KTH ROYAL INSTITUTE
OF TECHNOLOGY

Doctoral Thesis in Biotechnology

Exploring human variations by droplet barcoding

PONTUS HÖJER

Stockholm, Sweden 2024

Exploring human variations by droplet barcoding

PONTUS HÖJER

Academic Dissertation which, with due permission of the KTH Royal Institute of Technology, is submitted for public defence for the Degree of Doctor of Philosophy on Friday the 15th of March 2024, at 10:00 in Inghesalen, Widerströmska huset, Tomtebodavägen 18a, Solna, Sweden.

Doctoral Thesis in Biotechnology
KTH Royal Institute of Technology
Stockholm, Sweden 2024

TRITA-CBH-FOU-2024:7
ISBN 978-91-8040-840-0

Printed by: Universitetsservice US-AB, Sweden 2024

Abstract

Biological variations are being explored at ever-increasing rates through the rapid advancement of analytical techniques. Techniques like massively parallel sequencing empower scientists to accurately differentiate individuals' genetic compositions, cellular functionalities, and healthy tissue from diseased. The knowledge gained from these techniques brings us ever closer to grasping the complexities of life, contributing to human development. Still, to fully elucidate biological variations in different samples requires novel sensitive and high-throughput techniques, capable of placing everything in its correct context. One such technique gaining promise is droplet barcoding.

Droplet barcoding leverages emulsion droplets to segregate samples into their functional components, coupled with barcodes that can group tagged molecules following sequencing. This technique constitutes a versatile tool for studying biological variations in both the phenotype and genotype. This thesis leverages droplet barcoding to explore variations relating to human biology.

Droplet barcoding was used to study phenotype variations, looking at protein compositions in single extracellular vesicles (**Paper I**) and single cells (**Paper II**). **Paper I** studies extracellular vesicles which are naturally released from cells. They carry heterogeneous protein signatures that can inform about their cellular origin. Tens of thousands of extracellular vesicles were profiled, including approximately 25,000 from lung cancer patients. From these protein profiles, extracellular vesicles could be grouped into putative subtypes. **Paper II** presents a novel method

for studying single cells which was used to characterize blood-derived immune cells. The method enabled the identification of most major immune cell lineages.

Haplotype-resolved genetic variations were analyzed using a linked read sequencing method based on droplet barcoding. Linked-read sequencing conserves long-range information from short-read sequencing by co-barcoding subsections of long DNA fragments. **Paper III** presents an open-source pipeline (BLR) for whole genome haplotyping using linked reads. BLR generates accurate and continuous haplotypes, outperforming PacBio HiFi-based diploid assembly. We further show that integration with low-coverage long-read data can improve phasing accuracy in tandem repeats. With 10X Genomics linked reads, BLR generated more continuous haplotypes compared to other workflows. **Paper IV** applies linked read sequencing to reveal the haplotype complexities of cancer genomes. In two patients with colorectal cancer, we identified several large-scale aberrations impacting cancer-related genes. Additionally, several short somatic variants were found to impact nearly all oncogenic networks identified by TCGA. Demonstrating the importance of haplotype-resolved analysis for cancer genomics, one patient exhibited two nonsense mutations on separate haplotypes in the well-known colorectal cancer gene *APC*.

Sammanfattning

Biologiska variationer utforskas i allt snabbare grad, pådrivet av den snabba utvecklingen av analytiska tekniker. Tekniker som massiv parallellsekvensering möjliggör för forskare att noggrant särskilja individers genetiska sammansättningar, cellernas olika funktioner och frisk vävnad från sjuk. Vetskapen dessa tekniker medför ger oss allt djupare insikter om livsformers komplexitet som främjar mänsklig utveckling. Torts dessa framsteg kräver klarläggandet av biologiska variationer i olika prover nya känsliga tekniker med hög kapacitet, kapabla att placera information i dess rätta sammanhang. En särskilt lovande teknik är droppkodning.

Droppkodning utnyttjar emulsionsdropparnas förmåga att separera prover i dess funktionella komponenter kombinerat med DNA-koder för att gruppera märkta molekyler efter sekvensering. Denna teknik utgör ett mångsidigt verktyg för att studera biologiska variationer i både fenotyp och genotyp. Den här avhandlingen utforskar tekniker baserat på droppkodning för att analysera dessa variationer relaterat till mänsklig biologi.

Droppkodning användes i analys av fenotypvariationer genom att studera proteinsignaturer hos enskilda extracellulära vesiklar (**Artikel I**) samt enskilda celler (**Artikel II**). **Artikel I** studerar extracellulära vesiklar, vilka är partiklar som naturligt släpps ut från celler. Dessa vesiklar bär på heterogena protein-signaturer som kan informera om dess cellulära härkomst. I studien undersöks proteinsignaturer från tiotusentals extracellulära vesiklar, inklusive cirka 25 000 från lungcancerpatienter.

Utifrån dessa signaturer kunde extracellulära vesiklar sedan grupperas i potentiella subtyper. **Artikel II** presenterar en ny metod för att studera enskilda celler, som användes för att karakterisera immunceller från blod. Metoden möjliggjorde identifiering av de flesta stora immuncellspopulationerna.

Haplotyp-upplösta genotypvariationer analyserades med en metod för länkad sekvensering baserad på droppkodning. Länkad sekvensering möjliggör att vid sekvensering med kort läslängd bevara information över långa genomiska distanser genom DNA-kodning av små delar av långa DNA-fragment. **Artikel III** presenterar en pipeline (BLR) med öppen källkod för helgenoms haplotypning som använder data från länkad sekvensering. BLR genererar haplotyper med stor exakthet och kontinuitet som överträffar diploid genom-sammansättning (“assembly”) med PacBio HiFi data. Vi visar även att integrering med långa sekvenser med begränsad genomtäckning förbättra haplotypning i tandem-repetitiva genomregioner. Med 10X Genomics länkade sekvenser genererade BLR mer kontinuerlig haplotypning jämfört med andra analysflöden. **Artikel IV** tillämpar länkad sekvensering för att avslöja haplotypkomplexiteten hos cancergenom. Hos två patienter med tjocktarmscancer identifierades flera storskaliga variationer som överlappar cancerrelaterade gener. Dessutom hittades flera korta somatiska varianter som påverkade gener i nästan all onkogena nätverk identifierade av TCGA. En patient uppvisade två nonsensmutationer på separata haplotyper i den välkända tjocktarmscancergenen *APC*, vilket påvisar vikten av haplotyp-upplöst analys för cancergenomik.

Public defense

The public defense of this thesis will take place on the 15th of March 2024 at 10:00 in Inghesalen, Widerströmska huset, Tomtebodavägen 18a, Solna, Sweden for the degree of Doctor of Philosophy in Biotechnology.

RESPONDENT | M.Sc. Pontus Höjer
Department of Gene Technology, KTH Royal Institute of Technology, Science for Life Laboratory, Solna, Sweden.

FACULTY OPPONENT | Professor Lars Feuk
Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden.

EVALUATION COMMITTEE

Professor Jens Lagergren
Department of Computer Science, KTH Royal Institute of Technology, Science for Life Laboratory, Solna, Sweden.

Professor Adnane Achour
Department of Medicine, Karolinska Institute, Science for Life Laboratory, Solna, Sweden.

Associate Professor Linda Holmfeldt
Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden.

CHAIRMAN | Associate Professor Patrik Ståhl
Department of Gene Technology, KTH Royal Institute of Technology, Science for Life Laboratory, Solna, Sweden.

RESPONDENT'S SUPERVISOR | Professor Afshin Ahmadian
Department of Gene Technology, KTH Royal Institute of Technology, Science for Life Laboratory, Solna, Sweden.

RESPONDENT'S CO-SUPERVISOR | Associate Professor Pelin Sahlén
Department of Gene Technology, KTH Royal Institute of Technology, Science for Life Laboratory, Solna, Sweden.

List of papers

This thesis is based on the four papers (I-IV) listed below. The papers are available as appendices to this thesis.

Paper I. Mahsan Banijamali[*], Pontus Höjer[*], Abel Nagy, Petra Hååg, Elizabeth Paz Gomero, Christiane Stiller, Vitaliy O. Kaminskyy, Simon Ekman, Rolf Lewensohn, Amelie Eriksson Karlström, Kristina Viktorsson, Afshin Ahmadian.
Characterizing single extracellular vesicles by droplet barcode sequencing for protein analysis.
Journal of Extracellular Vesicles, e12277 (2022).
https://doi.org/10.1002/jev2.12277

Paper II. Pontus Höjer, Abel Nagy, Humam Siga, Jun Wang, Håkan Jönsson, Petter Brodin, Amelie Eriksson Karlström, Afshin Ahmadian.
Identification of major immune cell lineages with DBS-Pro.
Manuscript in writing

Paper III. Pontus Höjer, Tobias Frick, Humam Siga, Parham Pourbozorgi, Hooman Aghelpasand, Marcel Martin, Afshin Ahmadian.
BLR: a flexible pipeline for haplotype analysis of multiple linked-read technologies.
Nucleic Acids Research, **51** 11 (2023).
https://doi.org/10.1093/nar/gkad1010

Paper IV. Humam Siga, Pontus Höjer, Parham Pourbozorgi, Hooman Aghelpasand, Max Käller, Johan Hartman, Cecilia Williams, Afshin Ahmadian.
Resolving the haplotype complexity of colorectal cancer genomes with droplet barcode sequencing.
Manuscript in writing

* Authors contributed equally

Respondent's contribution

Paper I. Contributed to experimental design and development of analysis pipeline. Performed all data analysis and generated all data visualizations. Contributed to writing, preparing, editing and revision of the manuscript. M.B. and P.H. contributed equally to this work.

Paper II. Contributed to project conceptualization, experimental design and experiments. Performed all sequencing data analysis and generated all related data visualizations. Contributed to writing and preparing the manuscript.

Paper III. Contributed to project conceptualization and pipeline development. Performed all data analysis and generated all data visualizations. Contributed to writing, preparing, editing and revision of the manuscript.

Paper IV. Contributed to project conceptualization, data analysis, data visualization and writing of the manuscript.

Extended list of papers

Below is a list of publications, which I have contributed to, but are not the basis of this thesis.

Örjan Åkerborg, Rapolas Spalinskas, Sailendra Pradhananga, Anandashankar Anil, Pontus Höjer, Flore-Anne Poujade, Lasse Folkersen, Per Eriksson and Pelin Sahlén. **High-resolution regulatory maps connect vascular risk variants to disease-related pathways**. *Circulation: Genomic and Precision Medicine.* **12** e002353 (2019). https://doi.org/10.1161/CIRCGEN.118.002353

Pelin Sahlén, Rapolas Spalinskas, Samina Asad, Kunal Das Mahapatra, Pontus Höjer, Anandashankar Anil, Jesper Eisfeldt, Ankit Srivastava, Pernilla Nikamo, Anaya Mukherjee, Kyu-Han Kim, Otto Bergman, Mona Ståhle, Enikö Sonkoly, Andor Pivarcsi, Carl-Fredrik Wahlgren, Magnus Nordenskjöld, Fulya Taylan, Maria Bradley, Isabel Tapia-Páez, **Chromatin interactions in differentiating keratinocytes reveal novel atopic dermatitis– and psoriasis-associated genes**,*Journal of Allergy and Clinical Immunology.* **147** 5 (2021), https://doi.org/10.1016/j.jaci.2020.09.035

Contents

Preface

My work as a PhD student has been split between the development of new methods and the analysis of data generated by these methods. Both require careful arrangement of different tools and techniques to yield useful results. Some things have worked and are presented here. Others not so much … at least yet anyway (**fingers crossed**). This exploration with cycles of trial and error is central to scientific work. A similar process also occurs in the living organisms these studies are based on. New functions generated by the introduction of *variations*, followed by selection of the most useful ones.

This thesis is about the methods and technologies that allow us to explore biological variations, spanning individuals, cells and macromolecules. In focus are variations in the DNA sequence and protein content. All the work I will show here relates to human biology, therefore I have taken extra care to explain things in that context. The methods and technologies I present are however not limited to humans.

Before presenting my own contribution, I will give a introduction to get everyone familiarized with the subject. In the first chapter I will outline the origins and relevance of the variations observed in living organisms. Next I will cover some key technologies and how they can be combined to extract relevant biological information. Finally, I will give an overview of the analysis methods used to interpret the data generated by these technologies. Let us start at the beginning…

Pontus Höjer

Stockholm, February 2024

[…] I’m simply saying that life, uh… finds a way.

Dr. Ian Malcolm, *Jurassic Park* (1993)

CHAPTER 1.

On the origin of variations

In the beginning, there was nothing. At least nothing relevant to this thesis. But then, around 4 billion years ago, *life* emerged on Earth.[1] What exactly this looked like is still a matter of debate.[2] At some point it - whatever it looked like then - evolved into what we today call *prokaryotes*, small single-celled organisms.

The next milestone for life happened 2 billion years ago with the emergence of *eukaryotic* organisms.[1,3] With this development, life could transcend the limitations of a single cell, paving the way for the organization into multicellular organisms. Therein lay the foundations for the diverse and complex lifeforms we observe today, including us – humans – who made their entrance about 250,000 years ago[4].

While most of us can intuitively distinguish between inanimate and animate matter, defining life is no easy task. Looking at life as it appears on Earth we can observe multiple characteristics. Life creates and maintains

a state of order, resisting "decay into thermodynamical equilibrium (death)"[5]. This order is evident at various levels, ranging from single molecules to cells, from the organization of cells into tissues, and in the formation of organisms. Through metabolism, life transforms chemical materials to generate energy or build useful structures. Life replicates, passing information on to its offspring. Finally, life displays abundant *variations* which through evolutionary processes help it adapt and persist.

1.1. The central dogma

To comprehend the functions and characteristics of life, it is essential to delve into the molecular processes and foundations that govern it. At the core is the *central dogma*, a model describing the transfer of information from *DNA* to *RNA* to *proteins* (Figure 1).[6,7] Three main processes are involved in the central dogma: *replication*, *transcription* and *translation*. During replication, the DNA is duplicated into two copies. In transcription, DNA is copied as RNA. Finally, in translation, RNA serves as a template to generate proteins. Importantly, once information has been transferred to protein, it cannot be transferred back to DNA or RNA.

The central dogma, as presented here, applies to most life forms, with some exceptions. Some viruses, despite debate on whether they are considered alive, can perform replication on their RNA. It is also possible to transfer information from RNA to DNA through a process called *reverse transcription*. Nevertheless, this simplified model of the central dogma provides a useful framework for understanding the molecular basis of life.

Finally, the central dogma forms the basis for understanding the molecular underpinnings of variation. Variations are the source of diversity that

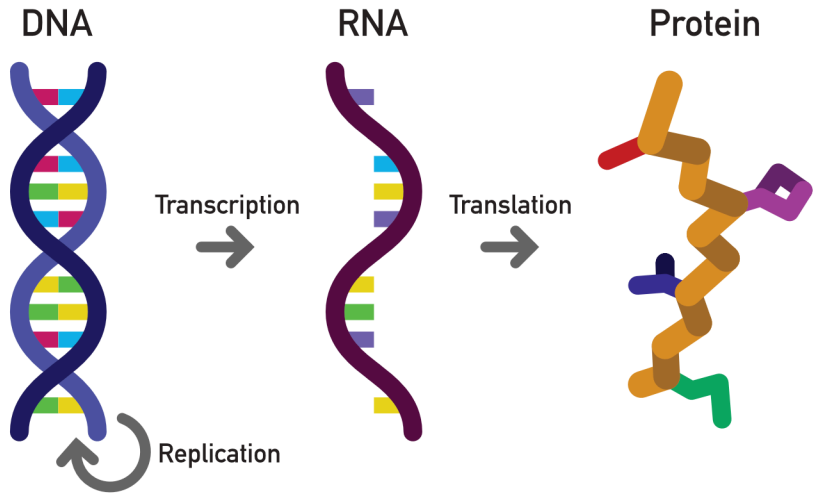

Figure 1: The central dogma of molecular biology shows the main flows of genetic information. DNA is duplicated through replication, allowing a copy to be passed on during cell division. Genes contained in the DNA are copied as a separate RNA molecule in transcription. The RNA is finally used to synthesize a protein in translation.

allows organisms to adapt to changing environments. In life, this is manifested in two interconnected forms, the *genotype* and the *phenotype*. The genotype is an organism's genetic material, predominantly in the form of DNA, containing heritable information. In contrast, the phenotype covers observable characteristics, primarily through the activities of proteins. While variations observed in the phenotype often reflect those in the genotype, this relationship is not always straightforward.[8] Nonetheless, the central dogma offers a framework for comprehending how variations are manifested and inherited.

1.2. The (macro)molecules of life

From the central dogma, we can identify three main types of macromolecules that are essential for life: DNA, RNA and proteins. These macromolecules are composed of smaller building blocks, each of which can be considered quite simple. DNA and RNA each contain four different nucleotides, while proteins are composed of 22 different amino acids. By combining these simple building blocks in different ways, a vast number of different macromolecules can be created.

Out of the three macromolecules of life, DNA provides the main storage for biological information, enabling both inheritance and persistence. It is often referred to as the *blueprint for life* because it contains most instructions required for the organism's life cycle. DNA is composed of the nucleotide bases adenine (`A`), cytosine (`C`), guanine (`G`) and thymine (`T`). These bases serve as a four-letter alphabet, encoding information by chaining them in a specific order, similar to the binary code (`0`s and `1`s) used in computers. In DNA this order is commonly referred to as the *sequence*. The nucleotide bases come in complementary pairs, where `A` binds to `T` and `C` binds to `G`. This means that the sequence of one chain can be used to deduce the sequence of the other. These chains are commonly referred to as *strands*. Together the two strands form the double-helix structure that DNA is famous for.[9] This makes DNA highly stable and capable of storing large amounts of information. The storage capacity is so huge that ~73 grams of DNA could store all global digital information available in the year 2018*.[10,11]

*All global digital information available in 2018 amounts to 33 zettabytes. That is *zetta* as in 10^{21} i.e. 33,000,000,000,000,000,000,000 bytes. A truly mind-numbing amount of data.

The DNA information is transferred by making copies which occurs in two ways, replication and transcription. During replication, the entire DNA sequence is duplicated to create a new, mostly identical (more on this in Chapter 1.6), DNA sequence. This process enables cell division, resulting in two cells sharing the same genetic information. During transcription, only a specific segment of the DNA sequence is copied and this time as an RNA molecule.

RNA is transcribed from a specific DNA sequence known as a *gene*. From a molecular standpoint, a gene can be defined as "a unit of DNA coding for an RNA molecule". The definition of a gene has varied over time[12], and it should be noted that the definition presented here, though adequate, is intentionally simplistic. Another bit of relevant nomenclature surrounding genes is *gene expression*. This refers to the utilization of information stored in the gene, i.e. transcription and possibly translation.

The RNA macromolecule exhibits several important differences compared to DNA. While they are chemically very similar, RNA contains uracil (`U`) bases instead of thymine (`T`) found in DNA. In contrast to DNA, RNA mostly occurs single-stranded and has a comparatively short lifespan. While the RNA sequence emerges as a DNA copy, it might end up with a different sequence before translation into a protein. Within the gene sequence are elements referred to as *introns* and *exons*. Through a process called *splicing* the introns are cut out and exons fused. Splicing can be done between different combinations of exons, creating multiple RNA *transcripts* from one gene. This process is called *alternative splicing*. In humans, the mean number of unique transcripts per gene is around 4[13], clearly contributing to added variation.

While the primary function of RNA is in translation, it also serves various other roles. The RNA molecule designated for translation into a protein is referred to as *messenger RNA* (mRNA). In translation, the mRNA serves as a template, where triplets of nucleotides called *codons* are matched to a particular amino acid. These amino acids are chained to form the protein. The translation process is actualized by a molecular machine known as the ribosome, composed of both proteins and RNA in the form of *ribosomal RNA* (rRNA). Other RNA molecules called *transfer RNA* (tRNA) are employed to link a specific amino acid to a specific codon. RNA also performs several other tasks, such as regulating gene expression.[14] The main takeaway here is that RNA is a versatile macromolecule that can perform many different tasks, both as a template for proteins and as a functional molecule in its own right.

The translation from nucleic acid to amino acid is an interesting process. As stated above a triplet of nucleotides, i.e. codon, is matched to a particular amino acid. Considering a triplet, there are $4^3 = 64$ possible combinations of nucleotides. However, there are only 22 amino acids used in translation. Some triplets encode different information, such as the position where translation should start and stop. However, there is also *degeneracy*, aka *redundancy*, in the code allowing multiple triplets to encode the same amino acid. As a result, the same protein sequence can be generated despite minor variations in the DNA sequence of the gene.

Proteins, the primary functional macromolecules of life, catalyze reactions, provide structure, and play a crucial role in transporting molecules. They are composed of up to 22 different amino acids linked together in the order dictated by the RNA template. For proteins to

become functional, they need to fold into a particular shape. The function is often linked to the protein shape. For example, prion disease is caused by a misfolded protein catalyzing more proteins to misfold in an often fatal chain reaction. This way, functional variation can be gained outside of what is encoded by the protein sequence. Proteins can also gain additional functionality by binding to other proteins or acquiring chemical modifications. Estimating the number of different proteins, or *proteoforms*, generated from a single gene is challenging[15], but analysis of the human H4 gene reported 74 distinct protein structures[16].

Besides the three mentioned macromolecules, others are also essential for life. *Lipids* are integral for the formation of cell membranes. *Carbohydrates* are used for energy storage and as structural components. These two will however not be covered in this thesis.

To summarize, much of the variations observed in life springs from these three macromolecules, DNA, RNA and proteins. Each molecule is built from a few simple blocks. While blocks are simple, they can be combined and modified in a vast number of ways.

1.3. Omes and omics

Before continuing on we should take care of a few more definitions. The complete set of these molecules as they appear in any biological context, such as cells, tissues, or individuals, might be referred to as an "*ome*". Thus from DNA, we have the *genome*, from RNA the *transcriptome* and from proteins the *proteome*. Similarly, the study of these systems is called *omics*, with corresponding subfields *genomics*, *transcriptomics* and *proteomics*.

There has also been a growing interest in the combined study of these "omes", so-called *multiomes*, with the growing field of *multiomics*.[17,18]

1.4. The cell

What connects the majestic blue whale to a microscopic bacterium? Any complex organism, such as the blue whale, can be broken down into smaller and smaller parts. At some point we reach the *cell* - the basic unit of life. This is the smallest unit that we consider to be alive. There have been a lot of cells since the origin of life, with an estimated 10^{39}–10^{40} cells that have existed on Earth.[19]

The cell can in essence be thought of as a bag of molecules. A 'bag' that can maintain order and drive processes. In most cells, this 'bag' is a lipid-bilayer membrane separating the outside from the inside. The membrane also provides a suitable interface to interact with the environment, sending or receiving signals, taking up nutrients, upholding electrochemical gradients and releasing waste.

The greatest division of cell organization is between prokaryotic and eukaryotic cells. In eukaryotic cells, DNA is stored in a membrane-enclosed compartment called the *nucleus*, absent in prokaryotes. The eukaryotic cells have several compartments called *organelles*. These include the *mitochondria*[*] – *the powerhouse of the cell* – responsible for energy production and the *endoplasmic reticulum* involved in protein production. Finally, unlike prokaryotic cells, eukaryotic cells can congregate into multicellular organisms.

[*]The mitochondria is believed to originate from a prokaryotic cell. According to the *symbiogensis* theory, a prokaryotic bacteria was engulfed by a prokaryotic archaea, forming the first eukaryotic cell.[20]

1.4.1. Different types, same DNA

Humans, along with other multicellular eukaryotic organisms, start as single cells. Over the course of its life, this cell divides to form more cells, each one dividing further and specializing to form our respective organs. In the end, a fully grown human male (weighing about 70 kg) has approximately 36 trillion cells.[21] All these cells, with a few exceptions such as red blood cells, practically contain the same DNA. This means that the same information is stored in all cells. It is however obvious that not all cells are identical. Humans have *neurons*, which branch out like tree roots to transmit electric signals, while *adepocytes* form swelling globules to store fat. Cells can differ in shape, size and function while sharing the same DNA.

Cells sharing the same DNA can exhibit different forms and functions by transcribing and translating different proteins. Remember how proteins are the main functional molecules? By relying on different sets of proteins the cells can thereby acquire different functions. In practice, this is achieved by expressing only a subset of all possible genes. Some genes are expressed in all cells, so-called *housekeeping genes*, while others are only expressed in a subset of cells. The level of gene expression is also relevant. A protein present in high amounts can have a different effect than one present in low amounts. Together this allows cells to acquire different functions.

Gene expression is a highly regulated process, controlled by a complex network of feedback loops and interactions. This allows cells to respond to changes in the environment and to coordinate with other cells. The same processes also allow for *homeostasis*, i.e. maintaining a particular

state despite changes in the environment. There are many mechanisms for this regulation, working at the DNA, RNA and protein levels.

When distinguishing between cells one often talks about *cell types*. A cell type is loosely defined as a group of cells that share a particular set of characteristics. For this purpose, gene expression is often used as a proxy. Thus an operational definition of cell type is cells having *similar* RNA and/or protein content. Similar is a key word as there are always stochastic variations in gene expression. Furthermore, the degree of similarity required to consider two cells to be of the same type could be set at different levels. For example, one overarching cell type in the immune system is T-cells. T-cells are commonly split into *helper* ($CD4^+$) and *cytotoxic* ($CD8^+$) T-cells, each of which could be split into subtypes and further, all the way to specific clonotypes. There is also a resolution aspect to cell type. Depending on how detailed gene expression is measured and how the similarity threshold is set, the number of cell types will vary.

Cells described as one "type" often display some level of phenotypic variations. Moreover, cells can change their characteristics over time, for example by epithelial-mesenchymal transition[22]. One way to view cells is the famous *Waddington landscape*, originally intended for describing cell differentiation.[23] Here cells are imagined as beads running down a rolling landscape with branching valleys and ridges. Valleys signify stable states for the cell. A perturbation, or "push" if you will, can send the cell up the ridge in some direction. Either the cell returns to the same state or, if the perturbation is sufficiently large, sends it over the ridge to a new valley. In this model the cell *state* is unstable but spans the entire landscape, providing a snapshot of the cell at a particular time.[24] The

Figure 2: Extracellular vesicles are small particles released from cells. They carry a rich cargo of various molecules, reflecting the cell of origin.

cell type is then the valley, signifying the actual functions of the cell. Thus some of the cell "types" usually defined might be transient cell states. Nevertheless, the operational cell type definition is a useful concept for describing the major functional groups of cells.

1.4.2. Extracellular vesicles

Cells regularly release a variety of items, among them extracellular vesicles (EVs) (Figure 2). The International Society for Extracellular Vesicles (ISEV) defines EVs as "particles naturally released from the cell that are delimited by a lipid bilayer and cannot replicate".[25] The naming for these particles is a subject of debate, with many referring to them as "exosomes".[26] Especially confusing is the nomenclature surrounding EV subtypes[27], but adoption of - mostly agreed upon - guidelines will hopefully aid with the situation[25]. Nonetheless, the biological relevance of EVs is significant as they are conserved among all domains of life.[28]

EVs are heterogeneous in terms of biogenesis, size and molecular content. Two major processes for generating EVs exist, either by budding from the plasma membrane one-by-one[29] or the endosomal system in

batches[30,31]. The size of EVs depends a bit on what types of EVs are considered but is generally in the nanometer range. EVs carry a rich cargo of various molecules. The most studied are proteins that are present both internally and on their surface.[32] Some EVs also carry nucleic acids including mRNA and DNA.[33] The EV cargo reflects the cell of origin, such as different breast cancer subtypes[34], but is also enriched for specific proteins and nucleic acids[35].

Various biological functions of EVs have been proposed and are still being explored.[35] EVs are known to be involved in cell-to-cell communication, transmitting information by interacting with the acceptor cell either on the cell surface or by being internalized.[36] For example, EVs have been involved in numerous roles relating to immune response.[37]

There are several clinical applications of EVs. For one, EVs are present in all bodily fluids making them interesting as biomarkers in liquid biopsies.[35] Liquid biopsies are appealing since patient sampling can be performed in a relatively non-invasive manner.[38] One example is the diagnosis of high-grade prostate cancer from urine-derived EVs, helping to inform whether an invasive prostate biopsy is warranted.[39,40]

1.5. The genome

The *genome* is the complete set of DNA present in a cell. Most of the genome is divided into one or multiple *chromosomes*. Some of the DNA is *extrachromosomal*, i.e. exists outside of the chromosomes, which in eukaryotic cells is found mainly in the mitochondria. For the purpose of the thesis, we will focus on the DNA present in the chromosomes.

Eukaryotic genomes are generally much larger than prokaryotes, presenting both a challenge and an opportunity. In humans, most cells contain roughly 2 meters worth of DNA.[41] This length needs to be packed into a cell nucleus that is only about 10 micrometers in size. To pack the DNA into this very small space the DNA is wrapped around histone proteins forming a structure called *chromatin*. The chromatin is further folded and packed to fit into the nucleus. This packing further allows for the control of gene expression. Dense packaging of chromatin renders the DNA inaccessible for transcription*, which can be controlled by chemical modifications on the DNA or histones.[42] The folding of chromatin also results in a hierarchical 3D organization, bringing genomically distant regions into spatial proximity that enables an additional layer of regulation.[43] Thus packing DNA into the nucleus allows eukaryotic cells to store and regulate large amounts of genetic information. This genome regulation constitutes its own "ome", the *epigenome*, but will not be the focus of this thesis.

1.5.1. The human genome

The human genome is divided across 24 different chromosomes totaling 3.1 billion bases[44], but in most human cells you will find 46 chromosomes. The reason for this is that the human genome is *diploid* meaning that it contains two sets of paired chromosomes. Most chromosome pairs are *homologous* meaning that they carry the same genes, the one exception being males that carry different sex chromosomes, i.e. X and Y. Not all cells are diploid as some are instead

*Chromatin is often classified based on how tight it is packed into condensed (*heterochromatin*) or open (*euchromatin*) chromatin.

haploid carrying only one set of chromosomes. These are the *gametes*, i.e. the sperm and egg cells, that upon conception form the diploid set of chromosomes we inherit. It is also possible to have more than two sets of chromosomes, a condition called *polyploidy*. While this is rare in animals, up to 40% of the hepatocytes in the human liver can be polyploid.[45] Finally some cells and individuals will have additional or missing chromosomes, a condition known as *aneuploidy*, often associated with diseases such as cancer.

Looking a bit deeper into the human genome we can see that it is made up of a wide variety of elements, many of which are still being explored. It contains about 40,000 genes*, of which ~20,000 are protein-coding.[13] These protein-coding genes only make up a small fraction of the genome (~1%).[46,47]

While only 1% of the human genome is protein coding, it does not mean that the rest is useless. Functional elements make up a significant fraction, how large depends on how you ascribe function.[48] Considering functional elements as those encoding products (i.e. proteins or RNAs) or displaying activities related to this production (including regulatory regions such as enhancers and promoters) account for ~8% of the human genome.[49] If we instead consider which parts of the genome are evolutionary conserved, i.e. lack significant variations due to detrimental effects, at least 10.7% could be considered functional.[50]

*This is based on the definition that genes are parts of the genome that are transcribed. In the NCBI RefSeq annotation, which includes e.g. non-transcribed pseudogenes, the number is ~60,000.

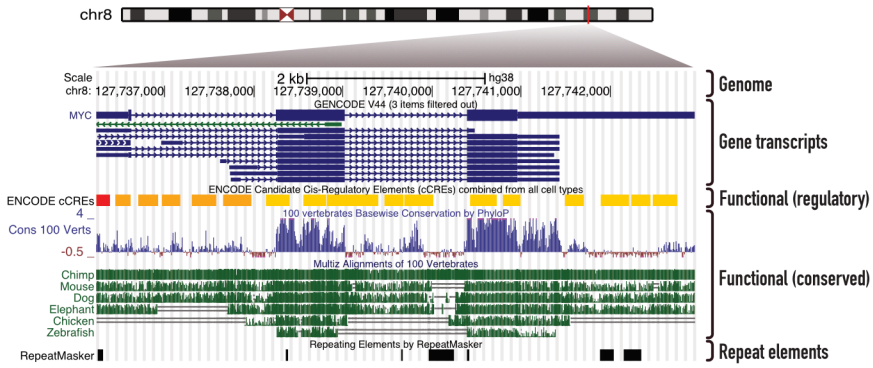

Figure 3: The mosaic structure of the human genome (GRCh38/hg38) surrounding the *MYC* gene on chromosome 8. Annotation with source in parenthesis includes from top to bottom; Transcripts (GENCODE v44), Functional elements including regulatory elements (ENCODE) and evolutionary conserved regions (PhyloP and multi-species alignments), and finally repeat elements (RepeatMasker). Image generated using UCSC Genome Browser[51].

Besides functional elements, more than half of the human genome is made up of *repeat elements*.[52] Repeat elements vary widely in both size and structure.[53] These elements span single bases (*homopolymers*) to chains of repeated patterns (*tandem repeats*) to *segmental duplications* encompassing thousands of bases. Together, these repeats constantly restructure the genome by expansion and contraction or by moving around from one place to the other (*transposition*), introducing variations that drive both evolution[54] and disease[55]. While repeat elements are found throughout the genome, they are enriched in the *centromeres* ("middle") and *telomeres* (ends) of chromosomes.

Finally, it should be noted that these functional and repeat elements often overlap. They are not mutually exclusive making the genome mosaic in structure (Figure 3).

1.6. Genetic variations

All human individuals have different genomes, including "identical" (monozygotic) twins.[56,57] The same is true for most other organisms. Even cells within the same organism differ in their genomes.[58] These differences between genomes are what we call *genetic variations* or *genetic variants*. The variants are manifested as changes in the DNA sequence of varying size, rate and impact.

Genetic variations are by nature dangerous to an organism as they could lead to loss of essential functions or - in the worst case - death. However, they are also a necessity for evolution and adaptation. Variations are the source of diversity that allows organisms to adapt to changing environments. Thus variations are a double-edged sword, both a necessity and a danger. The rate of such variations therefore needs to be carefully balanced.[59,60]

1.6.1. How variants are introduced

The mechanisms behind genetic variants are manifold.[61,59] In all organisms variants can get introduced due to errors in DNA replication[62] and DNA damage repair[63]. In Chapter 1.5 I introduced repeat elements, which can also cause variations during replication or by transposition.[64] Sexually reproducing organisms also have variants introduced during *meiosis*, for example, by interchanging parts of homologous chromosomes through *recombination*. Some immune cells also introduce variations into targeted sections of their genomes, increasing their phenotypic diversity to better fight pathogens.[65] One extreme example of rapid variation is *chromotripsis* where a chromosome

is shattered into hundreds of fragments and randomly reassembled in a single event.[66]

1.6.2. Variant classifications

In multicellular organisms, variants can be divided into *germline* and *somatic* variations. Germline variations are those that exist at the point of genesis such as, for humans, the variations present in the fertilized egg from which we originate, and thus occur throughout all descendant cells. Variants that emerge after genesis are considered somatic variations. These somatic variations accumulate with age[67] and are considered to be an important driver of cancer[68].

Genetic variants are often classified into different types (Table 1). They are also categorized based on size into *short variants* and *structural variants* (SVs). The threshold for this is somewhat arbitrarily set, and multiple thresholds have been used.[69] Variants below 50 base pairs (bp) are typically considered short. Short variants include SNVs, MNVs, and INDELs. Structural variants include insertions, deletions, inversions, and translocations. Large duplications and deletions are commonly referred to as *copy number variations* (CNVs). In some instances, multiple structural variants can be combined in complex rearrangements, e.g. during chromothripsis. Among the different types of variants, SNVs are the most numerous with about 4 million SNVs per human genome.[70] SVs on the other hand, while less numerous, affect more bases per genome[71,72], potentially having a higher phenotypic impact[73–75].

Table 1: Descriptions of different genetic variant types.

Variant type	Description
SNV	Single nucleotide variant. One nucleotide in a sequence is exchanged for another, e.g. A ⟶ G.
MNV	Multi nucleotide variant. Two or more nucleotides in a sequence are exchanged for other nucleotides.
INDEL	Group name to refer to insertion or deletion, usually reserved for small variants.
Insertion	One or more nucleotides are inserted into an existing sequence.
Deletion	One or more nucleotides are deleted from an existing sequence.
Duplication	A existing section of DNA is re-inserted into the genome. Depending on if the duplicated sequence is inserted next-to or away from the original sequence these are also respectively referred to as tandem and interspaced duplications.
Inversion	A section of DNA that is inverted in place.
Translocation	A section of DNA that is moved to another location in the genome.
CNV	Copy number variant. A section of DNA is duplicated or deleted.

1.6.3. Variant impact

The phenotypic effect of a variant depends on its type and context. For example, an SNV in a protein-coding sequence could lead to an amino acid change or encode the same amino acid due to codon degeneracy. Even deletions of a million bases can have no discernible phenotypic impact.[76] In addition to its genomic context, the effect of a variant is also dependent on the cellular context, for example, what other variants are present. Take, for example, a variant in a regulatory region connected to a specific gene. If that gene is not expressed, there is no phenotypic impact.

Somatic variations play a crucial role in driving cancer. Variants are acquired progressively during the lifespan, some possibly inherited, causing a single cell to proliferate out of control. The variants must endow the cell with a variety of capabilities, known as the *hallmarks of cancer*, which include uncontrolled cell growth and evasion of cell death.[77–79] While cells need to acquire multiple of these capabilities to become cancerous, each one provides the cell with a selective advantage. Variants that result in such an advantage are referred to as *driver mutations*.[80] A recent large study of different cancer types found that, on average, there are 4.6 driver mutations per tumor.[81] Of these mutations, 55% could be attributed to structural variants.[81]

1.6.4. The human diplome

One largely overlooked fact when studying genetic variants in humans is that they are not independent. Humans are diploid, meaning that they have two copies of most chromosomes*. This implies that each

*The one exception being the sex chromosomes in males.

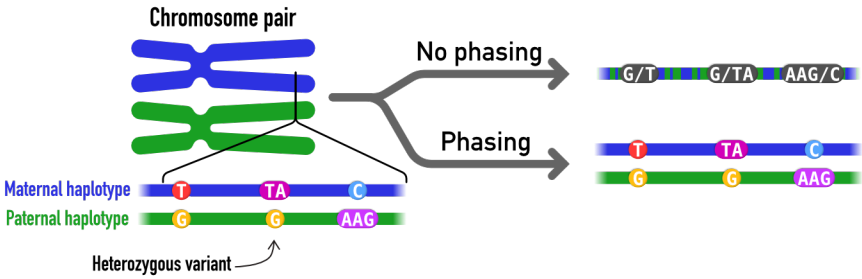

Figure 4: Genetic variants appear as two separate haplotypes in diploid organisms. These haplotypes can only be recovered through phasing.

variant can either appear on both (*homozygous*) or different (*heterozygous*) chromosomes. For example, a heterozygous variant in a gene might not have any effect if the other copy of the gene is functional and able to compensate. However, in cases of *haploinsufficiency*, losing functionality in one copy might be sufficient to alter the phenotype.[82]

One other aspect of this dependency is that variants are inherited and coexist on the same chromosome (Figure 4). This is referred to as *haplotypes*, of which we inherit one from each parent - a *maternal* and *paternal* haplotype. The process of determining which variants are inherited together is called *haplotyping* or *phasing*. Haplotyping is vital for understanding the connection between genotype and phenotype as variants that are linked together and, by extension, exert their effect together.[83–85] Consider, for example, the HLA genes, critical to immune system function.[86] These are located to a ~5 million base-pair region that is highly *polymorphic* (lots of variants) and exhibits high *linkage disequilibrium*[87], meaning that variants co-occur non-randomly. To accurately determine the HLA genotype it is therefore imperative to identify the haplotype.

CHAPTER 2.

The omics toolkit

As we have learned, there are multiple levels of variations in and between living organisms. The variations of particular relevance to this thesis are (1) variations of the genome between individuals and tissues and, (2) variations in protein expression between cells and extracellular vesicles. DNA molecules, cells and extracellular vesicles all represent different vehicles for variation, but often similar technologies can be applied to study them. Therefore, I will henceforth refer to them by a single word, *gestalts*.

Merriam-Webster defines *gestalt* as "something that is made of many parts and yet is somehow more than [...] the combination of its parts"[88]. Though the concept is not commonly applied in biology, I think it is suitable in this context. Here the gestalts can refer to DNA molecules, cells, EVs or other microscopic biological structures. The parts then refer to the measured "omes", i.e. genetic variants, gene expression, protein content, etcetera.

Omics measurements on gestalts require an array of different technologies. We need to extract the molecules of interest. Then we need to measure them to get data. Finally, we need to analyze the data to gain meaningful information. In this chapter, I will go through some of the technologies that can be used to get omics data. The analysis of the data recovered from these methods will be covered in Chapter 3.

2.1. Barcoding

One of the key technologies in omics is *barcoding*. Barcoding is the process of tagging something with a unique identifier, e.g. a barcode. The barcode allows tracking of the item through a specific process. An important application is *multiplexing*, where multiple objects are pooled, each with their own barcode that enables distinguishing between them. Similar concepts are used in different fields, ranging from retail to computer science dictionaries*.

Barcoding is especially useful in connection with sequencing.[89–91] Sequencing will be covered more in-depth in Chapter 2.4, but in short, it is the process of determining the order of nucleotides in a DNA molecule. In this process, the DNA molecules can be tagged with a *barcode*, a separate identifiable sequence of nucleotides, which is then read in sequencing along with the original DNA molecule. If you have DNA molecules from multiple samples you can tag each sample with a unique barcode and then pool everything. After sequencing you can then tell which sample a particular DNA sequence came from by looking up

* A dictionary is a collection of data, each of which can be accessed through a unique key (e.g. a barcode). Other names for the same thing include *associative arrays*, *maps* and *symbol tables*.

the barcode. This allows for multiple samples to be sequenced together, increasing throughput and reducing cost.

Though I am using *barcode* here there are many other related names for these unique sequences, their use depends a bit on context. In the context of barcoding individual samples before pooling, the sequence is often referred to as an *index*, and the process *indexing*. In the context of barcoding individual molecules for quantification or error correction, the sequence is referred to as a *UMI* (unique molecular identifier).[92,93] All words refer to the same concept, a unique sequence that can be used to identify something at some point in a process.

Barcode sequences can be either (1) predefined or (2) a pool of random sequences. In the first case, you can anticipate the sequences when the barcode is read during sequencing. For example, known sequences can be assigned to different samples, enabling the separation of samples based on their barcode sequence. Furthermore, the read barcode sequence might contain errors introduced after tagging or during sequencing. With predefined barcodes, the sequences can be designed to enable correction for some of these errors.[94,95] For example, in barcodes designed using a Hamming code with a minimum distance of three, a barcode `GATTAGA` read as `CATTACA` could be corrected to account for the substitution `G` ⟶ `C`.[96] In the second case, there is no *a priori* knowledge of the sequence. Unlike predefined barcodes, which need to be synthesized one by one, a population of different barcodes can be generated in a single synthesis run. This is achieved by adding multiple different bases at a single position along the synthesized sequence. These positions are called *degenerate*, containing two, three or four different bases. The most common use case

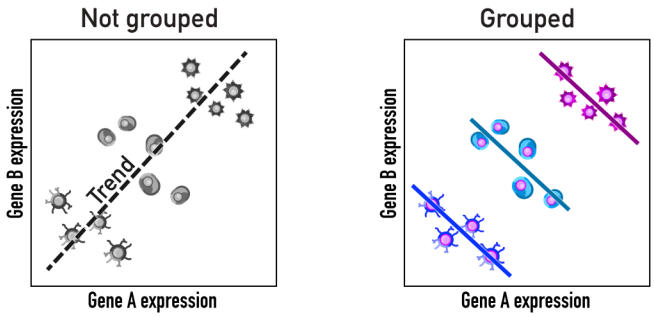

Figure 5: Simpson's Paradox[98] illustrated for two genes across cells. The overall trend is reversed when the data is split into subgroups.

for these is for barcoding molecules for quantitative purposes, e.g. UMIs. [92] Notably, error correction of this type of barcode is still possible but requires a different approach, such as sequence clustering.[97]

2.2. Compartmentalization

Historically, the study of cells has been largely limited to bulk measurements. Complex tissues would be ground down to a molecular smoothie, concealing internal heterogeneity and occasionally resulting in misinterpretations (Figure 5).[99] Cells, extracellular vesicles and chromosomes all appear as complex structures containing many different components. To learn about these structures, it is crucial to measure the molecules that belong together and distinguish them from those that do not. One approach to gaining this information is *compartmentalization*.

2.2.1. Methods for compartmentalization

In essence, compartmentalization involves isolating the gestalts comprising a biological system for measurement. In practice, this requires barcoding of the isolated gestalts by some technique. To isolate gestalts, one can either physically place them into separate compartments or

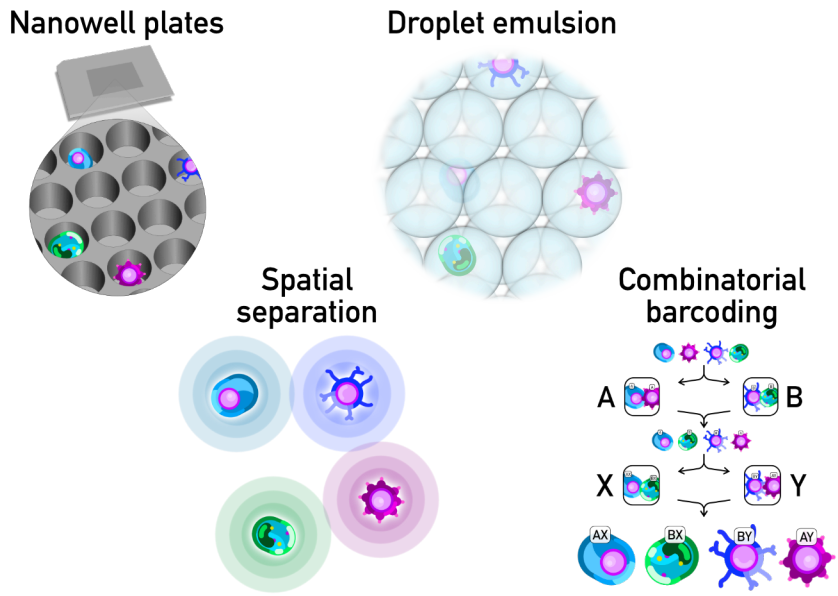

Figure 6: Different methods for high-throughput compartmentalization of gestalts.

rely on statistical methods for separation. There are many methods to physically separate gestalts, for example, picking[100,101], microfluidic valves[102,103] or flow sorting[104]. These approaches are however not feasible for large numbers of gestalts.[105] Instead one can rely on statistical approaches such as droplets to isolate gestalts at higher throughputs.

In the statistical approach, gestalts are randomly distributed over a larger number of compartments. By limiting the ratio of gestalts to compartments the risk of having more than one gestalt in a compartment is limited (more details in Chapter 2.2.3). In practice, there are multiple techniques to this, including *nanowell plates*, *spatial separation*, *combinatorial barcoding* and *droplet emulsions* (Figure 6).

Both plates and droplet emulsions employ a similar procedure for compartmentalization. Gestalts are randomly distributed over a large number of small, physically isolated compartments. In plates, gestalts are distributed over wells, typically by flowing a gestalt suspension over the wells and then allowing sedimentation in place.[106] Alternatively, a low-volume dispenser can be used.[107] In droplets, gestalts are instead encapsulated into droplet emulsions. Unlike plates which are physically constrained by the plate design, droplet generation can be dynamically scaled to the desired number of compartments. Various methods of generating droplets will be covered in Chapter 2.2.2.

Spatial separation involves generating virtual compartments by diluting a gestalt suspension to the point where the probability of having more than one gestalt within a certain distance is low. Although not widely used, this method could be quite accessible due to its lack of expensive instrument requirements. Examples of this approach include the immobilization of cells in a hydrogel[108] and of extracellular vesicles on the surface of plate wells[109].

In these three approaches, barcodes must also be distributed across the same compartments as the gestalts to capture their contents. These barcodes are distributed in various forms such as clonal beads[110,111], DNA nanoballs[109] or simple oligonucleotides[112–114]. In the last compartmentalization method, combinatorial barcoding, the barcodes are instead generated by the compartmentalization process.

Combinatorial barcoding*, involves generating virtual compartments through multiple rounds of splitting, barcoding and pooling.[115–117] Initially, a pool of gestalts is divided into multiple samples, each with multiple gestalts. Each gestalt is labeled with a sample-specific barcode. The samples are subsequently pooled, split again, and assigned a second unique barcode, resulting in gestalts having a combination of two barcodes. This process is reiterated until each gestalt possesses its own unique combination of barcodes, forming a gestalt-specific barcode. The advantage of this approach is its capability to process a very high number of gestalts, limited only by the number of barcodes and rounds. For instance, using a standard laboratory 96-well plate with 96 unique barcodes, over three rounds, one can generate nearly one million compartments†. However, a drawback is that it necessitates multiple rounds of splitting and pooling, a process that can be laborious and introduce bias. Nevertheless, it is a powerful approach that has been used in many different contexts, such as single-cell chromatin accessibility[118] and transcriptomics[119,120], chromatin 3D structure[121] and linked-read sequencing[122]. Additionally, it can be utilized in barcode generation.[123]

Combining different compartmentalization methods is also possible. One example of this is in the scifi-RNA-seq method for single-cell transcriptomics.[124] In this method, combinatorial barcoding is performed first using plate wells, then a second round in droplets. This allows for an increase in throughput per droplet.

*Combinatorial barcoding is also referred to as *combinatorial indexing*, *split-pool barcoding* or *split-pool indexing*.
†$96^3 = 884,736$ compartments

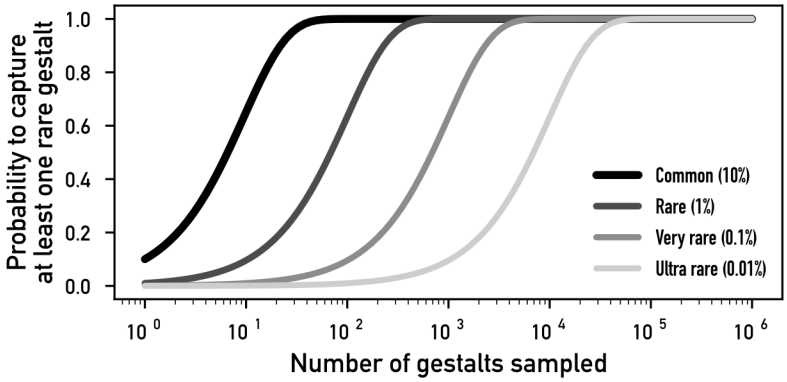

Figure 7: Capture of increasingly rare gestalts requires increasingly more gestalts to be sampled. The graph shows the probability of capturing a rare gestalt when sampling different numbers of gestalts. Rarity, i.e. frequency of gestalt type in population, is set a 10%, 1%, 0.1% and 0.01%.

Why is processing a high number of gestalts important? This is because many biological systems exhibit a heterogeneous composition, with gestalts appearing at varying frequencies. For example, in a tissue, there are many different cell types, some common, some rare. Measuring only a small number of cells might result in missing rare cell types. Increasing the number of gestalts enhances the likelihood of capturing rare gestalt types (Figure 7). Furthermore, increased sampling can also improve the accuracy of measurements by reducing the impact of technical noise, such as sequencing errors.

2.2.2. Droplet generation

Droplets, or rather emulsion droplets, are generated from two immiscible (un-mixable) fluids forming two distinct phases. One *continuous* phase, composed of oil, surrounds the droplets in the *dispersed* phase, mainly consisting of water. The dispersed (water) phase can in this case

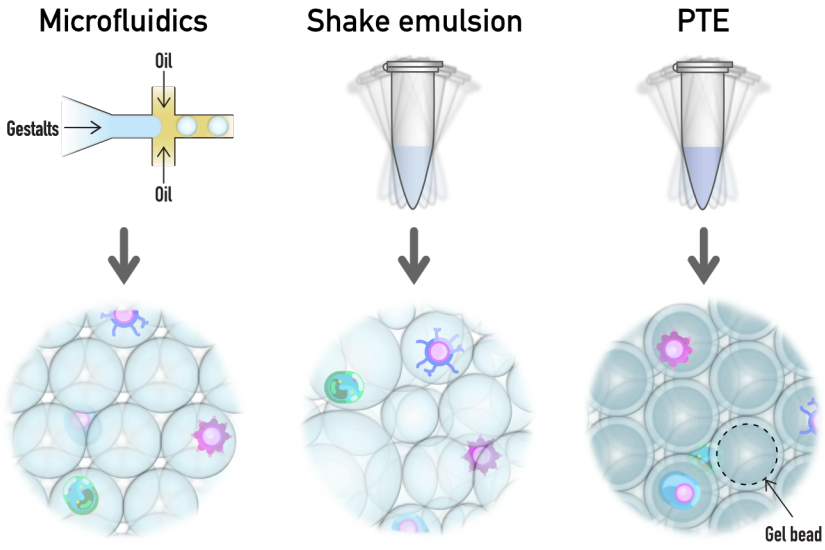

Figure 8: Illustrations of different methods for droplet generation. PTE stands for particle-templated emulsion.

contain the gestalts to be compartmentalized. To stabilize the droplets, a surfactant is commonly added to the oil phase.

Multiple approaches exist for generating droplets (Figure 8), with the most common being the use of *microfluidics*. Microfluidics involves a chip with carefully designed channels to control the flow of fluids and the formation of droplets.[125] The fluids are propelled through the channels to a point where they intersect, resulting in the generation of nano- or pico-liter droplets. These droplets can then be collected, for example, in a tube, for further processing. Numerous workflows exist for droplet generation using microfluidics, and several commercial solutions are available, including from companies such as 10X Genomics, Mission Bio, 1CellBio, and MGI.

The generation of droplets using microfluidics offers several advantages. Droplet size can be tuned depending on the application, for example by adjusting the flow rate of the fluids. The droplets are typically monodisperse, meaning they are of the same size, ensuring predictable encapsulation of gestalts. Moreover, droplets can be generated at a high rate, Clark and Abate demonstrated rates of 23,000 droplets per second*[126], enabling high throughput profiling. The microfluidics chip can also be designed to manipulate the different phases and droplets to create specialized workflows. For example, the chip may be designed so that cells and lysis buffer only mix right before encapsulation. The major drawback with microfluidics is the expensive setup, requiring specialized equipment. Nevertheless, microfluidics has been extensively used for single-cell omics.[127]

Alternatively, droplets can also be generated using *shake emulsions*. Here, two immiscible fluids are mixed by shaking vigorously, generating droplets through high shear forces. This approach is much simpler than microfluidics, requiring only a tube and a shaking device, e.g. a vortex. Furthermore, throughput scaling is no longer dependent on time, as with microfluidics, but rather on volume.[128] This means that the time to generate one thousand droplets is the same as for one billion droplets. Some drawbacks with shake emulsions include that generated droplets are *polydisperse*, i.e. not monodisperse, making encapsulation of gestalts less predictable. It is also harder to tune the droplet size. Despite these drawbacks, shake emulsions have been successfully used in many

*At a rate of 23,000 droplets per second it would take 43.5 seconds to generate one million droplets.

different applications, including in the first massively parallel sequencing system[129].

The shake emulsion approach was recently improved upon by Hatori, Kim and Abate, introducing the concept of *particle-templated emulsions* (PTE). [128] In this approach hydrogel beads of a fixed size were used in the dispersed phase forming a template for the droplets to form around. This enabled the generation of monodisperse droplets using a simple benchtop vortex. The droplet size can here also be controlled by changing the size of the hydrogel beads. The approach has been used in several different applications, including single-cell omics.[128,130]

2.2.3. Compartmentalization statistics

The compartmentalization of gestalts is a random process, but it can be modeled using the *Poisson distribution*[131]. This describes the probability of observing a certain number of events (e.g. gestalts) within a confined time-period or space (compartments). To use this distribution we need to know the rate of events λ, which for our purpose is the ratio of gestalts to compartments. The probability of observing k events is then given by the Poisson distribution:

$$P(k) = \frac{\lambda^k * e^{-\lambda}}{k!}$$

Using this one can aid in selecting an appropriate ratio of gestalts to compartments (λ). We can for example estimate the probability (p) for having multiple gestalts in occupied compartments, the so called *multiplet rate*:

$$p(\text{multiplet}) = \frac{p(\text{multiple gestalts})}{p(\text{not empty})} = \frac{1-P(0)-P(1)}{1-P(0)}$$

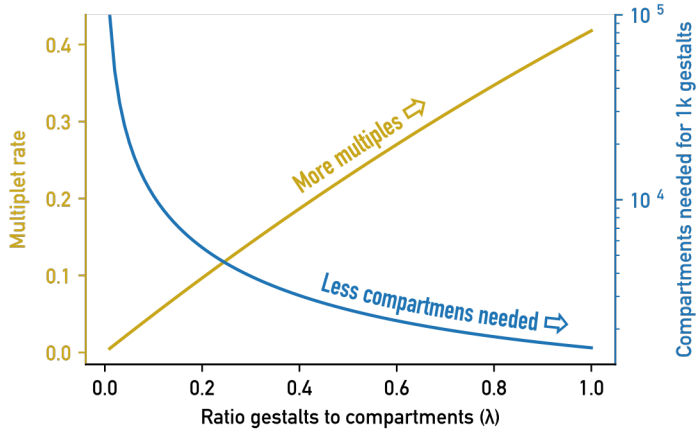

Figure 9: The relationship between the multiplet rate and the number of occupied compartments modeled on the Poisson distribution for λ (ratio gestalts to compartments) between 0 and 1. For example, for $\lambda = 0.1$ the multiplet rate is 4.92% and would require around 10,000 compartments to capture 1000 gestalts.

The multiplet rate is important if we aim to study single gestalts. Selecting a high λ will lead to a high multiplet rate. Similarly, selecting a low λ will lead to a lot of empty compartments, requiring more compartments to capture a desired number of gestalts. Thus selecting λ is a balancing act between these two extremes (Figure 9).

While the Poisson distribution in theory describes the distribution of gestalts over compartments, in practice other factors affect this distribution. For example, shake emulsion compartments can vary in size, skewing the distribution towards larger compartments. The compartment size also limits how many gestalts can fit inside. This fact can further be exploited, where matching the compartment size to the gestalt size can generate compartments containing predominantly single gestalts. This concept has been used for the compartmentalization of barcoded beads as

these are monodisperse, unlike e.g. cells, generating predominantly single bead compartments.[132,106]

2.3. Polymerase chain reaction

Polymerase chain reaction (*PCR*) is a fundamental technique in molecular biology, enabling us to rapidly generate large amounts of DNA from even a single fragment. As we learned in Chapter 1.2, DNA can replicate into two identical copies. This is done by the enzyme *DNA polymerase* which reads a DNA strand and synthesizes a new strand complementary to the original. A similar process to replication is used in the PCR to amplify a specific DNA sequence.

In PCR a DNA sequence is amplified by repeated temperature cycles of *denaturation*, *annealing* and *extension*. In denaturation, the double-stranded DNA is heated to the point that the strands separate. The now single-stranded DNA allows for a *primer*, a short complementary DNA sequence, to bind (anneal) to the DNA by cooling. After this, a polymerase can extend the primer along the other DNA strand by incorporating and chaining matching nucleotides. The process is then repeated for the desired number of cycles, each cycle doubling the amount of DNA leading to exponential amplification. 30 PCR cycles are, at least theoretically, enough to generate over a billion* copies of a single DNA sequence.

PCR is not a perfect process, introducing both errors and biases. The polymerase can for example incorporate the wrong nucleotide, or skip a nucleotide.[133] These errors can be introduced when amplifying DNA for sequencing, exacerbated over each additional PCR cycle. Importantly

*$2^{30} = 1,073,741,824$.

these errors introduce variations that were not present in the original DNA sequence and need to be accounted for in analysis. Furthermore, some sequences are amplified more efficiently than others. This can become a problem when performing PCR on a complex mixture of sequences, such as a genome, leading to a skewed representation.

2.4. Sequencing

A fundamental omics technology is sequencing. Sequencing is defined as the process of determining the order of the building blocks in DNA, RNA or proteins. In this section, I will focus on the sequencing of DNA. The output of this is a *read*, an inferred sequence of bases. Importantly, this read can contain errors, variants not found in the original molecule, introduced in the process of sequencing or before that.

DNA sequencing was first performed by Ray Wu in the early 1970s. [134,135] Though the method was not widely used, this inspired a method called *Sanger sequencing*[136,137] that is still used today. The first developed sequencing methods were rather limited in throughput. As such it has been largely replaced by *massively parallel sequencing* (MPS)* technologies able to sequence millions of bases in a single run. The first MPS technology was developed in 2005 by 454 Life Sciences. [129] Contemporary MPS technologies have evolved into two major groups, *short-read sequencing* and *long-read sequencing*, each with its own applications.

When comparing sequencing platforms it is common to evaluate their *read accuracy*. The *read accuracy* describes the rate at which a base in

*Massively parallel sequencing is also known as *next generation sequencing* (NGS).

the read DNA molecule is correctly called by the instrument. Accuracy is important for many applications, such as variant calling which is covered in Chapter 3.3 where errors can lead to untrue discoveries.

Since the introduction of MPS, there have been massive developments in how DNA, or other starting material, is prepared to extract sequencing information. This process is referred to as *library preparation*, where a *library* is a collection of DNA fragments ready for sequencing. In library preparation, the sample is manipulated over several steps to extract the requested information as a DNA library for sequencing. This process further needs to be tailored to the sequencing technology and platform that is used.

2.4.1. Sequencing platforms

Although 454 Life Sciences was the first to commercialize MPS sequencing, the field has long been dominated by Illumina (Figure 10). Illumina's sequencing-by-synthesis (SBS) chemistry, launched in 2006[138], remains the most widely used sequencing technology today.

Short-read sequencing platforms, including Illumina, are characterized by short reads (50-600 bp) and high accuracy (>99%)[139,140]. Over the years, these platforms have made sequencing progressively cheaper per base, costing about $10 per gigabase-pair (Gbp = 1,000,000,000 bp) in 2022.[141] The latest Illumina instrument, the NovaSeq X Plus, is stated to sequence up to 16 terabase-pairs (Tbp = 1,000,0000,000,000 bp) per run equivalent to about 128 human genomes*[142], priced at about $2 per Gbp[143]. Competitors are also driving prices lower, with both MGI and

*This number represents a run on a single flowcell and 30X coverage per human genome equaling ~120Gbp.

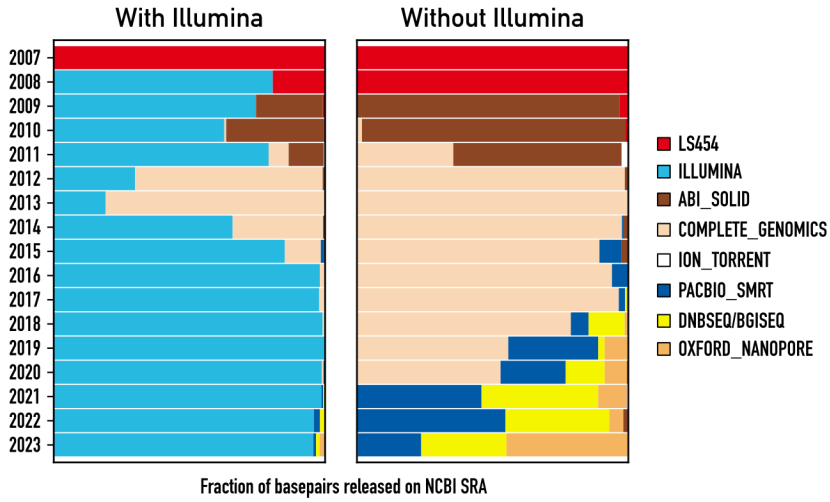

Figure 10: Illumina's dominance among sequencing platforms in human genomics from 2007 to 2023 shown as a fraction of base pairs released on NCBI's Sequence Read Archive (SRA). Data only considers human genome data and MPS platforms. Helicos BioSciences (HELICOS) was excluded due to low usage. Data acquired from ncbi.nlm.nih.gov/sra 2024-01-16 using query `("Homo sapiens"[orgn]) AND ("strategy wgs"[Properties])`.

newcomer Ultima Genomics stating prices of approximately $1 per Gbp for their latest instruments[144,145]

The short read sequencers are suitable for a lot of different applications, such as calling of short variants or quantification of transcripts. Some applications are however dependent on a wider "field of view" than these platforms provide. For example, characterization of transcript isoforms, *de novo* genome assembly, structural variant calling and haplotyping all benefit from long (>10,000 bp) reads. This has led to the development of long-read sequencing platforms such as *single-molecule real-time (SMRT) sequencing* by Pacific Biosciences (PacBio) and *nanopore sequencing* by Oxford Nanopore Technologies (ONT).

Long-read sequencing platforms have historically been characterized as low accuracy (<90%).[146,147] Recent progress has however improved accuracy greatly. PacBio introduced *HiFi* reads by sequencing the same 10-30 kilobase-pairs (kbp = 1,000 bp) DNA fragment multiple times to create a consensus sequence with high accuracy (>99%)*.[148] ONT has also seen a gradual improvement of their nanopore technology, refining the chemistry, base-calling algorithm and through the use of consensus sequencing†, increasing accuracy from ~60% to ~99%.[149–151] Nanopore also allows for *ultra-long* read sequencing generating reads longer than 100 kbp some even megabase pairs (Mbp = 1,000,000 bp) long.[152,153] Unlike most short-read platforms that get their signal from multiple copies of the same sequence, these methods are *single-molecule* making the improvement the all more impressive. Long reads are considerably more expensive compared to short-read sequencing.[154] With PacBio's latest instrument the cost is about $1000 per human genome or ~$11 per Gbp[155], 5-10 times the cost of emerging short-read sequencing platforms. Still, the advances in long-read sequencing are a great boon, enabling us for the first time to create complete haplotype-resolved genomes.[156,157]

*This process is called *circular consensus sequencing* (CCS) and generates CCS reads each containing the one or more subreads of the same DNA fragment. In each CCS read, the subreads are aligned to generate a consensus sequence which is quality filtered (> 99% accuracy) to get as HiFi reads.

†There are many versions of this approach, called *2D*, *1D*, *1D²* and *duplex* sequencing, but they all rely on reading both strands of the same DNA fragment to generate a higher accuracy consensus read.

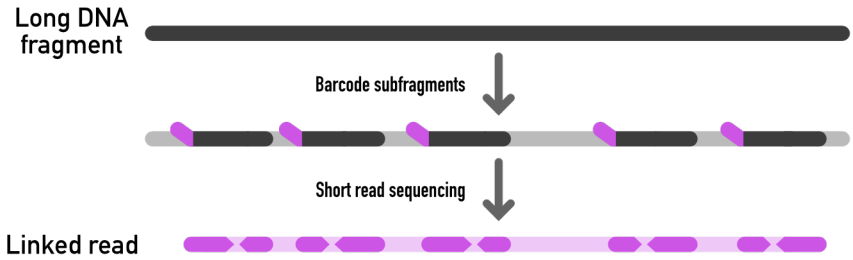

Figure 11: Illustration of linked-read sequencing. Long DNA molecules are fragmented and each subfragment is tagged with a barcode unique to the long DNA molecule. The barcoded subfragments are then sequenced and the reads can be linked together using the barcode information.

2.4.2. Linked-read sequencing

Alternative to performing long-read sequencing, various preparation methods can encode long-range information into a DNA fragment suitable for short-read sequencing. In *linked-read sequencing*, this is done by barcoding short DNA fragments from the same long DNA molecule (Figure 11).[158,159] There are also other approaches such as *Hi-C*[160] and *Strand-seq*[161,162], but these will not be covered here.

Linked-read sequencing is distinct from *synthetic long-read sequencing*. In synthetic long-read sequencing the objective is to regenerate the original molecule from the short reads. This is done by generating overlapping reads from the same molecule, with methods such as *Moleculo*[*][163], Illumina's newly launched *Complete Long Reads*[†] and Element Biosciences' *LoopSeq*. Since overlapping reads are used to reconstruct the original molecule before it can be used for further analysis, synthetic long-read sequencing requires a lot more reads. Linked-read

[*]Moleculo was historically sold by Illumina as *TruSeq Synthetic Long-Reads*.
[†]Illumina Complete Long Reads were initially marketed as *Infinity* reads.

sequencing on the other hand does not require overlapping reads on the same molecule, but rather uses barcodes to connect reads in analysis.

A wide selection of linked-read technologies have been developed over the years. Several technologies were developed by companies, including 10X Genomics' *GemCode* and *Chromium Genome* technologies[164,159], MGI's *stLFR*[149], Illumina's *CPT-seq*[158] and *CPTv2-seq*[165] and Universal Sequencing's *TELL-Seq*[166]. A few have also been developed by academic groups, including *Haplotagging*[167] and *DBS*[168].

The droplet barcode sequencing (DBS) approach uses droplet emulsions to generate linked reads. In DBS, long DNA molecules are fragmented with bead-linked transposomes. The transposomes both fragment and insert fixed sequences (*adapters*)* into the DNA at regular intervals, all while holding the DNA in place on the bead surface. The DNA-bead complexes are then encapsulated along with barcode oligonucleotides into droplets using the shake emulsion technique. In each droplet, a single barcode oligonucleotide is PCR amplified and then linked to the DNA fragment through the adapters. In **paper III**, the DBS method for generating linked reads has been further refined, along with the introduction of a new analysis pipeline. The DBS linked-read technology has further been applied to analyze cancer genomes in **paper IV**.

2.5. Technologies converged - single-gestalt omics

A field where all of these technologies have truly converged is *single-gestalt omics*. Single-gestalt omics involve studying the genomes,

*Adapter sequences, sometimes also referred to as simply *handles*, are short known sequences that usually delimit functional sequences in the read, such as barcodes and genomic sequences.

epigenomes, transcriptomes, and/or proteomes of individual gestalts. In this context, gestalts are commonly limited to cells, nuclei, or extracellular vesicles. The field has exploded in the last decade, in particular single-cell transcriptomics[169,170]. Studies have now achieved throughputs capable of characterizing entire organs, with atlases covering millions of cells[171]. With the maturation of single-cell transcriptomics, there has been growing interest in exploring other omics, including DNA, chromatin accessibility, and, of course, proteins.[172]

2.5.1. Single-gestalt proteomics

Proteomics brings many challenges compared to other omics fields.[173] Proteins cannot be amplified like DNA or RNA, meaning that the amount of protein at the start is all we get. Proteins are also highly complex molecules, with many different forms and modifications. A deep proteome analysis of a single cell type, HeLa, revealed approximately 14,000 proteoforms.[174] Furthermore, proteins have a wide dynamic range, with some proteins being present in millions of copies while others only in a few copies.[174] It is estimated that there are about 2-4 million proteins per femtoliter in a cell[175], which is about 100 million proteins for a HeLa cell* and 10,000 proteins if inferred for an EV†.

Several methods are currently being used to study the proteome of single gestalts, but two are at the current forefront, mass spectrometry and sequencing.[176] Mass spectrometry is the current gold standard for proteomics but is limited in throughput and sensitivity. For example, the

*Based on a HeLa cell volume of ~3,000 femtoliter.
†Based on an extracellular vesicle with a diameter of 100 nm.

recently published plexDIA method can quantify about 1,000 different proteins per cell, but only at a rate of about 144 cells per day.[177]

Unlike mass spectrometry, sequencing-based approaches have a great track record when it comes to gestalt throughput. Several single-cell sequencing methods for proteomics have been developed[176], including Abseq[113], Quantum barcoding (QBC)[178] and QBC2[179]. A lot of sequencing-based methods further integrate with other omics which is another benefit with this approach[180]. Examples are methods such as CITE-seq[181] and REAP-seq[182] integrating transcriptomics. For single-EVs, the field is still in its infancy, with only a few methods published[109,183,184].

Sequencing-based methods rely on a panel of affinity reagents, most commonly antibodies, that can bind specific protein targets. These affinity reagents are conjugated to a DNA oligonucleotide, containing a barcode sequence encoding the protein target information. Sometimes this oligonucleotide also includes a UMI for quantification. This oligonucleotide can be both amplified, to enhance sensitivity, and read through sequencing. Affinity reagents are however not without problems, often needing careful evaluation before integration into a panel. Currently available antibody panels, such as those provided by BioLegend, are limited to a few hundred proteins, a small fraction of the proteome. Another problem with affinity reagents is *cross-reactivity*, meaning that they exhibit some nonspecific binding to different proteins. This problem is further exacerbated when the panel size is expanded. Cross-reactivity can be mitigated by using multiple antibodies for the same protein,

for example by implementing proximity extension as in the SPARC method[185].

In **paper I** and **paper II** we have developed two associated sequencing-based methods for proteomics profiling of single EVs and single cells, respectively.

CHAPTER 3.

From bytes to biology

So you finally have your sequencing data, now what? In your computer sits a bunch of files containing millions - perhaps billions - of reads. How do you make sense of it? This is where *bioinformatics* comes in, combining biology and computer science to analyze biological data. In this chapter, I will go through some of the analysis steps when studying genome and single-gestalt sequencing data.

3.1. Sequencing data

That file you are looking at is probably a *FASTQ* file, the *de facto* standard format for storing sequencing data. FASTQ files contain both the reads and their associated quality. The quality is encoded for each read base using the *Phred score*[186], which is a logarithmic scale where the score Q is given by $-10 * \log(P)$ where P is the probability of the base being called wrong. For example, a base with a Phred score of 20 has a 1% chance of being wrong and with 30 it has a 0.1% chance. In practice, this score is

encoded using ASCII characters. Each read also has a unique identifier. A related format is *FASTA* which is missing sequence quality information. There are other formats for storing sequencing data at this stage, such as *uBAM* and *FAST5*.

In some cases the sequence is read from two positions, generating *paired-end* reads for each fragment, as opposed to *single-end* reads. These paired-end reads are either stored in the same (interleaved) or separate (paired) FASTQ files, but are linked together using the read identifier.

3.2. Pre-processing and quality control

The first step in any sequencing analysis is processing the raw data to make it suitable for downstream analysis. This usually involves separating the reads into their respective samples, trimming, filtering, and quality control.

If multiple samples are sequenced together, the first step is to separate the reads into their respective samples. This is done by *demultiplexing* the reads, where each read is assigned to a sample based on their barcode (index) sequence. Demultiplexing is usually handled by the sequencing instrument but can also be done separately, using software tools such as *bcl2fastq* for Illumina sequencing data.

Quality control is also done at this stage to ensure that the reads are of sufficient quality to proceed. Common software tools for this include *FastQC*[187] and *fastp*[188]. Typical quality metrics include Phred score distribution per-base and per-read, GC content distribution, duplication rate, and presence of overrepresented or adapter sequences. Some of the issues identified here can be corrected using trimming and filtering.

Trimming and filtering are performed to get a set of high-quality sequences to proceed with. Trimming removed bases from the start or the end of the read, removing low-quality bases and/or adapter sequences. If the read is deemed sufficiently bad it may also be appropriate to filter it out. This is done using software tools such as *Cutadapt*[189] and *fastp*[188]. After this, it is common to perform quality control again to ensure that the trimming and filtering were successful.

3.3. Genome analysis

The goal of genome analysis is to get information about genetic variants present in the studied genome. For this purpose, one often performs *whole genome sequencing* (WGS). In this process a DNA sample is extracted from the subject and, following different preparations, sequenced to generate reads. There are many approaches for this, some mentioned in Chapter 2.4. Furthermore, multiple methods can be combined to get a more complete picture balancing the strengths and weaknesses of each method.

How many reads do we need? This depends on the sequencing platform and the intended goal. A common practice is to sequence to a certain *coverage*, meaning the average number of times a base is sequenced. For example, at 30X coverage, which is a common target, each base has been sequenced on average 30 times. This would require about 90 Gbp of data for a human genome, corresponding to ~300 million Illumina reads (300 bp) or ~9 million PacBio HiFi reads (10 kbp). Coverage may need to be higher, e.g. $> 60X$ to detect rare somatic variants or to compensate for low-quality reads. Coverage may also be lower, e.g. <1X if the only goal is to detect large CNVs[190].

The process for uncovering genomic variation is usually broken into reference *alignment*, *variant calling* and, increasingly, *phasing*.

3.3.1. Human reference genome

Genetic variants only come to light when comparing reads against another sequence. Reference genomes are therefore generated to represent a particular species. They provide a common frame of reference for comparing different samples and between different studies. The latest NCBI RefSeq Release (v220) currently lists genomes from 141,099 different organisms.[191]

A draft of the human reference genome was first published in 2001[46,47] and has since been updated several times. The Genome Reference Consortium (GRC) was founded in 2007 as a collective effort to release high-quality reference genomes. Their latest release of the human genome, GRCh38 was first released in 2017[192], and the latest patch was released in 2022[193]. Though GRCh38 is the most widely used reference it is not without problems, containing multiple gaps and misassembled sequences.[194] The first complete human genome, T2T-CHM13v2.0, was recently released by the telomere-to-telomere (T2T) consortium.[194,44]

While a reference genome is meant to be a representation of a particular species, it is inherently unable to comprehensively represent and capture all variations in that species.[195] This can also lead to biases when studying genomes from different populations, a recent study found roughly 300 Mb additional DNA in individuals of African descent, missing from GRCh38.[196] For this purpose a *pangenome* reference has been suggested, a rich graph structure in which multiple genomes can be

integrated.[197] The Human Pangenome Reference Consortium (HPRC) was founded in 2019, aiming to generate a pangenome reference from 350+ individuals of diverse origins to capture the major human variations into a lasting reference.[198] A first draft of the human pangenome was published in 2023 by HPRC, containing diploid assemblies from 47 different individuals.[156] Despite these efforts a lot of work remains, for example, the development of computational tools and visualization software to empower scientists. At the moment, the use of a linear reference like GRCh38, though flawed, is the *de facto* standard for most genome analyses.

3.3.2. Calling variants

There are two main paths to calling variants using a linear reference such as GRCh38, either the reads are compared directly to it, or the reads are independently assembled before any comparison.[199] In the first approach reads are matched to a particular position on the reference in a process called *sequence alignment* or *sequence mapping*. For each position the reads are then compared to the reference and any differences are identified as variants, and scored based on read support. The other approach uses *de novo* assembly. Here the reads are stitched together to recreate the genome from scratch (i.e. *de novo*). Examples of contemporary assembly tools include *Verkko*[200] and *Hifiasm*[201]. Assembly forms *contigs*, which can then be aligned to a reference genome similar to the first approach. Variants are then identified directly based on differences in the contigs compared to the reference.

Choosing between these approaches usually depends on a multitude of factors. Firstly, though assembly can be performed with short reads, in

practice long reads are extremely helpful and will often yield much better results. Mapping is on the other hand less dependent on read length. Assembly is also a considerably more computationally expensive process than mapping. The biggest gain with assembly is that it can identify variants in difficult regions, such as highly repetitive regions, where mapping has low accuracy, especially with short reads. Mapping is on the other hand very accurate in "easy" genomic regions, i.e. with few repeats, that make up the majority of the genome. Thus the choice of approach is often a trade-off between accuracy, completeness and cost.

3.3.3. Alignment

Alignment or mapping is the process of finding the best matching position of a read in a reference genome*. As such it is highly important for any downstream analysis. This process is highly dependent on the characteristics of the read type.[202] Longer reads are easier to find a unique position for. A long read is however only guaranteed to have a unique position if it can cover the longest repeated genomic sequence. [203] Alignment is also dependent on the read error rate, where allowing for more errors makes it harder to find a unique position for the read. Another difficulty is that alignment should work for reads containing true variations, including large SVs. The number of reads also presents a challenge here as the number of reads to align can be in the billions, requiring computationally efficient tools. Altogether this presents a challenging problem, but several tools have been developed to solve it.

*The terms alignment and mapping are often used interchangeably but have distinct meanings. *Mapping* is the high-level location of the read in the reference while *alignment* refers to the actual placement of the read bases compared to the reference sequence.

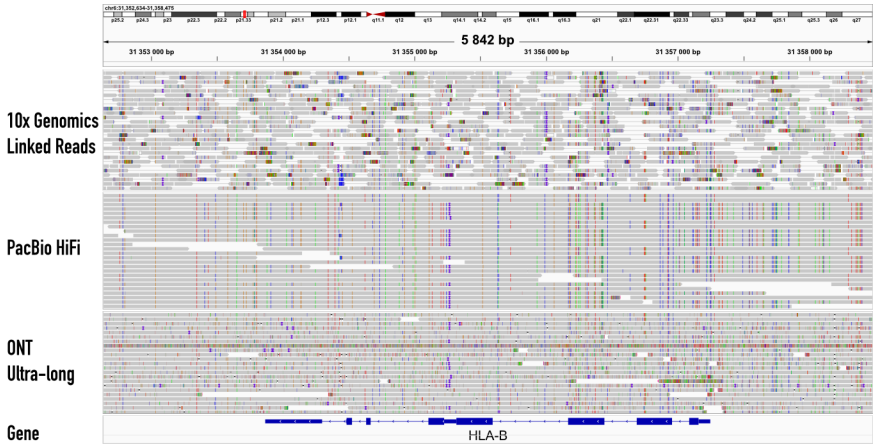

Figure 12: Alignment of different read types to the polymorphic region surrounding the *HLA-B* gene. Grey represents a perfect genome match with colors indicating differences in the reads either from variants or errors. 10X Genomics linked reads and long reads (PacBio HiFi and ONT Ultra-long) aligned to GRCh38 using *lariat*[208] (through Long Ranger) and *minimap2*[206] respectively. Data was acquired from GIAB for genome NA12878/HG001. Image generated using *IGV*[210].

There are many different alignment tools available, often tailored to a certain read type. The most common general-purpose tools for short reads are *BWA*[204] and *Bowtie2*[205] and for long reads *minimap2*[206] and *Winnowmap2*[207]. Long-read alignment tools can also be used for aligning assembled genomes. For linked reads there are purpose-built aligners such as *lariat*[208] and *EMA*[209]. Both are based on BWA but have been modified to optimize the placement of read pairs according to their molecule of origin, improving mapping in repetitive regions. [164,209]

The output from alignment is most commonly a *BAM* (binary alignment map) file, containing the alignment information for each read, including

the position, the mapping quality and a CIGAR string encoding how the bases were aligned. BAM file is a binary version of the *SAM* (sequence alignment map) format, making it both smaller and faster to read.[211] The BAM file can be processed in many different ways. For example, duplicate reads generated in PCR (see Chapter 2.3) can be identified using *Picard*[212], and reads can be filtered based on their mapping quality using *SAMtools*[211,213]. The aligned reads can also be visualized in a genome browser, such as *IGV*[210] (Figure 12).

3.3.4. Variant detection and filtration

Once the reads are aligned to a reference genome, variants are detected by finding differences. The variants then need to be filtered to separate true variants from artifacts. There have been many different tools developed for variant calling, often specific to the read type, variant size, and variant origin.

For short reads, variant calling is focused on short variants. The most common tools for finding short germline variants are *GATK HaplotypeCaller*[214] and *DeepVariant*[215]. To find somatic variants such as those occurring in tumors it is common to sequence both the tumor and normal (unaffected) tissues. These are analyzed jointly in tools like *Strelka2*[216], using the normal sample to filter out germline variants.

Detecting SVs with short reads is more challenging as large variants can often not be contained within the read length. For example, *GRIDSS*[217] uses read pair information such as split reads and discordant reads to extract and locally assemble reads to call SVs. Still, the ability to detect SVs is limited, with low sensitivity and many false positives.[218] Especially

larger insertions are difficult to capture.[219] The use of linked reads can somewhat improve the detection of SVs, for example with *LinkedSV*[220]. Long reads generally perform better for SV calling as they can span the entire variant. For example, *Sniffles*[219] and *NanoSV*[221] are efficient tools for detecting SVs using long reads.

Recently, long reads have been proven to detect short variants with SNV accuracies comparable to or even better than short reads.[222,223] Even high error-rate nanopore reads can compete with short reads using tools like *PEPPER-Margin-DeepVariant*, especially in difficult-to-map regions. [222] One weak spot for nanopore reads is the detection of short INDELs, which is significantly less accurate.[223,222]

3.3.5. Haplotype phasing

To get the full value of the variants found in a non-haploid genome it is important to know which appear on the same chromosome. For this, variant *phasing* needs to be performed.

There are many approaches to phasing. One approach is to perform *trio sequencing* where both parents are sequenced along with the child. This allows for variants to be phased based on the inheritance from the parents. This approach is however limited by the availability of parental DNA. Another approach is to use *population phasing* where haplotypes are inferred based on the variants found in a population.[224] This approach can however miss variants with low frequency in the population or, this is true for both approaches, miss *de novo* variants unique to the individual. Thus these approaches will yield at best limited haplotype information.

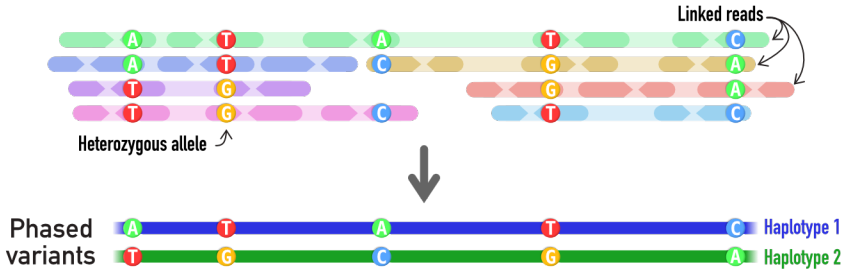

Figure 13: Visualization of variant phasing using long-range information, here from linked reads. Reads are aligned to a reference genome and variants are called. Heterozygous variants are then phased into haplotypes using long-range information.

[225] The best haplotype information is instead obtained by using *long-range information*.

Phasing using long-range information* relies on reads spanning multiple heterozygous variants. Using this information the path of the reads can be traced through the variants, allowing for variants to be phased into haplotypes (Figure 13). Long-range information can be acquired in several different ways as outlined in Chapter 2.4. The most common methods are long read sequencing followed by special library preparation techniques for short read sequencing such as linked-reads, Hi-C, and Strand-seq. Long and linked reads can generally cover more variants leading to more complete phasing.[218,226] They can also map more accurately to repetitive regions compared to Hi-C and Strand-seq. Hi-C and Strand-seq are however useful in hybrid approaches as they can span across centromeres, generating chromosome-scale haplotypes.[226,218,227,228]

*Sometimes referred to as *molecular haplotyping*.

In performing reference-based molecular haplotyping the reads need to be assigned to separate haplotypes, minimizing the number of erroneously assigned reads, known as the *minimum error correction* (MEC) problem. [229] This is a computationally challenging problem, but several tools have been developed to solve it. Examples include *WhatsHap*[230,231] and *HapCut2*[226]. Haplotype information can also be recovered without the use of a reference genome by performing diploid assembly. This is most commonly performed with long reads for *de novo* assembly. Hi-C, Strand-seq or trio information is then used to either split reads into haplotypes, e.g. *DipAsm*[232] or used to resolve the assembly graph, e.g. *Verkko*[200] and *Hifiasm*[201,233].

3.3.6. Phasing evaluation

To evaluate haplotype phasing, several metrics have been developed, often measuring *contiguity* and *accuracy*. Contiguity is measured as the length of the phased blocks[*], where a block is a contiguous region of the genome where variants are phased. A common metric is the block *N50* which is calculated by adding the block lengths in descending order taking the length of the block at 50% of the total phased assembly. Alternatively, the block *NG50* can be calculated by taking the length of the block where 50% of the genome is phased. Another metric is *AN50* which accounts for unphased variants in the block by multiplying the block length by the fraction of phased variants in the block.[234] The contiguity can also be shown as the *Nx-curve*, plotting the block length at different percentages of the phased assembly or genome. The benefit here is that it accounts for

[*]Many phasing contiguity metrics are borrowed from genome assembly evaluation as the phase block is in many ways comparable to a contig.

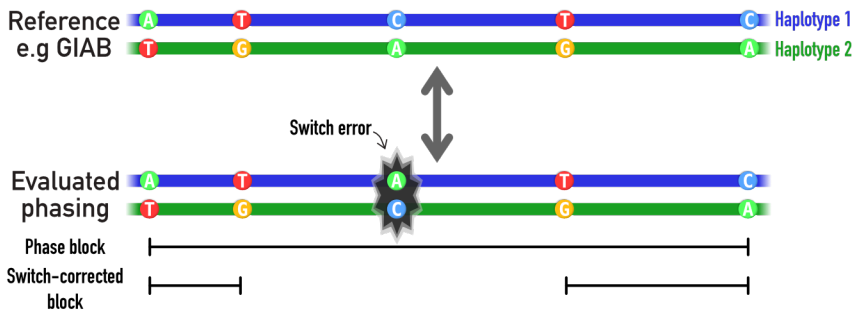

Figure 14: Visualization of switch errors in haplotypes. The top panel shows the true haplotypes, with the bottom panel showing the evaluated haplotypes. At the bottom, the extent of the phase blocks, before and after correcting for switch errors, is shown.

the full distribution of block lengths, N50 represents only a single point on this curve. To avoid a plot, the area under this curve can instead be calculated to get the *auN* metric.[235,236] This represents a more complete statistic for comparing contiguity.

Phasing accuracy is measured as the number of errors in the haplotypes, where an error is a variant that is assigned to the wrong haplotype, i.e. is switched (Figure 14). This requires a reference haplotype assembly to be evaluated, for example from Genome In A Bottle (GIAB) consortium[237]. The most common metric is the *switch error rate* which is calculated as the number of switches between haplotypes divided by the number of phased variants. Errors can also be divided depending on how many variants are affected into *long switch errors*, for multiple switched variants in a row, and *short switch errors*, for a single flipped variant*.

*Sometimes these types are called a *switch* and *flip* errors respective, for example when using `whatshap stats`[231].

Phasing contiguity and accuracy metrics can also be combined to get a single metric for phasing quality. For example, the *QAN50* metric is a version of AN50, where the blocks are split in two at switched positions. [234] In **Paper III** we use a similar approach to calculate switch-corrected N50 (QN50) and auN (auQN) metrics as well as plotting Nx-curves (QNx-curve).

3.4. Single-gestalt processing and analysis

The majority of single-gestalt processing and analysis can be broken down into two questions. How do I generate a count matrix? What can I do with a count matrix? It is clear that the count matrix is a central piece of this analysis. In the simplest terms, a count matrix is a table where each row represents a gestalt and each column a feature, some attribute of said gestalt. The features can be anything from genotypes to transcript or protein counts to chromatin accessibility sites. The values in the matrix are the counts of each feature for each gestalt. Using this matrix we can then perform a multitude of different analyses, answering various questions.

3.4.1. Extracting information

To convert raw sequencing data into a count matrix we need to extract different information from the reads. Each read contains both a gestalt barcode (e.g. cell barcode) and information about the feature. The gestalt barcode must be corrected for errors to group reads from the same gestalt, usually by comparing them to a list of known barcodes. The feature information needs to be extracted from the read, which can be done in several different ways depending on the read type and feature.

For example, in transcriptomics, the sequence needs to be mapped to a reference genome to identify the transcript. For proteomics, the protein information, also called antibody-derived tag (ADT), is encoded in a barcode separate from the gestalt barcode (see Chapter 2.5.1). This barcode is then compared against a list of barcodes used in the panel to identify the protein. In protocols using UMIs, these are also extracted and corrected. [238] For each gestalt and feature the number of reads/UMIs are then counted to generate the count matrix. This processing is typically part of a *pipeline* (see Chapter 3.5) using software tools like *CellRanger*[132] for 10X Genomics single-cell preparations.

3.4.2. Downstream analysis

Once the count matrix has been generated, downstream analysis can proceed. There are multiple tools for this purpose, a lot of which are implemented into analysis platforms such as *Seurat*[239] and *Scanpy*[240].

A good first step is to perform quality control across the gestalts and features. For the gestalts, we want to remove any that could be suspected to belong to the background (think droplets without gestalts). This can be done by removing gestalts with low overall feature count or few non-zero features. Sometimes features can also inform about the quality of the gestalt. For example, a gestalt with a high proportion of isotype control antibody is indicative of background.[241] Similarly features that are not expressed by the gestalts can be removed to reduce the dimensionality of the data. This is less common in single-gestalt proteomics where the number of features is usually limited*.

*ASAP-seq[242] used has the largest panel to date with 242 protein markers.

After the count matrix has been cleaned, the next step is normalization. Count variations between gestalts can for example be partially attributed to differences in sequencing depth rather than actual biology. There is also unwanted variation in the data, which in protein analysis mostly originate from free-floating antibodies and compartment-specific noise. [243] For this purpose, the *CLR* (centered log ratio) transformations have been used, in which counts are scaled using the geometric mean and then log-transformed.[181] An alternative approach is *dsb* where empty droplets are used to estimate and correct for the ambient noise.[243]

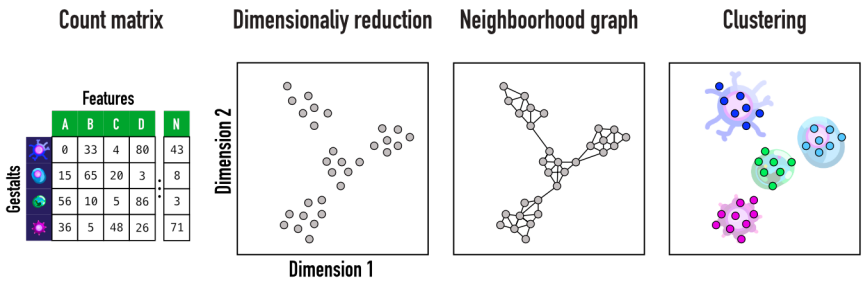

Figure 15: Overview of single-gestalt analysis from count matrix to clustering.

With the counts normalized, we can now investigate the data to find relationships between gestalts and features (Figure 15). Due to the high number of features this type of data suffers from the "*curse of dimensionality*". The meaning of this is that a lot of the variations observed in features are due to technical noise and randomness as opposed to meaningful information. Thus it is often appropriate to reduce the dimensionality of the data to find variations that constitute actual biological differences. In practice, this is done by *dimensionality reduction* which is a process of transforming the data into a low dimensional space

while preserving as much of the variation as possible. The most common method for this is linear reduction using *principal component analysis* (PCA). There are also non-linear dimensionality reduction methods such as *t-distributed stochastic neighbor embedding* (t-SNE)[244] and *uniform manifold approximation and projection* (UMAP)[244], which are especially useful to visualize the data in two dimensions.

A common analysis is the grouping of similar gestalts, for example, resolving different cell types among a population. For this, it is common to use graph-based approaches. The graph represents gestalts as nodes that connect to other nodes based on their similarity as measured by the distance between them. The distance is here a measure of the difference between two gestalts considering all their features or dimensions. For example, the length of a straight line connecting two dots on a paper (two-dimensional space) corresponds to the Euclidean distance. Once the graph has been constructed, it can then be used to cluster the gestalts into communities, for example using the *Leiden algorithm.*[245]. Here, gestalts in the same community (cluster) are more similar to each other than to gestalts in other communities. A word of caution is that the groups found by these algorithms are not necessarily biologically meaningful and can vary widely with parameter selection. Thus it is important to investigate the groups to determine if they are biologically meaningful, for example, based on known marker genes.

3.5. Pipelines

As we have seen, extracting information from a sequencing dataset often requires applying a long chain of software tools. This forms specific

computational workflows, often referred to as *pipelines*. These pipelines can be quite complex, sometimes requiring a lot of computational resources and time to run. For this, a high-performance computing (HPC) cluster or cloud computing is required to effectively run the analysis.

Pipelines come in many different forms, from simple bash scripts to complex workflow managers. While scripts are easy to construct they are not very flexible. As workflows grow to include more steps and run on more and more files this solution becomes increasingly untenable. Scripts can often only run locally, cannot resume from failed steps, lack documentation, and require manual installation, making them challenging to share and maintain. In extension, they make results hard, even impossible, to reproduce.[246] This is where workflow managers shine, providing *provenance* (tracking the steps and tools used to generate the results), *portability* (ability to run on different platforms e.g. HPCs, Cloud, local), *scalability* (handling resources and data of large size and quantity) and *re-entry* (ability to resume from the last successful step).[247,248]

Workflow managers break down the analysis into *rules* or *processes* that define the input, output, and commands to run. The manager then determines the order of execution based on the input/output dependencies between the rules. There are multiple different workflow managers to choose between, but three of the most popular are *Galaxy*[249], *Snakemake*[250,251] and *Nextflow*[252]. Galaxy caters more to people without much programming experience, providing a graphical user interface (GUI) for constructing workflows. Snakemake and Nextflow on the other hand are more geared towards people with

programming experience, using a *domain-specific language* (DSL) to construct workflows. While harder to learn, DSLs are more flexible and powerful than GUIs. Snakemake uses a Python-based DSL where the execution of rules is determined by working backward from the output files generated by each step. Nextflow instead uses a Groovy-based DSL where execution is determined by the input files. Both Snakemake and Nextflow have large active communities that provide a lot of support and resources. For example, the Nextflow-based nf-core community provides a large number of pipelines for different applications.[253]

Present investigation

What connects all papers present here is the technology used to study human variations - Droplet Barcode Sequencing (DBS). Here, gestalts and single barcode oligonucleotides are co-encapsulated in droplet emulsion. In each droplet the barcode, carrying one out of ~3.5 billion possible sequences, is PCR amplified and then linked to the "ome" under study. The barcode and ome-related sequences are then read using short-read sequencing. This method is highly flexible and affordable and can be used to study a wide range of gestalts, including single EVs (**Paper I**), single cells (**Paper II**) and long DNA molecules (**Papers III** and **IV**).

Papers I and **II** present novel methods to study surface proteomes. **Papers III** and **IV** rely on an existing method using DBS for linked-read sequencing[168]. **Paper III** presents an analysis pipeline for DBS linked reads to resolve haplotypes. In **paper IV** this pipeline was used for haplotyping colorectal cancer genomes to study somatic variations.

Paper I

Characterizing Single Extracellular Vesicles by Droplet Barcode Sequencing for Protein Analysis

Extracellular vesicles (EVs) released from multicellular organisms are highly heterogeneous. Variations on the level of single vesicles cannot be captured using current methods based on aggregate analysis. In this work, we have studied variations in the surface proteome of single small extracellular vesicles (sEVs), a subclass of EVs with a size below 200 nm. For this, we adopted the DBS technology for the study of proteins and single sEVs (Figure 16A). To study proteins we used antibodies conjugated

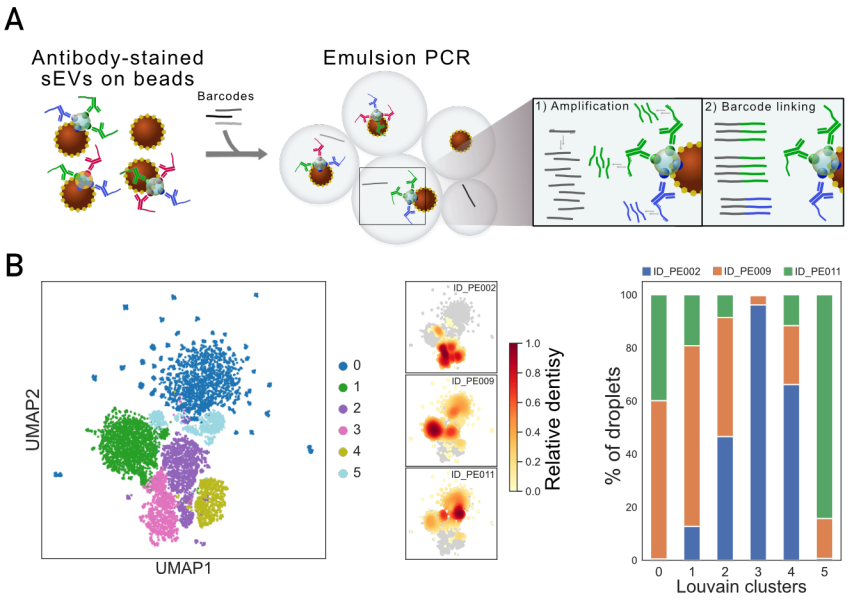

Figure 16: Profiling surface proteome of single sEVs. (**A**) Method workflow. (**B**) Heterogeneity in sEVs acquired from three lung cancer patients. UMAP plot (left) shows Louvain clustering of sEVs based on their surface protein profile. The relative density in the UMAP (middle) and proportions of the clusters (right) are also shown for each patient. *Adapted from Figure 1 and Figure 4C-E in Paper I.*

to an oligonucleotide that encodes both the antibody type and a UMI for quantification. sEVs were bound to magnetic beads to facilitate the removal of unbound antibodies. The sEVs were then compartmentalized using shake emulsions.

The method was validated using two hashing experiments to ensure that we did indeed capture single sEVs (A) on the beads and (B) in droplets. In both cases, we observed a low (<2%) mixing rate between the two groups.

We further validated the method using H1975 cell-line derived sEVs, collected over three separate time points. Using a panel of eleven

antibodies we analyzed ~50,000 sEVs. We confirmed that these sEVs on aggregate showed a high expression of CD9 and EGFR, previously observed in western blot and immuno-PCR. Being from the same cell line we expected the sEVs to be similar in profile, which was confirmed by principal component analysis.

Finally, we analyzed ~25,000 sEVs from malignant pleural effusion fluid from three non-small cell lung cancer patients using the same antibody panel. Using unsupervised Louvain clustering we were able to group the sEVs into putative subtypes, observing both patient-specific and shared subtypes (Figure 16B).

Paper II

Identification of Major Immune Cell Lineages with DBS-Pro

Cellular function and heterogeneity are inextricably linked to their protein expression. Here, the method from **paper I** was modified to be able to profile the surface proteome of single cells (Figure 17A). As cells are more fragile than sEVs we added a fixation step following staining and washing to prevent lysis. Shake emulsions were switched to microfluidics for monodisperse droplet generation. To enable better identification of empty droplets we further included an isotype control antibody in our panel.

In a proof-of-concept, this method was applied to profile peripheral blood mononuclear cells (PBMCs) using a panel of seven antibodies, including antibodies targeting CD3, CD4, CD8, CD19, CD14, and CD45. Despite low cell recovery, we were able to distinguish between major immune cell populations, such as $CD4^+$ T-cells, $CD8^+$ T-cells, monocytes and B-cells

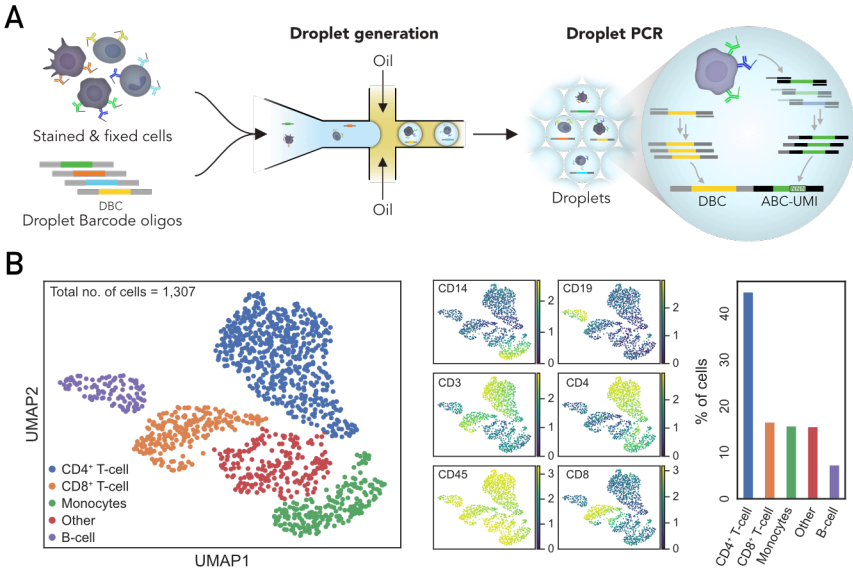

Figure 17: Profiling surface proteome on single cells (**A**) Method workflow following cell staining, washing and fixation. (**B**) Identification of major immune cell lineages from PBMCs. UMAP plot (left) shows annotated clusters generated using Leiden algorithm with log10 transformed counts of each marker (middle) and the relative proportions (right). Abbreviations: ABC = Antibody barcode, UMI = Unique molecular identifier. *Adapted from Figure 1 and Figure 3C in Paper II.*

at expected proportions (Figure 17B). While we recognize that further method improvement is necessary, it still presents a useful platform for single-cell proteomics.

Paper III

BLR: A Flexible Pipeline for Haplotype Analysis of Multiple Linked-read Technologies

Comprehensive capture of human genomic variations requires haplotype information. Linked-read sequencing is a promising approach, relying

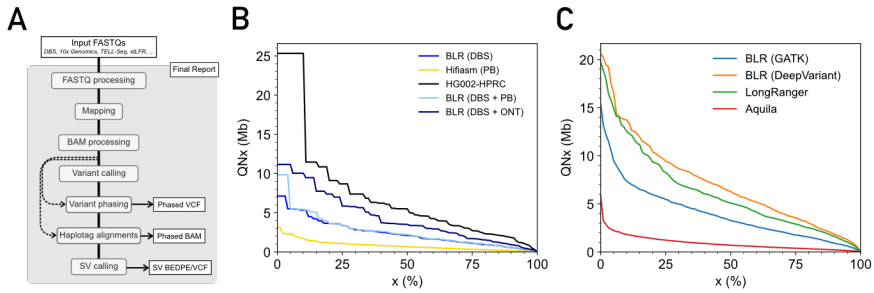

Figure 18: BLR pipeline for whole-genome haplotype analysis. (**A**) Pipeline overview. Plots show switch-error corrected phasing continuity on DBS (**B**) and 10X Genomics (**C**) linked reads. (**B**) BLR compared to Hifiasm diploid assembly with PacBio HiFi (PB) reads, the HPRC-HG002 reference assembly, and DBS integrated with 10X PB or 10X Oxford Nanopore ultra-long (ONT) reads. (**C**) BLR runs with GATK and DeepVariant calls compared to LongRanger and Aquila pipelines. *Adapted from Figure 1, Figure 2H and Figure 3C in Paper III.*

on cheap and accurate short reads while still preserving long-range information. In this work, we have developed BLR, an open-source pipeline for reference-based haplotype analysis using linked-read sequencing data (Figure 18A). The pipeline is designed to be flexible, allowing for the analysis of multiple linked-read technologies, including DBS, 10X Genomics, TELL-Seq and stLFR.

On DBS linked-reads from GIAB genome HG002 we showed that BLR generated highly continuous phasing with low (<0.2%) switch error rates, as compared to GIAB benchmark sets (v4.2.1) (Figure 18B). 98.6% of phased protein-coding genes were free of switch errors. Furthermore, large structural variant calls showed agreement when compared to calls from the Human Pangenome Reference Consortium (HPRC) HG002 reference assembly. Compared to diploid assembly using PacBio HiFi reads, BLR demonstrated improved phasing continuity when considering

switch errors. Additionally, we observed that incorporating long reads at low coverage enhanced phasing contiguity and reduced switch errors in tandem repeats.

We also analyzed linked reads generated using 10X Genomics, stLFR and TELL-Seq technologies with BLR. Running 10X Genomics linked reads through BLR and their own Long Ranger pipeline, the BLR phasing had a comparable switch error rate but more continuous phasing (Figure 18C). BLR further generated phase block lengths that were either longer or similar to those reported for TELL-Seq and stLFR linked reads, while maintaining low switch error rates.

Paper IV

Resolving the Haplotype Complexity of Colorectal Cancer Genomes with Droplet Barcode Sequencing

Cancer is to a large extent a disease of the genome. To resolve the cancer genome on a haplotype level, linked reads could be applied. Here we use DBS linked reads from paired normal and tumor tissue samples acquired from two colorectal cancer patients to haplotype-resolve somatic variants.

The samples were first run through BLR (**Paper III**) to generate haplotype-resolved genomes with haplotype blocks with a NG50 of 1.3 and 2.7 Mb. We then applied additional tools to study somatic variations at different scales, calling both CNVs, large (>2kb) SVs and short variants.

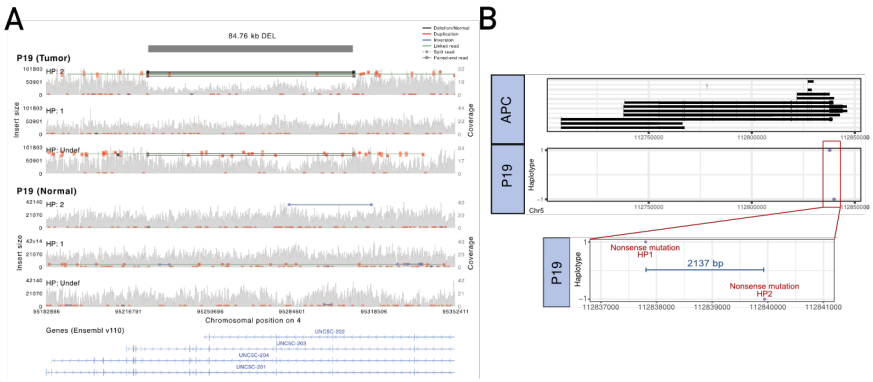

Figure 19: Haplotype-resolved analysis of colorectal cancer genomes. (**A**) Large deletion for patient P19 overlapping *UNC5C*. (**B**) Two nonsense mutations on separate haplotypes of established tumor suppressor gene *APC*. *Adapted from Figure 3B and Figure 4E in Paper IV.*

CNV calling revealed large aberrations on both samples, many of which are common in colorectal cancer. For example a gain of chromosome 8q arm containing the *MYC* oncogene, which was also confirmed to be upregulated by RNA-seq. From the large SV calls we found an 84 kb deletion overlapping *UNC5C* (Figure 19A), which is a putative tumor suppressor gene linked to cancer progression. We further found a large 31 kb inversion on chr20 that overlapped *SNHG17*, coding for a long non-coding RNA related to adverse CRC prognosis. Joint variant calling revealed 16,196 and 28,041 short somatic variants (mutations) in each patient. These affected genes in almost all TCGA-identified oncogenic pathways. We used the haplotype information to identify two-hit gene mutations. In one patient we found two nonsense mutations on separate haplotypes in *APC* (Figure 19B), a tumor suppressor gene well-known to colorectal cancer.

Future outlook

Single-EV analysis is quite a novel field and there are many avenues to explore. Sequencing-based approaches, like the one presented in **paper I**, show significant promise due to their ability to detect numerous proteins with high throughput. Analysis of single EVs is challenging as they carry a limited cargo compared to cells. However, this also limits the sequencing depth required to profile each vesicle, making analysis of millions of EVs considerably more affordable. Analysis at this scale might even enable systemic health insights from simple liquid biopsies. Another interesting avenue for development is to combine other omics modalities with single-EV proteomics, such as the detection of nucleic acids. RNA and DNA have been detected in bulk, but never in single EVs. This approach opens up new possibilities for comprehensive analysis with deeper insights into the molecular composition and nature of extracellular vesicles.

Unlike the single-EV field, contemporary sequencing-based single-cell methods are plenty and cover multiple "omes". While single-cell proteomics assays are less prevalent than transcriptomics, they are steadily gaining traction. Single-cell proteomics is predominantly performed in multiomics assays. While this is an information-rich approach, it is also expensive. Part of the cost is for library preparation with commercial kits being costly, often requiring dedicated single-purpose instruments. Our DBS approach presents a cheap alternative for library preparation. There are also multiple avenues for extending DBS to include other "omes". Back to the cost issue, the other major cost is for sequencing. This cost further needs to be scaled according to the number of cells and targets. Untargeted assays, such as polyA-

capture-based transcriptomics, are especially costly as they require a large number of reads per cell. While they enable the detection of novel targets, many reads are spent on high abundance and low-interest, e.g. housekeeping, targets. In contrast, targeted approaches, such as antibody-based proteomics, are cheaper as they require a smaller number of reads per cell. This approach makes analysis of large numbers of cells attainable. Therefore integration of DBS-based proteomics with targeted capture of relevant transcripts or genotypes is a promising avenue for future development.

The future of linked-reads sequencing currently hangs in the balance. The technology is definitely useful and has several applications besides the ones highlighted in this thesis. Compared to long reads, it is also cheaper and will likely remain so for the foreseeable future as Illumina and competitors continue to push down prices. The biggest hurdle is widespread adoption. 10X Genomics commercial solution was discontinued in 2019 leaving a gap in the market. While there are public protocols, such as DBS and Haplotagging and even other commercial solutions, such as Universal Sequencing's TELL-Seq and MGI's stLFR, they have not been widely used. This lack of adoption has further caused a slump in the development of linked-reads software tools while new tools for long reads are being developed at a rapid pace. Still, tools using linked reads are steadily being released and improved upon[*]. Tools like BLR (**Paper III**) certainly help in this regard. Within the current surge for long reads there might still be space for linked reads to expand. Only time will tell.

[*]I maintain an online list of linked-read related software (see github.com/pontushojer/awesome-linked-reads) which currently includes 67 different tools.

References

1. Betts H C, Puttick M N, Clark J W, Williams T A, Donoghue P C J, Pisani D. Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. Nat Ecol Evol. 2018;2:1556–62.

2. Joyce G F, Szostak J W. Protocells and RNA Self-Replication. Cold Spring Harbor Perspect Biol. 2018;10.

3. Brocks J J, Nettersheim B J, Adam P, Schaeffer P, Jarrett A J M, Güneli N, et al. Lost world of complex life and the late rise of the eukaryotic crown. Nature. 2023;618:767–73.

4. Bergström A, McCarthy S A, Hui R, Almarri M A, Ayub Q, Danecek P, et al. Insights into human genetic variation and population history from 929 diverse genomes. Science. 2020;367.

5. Schrödinger E. What is life? The physical aspect of the living cell and mind. Cambridge university press Cambridge; 1944.

6. Crick F. On protein synthesis. Symp Soc Exp Biol. 1958;12:138–63.

7. Crick F. Central Dogma of Molecular Biology. Nature. 1970;227:561–3.

8. Pontarotti G, Mossio M, Pocheville A. The genotype–phenotype distinction: from Mendelian genetics to 21st century biology. Genetica. 2022;150:223–34.

9. Watson J, Crick F. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature. 1953;171:737–8.

10. Reinsel D, Gantz J, Rydning J. IDC White Paper: The Digitization of the World - From Edge to Core [Internet]. 2018 Nov. Available from: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

11. Rutten M G T A, Vaandrager F W, Elemans J A A W, Nolte R J M. Encoding information into polymers. Nat Rev Chem. 2018;2:365–81.

12. Gerstein M B, Bruce C, Rozowsky J S, Zheng D, Du J, Korbel J O, et al. What is a gene, post-ENCODE? History and updated definition. Genome Res. 2007;17:669–81.

13. NCBI Homo sapiens Annotation Report Release GCF_000001405.40-RS_2023_10 [Internet]. 2023. Available from: https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Homo_sapiens/GCF_000001405.40-RS_2023_10

14. Morris K V, Mattick J S. The rise of regulatory RNA. Nat Rev Genet. 2014;15:423–37.

15. Aebersold R, Agar J N, Amster I J, Baker M S, Bertozzi C R, Boja E S, et al. How many human proteoforms are there?. Nat Chem Biol. 2018;14:206–14.

16. Dang X, Scotcher J, Wu S, Chu R K, Tolić N, Ntai I, et al. The first pilot project of the consortium for top-down proteomics: A status report. Proteomics. 2014;14:1130–40.

17. Method of the Year 2019: Single-cell multimodal omics. Nature Publishing Group; 2020. p. 1.

18. Zhu C, Preissl S, Ren B. Single-cell multimodal omics: the power of many. Nat Methods. 2020;17:11–4.

19. Crockford P W, On Y M B, Ward L M, Milo R, Halevy I. The geologic history of primary productivity. Curr Biol. 2023;33:4741–50.

20. Sagan L. On the origin of mitosing cells. J Theor Biol. 1967;14:255–74.

21. Hatton I A, Galbraith E D, Merleau N S C, Miettinen T P, Smith B M, Shander J A. The human cell count and size distribution. Proc Natl Acad Sci USA. 2023;120:e2303077120.

22. Kalluri R, Weinberg R A. The basics of epithelial-mesenchymal transition. J Clin Invest. 2009;119:1420.

23. Waddington C H. The strategy of the genes. George All & Unwin Ltd.; 1957.

24. Zeng H. What is a cell type and how to define it?. Cell. 2022;185:2739–55.

25. Théry C, Witwer K W, Aikawa E, Alcaraz M J, Anderson J D, Andriantsitohaina R, et al. Minimal information for studies of extracellular vesicles 2018 (MISEV2018): a position statement of the International Society for Extracellular Vesicles and update of the MISEV2014 guidelines. Journal of Extracellular Vesicles. 2018;7:1535750.

26. Witwer K W, Théry C. Extracellular vesicles or exosomes? On primacy, precision, and popularity influencing a choice of nomenclature. Journal of Extracellular Vesicles. 2019;8.

27. Gould S J, Raposo G. As we wait: coping with an imperfect nomenclature for extracellular vesicles. Journal of Extracellular Vesicles. 2013;2:20389.

28. Deatherage B L, Cookson B T. Membrane Vesicle Release in Bacteria, Eukaryotes, and Archaea: a Conserved yet Underappreciated Aspect of Microbial Life. Infect Immun. 2012;80:1948.

29. Anderson H C. Vesicles associated with calcification in the matrix of epiphyseal cartilage. J Cell Biol. 1969;41:59–2.

30. Harding C, Heuser J, Stahl P. Endocytosis and intracellular processing of transferrin and colloidal gold-transferrin in rat reticulocytes: demonstration of a pathway for receptor shedding. Eur J Cell Biol. 1984;35:256–63.

31. Pan B T, Teng K, Wu C, Adam M, Johnstone R M. Electron microscopic evidence for externalization of the transferrin receptor in vesicular form in sheep reticulocytes. J Cell Biol. 1985;101:942–8.

32. Pegtel D M, Gould S J. Exosomes. Annu Rev Biochem. 2019;88:487–14.

33. Pathan M, Fonseka P, Chitti S V, Kang T, Sanwlani R, Van Deun J, et al. Vesiclepedia 2019: a compendium of RNA, proteins, lipids and metabolites in extracellular vesicles. Nucleic Acids Res. 2019;47:516–9.

34. Wen S W, Lima L G, Lobb R J, Norris E L, Hastie M L, Krumeich S, et al. Breast Cancer-Derived Exosomes Reflect the Cell-of-Origin Phenotype. Proteomics. 2019;19:e1800180.

35. Kalluri R, LeBleu V S. The biology, function, and biomedical applications of exosomes. Science. 2020;367.

36. Mathieu M, Martin-Jaular L, Lavieu G, Théry C. Specificities of secretion and uptake of exosomes and other extracellular vesicles for cell-to-cell communication. Nat Cell Biol. 2019;21:9–7.

37. Théry C, Ostrowski M, Segura E. Membrane vesicles as conveyors of immune responses. Nat Rev Immunol. 2009;9:581–93.

38. Lone S N, Nisar S, Masoodi T, Singh M, Rizwan A, Hashem S, et al. Liquid biopsy: a step closer to transform diagnosis, prognosis and future of cancer treatments. Mol Cancer. 2022;21:1–2.

39. McKiernan J, Donovan M J, O'Neill V, Bentink S, Noerholm M, Belzer S, et al. A Novel Urine Exosome Gene Expression Assay to Predict High-grade Prostate Cancer at Initial Biopsy. JAMA Oncol. 2016;2:882–9.

40. McKiernan J, Donovan M J, Margolis E, Partin A, Carter B, Brown G, et al. A Prospective Adaptive Utility Trial to Validate Performance of a Novel Urine Exosome Gene Expression Assay to Predict High-grade Prostate Cancer in Patients with Prostate-specific Antigen 2–10ng/ml at Initial Biopsy. Eur Urol. 2018;74:731–8.

41. Piovesan A, Pelleri M C, Antonaros F, Strippoli P, Caracausi M, Vitale L. On the length, weight and GC content of the human genome. BMC Research Notes. 2019;12.

42. Klemm S L, Shipony Z, Greenleaf W J. Chromatin accessibility and the regulatory epigenome. Nat Rev Genet. 2019;20:207–20.

43. Bonev B, Cavalli G. Organization and function of the 3D genome. Nat Rev Genet. 2016;17:661–78.

44. Rhie A, Nurk S, Cechova M, Hoyt S J, Taylor D J, Altemose N, et al. The complete sequence of a human Y chromosome. Nature. 2023;621:344–54.

45. Wang M-J, Chen F, Lau J T Y, Hu Y-P. Hepatocyte polyploidization and its association with pathophysiological processes. Cell Death Dis. 2017;8:e2805.

46. Lander E S, Linton L M, Birren B, Nusbaum C, Zody M C, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–21.

47. Venter J C, Adams M D, Myers E W, Li P W, Mural R J, Sutton G G, et al. The Sequence of the Human Genome. Science. 2001;291:1304–51.

48. Kellis M, Wold B, Snyder M P, Bernstein B E, Kundaje A, Marinov G K, et al. Defining functional DNA elements in the human genome. Proc Natl Acad Sci USA. 2014;111:6131–8.

49. Abascal F, Acosta R, Addleman N J, Adrian J, Afzal V, Ai R, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583:699–10.

50. Christmas M J, Kaplow I M, Genereux D P, Dong M X, Hughes G M, Li X, et al. Evolutionary constraint and innovation across hundreds of placental mammals. Science. 2023;380:eabn3943.

51. Kent W J, Sugnet C W, Furey T S, Roskin K M, Pringle T H, Zahler A M, et al. The Human Genome Browser at UCSC. Genome Res. 2002;12:996–6.

52. Hoyt S J, Storer J M, Hartley G A, Grady P G S, Gershman A, Lima L G de, et al. From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. Science. 2022;376.

53. Storer J, Hubley R, Rosen J, Wheeler T J, Smit A F. The Dfam community resource of transposable element families, sequence models, and genome annotations. Mobile DNA. 2021;12:1–4.

54. Gemayel R, Cho J, Boeynaems S, Verstrepen K J. Beyond Junk-Variable Tandem Repeats as Facilitators of Rapid Evolution of Regulatory and Coding Sequences. Genes. 2012;3:461–80.

55. Hancks D C, Kazazian H H. Roles for retrotransposon insertions in human disease. Mobile DNA. 2016;7:1–8.

56. Bruder C E G, Piotrowski A, Gijsbers A A C J, Andersson R, Erickson S, Ståhl T D de, et al. Phenotypically Concordant and Discordant Monozygotic Twins Display Different DNA Copy-Number-Variation Profiles. Am J Hum Genet. 2008;82:763.

57. Li R, Montpetit A, Rousseau M, Wu S Y M, Greenwood C M T, Spector T D, et al. Somatic point mutations occurring early in development: a monozygotic twin study. J Med Genet. 2014;51:28–4.

58. Martincorena I, Campbell P J. Somatic mutation in cancer and normal cells. Science. 2015;349:1483–9.

59. Baer C F, Miyamoto M M, Denver D R. Mutation rate variation in multicellular eukaryotes: causes and consequences. Nat Rev Genet. 2007;8:619–31.

60. Zhang L, Vijg J. Somatic Mutagenesis in Mammals and Its Implications for Human Disease and Aging. Annu Rev Genet. 2018;52:397–19.

61. Escaramís G, Docampo E, Rabionet R. A decade of structural variants: description, history and methods to detect structural variation. Brief Funct Genomics. 2015;14:305–14.

62. Kunz B A, Ramachandran K, Vonarx E J. DNA Sequence Analysis of Spontaneous Mutagenesis in Saccharomyces cerevisiae. Genetics. 1998;148:1491–505.

63. Lieber M R. The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway. Annu Rev Biochem. 2010;79:181–11.

64. Bourque G, Burns K H, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al. Ten things you should know about transposable elements. Genome Biol. 2018;19.

65. Tonegawa S. Somatic generation of antibody diversity. Nature. 1983;302:575–81.

66. Stephens P J, Greenman C D, Fu B, Yang F, Bignell G R, Mudie L J, et al. Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. Cell. 2011;144:27–0.

67. Cagan A, Baez-Ortega A, Brzozowska N, Abascal F, Coorens T H H, Sanders M A, et al. Somatic mutation rates scale with lifespan across mammals. Nature. 2022;604:517–24.

68. Stratton M R, Campbell P J, Futreal P A. The cancer genome. Nature. 2009;458:719–24.

69. Berdan E L, Aubier T G, Cozzolino S, Faria R, Feder J L, Giménez M D, et al. Structural Variants and Speciation: Multiple Processes at Play. Cold Spring Harbor Perspect Biol. 2023;:a41446.

70. Auton A, Abecasis G R, Altshuler D M, Durbin R M, Abecasis G R, Bentley D R, et al. A global reference for human genetic variation. Nature. 2015;526:68–4.

71. Conrad D F, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010;464:704–12.

72. Pang A W, MacDonald J R, Pinto D, Wei J, Rafiq M A, Conrad D F, et al. Towards a comprehensive structural variation map of an individual human genome. Genome Biol. 2010;11:1–4.

73. Chiang C, Scott A J, Davis J R, Tsang E K, Li X, Kim Y, et al. The impact of structural variation on human gene expression. Nat Genet. 2017;49:692–9.

74. Scott A J, Chiang C, Hall I M. Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. Genome Res. 2021;31:2249–57.

75. Weischenfeldt J, Symmons O, Spitz F, Korbel J O. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet. 2013;14:125–38.

76. Nóbrega M A, Zhu Y, Plajzer-Frick I, Afzal V, Rubin E M. Megabase deletions of gene deserts result in viable mice. Nature. 2004;431:988–93.

77. Hanahan D, Weinberg R A. Hallmarks of Cancer: The Next Generation. Cell. 2011;144:646–74.

78. Hanahan D. Hallmarks of Cancer: New Dimensions. Cancer Discov. 2022;12:31–6.

79. Hanahan D, Weinberg R A. The Hallmarks of Cancer. Cell. 2000;100:57–0.

80. Bailey M H, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell. 2018;173:371–85.

81. Aaltonen L A, Abascal F, Abeshouse A, Aburatani H, Adams D J, Agrawal N, et al. Pan-cancer analysis of whole genomes. Nature. 2020;578:82–3.

82. Deutschbauer A M, Jaramillo D F, Proctor M, Kumm J, Hillenmeyer M E, Davis R W, et al. Mechanisms of Haploinsufficiency Revealed by Genome-Wide Profiling in Yeast. Genetics. 2005;169:1915–25.

83. Tewhey R, Bansal V, Torkamani A, Topol E J, Schork N J. The importance of phase information for human genomics. Nat Rev Genet. 2011;12:215–23.

84. Bansal V, Tewhey R, Topol E J, Schork N J. The next phase in human genetics. Nat Biotechnol. 2011;29:38–9.

85. Chan A P, Choi Y, Rangan A, Zhang G, Podder A, Berens M, et al. Interrogating the Human Diplome: Computational Methods, Emerging Applications, and Challenges. Haplotyping: Methods and Protocols. Springer US; 2023. p. 1–0.

86. Benacerraf B. Role of MHC Gene Products in Immune Regulation. Science. 1981;212:1229–38.

87. Jeffreys A J, Kauppi L, Neumann R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet. 2001;29:217–22.

88. Merriam-Webster.com. Gestalt [Internet]. 2024 [cited 2024Feb8]. Available from: https://www.merriam-webster.com/dictionary/gestalt

89. Binladen J, Gilbert M T P, Bollback J P, Panitz F, Bendixen C, Nielsen R, et al. The Use of Coded PCR Primers Enables High-Throughput Sequencing of Multiple Homolog Amplification Products by 454 Parallel Sequencing. PLoS One. 2007;2:e197.

90. Parameswaran P, Jalili R, Tao L, Shokralla S, Gharizadeh B, Ronaghi M, et al. A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. Nucleic Acids Res. 2007;35:e130.

91. Meyer M, Stenzel U, Myles S, Prüfer K, Hofreiter M. Targeted high-throughput sequencing of tagged nucleic acid samples. Nucleic Acids Res. 2007;35:e97.

92. Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods. 2012;9:72–4.

93. Kennedy S R, Schmitt M W, Fox E J, Kohrn B F, Salk J J, Ahn E H, et al. Detecting ultralow-frequency mutations by Duplex Sequencing. Nat Protoc. 2014;9:2586.

94. Hamady M, Walker J J, Harris J K, Gold N J, Knight R. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. Nat Methods. 2008;5:235–7.

95. Buschmann T, Bystrykh L V. Levenshtein error-correcting barcodes for multiplexed DNA sequencing. BMC Bioinf. 2013;14:1–0.

96. Hamming R W. Error detecting and error correcting codes. Bell System Technical Journal. 1950;29:147–60.

97. Zorita E, Cuscó P, Filion G J. Starcode: sequence clustering based on all-pairs search. Bioinformatics. 2015;31:1913–9.

98. Simpson E H. The Interpretation of Interaction in Contingency Tables. Journal of the Royal Statistical Society: Series B (Methodological). 1951;13:238–41.

99. Trapnell C. Defining cell types and states with single-cell genomics. Genome Res. 2015;25:1491.

100. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009;6:377–82.

101. Picelli S, Björklund Å K, Faridani O R, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods. 2013;10:1096–8.

102. Wang J, Fan H C, Behr B, Quake S R. Genome-wide Single-Cell Analysis of Recombination Activity and De Novo Mutation Rates in Human Sperm. Cell. 2012;150:402–12.

103. Xin Y, Kim J, Ni M, Wei Y, Okamoto H, Lee J, et al. Use of the Fluidigm C1 platform for RNA sequencing of single mouse pancreatic islet cells. Proc Natl Acad Sci USA. 2016;113:3293–8.

104. Jaitin D A, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, et al. Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. Science. 2014;.

105. Mereu E, Lafzi A, Moutinho C, Ziegenhain C, McCarthy D J, Álvarez-Varela A, et al. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. Nat Biotechnol. 2020;38:747–55.

106. Fan H C, Fu G K, Fodor S P A. Combinatorial labeling of single cells for gene expression cytometry. Science. 2015;347.

107. Goldstein L D, Chen Y-J J, Dunne J, Mir A, Hubschle H, Guillory J, et al. Massively parallel nanowell-based single-cell gene expression profiling. BMC Genomics. 2017;18:1–0.

108. Komatsu J, Cico A, Poncin R, Le Bohec M, Morf J, Lipin S, et al. RevGel-seq: instrument-free single-cell RNA sequencing using a reversible hydrogel for cell-specific barcoding. Sci Rep. 2023;13:1–1.

109. Wu D, Yan J, Shen X, Sun Y, Thulin M, Cai Y, et al. Profiling surface proteins on individual exosomes using a proximity barcoding assay. Nat Commun. 2019;10:1–0.

110. Klein A M, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. Cell. 2015;161:1187–201.

111. Macosko E Z, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 2015;161:1202–14.

112. Borgström E, Redin D, Lundin S, Berglund E, Andersson A F, Ahmadian A. Phasing of single DNA molecules by massively parallel barcoding. Nat Commun. 2015;6.

113. Shahi P, Kim S C, Haliburton J R, Gartner Z J, Abate A R. Abseq: Ultrahigh-throughput single cell protein profiling with droplet microfluidic barcoding. Sci Rep. 2017;7:1–2.

114. Briggs A W, Goldfless S J, Timberlake S, Belmont B J, Clouser C R, Koppstein D, et al. Tumor-infiltrating immune repertoires captured by single-cell barcoding in emulsion. bioRxiv [Preprint]; 2017.

115. Erlich Y, Chang K, Gordon A, Ronen R, Navon O, Rooks M, et al. DNA Sudoku—harnessing high-throughput sequencing for multiplexed specimen analysis. Genome Res. 2009;19:1243–53.

116. Prabhu S, Pe'er I. Overlapping pools for high-throughput targeted resequencing. Genome Res. 2009;19:1254–61.

117. Patterson N, Gabriel S. Combinatorics and next-generation sequencing. Nat Biotechnol. 2009;27:826–7.

118. Cusanovich D A, Daza R, Adey A, Pliner H A, Christiansen L, Gunderson K L, et al. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. Science. 2015;348:910–4.

119. Rosenberg A B, Roco C M, Muscat R A, Kuchina A, Sample P, Yao Z, et al. SPLiT-seq reveals cell types and lineages in the developing brain and spinal cord. Science (New York, NY). 2018;360:176.

120. Kuchina A, Brettner L M, Paleologu L, Roco C M, Rosenberg A B, Carignano A, et al. Microbial single-cell RNA sequencing by split-pool barcoding. Science. 2021;371.

121. Quinodoz S A, Ollikainen N, Tabak B, Palla A, Schmidt J M, Detmar E, et al. Higher-order inter-chromosomal hubs shape 3D genome organization in. Cell. 2018;174:744.

122. Adey A, Kitzman J O, Burton J N, Daza R, Kumar A, Christiansen L, et al. In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. Genome Res. 2014;24:2041–9.

123. Bose S, Wan Z, Carr A, Rizvi A H, Vieira G, Pe'er D, et al. Scalable microfluidics for single-cell RNA printing and sequencing. Genome Biol. 2015;16:1–6.

124. Datlinger P, Rendeiro A F, Boenke T, Senekowitsch M, Krausgruber T, Barreca D, et al. Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. Nat Methods. 2021;18:635–42.

125. Teh S-Y, Lin R, Hung L-H, Lee A P. Droplet microfluidics. Lab Chip. 2008;8:198–20.

126. Clark I C, Abate A R. Microfluidic bead encapsulation above 20 kHz with triggered drop formation. Lab Chip. 2018;18:3598–605.

127. Matuła K, Rivello F, Huck W T S. Single-Cell Analysis Using Droplet Microfluidics. Adv Biosyst. 2020;4:1900188.

128. Hatori M, Kim S C, Abate A R. Particle-Templated Emulsification for Microfluidics-Free Digital Biology. Anal Chem. 2018;90:9813–20.

129. Margulies M, Egholm M, Altman W E, Attiya S, Bader J S, Bemben L A, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437:376–80.

130. Clark I C, Fontanez K M, Meltzer R H, Xue Y, Hayford C, May-Zhang A, et al. Microfluidics-free single-cell genomics with templated emulsification. Nat Biotechnol. 2023;41:1557–66.

131. Poisson S D. Recherches sur la probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilités. Bachelier; 1837.

132. Zheng G X Y, Terry J M, Belgrader P, Ryvkin P, Bent Z W, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:1–2.

133. Potapov V, Ong J L. Examining Sources of Error in PCR by Single-Molecule Sequencing. PLoS One. 2017;12.

134. Wu R. Nucleotide sequence analysis of DNA: I. Partial sequence of the cohesive ends of bacteriophage λ and 186 DNA. Journal of Molecular Biology [Internet]. 1970;51:501–21. Available from: https://www.sciencedirect.com/science/article/pii/0022283670900045

135. Wu R, Taylor E. Nucleotide sequence analysis of DNA: II. Complete nucleotide sequence of the cohesive ends of bacteriophage λ DNA. Journal of Molecular Biology [Internet]. 1971;57:491–11. Available from: https://www.sciencedirect.com/science/article/pii/0022283671901057

136. Sanger F, Coulson A R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 1975;94:441–8.

137. Sanger F, Nicklen S, Coulson A R. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA. 1977;74:5463–7.

138. Bentley D R, Balasubramanian S, Swerdlow H P, Smith G P, Milton J, Brown C G, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456:53–9.

139. Fox E J, Reid-Bayliss K S, Emond M J, Loeb L A. Accuracy of Next Generation Sequencing Platforms. Next generation, sequencing & applications. 2014;1.

140. Stoler N, Nekrutenko A. Sequencing error profiles of Illumina sequencing instruments. NAR Genomics Bioinf. 2021;3:lqab19.

141. Wetterstrand K. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) [Internet]. 2024. Available from: https://www.genome.gov/sequencingcostsdata

142. Illumina. NovaSeq X and NovaSeq X Plus Sequencing Systems - Specification Sheet [Internet]. 2023 Nov. Available from: https://www.illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/novaseq-x-series-spec-sheet-m-us-00197/novaseq-x-series-specification-sheet-m-us-00197.pdf

143. Illumina. NovaSeq X Series Enables Broader, Deeper Sequencing [Internet]. 2024. Available from: https://www.illumina.com/systems/sequencing-platforms/novaseq-x-plus/applications/broad-sequencing.html

144. MGI. Complete Genomics Drops Genome Sequencing Price to Sub $100 at AGBT General Meeting-MGI-Leading Life Science Innovation [Internet]. 2023. Available from: https://en.mgi-tech.com/news/375

145. Almogy G, Pratt M, Oberstrass F, Lee L, Mazur D, Beckett N, et al. Cost-efficient whole genome-sequencing using novel mostly natural sequencing-by-synthesis chemistry and open fluidics platform. bioRxiv [Preprint]; 2022.

146. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-Time DNA Sequencing from Single Polymerase Molecules. Science. 2009;.

147. Mikheyev A S, Tin M M Y. A first look at the Oxford Nanopore MinION sequencer. Mol Ecol Resour. 2014;14:1097–102.

148. Wenger A M, Peluso P, Rowell W J, Chang P-C, Hall R J, Concepcion G T, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019;37:1155–62.

149. Wang Y, Zhao Y, Bollas A, Wang Y, Au K F. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol. 2021;39:1348–65.

150. Rang F J, Kloosterman W P, Ridder J de. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol. 2018;19:1–1.

151. Ni Y, Liu X, Simeneh Z M, Yang M, Li R. Benchmarking of Nanopore R10.4 and R9.4.1 flow cells in single-cell whole-genome amplification and whole-genome shotgun sequencing. Comput Struct Biotechnol J. 2023;21:2352.

152. Jain M, Koren S, Miga K H, Quick J, Rand A C, Sasani T A, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018;36:338–45.

153. Miga K H, Koren S, Rhie A, Vollger M R, Gershman A, Bzikadze A, et al. Telomere-to-telomere assembly of a complete human X chromosome. Nature. 2020;585:79–4.

154. Logsdon G A, Vollger M R, Eichler E E. Long-read human genome sequencing and its applications. Nat Rev Genet. 2020;21:597–14.

155. PacBio. PacBio Announces Revio, a Revolutionary New Long Read Sequencing System Designed to Provide 15 Times More HiFi Data and Human Genomes at Scale for Under 1,000 USD - PacBio [Internet]. 2022. Available from: https://www.pacb.com/press_releases/pacbio-announces-revio-a-revolutionary-new-long-read-sequencing-system-designed-to-provide-15-times-more-hifi-data-and-human-genomes-at-scale-for-under-1000

156. Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, et al. A draft human pangenome reference. Nature. 2023;617:312–24.

157. Gao Y, Yang X, Chen H, Tan X, Yang Z, Deng L, et al. A pangenome reference of 36 Chinese populations. Nature. 2023;619:112–21.

158. Amini S, Pushkarev D, Christiansen L, Kostem E, Royce T, Turk C, et al. Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. Nat Genet. 2014;46:1343–9.

159. Zheng G X Y, Lau B T, Schnall-Levin M, Jarosz M, Bell J M, Hindson C M, et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nat Biotechnol. 2016;34:303–11.

160. Lieberman-Aiden E, Berkum N L van, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science. 2009;326:289–93.

161. Falconer E, Hills M, Naumann U, Poon S S S, Chavez E A, Sanders A D, et al. DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. Nat Methods. 2012;9:1107–12.

162. Sanders A D, Falconer E, Hills M, Spierings D C J, Lansdorp P M. Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. Nat Protoc. 2017;12:1151–76.

163. Kuleshov V, Xie D, Chen R, Pushkarev D, Ma Z, Blauwkamp T, et al. Whole-genome haplotyping using long reads and statistical methods. Nat Biotechnol. 2014;32:261–6.

164. Marks P, Garcia S, Barrio A M, Belhocine K, Bernate J, Bharadwaj R, et al. Resolving the full spectrum of human genome variation using Linked-Reads. Genome Res. 2019;29:635.

165. Zhang F, Christiansen L, Thomas J, Pokholok D, Jackson R, Morrell N, et al. Haplotype phasing of whole human genomes using bead-based barcode partitioning in a single tube. Nat Biotechnol. 2017;35:852–7.

166. Chen Z, Pham L, Wu T-C, Mo G, Xia Y, Chang P L, et al. Ultralow-input single-tube linked-read library method enables short-read second-generation sequencing systems to routinely generate highly accurate and economical long-range sequencing information. Genome Res. 2020;30:898–9.

167. Meier J I, Salazar P A, Kučka M, Davies R W, Dréau A, Aldás I, et al. Haplotype tagging reveals parallel formation of hybrid races in two butterfly species. Proc Natl Acad Sci USA. 2021;118:e2015005118.

168. Redin D, Frick T, Aghelpasand H, Käller M, Borgström E, Olsen R-A, et al. High throughput barcoding method for genome-scale phasing. Sci Rep. 2019;9.

169. Svensson V, Vento-Tormo R, Teichmann S A. Exponential scaling of single-cell RNA-seq in the past decade. Nat Protoc. 2018;13:599–4.

170. Aldridge S, Teichmann S A. Single cell transcriptomics comes of age. Nat Commun. 2020;11.

171. Yao Z, Velthoven C T J van, Kunst M, Zhang M, McMillen D, Lee C, et al. A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. Nature. 2023;624:317–32.

172. Kashima Y, Sakamoto Y, Kaneko K, Seki M, Suzuki Y, Suzuki A. Single-cell sequencing techniques from individual to multiomics analyses. Exp Mol Med. 2020;52:1419–27.

173. Single-cell proteomics: challenges and prospects. Nat Methods. 2023;20:317–8.

174. Bekker-Jensen D B, Kelstrup C D, Batth T S, Larsen S C, Haldrup C, Bramsen J B, et al. An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. cels. 2017;4:587–99.

175. Milo R. What is the total number of protein molecules per cell volume? A call to rethink some published values. BioEssays. 2013;35:1050–5.

176. Bennett H M, Stephenson W, Rose C M, Darmanis S. Single-cell proteomics enabled by next-generation sequencing or mass spectrometry. Nat Methods. 2023;20:363–74.

177. Derks J, Leduc A, Wallmann G, Huffman R G, Willetts M, Khan S, et al. Increasing the throughput of sensitive proteomics by plexDIA. Nat Biotechnol. 2023;41:50–9.

178. O'Huallachain M, Bava F-A, Shen M, Dallett C, Paladugu S, Samusik N, et al. Ultra-high throughput single-cell analysis of proteins and RNAs by split-pool synthesis. Commun Biol. 2020;3:1–9.

179. Sheng J, Hod E A, Vlad G, Chavez A. Quantifying protein abundance on single cells using split-pool sequencing on DNA-barcoded antibodies for diagnostic applications. Sci Rep. 2022;12:1–1.

180. Vandereyken K, Sifrim A, Thienpont B, Voet T. Methods and applications for single-cell and spatial multi-omics. Nat Rev Genet. 2023;24:494–15.

181. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay P K, Swerdlow H, et al. Simultaneous epitope and transcriptome measurement in single cells. Nat Methods. 2017;14:865–8.

182. Peterson V M, Zhang K X, Kumar N, Wong J, Li L, Wilson D C, et al. Multiplexed quantification of proteins and transcripts in single cells. Nat Biotechnol. 2017;35:936–9.

183. Banijamali M, Höjer P, Nagy A, Hååg P, Gomero E P, Stiller C, et al. Characterizing single extracellular vesicles by droplet barcode sequencing for protein analysis. J Extracell Vesicles. 2022;11:12277.

184. Ko J, Wang Y, Sheng K, Weitz D A, Weissleder R. Sequencing-Based Protein Analysis of Single Extracellular Vesicles. ACS Nano. 2021;15:5631–8.

185. Reimegård J, Tarbier M, Danielsson M, Schuster J, Baskaran S, Panagiotou S, et al. A combined approach for single-cell mRNA and intracellular protein expression analysis. Commun Biol. 2021;4:1–1.

186. Ewing B, Green P. Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. Genome Res. 1998;8:186–94.

187. Andrews S. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data [Internet]. 2010. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc

188. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884--i890.

189. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. 2011;17:10–2.

190. Dong Z, Xie W, Chen H, Xu J, Wang H, Li Y, et al. Copy-Number Variants Detection by Low-Pass Whole-Genome Sequencing. Current Protocols in Human Genetics. 2017;94:8.

191. NCBI - RefSeq Release 220 [Internet]. 2023. Available from: https://ncbiinsights.ncbi.nlm.nih.gov/2023/09/11/refseq-release-220

192. Schneider V A, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts P A, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 2017;27:849.

193. NCBI - Homo sapiens genome assembly GRCh38.p14 [Internet]. 2022. Available from: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.40

194. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze A V, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376:44–3.

195. Ebert P, Audano P A, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder M J, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science. 2021;372.

196. Sherman R M, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat Genet. 2019;51:30–5.

197. Consortium C P-G. Computational pan-genomics: status, promises and challenges. Briefings Bioinf. 2018;19:118–35.

198. Miga K H, Wang T. The Need for a Human Pangenome Reference Sequence. Annu Rev Genomics Hum Genet. 2021;22:81–2.

199. Olson N D, Wagner J, Dwarshuis N, Miga K H, Sedlazeck F J, Salit M, et al. Variant calling and benchmarking in an era of complete human genome sequences. Nat Rev Genet. 2023;24:464–83.

200. Rautiainen M, Nurk S, Walenz B P, Logsdon G A, Porubsky D, Rhie A, et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. Nat Biotechnol. 2023;41:1474–82.

201. Cheng H, Concepcion G T, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18:170–5.

202. Reinert K, Langmead B, Weese D, Evers D J. Alignment of Next-Generation Sequencing Reads. Annu Rev Genomics Hum Genet. 2015;16:133–51.

203. Li W, Freudenberg J. Mappability and read length. Front Genet. 2014;5:110803.

204. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25:1754–60.

205. Langmead B, Salzberg S L. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.

206. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.

207. Jain C, Rhie A, Hansen N F, Koren S, Phillippy A M. Long-read mapping to repetitive reference sequences using Winnowmap2. Nat Methods. 2022;19:705–10.

208. Bishara A, Liu Y, Weng Z, Kashef-Haghighi D, Newburger D E, West R, et al. Read clouds uncover variation in complex regions of the human genome. Genome Res. 2015;25:1570–80.

209. Shajii A, Numanagić I, Whelan C, Berger B. Statistical Binning for Barcoded Reads Improves Downstream Analyses. Cell Syst. 2018;7:219–265.

210. Robinson J T, Thorvaldsdóttir H, Winckler W, Guttman M, Lander E S, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29:24–6.

211. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

212. Picard toolkit [Internet]. Broad Institute; 2019. Available from: http://broadinstitute.github.io/picard/

213. Danecek P, Bonfield J K, Liddle J, Marshall J, Ohan V, Pollard M O, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10:giab8.

214. Poplin R, Ruano-Rubio V, DePristo M A, Fennell T J, Carneiro M O, Auwera G A Van der, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv [Preprint]; 2018. p. 201178.

215. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36:983–7.

216. Kim S, Scheffler K, Halpern A L, Bekritsky M A, Noh E, Källberg M, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods. 2018;15:591–4.

217. Cameron D L, Schröder J, Penington J S, Do H, Molania R, Dobrovic A, et al. GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res. 2017;27:2050.

218. Chaisson M J P, Sanders A D, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun. 2019;10:1–6.

219. Sedlazeck F J, Rescheneder P, Smolka M, Fang H, Nattestad M, Haeseler A von, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018;15:461–8.

220. Fang L, Kao C, Gonzalez M V, Mafra F A, Silva R Pellegrino da, Li M, et al. LinkedSV for detection of mosaic structural variants from linked-read exome and genome sequencing data. Nat Commun. 2019;10:1–5.

221. Cretu Stancu M, Roosmalen M J van, Renkens I, Nieboer M M, Middelkamp S, Ligt J de, et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nat Commun. 2017;8:1–3.

222. Shafin K, Pesout T, Chang P-C, Nattestad M, Kolesnikov A, Goel S, et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. Nat Methods. 2021;18:1322–32.

223. Olson N D, Wagner J, McDaniel J, Stephens S H, Westreich S T, Prasanna A G, et al. PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions. Cell Genomics. 2022;2:100129.

224. O'Connell J, Sharp K, Shrine N, Wain L, Hall I, Tobin M, et al. Haplotype estimation for biobank-scale data sets. Nat Genet. 2016;48:817–20.

225. Snyder M W, Adey A, Kitzman J O, Shendure J. Haplotype-resolved genome sequencing: experimental methods and applications. Nat Rev Genet. 2015;16:344–58.

226. Edge P, Bafna V, Bansal V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. Genome Res. 2017;27:801.

227. Porubsky D, Garg S, Sanders A D, Korbel J O, Guryev V, Lansdorp P M, et al. Dense and accurate whole-chromosome haplotyping of individual genomes. Nat Commun. 2017;8:1–0.

228. Tourdot R W, Brunette G J, Pinto R A, Zhang C-Z. Determination of complete chromosomal haplotypes by bulk DNA sequencing. Genome Biol. 2021;22.

229. Lippert R, Schwartz R, Lancia G, Istrail S. Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem. Briefings Bioinf. 2002;3:23–1.

230. Patterson M, Marschall T, Pisanti N, Iersel L van, Stougie L, Klau G W, et al. WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. J Comput Biol. 2015;22:498–9.

231. Martin M, Patterson M, Garg S, Fischer S O, Pisanti N, Klau G W, et al. WhatsHap: fast and accurate read-based phasing. bioRxiv [Preprint]; 2016. p. 85050.

232. Garg S, Fungtammasan A, Carroll A, Chou M, Schmitt A, Zhou X, et al. Chromosome-scale, haplotype-resolved assembly of human genomes. Nat Biotechnol. 2021;39:309–12.

233. Cheng H, Jarvis E D, Fedrigo O, Koepfli K-P, Urban L, Gemmell N J, et al. Haplotype-resolved assembly of diploid genomes without parental data. Nat Biotechnol. 2022;40:1332–5.

234. Duitama J, McEwen G K, Huebsch T, Palczewski S, Schulz S, Verstrepen K, et al. Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of Single Individual Haplotyping techniques. Nucleic Acids Res. 2012;40:2041–53.

235. Li H. auN: a new metric to measure assembly contiguity [Internet]. 2022. Available from: https://lh3.github.io/2020/04/08/a-new-metric-on-assembly-contiguity

236. Salzberg S L, Phillippy A M, Zimin A, Puiu D, Magoc T, Koren S, et al. GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res. 2012;22:557–67.

237. Wagner J, Olson N D, Harris L, Khan Z, Farek J, Mahmoud M, et al. Benchmarking challenging small variants with linked and long reads. Cell Genomics. 2022;2.

238. Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res. 2017;27:491.

239. Satija R, Farrell J A, Gennert D, Schier A F, Regev A. Spatial reconstruction of single-cell gene expression data. Nature Biotechnology. 2015;33:495–2.

240. Wolf F A, Angerer P, Theis F J. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19.

241. Swanson E, Lord C, Reading J, Heubeck A T, Genge P C, Thomson Z, et al. Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. eLife. 2021;.

242. Mimitou E P, Lareau C A, Chen K Y, Zorzetto-Fernandes A L, Hao Y, Takeshima Y, et al. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. Nat Biotechnol. 2021;39:1246–58.

243. Mulè M P, Martins A J, Tsang J S. Normalizing and denoising protein expression data from droplet-based single cell profiling. Nat Commun. 2022;13:1–2.

244. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv [Preprint]; 2018.

245. Traag V A, Waltman L, Eck N J van. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep. 2019;9:1–2.

246. Stodden V, Seiler J, Ma Z. An empirical analysis of journal policy effectiveness for computational reproducibility. Proc Natl Acad Sci USA. 2018;115:2584–9.

247. Perkel J M. Workflow systems turn raw data into scientific knowledge. Nature. 2019;573:149–50.

248. Wratten L, Wilm A, Göke J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. Nat Methods. 2021;18:1161–8.

249. Goecks J, Nekrutenko A, Taylor J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010;11:1–3.

250. Mölder F, Jablonski K P, Letcher B, Hall M B, Tomkins-Tinch C H, Sochat V, et al. Sustainable data analysis with Snakemake. F1000Research. 2021;10:33.

251. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics. 2012;28:2520–2.

252. Di Tommaso P, Chatzou M, Floden E W, Barja P P, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35:316–9.

253. Ewels P A, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020;38:276–8.

Abbreviations

BLR	Barcode Linked Reads
bp	Base pairs
CLR	Centered log-ratio
CNV	Copy number variation
DBS	Droplet barcode sequencing
DSL	Domain specific language
DNA	Deoxyribonucleic acid
EV	Extracellular vesicle
GIAB	Genome in a Bottle
GRC	Genome Reference Consortium
GUI	Graphical user interface
HLA	Human leukocyte antigen
HPC	High-performance computing
HPRC	Human Pangenome Reference Consortium
INDEL	Insertion/deletion
MNV	Multiple nucleotide variant
MPS	Massively parallel sequencing
NCBI	National Center for Biotechnology Information
NGS	Next generation sequencing
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences
PBMC	Peripheral blood mononuclear cell
PCA	Principal component analysis
PCR	Polymerase chain reaction
RNA	Ribonucleic acid

sEV	Small extracellular vesicle
SNV	Single nucleotide variant
SV	Structural variant
t-SNE	t-distributed stochastic neighbor embedding
T2T	Telomere-to-Telomere Consortium
TCGA	The Cancer Genome Atlas
UCSC	University of California Santa Cruz
UMAP	Uniform manifold approximation and projection
UMI	Unique molecular identifier
WGS	Whole genome sequencing

Acknowledgements

While a PhD can sometimes be a solitary effort, I am happy to say that it was not a lonely one. I am forever grateful to the many people who have helped me along the way. Sincerely, thank you for your support, encouragement, inspiration, and guidance.

`Det kommer att funka!`

`Afshin, *Troubleshooting session* (#289)`

The quote above illustrates the closing of a typical troubleshooting session. Before this, we had a 2-minute-turned-2-hour meeting, exploring every possible angle on the whiteboard, and are now sitting there. Me, trying to find any possible reason why it will not work, you, notably optimistic. **Afshin**, you have an astonishing ability to stay positive through all my concerns and doubts. I have especially enjoyed mapping it all out together on the whiteboard, solving problems, unfolding methods, and exploring ideas. Thank you for always being open and providing valuable knowledge and insights.

There are many others I want to thank. **Pelin**, I have you to thank for introducing me to such a great place as **Alfa 3**. You brought me in for my master thesis and let me stay on to explore research and method development. I am grateful for the opportunity to work with you and learn from you. **Tobias**, you are truly someone with a great coffee-to-work balance (**finger guns**). We started on Alfa 3 around the same time and it was always a pleasure to work with you. You shared my interest

in programming and I think we both pushed each other to become better. I am grateful for the many coffee/beer sessions and the many hours of coding together. **Žaneta**, you are such a fun person and I have really enjoyed hanging out with you in and outside of work. Sneaking into Folkhälsomyndigheten for Thursday pancakes, having a beer on the beach in Miami, skiing in Orsa or Åre, and many other moments. Maybe one day I will finally learn to say Na zdraví with everyone before having the first sip of my beer. **Ludvig**, my deskmate and *old* classmate who always has some interesting story to share. I have enjoyed our many discussions on everything and nothing. I hope that you will one day forgive me for turning your back into an abstract art piece in the burning Miami sun. **Sami**, we always seem to meet around the coffee machine (the good one). I guess we are both as excited for a fresh brew! Your compassion and caring for other people is immense, so large it shines through your Finnish grumpiness. **Mahsan**, you are truly a machine in the lab and I have really enjoyed working with you through some difficult projects. Your dedication and hard work have been truly inspiring. **Humam**, one has to look hard to find a kinder person than you. Whenever I need help or advice, you have been happy to assist me. Thanks for being a great colleague and friend, and for reviewing this thesis. **Marcel**, you have been somewhat of a mentor in helping me to take my experience in coding and software development to the next level. I am grateful for your guidance and support. **Parham**, I admire your quest for adventure, a passion almost as strong as your hate for Apple products. **Hoomam**, you always give good insights on things. The saffron you gifted me made the best tahdig I have ever made.

To add to this, I am grateful to **Olof** for your review of my thesis and for providing invaluable feedback. Your comments and suggestions have been truly helpful.

Thanks to all my collaborators and co-authors **Ábel Nagy**, **Amelie Eriksson Karlsström**, **Jan Lindros**, **Fredrick Stridfeldt**, **Apurba Dev**, **Petter Brodin**, **Jun Wang**, **Kristina Viktorsson**, **Rolf Lewensohn**, **Petra Hååg**, **Christiane Stiller**, **Håkan Jönsson**. Your knowledge and feedback on the projects we have worked on these past years have been invaluable.

Thanks to **NGI Stockholm** for providing sequencing services and to **UPPMAX** for providing computational resources. Without these, the work in this thesis would not have been possible.

Cheers to all my former **NGI** colleagues **Fran**, **Johannes**, **Orlando**, **Fanny**, **Phil**, **Ellen**, **Liqun**, **Anand**, **Nemo**, **Christian**, **Maxime**, **Joel**, **Mattias**, **other Mattias**, **Sofia**, **Magdalena**, **Lina**, **Anna**, **Barham**, **Chuan**, **Veronica**, **Jun**, **Elísabet**, **Pär**, **Salvo**, **Remi**, **Helena** who have given me encouragement over these years. I hope you will forgive me for leaving the **dark side**.

Cheers to **Eva**, **Lovisa**, **Konstantin**, **Simon**, **Ludvig B**, **Linda**, **Leire**, **Enikö**, **Hailey**, **Mengxiao**, **Reza**, **Franzi**, **Kim**, **Christine**, **Filip**, **Maja**, **Annelie**, **Marco**, **Marcos**, **Solène**, **Julia**, **Hamid**, **Nayanika**, **Raphaël**, **Alfred**, **Fernando**, **Karl-Johan**, **Kristina**, **Markus**, **Xesus**, **Artemy**, **Javier**, **Zakaria**, **Lorena**, **Karin**, **Jian**, **Julie**, **Krysztof**, **José**, **Martí**, **Sofia**, **Yilin**, **Samu**, **Anastasiya**, **Anniina**, **Shuai**, **Adelina**, **Serhat**, **Fitz**, **Ioanna**, **Meike**, **Jörg**, **Alma**, **Rapolas**, **Linnar**, **David**, **Mikaela**, **Sara**,

Max, **Peter**, **Patrick**, **Stefania**, **Ian**, **Anders**, **Joakim** and all the other people who I have shared time with on **Alfa 3** and made it so enjoyable and memorable.

Till sist, **Bea**. Vi har varit tillsammans länge och jag är så glad för varje ögonblick. Tack för att du alltid finns där för mig och för att ge mig stöd och kärlek. Jag älskar dig så mycket … men nu kan jag inte skriva mer då **Aska** jätte-jätte-jättegärna vill ha dragkamp med mig. Och det kan jag ju inte säga nej till!

Appendix (Papers I-IV)