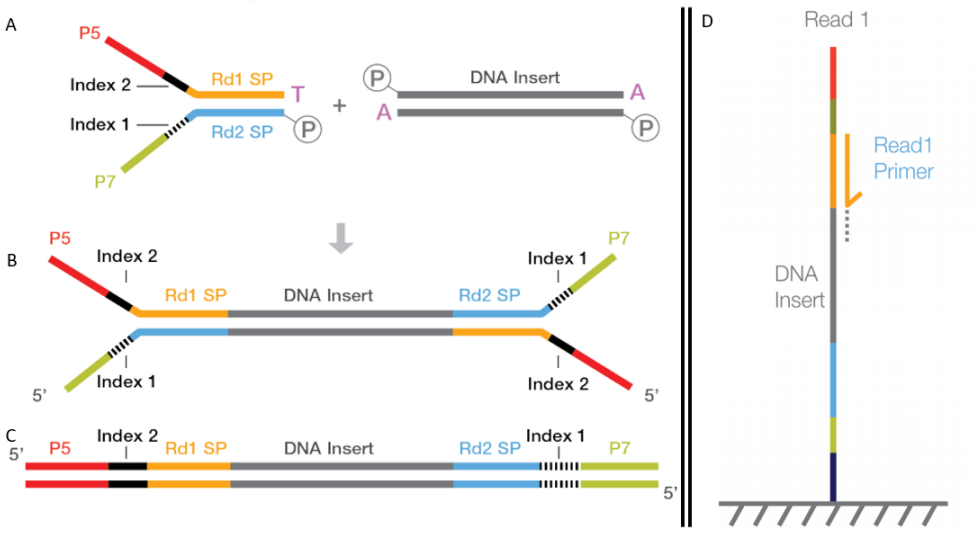

FIGURE 1.1: Adapter ligation during library preparation. The adapters are added to the DNA insert during library preparation. A. The DNA insert is prepared by adding an A-tail and phosphorylation. B. The adapter complex which includes the P5/P7 flow cell binding adapter is added to the DNA insert. C. 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"[15]

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 complex is ligated. This enabled the DNA to bind to the flow cell which is covered with the reverse complement of the adapter Figure 1.1.

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 1.2)[16, 17], which is based on the concept of alternating integration of florescently labelled nucleotides and imaging with a microscope (Figure 1.2) as well as multiplexing, where a DNA fragment is ligated to an index, which allows the sequencing of multiple samples at once [18, 19] as it is shown in Figure 1.1. This method allows highly accurate determination of the sequence of a DNA fragment and depending on the flow cell and sequencing machine allows to sequence a whole genome in just 24h.

1.3.3 Long read sequencing - the "third" generation sequencing

By now, multiple methods which broke free of the size limitations of NGS exist, which are commonly referred to as long read sequencing. Most of the current methods trade the very high accuracy of the second generation NGS methods for the capability of sequencing of sequencing huge continous strands of DNA (current record 2.3 Million bp [20]) 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[21]. 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 origin

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 (see 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 (see Figure 1.1)

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 [22, 23].

1.4.2 Variant calling - spot the distance

1.5 Lungcancer

1.6 Overview

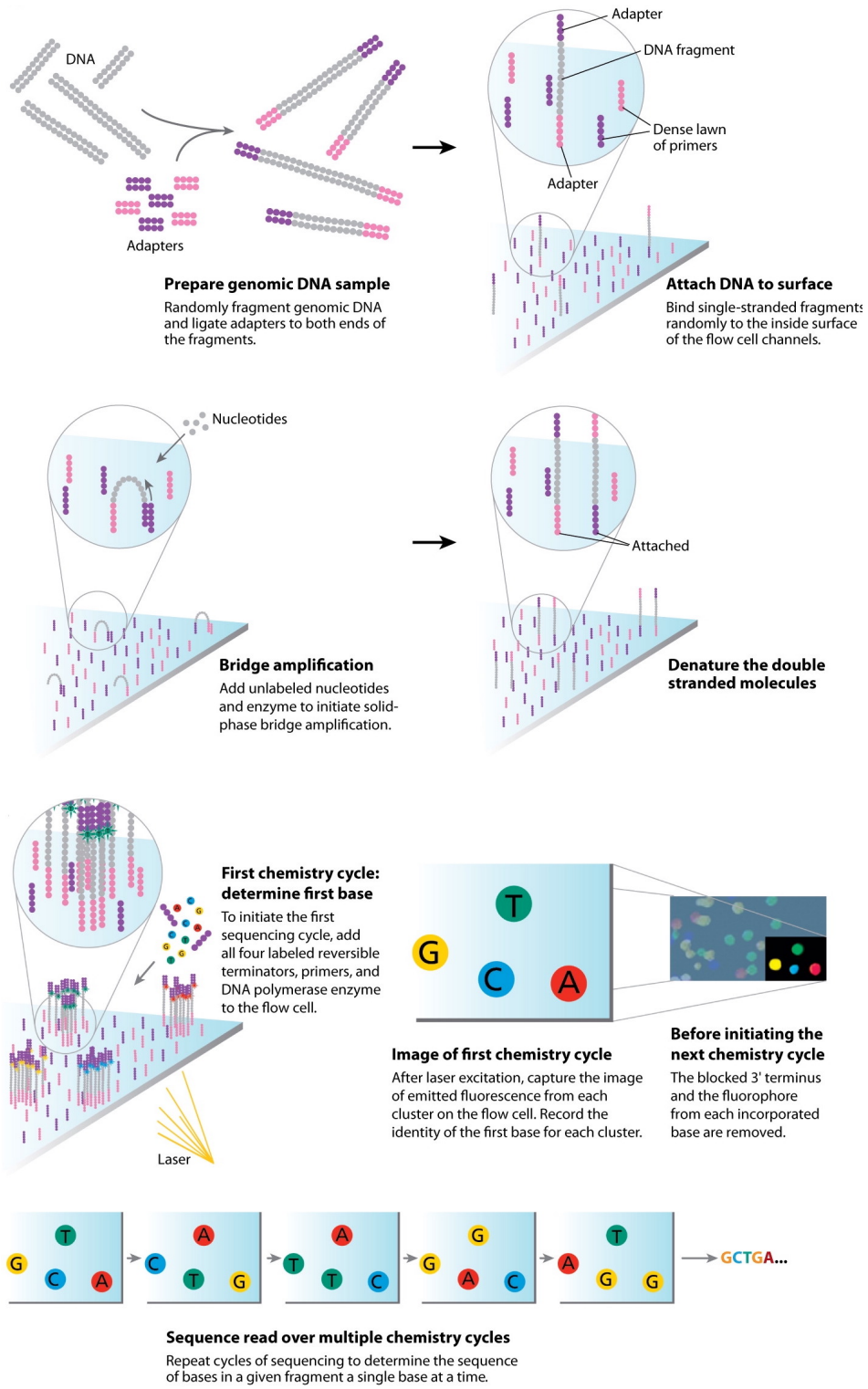

FIGURE 1.2: The Illumina sequencing-by-synthesis approach. Cluster strands created by bridge amplification are primed and all four fluorescently labeled, 3'-OH blocked nucleotides are added to the flow cell with DNA polymerase. The cluster strands are extended by one nucleotide. Following the incorporation 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[16]

Chapter 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.

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

2.4 Longitudinal analysis - something for the ages

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

Chapter 3

CASCADE - Late stage lung cancer in the spotlight

3.1 Introduction

3.2 Publication

3.3 Cohort analysis

3.4 Outlook

Chapter 4

MisMatchFinder - hope springs eternal

4.1 Introduction

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] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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).

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.