Chapter 1

Introduction

Upon fertilisation, the human zygote embarks upon a precisely orchestrated journey of cellular division, migration and differentiation that culminates in the formation of myriad specialised cells, tissues and organs that collaborate to form a new organism. The human body is composed of more than 30 trillion individual cells (Bianconi et al., 2013), with an enormous diversity in appearance, function, and localisation. For example, cells within the lining of the digestive tract are spatially confined and may exist for only a few days, while memory B lymphocytes can remain in circulation for decades. Despite this considerable diversity, the enormous collection of cells constituting a human body all originate from the same fertilised egg cell.

Understanding the path a cell takes from the zygote to its eventual developed form is a question at the heart of developmental biology. The life history of a cell can shed light on the manner in which cells are transformed into their respective cell types, the renewal and maintenance of tissues and the formation of whole organs. Furthermore, a deep knowledge of cell lineages can elucidate the origin of any disorder that is the result of human development or homeostasis going awry. Perhaps the most prominent example is the emergence of cancer, constituting a loss in the tight regulation of division, expansion and longevity established by normal human development.

An entire organism can be mapped onto a single family tree of cells, with the fertilised egg cell at its root and all the developed cells that currently or previously existed at its leaves. Combining the ancestries of many cells into one phylogeny allows a direct assessment of development in its entirety, from embryogenesis, through normal adult tissue homeostasis, to the potential emergence of abnormal expansions and carcinogenesis.

The research presented in this thesis uses DNA mutations that occur naturally after conception to reconstruct developmental phylogenies and the lineages of individual cells. This introduction provides an overview of the historical perspective and recent advances in

(1) lineage tracing in model organisms, (2) somatic mutagenesis in normal cells, and (3) human embryogenesis.

1.1 A tree of life in every organism

1.1.1 A historical perspective

The modern study of embryogenesis and lineage tracing was fuelled by three landmark scientific advances in the 18th and 19th century: (1) a gradual shift away from preformationism in favour of epigenesis as the dominant theory explaining embryogenesis, (2) the view that cells are a product of a division of a pre-existing cell, rather than spontaneously generated and (3) the theory of evolution and natural selection, and the ensuing debate on the relation between phylogeny and ontogeny.

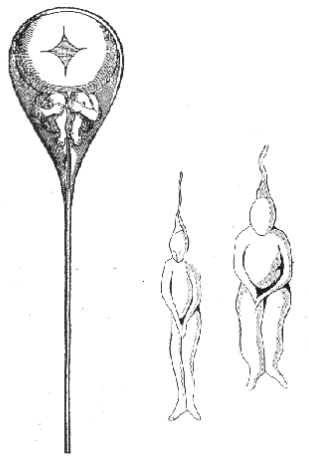

Figure 1.1 Drawing of a homunculus inside a sperm cell by Dutch mathematician and physicist Nicolaas Hartsoeker, 1695.

(1) The notion that an organism develops from a fertilised egg cell by division and differentiation (epigenesis) today seems obvious and incontestable, but historically, this was far from true (Needham and Hughes, 2015). The theory of epigenesis can be traced back to Aristotle in his work Περὶ ζῴων γενέσεως (*On the Generation of Animals*) and was elaborated further by Galen. However, epigenesis was not a generally accepted concept until well into the 19th century. For most of the intervening millennia, the theory of preformationism, the view that organisms develop from small, infinitesimal versions of themselves, was the dominant one. After Dutch microscopist (and staunch preformationist) Antonie van Leeuwenhoek discovered spermatozoa[1] in 1677, he postulated that these small vessels contained miniature versions of the animals they would seed. In other words, a human sperm cell would contain a homunculus, which already possessed all organs and characteristics of a full-grown human, including sperm cells with more homunculi, *ad infinitum* (**Fig. 1.1**). In essence, the preformationist theory postulates that all life was created at the same time. It is worth noting that, while this concept of "turtles all the way down" might appear absurd to a modern audience, Leibniz's and Descartes' view of infinite divisibility was still widely held at the time. In the second half of the 18th century, Prussian

[1] A literal translation of spermatozoon is "seed animal", as coined by embryologist Karl Ernst von Baer

physiologist Caspar Friedrich Wolff published two treatises (*Theoria Generationis* and *De Formatione Intestinorum*) which revived the concept of epigenesis (Roe, 2003). John Dalton's proof of the existence of the atom directly imposed a limit on the divisibility of matter and was irreconcilable with the notion of infinite homunculi. Hence, epigenesis replaced the preformation theory as the dominant view of human conception in the early 19th century.

(2) Another long-held view was that living organisms did not need to descend from pre-existing living organisms, and instead frequently arose spontaneously from non-living matter (Mazzarello, 1999). Again, this view had its origins in the natural philosophy of ancient Greece and was widely accepted for more than two millennia. Hugo von Mohl (1835) was the first to directly observe a cell division in plant cells, but this finding was not readily accepted nor generalised to animal cells. Matthias Schleiden (1838) postulated a centralised theory for plant cells, which was extended to animal cells by Theodor Schwann (1839). This cell theory had three tenets: (1) all living organisms consist of one or more cells, (2) the cell is the fundamental basic unit of life, and (3) cells form through crystallisation. While the first two still stand to this day, the latter was refuted decades later. Rudolf Virchow (1859) popularised the Latin dictum *omnis cellula e cellula* ("all cells from cells"), which decisively replaced the aforementioned third tenet in cell theory. In 1859, Louis Pasteur's famous experiment involving a swan neck flask finally disproved the notion of spontaneous generation altogether.

(3) Charles Darwin (1859) published his theory of evolution and natural selection, postulating that species arise from common ancestors. Besides the immediate implications of this new paradigm on the relationship between species, it also had a profound effect on thinking about the development of different organisms. In the decades prior to the publication of *On the Origin of Species*, embryologists, Karl Ernst von Baer among others, had begun comparing the developmental stages of different organisms, concluding that embryos from different vertebrates and invertebrates appeared to share common features. The formulation of the theory of natural selection functioned as a catalyst in connecting the study of the relationship between different types of organisms (phylogeny) and the study of the development of individual organisms from embryo to adult (ontogeny). Prominently, Ernst Haeckel (1866) embraced Darwin's writings and formulated his biogenetic law, which states that the ontogeny of the organism reflects its phylogeny.[2] In other words, before an organism can develop into its full adult self, it has to progress through the adult stages of 'lower' organisms from which it has evolved. His evidence included the observation that most animals go through similar developmental stages, such as the gastrula, a point he famously illustrated through comparative drawings in his work *Anthropogenie oder Entwicklungsgeschichte des Menschen*

[2]In fact, Haeckel himself coined the word 'phylogeny', a term that has extensively pervaded this dissertation.

(Haeckel, 1874). These drawings, depicting embryos from a variety of species such as fish, chickens and humans, drew much criticism in later decades for exaggerating the similarities between the different embryos. Wilhelm His (1888) rejected Haeckel's biogenetic theory and argued that early developmental stages of organisms do not represent adult versions of other organisms, but diverge during their ontogeny, essentially reverting to Von Baer's theories. This emerged as the dominant ontogenetic theory during the early 20th century and remains in place today.

Taken together, these three scientific advances led to the view that adult organisms develop from embryos by cell division and differentiation in a way that is specific to their species, but with broad similarities across different taxa. This view naturally leads to myriad scientific questions. What are the patterns of cell division and differentiation in the early embryo? How do these embryonic cells know which tissues and organs to form? Do these cells always follow the same pattern? What factors cause cellular differentiation and tissue morphogenesis to go awry and what can this tell us about malignant processes? The stage is set for the study of lineage tracing.

1.1.2 Early lineage tracing experiments

The earliest experiments attempting to trace the fate of individual cells through development relied on direct observation through light microscopy. An early endeavour to map the embryogenesis of an organism was performed by zoologist Charles O. Whitman (1878), who observed the fate of the zygote of the leech using this technique. Among the key discoveries in his pioneering research was that the development and patterning of cells after cleavage in the leech embryo was predetermined. Even after the earliest divisions, individual cells were already primed in terms of their eventual destination in the germ layers. The embryo as a whole was found to be tightly regulated and enforced the appropriate differentiation of each cell, rather than an autonomous and independent development of individual cells. This line of research was continued by Whitman himself, as well as his many students and colleagues, most prominently Edwin G. Conklin. In his work on the gastropod and its cell lineages, he discovered the emergence of the mesoderm as a single cell between the germ layers of endo- and ectoderm (Conklin, 1897).

The nematode *Caenorhabditis elegans* was a particularly popular organism for studying early development during the late 19th and early 20th century (Chitwood et al., 1937). In a similar fashion to leeches and gastropods, these nematodes exhibit highly determinate cell lineages in the early embryo. The advantage over the other invertebrates is that the adult *C. elegans* has a much lower number of cells, easing the burden of directly observing these lineages. This culminated in the complete mapping of the fate of every post-embryonic cell

by Sulston and Horvitz (1977), which was awarded the Nobel Prize for Physiology in 2002. Light microscopy-based lineage tracing was further extended to the vertebrate zebrafish (Kimmel et al., 1990), made possible by the transparency of its cells. Recent advances in light microscopy and computing have enabled automatic and precise methods of direct observation in the form of light sheet microscopy. This approach has been applied to the zebrafish (Keller et al., 2008), as well as the preimplantation mouse embryo (Strnad et al., 2016).

Microscopy-based lineage tracing has a number of limitations. While direct observation of embryogenesis is a tractable experiment for some species, it can only be applied to transparent organisms or developmental stages. Moreover, microscopic observation becomes unfeasible with a rapidly increasing number of cells or where the development of the organisms relies on an environment that is difficult to recreate *in vitro*, such as implantation in mammals.

Rather than retracing the origins of cells by directly observing their behaviour, the lineage of different cells can be determined by markers or barcodes that encode their developmental history and that can be retrospectively studied. In such experiments, cells need to be marked prospectively, at a single or a few time points. This imposes a limit on the capacity of the lineage tracing. Early studies traced cells by injecting them with dyes and radioactive material in order to visualise their fate and progeny (Kretzschmar and Watt, 2012). This approach has been applied to amphibian embryos (Vogt, 1929) and to mapping the development of the neural crest in chicken embryos (Serbedzija et al., 1989) and the neural plate in *Xenopus* embryos (Eagleson and Harris, 1990). Other compounds used for cell marking, such as horseradish peroxidase, cannot be excreted by cells and can only spread through cell division, after which it can be visualised using another compound. This has been applied to study the allocation of cells to the inner cell mass and trophectoderm in preimplantation mouse embryos (Balakier and Pedersen, 1982).

In recent decades, prospective cell marking by genetically modifying cells has become increasingly popular and potent, and has completely superseded dye-based lineage tracing (Kretzschmar and Watt, 2012). Genetic markers are generally more stable than dyes, as they are naturally replicated and passed on through cell division. Hence, these marks stay reliably confined to the progeny of the original cells that were barcoded. The earliest forms of this approach used retroviral vectors and transfection of reporter genes under specific promoters (Holland and Varmus, 1998; Lemischka et al., 1986). Cellular barcoding using genetic recombination techniques, such as Cre-LoxP, has also been widely applied for lineage tracing, notably to study the dynamics of stem cells (Barker et al., 2007; Morris et al., 2004; Pei et al., 2017). The usage of multicolour reporter constructs have greatly increased the granularity of lineage tracing (Yamamoto et al., 2009) and resulted in model systems such as the "confetti mouse" (Snippert et al., 2010).

1.1.3 Sequencing-based lineage tracing

The advent of next-generation sequencing techniques has revolutionised the life sciences and made it possible to answer biological questions on a previously unimaginable scale. A recent and particularly powerful technology is the sequencing of single cells, especially their RNA (Kolodziejczyk et al., 2015). By evaluating the expression profiles of single cells at different stages of development, it is possible to reconstruct the differentiation trajectories connecting distinct cell types. In this way, single-cell RNA sequencing can be used to study the pathways of differentiation that cells naturally undergo in various phases of embryogenesis (Nakamura et al., 2016; Xiang et al., 2020; Zhou et al., 2019), organogenesis (Pijuan-Sala et al., 2019), and adult tissue homeostasis (Giladi et al., 2018).

However, trajectories inferred from single-cell RNA sequencing rely on the identification of intermediate cell states, do not observe the cellular lineages directly and cannot answer quantitative questions about development, such as the number of progenitors responsible for a certain niche. Recently, single-cell RNA sequencing has been combined with genetic barcoding to perform large-scale lineage tracing experiments in model organisms. These methods rely on the CRISPR-Cas9 system to induce variable genetic scars at specific expressed sites in the genome, such that the genetic scars are detectable in the mRNA (Alemany et al., 2018; McKenna et al., 2016; Raj et al., 2018). Because of the high throughput of next-generation sequencing, these approaches are able to perform lineage tracing in entire organisms and resolve quantitative questions of development at an unprecedented scale, while simultaneously inferring cell types from the expression profiles. Within zebrafish, these studies have begun to shed light on the number of progenitor cells generating entire organs (McKenna et al., 2016), as well as the timing of divergence in bilaterally symmetric organs and dynamics of stem cells and tissue renewal (Alemany et al., 2018).

So far, I have discussed lineage tracing experiments in model organisms, either through direct observation via microscopy or through experimental modification via reporter constructions or specific genetic scarring. These approaches can only give insight into a limited period of development. Microscopy-based techniques can exclusively trace lineages as long as the observation lasts, with obvious further restrictions on the size of the specimen. Genetic editing approaches can only introduce cellular barcodes at discrete time points, and hence are confined to tracing the progeny of cells that were present and successfully marked at the time of experimental modification. Introducing lineage markers in multiple rounds only partially overcomes this limitation. Furthermore, it is difficult to exclude the possibility that experimental lineage tagging might impact cellular development. Besides these limitations, the invasiveness of genetic editing and other ethical concerns preclude the application of these approaches to studying human development.

1.2 Somatic mutations as natural markers

From fertilisation onwards, the cells of the human body naturally and continuously experience damage to their genome. This can be a consequence of either intrinsic causes, such as spontaneous deamination of methylated cytosines, or exposure to mutagens (Stratton et al., 2009) such as tobacco smoking or ultraviolet light. While the vast majority of DNA damage is repaired and the genome is replicated with extremely high fidelity, cells steadily acquire somatic mutations. The various categories of these mutations include single nucleotide variants (SNVs), double or multi-nucleotide variants (DNVs; MNVs), short insertions and deletions (indels), copy number variants (CNVs) and structural variants (SVs). Of these categories, the mutation rate of SNVs is the highest, estimated as two or three SNVs per cell doubling in the early embryo (Behjati et al., 2014; Ju et al., 2017), to 44 per year in human colonic crypts (Blokzijl et al., 2016; Lee-Six et al., 2019). Because of the continuous accumulation of these genetic scars, every cell will possess a near-unique set of somatic mutations.

The human genome is sufficiently large and the mutation rate during a single lifetime is low enough that we can invoke the infinite sites assumption. In other words, a given complement of mutations is only acquired once. Any mutations shared between two cells imply a shared developmental path, as they will be the progeny of the cell that gained that mutation. In line with this, the mutation rate is low enough so that the odds of "back mutations" are vanishingly small. In practice, somatic mutations can only be lost through chromosomal aberrations such as loss of heterozygosity.

Genetic relationships between cells are preserved throughout life such that early developmental patterns can be identified without the need to use embryonic or fetal material. In this section, I will review recent studies on the patterns of somatic mutagenesis in normal tissues and phylogenies that have been reconstructed such mutations. Beforehand, I will discuss the experimental or biological requirements to allow a reliable readout of somatic mutations in individual cells.

1.2.1 Genomic readouts of single cells

In order to leverage somatic mutations to study the relationship between different cells, it is necessary to obtain a read-out of the genome of those single cells. However, extracting the DNA from single cells and directly subjecting this to sequencing is not possible at the moment, due to the very low concentration of DNA. In practice, three methods are used to circumvent this limitation: (1) whole-genome amplification of single cell DNA, (2) *in vitro*

expansion of single cells into organoids or colonies and (3) laser capture microdissection (LCM) of naturally occurring clonal cell populations.

Single-cell genomics is the most direct method of obtaining mutational readouts (Lodato et al., 2015, 2018). However, this approach suffers from the loss of DNA content during the extraction (allelic dropout or the complete loss of entire segments of the genome) and a high load of artefactual mutations introduced during the whole-genome amplification. While this still allows for reliable detection of large-scale aneuploidies and CNVs (Cheng et al., 2011; Laks et al., 2019), it makes the calling and tracing of point mutations challenging.

Rather than artificially amplifying the DNA from a single cell, another approach is to amplify single cells into large aggregates of cellular progeny, such as cell lines, organoids (Behjati et al., 2014; Blokzijl et al., 2016; Yoshida et al., 2020), or (in the case of blood) colonies (Lee-Six et al., 2018; Osorio et al., 2018). In this way, DNA from the initial founder cell is amplified naturally into amounts suited to standard whole-genome sequencing. This removes the need for the error-prone step of whole-genome amplification. Mutations obtained *in vitro* can be distinguished from mutations present in the founder via their variant allele frequency (VAF), provided there was no clonal sweep during culturing. In addition, mutations shared between any two expanded populations should be unaffected by *in vitro* artefacts. However, not all cells are equally proficient at proliferating *in vitro*. Usually, only cells with a high replicative potential, such as stem cells, can be used as a basis for *in vitro* expansion and the success rate of these expansions varies dramatically between different tissues. This in turn introduces a bias in the potential tissues that can be used for such an experiment.

A third approach is based on LCM, which allows the targeted excision of specific cell populations or tissue units on a microscopic scale (Brunner et al., 2019; Ellis et al., 2021; Lawson et al., 2020; Lee-Six et al., 2019; Moore et al., 2020; Olafsson et al., 2020; Robinson et al., 2020). For example, LCM enables excision of cells belonging to one renal glomerulus or a single intestinal crypt from histological sections. Slight alterations in library preparation have made it possible to reliably sequence the DNA of as few as 100 cells from a single LCM cut (see Chapter 2). However, the tissue structures that are subjected to LCM and low-input whole-genome sequencing vary considerably in their clonality. Some structures such as colonic crypts or endometrial glands consistently represent the progeny of a single stem cell. In those cases, we can regard this as an *in vivo* expansion of the founder cell, in a similar vein to the *in vitro* expansions described above. However, not all tissue structures arise from single stem cells and some tissues, such as the sheets of epidermis or oesophageal epithelium, lack discrete structural units altogether. In practice, this confines the use of LCM-guided

whole-genome sequencing to those tissues that naturally possess morphologically defined clonal populations of cells.

One advantage of the LCM approach is the precise knowledge of the tissue structure (and hence often cell type composition) subjected to sequencing. More importantly, it is the only approach described here that retains the spatial information of sampling on a microscopic level. It allows sequencing of neighbouring units from the same slide, allows precise estimation of *in vivo* clone sizes in cases of expansions and allows for a spatial evaluation of developmental patterns.

1.2.2 Mutational processes in normal tissues

In addition to serving as a record of developmental phylogeny, each cell's unique set of mutations also reflects the specific mutational processes that have been at play during its life history. Distinct mutational processes, whether exogenous or endogenous, cause different patterns of mutations in the genome. These patterns, or mutational signatures, are most often displayed as probability distributions of the 96 different SNV categories defined by their trinucleotide contexts. However, these signatures can manifest in a wider nucleotide context or as indels, DNVs, and SVs as well. Initially, signatures of somatic mutagenesis were solely derived from cancer genomes and their aetiology has been linked to a variety of causes, such as defects in the DNA repair machinery, exposure to ultraviolet light, tobacco smoking or other mutagens, and cell-intrinsic DNA replication errors. This characterisation has led to a whole repertoire of mutational signatures, which is currently in its third iteration as maintained by the COSMIC database (Alexandrov et al., 2020).

Studies of the landscape of somatic mutations in normal tissues have revealed that the vast majority of SNVs in these cells can be attributed to a handful of mutational signatures: COSMIC reference signatures 1, 5 and 18 (**Fig. 1.2**). Signature 1 is characterised by C>T mutations at a CpG context and is the consequence of spontaneous deamination of methylated cytosine. This signature is ubiquitously present in all cancers and normal tissue and has been shown to accumulate in an age-dependent, clock-like fashion (Alexandrov et al., 2015). Signature 5 is similarly ubiquitous, but manifests as a flat, rather featureless signature. Its precise aetiology is currently unknown. While both signatures 1 and 5 appear to be present in all cells, the ratios between their exposures differ dramatically in different tissues. For example, colonic crypts have about equal exposures to signature 1 and 5 (Lee-Six et al., 2019), while signature 5 accounts for more than 80% of mutations in bronchial epithelium (Yoshida et al., 2020). It is unknown what precisely causes the difference in the ratio of signature 1 to 5. Lastly, signature 18 manifests in a wide variety of normal tissues, but at a much more limited incidence and exposure (Lee-Six et al., 2019; Moore et al., 2020;

Yoshida et al., 2020). Signature 18 is characterised by C>A mutations and has been linked to cellular stress and oxidative damage (Poetsch, 2020). Hence, this signature is mainly seen in normal tissues experiencing high levels of oxidative stress or undergoing rapid proliferation. However, it has also been observed in cell lines as an artefact of *in vitro* culturing (Petljak et al., 2019; Rouhani et al., 2016).

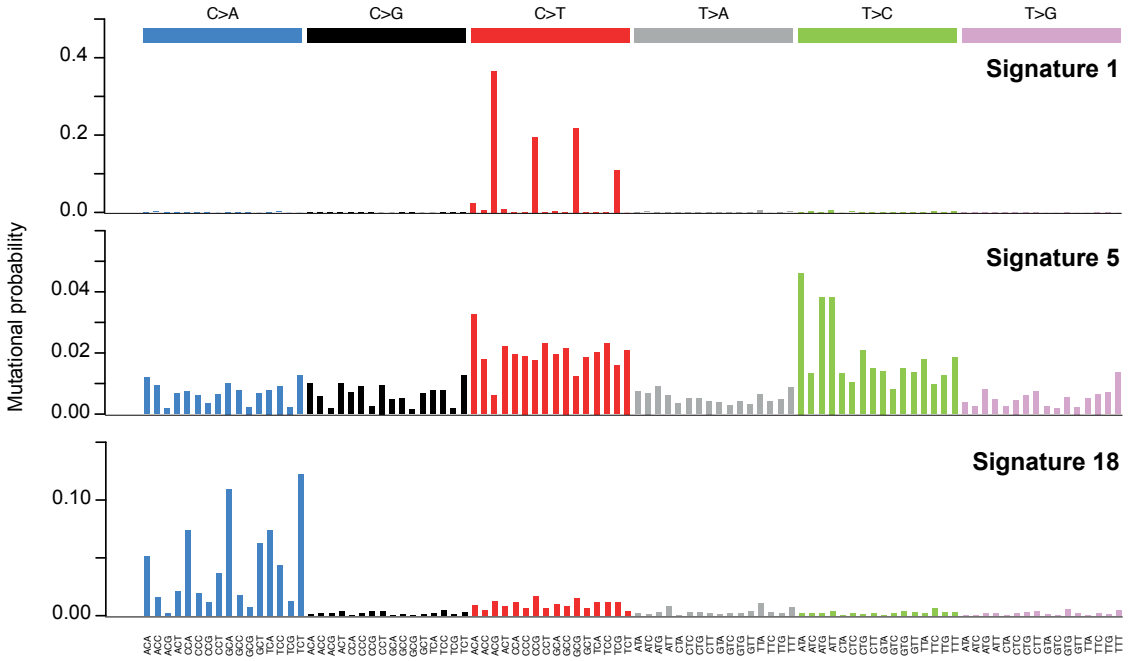

Figure 1.2 Bar plots of mutational probability for the three main mutational signatures involved in mutagenesis of normal tissues: signature 1 (spontaneous deamination of methylated cytosines), signature 5 (unknown aetiology), and signature 18 (oxidative stress).

While SNVs occur in abundance in normal tissues, indels and other types of somatic mutations are acquired at a much lower rate. Indel rates have been reported to be less than one tenth of the SNV rate in normal tissues, with DNVs at an even lower rate (Brunner et al., 2019; Lawson et al., 2020; Lee-Six et al., 2019; Moore et al., 2020). This is likely a consequence of the absence of a strong indel or DNV component in the few signatures governing normal mutagenesis, in contrast to e.g. signatures of ultraviolet light with a large proportion of DNVs (de Gruijl et al., 2001) and colibactin-induced mutagenesis with a large number of indels (Pleguezuelos-Manzano et al., 2020). CNVs and SVs are rarely identified in normal tissues, and are almost exclusively found in older individuals. This suggests that genomic instability is not a general feature of normal cells. This also indicates that the genomic integrity is sufficient to support the assumption that somatic point mutations will only rarely be lost due to chromosomal aberrations.

1.2.3 Early embryonic mutations and phylogenies

A small number of studies in the past six years have reconstructed developmental phylogenies from somatic mutations or investigated mutations arising in the early embryo. Here, I briefly summarise the findings of those studies in order to better compare and contrast the results presented in the subsequent chapters.

The first effort to reconstruct phylogenies of development from somatic mutations was a study on mouse organoids by Behjati et al. (2014). These organoids were derived from stomach, small bowel, large bowel and prostate from two mice in total, in addition to a bulk biopsy of the tail. This bulk sample represents a polyclonal aggregate of cells and can hence serve as a proxy to assess the contribution of each embryonic progenitor to the adult body. The variant allele frequency (VAF) of early embryonic mutations revealed that the first bifurcation in mouse development displayed an asymmetric contribution of the two daughter cells of the cell at the root (presumably the zygote) to the mice. The degree of asymmetry was approximately 2:1.

This early embryonic asymmetry was later recapitulated in humans through interrogation of matched blood samples from breast cancer patients by Ju et al. (2017). Rather than directly observing the phylogenies of development through genomes from single cells, this study used the VAF of embryonic variants in bulk blood samples to reconstruct the early asymmetries. In addition to corroborating the asymmetry of 2:1 observed in the mouse, this study reports an early mutation rate of approximately three variants per cell doubling which can largely be attributed to signatures 1 and 5.

The largest phylogeny of human cells published so far is a study by Lee-Six et al. (2018) on 140 *in vitro* expanded colonies of haematopoietic stem cells from a single patient. Using DNA from a buccal swab as a matched bulk to test body-wide embryonic contribution, this study also reported an early asymmetry of approximately 2:1.

It has been proposed that the cell allocation to the inner cell mass and the trophectoderm, the first lineage commitment in embryogenesis, causes this observed asymmetric contribution in mouse and human phylogenies (Behjati et al., 2014; Ju et al., 2017).

1.2.4 Driver mutations and cancer precursors

Most post-zygotic mutations occur in intergenic, non-coding regions of the genome and have very little or no impact on cellular phenotype. However, somatic mutations have the potential to profoundly alter the programming of individual cells if they occur in certain locations in the genome. Mutations may confer a positive selective advantage on the cell that harbours them by activating genes promoting cell proliferation (oncogenes) or inactivating genes

regulating and limiting cell growth and division (tumour suppressor genes). The sequential acquisition of such oncogenic (driver) mutations can transform normal, well-functioning cells into tumour cells.

The somatic mutation theory of cancer has been well-studied and substantiated over the past century (Boveri, 1914; Nowell, 1976; Stratton et al., 2009). A main aim of the cancer genomics studies of the past two decades has been to discover recurrent driver mutations in human cancers (Davies et al., 2002; ICGC TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020; Martincorena et al., 2017). These studies have identified many genes implicated in the pathogenesis of human cancers and revealed numerous recurrent hotspot mutations in oncogenes and inactivating or truncating mutations in tumour suppressor genes. However, statistical estimates of the number of coding sequences under selection in cancers indicate that only about half of cancer drivers fall in 369 known cancer genes that are unequivocally implicated in oncogenesis (Martincorena et al., 2017). The remaining drivers might have evaded detection due to low recurrence, leading to their omission from targeted cancer sequencing panels. More recently, the search for cancer-causing mutations has been expanded to include non-coding elements of the genome (Rheinbay et al., 2020) and heritable epigenetic modifications (Chatterjee et al., 2018)

However, most of the identified driver mutations are not sufficient to transform a healthy cell into a cancer cell. For example, endometrial glands have been shown to reliably harbour canonical driver mutations while still acting and appearing histologically like their unmutated neighbours (Moore et al., 2020). In fact, typically 50% of these glands will have a cancer driver at age 50, but the vast majority will never progress to a malignancy. In these tissues, the metamorphosis of normal cell into tumour cell appears to either require additional drivers or catastrophic genomic instability, such as large-scale chromosomal aberrations. The latter has rarely been found in normal tissues (Brunner et al., 2019; Lawson et al., 2020; Lee-Six et al., 2019; Moore et al., 2020). Such a genomic catastrophe might be what is needed to abruptly complete cellular dysregulation and realise the malignant potential of previously normal cells.

In some tissues, early signs of cancer manifest as large clonal expansions of normal cells that can be detected on an organ-wide level. Mutant clones are widespread throughout the normal human skin (Martincorena et al., 2015) and oesophagus (Martincorena et al., 2018), where their prevalence and size increases with age and mutagen exposure. Pre-malignant clonal expansion has been particularly well-described in blood, due to ease of representative sampling, and is termed clonal haematopoiesis. Clonal haematopoiesis is generally characterised by cells that are morphologically and phenotypically normal. The distinctive feature is that these cells have arisen from the same ancestral stem cell and

constitute a disproportionate amount of the blood cells in one individual (Challen and Goodell, 2020). Often these clones arise in the absence of an apparent canonical driver mutation (Zink et al., 2017). The fitness effect of these drivers is distinguishable from clonal growth due to genetic drift (Watson et al., 2020). Unsurprisingly, clonal haematopoiesis is associated with an increased risk of developing a haematological malignancy, though only a minority of affected individuals progress (Genovese et al., 2014; Jaiswal et al., 2014). Clonal haematopoiesis appears to be driven largely by endogenous mutagenesis and becomes ubiquitous with age (Zink et al., 2017).

Clonal expansions are thus a common feature of ageing normal tissues. Due to the constant accumulation of somatic mutations throughout life, it is possible to retrospectively reconstruct the development of these clones and time their emergence. This is a line of enquiry pursued in Chapter 3. This principle can also be applied to trace the origins of human cancers in the normal tissues from which they have arisen, which can reveal traces of pre-malignant clonal expansions. This is the focus of Chapter 4, which explores the embryonal origins of Wilms tumour.

1.3 Human embryogenesis and its bottlenecks

The final aim of this introductory chapter is to briefly review the early stages of human development and its lineage commitments. While our understanding of human embryogenesis has increased dramatically over the past two centuries, many questions about cellular trajectories and the mechanisms governing differentiation and tissue morphogenesis remain to be answered. Rather than providing an exhaustive narrative on all cellular processes in embryogenesis, this section will highlight certain aspects of early development that will feature in interpretation and discussion of the results in subsequent chapters.

1.3.1 Zygote, cleavage, and blastulation

The first step in human embryogenesis, after fertilisation, is a stage of rapid cell divisions without significant growth of the embryo as a whole. This phase is known as the cleavage stage. The axes and patterns of cell divisions during cleavage differ substantially between different taxa of the animal kingdom, with mammals exhibiting a rotational symmetry in the cleavage-stage embryo (Gulyas, 1975; Zernicka-Goetz, 2005). In contrast to the cleavage in non-mammalian model organisms, such as the zebrafish or chicken, cell divisions in the early human embryo are not fully synchronous (Milewski and Ajduk, 2017; Zernicka-Goetz,

2005), and hence the cell number does not increase strictly exponentially, which leads to the possibility of embryos with odd numbers of cells.

The zygote, by definition, is able to give rise to all embryonic and extraembryonic tissues, which is termed totipotency. This totipotency is proven to be retained by the early cells in the human embryo (blastomeres) until at least the 4-cell stage (De Paepe et al., 2014; Van de Velde et al., 2008), and likely into the 16-cell stage (De Paepe et al., 2013). Moreover, dye-based lineage tracing on a limited number of human embryos has indicated that generally all individual blastomeres in the two- to eight-cell stage contribute to both trophectoderm and inner cell mass (Mottla et al., 1995).

During the cleavage, the embryo transitions from relying on maternal proteins inherited from the oocyte to transcribing and translating its own genome. This process of maternal-to-zygotic transition starts with the well-coordinated zygotic genome activation around the eight-cell stage (Braude et al., 1988; Dobson et al., 2004), although a low level of zygotic transcription can be detected beforehand (Xue et al., 2013; Yan et al., 2013).

This handover of control to the genome of the embryo is underpinned by large-scale changes in the epigenetic landscape of the blastomeres (Eckersley-Maslin et al., 2018). Between the zygotic stage and the blastula stage, the human embryo is subjected to extensive DNA demethylation throughout the whole genome (Lee et al., 2014). This wave of demethylation occurs differently for the two sets of parental chromosomes, with maternal chromosomes being subjected to a slower, passive loss of methylation by cell division without methylation maintenance (Guo et al., 2014), while paternal chromosomes are actively and rapidly demethylated within the pronucleus before the completion of fertilisation (Guo et al., 2014; Iqbal et al., 2011). This landscape of global demethylation then persists until approximately the gastrula stage (Lee et al., 2014). Within this time frame, the methylation status of imprinted loci, i.e. with a consistent methylation pattern that leads to a parent-specific expression, are spared from the waves of de- and re-methylation (Lee et al., 2014; Weaver et al., 2009). These imprinted genes are often involved in proliferation, with the paternal and maternal methylation pattern promoting and inhibiting cell growth and division, respectively (Hurst and McVean, 1997; Maher et al., 2000). Loss of imprinting in these loci, either through uniparental disomy or an aberrant methylation patterns, frequently lead to over- or undergrowth syndromes such as Beckwith-Wiedemann syndrome (Maher et al., 2000; Weksberg et al., 2010). It is plausible that many of these imprinting disorders have their origin in the dynamic methylation mechanics of the early embryo.

At the eight-cell stage, likely as a result of the zygotic genome activation (Jukam et al., 2017), human blastomeres undergo a process called compaction (Iwata et al., 2014). This is the first morphological change of these blastomeres and involves the initiation of cell-to-cell

adhesion and other changes to the cell surface. The mechanism underpinning human embryo compaction remains unclear (Shahbazi, 2020). It is thought that compaction results in the first decision of cells to commit to the inner cell mass or the trophectoderm at the division between the eight- or 16-cell stage (Fleming, 1987; Morris et al., 2010; Strnad et al., 2016). However, much of this is based on mouse rather than human embryology. It appears that these lineage commitments occur largely on a spatial basis: the cells on the inside become the inner cell mass and the ones on the outside commit to the trophectoderm. The split between the inner cell mass and the trophectoderm represents the first lineage segregation of the human embryo. This segregation does not appear to be symmetric. In general, the cells allocated to trophectoderm outnumber the inner cell mass progenitors by a factor of four (Mottla et al., 1995). This signifies that, given an embryo consisting of 16 cells, roughly three would seed the inner cell mass and the remaining 13 would form the trophectoderm.

The commitment to trophectoderm and inner cell mass becomes more pronounced when the embryo transitions from the morula stage to the blastula stage, with a recognisable inner cavity filled with fluid, the blastocoel (**Fig. 1.3a**). The embryo is then referred to as a blastocyst.

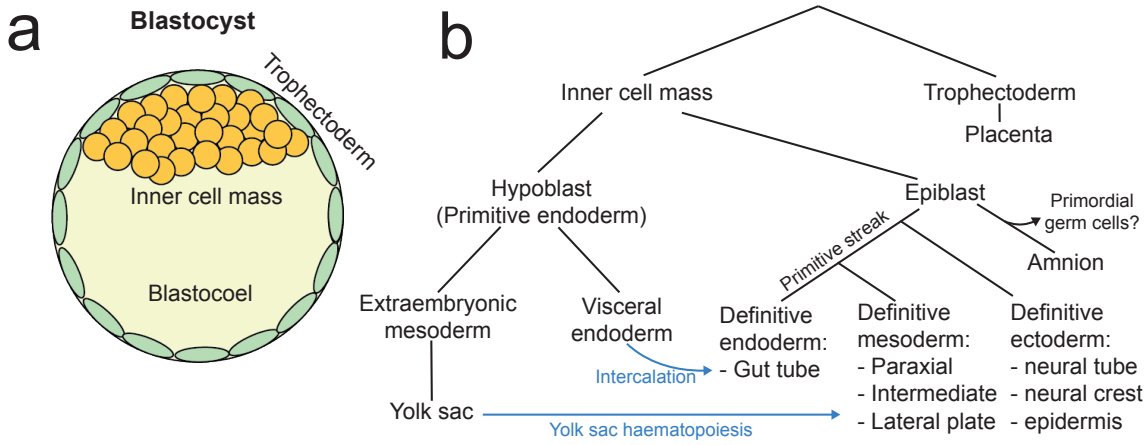

Figure 1.3 (a) Diagram representing the human embryo at the blastocyst stage (day 5 post-conception). (b) Overview of lineage commitments in the early human embryo, up until gastrulation and early organogenesis. Blue arrows indicate contribution of extraembryonic cells to embryonic lineages.

1.3.2 Symmetry breaking, gastrulation and extraembryonic intercalation

After formation of the blastula, which generally occurs on day 5 post-conception, the cells of the inner cell mass form a bilaminar embryonic disc and commit to one of two lineages: the hypoblast (also known as the primitive endoderm) and the epiblast (also known as the primitive ectoderm). The hypoblast forms the lining between the epiblast and the blastocoel and gives rise to the extraembryonic mesoderm and the yolk sac.

The cells that comprise the epiblastic layer will go on to form the definitive ectoderm and the primitive streak and hence the cells of the embryo proper. In addition, the epiblast will give rise to the amniotic ectoderm, which forms another membrane around the embryo. This membrane is known as the amnion and surrounds the amniotic cavity (Shahbazi, 2020). The precise origin of human primordial germ cells, the cells that will ultimately give rise to spermatocytes or oocytes, remains unclear, but they are thought to derive from the epiblast and sometimes more specifically from the amniotic epithelium (Kobayashi and Surani, 2018).

The hypoblast also plays an important role in the induction of a spatial axis in the epiblast and hence, the symmetry breaking of the epiblast disc. An analysis of human embryo single-cell RNA sequencing performed during my PhD suggests that the human hypoblast starts laying out the primordial body axis (anterior-posterior) on day 9 post-conception (Molè et al., 2021). At this stage, it is likely that cells in the epiblast, although not committed to a germ layer yet, might experience strong spatial biases in their lineage fates. In addition, the hypoblast appears to play a key role in the proliferation of the inner cell mass lineages by fibroblast growth factor-dependent signalling (Molè et al., 2021).

Upon gastrulation, which occurs approximately two weeks post-conception, the cells of the epiblast form the three major germ layers: endoderm, mesoderm and ectoderm. The cells of the ectoderm give rise to the neural tube and neural crest, and hence the human nervous system, as well as the epidermis. The cells of the endoderm form the gut tube and mainly produce the gastrointestinal tract and associated organs, such as the liver and pancreas. In addition, it gives rise to the lungs, the thyroid glands, and the prostate, among others. Lastly, the mesoderm commits to three different lineages: the paraxial mesoderm (producing the bones and skeletal muscle), the intermediate mesoderm (spawning kidneys, reproductive organs and the lower urinary tract) and lateral plate mesoderm (giving rise to the heart, blood, spleen and smooth muscles, among others).

However, tracing the lineage origins of the germ layers is further complicated by a process referred to as extraembryonic intercalation, in which previously extraembryonic tissues contribute cells to the definitive germ layers. This has primarily been shown in mouse

endoderm, which receives a widespread influx of primitive or visceral endoderm cells (Kwon et al., 2008; Viotti et al., 2014). In addition, it appears the extraembryonic contribution is most pronounced in the murine hindgut and least pronounced in the foregut (Pijuan-Sala et al., 2019), further adding a spatial dimension to this intercalation. In addition, part of the definitive mesoderm is thought to be derived from the extraembryonic mesoderm, mainly through blood cells that are produced by the hypoblast-derived yolk sac (Ferretti and Hadjantonakis, 2019). Extraembryonic intercalation has been shown in murine development, but remains to be fully generalised in humans, although there is evidence for human haematopoiesis in the yolk sac (Popescu et al., 2019).

This overview of lineage commitments in the early human embryo is visually depicted in **Fig. 1.3b**.

1.3.3 Aneuploidies in blastocysts and trophectoderm

It is worth noting that several studies have detected a large number of aneuploid cells in blastocysts, either through cytokinetic studies or single-cell sequencing efforts (Boué et al., 1975; Shahbazi et al., 2020; Voet et al., 2011). These findings contrast with the observation that normal adult human cells do not generally carry signs of aneuploidy. However, of all possible body-wide chromosomal losses and gains, only a handful have been observed in humans, most notably Down syndrome (trisomy 21), and a variety of syndromes affecting the sex chromosomes (Hassold et al., 1996). On the other hand, trisomy 16 is the most prevalent in human pregnancies, occurring in approximately one percent of conceptions, but always leads to a spontaneous abortion if present in all cells (Hassold et al., 1995). In total, it is estimated that a large proportion of human conceptions never progress to live birth, with an estimated success rate of 30%-40% (Larsen et al., 2013). In addition, it is estimated that over half of conceptions are lost prior to implantation (Wilcox et al., 2020).This suggests that the majority of chromosomal losses and gains detrimentally impact the normal course of embryogenesis in the cell they are present in, likely through arrested development or induction of apoptosis. This strong selection pressure prevents any aneuploid cells in the early embryo from contributing to the adult body, reconciling the high rate of aneuploidy in blastocysts with the low rate in normal adult tissues.

If this early aneuploidy is mosaic, the euploid cells are able to survive and form the embryo. Mosaic aneuploidies can arise from *de novo* acquisitions of chromosomal aberrations or post-zygotic reversals of aneuploidies, such as trisomic rescue (Los et al., 1998). One type of mosaic aneuploidy is confined placental mosaicism, where the placenta harbours an aneuploidy that is absent from the fetus (Kalousek and Dill, 1983). Confined placental mosaicism can be a consequence of an aforementioned trisomic rescue in the embryo

proper or an acquired aneuploidy in the placental lineage (Los et al., 1998; Sirchia et al., 1998). Confined placental mosaicism occurs in approximately one to two percent of human pregnancies (Hahnemann and Vejerslev, 1997). It is likely that this is due to aneuploidies lacking a sufficiently strong detrimental effect on cell proliferation in the trophectoderm, in contrast to the embryo proper.

1.4 Questions and outline of this dissertation

This introductory chapter has given an overview of the disciplines of lineage tracing, somatic mutagenesis and human embryology. These threads are woven together in the research presented in this dissertation, which follows the theme of leveraging naturally acquired mutations to trace the fate of individual cells through human development and in some cases, carcinogenesis. Using this approach, the research of this dissertation attempts to answer a wide variety of questions about cell dynamics in the early embryo, the embryonic spatial architecture and development of individual organs and tissues and the existence of pre-malignant precursor clones residing in normal tissues. Taken together, the analyses and results in this dissertation rely on a grand total of 1,086 whole-genome sequences. The presentation of this research is structured as follows:

- **Chapter 2: Materials and Methods.** This chapter contains all the information regarding the sampling of patients, sequencing of samples and the methodology of the analyses. A special emphasis is placed on the description of computational approaches, as the design of novel variant filters, phylogeny reconstruction strategies, and statistical frameworks were required for the work presented in this dissertation.

- **Chapter 3: Extensive phylogenies of human development** Large numbers of microdissected samples from three patients were used to reconstruct phylogenetic trees of human development. These trees reveal the earliest patterns of embryogenesis, such as an asymmetric contribution of early embryonic progenitors to the adult body and large-scale mosaic patterning of organs and tissues. In addition, these phylogenies show later adult clonal expansions in different normal tissues. This research has been published in *Nature* (Coorens et al., 2021b).

- **Chapter 4: Embryonal precursors of Wilms tumour** This chapter uses somatic mutations shared between a childhood kidney cancer, Wilms tumour, and normal renal samples to identify tissue-resident precursor lesions. These early aberrant clonal expansions, termed clonal nephrogenesis, are the consequence of hypermethylation of *H19*. This research has been published in *Science* (Coorens et al., 2019).

- **Chapter 5: Universal mosaicism of the human placenta** Somatic mutations identified in bulk placenta biopsies and microdissected trophoblast reveal the developmental architecture of this temporary organ. The human placenta is organised in large macroscopic, clonal patches, which carry considerable mutation burdens. Comparison of mutational patterns to umbilical cord indicated that trophectoderm and inner cell mass lineages can diverge at the earliest opportunity in development. Taken together, this naturally explains confined placental mosaicism, as illustrated by a case of trisomic rescue. This research has been published in *Nature* (Coorens et al., 2021c).