From Department of Cell and Molecular Biology
Karolinska Institutet, Stockholm, Sweden

CELL LINEAGE ANALYSIS IN HUMANS

Joanna Hård

Stockholm 2021

All previously published papers were reproduced with permission from the publisher.
Published by Karolinska Institutet.
Printed by Universitetsservice US-AB, 2021

ISBN 978-91-8016-301-9

Cell lineage analysis in humans

THESIS FOR DOCTORAL DEGREE (Ph.D)

By

Joanna Hård

The thesis will be defended in public at Nils Ringertz, Floor 3, Biomedicum, Karolinska Institutet, Solnavägen 9, 171 65 Solna. August 20th 2021.

Principal Supervisor:
Professor Jonas Frisén
Karolinska Institutet
Department of Cell and Molecular Biology

Co-supervisor:
Professor Per-Olof Berggren
Karolinska Institutet
Department of Molecular Medicine and Surgery

Opponent:
Professor Johan Jakobsson
Lund University
Wallenberg Neuroscience Center
Lund Stem Cell Center

Examination Board:
Professor Karin Loré
Karolinska Institutet
K2 Department of Medicine, Solna

Professor Anna Wedell
Karolinska Institutet
Department of Molecular Medicine and Surgery

Professor Henrik Boström
Royal Institute of Technology, KTH
Division of Software and Computer Systems

To my friends and family

ABSTRACT

Delineating a cell’s history, where it came from and what has happened to it, can provide clues as to how tissues and organs are formed and function in the healthy state and in disease. The gold standard for tracing the relationships between cells and their progeny, is performed through prospective labeling with dyes or genetics tags. In this procedure, individual cells are labeled in order to track their clonal progeny at a later time point. These methodologies are, however, not applicable to study the composition of cell lineages in humans. Recently, the development of technologies for single cell sequencing has opened up the possibility to infer cell lineage relationships through the analysis of naturally occurring somatic mutations. During normal cell division, some new random mutations occur, forming an evolving barcode, which carries information about its developmental relationship to other cells. As such, the history of a cell is written in its genome, and every acquired mutation gets passed on to daughter cells. Shared somatic mutations may thus be used to trace backward across cell lineages, and the life history of an organism.

The goal of this thesis is to explore the possibility of using genetic variation as cell lineage marks to compute the evolutionary history of human cells as they divide. This work involves the development of new experimental and analytical methods, and the application of these to study the origins and lineage relationships of human cell populations. The methods and results described here, are intended to provide a contribution towards future applications for cell lineage tracing in man.

LIST OF SCIENTIFIC PAPERS

I. Mikael Rydén[*], Mehmet Uzunel[*], **Joanna Hård**[*], Erik Borgström, Jeff E Mold, Erik Arner, Niklas Mejhert, Daniel P Andersson, Yvonne Widlund, Moustapha Hassan, Christina V Jones, Kirsty L Spalding, Britt-Marie Svahn, Afshin Ahmadian, Jonas Frisén, Samuel Bernard, Jonas Mattsson, Peter Arner. Transplanted bone marrow derived cells contribute to human adipogenesis. Cell Metabolism. 2015;22(3):408-17.

II. **Joanna Hård**[*], Ezeddin Al Hakim[*], Marie Kindblom[*], Åsa K. Björklund, Bengt Sennblad, Ilke Demirci, Marta Paterlini, Pedro Reu, Erik Borgström, Patrik L. Ståhl, Jakob Michaelsson, Jeff E. Mold and Jonas Frisén. Conbase: a software for unsupervised discovery of somatic mutations in single cells through read phasing. Genome Biology. 2019;20(1):68.

III. **Joanna Hård**, Jeff Mold, Carl-Johan Eriksson, Pietro Berkes, Åsa Björklund, Bengt Sennblad, Jack Kuipers, Katharina Jahn, Jakob Michaelsson, Jonas Frisén. Human memory and effector CD8 T cell development after vaccination. Manuscript.

[*] Co-first authors

CONTENTS

1 Introduction.. 1

1.1 Life is composed of cellular hierarchies 1

1.2 Tracing lines of succession 2

1.3 Towards lineage tracing in humans 5

1.3.1 Nuclear genomic mutations 6

1.3.2 Mitochondrial mutations.. 8

1.4 Single cell sequencing.. 10

1.4.1 Single cell DNA sequencing.. 10

1.4.2 Single cell RNA sequencing 12

1.5 Computational methods 13

1.5.1 Alignment and variant calling.. 13

1.5.2 Mutation discovery in nuclear DNA 14

1.5.3 Phylogenomics 16

1.5.4 Cell lineage reconstruction based on mitochondrial mutations 17

2 Present investigation..19

2.1 General aims.. 19

2.2 Specific aims 19

2.3 Paper I 20

2.4 Paper II.. 22

2.5 Paper III... 24

3 Summary and perspectives...27

4 Acknowledgement..29

5 References 31

LIST OF ABBREVIATIONS

CNV	Copy number variation
DNA	Deoxyribonucleic acid
L1	Long interspersed nuclear element 1
MDA	Multiple displacement amplification
mtDNA	Mitochondrial DNA
PCR	Polymerase chain reaction
RNA	Ribonucleic acid
scATACseq	Single cell assay for transposase-accessible chromatin high-throughput sequencing
scRNAseq	Single cell RNA sequencing
SNV	Single nucleotide variant
TCR	T cell receptor
WGA	Whole genome amplification
WTA	Whole transcriptome amplification

1 INTRODUCTION

1.1 LIFE IS COMPOSED OF CELLULAR HIERARCHIES

Human life begins with a single fertilized egg cell which, upon numerous divisions gives rise to daughter cells that differentiate and ultimately form an entire organism. All information required for becoming a human being is stored within the genome of the fertilized egg cell. The genome consist of a set of chromosomes made up of the organic chemical deoxyribonucleic acid (DNA). DNA is inherited from the parents and will be copied and passed on to daughter cells each time a cell divides. The hereditary information in the DNA is encoded by a sequence consisting of combinations of four chemical bases, that together with a sugar molecule and a phosphate form individual units called nucleotides. The nucleotides are organized into two linear strands that together form a double helix, and the molecules made by cells are encoded in the nucleotide sequence within genes, which represent distinct regions of the genome. Signals within the cell and from other cells, both spatially close as well as cells far away, will result in that genes are transcribed into a similar type of molecule known as ribonucleic acid (RNA) which, in turn, is translated into proteins that carry out the functions of the cell.

Minor differences between cells, manifested by complex networks of protein interactions, will influence gene expression and make cells acquire characteristics that distinguish them from others. This may be sufficient to set even neighboring cells into dramatically distinct developmental paths. During the early stages of the embryo development, it is widely believed that all cells retain the potential to become any cell type in the body. These are multipotent stem cells, defined by their ability to self-renew as well as generating progeny of different types (Gage 2000). One goal in developmental biology is to chart the structure of this process, often depicted as a tree consisting of cell lineages that contribute to the formation of tissues and organs (Laurenti and Gottgens 2018).

As development proceeds, the plasticity of stem cells gradually diminishes (Laurenti and Gottgens 2018). During development, many organs are seeded with cells that later in life will function as adult stem cells with a restricted ability to generate only the types of cells that are specific for the tissue in which they reside. The role of adult stem cells is to maintain tissue homeostasis, by replacing worn-out and damaged cells (Post and Clevers 2019). One goal in stem cell research and regenerative medicine involves the characterization of cell lineages that are formed in the adult. It is essential to understand how these complex processes are orchestrated since impaired tissue regeneration may result in a variety of diseases for which there are limited or no currently available treatments, including for example cancer, neurodegenerative disease, diabetes mellitus and traumatic injuries. A major focus is directed towards the identification of cells with the capacity to replace others, including adult stem cells, since these may be targeted for enhancing the regenerative potential of the human body. For example, established therapies for stimulating blood cell production for the treatment of hematological malignancies are available (Richard and Schuster 2002). Future therapeutic applications for regenerative medicine may achieve cell replacement also in other organs (Spalding, Bhardwaj et al. 2005). An essential step towards reaching this goal involves the identification of the origins of new cells and characterization of the composition of cell lineages in humans.

1.2 TRACING LINES OF SUCCESSION

The first description of cells dates back to 1665, from the book 'Micrographia' (Hooke 1665). Its author, Robert Hooke, examined cork and wood under a microscope and observed that the tissues were composed of structurally organized microscopical units which reminded him of the tiny rooms, or *cellula*, where monks lived. At the end of the 19th century, improved developments of light microscopy made it possible to obtain a more detailed view of the life of cells than what had previously been possible. By directly observing the behavior of cells as an organism develops, scientists discovered that new cells arise from pre-existing cells. By drawing maps of how one cell gives rise to two daughter cells, which in turn give rise to cell lineages, these early investigations demonstrated that cells had distinct fates and played specific roles in later development (Whitman 1887, Whitman 1887, Conklin 1905, Kretzschmar and Watt 2012). The realization that cell division results in cell lineages, which in turn contribute to the formation and maintenance of an organism, sparked an era of innovations aimed at providing tools to map the contribution of distinct lineages in tissues and to identify the cellular origins of these. Methods for cell lineage tracing aim to investigate the ancestry of cells that compose a multi-cellular organism.

Cell lineage analysis can be facilitated by tracking a marker that distinguishes a cell lineage from other cells in a tissue. In the 1920s, embryologist Walter Vogt performed a pioneering experiment where a colored substrate was introduced into individual cells of a developing embryo of the frog species *Xenopus* (Buckingham and Meilhac 2011, Kretzschmar and Watt 2012). This was done by placing a tiny agar chip containing the dye on top of the developing embryo, and cells present directly below the chip absorbed the dye into their cytoplasm. When the labeled cells divided, the dye would be transmitted to the daughter cells, and Vogt could thus track the progeny of the cell which were originally labeled. Using this strategy, Vogt performed a series of experiments in which individual cell lineages were labeled and tracked, which allowed him to observe how the 32-cell blastula *Xenopus* embryo was formed (Hsu 2015). One drawback with this technique was that the dye was diluted with every cell division, resulting in that the signal was lost after a certain number of divisions. In the 1980s Sulston and colleagues managed to reconstruct a 'fate map' of the entire cell lineage of the nematode Caenorhabditis elegans, by manually tracking and annotating dividing cells observed through a microscope (Sulston, Schierenberg et al. 1983). While this experiment represents a milestone in the lineage tracing field, direct observation is laborious and limited to small transparent organisms (Garcia-Marques, Espinosa-Medina et al. 2021).

To trace the progeny of cells in species with a larger number of cells, transplantation was used to generate chimeric animals. This enabled Spemann and Mangold to track donor cells in differentially pigmented newt species (Garcia-Marques, Espinosa-Medina et al. 2021). Another example involves the generation of chimeras of different species. In the 1970s, Nicole Le Douarin generated chick-quail chimeras to study the migration patterns and differentiation potential of neural crest cells. When transplanting quail cells into chicken embryos, quail-derived cells will be integrated into the developing tissues of the host chicken. Upon examination of the tissues by light microscopy, quail-derived cells could be distinguished from chicken-derived cells by staining the tissue with Feulgen DNA dye, which labels DNA in quail cells but not in chicken. The information obtained from these analyses demonstrated a remarkable migratory capacity of neural crest cells (Le Douarin 1973, Le Douarin and Teillet 1973)

One of the most notable studies taking advantage of transplantation to trace cell lineage was published by Weissmann and colleagues in 1988 (Spangrude, Heimfeld et al. 1988). In this experiment, a specific population of bone marrow cells was isolated from a mouse. The authors next set out to test whether the isolated cell population represented hematopoietic stem cells. This was done by transplanting the isolated cells into recipient mice subjected to lethal irradiation and observing whether the graft could produce blood cell progeny. Since the transplanted cells were able to reconstitute all blood cell types and promote long-term survival of recipient mice, the authors demonstrated that the isolated cell population should comprise hematopoietic stem cells. This study demonstrates the definition of adult stem cells, in their ability to self-renew and to generate progeny representing multiple cell types. Their work also sparked a large number of studies that focus on identifying and characterizing the role of stem cells in adult tissues (Spangrude, Heimfeld et al. 1988, Hsu 2015).

Related to the work on characterizing the plasticity of bone marrow-derived cells, Diane Krause and colleagues were able to track the progeny of a single transplanted bone marrow-derived cell from a male mouse through a series of transplantations into female recipient mice (Krause, Theise et al. 2001). This was made possible since male cells can be distinguished from female cells, based on the presence of a Y chromosome. Following transplantation, the tissues of the recipient mice were examined by microscopy analysis. In line with previous studies (Spangrude, Heimfeld et al. 1988), the transplanted female mice harbored blood cells containing a Y chromosome, demonstrating a donor-derived origin of these cells. Besides observing male donor-derived blood cells, the authors made the surprising discovery that donor-derived male DNA was also present in many non-hematopoietic tissues. The authors concluded that bone-marrow-derived cells have far more plasticity than previously thought, with an ability to differentiate into epithelial cells of diverse tissues, including for example the liver, lung, gastrointestinal tract, and skin (Krause, Theise et al. 2001). This finding would be of tremendous clinical importance, since it implies that the transplanted bone-marrow-derived cells were multipotent, with a capacity to give rise to many different cell types. Such cells could potentially be exploited to treat not only hematological malignancies but could also be targeted for the development of clinical treatments for genetic disease or tissue repair in non-hematopoietic organs (Krause, Theise et al. 2001).

However, subsequent studies revealed an alternative explanation for the presence of donor-derived DNA in non-hematopoietic cells following transplantation. Through detailed genetic analysis, multiple research groups demonstrated that cells containing donor-derived DNA, as determined by the presence of a Y chromosome, also contained recipient DNA. It is today well established that bone marrow-derived cells contribute to non-hematopoietic tissues through a mechanism by which transplanted donor cells fuse with recipient cells (Anderson, Gage et al. 2001, Alvarez-Dolado, Pardal et al. 2003, Vassilopoulos, Wang et al. 2003, Wang, Willenbring et al. 2003, Weimann, Johansson et al. 2003, Johansson, Youssef et al. 2008, Berry and Rodeheffer 2013). This event results in cells harboring genetic material from both the donor and the recipient. The observation that transplanted bone marrow-derived may contribute to non-hematopoietic tissues by cell fusion rather than differentiation, highlights the limitation of transplantation to define stem cells. In addition, it is important to note that cell behavior may be different in the setting of transplantation and under homeostatic conditions (Watt and Jensen 2009, Sun, Ramos et al. 2014). However, transplantation remains useful for studying for example migratory behavior and differentiation potential when cells are moved into a different tissue environment, as well as the use of xenografts allowing for analysis of the behavior of human cells in experimental animals (Garcia-Marques, Espinosa-Medina et al. 2021). In some instances, transplantation may be the only option to study cell fate. In Paper I, we investigated if bone marrow-derived cells could form adipocytes in human study subjects that had previously undergone a bone marrow- or peripheral blood stem cell transplantation. This experimental setup was selected due to the lack of available methodologies to trace the cellular origins of adipocytes in humans. To distinguish between cell fusion events and differentiation, we investigated the presence of polymorphic markers in donor and recipient-derived DNA in individual adipocytes in transplant recipients.

Since the 1990s, the rapid development of genetic tools has transformed lineage tracing methods. This enabled the introduction of inheritable markers into cells, including for example fluorescent proteins and genetic barcodes. The most widely used method for lineage tracing in experimental animals is the Cre-lox system (Orban, Chui et al. 1992, Blanpain and Fuchs 2006, Buckingham and Meilhac 2011, Hope and Bhatia 2011). This method relies on genetically modified mice, in which the expression of the DNA-cutting enzyme Cre can be controlled by a promoter gene that is specifically expressed in a cell type or tissue of interest. These experimental animals are subsequently crossed with other experimental animals that carry a fluorescent reporter gene that is expressed after activation by Cre. Since the expression of the reporter gene is inherited by daughter cells, this enables the tracking of migration, proliferation, and differentiation of specific cell types *in vivo*. The activity of Cre can furthermore be dynamically regulated by using external stimuli, including for example the estrogen analog Tamoxifen, and the induction of labeling of cells will only occur when Tamoxifen is applied (Metzger and Chambon 2001, Hsu 2015).

To increase the resolution in lineage tracing, 'brainbow' or 'confetti' mice were developed (Livet, Weissman et al. 2007). In these mice, Cre activation induces expression of a random set of fluorescent proteins which endows individual cells with different colors. This strategy has, for example, been used to demonstrate that intestinal crypts are monoclonal and to study developmental bias in mouse embryonic stem cells (Snippert, van der Flier et al. 2010, Tabansky, Lenarcic et al. 2013). Due to a restricted number of colors available to label cells, and a limited number of lasers to detect the fluorescent signals, these methodologies are limited by the number of founder cells that can be uniquely labeled.

To further increase the resolution of lineage tracing, contemporary methods take advantage of recent advances in genome engineering technologies and single cell sequencing. This allows for high throughput analyses of clonal dynamics within cell populations. This is done by introducing genomic modifications in the DNA of individual cells, using genome-editing technologies, such as the CRISPR–Cas9 system (Hsu, Lander et al. 2014). The introduced mutations can subsequently be discovered in the individual cells by DNA sequencing, which is the process of determining the order of nucleotides in the genetic code. Since daughter cells inherit the genome of their ancestors, cell lineage relationships can be reconstructed based on shared mutations.

1.3 TOWARDS LINEAGE TRACING IN HUMANS

Single cell sequencing has also opened up the possibility of inferring cell lineage relationships in humans. This can be done through analysis of the inheritance patterns of naturally occurring somatic mutations. Similar to the strategy of introducing mutations in experimental animals as described above, naturally occurring somatic mutations label the progeny of the cell in which it occurred (Figure 1).

Somatic mutations are alterations in the nucleotide sequence of the DNA, and may arise from DNA damage repair, DNA replication and mitosis (Zhang and Vijg 2018). Somatic variation may arise early in development or occur later in life. As such, somatic mutations are distinguished from germline mutations, which are present in the DNA of every cell since these were inherited from the parents. Somatic variation may vary from very large chromosomal aberrations and copy number variation (CNV) to smaller insertions, deletions and substitutions of nucleotides in the genetic code (Zhang and Vijg 2018). Somatic variation is generally considered to be a hallmark of disease, including cancer (Negrini, Gorgoulis et al. 2010), and in aging (Lopez-Otin, Blasco et al. 2013, Zhang and Vijg 2018). However, mutations frequently arise in genomic regions which do not contain genes and may not have a functional effect on the cell (Woodworth, Girskis et al. 2017, Zhang and Vijg 2018). Importantly, since somatic mutations are passed on to daughter cells, the patterns of somatic mutations will mirror cells' hierarchical relationships. Owing to advances in single cell genome sequencing, it is today possible to identify these natural lineage markers, and the genealogy of cells can be inferred by retrospectively deciphering the order of mutations that accumulate in cells as they divide (Lodato, Woodworth et al. 2015).

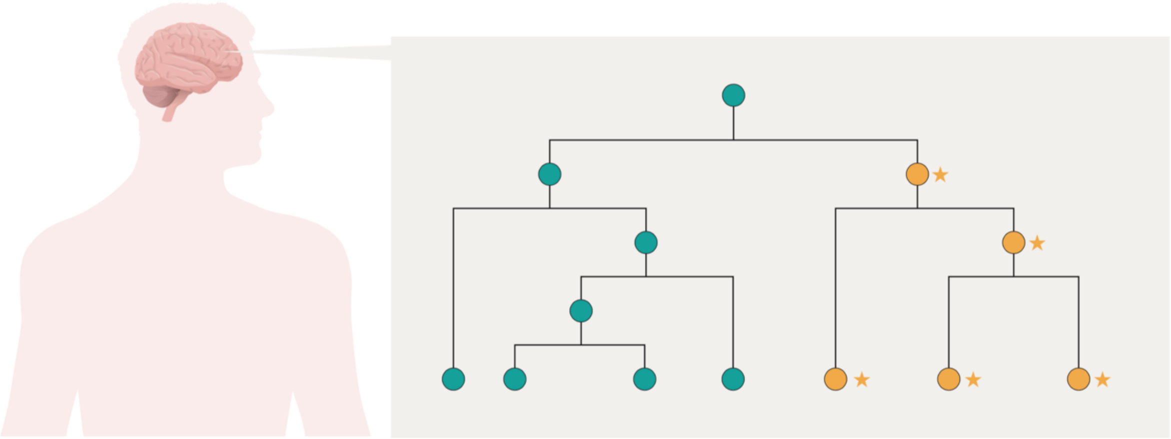

Figure 1. Somatic mutations, depicted as a star, arise in the genome of individual cells and are passed on to daughter cells during cell division. The lineage relationships of cells can be inferred by tracking the inheritance patterns of such mutations.

1.3.1 Nuclear genomic mutations

The first attempts to reconstruct cell lineage relationships based on somatic mutations were performed by two independent research groups that focused on areas of the genome with high mutation burden (Frumkin, Wasserstrom et al. 2005, Salipante and Horwitz 2006). These studies focused on microsatellites, which are repetitive stretches of DNA consisting of one or more base pairs that are repeated multiple times (Richard, Kerrest et al. 2008, Gulcher 2012). Due to the repetitive structure of microsatellites, these are prone to be altered when DNA is replicated prior to cell division. The errors result in an increase or decrease in length, which makes microsatellites variable between individuals and within individuals (Ellegren 2004). Owing to the instability of microsatellite repeats, these have been predicted to enable the reconstruction of the entire cell lineage tree of an organism by using computational methods for phylogenetic inference (Frumkin, Wasserstrom et al. 2005, Salipante and Horwitz 2006).

CNVs are genetic traits involving the number of copies of a particular genomic segment. These may range from 1 kilobase to several megabases in size (Zhang and Vijg 2018). CNVs are estimated to account for 12% of the human genome (Redon, Ishikawa et al. 2006). Interestingly, several groups have revealed extensive CNV between different cells in the same individual, which must have arisen as somatic events post-fertilization (Piotrowski, Bruder et al. 2008, Abyzov, Mariani et al. 2012, McConnell, Lindberg et al. 2013). CNV analysis in skin and brain has shown that a large proportion, 30-70% of skin cells and neurons harbor at least one large somatic CNV (Woodworth, Girskis et al. 2017). However, only a small number of CNVs have been found to be shared by multiple cells (Woodworth, Girskis et al. 2017).

Endogenous retroelements are a type of genetic component with the capability to move, or transpose. This occurs through a mechanism by which a DNA sequence is copied or cut out and then pasted into another location in the genome (Goodier 2016). Since somatic mobilization of retroelements, in particular long interspersed nuclear element 1 (L1), has been observed in humans (Woodworth, Girskis et al. 2017, Faulkner and Billon 2018), these could potentially be an interesting target for cell lineage reconstruction purposes. However, while retrotransposons are estimated to constitute approximately 40% of the human genome (Zhang and Vijg 2018), only a small number of these elements are actively mobilizing (Woodworth, Girskis et al. 2017) . Using available technologies it appears challenging to distinguish true L1 elements from false positives, leaving room for controversy with regards to the precise rate of L1 mobility (Evrony, Cai et al. 2012, Upton, Gerhardt et al. 2015).

Single nucleotide variants (SNVs) are substitutions of individual nucleotides in the genetic code. SNVs are considered to be major drivers of evolution and a major source for disease-causing mutations (Woodworth, Girskis et al. 2017). Since the vast majority of the genome is non-coding, SNVs are frequently occurring in regions that may not have a functional impact on the cell (Woodworth, Girskis et al. 2017). Estimations of the substitution rate range from 2 to 10 mutations per division (Lynch 2010, Martincorena and Campbell 2015, Woodworth, Girskis et al. 2017, Dou, Gold et al. 2018). Although the division rate of stem cells in adult tissues has been difficult to estimate and may vary widely between tissues, (Tomasetti and Vogelstein 2015), normal cells are expected to acquire hundreds to thousands of SNVs (Martincorena and Campbell 2015). Given their abundance and often neutral functionality, somatic SNVs represent a substantial source of genetic variation which can be utilized for cell lineage reconstruction (Woodworth, Girskis et al. 2017). Indeed, multiple reports have utilized somatic SNVs to reconstruct cell lineages in multiple organs in both the mouse and in humans, including stomach, intestine, prostate, brain, immune system, and liver (Behjati, Huch et al. 2014, Lodato, Woodworth et al. 2015, Hazen, Faust et al. 2016, Woodworth, Girskis et al. 2017, Lee-Six, Obro et al. 2018, Lodato, Rodin et al. 2018, Brunner, Roberts et al. 2019, Lee-Six, Olafsson et al. 2019).

While these initial attempts for lineage tracing in humans have relied on the detection of somatic variation in nuclear genomic DNA, these analyses typically require sequencing of the entire genome or the exome, which is the part of the genome containing genes. However, whole genome sequencing and exome sequencing remains costly and is difficult to apply at scale. In addition, single cell DNA sequencing data, is limited in that it does not provide information about the functions and identity of individual cells which, for example, can be obtained from gene expression analysis.

1.3.2 Mitochondrial mutations

Multiple reports have demonstrated the utility of mitochondrial mutations as cell lineage marks in the last decade, including recent reports demonstrating the feasibility of identifying mitochondrial variation at the single cell level (Taylor, Barron et al. 2003, Teixeira, Nadarajan et al. 2013, Walther and Alison 2016, Salas, Wiencke et al. 2018, Ludwig, Lareau et al. 2019, Xu, Nuno et al. 2019). A major advantage with targeting mitochondrial variation for cell lineage analysis approaches is that these can be detected by existing methods for single cell sequencing which also records information about the functions and identities of individual cells defined by their gene expression and epigenomic profiles. Such methodologies for data generation include single cell assay for transposase-accessible chromatin high-throughput sequencing (scATACseq) and single cell RNA sequencing (scRNAseq) (Ludwig, Lareau et al. 2019).

Mitochondria contain their own small genome, representing a 16.6kb-long circular DNA molecule that is transcribed in almost its entirety. The genes transcribed from the mitochondrial genome code for proteins that form the mitochondrial respiratory chain, which produces much of the chemical energy needed for the biochemical reactions occurring in a cell. Given the small genome size, it remains cost-effective to perform sequencing of mitochondrial DNA (mtDNA).

Each cell has multiple copies of mtDNA, with estimates ranging from 100-1000s per cell (Ludwig, Lareau et al. 2019). Mutations arise in individual mtDNA molecules and accumulate throughout human life (Wallace 1992). The mutation rate in mitochondrial genome is estimated to be substantially higher than for nuclear genomic DNA (Payne and Chinnery 2015, Ludwig, Lareau et al. 2019). Over time, the proportion of mtDNA molecules that harbor a specific mutation may increase in individual cells. Mitochondrial mutations can reach high levels of heteroplasmy, which is the fraction of mutated mitochondrial genomes within a cell (Ludwig, Lareau et al. 2019). Such increase in heteroplasmy of a particular mutation has been proposed to be the result of the combination of random genetic drift, constant turnover of mtDNA that is independent of cell division, and vegetative segregation (Figure 2) (Elson, Samuels et al. 2001, Wallace and Chalkia

2013, Stewart and Chinnery 2015). Although, many mutations in mtDNA are expected to have a neutral effect since cells can tolerate >80% heteroplasmy of pathogenic variants before defects in the respiratory chains are detected (Stewart and Chinnery 2015).

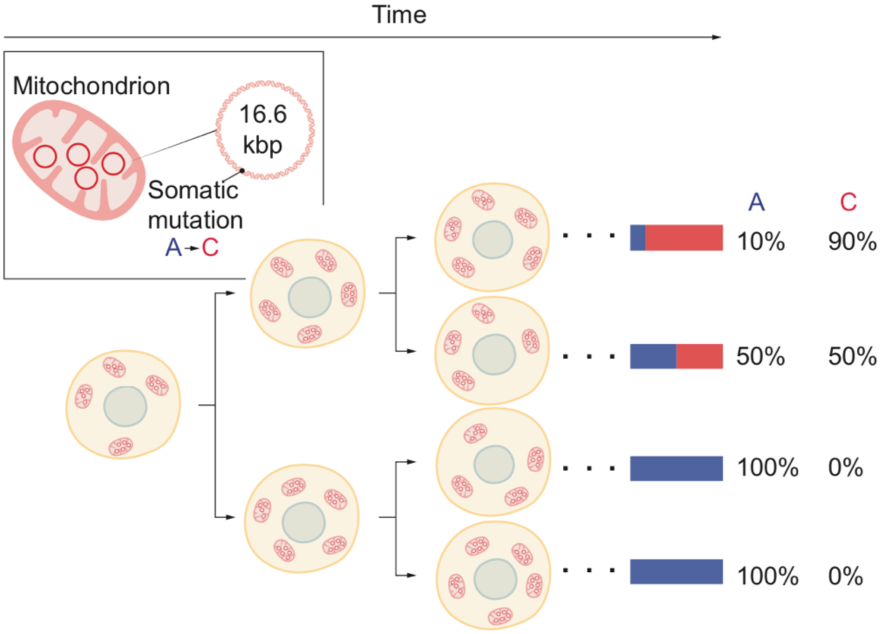

Figure 2. Each cell has multiple mitochondria. Each mitochondrion has multiple mtDNA molecules. Somatic mutations can arise in individual mtDNA and can reach high levels of heteroplasmy over time. Mutated mtDNA is inherited by daughter cells during cell division.

1.4 SINGLE CELL SEQUENCING

Identifying somatic mutations can be achieved by determining the order of nucleotides in the genetic code using DNA sequencing technologies, followed by computational analysis aiming to identify genomic locations where more than one nucleotide is observed. Since conventional methods for DNA sequencing are performed on a mixed population of cells, signals from rare somatic mutations will be concealed within the aggregated bulk data (Lahnemann, Koster et al. 2020). It is therefore very challenging to identify somatic mutations using such bulk sequencing approaches. To circumvent this issue, new technologies for single-cell sequencing attempts to obtain information about the genetic code in DNA or RNA from individual cells. The advantage of this approach is that somatic mutations present in only a subset of cells can be detected.

1.4.1 Single cell DNA sequencing

The amount of DNA present in a single cell is far from sufficient for performing a sequencing experiment. To obtain sufficient genetic material, single cells can be expanded in culture until there are enough cells to perform standard bulk sequencing (Welch, Ley et al. 2012, Behjati, Huch et al. 2014, Blokzijl, de Ligt et al. 2016, Hazen, Faust et al. 2016). However, during the culturing process, new mutations arise which may be challenging to distinguish from biologically relevant mutations present only in the founder cell (Dou, Gold et al. 2018). In addition, it is not always possible to obtain clonal colonies from single cells, which may result in biased cell loss (Dou, Gold et al. 2018). Moreover, culturing is not applicable for many cell types, including terminally differentiated post-mitotic cells such as neurons. A recently developed method demonstrates clonal expansion of neurons by single-cell nuclear transfer (SCNT) into dividing cells (Mizutani, Oikawa et al. 2015). However, SCNT requires substantial manual work and has been proposed to be associated with additional selection biases (Dou, Gold et al. 2018).

Another common approach to produce sufficient amounts of DNA for sequencing is to apply Whole Genome Amplification (WGA) which can generate micrograms of amplified DNA from the minute amounts of DNA present in a single cell. However, the WGA process introduces errors when the DNA is amplified. Such errors may be falsely interpreted as somatic mutations (Dou, Gold et al. 2018). Healthy, non-malignant cells are expected to be diploid, meaning that only two DNA molecules, - alleles -, exist at each locus. Every molecule that fails to be amplified inevitably leads to missing data. This effect, termed amplification bias, is represented by successful amplification in some genomic regions and failed amplification in other regions (Figure 3) (Dou, Gold et al. 2018, Lahnemann, Koster et al. 2020). Furthermore, amplification bias may lead to the scenario where only one of the two alleles was successfully amplified. This feature of single cell data, called allelic dropout, is particularly problematic since the failure to detect mutation signals may result in false negative variant calls. Multiple methodologies for WGA exist, each with a different error profile. It is important to recognize these differences, since certain WGA technologies may be optimal depending on the research question and genetic variation of interest (Lahnemann, Koster et al. 2020).

The polymerase chain reaction (PCR) is a widely used tool for amplifying specific regions in the genome (Saiki, Scharf et al. 1985, Mullis and Faloona 1987). PCR works by first heating the double-stranded DNA in order to obtain two single strands. Next, an enzyme called DNA polymerase builds two new strands using the original strands as templates. The initiation of DNA synthesis is determined by primers, which are short single-stranded nucleic acid sequences that bind to different targets in the DNA. Based on PCR, the first approach for amplifying the human genome was developed in the 1980s (Nelson, Ledbetter et al. 1989, Czyz, Kirsch et al. 2015). Several types of methods for WGA based on PCR have been developed since then (Telenius, Carter et al. 1992, Zhang, Cui et al. 1992, Sermon, Lissens et al. 1996, Dietmaier, Hartmann et al. 1999, Langmore 2002). PCR-based strategies for WGA rely on thermostable polymerases since the reaction is carried out under increasing and decreasing temperatures, in a cycling manner. Therefore, PCR-based strategies for WGA rely on thermostable polymerases, in order to remain stable during the maximum temperature reached during PCR cycling. However, all thermostable polymerases have relatively high error rates, which makes PCR-based approaches suboptimal for SNV detection (Lahnemann, Koster et al. 2020). An advantage of PCR-based approaches for WGS is that these achieve the most uniform coverage, and are therefore suitable for CNV analysis (Woodworth, Girskis et al. 2017).

Multiple displacement amplification (MDA) is a non-PCR-based method for WGA (Dean, Hosono et al. 2002). MDA can amplify the few femtograms of DNA found within a single bacterium to microgram amounts of material that can be readily used for sequencing. In MDA, DNA is amplified at a constant temperature using the enzyme phi29 DNA polymerase which has a proofreading activity and therefore delivers 1000-fold higher fidelity as compared to thermostable DNA polymerase (Tindall and Kunkel 1988, Esteban, Salas et al. 1993). Given the high fidelity of phi29 DNA polymerase, MDA is the method of choice for SNV calling (Woodworth, Girskis et al. 2017, Lahnemann, Koster et al. 2020), and was therefore used for WGA in Paper II in this thesis. One disadvantage with MDA is that this method suffers from stronger amplification bias as compared to PCR-based technologies (Picher, Budeus et al. 2016, Baumer, Fisch et al. 2018, Lahnemann, Koster et al. 2020).

MALBAC utilizes quasi-linear pre-amplification to reduce the amplification bias. The initial amplification depends on a set of primers consisting of a shared sequencing and a variable sequence that can bind to DNA even at low temperatures. The DNA is used as a template for a pre-amplification step that generates hairpin-shaped molecules, which are subsequently amplified into large quantities of DNA. A conceptually similar technique is the commercial application PicoPlex from Rubicon Genomics which was used for single cell WGA in Paper I in this thesis.

1.4.2 Single cell RNA sequencing

The behavior and function of a cell are determined by the genes it expresses. As such, sequencing of the RNA in individual single cells can provide information about the current state of a cell. Moreover, this can enable the identification of subpopulations of cells with distinct biological functions, both in the healthy state and in disease (Sandberg 2014, Kolodziejczyk, Kim et al. 2015). In recent years, the development of technologies for scRNAseq has revolutionized biological research, and heterogeneity in cell populations can now be dissected in ways that was previously not possible.

Smartseq (Ramskold, Luo et al. 2012) was developed at Karolinska Institutet for whole transcriptome amplification (WTA) and enables the detection of full-length transcripts. Further developments of this technology include smartseq2 (Picelli, Bjorklund et al. 2013) and smartseq3 (Hagemann-Jensen, Ziegenhain et al. 2020), which achieves improved sensitivity and full-length capture. Other methods for WTA include for example Quartz-Seq (Sasagawa, Nikaido et al. 2013) and CEL-seq (Hashimshony, Wagner et al. 2012). The main challenge in WTA is associated with handling the small amounts of RNA present in single cells. To simplify the preparation of transcriptome libraries, several methods based on microfluidics have been developed, including for example Drop-seq (Macosko, Basu et al. 2015) and inDrop (Klein, Mazutis et al. 2015), as well as platforms developed by vendors such as Chromium (10x Genomics), ddSEQ (Bio-Rad/Illumina), and Nadia (Dolomite) (Grun and van Oudenaarden 2015). Other challenges associated with scRNAseq library preparation include sample loss, amplification efficiency and amplification bias. (Kashima, Sakamoto et al. 2020). It is important to consider the differences between scRNAseq technologies in order to select suitable methods depending on the research questions addressed. A key distinction between different methods for scRNAseq is that some methods collect information about the full-length sequence of a transcript whereas others only amplify a set of nucleotides (Haque, Engel et al. 2017). While the latter strategy is sufficient for linking a transcript to a specific gene, full-length sequencing is preferable for mutation discovery. In Paper III, we set out to identify somatic mutations in transcripts originating from the mitochondrial genome, and we therefore used Smartseq2 which enables full-length sequencing (Picelli, Bjorklund et al. 2013).

1.5 COMPUTATIONAL METHODS

Both DNA and RNA sequencing is performed by determining, or reading, the order of nucleotides in the genetic code. In the case of RNA sequencing, RNA molecules are first converted into DNA molecules, termed cDNA, prior to sequencing (Mortazavi, Williams et al. 2008). The preparation steps prior to sequencing begin with the extraction of nucleic acids from cells. This is followed by fragmentation and the construction of a molecular library consisting of DNA or cDNA fragments. During the sequencing process, molecules from such libraries will be randomly sampled for sequencing. Each time a DNA molecule is read by the sequencing machine, a digital copy, termed read, is generated (Osborne, Barnes et al. 2001). The data in the read consists of the nucleotide sequence in the DNA fragment that was sequenced.

1.5.1 Alignment and variant calling

The first step of the computational analysis of DNA sequencing data involves sorting the reads according to the expected order of nucleotides in the human genome. This process, termed alignment or mapping, can be compared to putting together a giant puzzle where a vast number of reads are aligned back to the respective regions they likely originate from. This is done with the help of a reference genome file consisting of approximately 3 billion nucleotides, and is intended to be representative of the order of nucleotides in the human genome.

The reference genome file was generated as a result of The Human Genome Project (Lander, Linton et al. 2001). One driving goal of this effort was to develop a tool that could aid in future analyses aimed at providing increased knowledge about the human genome and to accelerate genomics research (Ballouz, Dobin et al. 2019). This has undeniably been successful since the reference genome is an essential tool for many types of computational analyses that involves human DNA sequencing data (Ballouz, Dobin et al. 2019). One of the major discoveries related to this work concerns the extensive presence of repetitive sequences throughout the human genome. The repetitiveness makes the sequencing machines introduce errors when reads are generated, and assembling of sequencing reads into the correct order becomes a computationally challenging problem (Treangen and Salzberg 2011). Moreover, many genomic regions differ significantly across the human population, resulting in ambiguities as to the precise location where reads belong in the genome which, in turn, result in alignment artifacts (Li 2014). This is further complicated by the fact that the vast majority of the human reference genome is based on DNA obtained from a single individual and may contain sequences as well as lack sequences that are present in other individuals (Ballouz, Dobin et al. 2019).

Genetic variation is detected by variant calling algorithms that take aligned read data as input and outputs positions where there are polymorphisms, defined as genomic location where at least one nucleotide differs from a reference sequence. However, because sequencing data suffer from errors, which include both mistakes introduced during sequencing as well as alignment artifacts, variant calling is difficult and there is often considerable uncertainty associated with the results (Nielsen, Paul et al. 2011). It is crucial to account for this uncertainty in downstream analyses, such as the identification of rare mutations (Nielsen, Paul et al. 2011).

1.5.2 Mutation discovery in nuclear DNA

Besides false positive variant calls arising from the sequencing process and alignment, there are a number of issues that are specific for single cell data. Ideally, single cell DNA sequencing data contains information about all types of genetic variation that can be exploited for retrospective cell lineage reconstruction, including for example genomic rearrangements, CNVs, insertions, deletions, and SNVs. However, the extensive amplification of single cell genomes that is required for sequencing, represents a major disturbing factor for identifying somatic mutations at the single cell level. WGA introduces amplification errors, amplification bias, allelic dropout and missing data (Woodworth, Girskis et al. 2017, Zhang and Vijg 2018). This presents serious challenges to computational analysis for the identification of somatic mutations (de Bourcy, Hou Y, Huang L, Estevez-Gomez).

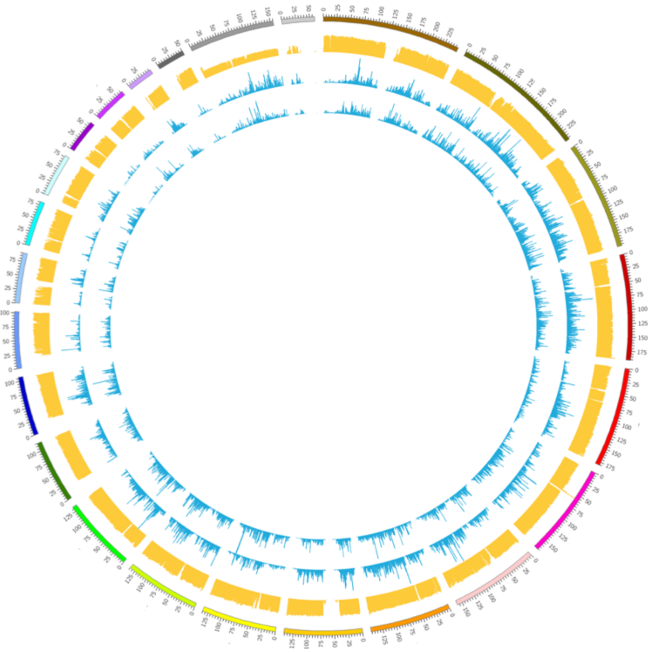

Figure 3. Circos plot showing sequencing read coverage in two MDA amplified single cell samples (blue tracks) and one unamplified bulk sample (yellow track). Each track represents a histogram showing the number of reads per genomic location, as indicated by the outer track representing individual chromosomes. The bulk sample displays even coverage across the genome. The single cells display uneven coverage across the genome.

In recent years, a number of new algorithms for tackling the unique features of single cell DNA sequencing data have been developed.
The first SNV variant caller specific for single cell data is Monovar, which handles errors and biases by leveraging sequencing information across the cell population (Zafar, Wang et al. 2016). In this process, Monovar predicts the presence or absence of variants in individual cells based on the posterior probability, which is a revised or updated mutation probability after considering the mutation profiles in other cells.

Multiple tools take advantage of heterozygous germline SNVs to identify and account for amplification bias. Germline variants are present in egg or sperm cells and are thus present in all cells of an organism. Heterozygous loci are positions where the two alleles harbor different nucleotides. Since the nucleotide differs in these loci, heterozygous germline SNVs can be used to distinguish reads originating from either of the two alleles of a diploid cell. SCcaller and SCAN-SNV account for local amplification bias by integrating data from neighboring heterozygous variants (Dong, Zhang et al. 2017). This assumes that WGA, in particular MDA, starts at random positions and amplifies long consecutive stretches of DNA (Dong, Zhang et al. 2017). The degree of local amplification bias in a particular genomic region can thus be predicted by considering the degree of bias in neighboring heterozygous SNVs. In Paper II, we propose an approach based on directly linking putative somatic SNVs to heterozygous germline SNVs. In genomics, this computational strategy is called phasing, and is typically used to identify patterns of neighboring genetic variation that is associated with health and disease (Tewhey, Bansal et al. 2011). We and others have shown read-backed phasing with heterozygous germline SNVs enable increased detection accuracy of somatic variant discovery and is particularly useful for data exhibiting high rates of allelic dropout (Dou, Gold et al. 2018, Bohrson, Barton et al. 2019, Hard, Al Hakim et al. 2019).

Computational tools exist for the analysis of large CNVs in single cells, including Ginkgo and Aneufinder (Garvin, Aboukhalil et al. 2015, Bakker, Taudt et al. 2016). Ginkgo analysis is performed directly via a web server, and computes CNV profiles by binning reads into segments of the genome, followed by analysis of read density in these genomic intervals (Garvin, Aboukhalil et al. 2015). Aneufinder also predicts CNVs based on significant difference in read count data, although the statistical methods to predict CNV profiles differ between Ginkgo and Aneufinder (Bakker, Taudt et al. 2016). In comparative analyses, Aneufinder and Ginkgo have been shown to generate concordant results in terms of CNV calls, although Aneufinder achieves better sensitivity for small CNV events while Ginkgo is more robust for noise in the data (Bakker, Taudt et al. 2016).

1.5.3 Phylogenomics

The inheritance patterns of somatic variation in single cells can be used to infer the somatic diversification of different lineages of single cells. However, substantial computational challenges are associated with single cell data, which requires the development of new analytical frameworks to model the evolution that occurs within one human being.

Such models of evolution may be described by a simple representation of the presence or absence of mutation events, which can be obtained by computational methods for single cell variant calling as described above (Lahnemann, Koster et al. 2020). To dissect clonal relationships, a number of clustering approaches have been developed which aim to identify groups of cells that share somatic mutations (Gawad, Koh et al. 2014, Roth, McPherson et al. 2016). Advantages with clustering methods is that these are easy to implement and fast to execute. One disadvantage with clustering-based approaches is that these rely on a simplified measure of distance considering only pairs of cells. As a consequence, the resulting tree represents a measure of the relatedness between cells but cannot be used for quantitative analysis. Moreover, while clustering methods are fast and straightforward, these may not perform well or fail completely for data that is burdened by errors and biases, which is often the case for single cell data (Kuipers, Jahn et al. 2017).

A different direction taken for predicting clonal relationships is based on phylogenetic inference which considers the underlying mechanisms of the evolutionary process as well as technical aspects of the data (Kuipers, Jahn et al. 2017, Lahnemann, Koster et al. 2020). These include probabilistic approaches that select trees based on how well they explain the observed data (Kuipers, Jahn et al. 2017). Traditionally, phylogenetics is used to reconstruct the evolutionary history of species and many methods for phylogenetic inference exist (Kapli, Yang et al. 2020). These methods evaluate observed traits, for example inherited mutations, and outputs a phylogeny, also known as a phylogenetic tree (Lahnemann, Koster et al. 2020). The tips of the tree diagram are called leaves or taxa, and represent the end, or present, of an evolutionary lineage (Yang and Rannala 2012). Such lineage is represented by the branches in the tree, whereas internal nodes represent lineage splitting events (de Queiroz 2013). When applying phylogenetic inference to visualize the clonal evolution of single cells, the leaves represent single cell samples, and the internal nodes represent the hypothetical common ancestors.

Tools for single cell phylogenetic inference include OncoNEM (Ross and Markowetz 2016) and SCITE (Jahn, Kuipers et al. 2016), which take binary input of presence or absence of mutations to compute the tree. These methods can handle multiple types of errors, including false negatives, false positives and missing data, assuming that it is more likely to observe false negatives than false positives (Lahnemann, Koster et al. 2020). Both OncoNEM and SCITE assume that data follows the infinite sites model, meaning that a mutation is expected to occur only once. (Kimura 1969). Approaches that allow for violations of the infinite sites assumptions include SiFit (Zafar, Tzen et al. 2019) and SPhyR (El-Kebir 2018). Leveraging the dependency structure among cells in a dataset can also be powerful for distinguishing true mutations from false positives (Lahnemann, Koster et al. 2020). This approach has been taken by a number of tools for variant calling and simultaneous reconstruction of cell lineage trees, including SciCloneFit and sciΦ (Singer, Kuipers et al. 2018, Zafar, Navin et al. 2019).

The key challenges for single cell phylogenetics methods include the design of models that are biologically realistic while being able to scale for datasets with an increasing number of cells and an increasing number of somatic mutation sites (Lahnemann, Koster et al. 2020). The space of possible phylogenetic trees grows extremely fast with the number of cells and mutations. Since probabilistic approaches need to search for a solution globally in this space, the computational problem quickly becomes intractable. Indeed, this is a common theme for probabilistic methods and a manifold of methods for approximate inference exists, each with different trade-offs (Bishop 2011). Taken together, while adding data from more cells will improve the resolution of phylogenetics analysis, probabilistic models for phylogenetic inference would still face the challenge of computational tractability, even when the data is perfect given the (super-)exponentially growing search space of possible trees (Lahnemann, Koster et al. 2020).

1.5.4 Cell lineage reconstruction based on mitochondrial mutations

Mitochondrial mutations are increasingly recognized as lineage markers for reconstructing clonal structures and have the advantage that these can be detected in scRNAseq and scATACseq data (Lareau, Ludwig et al. 2021). Such analyses can potentially decipher cell lineage relationships and simultaneously reveal cell identities. However, it is challenging to distinguish true mtDNA variation from amplification and sequencing errors introduced during library preparation steps (Kwok, Qiao et al. 2021). As a consequence, uninformative and noisy mtDNA variant calls may confound clonal inference and biological interpretation. While methods designed for somatic mutation discovery in nuclear genomic DNA can rely on assumptions regarding a diploid context, this is violated in the mitochondrial genome since multiple copies of mtDNA exist per cell and mutations may exist in any fraction between one of these mtDNA copies or all (Kwok, Qiao et al. 2021). The failure to sequence mtDNA copies that harbor variants may consequently lead to false negatives. The number of mtDNA copies per cell varies greatly between cell types and tissues, but the mechanisms by which mtDNA copy number is monitored and controlled are not well understood (O'Hara, Tedone et al. 2019). These uncertainties may ambiguate modeling the process of mtDNA mutations and the increasing level of heteroplasmy.

Very few computational methods are available for detecting genetic variation in mtDNA across single cell sequencing assays, in particular, for analyzing mitochondrial mutations in scRNAseq data (Kwok, Qiao et al. 2021, Miller, Lareau et al. 2021). Methods published in peer-reviewed journals include EMBLEM and mgatk, which rely on allele frequencies to detect somatic variants in mtDNA (Xu, Nuno et al. 2019, Lareau, Ludwig et al. 2021). This is done by considering the fraction of reads supporting more than one nucleotide. These methods were primarily designed for scATAC-seq, but may not be as robust for scRNAseq (Lareau, Ludwig et al. 2021).

With regards to inference of clonal relationships based on detected mitochondrial mutations, available methods report a binary representation of presence or absence of mutations, or through agglomerative hierarchical clustering to group cells that share mtDNA variants. Binary representation of presence and absence of mutations is limited in that perceived absence of mutation may be the result of a false negative from mutated alleles, for example due to low read coverage over variant loci. Another disadvantage of using agglomerative hierarchical clustering is the difficulty to determine the number of meaningful clusters of related cells. Moreover, this approach reports one binary tree, and does not account for alternative trees which may also explain the data. While an ideal tree construction algorithm would include also internal nodes, such algorithms are currently not available (Ludwig, Lareau et al. 2019). In Paper III, we address these issues by developing a probabilistic approach for tree construction, which uses a greedy search to iteratively add internal mutation nodes to create a tree with the highest likelihood. This allows us to evaluate the tree using the full set of cells, rather than being restricted to a pair-wise comparison.

2 PRESENT INVESTIGATION

2.1 GENERAL AIMS

This thesis has used single cell DNA and RNA sequencing to detect and utilize genetic variation for retrospective lineage reconstruction of human cell populations.

2.2 SPECIFIC AIMS

Paper I:
To investigate if bone marrow-derived stem cells can form adipocytes in bone marrow or peripheral blood stem cell transplant recipients.

Paper II:
To develop a computational method for the identification of somatic mutations which can be used to infer cell lineage relationships of human cell populations.

Paper III:
To develop a computational method for cell lineage reconstruction and simultaneous gene expression profiling, and to apply this method to investigate human memory and effector CD8+ T cell development after vaccination

2.3 PAPER I

In **Paper I**, we set out to investigate the origins of human adipocytes in bone marrow or peripheral blood stem cell transplant recipients, through analysis of germline variants in genomic DNA (Ryden, Uzunel et al. 2015).

White adipocytes are constantly replaced throughout life (Spalding, Arner et al. 2008). While this requires a source of precursor cells that can produce new adipocytes, the identity of such cells is not known. In transgenic mouse models, multiple studies suggest that new adipocytes are formed through the differentiation of cells that reside in close proximity to blood vessels that are dispersed throughout white adipose tissue (Crisan, Yap et al. 2008, Lin, Garcia et al. 2008, Zannettino, Paton et al. 2008). However, the origins of these precursor cells is not known, for example whether these are permanent residents or whether they originate from other organs and enter white adipose tissue via the circulation.

To characterize the origins of new adipocytes, a number of studies conducted in experimental animals have used of transplantation of bone marrow-derived cells expressing fluorescent proteins, to study if these cells can generate new adipocytes (Crossno, Majka et al. 2006, Sera, LaRue et al. 2009, Majka, Miller et al. 2012). Following transplantation, these studies report the presence of adipocytes expressing fluorescent proteins suggesting that transplanted bone marrow-derived cells have entered white adipose tissue and differentiated into adipocytes. However, the notion that bone marrow-derived cells, in particular hematopoietic stem cells, have the ability to contribute to non-hematopoietic lineages through differentiation is highly controversial (Anderson, Gage et al. 2001, Alvarez-Dolado, Pardal et al. 2003, Vassilopoulos, Wang et al. 2003, Wang, Willenbring et al. 2003, Weimann, Johansson et al. 2003, Johansson, Youssef et al. 2008, Berry and Rodeheffer 2013). Instead, it has been proposed that transplanted bone marrow-derived cells may fuse with recipient cells, resulting in cells with genetic material from both the donor and the recipient.

Since previous studies have focused on a possible contribution of bone marrow to mouse adipogenesis, we set out to investigate the origins of new adipocytes in humans. Given the lack of methodologies to distinguish cell lineages in humans, we studied adipocytes in patients that previously in life had received a bone marrow or peripheral blood stem cell transplantation. In these patients, cells derived from the donor can be distinguished from recipient cells by the presence of germline mutations in the DNA.

We began by investigating the fraction of donor and recipient-derived DNA in bulk populations of adipocytes. This showed that over the lifespan of an individual, ~10% of subcutaneous adipocytes originate from bone marrow-derived precursor cells.

We next set out to determine if the observed bone marrow contribution was resulting from cell fusion or differentiation. To accomplish this, we perform DNA sequencing of individual adipocytes in order to analyze the genetic composition at the single cell level. A major obstacle when working with adipocytes is their fragility and buoyancy in suspension. This makes adipocytes difficult to process as single cells. We circumvented these issues by developing a new experimental method in which single adipocytes were embedded in a low-temperature melting agarose followed by isolation by laser capture microdissection while collecting photomicrographs of all single cells included in the study. Following isolation, we performed WGA of each of the single cells. We next performed whole genome sequencing and/or exome sequencing of the single WGA libraries as well as DNA from bulk samples representing peripheral blood mono nuclear cells obtained from the donor and the recipient.

During computational analysis, we first identified homozygous germline variation from the donor and recipient bulk DNA. Homozygous germline variation are genomic locations where both alleles contain the same nucleotide, which may not represent the nucleotide that is commonly observed in the human population. In this case, we selected such positions that differed between the donor and the recipient. We next analyzed these positions in the single cell data, to determine if we observe DNA from the recipient, the donor or both. The strategy to exclusively analyze homozygous variants in single cells circumvented problems with potential false negatives resulting from allelic dropout, since signals from either allele would be informative.

Our findings demonstrate that ~90% of single adipocytes included in our study contain only recipient-derived DNA. This is consistent with our observations when analyzing bulk populations of adipocytes from these patients. We also find a small number of cells with genetic material from both the donor and the recipient, indicating that cell fusion may occur between bone marrow-derived cells and adipocytes or adipocyte precursors in humans. Interestingly, we also observe a small number of adipocytes only containing DNA that we can trace back to the donor. This finding suggests that, indeed, circulating bone marrow-derived progenitor cells can differentiate into adipocytes in humans.

It is important to note that using transplantation to define stem cells has certain limitations in that the setting of transplantation may be vastly different from homeostatic conditions, which may result in altered cell behavior. A definitive conclusion as to whether bone marrow-derived progenitors can generate adipocytes also in the physiological healthy state, requires the development of new technologies that allows for cell lineage of more closely related cells.

2.4 PAPER II

In **Paper II**, we set out to develop a computational method that would allow us to establish clonal relationships in closely related cells (Hard, Al Hakim et al. 2019). Recent studies demonstrate that cell lineage relationships can be inferred from analysis of the inheritance patterns of somatic mutations and that these are frequently occurring in nuclear genomic DNA of human cells (Behjati, Huch et al. 2014, Lodato, Woodworth et al. 2015, Hazen, Faust et al. 2016, Woodworth, Girskis et al. 2017, Lee-Six, Obro et al. 2018, Lodato, Rodin et al. 2018, Brunner, Roberts et al. 2019). We hypothesized that by performing whole genome sequencing of single cells, we can infer their relationships by identifying somatic mutations in the data and use this to resolve clonal relationships.

Variant calling in whole genome amplified single cells is challenging due to artifacts of various origins. These include false positive variants resulting from incorrect alignment of sequencing reads as well as errors introduced during amplification. False negative variant results from failed amplification and allelic dropout. To circumvent these issues, we developed Conbase, a method for the identification of clonal somatic mutations in whole genome sequencing data from single cells by taking advantage of read phasing. Genotyping is performed by linking each putative variant to adjacent, donor-specific heterozygous single nucleotide polymorphisms. This strategy mitigates effects of amplification errors and other aspects of bioinformatic analysis, including alignment artifacts that arise as a result of the limitation of using an incomplete reference genome with respect to the genome of the donor. Furthermore, this strategy allows us to determine the wild type genotype in unmutated cells, despite high rates of allelic dropout.

We validated the performance of Conbase by examining whole genome sequencing data from single cells with known lineage relationships. We began by analyzing somatic mutations in fibroblasts cultured while performing time-lapse movie recording followed by isolation of clonally related cells by laser capture microdissection. Unsupervised discovery of somatic mutations in single fibroblasts combined with unsupervised hierarchical clustering allowed us to unambiguously recapitulate clonal relationships. We next sought to test our method on somatic cells isolated from healthy human donors. To do this, we turned to T cells which represent an ideal system for studying cell lineage relationships, since clonally related T cells share a unique genomic sequence, the T cell receptor (TCR) which can be used as natural barcode for clonality (Burnet 1976). Here, we performed unsupervised hierarchical clustering using the presence or absence of clonal somatic mutations called by Conbase as input. We verified that cells that cluster together also shared the same TCR sequence. Finally, we performed an experimental validation by targeted sequencing of somatic mutations called by Conbase in additional clonally related cells. Our findings confirms that Conbase enables the identification of true somatic mutations in single cell data exhibiting high rates of allelic dropout.

While Conbase represents an advancement in the field, in particular for being useful for data exhibiting allelic dropout, this analysis is limited in two aspects. (1) Whole genome sequencing is costly, and analysis is therefore limited in terms of sample size. (2) Whole genome sequencing data does not contain information about cell type identity, and the experimental setup is required to include the collection of additional data, including for example protein expression. An interesting improvement of lineage tracing methods would enable cell lineage reconstruction and simultaneous cell type classification, for example through gene expression profiling using scRNAseq. This would, however, require the development of new methods that allow for the analysis of both of these read-outs.

2.5 PAPER III

In **Paper III**, we investigate the development of human memory and effector CD8+ T cells following vaccination. Here, we combine single cell RNA and DNA sequencing to study the origins of memory T cells through the analysis of the inheritance patterns of somatic variation at the single cell level.

CD8+ T cells play important roles in the acute defense against pathogens, but can also initiate a new immune response in case the pathogen is re-encountered in the future. These actions are carried out by clonally expanded populations of effector or memory T cells respectively. These cells are generated following activation of a clonally heterogeneous population of naïve T cells which rapidly divide and generate a diverse population of short-lived effectors and long-lived memory cell types (Kaech, Wherry et al. 2002). The ability of the immune system to remember pathogens is called long-term immunity. This can be achieved not only by infections but also by vaccines that train T cells to recognize a specific pathogen through the development of memory T cells.

Understanding the mechanisms that govern the ability of memory T cells to provide protection against future infections has recently gained substantial interest world-wide given the COVID-19 pandemic, including the quest to achieve long-term immunity by vaccination. One major unanswered question in this regard, involves the identification of the cellular origins of memory T cells. This is widely believed to be a key step towards enhancing the ability of vaccinations to generate clonally diverse populations of long-lived memory T cells with the highest potential to producing highly functional effector types when re-challenged (Ahmed, Bevan et al. 2009, Todryk 2018).

There is an active and ongoing debate concerning the origins of memory T cells and their relationship to effector T cells, and several models for T cell fate has been proposed (Ahmed, Bevan et al. 2009) (Figure 4). The fate restricted model suggests that naïve T cells give rise to daughter cells with either a memory or effector like phenotype which will be preserved upon further division (Buchholz, Flossdorf et al. 2013, Gerlach, Rohr et al. 2013). The linear model proposes that naïve T cells give rise to rapidly expanding effector cells upon activation, which in turn differentiate into long-lived memory cells over time (Akondy, Fitch et al. 2017, Youngblood, Hale et al. 2017). In contrast, the memory-to-effector model suggests that the naïve T cells upon activation give rise to long-lived memory T cells with the capacity to self-renew as well as give rise to fast-expanding but short-lived effector cells (Ahmed, Bevan et al. 2009). The asymmetric model suggests that naïve T cells give rise to distinct memory and effector cell lineages, which will maintain their phenotype upon further cell division (Chang, Palanivel et al. 2007).

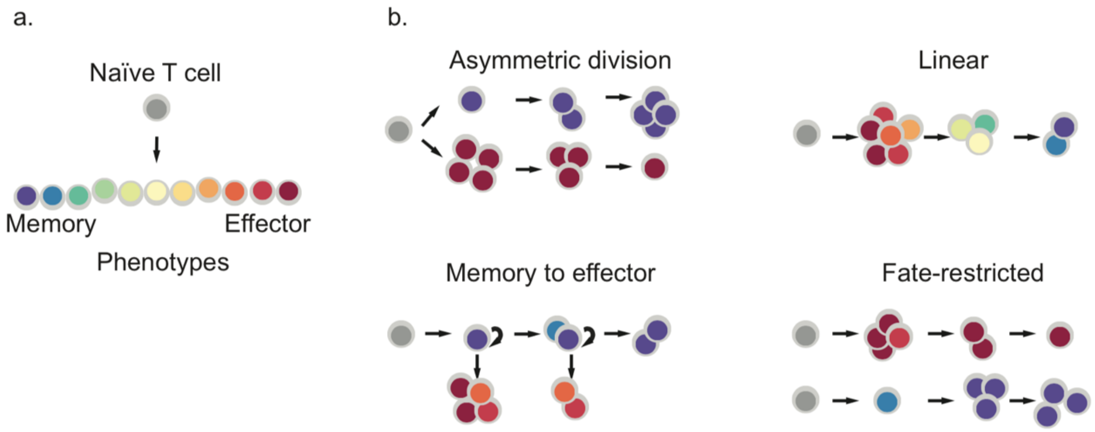

Figure 4. (a) Naïve T cells give rise to clonal populations of T cells with distinct memory and effector phenotypes. (b) Hypotheses for T cell fate.

To study the development of memory and effector CD8+ T cells in an immune response, we vaccinated human donors against yellow fever virus (YFV-17D), and collected antigen-specific CD8+ T cells at multiple time points after vaccination. Following collection, single cells were subjected to RNA or DNA sequencing. This allowed us to track the behavior of individual T cell clones in a longitudinal fashion.

To distinguish phylogenetic cell lineages in scRNAseq data, we developed a new computational method which enables cell lineage reconstruction based on mitochondrial mutations. This is done by a greedy search approach, where phylogenetic trees are built by inserting one mutation node at a time while iteratively scoring the resulting tree by the likelihood of observing the mutation data given the tree. This strategy opens up the possibility of obtaining a posterior distribution of trees given the mutation data. Such confidence interval of trees allows us to meaningfully compare different trees by their likelihood score and to explore trees with close-to-optimal scores. As compared to published methods for lineage tracing based on mitochondrial mutations (Xu, Nuno et al. 2019, Lareau, Ludwig et al. 2021), our approach generates trees in which the node and branch structure has a defined meaning and can be used for quantitative analysis, including for example relatedness between cells and mutation counts.

Using this strategy, we find that CD8+ T cells collected in the acute phase of the immune response had acquired a higher number of mutations as compared to memory T cells collected at late time points of the immune response. To further validate this finding, we quantified uniquely detected somatic SNVs in nuclear genomic DNA from single CD8+ T cells subjected to MDA and whole genome sequencing using a software based on read-backed phasing (Bohrson, Barton et al. 2019). This allowed us to compare the division histories of clonally related CD8+ T cell subsets, based on the assumption that somatic mutations are accumulated during cell division. In accordance with findings from analyzing mitochondrial mutations, we find that CD8+ T cells collected at early timepoints after vaccination harbor a higher number of somatic mutations also in nuclear genomic DNA, as compared to memory T cells, defined as being observed at late timepoints. This finding suggests that CD8+ memory T cells may not originate from rapidly expanding effector T cells, as previously proposed (Akondy, Fitch et al. 2017, Youngblood, Hale et al. 2017), but instead that CD8+ memory T cells are formed through asymmetric division or early memory specification in humans.

In ongoing analyses, we are aiming to measure the phylogenetic distance between clonally related CD8+ T cell populations, defined by their gene expression profiles and protein surface marker expression. The goal of these analyses is to characterize gene expression signatures of cells that share recent common ancestors. This will allow us to obtain a more detailed view of the evolutionary history of individual T cell clones, and has the potential to identify the cellular origins of human memory T cells. In addition to detailed analysis of phylogenetic trees generated based on the inheritance patterns of mitochondrial mutations, we will generate phylogenetic trees based on somatic SNVs detected in nuclear genomic DNA by applying the algorithm described in Paper II combined with available methods for phylogenetic inference including for example SCITE and Maximum Compatibility (Jahn, Kuipers et al. 2016, Cherry 2017, Hard, Al Hakim et al. 2019).

3 SUMMARY AND PERSPECTIVES

All the cells in our body were generated through cell division, forming cell lineages that trace back to a single fertilized egg cell. The developmental path of this process can be referred to as a cellular pedigree, a “family tree” at the cellular level. In this regard, such a cell lineage tree represents the division histories of cell lineages and their development into particular cell types. How a specific cell type is developed, and its relationship to other cells, are fundamental questions that may have important implications for how we view the human body in the healthy physiological state and in disease. The quest to understand the relationships between cell lineages is of interest for multiple fields in biology and medicine, including for example developmental biology, regenerative medicine and cancer research. Deciphering the division histories of cells can provide insight into tissue development and homeostasis, and how this is affected in disease. Whereas there are many ways to trace the history of cell lineages in experimental animals, for example by genetic labeling of cells and their progeny, it is vastly more challenging to perform these analyses in human tissues due to technical limitations to distinguish clonally related cells.

In the current work, I set out to develop a strategy to determine cell genealogy in humans, by taking advantage of genetic variation as natural barcodes for clonality. In **Paper I**, we applied this strategy in a rather extreme situation, in patients who had received an allogeneic hematopoietic stem cell transplant, with the goal of distinguishing host and donor cells. This revealed that donor cells participated in generating, in addition to blood cells, adipocytes in these patients. In this setting, we could rely on germline variation that distinguishes the host and donor DNA. However, when analyzing very closely related cells, it becomes much more challenging to distinguish the few mutations that are unique for individual cells from artifacts of various origins. In **Paper II,** we, therefore, developed a new computational method, Conbase, for the identification of somatic mutations at the single cell level. Conbase leverages phased read data from multiple samples in a dataset to account for errors and biases associated with single cell DNA sequencing data. We evaluated the performance of Conbase to identify clones of cells in humans by analyzing CD8+ T cells responding to a yellow fever vaccine. T cells are an ideal cell type for developing methodologies for lineage tracing, as they have an inherent clonal genetic mark by their TCR sequences, which served as a positive control for clonal identification. This enabled us to demonstrate that we had developed a method to identify the clonal relationship of individual cells in humans, opening a window into lineage tracing studies in humans. As an extension of this work, we next set out to perform cell lineage tracing at larger scales, with simultaneous cell type classification by gene expression profiling. In **Paper III**, we, therefore, developed a new computational method based on inheritance patterns of mitochondrial mutations that can be detected in single cell RNA sequencing data. We applied our methodology on human CD8+ T cells responding to a yellow fever vaccine. This allowed us to track the development of clonal CD8+ T cell lineages. Our results suggest that CD8+ memory T cells develop as distinct lineages with a slower division rate as compared to clonally related CD8+ effector T cells. This finding indicates that asymmetric division or early memory specification models are likely to account for memory T cell origins in humans.

The work described in this thesis has evolved in parallel with the rapid development of technologies for single cell sequencing. This has provided a continuously expanding toolkit for dissecting cellular heterogeneity, including new methodologies for cell lineage reconstruction. A particularly interesting extension of technologies for cell lineage tracing based on inheritance patterns of somatic mutations includes applications for spatial transcriptomics (Stahl, Salmen et al. 2016) which would enable classical histological evaluation, but with molecular imaging of the transcriptome with simultaneous cell lineage analysis. Such methodology will be broadly useful to investigate fundamental questions regarding the contribution of cell lineages across diverse tissues within a spatial context. With an increased understanding of tissue biology, lineage tracing is likely to be an important tool for studying development and regeneration also in the future. Analysis of somatic variation, therefore, has the potential to represent a paradigm shift in lineage tracing since it opens up the possibility for reaching the ultimate goal of assessing cell lineage relationships in humans. It will be very interesting to see how lineage tracing strategies will be further integrated in single cell analyses as well as the next generations of sequencing technologies, including for example spatial transcriptomics.

4 ACKNOWLEDGEMENT

I would like thank **Jonas Frisén** for giving me the opportunity to work on these unique and exciting projects. It has been tremendously inspiring to work with you. When the scientific projects were challenging, your supervision always helped me to see solutions and opportunities.

I would also like to express my gratitude to the wonderful and bright members in the Frisén lab. In particular, I would like to thank **Jeff Mold**, for supporting me from day one and for being my friend. Your knowledge and passion for science is a huge inspiration. Thank you **Moa Stenudd** for being a great friend and reliable source of various information during these years. **Embla Steiner**, you are one of the nicest persons I know and I really enjoyed our science discussions. **Camilla Engblom**, you are a super star scientist and top human being. **Johanna Classon**, thank you for all the nice moments and travels we have done together. **Carl-Johan Eriksson**, thank you for bringing positive energy and for spending your spare time to work with me. Thank you also **Leonie Von Berlin**, **Marta Paterlini**, **Margherita Zamboni**, **Ionut Dumitru, Sarantis Giatrellis, Michael Ratz, Enric Llorens Bobadilla, Helena Lönnqvist, Giuseppe Santopolo, Kanar Alkass, Mathew Tata**, **Qirong Lin**, **Marie Kindblom, Ezeddin Al Hakim, Jens Magnusson, Pedro Reu, Maggie Yeung, Sofia Zdunek, Hanna Sabelström, Olaf Bergman, James Chell, Patrik Ståhl** and past members and students of the Frisén labs for making these years so great!

Many thanks to my collaborators for your interest in my work and for welcoming me into your labs, in particular **Afshin Ahmadian**, **Erik Borgström**, **Mikael Rydén**, **Peter Arner, Jakob Michaelsson**, **Kirsty Spalding**, **Adam Ameur, Katharina Jahn and Jack Kuipers** as well as the Spatial Transcriptomics group at KTH.

Finally, I would like to thank my lovely family, extended family, friends and Pietro.

5 REFERENCES

Abyzov, A., J. Mariani, D. Palejev, Y. Zhang, M. S. Haney, L. Tomasini, A. F. Ferrandino, L. A. Rosenberg Belmaker, A. Szekely, M. Wilson, A. Kocabas, N. E. Calixto, E. L. Grigorenko, A. Huttner, K. Chawarska, S. Weissman, A. E. Urban, M. Gerstein and F. M. Vaccarino (2012). "Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells." Nature **492**(7429): 438-442.

Ahmed, R., M. J. Bevan, S. L. Reiner and D. T. Fearon (2009). "The precursors of memory: models and controversies." Nat Rev Immunol **9**(9): 662-668.

Akondy, R. S., M. Fitch, S. Edupuganti, S. Yang, H. T. Kissick, K. W. Li, B. A. Youngblood, H. A. Abdelsamed, D. J. McGuire, K. W. Cohen, G. Alexe, S. Nagar, M. M. McCausland, S. Gupta, P. Tata, W. N. Haining, M. J. McElrath, D. Zhang, B. Hu, W. J. Greenleaf, J. J. Goronzy, M. J. Mulligan, M. Hellerstein and R. Ahmed (2017). "Origin and differentiation of human memory CD8 T cells after vaccination." Nature **552**(7685): 362-367.

Alvarez-Dolado, M., R. Pardal, J. M. Garcia-Verdugo, J. R. Fike, H. O. Lee, K. Pfeffer, C. Lois, S. J. Morrison and A. Alvarez-Buylla (2003). "Fusion of bone-marrow-derived cells with Purkinje neurons, cardiomyocytes and hepatocytes." Nature **425**(6961): 968-973.

Anderson, D. J., F. H. Gage and I. L. Weissman (2001). "Can stem cells cross lineage boundaries?" Nat Med **7**(4): 393-395.

Bakker, B., A. Taudt, M. E. Belderbos, D. Porubsky, D. C. Spierings, T. V. de Jong, N. Halsema, H. G. Kazemier, K. Hoekstra-Wakker, A. Bradley, E. S. de Bont, A. van den Berg, V. Guryev, P. M. Lansdorp, M. Colome-Tatche and F. Foijer (2016). "Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies." Genome Biol **17**(1): 115.

Ballouz, S., A. Dobin and J. A. Gillis (2019). "Is it time to change the reference genome?" Genome Biol **20**(1): 159.

Baumer, C., E. Fisch, H. Wedler, F. Reinecke and C. Korfhage (2018). "Exploring DNA quality of single cells for genome analysis with simultaneous whole-genome amplification." Sci Rep **8**(1): 7476.

Behjati, S., M. Huch, R. van Boxtel, W. Karthaus, D. C. Wedge, A. U. Tamuri, I. Martincorena, M. Petljak, L. B. Alexandrov, G. Gundem, P. S. Tarpey, S. Roerink, J. Blokker, M. Maddison, L. Mudie, B. Robinson, S. Nik-Zainal, P. Campbell, N. Goldman, M. van de Wetering, E. Cuppen, H. Clevers and M. R. Stratton (2014). "Genome sequencing of normal cells reveals developmental lineages and mutational processes." Nature **513**(7518): 422-425.

Berry, R. and M. S. Rodeheffer (2013). "Characterization of the adipocyte cellular lineage in vivo." Nat Cell Biol **15**(3): 302-308.

Bishop, C. (2011). Pattern Recognition and Machine Learning, Springer-verlag New York inc.

Blanpain, C. and E. Fuchs (2006). "Epidermal stem cells of the skin." Annu Rev Cell Dev Biol **22**: 339-373.

Blokzijl, F., J. de Ligt, M. Jager, V. Sasselli, S. Roerink, N. Sasaki, M. Huch, S. Boymans, E. Kuijk, P. Prins, I. J. Nijman, I. Martincorena, M. Mokry, C. L. Wiegerinck, S. Middendorp, T. Sato, G. Schwank, E. E. Nieuwenhuis, M. M. Verstegen, L. J. van der Laan, J. de Jonge, I. J. JN, R. G. Vries, M. van de Wetering, M. R. Stratton, H. Clevers, E. Cuppen and R. van Boxtel (2016). "Tissue-specific mutation accumulation in human adult stem cells during life." Nature **538**(7624): 260-264.

Bohrson, C. L., A. R. Barton, M. A. Lodato, R. E. Rodin, L. J. Luquette, V. V. Viswanadham, D. C. Gulhan, I. Cortes-Ciriano, M. A. Sherman, M. Kwon, M. E. Coulter,

A. Galor, C. A. Walsh and P. J. Park (2019). "Linked-read analysis identifies mutations in single-cell DNA-sequencing data." Nat Genet **51**(4): 749-754.

Brunner, S. F., N. D. Roberts, L. A. Wylie, L. Moore, S. J. Aitken, S. E. Davies, M. A. Sanders, P. Ellis, C. Alder, Y. Hooks, F. Abascal, M. R. Stratton, I. Martincorena, M. Hoare and P. J. Campbell (2019). "Somatic mutations and clonal dynamics in healthy and cirrhotic human liver." Nature **574**(7779): 538-542.

Buchholz, V. R., M. Flossdorf, I. Hensel, L. Kretschmer, B. Weissbrich, P. Graf, A. Verschoor, M. Schiemann, T. Hofer and D. H. Busch (2013). "Disparate individual fates compose robust CD8+ T cell immunity." Science **340**(6132): 630-635.

Buckingham, M. E. and S. M. Meilhac (2011). "Tracing cells for tracking cell lineage and clonal behavior." Dev Cell **21**(3): 394-409.

Burnet, F. M. (1976). "A modification of Jerne's theory of antibody production using the concept of clonal selection." CA Cancer J Clin **26**(2): 119-121.

Chang, J. T., V. R. Palanivel, I. Kinjyo, F. Schambach, A. M. Intlekofer, A. Banerjee, S. A. Longworth, K. E. Vinup, P. Mrass, J. Oliaro, N. Killeen, J. S. Orange, S. M. Russell, W. Weninger and S. L. Reiner (2007). "Asymmetric T lymphocyte division in the initiation of adaptive immune responses." Science **315**(5819): 1687-1691.

Cherry, J. L. (2017). "A practical exact maximum compatibility algorithm for reconstruction of recent evolutionary history." BMC Bioinformatics **18**(1): 127.

Conklin, E. G. (1905). "The organization and cell lineage of the ascidian egg." J. Acad. Nat. Sci. (Philadelphia)(13): 1-119.

Crisan, M., S. Yap, L. Casteilla, C. W. Chen, M. Corselli, T. S. Park, G. Andriolo, B. Sun, B. Zheng, L. Zhang, C. Norotte, P. N. Teng, J. Traas, R. Schugar, B. M. Deasy, S. Badylak, H. J. Buhring, J. P. Giacobino, L. Lazzari, J. Huard and B. Peault (2008). "A perivascular origin for mesenchymal stem cells in multiple human organs." Cell Stem Cell **3**(3): 301-313.

Crossno, J. T., Jr., S. M. Majka, T. Grazia, R. G. Gill and D. J. Klemm (2006). "Rosiglitazone promotes development of a novel adipocyte population from bone marrow-derived circulating progenitor cells." J Clin Invest **116**(12): 3220-3228.

Czyz, Z. T., S. Kirsch and B. Polzer (2015). "Principles of Whole-Genome Amplification." Methods Mol Biol **1347**: 1-14.

de Queiroz, K. (2013). "Nodes, branches, and phylogenetic definitions." Syst Biol **62**(4): 625-632.

Dean, F. B., S. Hosono, L. Fang, X. Wu, A. F. Faruqi, P. Bray-Ward, Z. Sun, Q. Zong, Y. Du, J. Du, M. Driscoll, W. Song, S. F. Kingsmore, M. Egholm and R. S. Lasken (2002). "Comprehensive human genome amplification using multiple displacement amplification." Proc Natl Acad Sci U S A **99**(8): 5261-5266.

Dietmaier, W., A. Hartmann, S. Wallinger, E. Heinmoller, T. Kerner, E. Endl, K. W. Jauch, F. Hofstadter and J. Ruschoff (1999). "Multiple mutation analyses in single tumor cells with improved whole genome amplification." Am J Pathol **154**(1): 83-95.

Dong, X., L. Zhang, B. Milholland, M. Lee, A. Y. Maslov, T. Wang and J. Vijg (2017). "Accurate identification of single-nucleotide variants in whole-genome-amplified single cells." Nat Methods **14**(5): 491-493.

Dou, Y., H. D. Gold, L. J. Luquette and P. J. Park (2018). "Detecting Somatic Mutations in Normal Cells." Trends Genet **34**(7): 545-557.

El-Kebir, M. (2018). "SPhyR: tumor phylogeny estimation from single-cell sequencing data under loss and error." Bioinformatics **34**(17): i671-i679.

Ellegren, H. (2004). "Microsatellites: simple sequences with complex evolution." Nat Rev Genet **5**(6): 435-445.

Elson, J. L., D. C. Samuels, D. M. Turnbull and P. F. Chinnery (2001). "Random intracellular drift explains the clonal expansion of mitochondrial DNA mutations with age." Am J Hum Genet **68**(3): 802-806.

Esteban, J. A., M. Salas and L. Blanco (1993). "Fidelity of phi 29 DNA polymerase. Comparison between protein-primed initiation and DNA polymerization." J Biol Chem **268**(4): 2719-2726.

Evrony, G. D., X. Cai, E. Lee, L. B. Hills, P. C. Elhosary, H. S. Lehmann, J. J. Parker, K. D. Atabay, E. C. Gilmore, A. Poduri, P. J. Park and C. A. Walsh (2012). "Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain." Cell **151**(3): 483-496.

Faulkner, G. J. and V. Billon (2018). "L1 retrotransposition in the soma: a field jumping ahead." Mob DNA **9**: 22.

Frumkin, D., A. Wasserstrom, S. Kaplan, U. Feige and E. Shapiro (2005). "Genomic variability within an organism exposes its cell lineage tree." PLoS Comput Biol **1**(5): e50.

Gage, F. H. (2000). "Mammalian neural stem cells." Science **287**(5457): 1433-1438.

Garcia-Marques, J., I. Espinosa-Medina and T. Lee (2021). "The art of lineage tracing: From worm to human." Prog Neurobiol **199**: 101966.

Garvin, T., R. Aboukhalil, J. Kendall, T. Baslan, G. S. Atwal, J. Hicks, M. Wigler and M. C. Schatz (2015). "Interactive analysis and assessment of single-cell copy-number variations." Nat Methods **12**(11): 1058-1060.

Gawad, C., W. Koh and S. R. Quake (2014). "Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics." Proc Natl Acad Sci U S A **111**(50): 17947-17952.

Gerlach, C., J. C. Rohr, L. Perie, N. van Rooij, J. W. van Heijst, A. Velds, J. Urbanus, S. H. Naik, H. Jacobs, J. B. Beltman, R. J. de Boer and T. N. Schumacher (2013). "Heterogeneous differentiation patterns of individual CD8+ T cells." Science **340**(6132): 635-639.

Goodier, J. L. (2016). "Restricting retrotransposons: a review." Mob DNA **7**: 16.

Grun, D. and A. van Oudenaarden (2015). "Design and Analysis of Single-Cell Sequencing Experiments." Cell **163**(4): 799-810.

Gulcher, J. (2012). "Microsatellite markers for linkage and association studies." Cold Spring Harb Protoc **2012**(4): 425-432.

Hagemann-Jensen, M., C. Ziegenhain, P. Chen, D. Ramskold, G. J. Hendriks, A. J. M. Larsson, O. R. Faridani and R. Sandberg (2020). "Single-cell RNA counting at allele and isoform resolution using Smart-seq3." Nat Biotechnol **38**(6): 708-714.

Haque, A., J. Engel, S. A. Teichmann and T. Lonnberg (2017). "A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications." Genome Med **9**(1): 75.

Hard, J., E. Al Hakim, M. Kindblom, A. K. Bjorklund, B. Sennblad, I. Demirci, M. Paterlini, P. Reu, E. Borgstrom, P. L. Stahl, J. Michaelsson, J. E. Mold and J. Frisen (2019). "Conbase: a software for unsupervised discovery of clonal somatic mutations in single cells through read phasing." Genome Biol **20**(1): 68.

Hashimshony, T., F. Wagner, N. Sher and I. Yanai (2012). "CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification." Cell Rep **2**(3): 666-673.

Hazen, J. L., G. G. Faust, A. R. Rodriguez, W. C. Ferguson, S. Shumilina, R. A. Clark, M. J. Boland, G. Martin, P. Chubukov, R. K. Tsunemoto, A. Torkamani, S. Kupriyanov, I. M. Hall and K. K. Baldwin (2016). "The Complete Genome Sequences, Unique Mutational Spectra, and Developmental Potency of Adult Neurons Revealed by Cloning." Neuron **89**(6): 1223-1236.

Hope, K. and M. Bhatia (2011). "Clonal interrogation of stem cells." Nat Methods **8**(4 Suppl): S36-40.

Hsu, P. D., E. S. Lander and F. Zhang (2014). "Development and applications of CRISPR-Cas9 for genome engineering." Cell **157**(6): 1262-1278.

Hsu, Y. C. (2015). "Theory and Practice of Lineage Tracing." Stem Cells **33**(11): 3197-3204.

Jahn, K., J. Kuipers and N. Beerenwinkel (2016). "Tree inference for single-cell data." Genome Biol **17**: 86.
Johansson, C. B., S. Youssef, K. Koleckar, C. Holbrook, R. Doyonnas, S. Y. Corbel, L. Steinman, F. M. Rossi and H. M. Blau (2008). "Extensive fusion of haematopoietic cells with Purkinje neurons in response to chronic inflammation." Nat Cell Biol **10**(5): 575-583.
Kaech, S. M., E. J. Wherry and R. Ahmed (2002). "Effector and memory T-cell differentiation: implications for vaccine development." Nat Rev Immunol **2**(4): 251-262.
Kapli, P., Z. Yang and M. J. Telford (2020). "Phylogenetic tree building in the genomic age." Nat Rev Genet **21**(7): 428-444.
Kashima, Y., Y. Sakamoto, K. Kaneko, M. Seki, Y. Suzuki and A. Suzuki (2020). "Single-cell sequencing techniques from individual to multiomics analyses." Exp Mol Med **52**(9): 1419-1427.
Kimura, M. (1969). "The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations." Genetics **61**(4): 893-903.
Klein, A. M., L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz and M. W. Kirschner (2015). "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells." Cell **161**(5): 1187-1201.
Kolodziejczyk, A. A., J. K. Kim, V. Svensson, J. C. Marioni and S. A. Teichmann (2015). "The technology and biology of single-cell RNA sequencing." Mol Cell **58**(4): 610-620.
Krause, D. S., N. D. Theise, M. I. Collector, O. Henegariu, S. Hwang, R. Gardner, S. Neutzel and S. J. Sharkis (2001). "Multi-organ, multi-lineage engraftment by a single bone marrow-derived stem cell." Cell **105**(3): 369-377.
Kretzschmar, K. and F. M. Watt (2012). "Lineage tracing." Cell **148**(1-2): 33-45.
Kuipers, J., K. Jahn and N. Beerenwinkel (2017). "Advances in understanding tumour evolution through single-cell sequencing." Biochim Biophys Acta Rev Cancer **1867**(2): 127-138.
Kwok, A. W. C., C. Qiao, R. Huang, M.-H. Sham, J. W. K. Ho and Y. Huang (2021). "MQuad enables clonal substructure discovery using single cell mitochondrial variants." bioRxiv: 2021.2003.2027.437331.
Lahnemann, D., J. Koster, E. Szczurek, D. J. McCarthy, S. C. Hicks, M. D. Robinson, C. A. Vallejos, K. R. Campbell, N. Beerenwinkel, A. Mahfouz, L. Pinello, P. Skums, A. Stamatakis, C. S. Attolini, S. Aparicio, J. Baaijens, M. Balvert, B. Barbanson, A. Cappuccio, G. Corleone, B. E. Dutilh, M. Florescu, V. Guryev, R. Holmer, K. Jahn, T. J. Lobo, E. M. Keizer, I. Khatri, S. M. Kielbasa, J. O. Korbel, A. M. Kozlov, T. H. Kuo, B. P. F. Lelieveldt, Mandoiu, II, J. C. Marioni, T. Marschall, F. Molder, A. Niknejad, L. Raczkowski, M. Reinders, J. Ridder, A. E. Saliba, A. Somarakis, O. Stegle, F. J. Theis, H. Yang, A. Zelikovsky, A. C. McHardy, B. J. Raphael, S. P. Shah and A. Schonhuth (2020). "Eleven grand challenges in single-cell data science." Genome Biol **21**(1): 31.
Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C.

Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki and C. International Human Genome Sequencing (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Langmore, J. P. (2002). "Rubicon Genomics, Inc." Pharmacogenomics **3**(4): 557-560.

Lareau, C. A., L. S. Ludwig, C. Muus, S. H. Gohil, T. Zhao, Z. Chiang, K. Pelka, J. M. Verboon, W. Luo, E. Christian, D. Rosebrock, G. Getz, G. M. Boland, F. Chen, J. D. Buenrostro, N. Hacohen, C. J. Wu, M. J. Aryee, A. Regev and V. G. Sankaran (2021). "Massively parallel single-cell mitochondrial DNA genotyping and chromatin profiling." Nat Biotechnol **39**(4): 451-461.

Laurenti, E. and B. Gottgens (2018). "From haematopoietic stem cells to complex differentiation landscapes." Nature **553**(7689): 418-426.

Le Douarin, N. (1973). "A biological cell labeling technique and its use in expermental embryology." Dev Biol **30**(1): 217-222.

Le Douarin, N. M. and M. A. Teillet (1973). "The migration of neural crest cells to the wall of the digestive tract in avian embryo." J Embryol Exp Morphol **30**(1): 31-48.

Lee-Six, H., N. F. Obro, M. S. Shepherd, S. Grossmann, K. Dawson, M. Belmonte, R. J. Osborne, B. J. P. Huntly, I. Martincorena, E. Anderson, L. O'Neill, M. R. Stratton, E. Laurenti, A. R. Green, D. G. Kent and P. J. Campbell (2018). "Population dynamics of normal human blood inferred from somatic mutations." Nature **561**(7724): 473-478.

Lee-Six, H., S. Olafsson, P. Ellis, R. J. Osborne, M. A. Sanders, L. Moore, N. Georgakopoulos, F. Torrente, A. Noorani, M. Goddard, P. Robinson, T. H. H. Coorens, L. O'Neill, C. Alder, J. Wang, R. C. Fitzgerald, M. Zilbauer, N. Coleman, K. Saeb-Parsy, I. Martincorena, P. J. Campbell and M. R. Stratton (2019). "The landscape of somatic mutation in normal colorectal epithelial cells." Nature **574**(7779): 532-537.

Li, H. (2014). "Toward better understanding of artifacts in variant calling from high-coverage samples." Bioinformatics **30**(20): 2843-2851.

Lin, G., M. Garcia, H. Ning, L. Banie, Y. L. Guo, T. F. Lue and C. S. Lin (2008). "Defining stem and progenitor cells within adipose tissue." Stem Cells Dev **17**(6): 1053-1063.

Livet, J., T. A. Weissman, H. Kang, R. W. Draft, J. Lu, R. A. Bennis, J. R. Sanes and J. W. Lichtman (2007). "Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system." Nature **450**(7166): 56-62.
Lodato, M. A., R. E. Rodin, C. L. Bohrson, M. E. Coulter, A. R. Barton, M. Kwon, M. A. Sherman, C. M. Vitzthum, L. J. Luquette, C. N. Yandava, P. Yang, T. W. Chittenden, N. E. Hatem, S. C. Ryu, M. B. Woodworth, P. J. Park and C. A. Walsh (2018). "Aging and neurodegeneration are associated with increased mutations in single human neurons." Science **359**(6375): 555-559.
Lodato, M. A., M. B. Woodworth, S. Lee, G. D. Evrony, B. K. Mehta, A. Karger, S. Lee, T. W. Chittenden, A. M. D'Gama, X. Cai, L. J. Luquette, E. Lee, P. J. Park and C. A. Walsh (2015). "Somatic mutation in single human neurons tracks developmental and transcriptional history." Science **350**(6256): 94-98.
Lopez-Otin, C., M. A. Blasco, L. Partridge, M. Serrano and G. Kroemer (2013). "The hallmarks of aging." Cell **153**(6): 1194-1217.
Ludwig, L. S., C. A. Lareau, J. C. Ulirsch, E. Christian, C. Muus, L. H. Li, K. Pelka, W. Ge, Y. Oren, A. Brack, T. Law, C. Rodman, J. H. Chen, G. M. Boland, N. Hacohen, O. Rozenblatt-Rosen, M. J. Aryee, J. D. Buenrostro, A. Regev and V. G. Sankaran (2019). "Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics." Cell **176**(6): 1325-1339 e1322.
Lynch, M. (2010). "Rate, molecular spectrum, and consequences of human mutation." Proc Natl Acad Sci U S A **107**(3): 961-968.
Macosko, E. Z., A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev and S. A. McCarroll (2015). "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets." Cell **161**(5): 1202-1214.
Majka, S. M., H. L. Miller, T. Sullivan, P. F. Erickson, R. Kong, M. Weiser-Evans, R. Nemenoff, R. Moldovan, S. A. Morandi, J. A. Davis and D. J. Klemm (2012). "Adipose lineage specification of bone marrow-derived myeloid cells." Adipocyte **1**(4): 215-229.
Martincorena, I. and P. J. Campbell (2015). "Somatic mutation in cancer and normal cells." Science **349**(6255): 1483-1489.
McConnell, M. J., M. R. Lindberg, K. J. Brennand, J. C. Piper, T. Voet, C. Cowing-Zitron, S. Shumilina, R. S. Lasken, J. R. Vermeesch, I. M. Hall and F. H. Gage (2013). "Mosaic copy number variation in human neurons." Science **342**(6158): 632-637.
Metzger, D. and P. Chambon (2001). "Site- and time-specific gene targeting in the mouse." Methods **24**(1): 71-80.
Miller, T. E., C. A. Lareau, J. A. Verga, D. Ssozi, L. S. Ludwig, C. E. Farran, G. K. Griffin, A. A. Lane, B. E. Bernstein, V. G. Sankaran and P. van Galen (2021). "Mitochondrial variant enrichment from high-throughput single-cell RNA-seq resolves clonal populations." bioRxiv: 2021.2003.2008.434450.
Mizutani, E., M. Oikawa, H. Kassai, K. Inoue, H. Shiura, R. Hirasawa, S. Kamimura, S. Matoba, N. Ogonuki, H. Nagatomo, K. Abe, T. Wakayama, A. Aiba and A. Ogura (2015). "Generation of cloned mice from adult neurons by direct nuclear transfer." Biol Reprod **92**(3): 81.
Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer and B. Wold (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." Nat Methods **5**(7): 621-628.
Mullis, K. B. and F. A. Faloona (1987). "Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction." Methods Enzymol **155**: 335-350.
Negrini, S., V. G. Gorgoulis and T. D. Halazonetis (2010). "Genomic instability--an evolving hallmark of cancer." Nat Rev Mol Cell Biol **11**(3): 220-228.
Nelson, D. L., S. A. Ledbetter, L. Corbo, M. F. Victoria, R. Ramirez-Solis, T. D. Webster, D. H. Ledbetter and C. T. Caskey (1989). "Alu polymerase chain reaction: a method for

rapid isolation of human-specific sequences from complex DNA sources." Proc Natl Acad Sci U S A **86**(17): 6686-6690.
Nielsen, R., J. S. Paul, A. Albrechtsen and Y. S. Song (2011). "Genotype and SNP calling from next-generation sequencing data." Nat Rev Genet **12**(6): 443-451.
O'Hara, R., E. Tedone, A. Ludlow, E. Huang, B. Arosio, D. Mari and J. W. Shay (2019). "Quantitative mitochondrial DNA copy number determination using droplet digital PCR with single-cell resolution." Genome Res **29**(11): 1878-1888.
Orban, P. C., D. Chui and J. D. Marth (1992). "Tissue- and site-specific DNA recombination in transgenic mice." Proc Natl Acad Sci U S A **89**(15): 6861-6865.
Osborne, M. A., C. L. Barnes, S. Balasubramanian and D. Klenerman (2001). "Probing DNA Surface Attachment and Local Environment Using Single Molecule Spectroscopy." J. Phys. Chem. B(105): 3120-3126.
Payne, B. A. and P. F. Chinnery (2015). "Mitochondrial dysfunction in aging: Much progress but many unresolved questions." Biochim Biophys Acta **1847**(11): 1347-1353.
Picelli, S., A. K. Bjorklund, O. R. Faridani, S. Sagasser, G. Winberg and R. Sandberg (2013). "Smart-seq2 for sensitive full-length transcriptome profiling in single cells." Nat Methods **10**(11): 1096-1098.
Picher, A. J., B. Budeus, O. Wafzig, C. Kruger, S. Garcia-Gomez, M. I. Martinez-Jimenez, A. Diaz-Talavera, D. Weber, L. Blanco and A. Schneider (2016). "TruePrime is a novel method for whole-genome amplification from single cells based on TthPrimPol." Nat Commun **7**: 13296.
Piotrowski, A., C. E. Bruder, R. Andersson, T. Diaz de Stahl, U. Menzel, J. Sandgren, A. Poplawski, D. von Tell, C. Crasto, A. Bogdan, R. Bartoszewski, Z. Bebok, M. Krzyzanowski, Z. Jankowski, E. C. Partridge, J. Komorowski and J. P. Dumanski (2008). "Somatic mosaicism for copy number variation in differentiated human tissues." Hum Mutat **29**(9): 1118-1124.
Post, Y. and H. Clevers (2019). "Defining Adult Stem Cell Function at Its Simplest: The Ability to Replace Lost Cells through Mitosis." Cell Stem Cell **25**(2): 174-183.
Ramskold, D., S. Luo, Y. C. Wang, R. Li, Q. Deng, O. R. Faridani, G. A. Daniels, I. Khrebtukova, J. F. Loring, L. C. Laurent, G. P. Schroth and R. Sandberg (2012). "Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells." Nat Biotechnol **30**(8): 777-782.
Redon, R., S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. Gonzalez, M. Gratacos, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer and M. E. Hurles (2006). "Global variation in copy number in the human genome." Nature **444**(7118): 444-454.
Richard, G. F., A. Kerrest and B. Dujon (2008). "Comparative genomics and molecular dynamics of DNA repeats in eukaryotes." Microbiol Mol Biol Rev **72**(4): 686-727.
Richard, S. and M. W. Schuster (2002). "Stem cell transplantation and hematopoietic growth factors." Curr Hematol Rep **1**(2): 103-109.
Ross, E. M. and F. Markowetz (2016). "OncoNEM: inferring tumor evolution from single-cell sequencing data." Genome Biol **17**: 69.
Roth, A., A. McPherson, E. Laks, J. Biele, D. Yap, A. Wan, M. A. Smith, C. B. Nielsen, J. N. McAlpine, S. Aparicio, A. Bouchard-Cote and S. P. Shah (2016). "Clonal genotype and population structure inference from single-cell tumor sequencing." Nat Methods **13**(7): 573-576.
Ryden, M., M. Uzunel, J. L. Hard, E. Borgstrom, J. E. Mold, E. Arner, N. Mejhert, D. P. Andersson, Y. Widlund, M. Hassan, C. V. Jones, K. L. Spalding, B. M. Svahn, A.

Ahmadian, J. Frisen, S. Bernard, J. Mattsson and P. Arner (2015). "Transplanted Bone Marrow-Derived Cells Contribute to Human Adipogenesis." Cell Metab **22**(3): 408-417.

Saiki, R. K., S. Scharf, F. Faloona, K. B. Mullis, G. T. Horn, H. A. Erlich and N. Arnheim (1985). "Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia." Science **230**(4732): 1350-1354.

Salas, L. A., J. K. Wiencke, D. C. Koestler, Z. Zhang, B. C. Christensen and K. T. Kelsey (2018). "Tracing human stem cell lineage during development using DNA methylation." Genome Res **28**(9): 1285-1295.

Salipante, S. J. and M. S. Horwitz (2006). "Phylogenetic fate mapping." Proc Natl Acad Sci U S A **103**(14): 5448-5453.

Sandberg, R. (2014). "Entering the era of single-cell transcriptomics in biology and medicine." Nat Methods **11**(1): 22-24.

Sasagawa, Y., I. Nikaido, T. Hayashi, H. Danno, K. D. Uno, T. Imai and H. R. Ueda (2013). "Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity." Genome Biol **14**(4): R31.

Sera, Y., A. C. LaRue, O. Moussa, M. Mehrotra, J. D. Duncan, C. R. Williams, E. Nishimoto, B. A. Schulte, P. M. Watson, D. K. Watson and M. Ogawa (2009). "Hematopoietic stem cell origin of adipocytes." Exp Hematol **37**(9): 1108-1120, 1120 e1101-1104.

Sermon, K., W. Lissens, H. Joris, A. Van Steirteghem and I. Liebaers (1996). "Adaptation of the primer extension preamplification (PEP) reaction for preimplantation diagnosis: single blastomere analysis using short PEP protocols." Mol Hum Reprod **2**(3): 209-212.

Singer, J., J. Kuipers, K. Jahn and N. Beerenwinkel (2018). "Single-cell mutation identification via phylogenetic inference." Nat Commun **9**(1): 5144.

Snippert, H. J., L. G. van der Flier, T. Sato, J. H. van Es, M. van den Born, C. Kroon-Veenboer, N. Barker, A. M. Klein, J. van Rheenen, B. D. Simons and H. Clevers (2010). "Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells." Cell **143**(1): 134-144.

Spalding, K. L., E. Arner, P. O. Westermark, S. Bernard, B. A. Buchholz, O. Bergmann, L. Blomqvist, J. Hoffstedt, E. Naslund, T. Britton, H. Concha, M. Hassan, M. Ryden, J. Frisen and P. Arner (2008). "Dynamics of fat cell turnover in humans." Nature **453**(7196): 783-787.

Spalding, K. L., R. D. Bhardwaj, B. A. Buchholz, H. Druid and J. Frisen (2005). "Retrospective birth dating of cells in humans." Cell **122**(1): 133-143.

Spangrude, G. J., S. Heimfeld and I. L. Weissman (1988). "Purification and characterization of mouse hematopoietic stem cells." Science **241**(4861): 58-62.

Stahl, P. L., F. Salmen, S. Vickovic, A. Lundmark, J. F. Navarro, J. Magnusson, S. Giacomello, M. Asp, J. O. Westholm, M. Huss, A. Mollbrink, S. Linnarsson, S. Codeluppi, A. Borg, F. Ponten, P. I. Costea, P. Sahlen, J. Mulder, O. Bergmann, J. Lundeberg and J. Frisen (2016). "Visualization and analysis of gene expression in tissue sections by spatial transcriptomics." Science **353**(6294): 78-82.

Stewart, J. B. and P. F. Chinnery (2015). "The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease." Nat Rev Genet **16**(9): 530-542.

Sulston, J. E., E. Schierenberg, J. G. White and J. N. Thomson (1983). "The embryonic cell lineage of the nematode Caenorhabditis elegans." Dev Biol **100**(1): 64-119.

Sun, J., A. Ramos, B. Chapman, J. B. Johnnidis, L. Le, Y. J. Ho, A. Klein, O. Hofmann and F. D. Camargo (2014). "Clonal dynamics of native haematopoiesis." Nature **514**(7522): 322-327.

Tabansky, I., A. Lenarcic, R. W. Draft, K. Loulier, D. B. Keskin, J. Rosains, J. Rivera-Feliciano, J. W. Lichtman, J. Livet, J. N. Stern, J. R. Sanes and K. Eggan (2013). "Developmental bias in cleavage-stage mouse blastomeres." Curr Biol **23**(1): 21-31.

Taylor, R. W., M. J. Barron, G. M. Borthwick, A. Gospel, P. F. Chinnery, D. C. Samuels, G. A. Taylor, S. M. Plusa, S. J. Needham, L. C. Greaves, T. B. Kirkwood and D. M. Turnbull (2003). "Mitochondrial DNA mutations in human colonic crypt stem cells." J Clin Invest **112**(9): 1351-1360.

Teixeira, V. H., P. Nadarajan, T. A. Graham, C. P. Pipinikas, J. M. Brown, M. Falzon, E. Nye, R. Poulsom, D. Lawrence, N. A. Wright, S. McDonald, A. Giangreco, B. D. Simons and S. M. Janes (2013). "Stochastic homeostasis in human airway epithelium is achieved by neutral competition of basal cell progenitors." Elife **2**: e00966.

Telenius, H., N. P. Carter, C. E. Bebb, M. Nordenskjold, B. A. Ponder and A. Tunnacliffe (1992). "Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer." Genomics **13**(3): 718-725.

Tewhey, R., V. Bansal, A. Torkamani, E. J. Topol and N. J. Schork (2011). "The importance of phase information for human genomics." Nat Rev Genet **12**(3): 215-223.

Tindall, K. R. and T. A. Kunkel (1988). "Fidelity of DNA synthesis by the Thermus aquaticus DNA polymerase." Biochemistry **27**(16): 6008-6013.

Todryk, S. M. (2018). "T Cell Memory to Vaccination." Vaccines (Basel) **6**(4).

Tomasetti, C. and B. Vogelstein (2015). "Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions." Science **347**(6217): 78-81.

Treangen, T. J. and S. L. Salzberg (2011). "Repetitive DNA and next-generation sequencing: computational challenges and solutions." Nat Rev Genet **13**(1): 36-46.

Upton, K. R., D. J. Gerhardt, J. S. Jesuadian, S. R. Richardson, F. J. Sanchez-Luque, G. O. Bodea, A. D. Ewing, C. Salvador-Palomeque, M. S. van der Knaap, P. M. Brennan, A. Vanderver and G. J. Faulkner (2015). "Ubiquitous L1 mosaicism in hippocampal neurons." Cell **161**(2): 228-239.

Vassilopoulos, G., P. R. Wang and D. W. Russell (2003). "Transplanted bone marrow regenerates liver by cell fusion." Nature **422**(6934): 901-904.

Wallace, D. C. (1992). "Mitochondrial genetics: a paradigm for aging and degenerative diseases?" Science **256**(5057): 628-632.

Wallace, D. C. and D. Chalkia (2013). "Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease." Cold Spring Harb Perspect Biol **5**(11): a021220.

Walther, V. and M. R. Alison (2016). "Cell lineage tracing in human epithelial tissues using mitochondrial DNA mutations as clonal markers." Wiley Interdiscip Rev Dev Biol **5**(1): 103-117.

Wang, X., H. Willenbring, Y. Akkari, Y. Torimaru, M. Foster, M. Al-Dhalimy, E. Lagasse, M. Finegold, S. Olson and M. Grompe (2003). "Cell fusion is the principal source of bone-marrow-derived hepatocytes." Nature **422**(6934): 897-901.

Watt, F. M. and K. B. Jensen (2009). "Epidermal stem cell diversity and quiescence." EMBO Mol Med **1**(5): 260-267.

Weimann, J. M., C. B. Johansson, A. Trejo and H. M. Blau (2003). "Stable reprogrammed heterokaryons form spontaneously in Purkinje neurons after bone marrow transplant." Nat Cell Biol **5**(11): 959-966.

Welch, J. S., T. J. Ley, D. C. Link, C. A. Miller, D. E. Larson, D. C. Koboldt, L. D. Wartman, T. L. Lamprecht, F. Liu, J. Xia, C. Kandoth, R. S. Fulton, M. D. McLellan, D. J. Dooling, J. W. Wallis, K. Chen, C. C. Harris, H. K. Schmidt, J. M. Kalicki-Veizer, C. Lu, Q. Zhang, L. Lin, M. D. O'Laughlin, J. F. McMichael, K. D. Delehaunty, L. A. Fulton, V. J. Magrini, S. D. McGrath, R. T. Demeter, T. L. Vickery, J. Hundal, L. L. Cook, G. W. Swift, J. P. Reed, P. A. Alldredge, T. N. Wylie, J. R. Walker, M. A. Watson, S. E. Heath, W. D. Shannon, N. Varghese, R. Nagarajan, J. E. Payton, J. D. Baty, S. Kulkarni, J. M. Klco, M. H. Tomasson, P. Westervelt, M. J. Walter, T. A. Graubert, J. F. DiPersio, L. Ding, E. R. Mardis and R. K. Wilson (2012). "The origin and evolution of mutations in acute myeloid leukemia." Cell **150**(2): 264-278.

Whitman, C. O. (1887). "A contribution to the history of the germ-layers in Clepsine." Journal of Morphology **1**: 105–182.

Whitman, C. O. (1887). "The embryology of Clepsine." Q J Microscop Sci (NS) **18**: 215–315.

Woodworth, M. B., K. M. Girskis and C. A. Walsh (2017). "Building a lineage from single cells: genetic techniques for cell lineage tracking." Nat Rev Genet **18**(4): 230-244.

Xu, J., K. Nuno, U. M. Litzenburger, Y. Qi, M. R. Corces, R. Majeti and H. Y. Chang (2019). "Single-cell lineage tracing by endogenous mutations enriched in transposase accessible mitochondrial DNA." Elife **8**.

Yang, Z. and B. Rannala (2012). "Molecular phylogenetics: principles and practice." Nat Rev Genet **13**(5): 303-314.

Youngblood, B., J. S. Hale, H. T. Kissick, E. Ahn, X. Xu, A. Wieland, K. Araki, E. E. West, H. E. Ghoneim, Y. Fan, P. Dogra, C. W. Davis, B. T. Konieczny, R. Antia, X. Cheng and R. Ahmed (2017). "Effector CD8 T cells dedifferentiate into long-lived memory cells." Nature **552**(7685): 404-409.

Zafar, H., N. Navin, K. Chen and L. Nakhleh (2019). "SiCloneFit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data." Genome Res **29**(11): 1847-1859.

Zafar, H., A. Tzen, N. Navin, K. Chen and L. Nakhleh (2019). "Comments on the model parameters in "SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models"." Genome Biol **20**(1): 95.

Zafar, H., Y. Wang, L. Nakhleh, N. Navin and K. Chen (2016). "Monovar: single-nucleotide variant detection in single cells." Nat Methods **13**(6): 505-507.

Zannettino, A. C., S. Paton, A. Arthur, F. Khor, S. Itescu, J. M. Gimble and S. Gronthos (2008). "Multipotential human adipose-derived stromal stem cells exhibit a perivascular phenotype in vitro and in vivo." J Cell Physiol **214**(2): 413-421.

Zhang, L., X. Cui, K. Schmitt, R. Hubert, W. Navidi and N. Arnheim (1992). "Whole genome amplification from a single cell: implications for genetic analysis." Proc Natl Acad Sci U S A **89**(13): 5847-5851.

Zhang, L. and J. Vijg (2018). "Somatic Mutagenesis in Mammals and Its Implications for Human Disease and Aging." Annu Rev Genet **52**: 397-419.