Lineage tracing of normal human development and childhood cancers

Tim Coorens

Wellcome Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Clare Hall

September 2020

I would like to dedicate this dissertation to family and friends close by and far away, who have never stopped making me smile and laugh.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text. This dissertation does not exceed 60,000 words in length.

Tim Coorens
September 2020

Lineage tracing of normal human development and childhood cancers

Tim Coorens

Summary

From fertilisation onwards, the cells of the human body continuously experience damage to their genome, either from intrinsic causes or from exposure to mutagens. While the vast majority of DNA damage is repaired and the genome is replicated with extremely high fidelity, cells steadily acquire single nucleotide variants throughout life. Since cells pass these genetic changes on to their descendants, mutations shared between any two cells therefore imply a shared developmental path. In essence, these somatic mutations connect all cells together into one large phylogenetic tree of human development with the zygote at the root.

Reconstructing phylogenies of human development requires readouts of somatic mutations present in single cells. Recently, low-input whole-genome sequencing following laser-capture microdissection has allowed us to reliably call somatic mutations in distinct single-cell derived physiological units, such as colonic crypts and endometrial glands, while retaining spatial information on a microscopic level. In this way, I reconstructed large-scale phylogenies of cells from many different organs of three individuals. These phylogenetic trees recapitulate the early stages of embryonic development and asymmetric cell allocation in the blastocyst, as well as later clonal expansions such as benign prostatic hyperplasia and neoplastic polyp formation.

In a similar way, I also used somatic mutations to investigate the emergence of paediatric cancer, which is thought to be closely linked to aberrations in development. In the context of phylogenetic analyses of tumours, mutations shared between childhood cancers and different normal tissues can shed light on the embryonic lineage of tumours and may reveal the precise juncture at which tumours began to form. Accordingly, I studied the origin of Wilms tumour, the most common childhood cancer of the kidney. I discovered that these tumours often arise from large tissue-resident precursor clones residing in the normal kidney. These embryonal precursors represent an early clonal expansion driven by *H19* hypermethylation.

Lastly, using somatic mutations I discovered that the human placenta is made up of large clonal patches of closely related trophoblast cells. Comparing early embryonic mutations between placental lineages and umbilical cord DNA, which is derived from the inner cell mass, revealed that in approximately half of the cases, a trophectodermal lineage shares no somatic mutations with the umbilical cord. Furthermore, in a quarter of cases, the umbilical cord is entirely derived from a progenitor later than the zygote. This indicates a natural early segregation between these lineages and a pathway to generate confined placental mosaicism.

This dissertation as a whole provides a new framework to study normal and aberrant human development from whole-genome sequencing. The ability to reconstruct developmental lineages retrospectively can answer fundamental questions about human development and carcinogenesis.

Acknowledgements

The work described in this thesis would have been impossible without the support of many people. First and foremost, I would like to thank my *de facto* triumvirate of supervisors: Professor Mike Stratton, for his deep insights, endless patience and delightful deliberations on Gibbon's *Decline and Fall* and Procopius' *Secret Histories*; Dr Iñigo Martincorena, for his invaluable analytic mind, steadfast support and imbuing me with a strong affinity for the beta-binomial distribution; Dr Sam Behjati, for his unwavering enthusiasm, endless ideas and the lengthy dialogues about right and wrong in the use of colour in figures.

I would also like to thank the other members of my thesis advisory committee, Dr Peter Campbell, Professor Elizabeth Murchison, Professor Magdalena Zernicka-Goetz, Dr Moritz Gerstung and Dr Nick Goldman, for all the helpful insights, discussions and comments along the road.

I would never have been able to perform this work without the support of the Wellcome Trust Mathematical Genomics and Medicine PhD programme, and I am very grateful to the directors and my fellow students over the years for making this programme so special.

I am especially indebted to the administrative staff, IT team and lab support that make the Sanger Institute run smoothly every day and made this research possible on numerous fronts. A special thanks to Jo Jones, Carl Logan and Wendy McLaughlin, the true unsung heroes of this work, whose impeccable handling of chaotic agendas and enquiries small and large helped me enormously. I am very grateful to all the colleagues I closely worked with over the years, especially Dr Luiza Moore, Dr Taryn Treger, Dr Thomas Oliver, Dr Raheleh Rahbari, Dr Roser Vento-Tormo and Dr Philip Robinson.

I am extremely fortunate to have to thank a small legion of friends, without whom this PhD would not nearly have been so much fun. At the risk of committing the heinous crime of omission, I would like to give a few friends a special thanks for extensively proofreading this dissertation, providing me with much-needed distractions and helping with the crosswords: Jannat Ijaz, Dr Pantelis Nicola, Dr Grace Collord, Dr Alex Cagan, Dr Timothy Butler, Dr Heather Machado, Andrew Lawson, Luke Harvey, Dr Patrick McClanahan, Dr Sophie Allcock, Timo Haber and Natalie Shatashvili.

It almost goes without saying that I want express my deepest gratitude to my family back in the Netherlands, to whom I owe everything and who have supported me every step along the way, especially my parents, Peter and Sandra, my siblings Jodie and Jim, my step-parents, Paul and Angelique, and of course, my grandparents.

And of course I would like to thank you, dear reader, for opening this dissertation and having a look at this work. With all sincerity, I hope you will enjoy this as much as I did.

Contents

List of Figures

List of Tables

Nomenclature

Greek Symbols

ρ	The overdispersion parameter in a beta-binomial distribution

Acronyms / Abbreviations

AIC	Akaike information criterion
AML	Acute myeloid leukaemia
BIC	Bayesian information criterion
ccRCC	Clear cell renal cell carcinoma
CMN	Congenital mesoblastic nephroma
cnn-LOH	Copy number neutral loss of heterozygosity
CNV	Copy number variant
DNV	Double nucleotide variant
Indel	Insertion and deletion
LCM	Laser capture microscopy/microdissection
LOH	Loss of heterozygosity
LOI	Loss of imprinting
MDS	Myelodysplastic syndrome
MNV	Multinucleotide variant
MRT	Malignant rhabdoid tumour

PGC	Primordial germ cell
RNA	Ribonucleic acid
SBS	Single base substitution
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SV	Structural variant
TGS	Targeted genome sequencing
VAF	Variant allele frequency
WES	Whole-exome sequencing
WGS	Whole-genome sequencing

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.

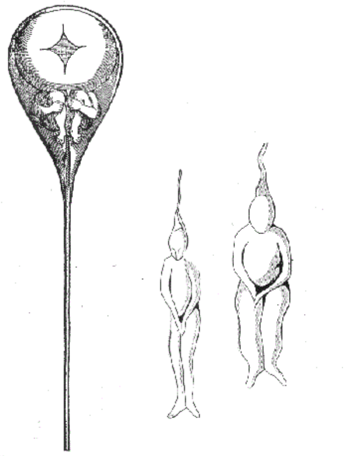

Figure 1.1 Drawing of a homunculus inside a sperm cell by Dutch mathematician and physicist Nicolaas Hartsoeker, 1695.

(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).

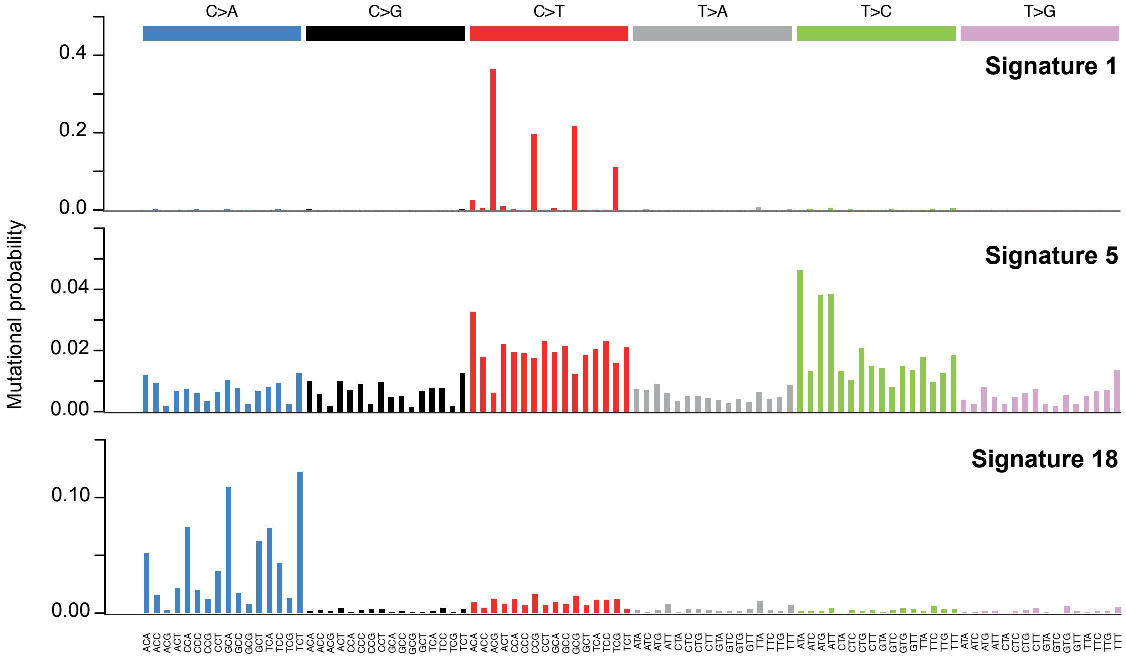

Figure 1.2 Bar plots of mutational probability for the three main mutational signatures involved in mutagenesis of normal tissues: signature 1 (spontaneous deamination of methylated cytosines), signature 5 (unknown aetiology), and signature 18 (oxidative stress).

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.

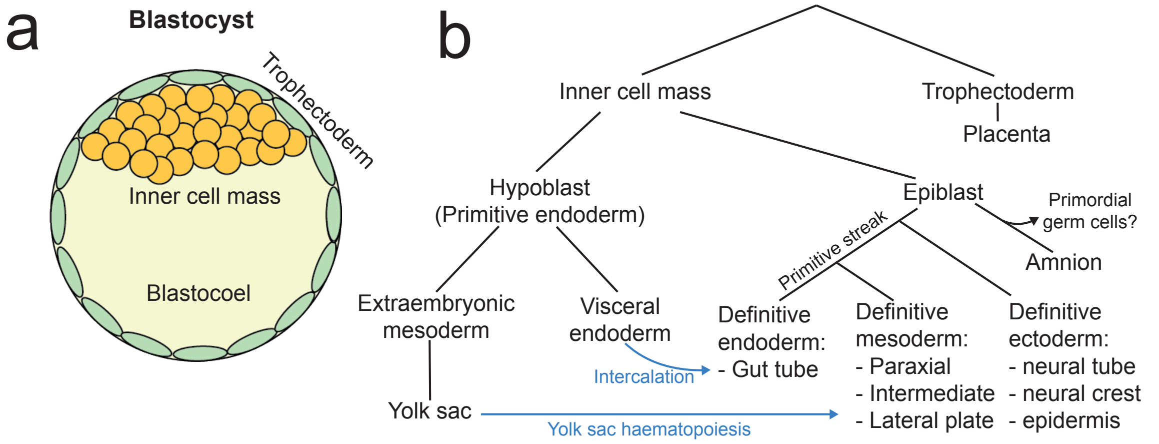

Figure 1.3 (a) Diagram representing the human embryo at the blastocyst stage (day 5 post-conception). (b) Overview of lineage commitments in the early human embryo, up until gastrulation and early organogenesis. Blue arrows indicate contribution of extraembryonic cells to embryonic lineages.

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).

Chapter 2

Materials and methods

Before presenting the individual studies and results that comprise this dissertation, this chapter will cover the strategies and methodologies behind sampling, sequencing and analyses. The contribution of others to the research presented in this thesis is explicitly listed here. A special emphasis is placed on the methods and algorithms involved in variant filtering, reconstruction of phylogenies, and other mutational analyses, as they constitute a major piece of work undertaken by me during the PhD. This chapter expands on analytical concepts, statistical principles and methods of calculation referred to in subsequent chapters.

2.1 Samples and sequencing

2.1.1 Ethics, patient sampling and data availability

The samples presented throughout this dissertation were obtained from numerous studies with their own patient basis and ethical approval. In short:

- **Chapter 3**: samples subjected to LCM were obtained from three rapid autopsies (NHS National Research Ethics Service reference 13/EE/0043). DNA sequencing data are deposited in the European Genome-Phenome Archive (EGA) with accession code EGAS00001003021. Detail of samples, including sequencing platform and coverage, can be found in the supplementary material of the associated paper (Coorens et al., 2021b).
- **Chapter 4**: adult kidney samples were obtained through the ‘Evaluation of biomarkers in urological disease’ study (NHS National Research Ethics Service reference 03/018). Children’s samples were acquired from patients enrolled in the ’Investigating how childhood tumours and congenital disease develop’ (NHS National Research Ethics

Service reference 16/EE/0394) or through the UK IMPORT study (NHS National Research Ethics Service reference 12/LO/0101). Samples subjected to LCM were obtained from a rapid autopsy (NHS National Research Ethics Service reference 13/EE/0043) and two patients with renal clear cell carcinoma (NHS National Research Ethics Service reference 16/WS/0039). Normal kidney biopsies were also obtained from a declined transplant donor (NHS National Research Ethics Service reference 15/EE/0152). Raw sequencing data have been deposited in EGA under study ID EGAD00001004774. Detail of samples, including sequencing platform and coverage, can be found in the supplementary material of the associated paper (Coorens et al., 2019).

- **Chapter 5**: all the samples were obtained from the Pregnancy Outcome Prediction (POP) study, a prospective cohort study of nulliparous women attending the Rosie Hospital, Cambridge (UK) for their dating ultrasound scan between January 14, 2008, and July 31, 2012. The study has been previously described in detail (Pasupathy et al., 2008; Sovio et al., 2017). Ethical approval for this study was given by the Cambridgeshire Research Ethics Committee (reference number 07/H0308/163) and all participants provided written informed consent. DNA sequencing data are deposited in EGA with accession code EGAD00001006337. Detail of samples, including sequencing platform and coverage, can be found in the supplementary material of the associated paper (Coorens et al., 2021c).

2.1.2 Laser capture microdissection

Excision of cell clusters by LCM was performed on a LMD7000 microscope (Leica) into a skirted 96-well PCR plate. Lysis of cells was done using 20µl proteinase-K Picopure DNA Extraction kit (Arcturus), after which LCM cuts were incubated at 65 °C for 3 hours followed by proteinase denaturation at 75 °C for half an hour. Afterwards, samples were stored at -20 °C in anticipation of DNA library preparation.

The DNA library preparation of microdissected tissue samples was undertaken using a bespoke low-input enzymatic-fragmentation-based library preparation method as described in a large number of recent LCM-based studies (Brunner et al., 2019; Ellis et al., 2021; Lee-Six et al., 2019; Moore et al., 2020; Olafsson et al., 2020; Robinson et al., 2020). This method was developed at the Wellcome Sanger Institute and spearheaded by Dr Peter Ellis recently (Ellis et al., 2021). This method was used as it allows for high quality DNA library preparation from very low starting quantity of material, without the need to include the error-prone step of whole-genome amplification associated with single-cell DNA sequencing.

The number of cells that are typically sampled differs per tissue and cutting strategy, but approximately ranges from 100 to 1000.

LCM cutting of biopsies for the research presented in this dissertation was performed by the following individuals: Dr Luiza Moore (Chapters 3 and 4), Dr Philip Robinson (Chapter 3) and Dr Thomas Oliver (Chapter 5).

2.1.3 Whole-genome sequencing

Short-insert (500bp) genomic libraries were constructed and 150 base pair paired-end sequencing clusters were generated on the Illumina HiSeq XTEN platform (Chapters 3 and 4) or the Illumina Novaseq platform (Chapter 5), in accordance with Illumina no-PCR library protocols. All DNA sequences were aligned to the GRCh37d5 reference genome using the Burrows-Wheeler algorithm (BWA-MEM) (Li and Durbin, 2009).

2.2 Somatic variant calling and filtering

As previously noted, traditionally, calling of somatic variants was largely based on a pairwise comparison between the sample of interest, often a tumour, and an utterly polyclonal, normal sample, often blood. Any variants present in both are considered to be inherited, while variants present in the sample of interest only will be acquired during life. However, genuine early embryonic variants, acquired during the first few divisions of life, will also be present in the matched normal sample, even if it is polyclonal. Given the importance of early post-zygotic mutations for lineage tracing of embryonic dynamics, a different approach to variant calling was necessary.

Single nucleotide variants (SNVs) were called using the in-house CaVEMan (Cancer Variants through Expectation Maximisation) algorithm (Jones et al., 2016), a naïve Bayesian classifier. Traditionally, it is used to call somatic variants in one sample, usually tumour, by using another sample from the same patient as a matched normal. However, to preserve the early mutations that will be present in normal bulk samples, I used an unmatched normal sample. This alignment file was created *in silico* by generating reads from the human reference genome (GRCh37). In effect, this causes all nucleotide deviations with respect to the reference genome to be reported, including a large amount of inherited single nucleotide polymorphisms (SNPs). However, CaVEMan automatically employs its flags and filters when calling variants, even in an unmatched setting. One of such filters is the exclusion of a putative SNV that is present in a large panel of normal samples. This excludes most of the germline SNPs from subsequent analysis and brings their total number down from approximately 4-5 million (1000 Genomes Project Consortium, 2015) to 30,000-40,000 in most unmatched runs.[1] The latter represents rare inherited SNPs. In addition to the default CaVEMan filters, putative SNVs were forced to have a mean mapping score (ASMD) of at least 140 and fewer than half supporting reads being clipped (CLPM=0). The same approach was taken for calling indels by the Pindel algorithm (Ye et al., 2018b). Pindel then applies its own post-calling filters, such as requiring a variant to be present in both strand orientations and a minimum mappability score. In addition, for putative indel calling, I enforced a minimum quality score of 300.

Where whole-genome sequencing data was obtained from LCM samples through the low-input pipeline, an additional set of filters was used to remove specific artefacts. These artefacts include spurious variant calls due to the formation of cruciform DNA and the double counting of variants due to a negative insert size. These filters were developed by Dr Mathijs Sanders and can be found on GitHub (https://github.com/MathijsSanders/SangerLCMFiltering).

Copy number variants (CNVs) were called using the ASCAT (Allele-Specific Copy number Analysis of Tumours) algorithm (Van Loo et al., 2010) and the Battenberg algorithm (Nik-Zainal et al., 2012). Where parental genomes were available, the standard pipeline of Battenberg was modified by only including sites where SNPs could unequivocally be assigned to either parent. This allows for full phasing of the chromosome (further increasing the sensitivity of the algorithm) and assignment of any CNV to one of two parental alleles. Since both ASCAT and Battenberg rely on the sharedness of SNPs, they cannot be run against the *in silico*-generated unmatched normal sample. To detect potential early embryonic CNVs, both ASCAT and Battenberg were run on all LCM samples against bulk samples (e.g. brain (Chapter 3), blood (Chapter 4) or umbilical cord (Chapter 5) and alternatively, against an LCM sample known to be derived from the other branch of the first split in the phylogeny, and hence genetically unrelated on a post-zygotic level. No early embryonic CNVs were present in any of the phylogenies throughout this thesis.

To detect structural variants (SVs), I used the BRASS (BReakpoint AnalySiS) algorithm (Nik-Zainal et al., 2016), which relies on discovery through discordantly mapped reads and confirmation by local reassembly of the breakpoints of the SV. The strategy for matching samples was identical to the strategy described for ASCAT and Battenberg. Again, no early SVs were identified in any of the phylogenies, with the exception of the phylogenies of trophoblast clusters described in Chapter 5.

[1] These numbers can vary substantially between different patients. One cause of this is genetic proximity to the human reference genome and the unmatched normal panel employed by CaVEMan and hence, the ethnicity of the patient.

2.2.1 Filtering germline variants

Because the variant calling is performed without a matched sample, germline variants will still be present in the output from CaVEMan. Fortunately, in the studies presented throughout this dissertation, multiple, different normal samples from the same patient have been sequenced. Hence, the aggregated depth of coverage across samples from one patient is rather high and usually exceeds 100x. Here, the distinction in body-wide VAF between germline variants (0.5) and early post-zygotic variants (mostly <0.4) allows us to distinguish between the two.

Given that inherited variants are expected to be present at least at a heterozygous level in all cells, a one-sided exact binomial test can be used on the aggregated counts of reads supporting the variant and the total depth at that site. Basically, this tests whether the observed variant counts are likely to have come from a germline distribution (given the total depth), or whether it is more probable to come from a distribution with a lower true VAF. For sex chromosomes in male patients, the binomial probability (true VAF) for comparison was set to 0.95 rather than 0.5. This results in a probability per variant whether it is inherited. I correct these values for multiple testing using a Benjamini-Hochberg correction (Benjamini and Hochberg, 1995). Any variant with a corrected p-value less than 10^{-5} is categorised as a putative somatic variant.[2]

The underlying assumption that all germline variants will at least be heterozygously present in all cells of the body will be violated if a sizeable proportion of the samples considered has a copy number variation. Therefore, it is best practice to exclude at least the copy number altered regions in the affected samples from consideration. This has been applied to all variant calling presented in this dissertation.

2.2.2 Filtering shared artefacts

To filter out recurrent artefacts, I fit a beta-binomial distribution to the number of variant supporting reads and total number of reads across samples from the same patient. For every somatic SNV, I determine the maximum likelihood over-dispersion parameter (ρ) in a grid-based way (ranging the value of ρ from 10^{-6} to $10^{-0.05}$). A low over-dispersion captures artefactual variants as they appear seemingly randomly across samples and can be modelled as drawn from a binomial distribution. In contrast, true somatic variants will be present at a high VAF in some, but not all genomes, and are thus best represented by a beta-binomial distribution with a high over-dispersion. To distinguish artefacts from true variants, I use $\rho = 0.1$ as a threshold, below which variants were considered to be artefacts. The code

[2] This threshold corresponds to only allowing a few true germline variants to falsely pass this filter, as their number will be in the order of 10^5.

for this filtering approach is an adaptation of the Shearwater variant caller (Gerstung et al., 2014).

While the number of variants that are filtered out in this step is usually rather modest (in most cases this will be fewer than a 1,000 mutations), it greatly enhances the ability to reconstruct genuine phylogenies of the samples. This over-dispersion filter will remove any inherited mutations that have falsely passed the filters so far (as they will be completely under-dispersed) and also removes recurrent artefacts that might wrongly be interpreted as mutations occurring early in development.

2.2.3 Distinguishing true presence from sequencing noise

In chapters 4 and 5, in cases of low frequency support for SNVs, it necessary to distinguish true presence of these mutations in the samples from support due to sequencing noise. The low VAF is due the variant originally being called in one sample (e.g. tumour) but with some support for these variants in bulk samples, which usually represent large polyclonal aggregates of cells. As with the artefact filter presented above, the distinction of true somatic variants from base-specific errors was done using a beta-binomial model derived from the Shearwater variant caller (Gerstung et al., 2014). The crucial difference, however, is that in this case, a large set of normal samples (unrelated to the patient in question) is used to obtain a locus-specific error rate. Per variant, I then calculated the probability of the observed supporting read counts being drawn from a beta-binomial distribution with that error rate. These p-values were corrected for multiple testing using the Benjamini-Hochberg correction (Benjamini and Hochberg, 1995). If this corrected p-value was less than 0.001, the observed support for the variant was very unlikely to be the consequence of site-specific noise, such as due to sequencing errors. This particular threshold was chosen to minimise the false positive rate. Subsequently, all variants were visually inspected using the genome browser Jbrowse (Buels et al., 2016) to exclude further sequencing or mapping artefacts.

2.3 Binomial mixture models

There are many situations when the somatic variants called in a certain sample consist of different populations that are mixed together. In order to be distinguishable by this mixture model, these different populations of variants should reflect a difference in VAF, i.e. the probability of finding a mutant read during sequencing.

Mixture models will try to cluster observations together according to their underlying probability. The manifestation of somatic mutations in sequencing data can ultimately be seen as a binomial draw. Successes reflect seeing a read containing the variant, the number of trials correspond to the total depth at a site, and the underlying binomial probability is the true VAF of the mutation. Therefore, I have used binomial mixture models as an effective tool for different purposes throughout this dissertation. Another common choice for the base distribution in a mixture model is the Dirichlet process (Brunner et al., 2019; Nik-Zainal et al., 2012).

For each cluster, the optimal binomial probability and mixing proportion is estimated using an expectation-maximisation algorithm.

Regardless of the underlying probability distribution, the mixture model will need to know *a priori* into how many clusters to divide the data. Most often, the exact number of detectable clusters or clones within a sample will be unknown. Whole-genome sequencing data at a typical depth of 30x will not allow for the detection of variants with very low VAF, therefore the number of detectable clones within one sample is low, typically ranging from one to five [3]. Therefore, a range of cluster numbers to be considered can be passed to the mixture model.

The mixture model can be run with all these parameters and afterwards select the most optimal fit using either the Akaike information criterion (AIC) or the Bayesian information criterion (BIC). Both measure the likelihood of the model but penalise the addition of clones with a different parameter. In the case of the BIC, the penalisation depends on the number of observations and will effectively avoid over-splitting the data when using large number of mutations, which is why I have used it as a preferred choice.

2.3.1 Truncated binomial distributions

It is often the case with variant calling algorithms that there is a threshold for the number of reads supporting a variant before it is taken into consideration. This is mostly because of practical reasons: variants with low support will be abundant in the genome due to sequencing errors and noise in alignments leading to an increased run time of the algorithm whole increasing the rate of artefacts. For CaVEMan, the minimum support for a variant is hard-coded to be four reads.

This minimum number of supporting reads poses a problem when using a binomial mixture model on variants from a sample. In effect, part of the full binomial model will be censored, as observations from the lower tail of the distribution are disallowed. This will lead to an erroneous fitting of distributions, especially for clusters with a binomial probability

[3] The term 'clone' can be confusing, as every single cell within one sample can be thought of as a 'clone'. Within this dissertation, the word 'clone' will denote a large group of cells with a set of shared mutations (detectable by whole-genome sequencing), a consequence of normal development or aberrant clonal expansions.

approaching the edge of the censored distribution. The need for adjusted binomials also increases with lower depth of coverage, where the minimum threshold for support will more often be imposed on genuine variants. To fix this, I have used truncated binomial distributions instead of full ones, where appropriate. In effect, this curtails the entire binomial probability distribution below a set threshold (here taken to be four reads), after which the remaining distribution is re-normalised.

2.3.2 Estimating clonality

A mixture model can be applied when multiple different clones can inhabit the same sample, especially if the sample was a small section of tissue without much histological structure, such as skin or oesophageal epithelium. In this case, the model will try to separate the overall variants into the clones they could have arisen from, each with their own probability (VAF) and proportion (the amount of variants a clone contributes). Note that the latter proportion is the proportion of variants explainable by a clone, not the proportion of cells in a sample inhabiting that clone. The latter measure can be obtained by multiplying the VAF of a clone by two and is extremely useful to determine the clonal origin of any sample and whether two clones are nested or parallel. For example, clone A has an estimated VAF of 0.45 and clone B 0.15, so that the proportion of cells belonging to clone A is 90% and for clone B 30%. Simply because the sum of these proportions would exceed the total of the sample (120% is impossible), cells belonging to clone B must also belong to clone A. In other words, clone B must be a subclone of clone A. This logic is generally referred to as the pigeon hole principle.

Determining the largest clone in a given sample is straightforward with this framework. To reconstruct phylogenies we must be sure that our set of variants represent those coming from distinct single-cell derived clones and not mixtures that could be parallel, i.e. sharing no genetic ancestry. As soon as any clone accounts for more than half of the cells in a sample, we can be sure that this signal is coming from a single ancestor, since two overlapping clones at 51% cannot co-exist. However, theoretically, as soon as a clone only accounts for (less than) half, it becomes possible for two clones to overlap, confuse the signal, and violate the assumptions of the phylogeny reconstruction. Therefore, only samples containing a clone with an estimated peak VAF higher than 0.25 are used for tree building.

In practice, many samples will have a degree of contamination from a polyclonal source, such as stroma. In such cases, the total clonal reconstruction will add up to less than 100%, because variants belonging to the polyclonal contaminant will have a true VAF of far below the detection limit in a typical whole-genome sequencing experiment. This contamination decreases the VAF of the largest clone and can therefore affect the estimated clonality of the sample. For example, with 20% stromal contamination, a clone comprising 60% of cells of interest will only account for 48% of the total sample, and hence, will not make the cut to be included in initial phylogeny reconstruction.

2.3.3 Estimating tumour contamination in normal samples

The mixture model can also be applied when investigating tumour variants in normal tissues, such as employed in Chapter 4. Variants can be shared between normal, bulk samples and tumours due to a shared development, but also due to a contamination of cancer cells in the normal sample. These two scenarios can be distinguished from one another by a closer investigation of the clonal structure of such shared variants.

In this case, the variants under consideration are restricted to those that are clonal in the tumour. Most likely, if the normal sample, usually blood or bulk kidney in the context of this thesis, contained a degree of contamination from the tumour, all of these clonal tumour variants would present with the same, underlying VAF. This VAF is equal to half the contamination rate, similar to the clone size estimates in the previous section. However, any variants truly present in normal cells due to the consequences of development will be present at VAFs higher than the background contamination rate.

The mixture model will try to separate the tumour variants in the normal sample into distinct clusters as before. The clone or cluster with the lowest VAF must correspond to an upper bound of the background infiltration of tumour cells. Therefore, excluding this cluster will effectively exclude tumour contamination. Any variants remaining likely reflect shared early development or aberrant pre-malignant clonal expansions in normal tissues.

2.3.4 Timing of copy number gains

A further use of the mixture model is to time chromosomal duplication events. To time the occurrence of such copy number gains, it is necessary to know the ratio between somatic variants that happened before and after the event. Naturally, the accuracy of such a ratio depends on the abundance of variants within the region of duplication and as such, works best when employed on clonal samples, most often tumours.

Generally, somatic variants in gained regions will fall into two groups: those having occurred on the duplicated chromosome prior to the event (e.g. present on two out of three copies in the clone), and those either acquired on the non-duplicated chromosome before the event or on any afterwards (e.g. present on one of three copies). A third group of variants can be present if the clone has a considerable amount of subclonal diversification.

In cases of simple duplication (2:1) or copy number-neutral loss of heterozygosity (2:0), these events can be timed using the following relation:

$$T = \frac{C_M + C_P}{max(C_M, C_P) + P_{ND}/P_D} \tag{2.1}$$

Where C_M and C_P are the maternal and paternal copy number, respectively, and P_D and P_{ND} the proportion of mutations assigned to the duplicated or non-duplicated copy number. The time point of duplication (T) will be a number between 0 and 1, the former representing the zygote and the latter the most recent common tumour ancestor.

The proportion of duplicated (P_D) and non-duplicated (P_{ND}) mutations can be estimated using the mixture model on the variant supporting counts and total counts in the region of the gain. The binomial probabilities of the components are then compared to the expected VAFs resulting from the different copy number states. The expected VAFs must be corrected for the purity of the tumour sample. If the estimated binomial probability of a cluster was sufficiently close to the estimated purity-corrected VAF, the proportion of that cluster was taken to be P_D or P_{ND}. The usual threshold for acceptance is 0.05.

The uncertainty around the point estimate of the time of duplication will depend on the number of mutations that fall within the region. Confidence intervals around that estimate are therefore most easily obtained using a Poisson test on the rounded duplicated and non-duplicated mutation counts. These are simply obtained by multiplying the estimated proportion of duplicated and non-duplicated mutations with the total number of mutations in the region.

To discern the likelihood of two gains occurring at the same time, I used the Poisson test again to compare the sets of duplicated and non-duplicated counts from two different copy number gains in the same patient.

2.4 Phylogenetic tree reconstruction

Many different algorithms have been developed to reconstruct phylogenetic trees based on DNA sequences. These character-based algorithms rely on different approaches: maximum parsimony, maximum likelihood, or Bayesian inference (Yang and Rannala, 2012). Maximum parsimony-based algorithms seek to produce a phylogeny that requires the least amount of discrete changes on the tree. Because of this minimisation of nucleotide changes, this approach implicitly assumes that mutations are likely to occur only once. Hence, maximum parsimony will produce erroneous phylogenies when there is a high likelihood of recurrent or reversal mutations, such as with long divergence times or high mutation rates (Swofford et al., 2001). Phylogenetic tree algorithms relying on maximum likelihood or Bayesian inference are model-based, i.e. they require a specific notion of the parameters governing genetic sequence evolution to calculate either distances or likelihoods. Oftentimes, this involves a general time-reversible model of sequence evolution (Tavaré, 1986). The maximum likelihood approach will attempt to identify the phylogeny with the highest log-likelihood of the data given the underlying model of sequence evolution, while Bayesian inference methods will seek the posterior distribution with the highest probability. All these approaches have been widely applied to the reconstruction of phylogenetic trees between species or individuals (Yang and Rannala, 2012).

However, the task of constructing a phylogeny of somatic cells derived from a single individual is fundamentally different from reconstructing species trees in three ways: (1) precise reconstruction of the ancestral state, (2) the lack of time-reversibility of mutation rates and (3) the low number of mutations compared to the size of the genome.

(1) In contrast to the unknown ancestral genetic state of multiple species, the ancestral DNA sequence at root of the tree (i.e. the zygote) can readily be inferred from the data. Since all cells in the body are derived from the fertilised egg cell, any post-zygotic mutation will be present in a subset, and not all leaves of tree. Hence, the genetic sequence at the root of the tree is defined by the absence of all of these mutations, which is in essence the nucleotide sequence of the reference genome for the variant sites. This simple observation effectively anchors the phylogeny in time.

(2) Somatic mutation rates of specific nucleotide changes are not time-reversible. In order to accommodate the uncertainty in the ancestral state and the direction of nucleotide substitutions, model-based phylogeny reconstruction has relied on a time-reversible model of nucleotide changes (Tavaré, 1986). In principle, this states that the probability of a certain substitution (e.g. C>T) is equal to its inverse (T>C). In somatic mutagenesis, since the direction of change is known, assuming general reversibility of mutational probabilities fails to acknowledge the genuine discrepancies in the likelihood of certain (trinucleotide) substitutions. For example, a C>T mutation in a CpG context is much more probable than a T>C at TpG due to the specific mutational processes acting on the genome, in this case, spontaneous deamination of methylated cytosine (signature 1).

(3) When taking into account the size of the human genome, the number of mutations that are informative for purposes of phylogeny reconstruction, i.e. SNVs shared between two or more samples, is relatively low compared to the settings of phylogenies of species or individuals within a species. As mentioned in the previous chapter, the mutation rate can be considered low enough that it is possible to assume an infinite sites model. In short, this

means that the odds of recurrent mutations at the same site or reversals of those nucleotide changes ("back mutations") are vanishingly small. Because of this, a mutation shared between multiple samples can generally be assumed to represent a single event in an ancestral cell that has been retained in all its progeny. Exceptions to this assumption are discussed in more detail in Chapter 3.

Because of the reasons outlined above, out-of-the-box, model-based phylogeny reconstruction algorithms relying on a maximum likelihood approach, such as RAxML (Stamatakis, 2014), or a Bayesian approach, such as BEAST (Bouckaert et al., 2014), did not perform well in the early exploratory phase of this PhD. Instead, these algorithms were consistently outperformed by maximum parsimony-based algorithms, such as PHYLIP (Felsenstein, 1993), PAUP (Swofford, 2001), Phangorn (Schliep, 2011) and MPBoot (Hoang et al., 2018). Out of the latter group of algorithms, MPBoot performed the best due to its quick run-time and hence, I have used it as the basis for phylogenetic tree reconstruction in this dissertation and other recent studies (Lawson et al., 2020; Moore et al., 2020; Olafsson et al., 2020; Robinson et al., 2020; Yoshida et al., 2020).

To create the phylogenies presented in this dissertation, I ran MPBoot with default parameters (1,000 bootstrap iterations) on concatenated nucleotide sequences of variant sites. I use the VAF to inform the state of the mutations: '1' (present) for VAF ≥ 0.3, '0' (absent) for VAF ≤ 0.1 and '?' (unknown) for 0.1<VAF<0.3 to allow uncertainty due to noise in the VAF. To ensure the proper anchoring of the root of the phylogeny, I included an artificial nucleotide sequence composed of the reference genome bases at variant sites to simulate the ancestral state at the zygote. This outputs a tree topology, i.e. the structure of bifurcations between samples, but does not inform which mutations reside on which branch of the tree.

A considerable source of uncertainty in our data is introduced by the depth of sequencing, which can be variable across samples and variant sites. In using the VAF to determine the character state of a variant site, the information of the number of counts supporting the variant and the read depth at the variant loci is lost. Therefore, the mapping of mutations to branches is done by calculating likelihood of the observed read counts of a mutated site for all tree branches, given a binomial probability of 0.5 if the mutation is present and 0 if absent. Rather than mutations being 'soft' assigned with a probability on each branch, mutations are 'hard' assigned to the most likely branch. This approach is implemented in the 'treemut' R package. In addition, this method calculates a probability of the variant not adhering to the tree structure at all, which can be used as a basis for filtering variants further. I applied a cut-off of 0.01 to include variants in the phylogeny. SNVs that do not adhere to the tree topology can be investigated further in case they represent genuine recurrent mutations. Usually, they represent artefactual variants that wrongly passed the filters. Because MPBoot only outputs bifurcations and not multifurcations, any branch with a length of 0 after SNV fitting is collapsed into a polytomy.

In addition to the algorithms described above, I used the ape (Paradis et al., 2004) and ggtree (Yu et al., 2017) packages for analysis and visualisation of phylogenetic trees in R.

2.5 Miscellaneous methods

2.5.1 Methods pertaining to chapter 3

Estimating asymmetry and likelihood ratio test

The number of variant supporting reads (NV) and total reads (NR) for mutations on the major are modelled with a binomial distribution, with underlying probability p, while those on the minor branch have probability $1-p$. The likelihood of the model is then given as the binomial probability of seeing NV successes out of NR trials, given p (or $1-p$). A grid-based approach was used to calculate the maximum likelihood estimates of the binomial probabilities in each scenario. Confidence intervals around this estimate were calculated based on the profile likelihood.

We used a likelihood ratio test to calculate whether two bulk samples have a significantly different asymmetry. The null model is that the underlying binomial probability (asymmetry) is the same in the two bulk samples ($p1 = p2$; joint maximum likelihood estimate for the samples), while the alternative model is that these are different ($p1 \neq p2$; separate maximum likelihood estimate for each sample).

Mutation rate in early embryogenesis

The calculation of the mutation rate from the phylogenetic trees presented in Chapter 3 relies on the mutations per branch following a Poisson distribution, with the mutation rate itself as its only parameter (λ). The probability of observing a certain number of mutations in a branch (k) is then given by:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{2.2}$$

The rate parameter in a Poisson distribution is also the mean and the variance of the distribution. Given the absence of polytomies in the early branching events of the tree, the mutation rate can be calculated as the total number of mutations observed in the first two generations divided by the number of branches in those generations, the latter of which is equal to 6.

This results in the estimates per patient displayed in Table 2.1, with estimated 95% confidence intervals (CI).

	Number of SNVs	Rates	CI_{lower}	CI_{upper}
PD28690	14	2.3	1.3	3.9
PD43850	10	1.6	0.8	3.1
PD43851	19	3.2	1.9	4.9
Combined	43	2.4	1.7	3.2

Table 2.1 Mutation rate estimates in the first two cell generations

For the subsequent generations, where the occurrence of polytomies becomes so frequent, it is more straightforward to estimate the mutation rate by translating the observed number of branches involved in a multifurcation into the necessary frequency of branches with no mutations.[4]. Then, the following relation will hold:

$$P(X=0) = \frac{\lambda^0 e^{-\lambda}}{0!} = e^{-\lambda} \rightarrow \lambda = -\ln\left(\frac{n_{missed\_divisions}}{n_{total\_branches}}\right) \tag{2.3}$$

Following this, the calculations for the mutation rates of the third and fourth observed generation are displayed in Table 2.2.

	Number of missed divisions	Number of branches	Rates	CI_{lower}	CI_{upper}
PD28690	42	68	0.48	0.32	0.65
PD43850	16	42	0.97	0.67	1.26
PD43851	20	38	0.64	0.39	0.90
Combined	78	148	0.64	0.51	0.77

Table 2.2 Mutation rate estimates in subsequent cell generations

Targeted sequencing of embryonic variants

Following the identification of early emrbyonic variants in PD28690, a custom bait set for the Agilent SureSelectXT platform was designed using Agilent's online tool. Repetitive regions were masked using the criterion of least stringency, as defined by Agilent. Each embryonic variant was covered by one tile. In addition to early variants, testis-specific somatic mutations, as well as heterozygous SNPs were included in the bait set. The former was included to assess the split between the germline and the somatic tissues, while the latter was used to

[4] The number of missed divisions is simply the number of branches involved in a polytomy minus two. E.g. a trifurcation is the result of a single missed division; a four-way polytomy of two missed divisions and so on.

assess any biases in the performance of the pulldown. DNA libraries from 86 bulk samples were made and subsequently hybridised with the baits. The sequencing was performed on the Illumina HiSeq 2500 platform.

Soft cosine similarity

To compare the similarity between two vectors of VAFs of variants positioned on a phylogenetic tree, I calculated a soft cosine similarity, which includes a specific similarity term to incorporate the dependence of observations. That is, variants on the same branch of the phylogeny will convey the same information, while variants of different branches of the tree will provide parallel information on lineage composition. As such, an interaction term $s_{i,j}$ is worked into the following definition of the soft cosine similarity:

$$soft\_cosine(a,b) = \frac{\sum_{i,j}^{N} s_{ij} a_i b_j}{\sum_{i,j}^{N} s_{ij} a_i a_j \sum_{i,j}^{N} s_{ij} b_i b_j} \tag{2.4}$$

In this relation, a and b refer to two bulk samples and their vector of VAFs of embryonic mutations of length N, which are indexed by i and j. These VAFs are log-transformed. The similarity metric $s_{i,j}$ is defined as follows:

$$s_{i,j} \begin{cases} 1, & \text{if } br_i = br_j. \\ 0, & \text{if } br_i \neq br_j. \end{cases} \tag{2.5}$$

Note that if the similarity metric $s_{i,j}$ is 1 for $i = j$ and 0 otherwise, the formulation of the regular cosine similarity is retrieved.

To use the matrix of soft cosine similarities as a basis for clustering, I converted it to a soft cosine distance matrix by subtraction from 1 (i.e. if the soft cosine similarity between a and b is 0.7, the distance will be 0.3) . Hierarchical clustering was then performed in R using the "complete" method.

2.5.2 Methods pertaining to chapter 4

Methylation arrays and analysis

Data on genome-wide methylation status was obtained using the Illumina Infinium MethylationEPIC BeadChip microarray kit. Samples were selected based on the availability of DNA. Data was processed in R using the minfi package (Fortin et al., 2017). Comparisons were made based on the beta score, which is the ratio of intensities between methylated and unmethylated alleles.

RNA sequencing

RNA libraries were sequenced on the Illumina HiSeq 4000 platform. Reads were aligned using STAR (Dobin et al., 2013) and read counts of genes were obtained using HTSeq (Anders et al., 2015). Differential expression analysis was then performed in R using the DESeq2 package (Love et al., 2014). No significant differential expression between kidney samples with and without clonal nephrogenesis was identified.

Whole-exome sequencing

15 out of 17 normal tissues found to contain clonal nephrogenesis were submitted for further whole-exome sequencing on the Illumina HiSeq 4000 platform. Interrogation of identified coding SNVs in tumours (see table S3 and table S4), and testing of presence or absence of these variants using a site-specific error rate as previously described, was undertaken. No support for a driver mutation in a tumour was found in the corresponding normal.

Interrogation of coding indels yielded one mutant read of a driver (in-frame insertion in *MLLT1*) in PD40713c. However, as this read contained additional sequence variation (that was absent from corresponding mutant tumour reads) and could not be validated in cDNA reads, I considered it to be artefactual.

2.5.3 Methods pertaining to chapter 5

Exclusion of maternal contamination

To exclude the possibility of any remaining maternal DNA in the placenta to skew results on mutation burden and clonality, I used maternal SNPs to quantify contamination. For each pregnancy, I randomly picked 5,000 rare germline variants (i.e. left in by the common SNP filter in CaVEMan) found in mother but not in umbilical cord. All these variants passed other CaVEMan flags, did not fall in regions of low depth (on average, below 35), and were present at a VAF greater than 0.35 in mother. Their VAFs in all individual placental samples, microdissections and biopsies, is displayed in Extended Data Fig.1. No sample had a level of support for maternal SNPs that exceeded the expectations for sequencing noise (0.1%), excluding maternal contamination as a plausible origin for any observations made here.

Mutational signature extraction and fitting

To identify possibly undiscovered mutational signatures in human placenta, I ran the hierarchical Dirichlet process (HDP, see https://github.com/nicolaroberts/hdp) on the 96 trinucleotide counts of all microdissected samples, divided into individual branches. To avoid over-fitting, branches with fewer than 50 mutations were not included in the signature extraction. HDP was run with individual patients as the hierarchy, in twenty independent chains, for 40,000 iterations, with a burn-in of 20,000.

Besides the usual flat noise signature (Component 0) that is usually extracted, only one other signature emerged (Component 1) from the signature extraction. Deconvolution of that signature revealed it could be fully explained by a combination of reference signatures 1, 5, and 18, all of which have been previously reported in normal tissues. Because of the lack of novel signatures in this data set, the remainder of mutational signature analysis was performed by fitting this set of three signatures to trinucleotide counts using the R package deconstructSigs (v1.8.0) (Rosenthal et al., 2016).

Sensitivity correction of mutation burden

To compensate for the effects of sequencing coverage and low clonality on the final mutation burden per sample, I estimated the sensitivity of variant calling. For each sample, I generated an *in silico* coverage distribution by drawing 100,000 times from a Poisson distribution with the observed median coverage of the sample as its parameter. For each coverage simulation, I calculated the probability of observing at least four mutant reads for SNVs or five for indels (the minimum depth requirement for our CaVEMan and Pindel calls respectively) with the underlying binomial probability given by the observed median VAF of the sample. The average of all these probabilities then represents the sensitivity of variant calling. Final mutation burdens were then obtained by dividing the observed number of mutations by the estimated sensitivity.

Genetic proximity scores

To measure the genetic proximity between any two trophoblast clusters or mesenchymal cores from the same biopsy, I used the following equation:

$$sim_{i,j} = \frac{mut_{sharedi,j}}{(mut_{tot,i} + mut_{tot,j})/2} \tag{2.6}$$

Or simply, the fraction of shared mutations between samples i and j divided by their average total mutation burden. The resulting number reflects how much of in utero development was shared between these samples.

However, control data of normal human colon (Lee-Six et al., 2019) and endometrium (Moore et al., 2020) were obtained from adults and their phylogenetic histories will reflect postnatal tissue dynamics as well. To obtain a proxy for the sharedness due to development *in utero*, I only considered a pair of samples i and j, if they did not split at a mutational time inconsistent with early development. I set this threshold for both colon and endometrium at 100 mutations, a very rough estimate of the maximum burden at birth in these tissues given preliminary studies. Consequently, instead of dividing the number of early shared mutations by the average burden, for adult tissues, these were divided by 100. The latter number was taken as an approximation of the maximum burden of endometrial and colonic cells at birth, assuming a higher mutation rate *in utero* than during adult life (Kuijk et al., 2019).

Chapter 3

Extensive phylogenies of normal human development

3.1 Introduction

The somatic mutations present in a cell encode its life history and developmental path from the fertilised egg cell onwards. Recent advances in next-generation sequencing now allow for the retrospective reconstruction of human ontogeny. In this chapter, this principle is used to reconstruct extensive phylogenetic trees of human development by whole-genome sequencing a large number of cells derived from 24 different tissues of three individuals. For the first time, the phylogeny will be constructed from human samples from multiple different organs and germ layers, and is supplemented by a large number of DNA sequencing data of bulk biopsies. From these phylogenies and the patterns of mutations across biopsies, we can then deduce cell dynamics in the early embryo, such as the asymmetric contribution of the first lineages and the spatial pattern of embryonic mosaicism. In addition, we can detect later clonal expansions in adult life, possibly driven by canonical cancer driver mutations. Furthermore, the construction of the phylogenies allows for the direct assessment of the validity of the infinite sites assumption and evaluation of any mutations in violation of the tree topology, such as recurrent somatic mutations.

3.2 Experimental design

Samples studied in this chapter were obtained from three individuals, all subjected to rapid autopsies. The cohort consists of a 78-year old male who died of oesophageal adenocarcinoma (PD28690), a 54-year old female, who died as a result of traumatic injuries (PD43850) and a 47-year old male, who succumbed to acute coronary syndrome (PD43851). Biopsies from PD28690 were collected within six hours of death, while samples from PD43850 and PD43851 were taken within ten hours. Since somatic mutations are stable as soon as they are incorporated into the genome, the age of the patient should not compromise our ability to reconstruct early developmental patterns. The same is true for the specific cause of death, as long as the samples sequenced are not infiltrated by metastatic tumour cells. Large clonal expansions later in life might distort the developmental history of a tissue on a microscopic level, but these will bear distinct mutational markers and thus be detectable by the same means used to reconstruct early development.

To obtain genomic readouts suitable for phylogeny reconstruction while simultaneously retaining the spatial information of individual samples on a microscopic level, we used laser capture microdissection (LCM) as a basis for the study presented here. These LCM cuts consist of a few hundred cells each and therefore provide a much more refined genomic readout than a conventional bulk sample, especially if these cells are derived from a small set of stem cells. To accommodate the low yield of DNA from an LCM cut, we used a novel method of low-input DNA library preparation developed at the Wellcome Sanger Institute (Brunner et al., 2019; Lee-Six et al., 2019; Moore et al., 2020; Olafsson et al., 2020; Robinson et al., 2020). In total, we used this low-input method of whole-genome sequencing on a total of 393 LCM cuts from 25 different tissues in PD28690, 67 from 12 tissues in PD43850 and 110 from 12 tissues in PD43851.

These samples then form the basis for the reconstruction of phylogenies, which allow us to assess the asymmetric contribution of embryonic progenitors, the most recent common ancestors of individual tissues and the spatial patterning of developmental mosaicism on a histological level. The embryonic variants then form a basis for targeted re-sequencing in a variety of large polyclonal aggregates of cells, which resolves organ-level heterogeneity in developmental paths. In addition, the continuous accumulation of somatic mutations makes it possible to discover later microscopic and macroscopic clonal expansions, such as neoplastic polyp formation and benign prostatic hyperplasia in PD28690. Lastly, from the phylogenies we can deduce any mutations that do not adhere to the tree structure, either by selective recurrence, chance or other biological processes. Taken together, the research presented in this chapter highlight the power of next-generation sequencing in resolving questions of human development and displays the potential of lineage tracing using somatic mutations.[1]

[1] The exploration of the data in terms of their mutational landscape itself, e.g. burden and exposure to mutational signatures, is part of another research project and hence will only feature in a very minor way in this dissertation.

3.3 The clonality of LCM samples

Excision by LCM is performed on histology sections, which display the tissue as close as possible to its state *in vivo*. Tissue architecture varies dramatically across different organs. Hence, LCM dissection strategy depends on the tissue and its natural histological architecture. For example, in prostate biopsies, the small terminal glands constitute the smallest tissue unit amenable to LCM with an equivalent role for glomeruli in the kidney. Other tissues consist of networks of tubules and a small cross-section of one would be considered the fundamental unit in that tissue, such as seminiferous tubules in testes or proximal and distal tubules in the kidney. However, some tissues, such as epithelial sheets in skin and oesophagus, lack distinct features delineating these units. In these cases, a small strip of the epithelial sheet was subjected to microdissection.

In order to facilitate phylogenetic reconstruction, LCM biopsies should ideally represent monoclonal cell populations derived from a single cell. The clonal organisation of most of the tissue units sequenced here was not known prior to this experiment. However, the distribution of the variant allele frequencies (VAFs) of somatic mutations readily reflects the clonality of the cell population (**Fig. 3.1a**). The sequencing of a pure, monoclonal population of cells results in a VAF distribution centred around 0.5 (**Fig. 3.1b**), while a polyclonal aggregate of cells would exhibit VAFs close to the detection limit, generally lower than 0.1 (**Fig. 3.1d**). Oligoclonal samples behave in between these two extremes, and can exhibit VAF distributions corresponding to distinct clones within each sample (**Fig. 3.1c**). To reliably build phylogenetic trees from this data, LCM cuts must exhibit a clone at a VAF of 0.25 or higher, corresponding to a majority of cells within that sample.[2] In this way, it is safe to assume this must be the mutational legacy of a single cell. After all, there cannot be two distinct clones accounting for a majority.

To evaluate clonality in all LCM samples from the three patients, I called single nucleotide variants (SNVs) against a polyclonal bulk sample obtained from brain. This approach would obscure any mutations that are truly shared between most samples, i.e. the mutations obtained in the first few divisions of life. The strategy for identifying early embryonic mutations is described in the subsequent sections. In the first instance, a direct matched analysis enables assessment of the clonality of individual samples and provides a criterion for the inclusion of samples into the final set. This set of mutations was then subjected to a binomial mixture model to decompose the VAF landscape into individual clones. This is described in more detail in the Methods chapter.

[2] The VAF of a mutation is a direct readout of the proportion of cells carrying that mutation, which is obtained by multiplying the VAF by the ploidy, usually 2. Hence, a VAF of 0.25 corresponds to 50% of cells.

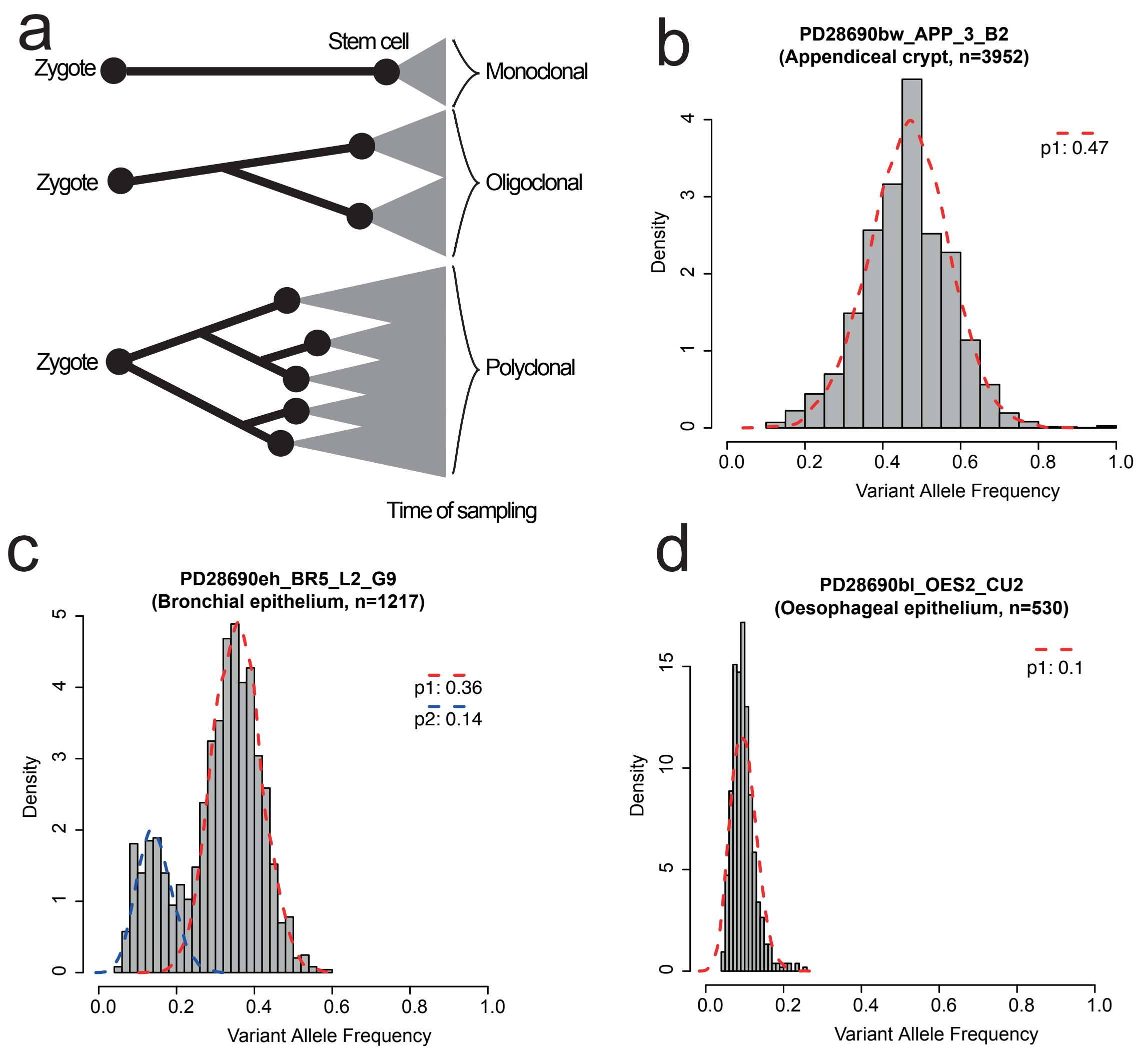

Figure 3.1 (a) Schematic of three different progenitor or stem cell contributions to the eventual sample. Monoclonal samples consist of the progeny of one cell, while oligo- and polyclonal are derived from a few and many progenitors, respectively. VAF histograms and binomial decompositions for a monoclonal (b), oligoclonal (c) and polyclonal (d) sample. Red and blue dashed lines indicate clonal decomposition through a binomial mixture model, with the estimated peak VAF of clones indicated in the legend. The number indicated in the title of each histogram is the SNV burden.

In total, the clonality and depth of sequencing[3] was sufficiently high for 187 samples in PD28690, 62 samples in PD43850, and 106 for PD43851. Some tissue structures consistently yielded clonal samples, such as intestinal crypts, prostate glands, and seminiferous tubules. For other structures, notably strips of epithelium in skin and oesophagus, the clonality was of

[3] A minimum mean depth of 10x.

a stochastic nature and differed wildly between cuts. Other tissues, such as visceral fat, never yielded a sample of sufficient clonality to warrant inclusion.

3.4 Reconstruction of phylogenies

The remaining clonal samples were used as a basis for the construction of phylogenetic trees. However, a classic matched variant calling approach, for example when calling mutations in tumours by comparing to blood, will obscure some of the early embryonic mutations by virtue of their presence even in polyclonal aggregates of cells. Therefore, it is necessary to perform the variant calling in an unmatched fashion. The pipeline I devised for this is described in more detail in Chapter 2. Briefly, SNVs were called against an *in silico* generated alignment file based on the human reference genome. In this way, both germline and somatic variants are called. For PD28690, a total of 349,386 SNVs were called. To separate germline and somatic variants, I use an exact binomial test on the sum of the site depth and variant counts across all samples from the same patient. By pooling all sample counts, the aggregate depth allows for precise discrimination between germline variants (distributed tightly around 0.5) and early embryonic variants at a potentially high VAF (e.g. 0.4-0.45). For PD28690, 35,934 variants were classified as germline and filtered out (10.3% of called variants).

Subsequently, a filter based on the beta-binomial distribution enabled elimination of recurring artefacts. In short, this step evaluates how a shared variant is distributed across samples from the same patient. If it is truly somatic, it will be present at a high VAF in a subset of samples, but absent from others. This translates to a highly over-dispersed beta-binomial distribution. Alternatively, if the variant is artefactual in origin, it will be present in a few reads across many samples, resulting in a distribution more akin to a classical binomial one, i.e. with a low over-dispersion factor. Hence, quantifying the degree of over-dispersion for a given shared variant allows for its classification as either a putative artefact or a genuine variant. For PD28690, a total of 3,641 variants were filtered out this way (1.2% of putative somatic variants). Variants were also filtered out when they resided in regions of the genome with consistently low or high sequencing depth across all samples. These sites are likely to be affected by difficulties in mapping or other artefacts. For PD28690, this amounted to 4,299 variants (1.4% of putative somatic variants).[4]

The remaining SNVs were then used as a basis for the maximum parsimony tree reconstruction. In the case of PD28690, 306,098 variants remained for the tree building step. Note that this reconstruction only informs the topology of the tree. The assignment of mutations onto individual branches is done in a separate step. In this way, it is possible to directly assess

[4] 586 SNVs failed both the beta-binomial filter and occurred in regions of consistently high or low coverage.

which variants violate the structure of the phylogeny, due to their possibly artefactual nature or a genuine biological effect. In addition, this step filters out low VAF mutations in single samples, as the mutation mapping algorithm will assume a high clonality for a given branch.

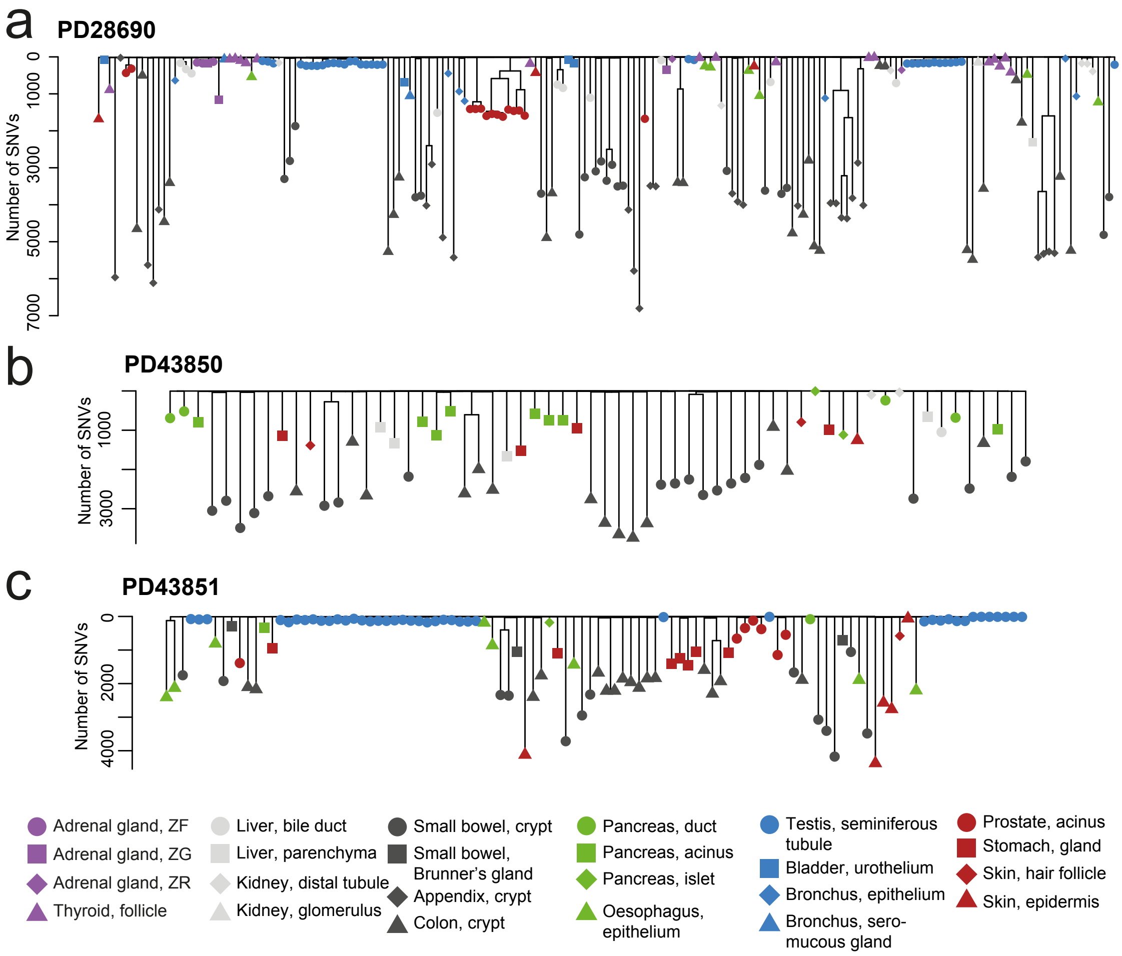

Figure 3.2 Phylogenetic trees of clonal LCM cuts of PD28690 (a), PD43850 (b) and PD43851 (c). The branch length conveys the number of SNVs acquired in that lineage. The shape and colour of the dots reflect the tissue type of the sample as per the legend.

The resulting phylogenies display the post-zygotic genetic relationship of all considered samples per patient (**Fig. 3.2**). When the branch length reflects the number of SNVs acquired in that lineage, all three trees exhibit a “comb-like” structure. In other words, most genetic sharing occurs in a short amount of time at the top of the tree, after which most SNVs are acquired privately in each sample. The terminal branch length reflects the mutation burden for the last single cell ancestor that generated the majority of cells in that sample. This cell could have been present many years prior to sampling, and the time delay likely differs per tissue. For example, the turnover time for an intestinal crypts is in the range of a few years (Nicholson et al., 2018), while most other tissues are estimated to have a much longer turnover time. Because the delay between the existence of the progenitor and time of sampling is unknown and variable, the estimates of the mutation burden do not reflect the mean number of mutations per cell at the age of sampling. In line with previous estimates, intestinal crypts exhibit a high mutation burden, while seminiferous tubules are largely spared from heavy mutagenesis. The latter finding is consistent with estimates of the paternal age effect in *de novo* germline mutations from trio studies (Rahbari et al., 2016). This reinforces that the main cellular component sequenced in these LCM cuts of seminiferous tubules are the spermatogonial cells that ultimately give rise to the spermatocytes.

3.5 Embryonic asymmetries

The previous visualisation of the phylogenies of normal tissues obscured the dynamics of embryogenesis, as this unfolds when few SNVs have been generated. These patterns re-emerge, however, when reducing the phylogenetic tree to its fundamental topology, in essence by setting a unit branch length. Every node in the resulting phylogeny depicts a distinct progenitor cell that existed at one time during human development or adult life. As most of the splits in the phylogenies occur after very few mutations have been acquired, the majority of nodes in the three trees are embryonic progenitors.

In all three cases, the tree starts in a bifurcating fashion for roughly two generations before preferentially generating large multifurcations or polytomies. These polytomies can be explained by cell divisions occurring without an observed mutation, either through an inability to detect these or their inexistence. The former explanation is unlikely given the repetition of this pattern across these patients. If later cell divisions genuinely have a higher probability of being ‘silent’, this must reflect a change in the mutation rate during early embryogenesis. Previous studies have observed a high mutation rate in the first two cell divisions, which subsequently decreases (Bae et al., 2018; Ye et al., 2018a). The time of the reduction of the mutation rate coincides with zygotic genome activation, when the translation machinery of the embryo starts generating proteins *de novo* rather than relying on oocyte proteins. It is plausible that this new abundance of embryonic proteins bolsters the DNA repair machinery, which had become diluted during the rapid cleavage of early embryogenesis. For our three patients, I estimate a mutation rate in the first two generations of approximately 2.4 per branch (95% confidence interval: 1.7-3.2), while it drops to a mean of 0.64 for subsequent generations (95% confidence interval: 0.51-0.77). See Chapter 2, section 2.5.1, for more information on the methods behind these estimates.

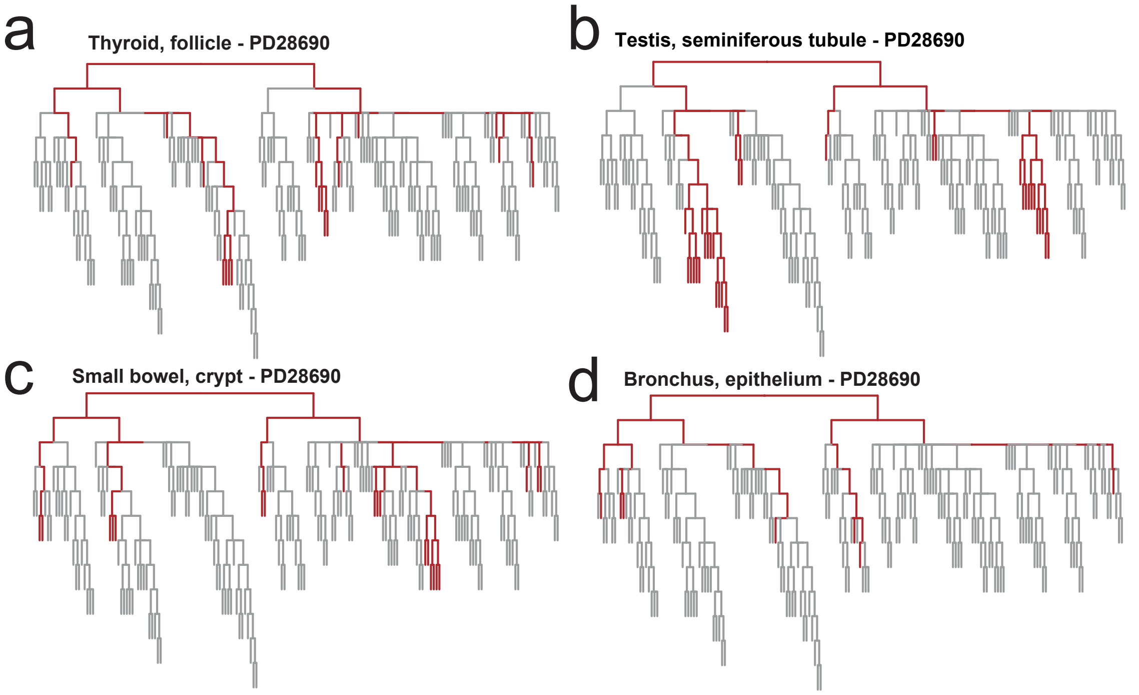
a
Thyroid, follicle - PD28690
b
Testis, seminiferous tubule - PD28690
c
Small bowel, crypt - PD28690
d
Bronchus, epithelium - PD28690

Figure 3.3 Phylogenetic trees with unit branch lengths for PD28690, showing the coalescence (red) of all samples from four tissues types: thyroid follicles (a), seminiferous tubules (b), small bowel crypts (c) and bronchal epithelium (d). The most recent common ancestor for all these tissues is the root of the tree.

In PD28690, the most recent common ancestor for all individual tissue types was the root of the tree (**Fig. 3.3**). Presumably, the root of the tree can be equated to the zygote, although this assumption will be explored in more depth in Chapter 5. The commonality of the most recent common ancestor indicates that no tissue studied here has a post-zygotic monophyletic origin. In other words, no tissue sequenced here is entirely derived from a single cell beyond the cell that gave rise to the entirety of the embryo. For example, following the lineages of intestinal crypts derived from the small bowel (**Fig. 3.3c**), we see that both daughter nodes and all four granddaughter nodes give rise to at least one crypt. After this stage, individual lineages on the tree are no longer observed to contribute to the small bowel, but this is undoubtedly influenced by the number of biopsies and LCM cuts taken. This picture of a polyphyletic origin for all tissues is in line with lineage commitment occurring during a later stage of embryogenesis with a much larger cell population, as is the case in gastrulation, organogenesis and beyond. This pattern is consistent with the phylogenetic trees of PD43850 and PD43851.

Previous studies on somatic phylogenies of mouse and human observed an asymmetric contribution of the two daughter cells of the presumed zygote, such that one contributes twice

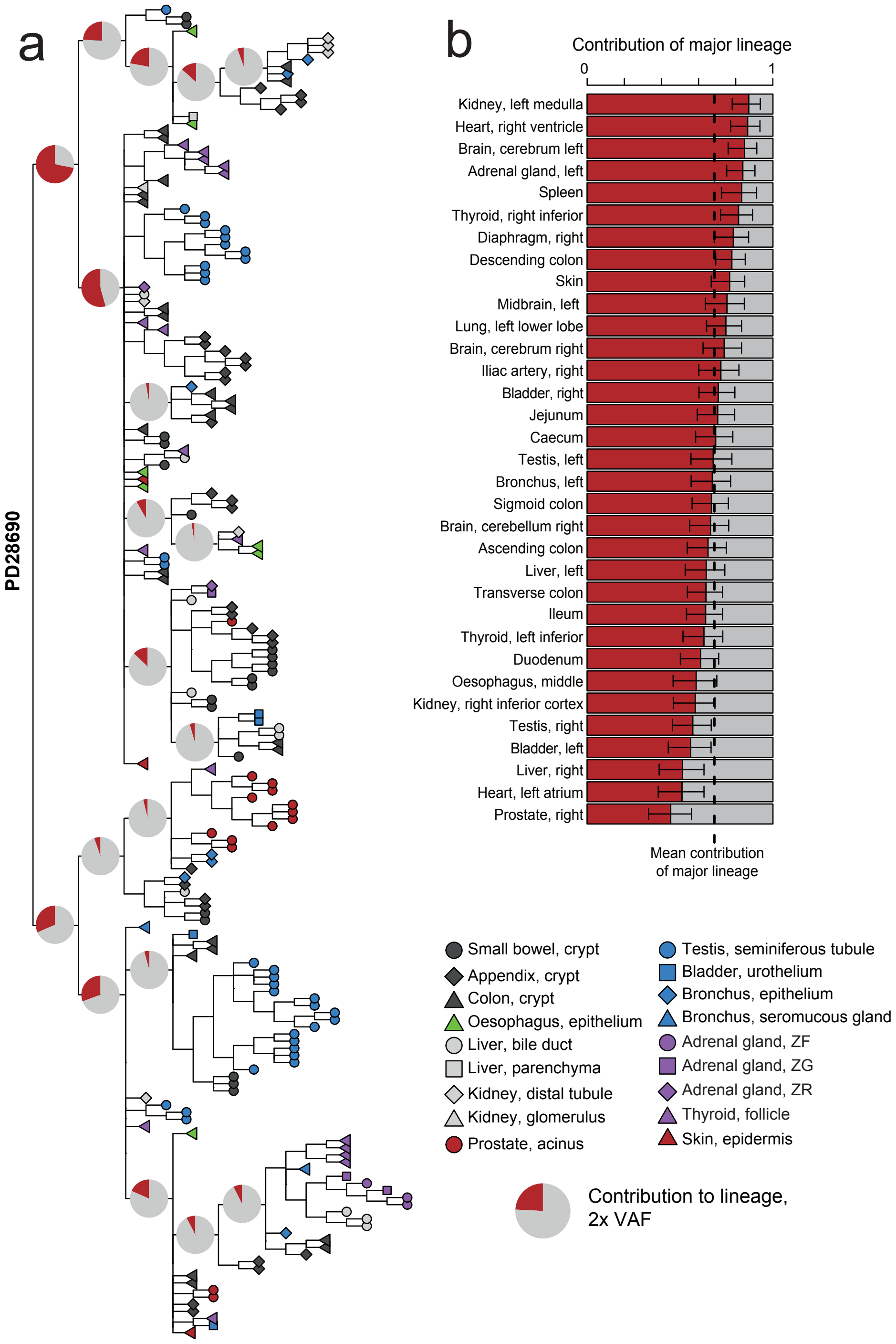

Figure 3.4 (a) Phylogenetic tree for PD28690. The shape and colour of the labels indicate tissue type. Pie charts show the mean contribution (twice the VAF) of that lineage to the 33 bulk biopsies of this patient. (b) The contribution of the major lineage to each bulk biopsy.

as much to the developed body as the other (Behjati et al., 2014; Ju et al., 2017; Lee-Six et al., 2018). It has been postulated that this is a possible consequence of differential cell allocation to the trophectoderm and inner cell mass during blastulation.

To assess the contribution of early embryonic progenitors to the developed body, I recounted all somatic variants contained in the phylogenies in whole-genome sequencing data from 33 bulk samples from 18 different organs for PD28690 (**Fig. 3.4**). Many of these biopsies are derived from organs that have no representative tissue in the set of clonal LCM samples, such as the post-mitotic brain and heart. Hence, the presence of previously identified variants confirms the pre-gastrulation origin of these SNVs. Mutations in the two daughter reconstructed cells exhibit different mean VAFs: 0.36 (the major lineage) and 0.16 (the minor lineage).This corresponds to an asymmetric contribution of 69:31 (95% CI: 66.7-70.2). This asymmetry is reflected in most individual bulk samples as well, meaning that all were derived from a large population of ancestral cells. The observation that these VAFs sum up near 0.5 indicates that the lineages identified in the phylogeny construction fully account for all samples sequenced in this patient.

There is an establishment of further asymmetries in subsequent generations of PD28690. The major lineage splits into two unequally contributing lineages with mean VAFs of 0.27 and 0.12, while the minor lineage splits in 0.15 and 0.02. However, a single, simple bottleneck of cell allocation to the trophectoderm or inner cell mass would not account for these later asymmetries. Therefore, this observation suggests that multiple, successive bottlenecks shape the quantitative contribution of individual embryonic progenitors to the adult body.

For PD43850, we estimate early lineage contribution using a single polyclonal bulk sample derived from brain, an organ unrepresented in the phylogeny. PD43850 exhibits a more modest asymmetry than PD28690 (60:40, 95% CI: 43.9-74.5; **Fig. 3.5a**), although this could be due to the paucity of available bulk samples.

For PD43851, bulk samples from both brain and colon were available. In the primarily ectoderm-derived brain sample, the stark asymmetry (93:7, 95% CI: 87.9-96.1) indicates it was almost entirely derived from one of the two reconstructed cells in the first generation (**Fig. 3.5b**). However, in the primarily endoderm-derived colon sample, the asymmetry is 81:19 (95% CI: 74.3-87.3), significantly different from the observed asymmetry in brain (p=0.004, likelihood ratio test, Methods). Other than brain, the only other ectodermal tissue sampled in the phylogeny of this individual is epidermis, microdissections of which are exclusively derived from the major lineage. By contrast, endoderm-derived microdissections are distributed on the phylogeny (81:19) in line with the asymmetry in colon (p=0.86), but not brain (p=0.005), further confirming the difference in asymmetry observed in bulk samples. Interestingly, the minor lineage is ancestral to the large majority of seminiferous tubule microdissections, most from a distinct daughter lineage, and their distribution (major lineage:minor lineage 35:65) is very different from the asymmetry in brain and colon (p<10-8 for both comparisons, binomial test). Subsequently, the first split in the major lineage of the brain exhibits a similar asymmetry (64:36), while this amounts to 50:50 in the colon. This difference in asymmetry between multiple tissue types is a further indication that multiple bottlenecks beyond the first shape the contribution of embryonic progenitors to different tissues by their origins from distinct cell populations. This effect is especially pronounced in the primordial germ cell-derived seminiferous tubules.

It is plausible that the stark difference asymmetry observed between the brain and colon sample is due to their developmental histories. Either on a biopsy, organ or germ layer level, it is conceivable that the brain has a more restricted set of progenitors than the rest of the body. It has been shown that there is a mixing of previously extraembryonic cells with embryonic cells to form definitive endoderm and mesoderm (Ferretti and Hadjantonakis, 2019; Nowotschin and Hadjantonakis, 2020). There is currently no evidence of such a re-entry of extraembryonic lineages in ectoderm. In addition, although the exact origin of human primordial germ cells remains unclear, it is thought that they segregate prior to gastrulation and possibly arise from extraembryonic tissues as well (Kobayashi and Surani, 2018). Thus, these results are consistent with a scenario in which all cells undergoing gastrulation have been derived from one of two daughter cells of the zygote, while this is not true for the extraembryonic hypoblast and primordial germ cell lineages. Subsequently, this exclusivity is maintained in ectoderm but lost in the intercalated germ layers and lineages of endoderm and mesoderm, and primordial germ cells.

Early cells committing to extraembryonic lineages would result in cell divisions not being observed. This means that in actuality early branches represent multiple cell divisions. As the cell allocation to trophectoderm and inner cell mass is a proposed cause for the early asymmetries in the phylogenies, it follows that the major lineage represents the branch from the zygote with more inner cell mass progenitors than the minor lineage. In general, this would result in more divisions being detected in the major lineage than in minor lineage. Because of this, it is expected to find more SNVs on the branch constituting the minor lineage and fewer on the branch of the major lineage. Indeed, I observe an average of 1.7 SNVs on the first major branch, while the mean burden of the minor branch is 4.3. This lends more weight to the assumption that the asymmetry is due to cell allocation. If it were purely due to a differential potential of cells (i.e. the progeny of the minor lineage divides more slowly than the progeny of the major lineage), we would not observe a considerable difference in mutation burden in the cell division leading up to these cells.

However, cell allocation at a single point in time (i.e. the lineage commitment at the blastula stage) cannot be the sole explanation for the observed embryonic asymmetries. A differential allocation of cells can explain the first observed asymmetry of 2:1, but fails to explain subsequent asymmetric contributions, as seen in generations beyond the first in the three phylogenies presented here. After all, the progenitor cells of the inner cell mass would all contribute equally after the segregation. Instead, it is likely that later lineage commitments, such as the segregation of hypoblast or amnion, further tunes the embryonic contributions of individual cells. In addition, cells in the embryo abandon the semi-synchronous divisions in the blastula stage, which makes it plausible that certain cells proliferate more rapidly than others. This could have its origin in emergent spatial effects prior to and during gastrulation, such as proximity to the hypoblast, which we have shown to induce proliferation in the epiblast via the fibroblast growth factor pathway around day 9 of embryogenesis (Molè et al., 2021).

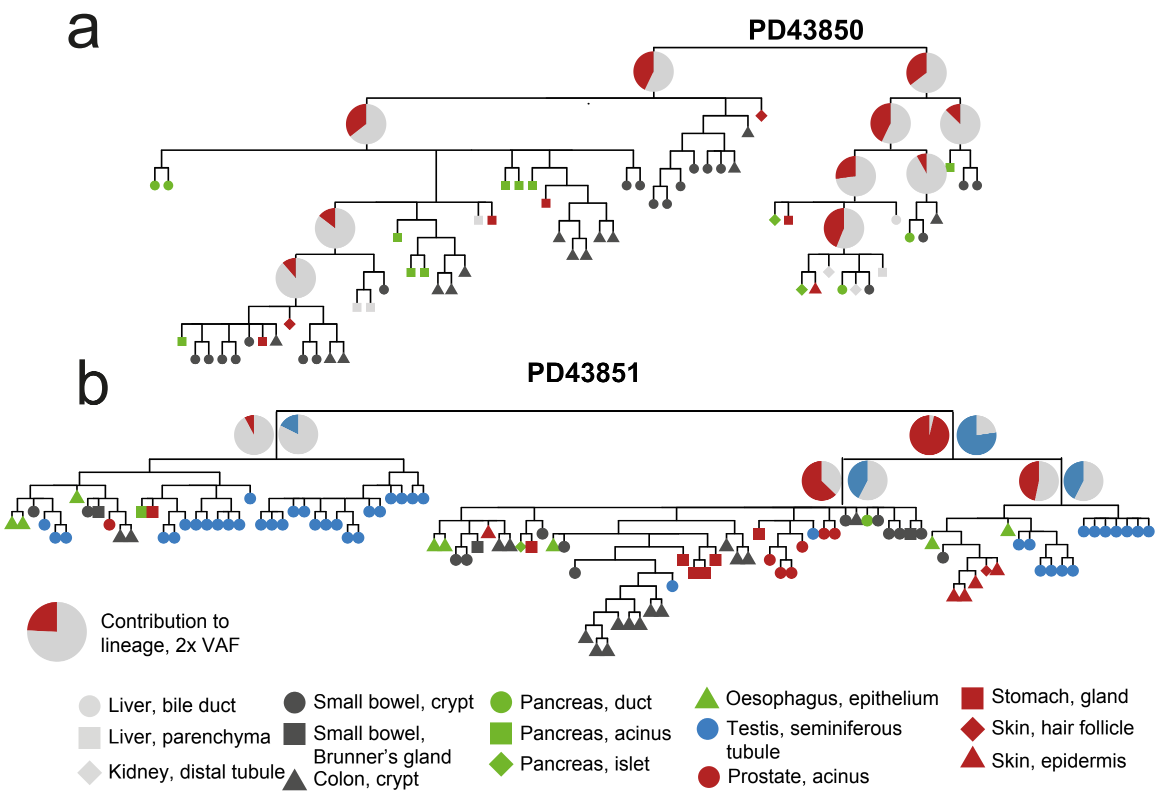

Figure 3.5 Phylogenetic tree topology for PD43850 (a) and PD43851 (b). The shape and colour of the tip labels reflect the tissue type of the sample as per the legend. Pie charts show the mean contribution (twice the VAF) to a bulk brain sample of PD43850, or a bulk brain (red) or colon (blue) sample of PD43851. Only lineages with a mean VAF over 0.02 are shown.

3.6 Targeted sequencing and organ-wide mosaic patterns

The previous sections covered the discovery of somatic variants, many of which were embryonic in origin, to reconstruct phylogenetic trees of human development. When the topology of the tree and the SNVs that encode early development are known, it is straightforward to specifically interrogate these sites to assess the contribution of different embryonic lineages to previously uninformative samples such as bulk biopsies and polyclonal LCM samples. The next few sections will delve into this approach.

A first experiment to this end involved designing a bait set as a basis for targeted re-sequencing of sites of interest for PD28690. These included the early embryonic variants that constitute the top of this patient's phylogeny. To assess questions of the role of laterality in development, as well as the broad spatial patterning of the development of individual organs, 86 different bulk samples of this patient were subjected to targeted sequencing. These 86 biopsies included the 33 previously subjected to whole-genome sequencing. In addition, 40 of these biopsies were derived from various regions of the brain, including the cerebellum, cerebrum and midbrain.

The VAF of embryonic mutations in individual bulk samples reflect the developmental bottlenecks and the population of founder cells, even if the mutations themselves were acquired prior to lineage commitment. In other words, even if the discovery of embryonic variants did not yield SNVs specific to the brain (for example), a sharing of a specific VAF pattern of pre-gastrulation mutations might indicate a proximity of development between two brain biopsies.

A direct comparison of VAFs between two samples fails to account for the dependence of the observations. Comparing the similarity of two vectors of observations is a common problem. For example, comparison of two mutational spectra or signatures, in effect two vectors of 96 values, is often done through the cosine similarity metric. In this example, the frequency or probability of C>T mutations in a CCT context is not directly influenced by a different trinucleotide change such as C>T at CCC. However, the mutations on a phylogeny are not independent of one another. SNVs acquired on the same branch will convey the same information, while SNVs on parallel branches will represent complementary lineages. Therefore, it is necessary that each VAF measurement is weighted by a similarity metric of two mutations, i.e. their distance on the phylogeny. The developmental proximity between two samples is then a version of the soft cosine similarity metric, which is employed in the field of natural language processing (Sidorov et al., 2014). Please refer to the Chapter 2, section 2.5.1, for an in-depth explanation of the calculations.

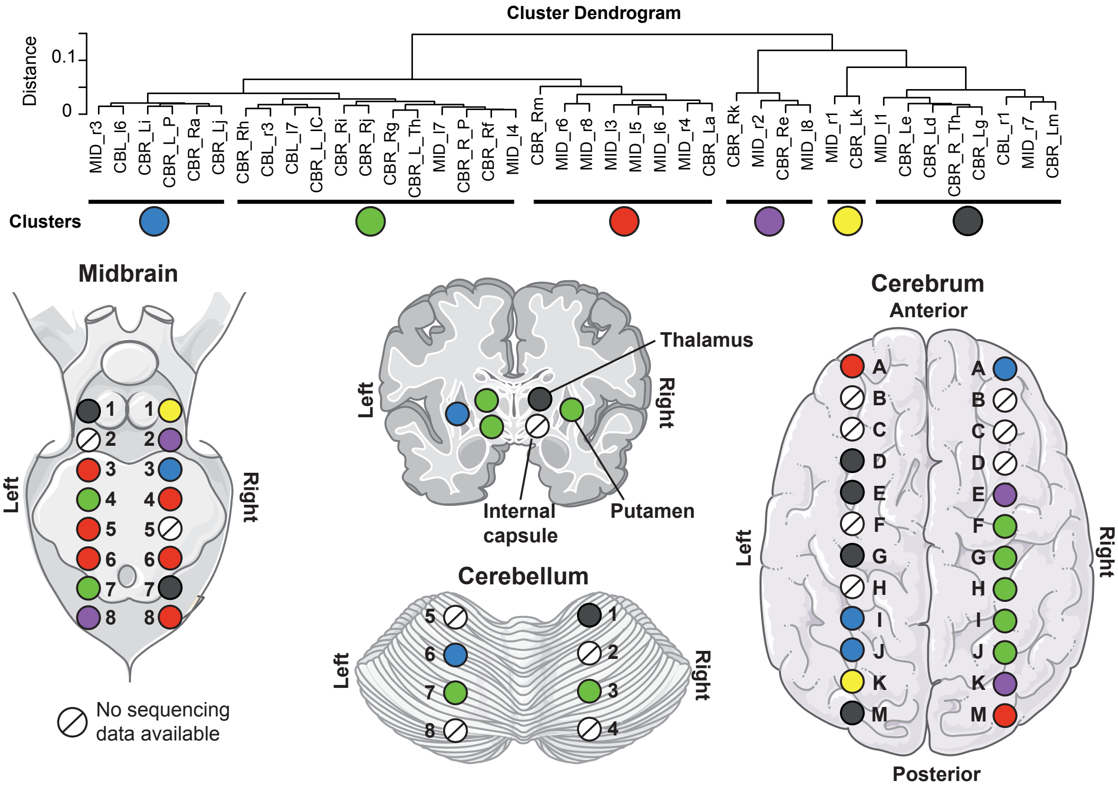

Figure 3.6 Cluster dendrogram of soft cosine similarity of VAFs of early embryonic mutations in bulk brain samples of PD28690. The data is split into six clusters, which are spatially displayed in the various brain regions. CBL=cerebellum, CBR=cerebrum, IC=internal capsule, L=left, MID=midbrain, P=putamen, R=right, Th=thalamus. Parts of the figure are composed of pictures from Servier Medical Art. Servier Medical Art by Servier is licensed under a Creative Commons Attribution 3.0 Unported License (https://creativecommons.org/licenses/by/3.0/).

Clustering of bulk samples subjected to targeted sequencing did not result in germ layer-specific clusters, as one might have expected. Given the lack of monophyletic origins for individual tissues, it is clear that no single post-zygotic cells account for the individual germ layers. Moreover, the global conservation of the embryonic asymmetries of the first split suggests that germ layers are generally derived from a cell population of sufficient size as to not impose a tight bottleneck that would distort that asymmetry. Combined with the majority of embryonic mutations re-sequenced being of pre-gastrulation origin, it follows that different organs or germ layers do not segregate into different clusters. Although the segregation between the large cell populations of the germ layers might remain out of sight, clustering within germ layers or organs might reveal patterns due to spatial effects in the early embryo or small pools of progenitors, as has been noted for the murine heart (Abu-Issa et al., 2004).

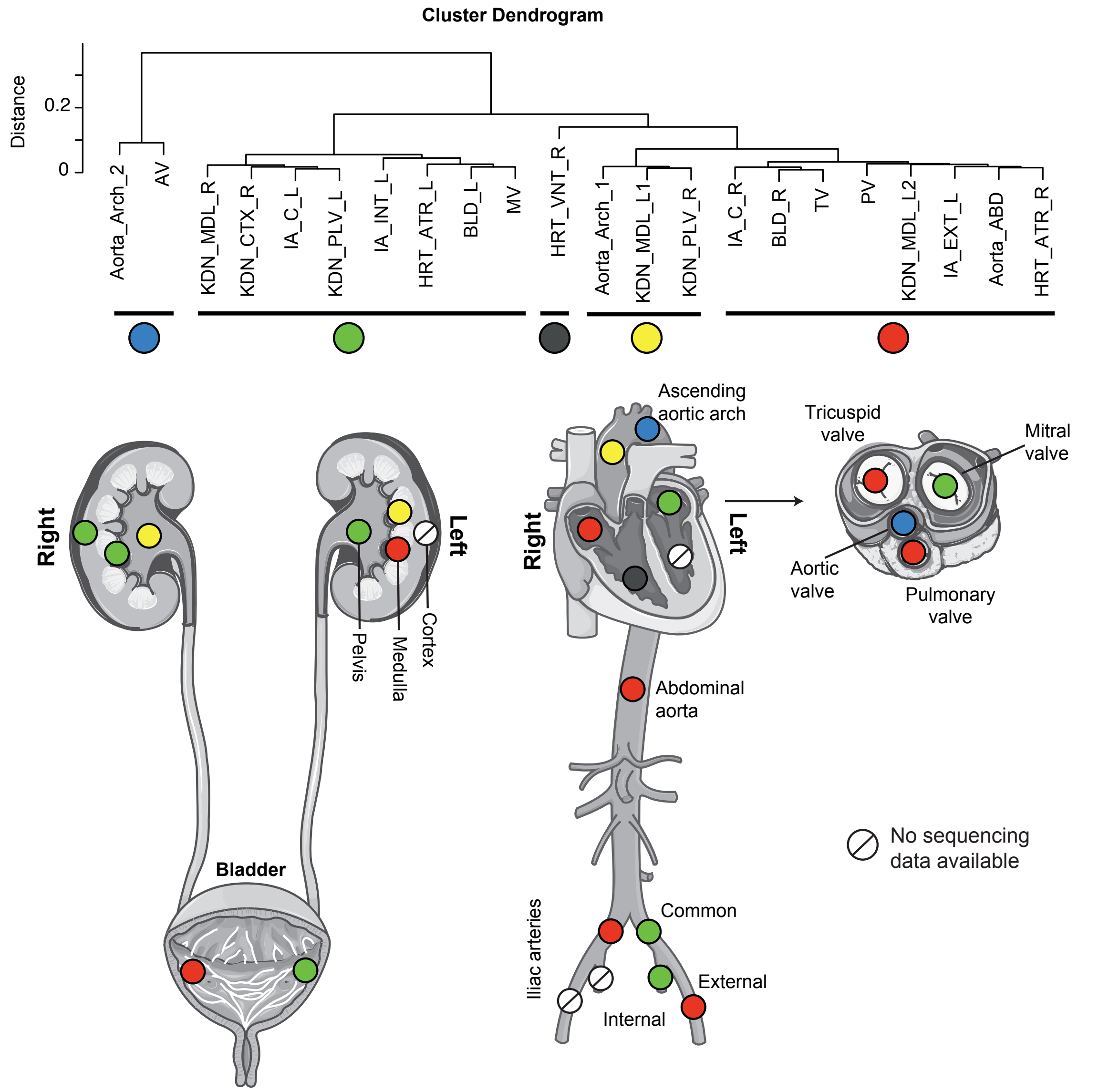

Figure 3.7 Cluster dendrogram of soft cosine similarity of VAFs of early embryonic mutations in mesodermal samples of PD28690. The data is split into five clusters, which are spatially displayed in the various organs and tissues derived from the mesoderm. ABD=abdominal, ATR=atrium, AV=aortic valve, BLD=bladder, CTX=cortex, EXT=external, IA=iliac artery, INT=internal, KDN=kidney, L=left, MV=mitral valve, PLV=pelvis, PV=pulmonary valve, R=right, TV=tricuspid valve, VNT=ventricle. Parts of the figure are composed of pictures from Servier Medical Art. Servier Medical Art by Servier is licensed under a Creative Commons Attribution 3.0 Unported License (https://creativecommons.org/licenses/by/3.0/).

In particular, the large number of samples obtained from the brain allows for visualisation of the large-scale mosaic pattern of this organ. In total, 22 cerebral samples, 4 samples from cerebellum, and 14 midbrain samples were subjected to targeted sequencing. Clustering based on the soft cosine similarity revealed that all 40 brain samples fell into roughly six different clusters (**Fig. 3.6**). Strikingly, adjacent biopsies in the cerebrum and midbrain often belong to the same cluster. It is plausible that this is a consequence of a shared developmental path and the seeding of those areas by similar populations of cells. Moreover, the different regions of the brain show a heterogeneity in the prevalence of different clusters, with one particular cluster being prominently present in the midbrain. The correlation between the spatial and genetic distance was significant in the cerebrum but not the midbrain (p=0.039 and p=0.27, Mantel test). This means that, despite these bulk samples being taken millimetres apart and being composed of many different cell types, there is greater sharing of a specific VAF pattern of pre-gastrulation mutations between neighbouring regions of the cerebrum compared to distant regions. Taken together, this data indicates that developmental bottlenecks and lineage commitments generate a spatial effect in the mosaicism of the brain, especially in the cerebrum.

In addition, the same analysis was performed on 22 bulk samples obtained from a variety of mesodermal organs, including bladder, heart and kidney (**Fig. 3.7**). Five clusters separated the data, but no clear organ-wide patterning emerged, including in the difference between the left and right kidney. A denser sampling strategy, such as was employed in the brain, combined with deeper sequencing and more extensive discovery of peri- and post-gastrulation mutations, would benefit the detection of possible organ-wide spatial patterns in the mesoderm.

3.7 Spatial genomics of mosaicism and embryonic patches

As a first experiment, I combined the genetic information from whole-genome sequencing with the spatial information from histological sections to directly visualise the mosaic patterns laid out early in development. In the case of intestinal crypts derived from the jejunum and ileum sections of the small intestine (**Fig. 3.8**), this revealed that stretched of crypts can arise from the same embryonic progenitor, as seen in samples labelled SB2_C11, SB2_E11, SB2_F11 and SB2_G11 (coloured in blue) in ileum, as well as SB1_G8, SB1_H8 and SB1_A9 (coloured in red) in jejunum. It is worth noting that only SB2_F11 and SB2_G11 appear to be related through a crypt fission event later in life, indicating that the genetic relationship observed between these crypts has its root in the embryonic seeding of that section of the intestine. In contrast to large embryonic patches, instances of neighbouring crypts being derived from the other lineage of the first dichotomy are also captured, as is the case for samples SB2_H10 and SB2_F10 contrasting with SB2_G10.

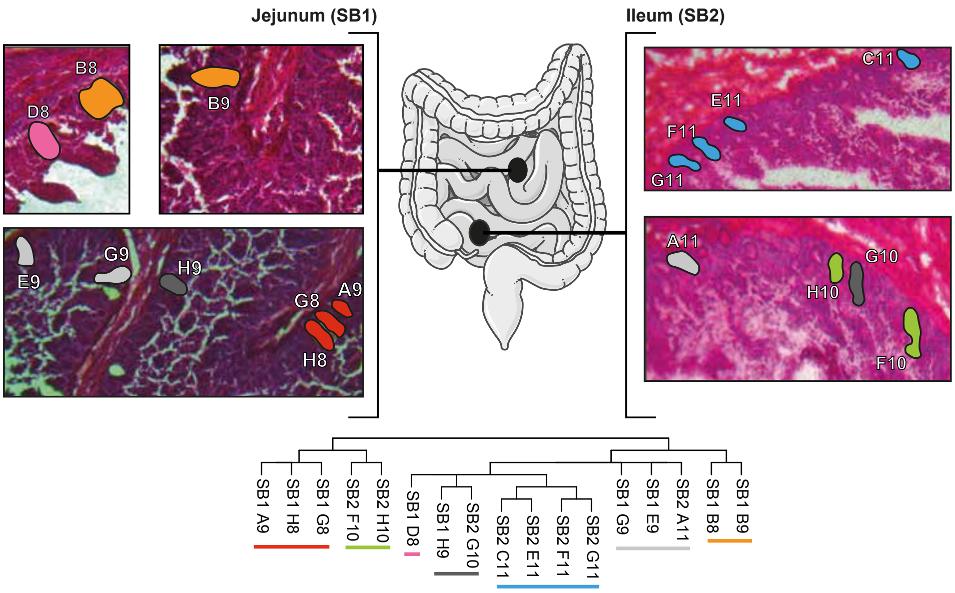

Figure 3.8 Histology sections of ileum and jejunum of PD28690 coloured by location on the phylogeny. Parts of the figure are composed of pictures from Servier Medical Art. Servier Medical Art by Servier is licensed under a Creative Commons Attribution 3.0 Unported License (https://creativecommons.org/licenses/by/3.0/).

To assess what generally happens in the development of colonic crypts, I supplemented the data from our individuals with a published data set from our lab (Lee-Six et al., 2019), employing the same sampling and sequencing strategy on a further 326 crypt from 36 patients. I subjected this data to the exact same variant calling procedures. Because of the sampling strategy, I was able to quantify the distance between pairs of crypts from the same histology section and number of mutations shared from the sequencing data, such as exemplified in the phylogeny and histology sections of colonic crypts in PD43851 as shown below. This shows that these crypts form small clades with neighbouring crypts, sharing between 20 and 45 mutations in this case (**Fig. 3.9a,b**). To avoid incorporating later crypt fission events, I excluded pairs of crypts sharing more than 200 SNVs. I then looked at the distribution of physical distances and genetic sharedness (**Fig. 3.9c**).

As was the case with small bowel crypts, many pairs of crypts from the same section share none or fewer than 15 SNVs. These crypts derive from different daughter cells of the zygote or from progenitors that have shared only a few post-zygotic cell divisions. This sharing will have occurred in the very early embryo, prior to organogenesis, and is therefore uninformative for the estimation of embryonic patch sizes. There is a distribution of shared mutations above 15 SNVs (**Fig. 3.9d**), which reflects the sharing as a consequence of embryonic patch formation.

For the distribution above 15 SNVs, the median number of SNVs shared between a pair of crypts is 27. This is an estimate of the mutation burden of the founder cell of the embryonic patch. Data from single cell-derived organoids of fetal intestinal stem cells (Kuijk et al., 2019) suggest a weekly mutation accumulation of 3-4 SNVs, suggesting the founding of an embryonic patch of crypts happens around 7-9 weeks post conception. This is also in line with data on fetal blood colonies, which possess an average mutation burden of 25.5 SNVs at week 8 (Spencer Chapman et al., 2020). After this time point, it seems the colon progenitor cells have specified and there is no further mixing.

To estimate the crypt patch size from our data, we simulated a simple model of spatial sampling and used approximate Bayesian computation to evaluate the posterior distribution of radii of embryonic patches, which we assume to be circular in shape. Briefly, the model works as follows:

1. Sample an embryonic patch size radius (r) from a uniform distribution between 2 and 20 crypts (in steps of 0.01) (**Fig. 3.9e**).

2. Uniformly sample an integer point within a circle with radius r, centred around the origin.

3. Sample an integer distance from that point. To account for our bias in sampling, this distance is drawn from the a distribution with the same distance frequencies as our data.

4. Evaluate whether the second point falls within the circle of radius r.

5. Repeat this as many times as there were observations in the original data set. For each distance evaluated, return the counts of crypt pairs belonging to the same patch, which will act as the summary statistics. To avoid noisy observations, we included distances for which we had more than 10 observations, corresponding to the range of distances of 1 to 17 crypts.

This simulation was run 50,000 times. We approximate the posterior distribution of the patch size radius by accepting the 5% simulations with the lowest Euclidean distance to the original sharing counts per distance (**Fig. 3.9f**). This was followed by a neural network regression method, to gain a more precise estimate (**Fig. 3.9g,h**). The posterior distribution estimated from rejection alone has a mean radius of 8.68 (95% confidence interval: 6.03 -11.64), which translates to embryonic patch size of 237 (95% confidence interval: 114 - 425).

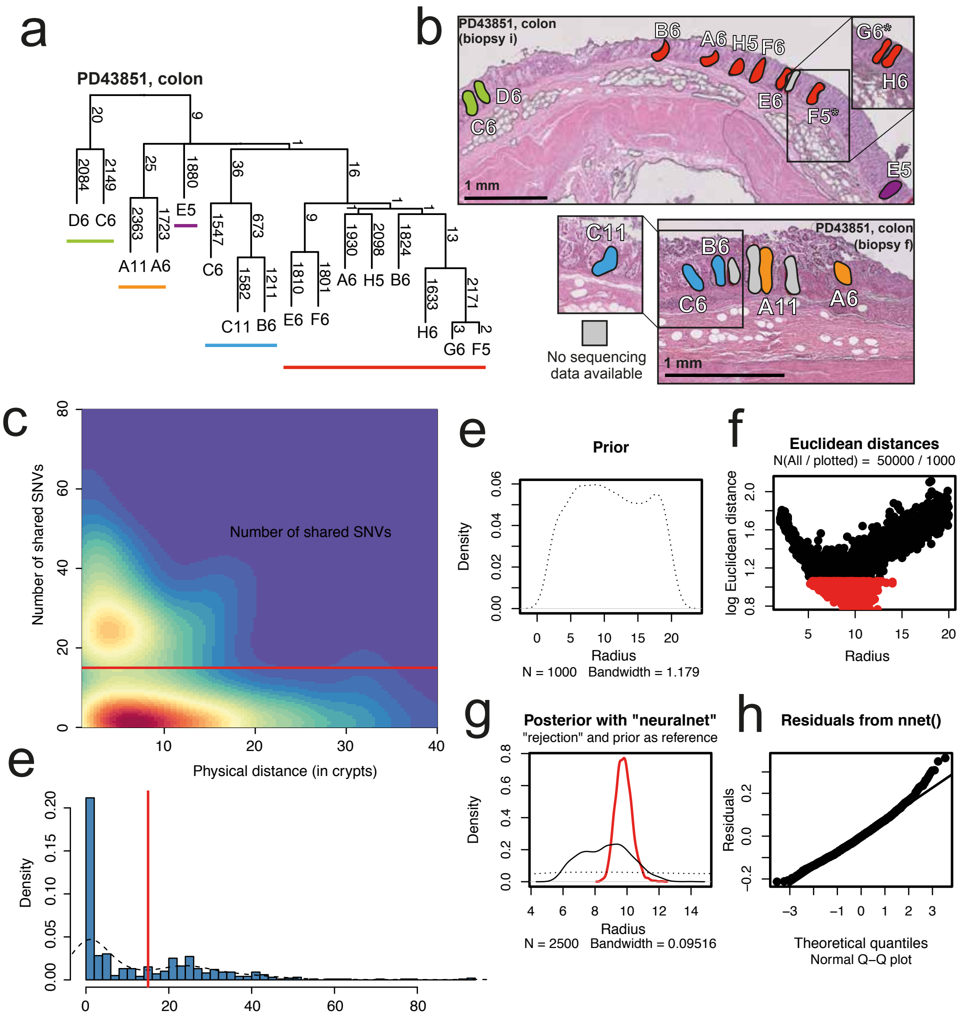

Figure 3.9 (a) Phylogeny of colonic crypts in PD43851, with coloured lines indicating distinct embryonic patches, each diverged after 20-46 SNVs. (b) Histology sections of samples shown in (a). Connected lines and boxes indicate z-stacks, with the asterisk indicating these samples were taken from different sections of the same crypt. (c) Kernel smoothed 2D histogram of the linear distance (in number of crypts) and number of shared SNVs between any two crypts from the same biopsy. The red line at an shared SNV burden of 15, above which crypts were taken to be from the same embryonic patch. (d) Histogram of number of SNVs shared between all pairs of crypts showing a bimodal distribution on either side of an SNV burden of 15 (red line). (e) Density plot of the prior distribution of the embryonic patch size radius. (f) Plot of the radius versus the Euclidean distance in summary statistics between the simulations and our observed data. Red dots are those within the 5% closest simulations and are accepted. (g) Density plot of the prior distribution (dashed line), the posterior distribution from the rejection method (black line) and the posterior distribution from the neural network regression (red line) of the embryonic patch size radius. (h) A QQ-plot of the residuals of the neural network regression.

The weighted posterior distribution estimated from the neural network regression method has a mean radius of 9.79 (95% confidence interval: 8.94 – 10.82), which indicates an embryonic patch size of 301 (95% confidence interval: 251 - 368).

The estimated size of the embryonic patches in colon we obtained from our sequencing data is consistent with the estimate obtained through X-inactivation studies (Novelli et al., 2003).

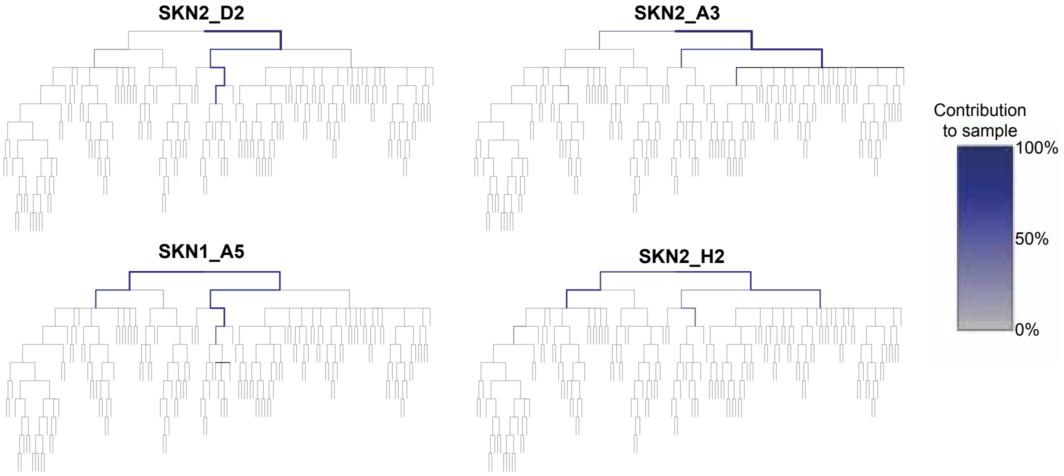

Figure 3.10 Phylogenetic trees with unit branch lengths for four polyclonal samples of epidermis of PD28690, showing the contribution (blue) of early embryonic progenitors in the phylogeny to the sample.

Although highly ordered tissues with distinct microscopic structures, such as intestinal epithelium, lend themselves to this type of analysis, echoes of embryonic patterning can be seen in other adult tissues. Polyclonal samples do not adhere to the phylogeny as a single leaf and were thus omitted from tree building, but their internal clonal architecture will obey the phylogenetic tree. Hence, it is possible to decompose polyclonal samples into the constituent embryonic lineages that gave rise to that sample by virtue of the VAFs of mutations identified in the phylogeny. In this way, even for polyclonal samples, it is possible to resolve the contributions of embryogenesis. Here, I illustrate this point by using epidermal LCM cuts from PD28690 (**Fig. 3.10**). The VAF patterns of mutations on branches in the early phylogeny revealed that, while some of the epidermal samples are aggregates of multiple embryonic clones and represent a mixing of different seeding events, other samples, such as SKN2_D2, seem to adhere to a single clonal lineage in the early embryo. However, this skin sample was excluded from the phylogeny reconstruction step due to an insufficient degree of clonality as apparent from its overall set of mutations. Its ‘embryonic’ clonality indicates that while this cut of epidermis does not have a single recent ancestor that accounts for the majority of cells, it does have an original, developmental founder cell, the progeny of which extensively pervades this sample.

3.8 Separation of primordial germ cells from somatic lineages

The primordial germ cells are the first to segregate into a separate lineage, which is thought to occur prior to or around gastrulation (Kobayashi and Surani, 2018). These cells only give rise to the oocytes and spermatogonia, the latter of which contributes the majority of cells in seminiferous tubules. Therefore, any mutations arising after the segregation of the germ line should not be seen in any sample derived from the three major germ layers. This allows us to estimate the timing of separation as well as the number of primordial germ cells giving rise to the data. In a hypothetical experiment where testes are sampled exhaustively and SNVs are recalled in an extensive set of somatic tissues, this could enable estimation of the total population size of the primordial germ cells.

On average, seminiferous tubules shared 7.0 mutations with any other lineage in PD28690 and 8.7 in PD43851. This indicates that an appreciable proportion of *de novo* germline variants arise during the first cell divisions of life, prior to primordial germ cell commitment. However, in PD43851, a commitment of progenitors to the primordial germ cell lineage was observed as early as the second observed division. This suggests that the dynamics and commitments in the early embryo are able to generate an early segregation of primordial germ cells and gastrulating cells, which can translate into early mosaicism that is confined to either lineage. Samples of seminiferous tubules clustered into three and five clades in PD43851 and PD28690, respectively.

Targeted re-sequencing of testis-specific variants in PD28690 confirm the observed segregation of primordial germ cells in the phylogeny. Excluding the two bulk biopsies derived from testes, there is no contribution of somatic mutations found exclusively in seminiferous tubules. This confirms that the observed split between germline and soma in the phylogenetic trees reflect the true segregation of primordial germ cells. Of course, the observed number of progenitors and the mutational timing of the split only confers a lower bound on both estimates. It is possible that primordial germ cells share a longer developmental path with extraembryonic lineages than cells that eventually constitute the embryo, such as through differentiation of epiblast cells into amnion before commitment to the germ cell lineage in turn.

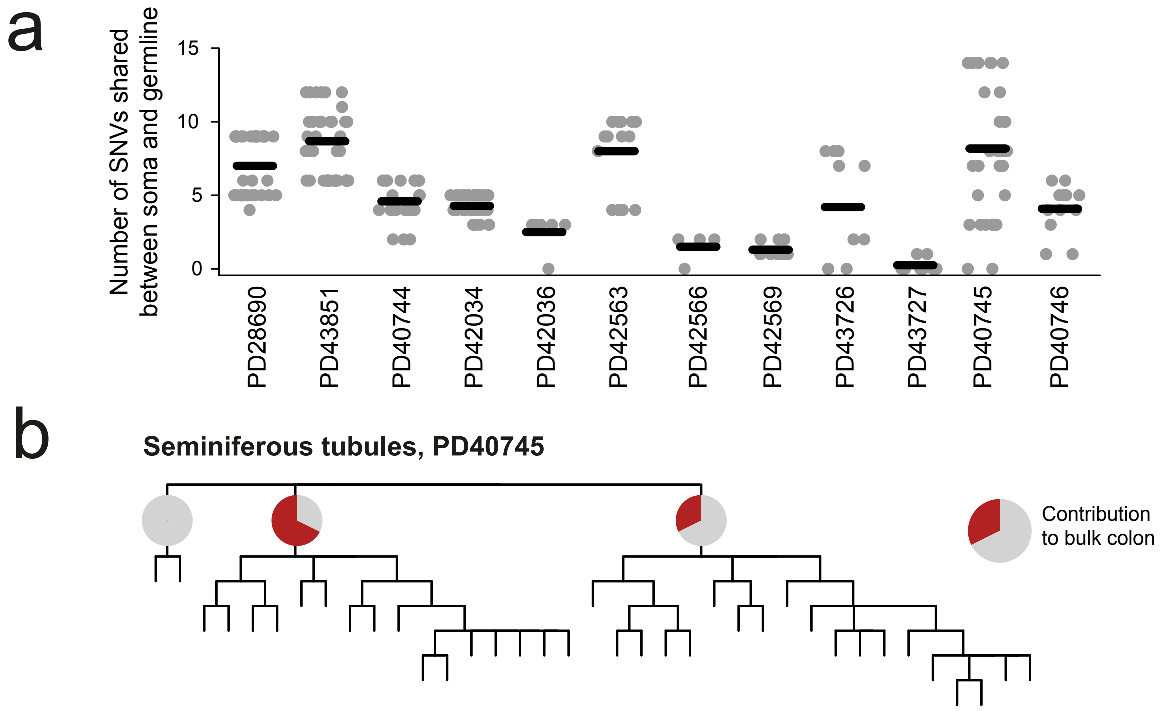

Figure 3.11 (a) Number of SNVs shared between any seminiferous tubule microdissection and another microdissection in PD28690 and PD43851. In other patients, number of SNVs shared between seminiferous tubule microdissections and matched bulk sample. (b) Phylogeny of seminiferous tubules of PD40735, with early lineage contribution to bulk colon in pie charts, showing a lineage of seminiferous tubules undetectable in bulk colon.

Further sequence data from 162 microdissections of seminiferous tubules from 11 individuals from whom colon or blood samples were also available (Coorens et al., 2021b) showed that, on average, seminiferous tubules shared 4.5 mutations with the bulk sample (**Fig. 3.11a**). Furthermore, in five out of twelve individuals, a subset of seminiferous tubules shared no SNVs whatsoever with their matched bulk sample. For example, in PD40745, the root of the phylogeny splits into three lineages of seminiferous tubules with only two making a detectable contribution to bulk colon (**Fig. 3.11b**). Similar patterns of early segregation have been observed for tissues with an extraembryonic origin or contribution, such as placenta (see Chapter 5, Coorens et al. (2021c)) or fetal blood (Spencer Chapman et al., 2020). This suggests that, due to later embryonic bottlenecks, a subset of primordial germ cells genetically segregated from the cells generating the three major germ layers after the first cell divisions of life.

It is hypothesised that the human PGCs originate from the posterior epiblast or nascent amnion (Kobayashi and Surani, 2018), which develops soon after the formation of the inner cell mass and implantation. Thus, the results here are consistent with the amnion as a site of human primordial germ cell specification. This extraembryonic contribution could also explain the strong deviation in asymmetry observed in the seminiferous tubules of PD43851 when compared to colon and brain.

3.9 Clonal expansions, benign prostatic hyperplasia and polyp formation

So far, we have mostly considered SNVs as marks of embryogenesis and early development. However, since DNA damage occurs all throughout life, somatic mutations are constantly acquired by cells and hence provide barcodes for continuous lineage tracing. As such, SNVs can be used to trace cellular dynamics later in life, including aberrant clonal expansions and emergent neoplasia.

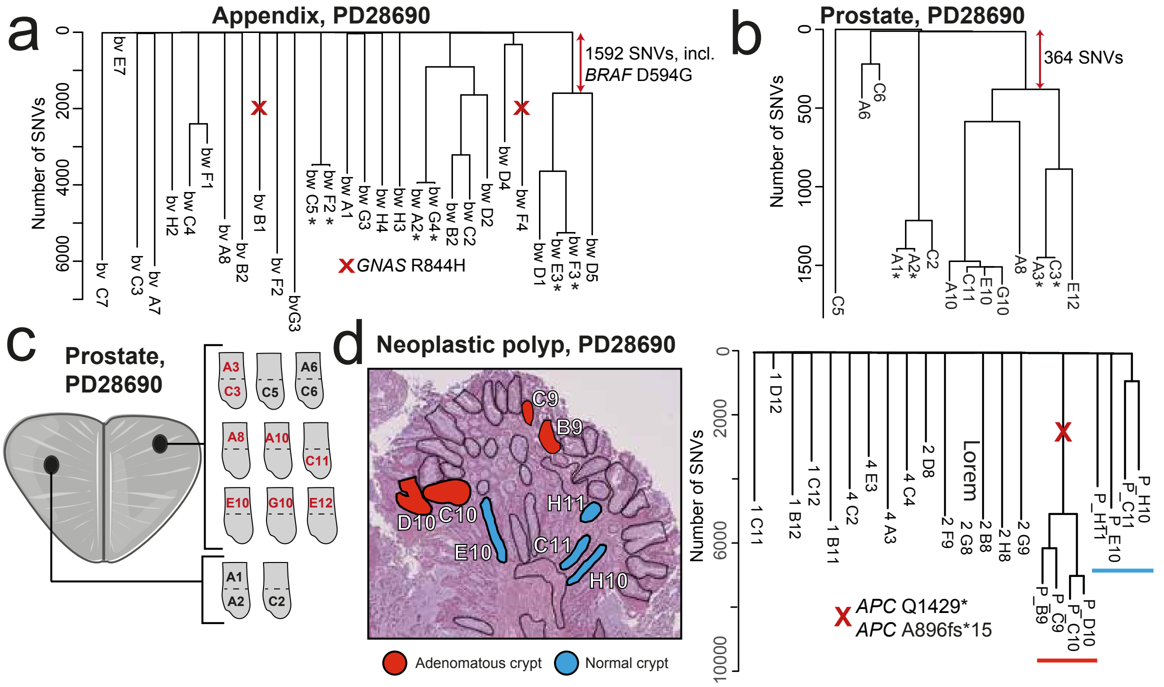

Figure 3.12 (a) Phylogenetic tree for appendiceal crypts in PD28690, with annotated cancer driver mutations. An asterisk indicates the two neighbouring crypts were taken as biological replicates of one another. (b) Phylogeny and sampling overview for prostatic acini in PD28690, showing widespread benign prostatic hyperplasia in one biopsy. (c) Histology and sampling overview alongside the phylogeny for a microscopic polyp in the colon of PD28690. Parts of the figure are composed of pictures from Servier Medical Art. Servier Medical Art by Servier is licensed under a Creative Commons Attribution 3.0 Unported License (https://creativecommons.org/licenses/by/3.0/).

As noted previously, the phylogenetic trees exhibit a comb-like structure when using the number of SNVs as the branch length, where most splits are obscured due to their embryonic nature. However, in certain tissues, the phylogenies show bifurcations after an appreciable number of SNVs have already been acquired. These mostly represent crypt fission events in intestinal samples (Lee-Six et al., 2019), including a notable expansion driven by a *BRAF* D594G hotspot mutation in appendiceal crypts from PD28690 (**Fig. 3.12a**). Assuming SNVs have occurred in a clock-like fashion (Alexandrov et al., 2015), this driver mutation was acquired prior to the age of approximately 23.[5] In addition, two appendiceal crypts carried a *GNAS* R844H mutation, which was the result of independent acquisition rather than shared development according to the phylogeny. This is further reinforced by the two crypts being excised from two different biopsies, since it is unlikely a large-scale expansion would be confined to only one crypt in each biopsy.

In addition, glands sampled from different regions of the prostate in PD28690 showed an intricate phylogenetic relationship (**Fig. 3.12b**). The mutation burden at which these splits happen is inconsistent with an occurrence during early development, as the mutational spectra of the prostatic acini appear to be solely due to intrinsic mutagenesis. This branching pattern signals an extensive clonal expansion in this organ. This is consistent with benign prostatic hyperplasia, a condition frequently affecting men aged over 60 (Berry et al., 1984). Assuming a linear mutation rate, the earliest post-developmental bifurcation in prostate can be timed to an age of approximately 19 years.[6]

Besides benign hyperplasia in the prostate, we also identified a microscopic polyp in PD28690, an early signal of transformation of normal colonic epithelium into colorectal carcinoma (**Fig. 3.12c**). The four crypts sampled from the polyp itself exhibited modest but marked increases in mutation burden compared to other, normal colonic crypts, including the four unmutated crypts excised from the same section. In addition, the truncal branch of the polyp crypts harboured two different mutations in the *APC* gene, one of which was a nonsense mutation (Q1429*) and the other a truncating frameshift insertion. These two hits effectively inactivate *APC* biallelically. The *APC* gene has been widely implicated in colorectal cancer and its inactivation is thought of as the first step in the progression of normal crypt to cancer (Fearon and Vogelstein, 1990; Fodde et al., 2001; Martincorena et al., 2017; Sparks et al., 1998). No copy number aberrations were found in the crypts of the polyp. It is worthwhile to note that all eight crypts, normal or adenomatous, from this section arise from the same embryonic lineage and share a total of 19 SNVs.

[5] The mutation burden at the split of the branch carrying the driver is 1592, the mean terminal burden of its leaves is 5330. The ratio of these numbers times the age of the patient (78) is equal to 23.3.

[6] Similar to previous calculation, with the burden of the branch in question 364 and the mean burden of its leaves 1531.

3.10 Recurrent SNVs and the infinite sites model

The assumption that somatic mutations generally represent unique events and do not happen twice is known as the infinite sites model and forms the cornerstone for maximum parsimony, the phylogenetic tree building method employed in this dissertation. The SNVs identified in this research and the constructed phylogenies can confirm or refute the validity of this assumption. One exception to the infinite sites model comes in the form of selected driver mutations, such as variants affecting hotspot regions in oncogenes. Indeed, in PD28690 we observe a recurrent mutation in *GNAS* (R844H) acquired independently in two appendiceal crypts.

In addition, a small number of SNVs in non-coding regions present as recurrent mutations in the same patient at the same site. This can be either the same nucleotide change, i.e. a mutation violating the topology of the tree, or a different one. While the human genome is large, it is not infinite. This sequencing study catalogued over 100,000 SNVs in each individual studied. While this number is not yet sufficiently high to saturate the genome and invalidate all inferences on the timing of the acquisition of the SNV, it is enough to occasionally encounter independent SNVs at the same site or even the exact same SNV altogether. This probability can be quantified using a birthday paradox-inspired collision model (Suzuki et al., 2006).[7] For ease of argument, we take the human genome to have three billion sites, all equally mutable. PD43850 and PD43851 each have slightly over 100,000 somatic mutations called, while PD28690 has approximately 300,000 in total. The probability of observing a site being mutated twice within the entire set of mutations is then 81% for PD43850 and PD43851, and approximately 100% for PD28690. If we require the exact same mutation to occur, i.e. increase the number of possible mutations to nine billion, these probabilities become 43% and 99%, respectively.

SNVs at the same site, but with a different substitution, do not necessarily represent independent events. In PD43851, two different SNVs occupy the same site (Chr22: 46,992,207) in neighbouring branches, a G>T substitution in one clade of epidermal LCM cuts and a G>C in the adjacent skin epidermis sample. It is possible that this is a consequence of an unrepaired DNA lesion passed on after cell division, which is subsequently repaired differentially in the two daughter clades. This would be an *in vivo* observation in normal tissues of the recently proposed DNA lesion segregation in cell lines and cancer (Aitken et al., 2020).

[7] The original problem asks the question how many people are required before the probability of any pair sharing a birthday equals 50%, to which the answer is 23 people. In a general setting, the probability of any number drawn from a uniform distribution being repeated is given by $p \approx 1 - (\frac{d-1}{d})^{n(n-1)/2}$, with d the number of draws (SNVs) and n the size of the uniform distribution (human genome) (Suzuki et al., 2006).

It is important to note that this does not invalidate the overall maximum parsimony assumption nor the reconstructed phylogenies. Mutations occurring at the same site, but with different consequences, would be considered independent events. The occasional recurrent, independent mutations violating the phylogeny will be vastly outweighed by the numerous uniquely acquired mutations delineating human development. Hence, a low number of departures from the infinite sites assumption will not affect the overall tree topology. However, this might increasingly become a problem as the capacity for sequencing is increased and these lineage tracing efforts using somatic mutations are done at an even larger scale.

3.11 Other types of mutations and recurrent loss of the Y chromosome

So far, SNVs in the nuclear genome have been the only type of somatic mutations used to infer phylogenies. This is due both to their relatively high rate of occurrence and their ease of discovery and validation. Recently, mitochondrial SNVs have been used for lineage tracing of human cells Ludwig et al. (2019). Since we can place mitochondrial SNVs on the phylogenies already constructed with nuclear SNVs, this allows us to evaluate the utility of mitochondrial SNVs for lineage tracing of embryonic development. In general, I observed four distinct patterns of sharing of mitochondrial mutations. 1) Mitochondrial mutations present in a single clade of closely related samples. These represent genuine mitochondrial variants acquired during a shared trajectory of these samples. In our dataset, this exclusively occurs when crypts have undergone one or more fission events (**Fig. 3.13a**). 2) Mitochondrial mutations present in many (but not all) samples from the patient at a variable level. These samples do not belong to a specific clade, but span the entire phylogeny. Therefore, it is most likely these mutations were already present in the zygote, but at a heteroplasmic level (**Fig. 3.13b**). 3) Mitochondrial mutations that are shared between different samples or clades derived from the same biopsy. It is possible that different lineages within the same biopsy may acquire the same mutation. However, as these are mixed cell populations and these mutations are present at a low VAF (<0.10), it is more likely this sharedness reflects the sharing of a (sub)clone or stromal contamination (**Fig. 3.13c**). 4) Mitochondrial mutations that are recurrent in samples from different tissues. These are not caused by shared clones or stromal contamination and likely represent the independent acquisition of the same mutation in different lineages (**Fig. 3.13d**). Therefore, none of the called mitochondrial mutations provided further phylogenetic information on embryonic development.

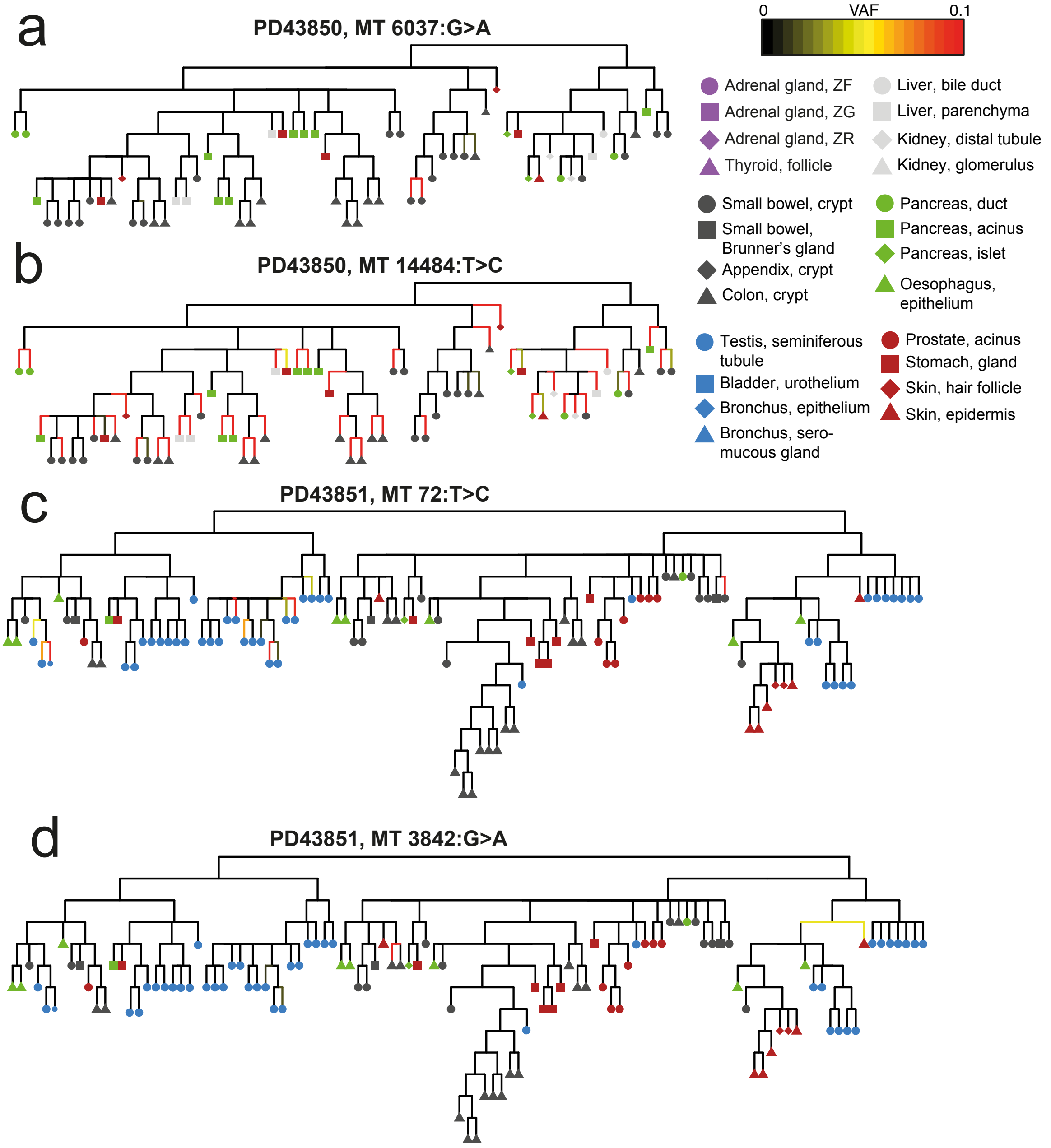

Figure 3.13 Phylogenies of nuclear SNVs with the VAF of mitochondrial mutations overlaid on them, showing a late shared SNVs (a), an SNV that was heteroplasmic in the zygote (b), an SNV that is consistent with a shared subclone or stromal contamination (c) and an SNV recurrently acquired in samples from different tissues (d)

Indels are acquired at a much lower rate than SNVs, often only at a tenth of the substitution rate. In addition, indel detection is less sensitive and specific, mainly due to their greater impact on read mappability. Hence, indel calls are frequently affected by high rates of false positives and shared artefacts. Calling and filtering of indels did not yield convincing embryonic variants nor improved the resolution of polytomies in the SNV tree. As a result, indels were disregarded in further analysis.

Copy number variants (CNVs) and structural variants (SVs) are large-scale chromosomal aberrations. Whole or partial loss of chromosomes has the potential to erase previously acquired somatic mutations and hence obscure the developmental life history of the sample. Within this cohort, the small number of CNVs and SVs were limited to individual samples and hence acquired later in life, rather than being embryonic in origin. The paucity of aneuploidies in normal tissues stands in stark contrast to the observation of widespread chromosomal abnormalities in the early embryo.

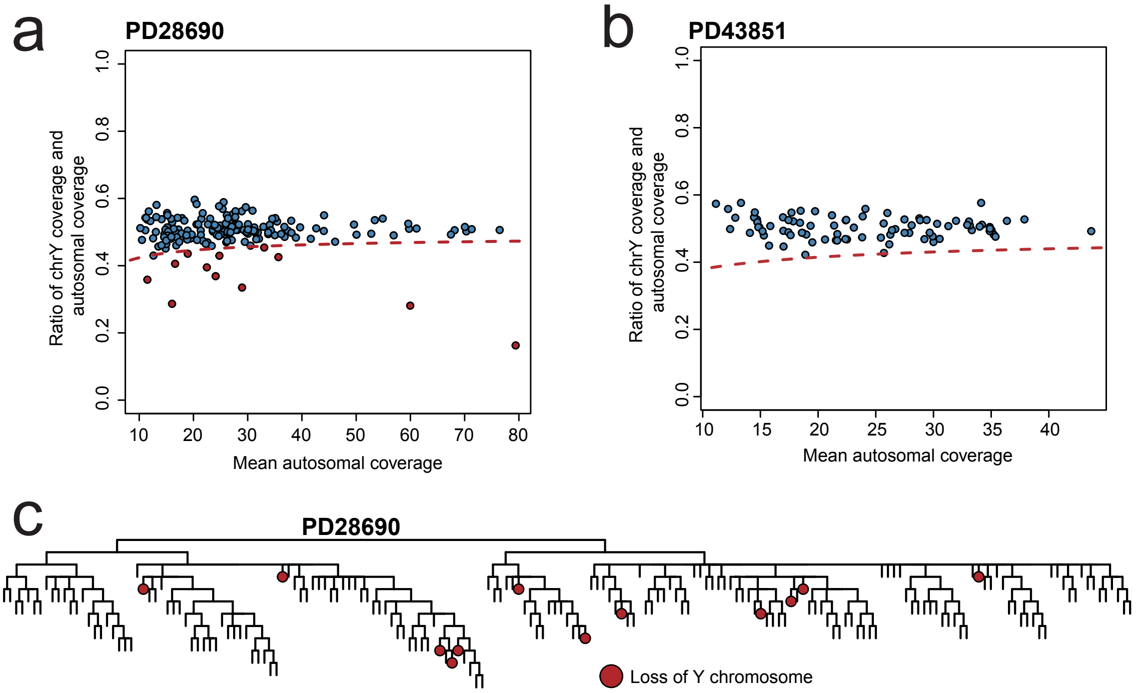

Figure 3.14 Scatterplot showing the ratio between the mean Y-chromosomal coverage and autosomal coverage against the mean autosomal coverage for all samples of PD28690 (a) and PD43851 (b). The dashed red lines indicates the 95% confidence interval around an expected ratio of 0.5. Red-coloured dots indicate samples with significant evidence of loss of the Y chromosome. (c) Phylogeny of PD28690 with samples exhibiting LOY marked in red, indicating all LOY events are acquired independently.

Interestingly, the only recurrent CNV that affected multiple tissues was a loss of the Y chromosome (LOY) in PD28690. Within this patient, 12 out of 187 (6%) LCM cuts exhibited a significant depletion of Y chromosomal coverage compared to autosomal coverage (**Fig. 3.14a**), indicating the presence of LOY. This LOY can affect a subset of cells within these samples or be pervasive throughout all cells in the microdissection. No LOY was detected in samples from the other male patient, PD43851 (**Fig. 3.14b**). Closer inspection of the LOY events in PD28690 revealed that all twelve instances represented independent chromosomal losses as none of the affected samples shared a private coalescent branch (**Fig. 3.14c**). Strikingly, LOY affected multiple samples of bladder urothelium, bile duct, the zona glomerulosa of the adrenal gland and renal distal tubules, hinting at a tissue-specific propensity for losing the entire Y chromosome. The incidence of LOY been found to increase with age (Jacobs et al., 1963) and has been associated with a wide variety of adverse health outcomes, including cancer (Forsberg et al., 2014), Alzheimer's disease (Dumanski et al., 2016) and all-cause mortality (Loftfield et al., 2018) and can be used as a proxy for genomic instability (Thompson et al., 2019).

3.12 Conclusion

The research presented in this chapter exemplifies the use of somatic mutations as markers for lineage tracing in human development. By constructing phylogenetic trees from samples of many normal tissues, it is possible to follow the life history of a cell from its first beginning as a zygote, through embryogenesis and early development, all the way to its final destination as a differentiated adult cell. Within this framework, it is possible to assess the quantitative contribution of individual embryonic progenitors to the developed adult body, showing a universal but variable degree of asymmetry. Combining the information from the presence of somatic mutations with the spatial information from sampling and microdissection can resolve the mosaic patterns that emerge from developmental processes, both on a microscopic scale as well as on an organ-wide level, such as the brain. Beyond the early development, this approach also allows for the research of later clonal expansions, many of which have an increased potential to transform into cancers, and other age-related genomic abnormalities, such as widespread loss of the Y chromosome.

This study can be seen as the next step in a long line of lineage tracing experiments in humans and model organisms, all contributing to our understanding of the fundamental human ontogeny. It is likely that future work will involve much higher numbers of samples for a single individual, possibly combining different approaches such as LCM and organoid generation to increase the number of different tissue types that can be used. Such a study can answer both qualitative and quantitative questions of cell dynamics and differentiation in later developmental stages, such as gastrulation and organogenesis, with much more granularity. Somatic mutations have the property of being natural, continuously acquired genetic scars and as such can be used for retrospective analysis. Hence, they can inform on intervals in human development that have been challenging to study, in particular the development of the embryo and fetus in the first trimester, without the need for embryonic or fetal material.

This chapter provides a fundamental background to the research presented in the next two chapters. An understanding of normal development is crucial to identify and appreciate processes that go awry during the early stages of life and potentially result in large embryonal clonal expansions. This will be the focus of Chapter 4. In addition, while the lineage tracing presented so far has yielded insights into early human development, the phylogeny has been confined to tissues derived from the embryo proper. This means that a substantial part of the phylogeny of the early embryo is missing. These missing lineages will have given rise to extraembryonic tissues, such as the yolk sac or placenta. The allocation of cells to the trophectoderm and inner cell mass is thought to cause the observed embryonic asymmetry, but so far this has only been inferred from inner cell mass-derived tissues. The research in Chapter 5 is based on sequencing a large number of human placentas, along with a representative tissue from the inner cell mass, to study the allocation of cells in the first differentiation event and directly observe roots of the embryonic asymmetry.

Chapter 4

Embryonal precursors of Wilms tumour

4.1 Introduction

The development of the human through the stages of zygote, embryo, fetus and adult organism constitutes myriad precisely orchestrated cascades of division, differentiation and migration. The analyses in the previous chapter leveraged somatic mutations to trace the journey of a normal cell from the fertilised egg to its eventual adult destination, exploring the developmental relationships between normal cells and tissues. The somatic mutations I used to infer cellular ancestries represented passive marks of cellular division with 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. Most notably, the sequential acquisition of such driver mutations can transform normal, well-functioning cells into tumour cells.

The introduction of errors into the genome happens continuously, but at a low rate such that the accumulation of the required set of driver mutations usually requires many decades. Childhood cancers, however, lack the long period of mutation accumulation prior to their emergence. In some childhood tumours an inherited germline mutation constitutes the founding driver mutation. A classic example of this is retinoblastoma and the gene that carries its name (*RB1*) (Knudson, 1971). Both copies of this tumour suppressor need to be inactivated to enable cancer formation. Frequently in these patients, the first hit is a disrupting mutation in this gene supplied by one of their parents, either acquired during their own lifetime or inherited in turn. This is effectively the first driver mutation in these children and severely increases the probability of malignancy through inactivation of the second copy. In fact, between 60 and 75% of children with a hereditary form of retinoblastoma develop tumours bilaterally (Knudson, 1971).

In addition to inherited drivers, somatic mutations arising early in development have the potential to act as a first hit in many cancers (Narod and Lenoir, 1991). Because such genomic alterations are automatically present in many, but not all cells of the individual, they can establish a mosaic predisposition throughout large parts of the body. Unlike inherited events, post-zygotic mutations have the ability to create a differential fitness landscape across different embryonic progenitors if the mutation in question has an oncogenic effect. In other words, if all cells carry a driver mutation, none of them have an advantage over one another. However, if only some cells harbour a genomic alteration that increase their fitness, they have the opportunity to outcompete their unmutated neighbours. In such a scenario, the imbalance of potency can potentially disrupt the physiological course of development and create large precursors lesions spreading through different organs.

Many paediatric cancers exhibit a close link to developing cells. While adult cancers might hijack mechanisms of early development, in the form of dedifferentiation and replicative immortality, paediatric tumours appear to be a consequence of cells in arrested development, which are unable to differentiate. This notion is mainly derived from the histologically undifferentiated appearance of paediatric tumour cells. More recently, this has been fortified by transcriptomic studies, characterising childhood cancer cells as having a fetal potential akin to developing cells (Young et al., 2018). In addition, paediatric malignancies have a distinct underlying suite of potential driver genes, many of which have functions intricately linked to development and differ from those found in adult cancers (Grobner et al., 2018).

In a similar vein to clonal haematopoiesis, paediatric cancers might arise from precursor clones that have their origin in aberrant homeostasis or development. As with normal development, somatic mutations can be used to retrace the shared ancestry of tumour and surrounding normal cells. In addition, the possibility of an early driver might be confirmed by such a lineage tracing exercise.

Although a deep phylogeny needs many single cell-derived readouts of somatic mutations, the question of the relationship between tumour and normal tissue can be answered without the need to experimentally obtain such samples. The cancer itself represents a single clonal lineage, from zygote to the founder cell of the tumour. If the mutations that delineate the developmental trajectory of the cancer are also found in normal cells, they will have shared that part of development with the tumour. The proportion of such normal cells harbouring this shared ancestry is naturally encoded in the VAF of these mutations. Therefore, large clonal expansions can be picked up in traditional bulk DNA sequencing experiments with relative ease.

Among the childhood cancers conventionally considered embryonal, i.e. morphologically resembling fetal tissue, is Wilms tumour, also known as nephroblastoma. It is the most common renal cancer in children (Breslow et al., 1993). Most cases of Wilms tumour will manifest as sporadic, unilateral tumours. Early studies assumed that bilateral Wilms tumours originate in much the same fashion as bilateral retinoblastoma, i.e. as a consequence of an inherited predisposing first hit in a tumour suppressor (Treger et al., 2019). Intriguingly, a large French epidemiological study found that a family history of Wilms tumour was no more common in bilateral than in unilateral cases (Bonaiti-Pellie et al., 1992), raising the possibility that an early, post-zygotic event rather than an inherited driver mutation might play a prominent role in carcinogenesis. In order to investigate this hypothesis, I used somatic mutations derived from multi-site biopsies to determine the phylogenetic relationships between Wilms tumour and normal tissues (kidney and blood).

4.2 Detecting early clones in normal kidneys

The study of somatic mutations in early embryonic development from the previous chapter has revealed that large, polyclonal aggregates of cells, such as bulk biopsies, generally obey the same early asymmetries as the whole body. No organ or tissue is derived from a single precursor later than the most recent common ancestor of the entire body, which I presume to be the zygote. Therefore, if the presence of certain mutations and their VAFs differ between different bulk biopsies, the possibility arises that these represent large, aberrant clonal expansions.

To investigate patterns of mutation sharing, the Wilms tumour(s) of 23 patients in total were subjected to whole-genome sequencing, along with at least one sample of normal kidney and blood. For eight of the patients, blood samples from both parents were sequenced as well in order to evaluate and identify any deleterious *de novo* germline mutations. In one case, it was possible to sample a nephrogenic rest, a benign lesion sharing some morphological features with Wilms tumour. Calling and filtering of single nucleotide variants (SNVs) was performed in an unmatched fashion, as described in Chapter 2, section 2.2.

The initial discovery cohort consisted of three children that had unilateral Wilms tumour. The unmatched variant calling revealed a subset of mutations found in tumour that were present across all normal samples as well. These SNVs are presumed to represent the post-zygotic mutations that are acquired during the first few cell divisions of life (early embryonic mutations). However, in two out of three cases, I also identified SNVs that were shared between one or more renal samples and the Wilms tumour, but absent from the blood. This is best illustrated by PD37272, from whom renal pelvis, medulla and cortex were sampled (**Fig.**

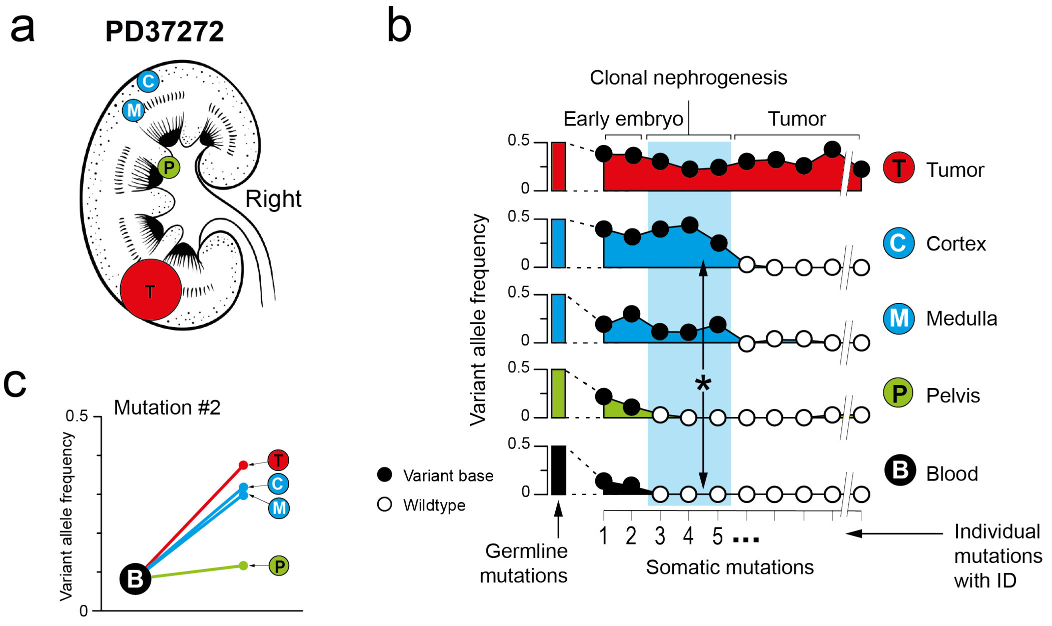

Figure 4.1 (a) Overview of tissue sampling in the kidney of PD37272. (b) Classification and overview of early somatic mutations. If the mutation is present in tumour, kidney, and blood, it is classified as early embryonic (mutation #1-#2). If it is present in kidney samples and tumour only, it is clonal nephrogenic (mutations #3-#5, marked by an asterisk). If it is only in the tumour, it is labelled as such. White and black circles indicate whether the observed VAF is insignificant (white) or significant (black), $p < 0.001$ (test of presence using beta-binomial overdispersion, Methods) (c) The VAF for the last embryonic mutation in kidney samples and tumour compared with blood. From Coorens et al. (2019). Reprinted with permission from AAAS.

4.1a). Intriguingly, three SNVs were shared between the cortex, medulla and tumour that were absent from pelvis and blood (**Fig. 4.1b**). In this case, blood was deeply sequenced, to a genome-wide coverage of 106x, thus minimising the probability of missing the variant by chance. The VAF of one of these mutations was as high as 0.44 in renal cortex, suggesting that a remarkable 88% of cells in the biopsy were derived from the progenitor cell that acquired that SNV. I use the term 'clonal nephrogenesis' to describe this phenomenon in which a substantial proportion of normal kidney cells derives from a single cell that existed later than the zygote. Beyond the pattern of sharing between different tissues, the VAF of embryonic mutations is higher in cortex and medulla, which contain the clonal nephrogenesis, further confirming a more pronounced shared ancestry between these biopsies and the tumour (**Fig. 4.1c**).

It is conceivable that these variants shared between normal samples and the cancer could represent contaminating tumour cells in the normal kidney that are not visible on histology sections or due to cross-contamination of DNA during extraction, library preparation or sequencing. This explanation is implausible, as contamination would manifest as a sharing of all tumour mutations at a VAF consistent with the contamination rate. Given the absence of any morphological evidence of tumour contamination on histology, the contamination rate would have had to be low, inconsistent with the observed high VAF of shared mutations. Moreover, the VAF of these clonal nephrogenesis mutations gradually decreased along an anatomical gradient. This pattern is much more consistent with a developmental history of clonal nephrogenesis. I also further statistically excluded the possibility of infiltration or contamination by tumour DNA using a binomial mixture model applied to all normal samples. This approach is outlined in the Chapter 2, section 2.3.3.

To evaluate whether clonal nephrogenesis is a frequent antecedent of Wilms tumour, another 20 cases were studies. Four of these cases were bilateral Wilms tumours, for which bilateral biopsies of normal kidney tissue were also obtained. For one of the 16 cases with unilateral Wilms tumour, it was possible to sequence ten biopsies of five normal renal tissues **Fig. 4.5**). As before, I identified SNVs constituting the tumour lineage and evaluated their VAF in the normal biopsies. If there was at least one mutation present in a renal sample and tumour, but absent from blood, I counted the case as one harbouring clonal nephrogenesis.

Within the entire group of 23 patients (both the discovery and validation cohort), I identified mutations signifying clonal nephrogenesis in 10 of 19 children with unilateral disease (53%) and in all four children with bilateral cancers (**Fig. 4.2a**). The presence of such nephrogenic clones was confirmed by a significant ($p < 0.01$, Wilcoxon signed-rank test) inflation of VAFs of early post-zygotic mutations (**Fig. 4.2b**). Importantly, none of the normal renal samples harboured any copy number variants, which might alter the VAFs of such early embryonic mutations.

Collectively, these observations hint that the renal cortex and medulla are derived from a pool of progenitors more closely related to the tumour than renal pelvis and blood. The developmental segregation between the corticomedullary lineage and the blood and renal pelvis is known. Both blood and kidney cells are mesodermal in origin, but segregate from one another soon after gastrulation (Barresi and Gilbert, 2020). The renal pelvis is derived from the ureteric bud, unlike the renal corticomedullary lineages, which have their origin in the nephrogenic cords. This course of development provides an opportunity for a genomic alteration in the corticomedullary lineage to cause a localised clonal expansion without affecting the blood and renal pelvis. However, it is unclear whether this pattern of early SNVs genuinely represent a *bona fide* aberrant nephrogenic clone or is a consequence of natural bottlenecks during the development of the urinary tract. In other words, is clonal nephrogenesis a ubiquitous phenomenon also prevalent amongst patients without Wilms tumours or other renal malignancies?

4.3 Nephrogenic clones are exclusive to individuals with Wilms tumour

Whether clonal nephrogenesis is an aberrant feature of kidneys harbouring Wilms tumours or whether it represents the normal clonal architecture of human nephrons is unclear at this point. This was investigated further using three approaches: by looking at the clonal architecture of normal human kidneys by laser capture microdissection (LCM), whether patients with other paediatric or adult kidney cancer share this phenomenon, and whether the distribution of non-tumour variants in Wilms tumour kidneys is distinct from those derived from other kidneys.

First, to assess whether the normal human kidney has clonal units, glomeruli (n = 7) and proximal and distal tubules (n = 15) were excised from the kidneys of three individuals (**Fig. 4.2c**): the warm autopsy patient described in the previous chapter (PD28690) and two patients who underwent nephrectomies for clear cell renal carcinoma (ccRCC). Neither tubules nor glomeruli had a VAF distribution centred above 0.25 (**Fig. 4.2d,e**), signifying that they do not generally represent monoclonal units as seen in, for example, endometrial glands or colonic crypts. This fact also becomes apparent when taking into account the experiment outlined in the previous chapter, where the majority of renal samples were of an insufficient level of monoclonality to be included in the phylogeny reconstruction.

Second, I assessed whether mutations were commonly shared between other renal tumours and surrounding normal kidney tissue. I studied childhood congenital mesoblastic nephroma (CMN; two tumours and six normal kidney samples), childhood renal malignant rhabdoid tumour (MRT; one cancer and one normal kidney sample), and adult ccRCC (eight cancers, including one bilateral case, and 15 normal tissue samples).

Applying the same unmatched variant calling strategy as before, I sought early post-zygotic mutations shared between tumour, normal kidney tissues, and blood. However, none of these cases harboured mutations that were shared only between neoplastic and normal renal samples but absent from blood (**Fig. 4.2f**). This shows that this pattern of mutation sharing, i.e. clonal nephrogenesis, is specific for Wilms tumours and highly prevalent in this group ($p<0.001$, Fisher's exact test). The absence of clonal nephrogenesis in CMN and MRT may be attributable to their origins in different developmental lineages within the kidney. However, ccRCC and Wilms tumour are both thought to arise from proximal tubular cells (Hohenstein et al., 2015; Kovacs et al., 1997; Treger et al., 2019). If normal embryological clonal dynamics typically generated large clonal expansions, I would have expected to find clonal nephrogenesis in ccRCC cases as well, since the emergence of the tumour in adulthood would not obliterate the developmental history of surrounding normal cells.

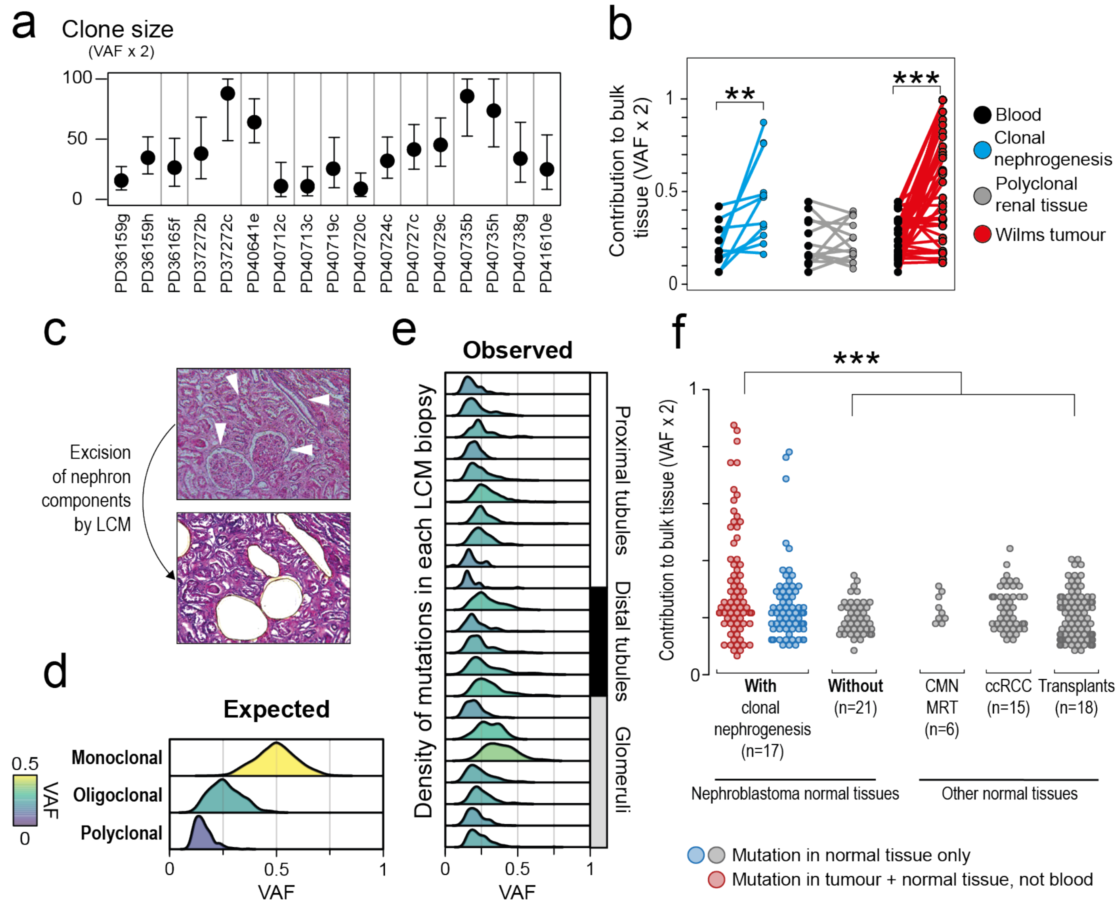

Figure 4.2 (a) Sizes of nephrogenic clones in normal renal samples as predicted by twice the VAF of the most prominent nephrogenic mutation. (b) Plot showing the contribution of the last embryonic mutation in tumour (red) and in samples with (blue) and without clonal nephrogenesis (grey), alongside the contribution to blood (black). The increase was significant in clonal nephrogenesis and tumour samples, but not in renal tissues without clonal nephrogenesis (**p<0.01;***p<0.001, Wilcoxon signed-rank test). (c) Histology images showing components (arrow heads) of the human nephron excised by LCM. (d) VAF simulations to derive expected distributions depending on clonality of a tissue; monoclonal origin (peak VAF 0.5), oligoclonal origin (peak VAF 0.3), or polyclonal origin (peak VAF 0.1). (e) VAF distributions for 22 microdissected samples (10 proximal tubules, 5 distal tubules, and 7 glomeruli) from 3 patients, 1 rapid autopsy donor and 2 ccRCC patients. Colour indicates the underlying maximum likelihood peak VAF as predicted by a truncated binomial mixture model (see Methods). (f) Mutations present in samples obtained from normal kidneys but absent in matched blood. Only in Wilms tumour were some of these mutations shared with the corresponding tumour. In the presence of clonal nephrogenesis, the VAF distribution of these mutations was significantly elevated (***p<0.001, Wilcoxon rank-sum test). From Coorens et al. (2019). Reprinted with permission from AAAS.

As a third approach, I interrogated all mutations of normal kidney tissues listed thus far, supplemented by an additional 18 biopsies obtained from bilateral kidneys that had been declined for transplantation. I analysed somatic mutations present in kidney tissue and absent from non-renal tissue, irrespective of whether they were shared with tumours. Collectively, these analyses of 77 normal kidney biopsies revealed that variants of tissues without clonal nephrogenesis have a significantly lower VAF distribution than clonal nephrogenesis mutations ($p<0.001$, Wilcoxon rank-sum test, **Fig. 4.2f**). Many of the SNVs absent in tumour but present in kidneys containing clonal nephrogenesis will have been generated by alternative lineages after the initial clonal expansions. These are cells that are descended from the initial founder cell but have split from the lineage that will have generated the eventual tumour founder.

Taken together, these results indicate that these nephrogenic clones represent an abnormal state of kidney development, intimately associated with Wilms tumour pathogenesis.

4.4 The driver of clonal nephrogenesis

So far, I have identified the peculiar phenomenon of early embryonal clonal expansions that frequently precede Wilms tumour development. An explanation of genetic drift passively causing such early expansions is unlikely given the specificity to Wilms tumour patients and the short time frame in which genetic drift would need to act to manifest in childhood. In this instance, a more likely explanation for a cell generating a clonal expansion is the acquisition of a selective advantage. In other words, the founder cell must obtain a gain of fitness in order to have a relative advantage over unmutated cells. Therefore, to understand the mechanism of clonal nephrogenesis, it is essential to identify the potential driver event causing this abnormal feature of development.

Surprisingly, almost all SNVs identified as being part of nephrogenic clones fell in non-coding regions (64 out of 66).The remaining two mutations did fall in gene regions and generated a missense change, but in genes (*REXO1* and *R3HDM2*) that have not been implicated in carcinogenesis. These missense mutations were predicted to be benign according to the Variant Effect Predictor algorithm (McLaren et al., 2016). Moreover, none of these SNVs recurred in a similar site or region in the genome, strongly hinting that either a diverse range of genomic changes could generate clonal nephrogenesis, or, perhaps more plausibly, that all the identified genetic alterations were passenger mutations. In the case of the latter, the driver event at the root of clonal nephrogenesis may not be a genetic change.

Whole-genome sequences of Wilms tumour and normal samples were further supplemented by methylation data from arrays and by RNA sequencing to assess the transcriptomic

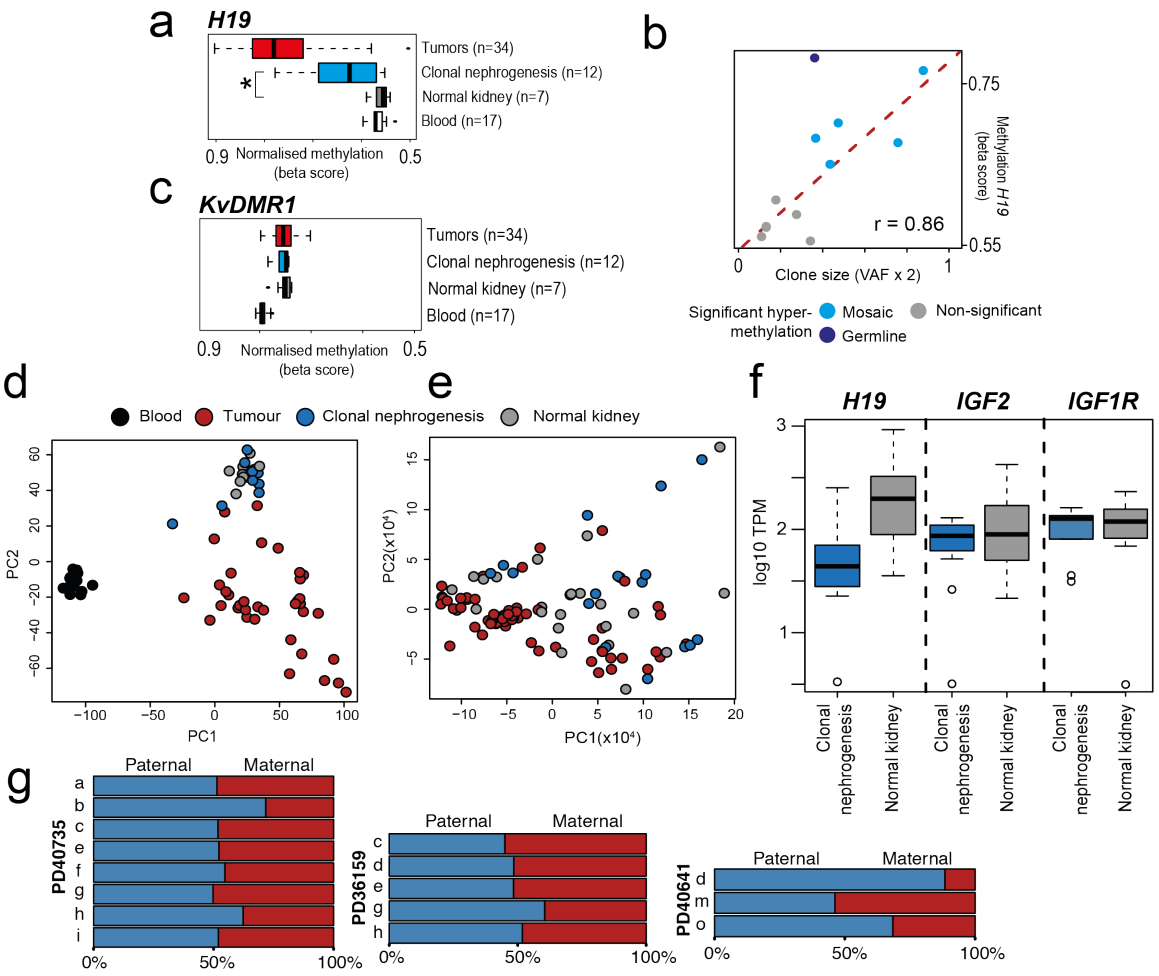

Figure 4.3 (a) Group-level methylation beta values of *H19* (*p<0.05, Wilcoxon rank-sum test). (b) Relationship between predicted clone sizes from nephrogenic mutation and the methylation level of *H19*. The dark blue dot represents PD40738g, which is affected by germline H19 hypermethylation (omitted from correlation and linear regression). The light blue dots indicate samples with clonal nephrogenesis and significant deviation from the background methylation distribution of *H19* as obtained from normal kidney samples without clonal nephrogenesis. Grey dots indicate significant deviation from the background. (c) Group-level methylation levels of KvDMR1 (non-significant, Wilcoxon rank-sum test). Plots of principal component analysis of methylation data from arrays (d) and transcriptome data from RNA sequencing (e). (f) Expression levels of *H19*, *IGF2* and *IGF1R* in kidney samples with and without clonal nephrogenesis. (g) Parental allele-specific expression pattern of *IGF2* in patients with *H19* hypermethylation. From Coorens et al. (2019). Reprinted with permission from AAAS.

profiles of these samples (see Chapter 2, section 2.5.2). Strikingly, I found significant hypermethylation of the *H19* locus in seven out of twelve normal kidney samples containing nephrogenic clones (**Fig. 4.3a,b**). *H19* hypermethylation was absent from blood and other non-clonal renal tissues, bar the blood of a child with Beckwith-Wiedemann syndrome (PD40738), often caused by this hypermethylation body-wide (Weksberg et al., 2010). Hypermethylation of *H19* is a well-established driver event in Wilms tumour pathogenesis (Charlton et al., 2015; Moulton et al., 1994; Okamoto et al., 1997). The *H19* locus itself encodes a long non-coding RNA that suppresses the expression of growth-promoting genes, such as *IGF2*, that all reside in close proximity to the locus itself, on chromosome 11p15. Furthermore, the *H19* locus falls within one of the few strongly imprinted regions in the human genome (Giannoukakis et al., 1993; Zhang and Tycko, 1992). Normally, the maternal copy of this region is unmethylated, while the paternal one is always methylated. Hence, this hypermethylation event is more appropriately described as loss of imprinting (LOI).

The degree of hypermethylation of *H19* strongly correlated with the VAF delineating the nephrogenic clones (**Fig. 4.3b**), indicating that hypermethylation was present in the founding cell of the clone and pervaded the clone in its entirety. In the five samples with clonal nephrogenesis, but without significant *H19* hypermethylation, it is likely that the size of the clone was small enough to preclude detection of such focal LOI by methylation arrays. Alternatively, I cannot exclude the possibility that these clones harbour distinct, unrecognised genetic or epigenetic driver events.

In addition to *H19*, hypermethylation of the *KvDMR1* locus (chromosome 11p15.5) is also able to cause Beckwith-Wiedemann (Weksberg et al., 2010). However, in those cases, the predisposition to Wilms tumours is only minimal (Treger et al., 2019). Interrogating the methylation status of the *KvDMR1* locus, I determined that its imprinting remained intact in clonal nephrogenesis (**Fig. 4.3c**). This further substantiates the specificity of the LOI as an epigenetic mutation to generate Wilms tumours, effectively by a mosaic version of the Beckwith-Wiedemann syndrome.

Besides LOI of *H19*, I was unable to identify any further driver events accounting for clonal nephrogenesis, despite using whole-exome sequencing to re-interrogate coding mutations in 15 of 17 tissues containing nephrogenic clones. Global gene expression profiles, including the prevalence of fetal transcripts, did not differ between normal renal tissues that did or did not display clonal nephrogenesis (**Fig. 4.3d**). Similarly, global methylation patterns did not differ between these two groups (**Fig. 4.3e**). However, while the expression of *H19* differed between renal samples with and without clonal nephrogenesis, the level of expression of *IGF2* and *IGF1R* remained unchanged between these groups (**Fig. 4.3f**). By assessing the prevalence of parent-specific SNPs in the *IGF2* locus, we could determine that both parental alleles of *IGF2* were expressed in samples with hypermethylation of *H19* (**Fig. 4.3g**). This is in contrast to cells with proper imprinting of this locus, which should show monoallelic expression. This hints at the ability of the cells with LOI of *H19* to eventually neutralise the up-regulation in *IGF2*-signalling and effectively overcome the overgrowth syndrome.

4.5 Timing the early expansion

From the large contribution of nephrogenic clones to normal renal tissues, we can deduce that the hypermethylation of *H19* must have happened during the development of these organs. The exact timing of this event, however, is much more challenging to pinpoint. In two cases of bilateral Wilms tumour (PD36159 and PD40735) it is noteworthy that the nephrogenic clones span both kidneys, as left and right tumour and normal samples all share clonal nephrogenesis as a common ancestor (**Fig. 4.4a,c**). Therefore it stands to reason that in these cases the LOI of *H19* must have happened prior to the segregation of both kidney primordia, soon after gastrulation (Short and Smyth, 2016).

The observation that clonal nephrogenesis must have been an early event is further reinforced by the size of the clone in the normal kidney samples. In PD40745, two mutations delineating clonal nephrogenesis are found in both kidneys, accounting for 63% of cells on the left and 86% of cells on the right (**Fig. 4.4b**). Strikingly, the mutations identified next separate into branches specific to the right and left kidney, raising the possibility that this lateral split was established a few cell divisions after the original expansion. The size of these clones signifies a remarkable deviation from the early embryonic asymmetry seen in blood, where the first mutation giving rise to kidney and Wilms tumour only accounts for 20%-25% of cells in blood. The remaining 75%-80% of blood cells is delineated by an early embryonic mutation only accounting for 10-20% in normal kidney samples. This reversal of early embryonic asymmetry underscores how aberrant these expansions truly are, particularly, when compared to patterns found the individual in Chapter 3, in whom the early embryonic asymmetry is maintained across tissues throughout a 78 year lifespan.

In another patient with bilateral Wilms tumour, PD40378, all five left tumour samples, but not the right tumour, were related to a clonal expansion in the left kidney (**Fig. 4.4e**). However, it is of note that this patient was diagnosed with a germline Beckwith-Wiedemann syndrome. Notably, the tumour on the right did not appear to arise from an early clone. It is possible that the precursor clone for this cancer only resided in the right of the kidney, from which no normal tissue was available for sampling. The finding that a germline LOI of *H19* grants the potential imbalance to generate precursor clones is surprising, since presumably all cells would carry the hypermethylation and thus have an equal fitness. The mechanism underpinning these clonal expansion therefore remains unclear in this instance.

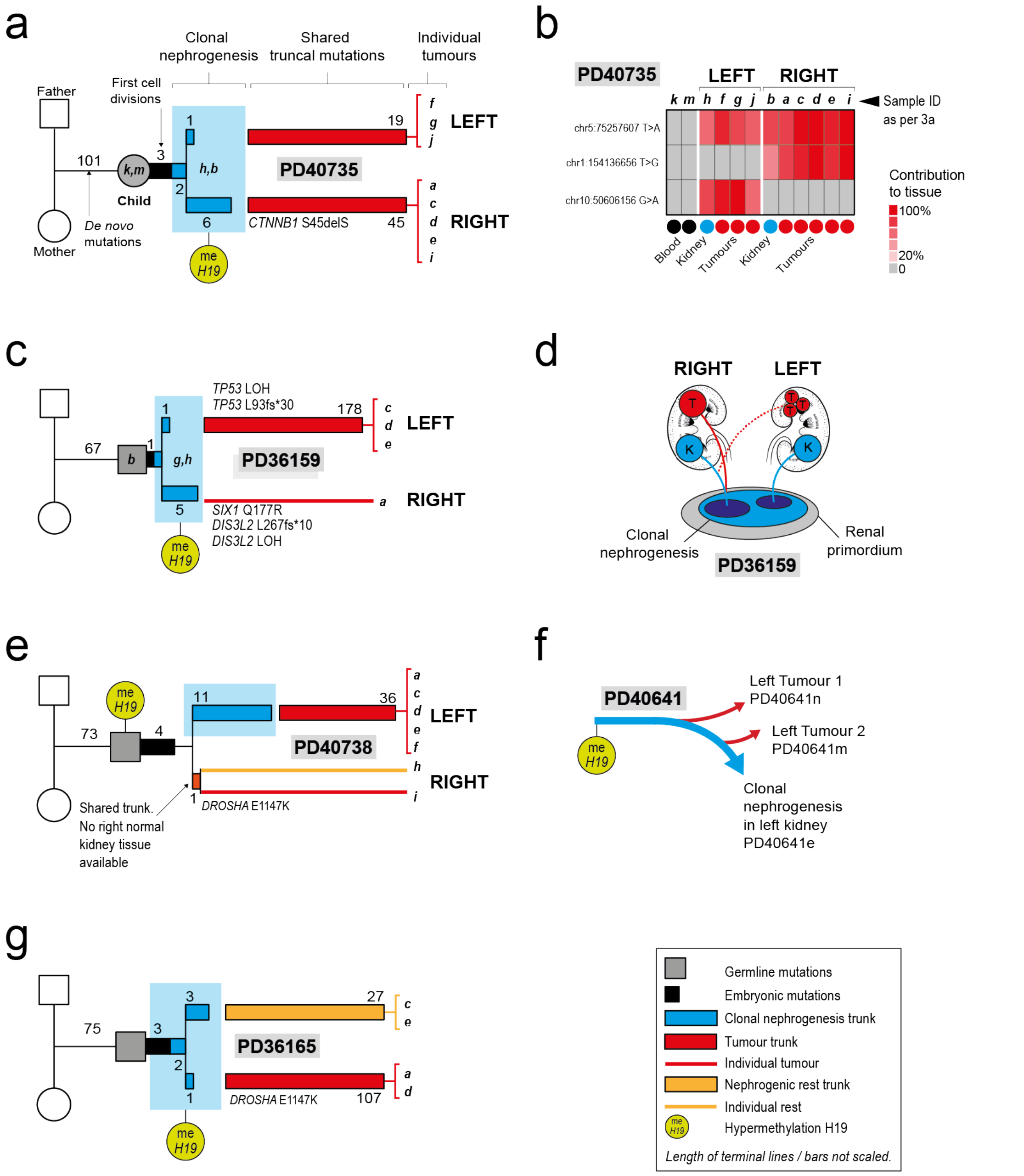

Figure 4.4 (a,c,e,g) For each tumour, the phylogeny is shown including *de novo* germline mutations, embryonic mutations, mutations demarcating clonal nephrogenesis, and tumour mutations. Numbers refer to the number of substitutions defining each developmental branch. Truncal driver events are detailed. (b) Heatmap showing the contribution of a mutation to a sample (as per legend). The pattern of shared mutations reveals a split between left and right kidney, in both tumour and normal samples. (d) As revealed by the shared mutations, the left tumour is more closely related to the right branch of clonal nephrogenesis than to the left in PD36159. (f) Two mutations indicate the independent emergence of tumours at different time points from the nephrogenic clone in PD40641. (g) Tumour and nephrogenic rest in PD36165 both originated from clonal nephrogenesis despite being situated at opposing kidney poles. From Coorens et al. (2019). Reprinted with permission from AAAS.

In the case of unilateral tumours, the timing of the initiation of clonal nephrogenesis remains unclear and is more difficult to establish. It may have evolved before the kidney was formed or thereafter, followed by a "clonal sweep" of clonal nephrogenesis replacing kidney tissue. However, it is probable that the size of the nephrogenic clone hints at the original timing of the *H19* LOI. In PD37272, the SNV demarcating the expansion accounts for 88% of the cortex sample, while also being pervasive in the medullary biopsy. Of course, whether this clone is also present in the contralateral kidney is unknown. Nevertheless, this is in clear contrast with PD41750, where the nephrogenic clone was only found in one cortical sample, out of the ten normal renal samples in total (**Fig. 4.5**). In this cortical sample, the clone accounted for only 25% of cells. It is conceivable that this means the hypermethylation of *H19* occurred at a later point in development and therefore the clonal expansion is only prevalent locally.

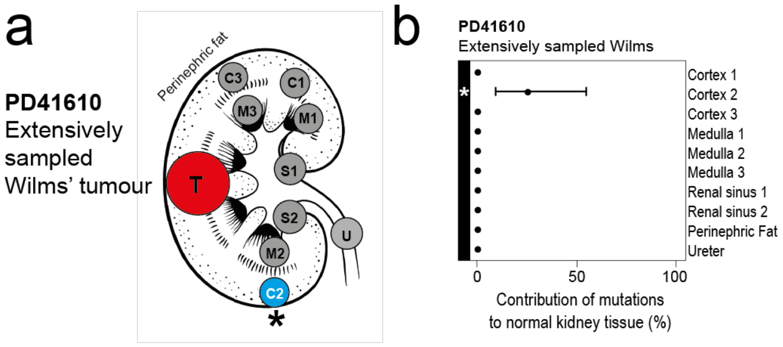

Figure 4.5 (a) Sampling overview for PD41610, the extensively sampled nephroblastoma kidney. This patient shows localised clonal nephrogenesis as depicted by the estimated nephrogenic clone size (twice the VAF) only contributing to a single renal cortical sample (b), highlighted in blue and marked by an asterisk in (a) and (b). From Coorens et al. (2019). Reprinted with permission from AAAS.

The time at which imprinting of *H19* was lost will have a direct impact not only on the predisposition to Wilms tumours, but to other embryonal cancers as well. Hepatoblastoma is the second most common cancer associated with germline Beckwith-Wiedemann, and it appears probable that it can arise as a consequence of mosaic overgrowth as well. More rarely, rhabdomyosarcoma, neuroblastoma and adrenocortical carcinoma are associated with this overgrowth syndrome as well. In a subset of cases, the timing of the emergence of the nephrogenic clone might very well be prior to the commitment of the cell to renal development. In fact, if *H19* imprinting is lost in a cell prior to gastrulation, it is conceivable that the predisposing effect might be mosaic across germ layers, generating a field effect of cancer risk in more than one organ.

4.6 Sufficiency of *H19* hypermethylation

I have now established that a significant proportion of clonal nephrogenesis is most likely driven by an early cell losing the imprinting pattern of *H19*, which causes expansions of these cells through lost suppression of *IGF2*-mediated signalling. This results in large sections of renal tissue effectively harbouring a first driver in the transformation of normal kidney cell into Wilms tumour. However, the sequencing data from tumours reveals that in the majority of cases, the Wilms tumour acquired more drivers prior to the formation of the malignancy, while no additional drivers were identified in the normal tissue. A driver mutation in *CTNNB1* was found in three tumours, while the *DROSHA* E1147K mutation was found in two cases. Other identified drivers include hits in *TP53* (followed by loss of heterozygosity), *WHSC1* and *SIX1*. In five cases with clonal nephrogenesis, I did not identify any additional driver events beyond the LOI of *H19*, raising the possibility of cryptic genetic driver events (Martincorena et al., 2017) or other epigenomic alterations that might have triggered the formation of the cancer.

Insight into the necessity of additional drivers is perhaps gained best by comparing the genome of the benign nephrogenic rest to the Wilms tumours in PD36165 and PD40738 (**Fig. 4.4e,g**). In these cases, both nephrogenic rest and tumour originate from a common nephrogenic clone. Neither of the nephrogenic rests carry any plausible driver mutation in addition to LOI of *H19*. However, in both cases, the cancer on the ipsilateral side of the nephrogenic rest only differs from the benign lesion by the acquisition of the *DROSHA* E1147K. This suggests that this additional driver is necessary to transform a nephrogenic rest into a full-blown Wilms tumour.

Two tumours found on the left side of PD40641 were revealed to have originated independently from two related but different cells belonging to the same nephrogenic clone (**Fig. 4.4f**). This difference was established by two SNVs only, indicating the transformation events must have occurred soon after one another. Nevertheless, this recurrence points at a sustained potential of the nephrogenic clone to spawn Wilms tumours.

A particularly intriguing aspect of clonal nephrogenesis is the large proportion of morphologically and functionally normal cells carrying the hypermethylation of *H19*. Since there appears to be no genomic difference between these cells and nephrogenic rests, there might be a mechanism that ensures the normal and regular differentiation of these cells into function units of the kidney. After all, nephrogenic rests are essentially clusters of undifferentiated renal cells residing in the kidney. The question arises whether differentiated renal cells with *H19* hypermethylation retain the potential to transform into Wilms tumours or whether only nephrogenic rests have such a capability.

Of note, the incidence of Wilms tumour plummets after age 6 and is essentially zero beyond age 10 (Breslow et al., 1988). If clonal nephrogenesis represented a lifelong predisposition to nephroblastoma, such tumours would continue to appear throughout the entirety of childhood and long after. However, the absence of such prolonged increase in risk suggests that the predisposing effect of mosaic *H19* LOI is transient. In other words, the epigenetically primed "neoplasia-ready" (Feinberg et al., 2006) cells of the nephrogenic clone lose their malignant potential over time. This is further reinforced by the lack of a clear distinction of the global transcriptome and methylome between normal renal samples with and without clonal nephrogenesis, as mentioned before. This suggests that over time even cells carrying this overgrowth driver somehow repress its effect and become indistinguishable, both in terms of their methylation and expression landscape, from renal cells that never lost this imprinting to begin with. The exact mechanism by which these precursors differentiate over time remains unknown. It is conceivable that this occurs via the conventional pathways operating in renal development, but at a lower rate due to the imprinting loss. In such a scenario, over years, one would only be left with pockets of undifferentiated cells, in essence nephrogenic rests, which slowly disappear with age. With differentiation and specialisation, the carcinogenic potential of these cells is lost, perhaps since these cells no longer rely on the H19-IGF2 pathway.

4.7 Loss of imprinting versus loss of heterozygosity

LOI of *H19* appears to lie at the heart of clonal nephrogenesis, but loss of heterozygosity (LOH) of the same locus, while having the same genomic effect, does not seem to be the dominant pathway in these early expansions. None of the normal renal samples with detected clones have any copy normal abnormalities, including 11p LOH. Moreover, out of the 18 unique tumours originating from a nephrogenic clone, only one also exhibited 11p15 LOH (**Fig. 4.6**). However, of the nine tumours not preceded by clonal nephrogenesis, five exhibited LOH of 11p15. This difference is significant (p<0.01, Fisher's exact test) and indicates that LOH and LOI of 11p15 are generally alternative pathways to generating Wilms tumours. In total, only four of 27 unique tumours had neither LOI nor LOH of *H19*. This high prevalence demonstrates that dysregulating the imprinting pattern of the *H19* locus is a key driver of Wilms tumour.

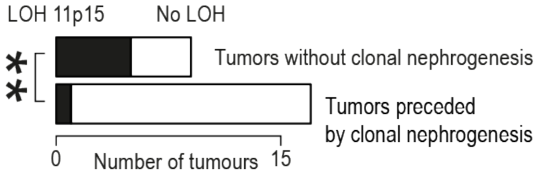

Figure 4.6 Comparison of incidence of copy neutral LOH of 11p15 in this nephroblastoma cohort. Out of the 18 unique tumors that originate from a nephrogenic clone, only one exhibits LOH of 11p15. Out of the nine tumours that did not have such a clone, five have 11p15 LOH (**p<0.01, Fisher's exact test). From Coorens et al. (2019). Reprinted with permission from AAAS.

Recently, we have observed cases of Wilms tumour with paired normal kidney cells that exhibit uniparental disomy of 11p, clearly showing clonal nephrogenesis can be driven by LOH of *H19*. However, the difference in prevalence is striking. Even though both LOI and LOH achieve the same phenotypic result, the underlying rate or tolerance of the different mutational processes, an epimutation versus mitotic recombination, might explain why these mechanisms predominate at different developmental stages. The observation that all bilateral tumours originated through LOI of *H19* and not LOH, and that we can, in a subset of these, time the emergence of the nephrogenic clone to early embryogenesis supports this hypothesis. A higher prevalence of LOI in the early embryo might be expected during its remodelling of the methylation landscape after the first few divisions of life (Eckersley-Maslin et al., 2018). It is possible that an imprinting error at this time happens more frequently than uniparental disomy from mitotic recombination, rendering LOI the commoner pathway to generate embryonal precursors to Wilms tumour.

4.8 Conclusion

The findings presented in this chapter show that early clonal expansions in histologically and functionally normal kidney tissue are an aberrant outcome of renal development that commonly antedates Wilms tumour. These nephrogenic clones are the consequence of hypermethylation of *H19*, effectively producing the phenotype of a mosaic Beckwith-Wiedemann syndrome. Clonal nephrogenesis has the ability to generate histologically and functionally normal kidney cells, which occupied the majority of renal biopsies in the most pronounced cases.

The extent of clonal nephrogenesis might be a marker of malignant potential and inform on the risk of cancer occurrence, at least in childhood, presumably while pockets of undifferentiated precursor cells remain sown in the normal tissue bed. This information could potentially be used to guide treatment of Wilms tumour patients and surveillance for relapse. Moreover, if it were possible to manipulate the neoplastic potential of clonal nephrogenesis by inducing differentiation, prevention of Wilms tumour in could become feasible. Conceivably, this would also apply to individuals diagnosed with full *H19*-driven Beckwith-Wiedemann syndrome.

Collectively, these findings demonstrate that Wilms tumour frequently represents an insurrection on the background of a premalignant tissue bed, rather than a clearly demarcated neoplasm in an otherwise normal polyclonal kidney. It is highly likely that embryonal clonal expansions, possibly also driven by epigenetic mechanisms rather than genetic changes, may be a common phenomenon in the emergence of childhood cancers.

Chapter 5

Universal mosaicism of the human placenta

5.1 Introduction

In the previous chapters, I have established that somatic mutations serve as a readout of the developmental history of the cells they are present in. I have shown the recapitulation of early embryonic dynamics in normal human development, as well as early clonal expansions as precursors to Wilms tumours. However, all research up until this point was performed on tissues and cells derived from the embryo proper, a logical consequence of confinement to postnatal sampling. Many of the early lineage commitments during embryogenesis will pertain to extraembryonic tissues, derived from the trophectoderm, hypoblast or amnion. In fact, our current explanation for the observed asymmetric contribution in the first cell division is the cell allocation to inner cell mass and trophectoderm. Therefore, the final study presented in this dissertation aims to further elucidate cell dynamics in embryogenesis by sequencing placental biopsies in conjunction with the umbilical cord.

Besides answering questions about early embryogenesis, lineage tracing within the placenta might also reveal its natural course of development. The human placenta is a temporary organ whose dysfunction contributes substantially to the global burden of disease (Brosens et al., 2011). Among its many peculiarities is the occurrence of chromosomal aberrations confined to the placenta, which are absent from the new-born infant (Kalousek and Dill, 1983). Confined placental mosaicism affects one to two percent of pregnancies (Hahnemann and Vejerslev, 1997). This mosaicism may pervade both components of placental villi, the trophectoderm or the inner cell mass-derived mesenchyme, alone or in combination.

The genetic segregation of placental biopsies in confined placental mosaicism suggests that bottlenecks may exist in early development that provide opportunities for genetically separating placental and fetal lineages. It is conceivable that these are physiological genetic bottlenecks underlying the normal somatic development of placental tissue. Alternatively, genetic segregation may represent pathological perturbation of the normal clonal dynamics of early embryonic lineages. For example, it has been suggested that confined placental mosaicism represents a depletion from the fetus-forming inner cell mass of cytogenetically abnormal cells, commonly found in early embryos (Los et al., 1998).

Definition of clinical group	**Number of placentas (bulk)**	**Number of placentas (LCM)**
Birth weight <3rd percentile. Low first trimester PAPP-A.	3	0
Birth weight <3rd percentile. Top decile uterine artery PI (plus bilateral notches if possible).	3	0
Birth weight <3rd. Top decile umbilical artery PI.	3	0
Birth weight <3rd. Lowest decile ACGV.	3	0
Birth weight >97th. Highest decile ACGV. No gestational diabetes.	3	0
Severe preeclampsia and normal birth weight (40-60th percentile), plus sFLT1:PlGF >38 at 36 weeks.	3	0
Any preeclampsia and birth weight <3rd percentile.	3	0
Healthy controls. Birth weight from 40 to 60th percentile, normal PAPP-A and Dopplers, normal ACGV. No preeclampsia.	13	5
Placenta >97th percentile and birth weight range from 39th to 74th percentile	3	0
Total	37	5

Table 5.1 Overview of different clinical groups in the placenta cohort. ACGV=abdominal circumference growth velocity, PAPP-A=Pregnancy Associated Plasma Protein-A, PI=pulsality index, PIGF=Placental growth factor, sFLT1=soluble fms-like tyrosine kinase 1

Lastly, the temporary nature of the placenta might be reflected in the level of mutagenesis that manifests in the organ. The short lifespan of the placenta means mutations have a very limited time to have a phenotypic effect and cause neoplastic lesions of consequence. Concordantly, the rates of placental cancer (choriocarcinoma) are extremely low (Smith et al., 2003). Because of the lack of a selective pressure against a high mutation rate, it is plausible to find many more somatic mutations in placenta than expected in embryonic tissues at the equivalent age.

Here, I studied the development of the placenta and the split between the inner cell mass and trophectoderm by whole-genome sequencing of bulk placenta samples, as well as laser capture microdissections of distinct components of placental villi. In addition, I reconstructed phylogenies of microdissected clusters of trophoblast and mesenchymal cores (**Fig. 5.1**). All of the tissues had been curated by the Pregnancy Outcome Prediction study, a prospective collection of placental tissue and extensive clinical data, including histological assessment of individual biopsies (Pasupathy et al., 2008; Sovio et al., 2017). I included placentas from normal pregnancies and from pregnancies associated with a range of abnormal parameters, such as preeclampsia or low birth weight (**Table 5.1**).

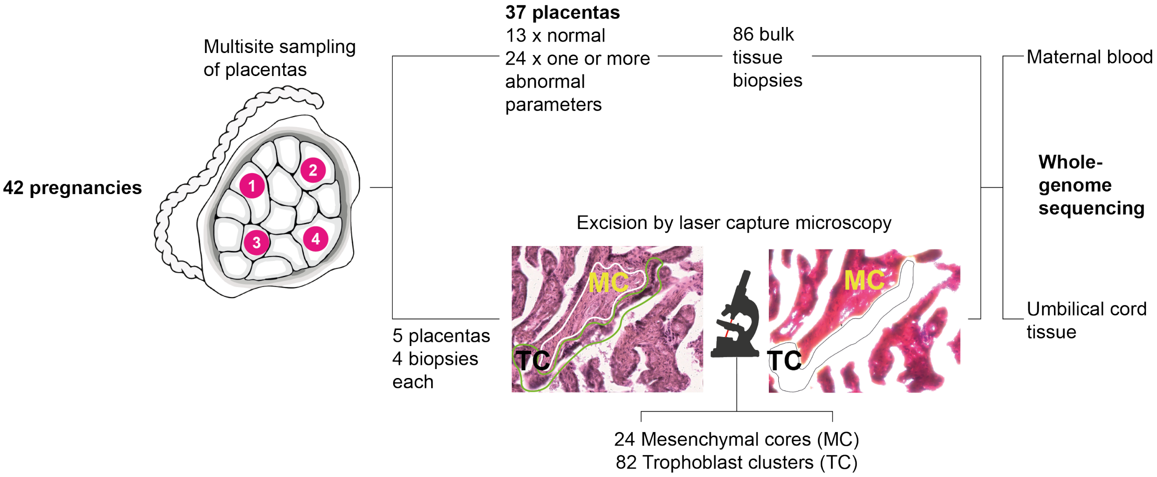
42 pregnancies
Multisite sampling of placentas
1
2
3
4
37 placentas
13 x normal
24 x one or more abnormal parameters
86 bulk tissue biopsies
Maternal blood
Whole-genome sequencing
Umbilical cord tissue
Excision by laser capture microscopy
MC
TC
MC
TC
5 placentas
4 biopsies each
24 Mesenchymal cores (MC)
82 Trophoblast clusters (TC)

Figure 5.1 Workflow detailing the experimental design with photomicrograph demonstrating microdissection of trophoblast components. TC=trophoblast clusters, MC=mesenchymal core.

5.2 Placental biopsies contain clonal populations

The starting point of this study were whole genome sequences of 86 placental biopsies, obtained from 37 term placentas along with inner cell mass-derived umbilical cord tissue and maternal blood. Placental and umbilical cord biopsies were washed in phosphate-buffered saline to remove maternal blood. Placental biopsy contamination with maternal blood was excluded by screening placental genome sequences for low-level maternal germline variants (see Chapter 2, section 2.5.3).

SNVs called in these placental biopsies revealed a high burden and an unusual degree of clonality (**Fig. 5.2**). On average, each placental bulk sample harboured 145 SNVs, with the burden ranging from 38 to 259. In addition, the VAF profile of these samples generally revealed the presence of a large clone residing in these tissues. The median VAF within these samples was 0.24 on average, ranging from 0.15 to 0.44. This indicates that generally these large clones account for approximately half of the cells in these biopsies. As demonstrated in previous chapters, solid organ biopsies are consistently polyclonal and only harbour a handful of mutations acquired during the earliest cell divisions of life. Clonal nephrogenesis in Wilms tumour patients represents an exception to this. However, in those cases, only a few mutations signal the presence of the early expansion. This falls short of the high burden and clonality in bulk placenta samples.

Neither the observed median VAF nor the mutation burden differed between normal placentas and those with an abnormal feature, such as growth restriction. This indicates that the dynamics generating these clones are a general feature of physiological placental development.

Analysis of other classes of somatic mutations, indels and copy number changes confirmed the clonal composition of biopsies. Remarkably, 41 out of 86 biopsies harboured at least one copy number change (gain or loss; median size per unique segment, 73.6 kb). However, only one copy number variant, a trisomy of chromosome 10, would have been detectable by clinical karyotyping of chorionic villi. Comparing somatic changes between

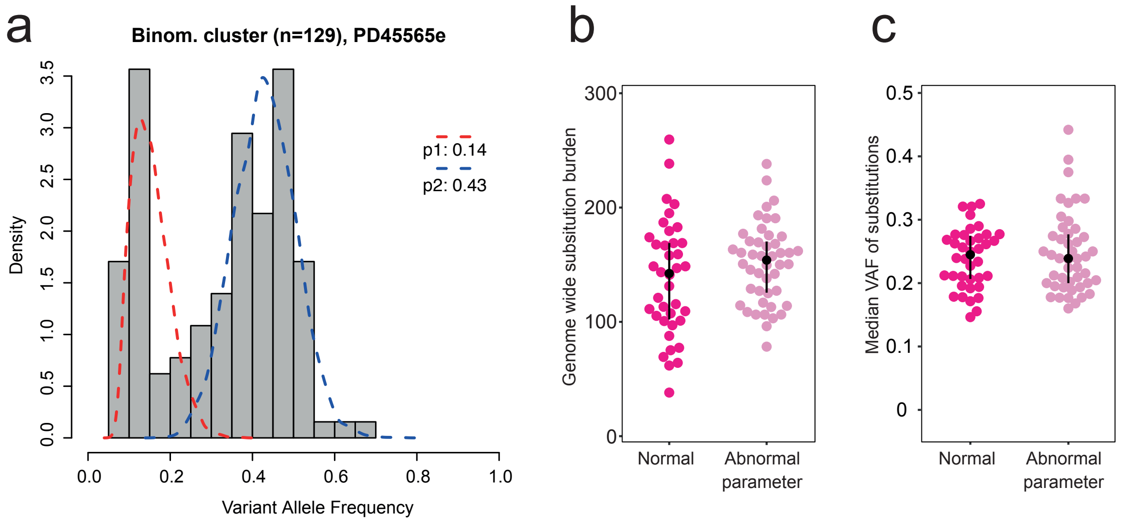

Figure 5.2 (a) VAF distribution within one placental bulk sample (PD45565e). Red and blue dashed lines indicate clonal decomposition through a binomial mixture model, with the estimated VAF per cluster indicated in the legend. (b) SNV burden of each placental biopsy, adjusted for coverage and median VAF. An abnormal pregnancy is defined by the deviation of one or more clinically validated markers from their normal range over the course of pregnancy. (c) Median variant allele frequency of SNVs in each placental biopsy.

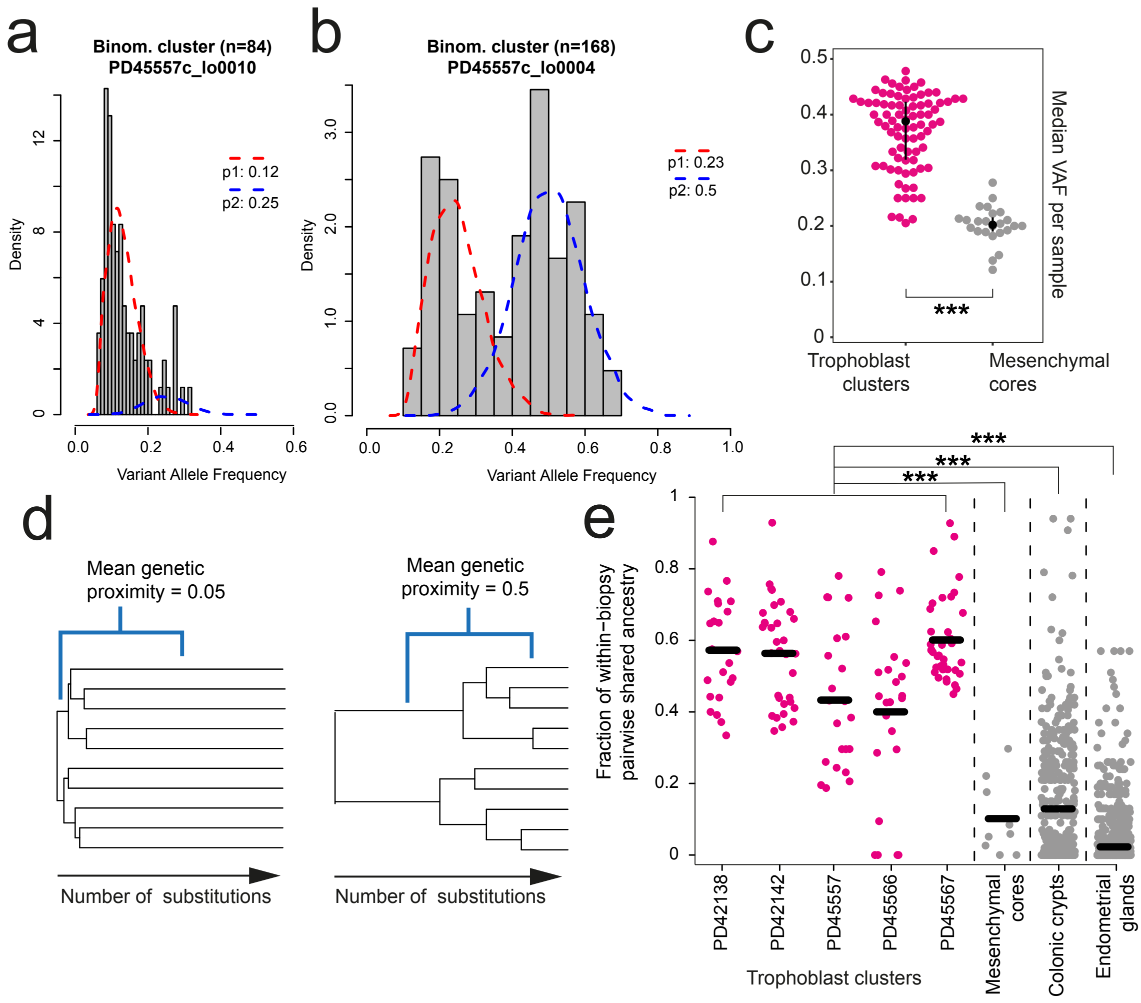

Figure 5.3 Histograms of VAF distribution, overlaid with components as identified by a binomial mixture model for a mesenchymal core sample (a) and a trophoblast cluster (b). ‘p’ indicates the estimated peak VAF of the components. (c) Comparison of the median substitution VAF between microdissected trophoblast and mesenchymal cores. (*** indicates p<0.001, Wilcoxon rank-sum test) (d) Genetic proximity scores were calculated as the fraction of shared mutations of a pair of samples divided by their mean total mutation burden. For example, a mean score of 0.05 conveys little sharing, while 0.5 signifies a longer shared development. (e). Genetic proximities across trophoblast clusters and mesenchymal cores from the same placental biopsies and data from colonic crypts (Lee-Six et al., 2019) and endometrial glands (Moore et al., 2020). Each dot represents the comparison of two of the same histological unit (e.g., two trophoblast clusters) from the same biopsy. To avoid including adult clonal expansions, bifurcations in phylogenies after 100 post-zygotic mutations were not considered for colon and endometrium. (*** indicates p<0.001, Wilcoxon-rank sum test)

multiple biopsies from the same placenta showed that the majority were unique to the given sample. This suggests that each biopsy represents a genetically independent unit. Of note, placental biopsies had been obtained from separate quadrants of the placenta, several centimetres apart, and therefore they represent distinct lobules in the organ. Therefore, these observations indicate that placental biopsies inherently possessed confined, mosaic genetic alterations.

5.3 Trophoblast clusters are closely related clonal units

The large clones residing in placental tissues could be due to one of the two main components in chorionic villi, namely the trophoblast or the inner cell mass-derived mesenchymal cores. To investigate the cellular origin of these clones in placental biopsies, 82 trophoblast clusters and 24 mesenchymal cores were excised using LCM from the term placentas of five normal pregnancies. These were then subjected to low-input library preparation and whole-genome sequencing.

Calling SNVs revealed that mesenchymal cores generally exhibited a polyclonal VAF distribution (**Fig. 5.3a**). In contrast, clusters of trophoblast exhibited more elevated VAF distributions, with a distinct monoclonal architecture (**Fig. 5.3b**). Concordantly, the median VAF observed per LCM cut was significantly higher in trophoblast clusters compared to mesenchymal cores ($p<0.001$; Wilcoxon rank-sum test) (**Fig. 5.3c**). Hence, the mosaic clonal architecture observed in bulk sample most likely emanates from the trophoblast. In other words, the SNVs identified in placental biopsies are largely accumulated by the trophectodermal lineage, rather than the inner cell mass lineage.

This conclusion is further reinforced by studying the genetic relationship between trophoblast derivatives and mesenchymal cores from the same biopsies. From the reconstructed phylogenetic trees a pairwise genetic proximity scores of microdissections of the two components was calculated. This score was defined as the fraction of shared mutations out of the total mutation burden of the pair (**Fig. 5.3d**; see Chapter 2, section 2.5.3). A low genetic proximity score for pairs of trophoblast clusters or of mesenchymal cores from the same biopsy would indicate that the pool of precursor cells forming these diverged early in development. Conversely, a high score would suggest that histological units within each patch of tissue arose from only a few precursor cells with a relatively long shared ancestry. This analysis revealed a significant difference in the developmental clonal composition between trophoblast clusters and mesenchymal cores ($p < 0.001$; Wilcoxon rank-sum test) (**Fig. 5.3e**). On average, within each biopsy, pairs of trophoblast clusters shared 53% of somatic mutations, indicating a long, joint developmental path of these cells. In contrast, pairs of mesenchymal cores from the same biopsy exhibited a mean genetic proximity of 10% and thus a short, shared phylogeny, in line with other inner cell mass-derived tissues, such as colon and endometrium (**Fig. 5.3e**). These observations suggest that large expansions of single trophoblastic progenitors underpin the normal clonality and observed confined mosaicism of placental biopsies.

The monoclonal organisation of trophoblast clusters provided the opportunity to examine mutational processes that forged placental tissue in detail. Examining the burden of SNVs of individual trophoblast clusters further, I found an average of 192 variants per cluster.

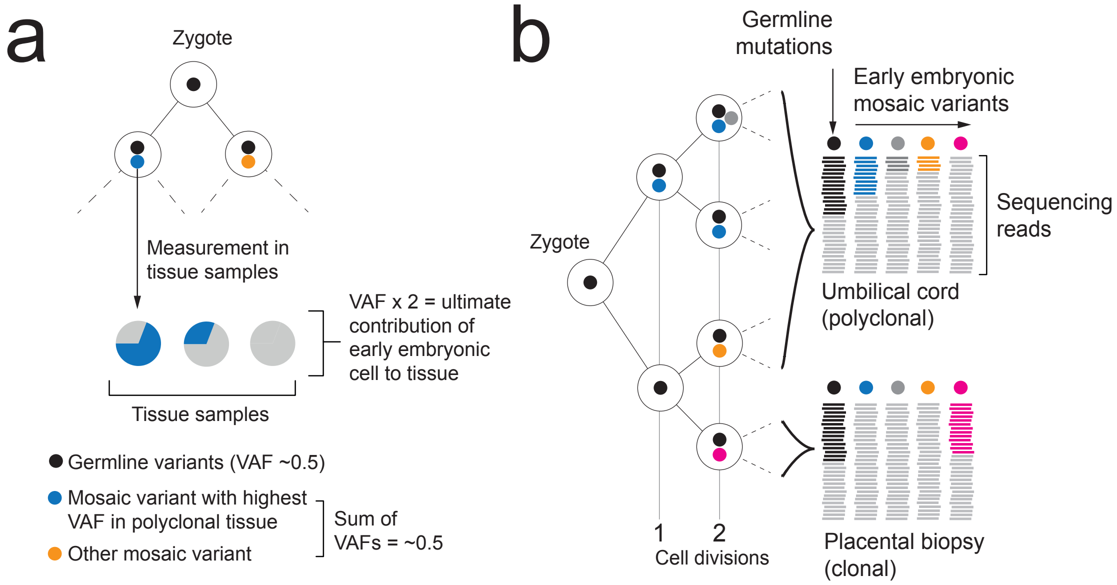
a
Zygote
Measurement in tissue samples
VAF x 2 = ultimate contribution of early embryonic cell to tissue
Tissue samples
● Germline variants (VAF ~0.5)
● Mosaic variant with highest VAF in polyclonal tissue
● Other mosaic variant
Sum of VAFs = ~0.5
b
Germline mutations
Early embryonic mosaic variants
Zygote
Sequencing reads
Umbilical cord (polyclonal)
Placental biopsy (clonal)
1
2
Cell divisions

Figure 5.4 (a) Schematic depicting the detection of the earliest post-zygotic mutations and the estimation of contribution to samples from their variant allele frequencies. (b) Hypothetical lineage tree of the early embryo showing how measurements of VAF may relate to cell divisions.

5.4 Biases in cell allocation to trophectoderm and inner cell mass

So far, I have established that the placenta encompasses large clonal populations of trophectodermal cells. It is most likely that these clones are seeded by single cells very early in development. After the initial seeding there appears to be little to no migration of trophectodermal cells along in the developing placenta. Thus, post-zygotic acquisitions of an aneuploidy in this lineage can result in clinically detectable, trophoblastic confined placental mosaicism. In addition to the committed lineage of trophoblast progenitors, it is possible to investigate the patterns of early, pre-blastulation cell differentiation decisions by identifying mutations shared with the umbilical cord, a tissue entirely derived from the inner cell mass (**Fig. 5.4**).

The most straightforward assessment of this early split is to reconstruct phylogenies of trophoblast clusters. By interrogating the presence of variants in the phylogeny in the umbilical cord genomes I found three different configurations. (1) In PD45566 and PD45567, the trophoblast clusters and the umbilical cord share a most recent common ancestor, the root of the tree, which is most likely the zygote (**Fig. 5.5a**). The umbilical cord exhibits an early asymmetry in line with the expectation, of approximately 2:1 (Behjati et al., 2014; Ju et al., 2017; Lee-Six et al., 2018). (2) In PD45557, one lineage of trophoblast clusters does not share any SNVs with the umbilical cord (**Fig. 5.5b**). However, not all cell divisions are resolved as the tree starts with a trifurcation, obscuring definitive conclusions about the most recent common ancestor of either lineage. The umbilical cord of PD45557 still exhibits the expected asymmetry of 2:1. (3) In PD42138 and PD42142, there was a complete separation between trophectodermal samples and the umbilical cord, without any shared SNVs between them (**Fig. 5.5c**). For PD42138, this indicates that the most recent common ancestor for all trophoblast clusters and for the umbilical cord are two distinct cells later than the zygote, presumably the first generation of daughter cells. The phylogeny of PD42142 starts with a trifurcation, but nevertheless, the most recent common ancestor for the umbilical cord must be a cell later than the zygote.

These three patterns are recapitulated in the bulk genomes as well (**Fig. 5.5d**). In about half of pregnancies (17/37), the earliest post-zygotic mutation exhibited an asymmetric VAF across inner cell mass and trophectoderm lineages, without genetically segregated placental samples in this configuration. In about a quarter of pregnancies (11/37), I found that one placental biopsy did not harbour the early embryonic mutations shared between umbilical cord and other placental biopsies. In other words, that trophectodermal lineage shared no post-zygotic genetic ancestry with the inner cell mass. In the remaining quarter (9/37) of pregnancies, the early cell allocation generated a complete separation of all placental tissues from umbilical cord samples (**Fig. 5.5e**). Taken together, this data suggests that in about half of placentas, there is evidence for a trophectodermal lineage sharing no mutations with the inner cell mass. Consequently, genomic alterations that pre-exist in the zygote, or arise within the first few cell divisions, may segregate between the placenta and fetal lineages by virtue of the natural separation between these lineages.

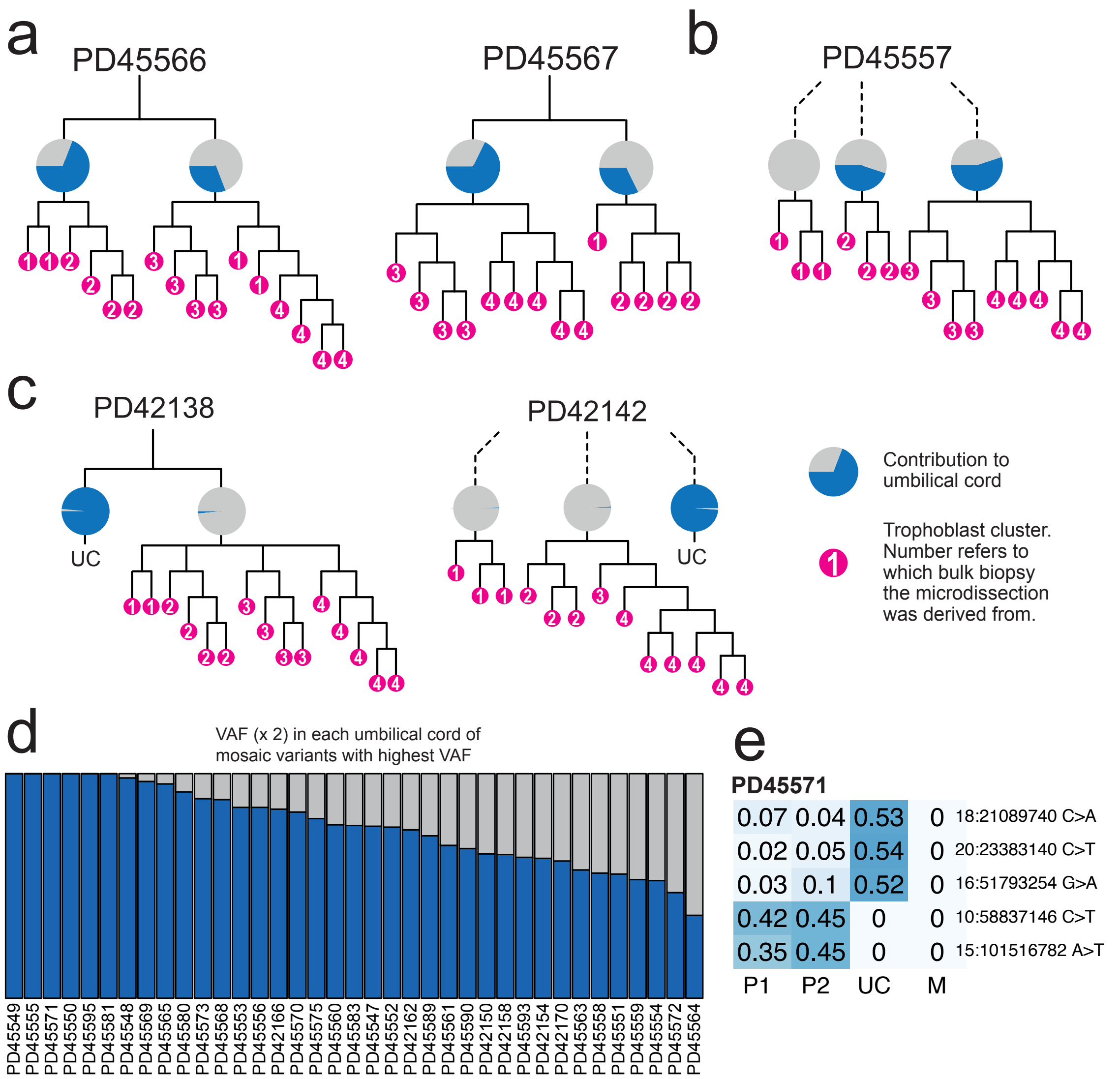

Figure 5.5 (a) Early trees of trophoblast clusters of PD45566 and PD45567, with the contribution of lineages to the umbilical cord colored in blue in pie charts. The umbilical cord exhibits an asymmetric contribution of the daughter cells of the zygote. (b) Early cellular contribution in PD45557 shows separation of one placental lineage. (c) In PD42138 and PD42142 the placental and umbilical cord lineages do not share any early embryonic mutations. (d) The contribution of the major lineage to the umbilical cord as calculated from the embryonic mutation with the highest VAF. (e) Heatmap of early embryonic variants in PD45571 showing a complete split between placenta (P1 and P2) and umbilical cord (UC). SNVs are absent from the maternal blood (M).

5.5 A reversal of trisomy 10

This potential for a full segregation between trophectodermal and inner cell mass-derived lineages is most strikingly exemplified by PD45581. One of the two placental biopsies sequenced for this patient harbours a trisomy of chromosome 10 (PD45581c). This trisomy is absent in the other placental biopsy (PD45581e), as well as the umbilical cord (PD45581e), both of which displayed a regular disomic pattern for chromosome 10 (**Fig. 5.6a**). From interrogating common SNP sites on this chromosome across all samples from this patient and the whole-genome data of maternal blood, it became apparent that the trisomy consisted of two maternal copies and one paternal one. However, the disomic samples did not carry one paternal and one maternal copy as expected, but rather two non-identical maternal copies. This was evident from SNPs present in the trisomy but absent from the disomy, as well as SNPs homozygous in the disomy, but less than homozygous in the trisomy (**Fig. 5.6b**). If the disomy had evolved into a trisomy by a conventional gain of a whole chromosome, neither of these patterns would have manifested.

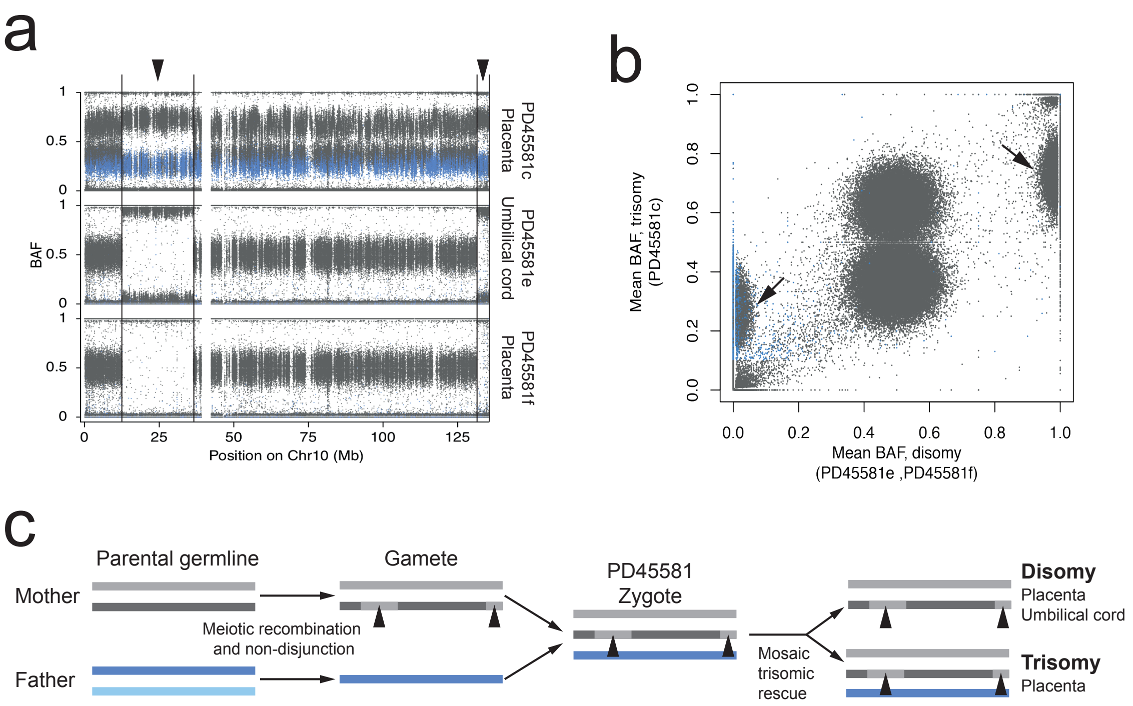

Figure 5.6 (a) B-allele frequency (BAF) of germline SNPs on chromosome 10 in PD45581, showing a trisomy in PD45581c. SNPs absent from mother are colored in blue. (b) Mean BAF of samples with trisomy 10 (PD44581c) versus the mean BAF in disomic samples (PD45581e, PD45581f). SNPs absent from mother are colored in blue. The clusters of variants absent or homozygously present in disomic samples, but distributed around 0.33 and 0.66 (arrows) in the trisomy refute the possibility of a somatic duplication. (c) Overview of genomic events in PD45581 and parents leading to the observed mosaic trisomic rescue. The arrowheads highlight areas of two genotypes in PD45581c due to meiotic recombination in the mother.

Instead, all evidence points to the zygote starting out with a trisomy consisting of one paternal and two maternal copies (**Fig. 5.6c**). Subsequently, the paternal copy was mosaically lost, resulting in a subset of cells with a normal disomic profile. Hence, this is a direct observation of a trisomic rescue. The origin of the trisomy as maternal meiotic non-disjunction is further reinforced by two large regions of homologous recombination, which appear fully homozygous in disomic samples (**Fig. 5.6a**).

A full trisomy 10 is lethal during embryogenesis (Hassold et al., 1996). It stands to reason that the embryo was only able to develop because a subset of cells in the inner cell mass-derived lineages had reverted back to disomy. The low number of embryonic mutations that were found at a clonal level in the umbilical cord hint that the initial trisomic rescue event must have happened within the first few cell divisions of life. This is consistent with the initial event occurring during the cleavage stage of embryogenesis. The reportedly high level of genomic instability due to rapid divisions at this time might aid aneuploid embryonic cells to self-correct.

The trophectodermal line, however, does not seem to harbour the same strong selective pressure against aneuploidies as the embryo proper. Because of this, cells containing the trisomy of chromosome 10 are still abundant in the placenta. It is plausible that many clinical cases of confined placental mosaicism represent trisomic rescues.

5.6 Processes and impact of placental mutagenesis

By deconvoluting the mutational signatures that contribute to the profile of SNVs seen in bulk biopsies and trophoblast microdissections, it is possible to determine the mutagenic processes operating in the human placenta. Three different reference signatures contributed to the mutagenesis in placenta: signatures 1, 5, and 18 (Alexandrov et al., 2020). Signature 1 consists of C>T mutations at CpG sites, and corresponds to spontaneous deamination of methylated cytosines. Signature 5 is a rather flat signature, without prominent features. Both signatures 1 and 5 are ubiquitous in human tissues and accumulate throughout life (Alexandrov et al., 2015). In contrast, signature 18 has been identified infrequently in normal tissues. It is characterised by C>A variants and has been associated with reactive oxygen species and oxidative stress (Poetsch, 2020). In placental biopsies and microdissected trophoblast clusters, signature 18 contributed 43% of all SNVs (**Fig. 5.7a,b**). In comparison, in normal human colorectal crypts, the normal tissue with the highest prevalence of signature 18 mutations described to date, it contributed an average of 13% of substitutions (Lee-Six et al., 2019). Within the confines of limited sampling size of this cohort, I did not observe a significant difference in signature composition between normal placentas and those with an abnormal parameter.

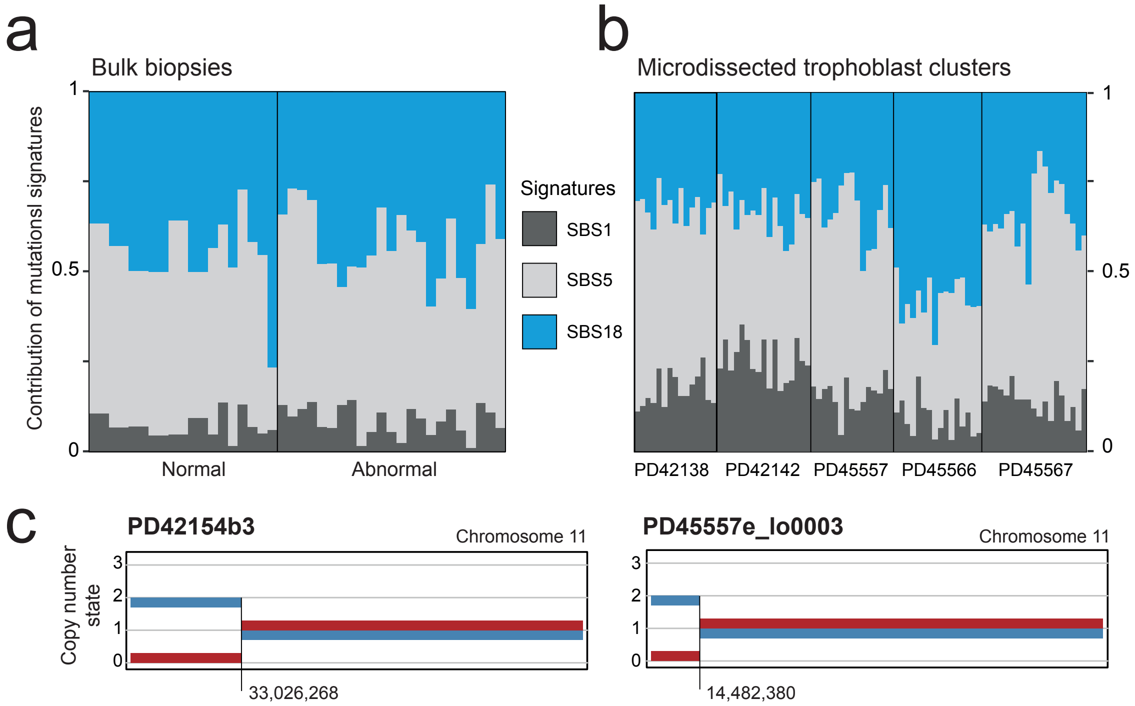

Figure 5.7 Proportion of mutational signatures 1, 5 and 18 attributed to placental bulk samples (a) and microdissected trophoblast clusters (b). (c) Amplification of the paternal allele through copy number neutral loss of heterozygosity of 11p in one bulk placental sample (PD42154b3) and one trophoblast cut (PD45557e_lo0003).

It is conceivable that the high exposure to signature 18 in the human placenta is entirely due to its unique function as the supplier of oxygen *in utero* and the rapid cell division required for full development of the organ. An alternative explanation stems from the temporary nature of the placenta. Somatic mutations arising in the trophectodermal lineage have a very limited timeframe to have a negative impact on the fitness of the offspring, as the very low incidence of malignant neoplasia illustrates. It stands to reason that the placenta tolerates most somatic mutations, in line with its tolerance of aneuploidies. Therefore, it might have a lower activity of the DNA repair machinery than tissues derived from the embryo proper, where mutations can impact phenotype for an entire lifetime. In line with this, the mutation burden attributable to signatures 1 and 5 in trophoblasts is at least 109 SNVs[1] for the duration of a pregnancy, corresponding to a mutation rate of approximately 146 a year. This is several times higher than any mutation rate reported for an adult normal

[1] The mean burden of trophoblast clusters is 192, the mean attribution to signatures 1 and 5 is 57%, so the burden due to signatures 1 and 5 is their product.

tissue, the highest of which is colonic epithelium with a rate of 43.6 SNVs a year (Lee-Six et al., 2019).

Annotating functional consequences of all somatic variants found in bulk biopsies and trophoblast samples, indicated that most changes were unlikely to have any impact on phenotype. The majority (42 of 81 unique variants) of copy number changes lay within fragile sites. Interestingly, two placentas out of 42 harboured copy number neutral loss of heterozygosity of chromosome 11p15. In both cases, this constituted uniparental disomy of the paternal allele (**Fig. 5.7c**). The 11p15 region contains the imprinted loci, among which the *H19* locus discussed at length in the previous chapter. Paternal amplification of this region underpins the overgrowth syndrome, Beckwith-Wiedemann, while maternal amplification would result in Silver-Russell syndrome. Both have been implicated in placental disease, mainly in pregnancy-induced hypertension and intrauterine growth restriction, respectively (Angiolini et al., 2011; Bourque et al., 2010; Fowden et al., 2006). Since these imprinted regions can be affected by a loss of imprinting rather than a loss of heterozygosity, it is likely that a proportion of placental cells might harbour an abnormal methylation landscape of the 11p15 imprinted sites.

5.7 Conclusion

In this exploration of the somatic genomes of human placentas, I identified genetic bottlenecks at different developmental stages that confined placental tissues genetically. Most prominently, every placental biopsy that I examined represented an independent clonal trophoblast unit, suggesting that mosaicism represents the inherent trophectodermal clonal architecture of human placentas. During the first few divisions of life, the prospective split between trophectoderm and inner cells segregates placental tissues from the embryo proper, genetically isolating trophoblast lineages. Together, these bottlenecks may represent developmental pathways through which cytogenetically abnormal cells phylogenetically and spatially separate. This renders them detectable by genomic assays utilized in the clinical assessment of chorionic villi. The findings thus provide plausible, physiological developmental routes through which confined placental trophoblast mosaicism may arise. I suspect that as our understanding of the clonal dynamics of human embryonic lineages grows, it is possible to find additional bottlenecks that account for placental mosaicism affecting mesenchymal lineages also.

The landscape of somatic mutations in placental biopsies and at the level of individual trophoblast units was unusual compared to other normal human tissues studied to date. In particular, placental tissue was an outlier with regards to the abundance of signature

18 mutations. It is possible that these somatic genetic peculiarities represent the specific challenges that trophoblast lineages undergo during placental growth, such as the approximate threefold rise in changes in the local oxygen tension of blood surrounding the villi between eight and twelve week's gestation (Jauniaux et al., 2000). Furthermore, it may be conceivable that as a temporary, ultimately redundant organ, some of the mechanisms protecting the somatic genome elsewhere do not operate in placental trophoblasts.

It is possible that genomic placental alterations contribute to the pathogenesis of placental dysfunction, which is a key determinant of the "Great Obstetrical Syndromes", such as preeclampsia, fetal growth restriction and stillbirth. Previous studies associating confined placental mosaicism with these syndromes have yielded conflicting results (Amor et al., 2006; Baffero et al., 2012; Grati et al., 2020; Jauniaux et al., 2000; Kalousek et al., 1991; Toutain et al., 2018). This study may explain these discrepancies, as the genomic alterations I observed were not uniformly distributed across multiple biopsies from the same placenta. Larger scale systematic studies of the genomic architecture of the human placenta in health and disease might establish the role of placental genomic aberrations in driving placenta-related complications of human pregnancy.

Chapter 6

Conclusions and future perspectives

6.1 Summary of the main findings

The research presented in this dissertation leveraged naturally occurring somatic mutations as passive marks of cell division and hence as a direct bar coding of the developmental history of a cell. By reconstructing phylogenies from those mutations, I was able to investigate fundamental processes of human development as well as the emergence of cancer via a precursor lesion. Briefly, the main results per chapter are:

1. **Extensive phylogenies of normal human development reveal asymmetries in embryonic lineages, spatial patterning of mosaicism and clonal expansions in adult life.** While the asymmetry of contribution of the first cell division is approximately 2:1 and is conserved in most tissues, the discrepancy in asymmetry between different samples from the same patients points at a role for genetic bottlenecks after blastulation. This is strongly exemplified by a bulk brain sample from one patient, which is almost exclusively derived from one of two zygotic lineages, unlike the bulk colon sample from the same patient. From targeted re-sequencing of embryonic variants, I was able to show a large-scale spatial pattern in human brain corresponding to heterogeneous mosaicism. The phylogenies and spatial data revealed the typical embryonic patch size for human colonic crypts, as well as later clonal expansions in normal tissues such as the prostate, colon and appendix, in some cases due to cancer driver mutations. In addition, the pattern of sharing between seminiferous tubule sections and samples derived from other tissues indicates an extraembryonic origin for human germ cells, most likely the amnion.

2. **Wilms tumours often arise from an embryonal precursor lesion, residing in the normal kidney.** By interrogating tumour variants in normal renal tissues, I identified

a detectable, pre-neoplastic clone in over half of the Wilms tumour. This precursor appears to be driven by hypermethylation of *H19*, effectively resulting in a mosaic Beckwith-Wiedemann syndrome. However, kidneys containing cells with this hypermethylation appear functionally and morphologically normal. From timing the occurrence in bilateral tumours, it is likely that the initial hypermethylation event happens early in development.

3. **The human placenta consists of large clonal patches, which can segregate from the inner cell mass-derived lineages at the earliest stages.** Large bulk biopsies of the placenta contain clonal populations, which can be traced back to its trophectodermal component. Trophoblast clusters excised from the same biopsy are closely related to one another. Comparing early embryonic mutations between placental lineages and umbilical cord DNA, which is derived from the inner cell mass, revealed that in approximately half of the cases a placental lineage has no post-zygotic relation with the umbilical cord. Furthermore, in a quarter of cases, the umbilical cord is entirely derived from a progenitor later than the zygote, facilitating a natural segregation and pathway to generate confined placental mosaicism.

6.2 Future perspectives and ongoing work

Taken together, these three lines of research illustrate the potential of somatic mutations to answer fundamental questions about human developmental processes and the origins of cancer. The findings from the individual chapters have profound implications and lead to interesting future work.

Lineage tracing of normal development has highlighted the existence of mosaic patterns of embryonic mutations on a tissue- and organ-level, which reflect lineage commitments and segregations during embryogenesis. Future studies can use such patterns to reconstruct the growth of individual organs, such as exemplified by the placental study in Chapter 5, but also to investigate expansions and tissue renewal in case of injury. It is plausible that, as the cost of whole-genome sequencing decreases even further, that future lineage tracing studies can be based on even more samples per patient and reveal patterns of human embryogenesis currently obscured to us, such as the precisely localising timing the emergence of the primordial germ cells and the providing evidence for the existence of extraembryonic intercalation in humans. Furthermore, to complement the phylogeny of the embryo proper, genomes derived from extraembryonic tissues, such as the placenta or umbilical cord might further elucidate the lineage commitments between the trophectoderm and inner cell mass, and the epiblast and hypoblast, respectively.

The emergence of Wilms tumour from large precursor clones in the normal kidney has profound implications on the screening, treatment and further surveillance of Wilms tumour patients, especially if aberrant *H19* methylation status is detectable non-invasively. In particular, the discovery of large fields of normal renal cells carrying *H19* hypermethylation leads to important questions: (1) do differentiated renal cells still have the potential to transform into Wilms tumours, (2) how are these cells able to overcome the loss of imprinting, (3) what role do nephrogenic rests have in the origin of Wilms tumour and do they differentiate or disappear naturally over time and (4) how widespread is a mosaic Beckwith-Wiedemann syndrome in the human population.

The results from the placental sequencing project emphasise the common presence of aneuploidy in the human placenta, best exemplified by a case of trisomic rescue of chromosome 10. It is plausible that a significant proportion of the human population has been subjected to trisomic rescue without knowing, given the high prevalence of confined placental mosaicism and the observation that trisomic rescues could explain a significant part of this phenomenon. This naturally leads to the question of how often this affects the general population, and whether we can detect any chromosome-level biases in frequency of aneuploidy. If the trisomic rescue event would lead to a completely normal chromosomal landscape, where the disomy consists of one maternal and one paternal copy, the copy number profile appears completely normal and the reversal of aneuploidy is undetectable. However, if the the trisomic rescue leads to two copies from the same parent being retained, the meiotic recombination leads to regions of haploidisation in the genome, which are easily detectable through conventional whole-genome sequencing. The latter should happen in approximately a third of all cases, if all three chromosomes have an equal probability of being lost.

Beyond the immediate ramifications and direct follow-ups to the research presented here, the methodology of lineage tracing using somatic mutations can be applied to answer a wide variety of other biological questions. The section below will cover a few of the exciting avenues of research following this principle, some of which are ongoing work.

6.2.1 Other childhood cancers, bilateral tumours and secondary malignancies

The initial discovery of an early clonal expansion predisposing to Wilms tumours prompts questions about the generality of this phenomenon. That is, do other childhood cancers arise similarly? Can early *H19* hypermethylation lead to other tumours? One piece of ongoing work focuses on malignant rhabdoid tumour (MRT) and the associated normal tissue, i.e. nerves of the kidney hilum for renal MRT or spinal nerve roots in the case of spinal MRT

(Custers et al., 2021). Preliminary analysis has indicated that normal nervous tissue carries large precursors clones to the MRT, with a mutation burden and clonal composition even more pronounced than the clonal nephrogenesis observed in Wilms tumour. In addition, both normal cells and tumour have inactivated both copies of *SMARCB1*, a tumour suppressor unequivocally implicated in MRT pathogenesis (Margol and Judkins, 2014). No discernible genetic events with a phenotypic consequence distinguish the normal clone from the tumour, suggesting that a non-genetic event might trigger the final transformation from normal cell to tumour.

Beckwith-Wiedemann syndrome predisposes to a number of different embryonal childhood cancers besides Wilms tumour, leading to the possibility that the *modus operandi* of tumour formation is equivalent to clonal nephrogenesis. A preliminary analysis of a single case of hepatoblastoma (an embryonal childhood liver cancer) from a patient diagnosed with Beckwith-Wiedemann syndrome, confirmed the presence of a precursor clone in normal liver, but many more cases are required to investigate the recurrence and robustness of these patterns. This is somewhat hampered by the rarity of hepatoblastoma.

Somatic mutations can be particularly effective when studying the origins of bilateral cancers. Comparison of the mutational patterns between the two lesions can distinguish whether (1) one is a metastasis of the other, (2) both tumours emerge from a common clone or (3) the two tumours have emerged independently (Foulkes and Polak, 2020). For example, the bilateral Wilms tumour cases in this dissertation represent the second scenario, as they arise from a common root of clonal nephrogenesis. However, two cases of bilateral neuroblastoma studied during my PhD revealed that the bilateral lesions only share a few SNVs, all of which shared with bulk blood (Coorens et al., 2020). This indicates an early divergence of tumour lineages during the first few cell divisions of life and hence, an independent emergence of both tumours. This is likely a consequence of germline predisposition mutations. Interestingly, the two tumours often exhibited similar genomic events, such as loss of the second copy of *SMARCA4*.

A last piece of ongoing work studies the origins of secondary acute myeloid leukaemia (AML) or myelodysplastic syndrome (MDS) in two patients after chemotherapy treatment for neuroblastoma and an autologous haemotopoietic stem cell transplant (Coorens et al., 2021a). From the patterns of somatic mutations, it was possible to discern that the primordial clone of the AML and MDS was already present in the stem cell populations prior to transplant and constitutes *bona fide* clonal haematopoiesis. Moreover, the somatic mutations were caused by mutational signatures attributable to platinum-based chemotherapy (Pich et al., 2019), further underpinning the role of the treatment in emergence of these secondary cancers.

This ongoing work further illustrates the versatility and ubiquity in which somatic mutations can be used to answer questions about the origins of human cancers. While the work of my PhD has focused on childhood cancers, these approaches are directly generalisable to adult cancers.

6.2.2 Further methodological advancements

The ongoing decrease in costs for DNA sequencing can facilitate deeper and more extensive sampling strategies to reconstruct phylogenetic trees of normal and aberrant human development. To improve the detection of early somatic variants and the building of phylogenies in such large data sets, two areas of substantial and necessary methodological advancement some to mind.

Firstly, the approach currently taken to call somatic variants is based on the old paradigm of a comparison between a tumour and a normal sample. As an organic extension to this, the somatic variant calling in normal tissues is currently built on algorithms with that paradigm in mind, resulting in the use of an *in silico*-created unmatched normal sample. However, this completely neglects the potential for leveraging the many samples acquired from the same patient. Instead, this wealth of information is only used in the filtering steps I devised, after combining the calls from all pairwise comparisons. In the long run, it might be more time-efficient and computationally powerful to directly combine the genomic information and patterns of mutations across the many samples from the same patient in the initial variant calling stage. Such an amalgamation of whole-genome sequencing data might result in calling of variants with more sensitivity to low VAF mutations by combining variant reads across samples, while simultaneously better handling recurrent artefactual calls and increasing the specificity of calling.

Secondly, while the infinite sites assumption still holds and the maximum parsimony approach for phylogeny was the optimal choice within the confines of this dissertation, the increasing size of data sets and called somatic variants might lead to insurmountable violations of that assumption. As explored in Chapter 3, section 3.10, even at the current scale of sampling and sequencing, mutations re-occur independently within the phylogeny by chance. In addition, algorithms based on maximum parsimony do not scale optimally and very large sets of samples might render this approach intractable. Hence, in the near future, it might be desirable to develop a novel phylogenetic tree building algorithm based on maximum likelihood or Bayesian inference that is specifically tailored to the unique aspects of somatic phylogenetics. Since these approaches require explicit models of genomic sequence evolution, it would provide an opportunity to incorporate our prior knowledge on mutational signatures.

The proposed methodological advancements are by no means straightforward exercises, but they will aid in maturing the field of ‘somatic phylogenetics’ by further adding robustness to the approaches employed throughout this dissertation.

6.2.3 Closing remarks

This dissertation has explored the vast potential of using somatic mutations to trace the lineages of individual cells to elucidate fundamental processes of human development and the emergence of cancer. The studies presented here, however, only scratch the surface of the myriad problems that can be addressed with the approaches outlined in this thesis. The ubiquity of somatic mutations lends an enormous versatility to the unresolved questions they can answer. Without a doubt, the future of this field holds many exciting opportunities.

Bibliography

1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

Abu-Issa, R., Waldo, K., and Kirby, M. L. Heart fields: one, two or more? *Developmental biology*, 272(2):281–285, 2004.

Aitken, S. J., Anderson, C. J., Connor, F., et al. Pervasive lesion segregation shapes cancer genome evolution. *Nature*, 583(7815):265–270, 2020.

Alemany, A., Florescu, M., Baron, C. S., Peterson-Maduro, J., and van Oudenaarden, A. Whole-organism clone tracing using single-cell sequencing. *Nature*, 556(7699):108–112, 2018.

Alexandrov, L. B., Jones, P. H., Wedge, D. C., et al. Clock-like mutational processes in human somatic cells. *Nat Genet*, 47(12):1402–7, 2015.

Alexandrov, L. B., Kim, J., Haradhvala, N. J., et al. The repertoire of mutational signatures in human cancer. *Nature*, 578(7793):94–101, 2020.

Amor, D. J., Neo, W. T., Waters, E., et al. Health and developmental outcome of children following prenatal diagnosis of confined placental mosaicism. *Prenat Diagn*, 26(5):443–8, 2006.

Anders, S., Pyl, P. T., and Huber, W. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–9, 2015.

Angiolini, E., Coan, P. M., Sandovici, I., et al. Developmental adaptations to increased fetal nutrient demand in mouse genetic models of Igf2-mediated overgrowth. *The FASEB Journal*, 25(5):1737–1745, 2011.

Bae, T., Tomasini, L., Mariani, J., et al. Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. *Science*, 359(6375):550–555, 2018.

Baffero, G. M., Somigliana, E., Crovetto, F., et al. Confined placental mosaicism at chorionic villous sampling: risk factors and pregnancy outcome. *Prenatal diagnosis*, 32(11):1102–1108, 2012.

Balakier, H. and Pedersen, R. Allocation of cells to inner cell mass and trophectoderm lineages in preimplantation mouse embryos. *Developmental biology*, 90(2):352–362, 1982.

Barker, N., van Es, J. H., Kuipers, J., et al. Identification of stem cells in small intestine and colon by marker gene Lgr5. *Nature*, 449(7165):1003–7, 2007.

Barresi, M. J. F. and Gilbert, S. F. *Developmental biology*. Sinauer Associates, New York, twelfth edition. edition, 2020.

Behjati, S., Huch, M., van Boxtel, R., et al. Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature*, 513(7518):422–425, 2014.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 57(1):289–300, 1995.

Berry, S. J., Coffey, D. S., Walsh, P. C., and Ewing, L. L. The development of human benign prostatic hyperplasia with age. *The Journal of urology*, 132(3):474–479, 1984.

Bianconi, E., Piovesan, A., Facchin, F., et al. An estimation of the number of cells in the human body. *Ann Hum Biol*, 40(6):463–71, 2013.

Blokzijl, F., de Ligt, J., Jager, M., et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature*, 538(7624):260–264, 2016.

Bonaiti-Pellie, C., Chompret, A., Tournade, M. F., et al. Genetics and epidemiology of Wilms' tumor: the French Wilms' tumor study. *Med Pediatr Oncol*, 20(4):284–91, 1992.

Bouckaert, R., Heled, J., Kühnert, D., et al. Beast 2: a software platform for bayesian evolutionary analysis. *PLoS Comput Biol*, 10(4):e1003537, 2014.

Boué, J., Boué, A., and Lazar, P. Retrospective and prospective epidemiological studies of 1500 karyotyped spontaneous human abortions. *Teratology*, 12(1):11–26, 1975.

Bourque, D., Avila, L., Penaherrera, M., Von Dadelszen, P., and Robinson, W. Decreased placental methylation at the H19/IGF2 imprinting control region is associated with normotensive intrauterine growth restriction but not preeclampsia. *Placenta*, 31(3):197–202, 2010.

Boveri, T. *Zur Frage der Entstehung maligner Tumoren*. Fischer, 1914.

Braude, P., Bolton, V., and Moore, S. Human gene expression first occurs between the four-and eight-cell stages of preimplantation development. *Nature*, 332(6163):459–461, 1988.

Breslow, N., Beckwith, J. B., Ciol, M., and Sharples, K. Age distribution of Wilms' tumor: report from the National Wilms' Tumor Study. *Cancer Res*, 48(6):1653–7, 1988.

Breslow, N., Olshan, A., Beckwith, J. B., and Green, D. M. Epidemiology of Wilms tumor. *Med Pediatr Oncol*, 21(3):172–81, 1993.

Brosens, I., Pijnenborg, R., Vercruysse, L., and Romero, R. The "Great Obstetrical Syndromes" are associated with disorders of deep placentation. *American journal of obstetrics and gynecology*, 204(3):193–201, 2011.

Brunner, S. F., Roberts, N. D., Wylie, L. A., et al. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature*, 574(7779):538–542, 2019.

Buels, R., Yao, E., Diesh, C. M., et al. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome biology*, 17(1):1–12, 2016.

Challen, G. and Goodell, M. A. Clonal hematopoiesis: Mechanisms driving dominance of stem cell clones. *Blood*, 2020.

Charlton, J., Williams, R. D., Sebire, N. J., et al. Comparative methylome analysis identifies new tumour subtypes and biomarkers for transformation of nephrogenic rests into wilms tumour. *Genome Med*, 7(1):11, 2015.

Chatterjee, A., Rodger, E. J., and Eccles, M. R. Epigenetic drivers of tumourigenesis and cancer metastasis. *Semin Cancer Biol*, 51:149–159, 2018.

Cheng, J., Vanneste, E., Konings, P., et al. Single-cell copy number variation detection. *Genome biology*, 12(8):R80, 2011.

Chitwood, B. G., Chitwood, M. B., et al. An introduction to nematology. *An introduction to nematology.*, 1937.

Conklin, E. G. *The embryology of Crepidula: a contribution to the cell lineage and early development of some marine gasteropods*. Ginn, 1897.

Coorens, T. H. H., Treger, T. D., Al-Saadi, R., et al. Embryonal precursors of Wilms tumor. *Science*, 366(6470):1247–1251, 2019.

Coorens, T. H. H., Farndon, S. J., Mitchell, T. J., et al. Lineage-independent tumors in bilateral neuroblastoma. *New England Journal of Medicine*, 383(19):1860–1865, 2020.

Coorens, T. H. H., Collord, G., Lu, W., et al. Clonal hematopoiesis and therapy-related myeloid neoplasms following neuroblastoma treatment. *Blood, The Journal of the American Society of Hematology*, 137(21):2992–2997, 2021a.

Coorens, T. H. H., Moore, L., Robinson, P. S., et al. Extensive phylogenies of human development inferred from somatic mutations. *Nature*, 2021b.

Coorens, T. H. H., Oliver, T. R. W., Sanghvi, R., et al. Inherent mosaicism and extensive mutation of human placentas. *Nature*, 592:80–85, 2021c.

Custers, L., Khabirova, E., Coorens, T. H. H., et al. Somatic mutations and single-cell transcriptomes reveal the root of malignant rhabdoid tumours. *Nature communications*, 12 (1):1–11, 2021.

Darwin, C. *On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life*. J. Murray, London, 1859.

Davies, H., Bignell, G. R., Cox, C., et al. Mutations of the BRAF gene in human cancer. *Nature*, 417(6892):949–54, 2002.

de Gruijl, F. R., van Kranen, H. J., and Mullenders, L. H. UV-induced DNA damage, repair, mutations and oncogenic pathways in skin cancer. *Journal of Photochemistry and Photobiology B: Biology*, 63(1-3):19–27, 2001.

De Paepe, C., Krivega, M., Cauffman, G., Geens, M., and Van de Velde, H. Totipotency and lineage segregation in the human embryo. *Molecular human reproduction*, 20(7):599–618, 2014.

De Paepe, C., Cauffman, G., Verloes, A., et al. Human trophectoderm cells are not yet committed. *Human reproduction*, 28(3):740–749, 2013.

Dobin, A., Davis, C. A., Schlesinger, F., et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

Dobson, A. T., Raja, R., Abeyta, M. J., et al. The unique transcriptome through day 3 of human preimplantation development. *Human molecular genetics*, 13(14):1461–1470, 2004.

Dumanski, J. P., Lambert, J. C., Rasi, C., et al. Mosaic Loss of Chromosome Y in Blood Is Associated with Alzheimer Disease. *American Journal of Human Genetics*, 98(6): 1208–1219, 2016.

Eagleson, G. W. and Harris, W. A. Mapping of the presumptive brain regions in the neural plate of *Xenopus laevis*. *Journal of neurobiology*, 21(3):427–440, 1990.

Eckersley-Maslin, M. A., Alda-Catalinas, C., and Reik, W. Dynamics of the epigenetic landscape during the maternal-to-zygotic transition. *Nature Reviews Molecular Cell Biology*, 19(7):436–450, 2018.

Ellis, P., Moore, L., Sanders, M. A., et al. Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. *Nature Protocols*, 16(2):841–871, 2021.

Fearon, E. R. and Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell*, 61(5): 759–767, 1990.

Feinberg, A. P., Ohlsson, R., and Henikoff, S. The epigenetic progenitor origin of human cancer. *Nat Rev Genet*, 7(1):21–33, 2006.

Felsenstein, J. *PHYLIP (phylogeny inference package), version 3.5 c*. Joseph Felsenstein., 1993.

Ferretti, E. and Hadjantonakis, A. K. Mesoderm specification and diversification: from single cells to emergent tissues. *Curr Opin Cell Biol*, 61:110–116, 2019.

Fleming, T. P. A quantitative analysis of cell allocation to trophectoderm and inner cell mass in the mouse blastocyst. *Developmental biology*, 119(2):520–531, 1987.

Fodde, R., Smits, R., and Clevers, H. APC, signal transduction and genetic instability in colorectal cancer. *Nature Reviews Cancer*, 1(1):55–67, 2001.

Forsberg, L. A., Rasi, C., Malmqvist, N., et al. Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. *Nature Genetics*, 46(6): 624–628, 2014.

Fortin, J. P., Triche, J., T. J., and Hansen, K. D. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics*, 33(4):558–560, 2017.

Foulkes, W. D. and Polak, P. Bilateral tumors - inherited or acquired? *N Engl J Med*, 383(3): 280–282, 2020.

Fowden, A., Sibley, C., Reik, W., and Constancia, M. Imprinted genes, placental development and fetal growth. *Hormone Research in Paediatrics*, 65(Suppl. 3):50–58, 2006.

Genovese, G., Kahler, A. K., Handsaker, R. E., et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N Engl J Med*, 371(26):2477–87, 2014.

Gerstung, M., Papaemmanuil, E., and Campbell, P. J. Subclonal variant calling with multiple samples and prior knowledge. *Bioinformatics*, 30(9):1198–1204, 2014.

Giannoukakis, N., Deal, C., Paquette, J., Goodyer, C. G., and Polychronakos, C. Parental genomic imprinting of the human IGF2 gene. *Nat Genet*, 4(1):98–101, 1993.

Giladi, A., Paul, F., Herzog, Y., et al. Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis. *Nat Cell Biol*, 20 (7):836–846, 2018.

Grati, F. R., Ferreira, J., Benn, P., et al. Outcomes in pregnancies with a confined placental mosaicism and implications for prenatal screening using cell-free DNA. *Genetics in medicine: official journal of the American College of Medical Genetics*, 22(2):309, 2020.

Grobner, S. N., Worst, B. C., Weischenfeldt, J., et al. The landscape of genomic alterations across childhood cancers. *Nature*, 555(7696):321–327, 2018.

Gulyas, B. J. A reexamination of cleavage patterns in eutherian mammalian eggs: rotation of blastomere pairs during second cleavage in the rabbit. *Journal of Experimental Zoology*, 193(2):235–247, 1975.

Guo, F., Li, X., Liang, D., et al. Active and passive demethylation of male and female pronuclear DNA in the mammalian zygote. *Cell stem cell*, 15(4):447–459, 2014.

Haeckel, E. *Generelle Morphologie der Organismen : allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von Charles Darwin reformirte Descendenz-Theorie*. G. Reimer, Berlin, 1866.

Haeckel, E. *Anthropogenie oder Entwickelungsgeschichte des Menschen*. W. Engelmann, 1874.

Hahnemann, J. M. and Vejerslev, L. O. European collaborative research on mosaicism in CVS (EUCROMIC)—fetal and extrafetal cell lineages in 192 gestations with CVS mosaicism involving single autosomal trisomy. *American journal of medical genetics*, 70 (2):179–187, 1997.

Hassold, T., Merrill, M., Adkins, K., Freeman, S., and Sherman, S. Recombination and maternal age-dependent nondisjunction: molecular studies of trisomy 16. *Am J Hum Genet*, 57(4):867–74, 1995.

Hassold, T., Abruzzo, M., Adkins, K., et al. Human aneuploidy: incidence, origin, and etiology. *Environmental and molecular mutagenesis*, 28(3):167–175, 1996.

His, W. *On the Principles of Animal Morphology: Letter to Mr. John Murray*. 1888.

Hoang, D. T., Vinh, L. S., Flouri, T., et al. MPBoot: fast phylogenetic maximum parsimony tree inference and bootstrap approximation. *BMC evolutionary biology*, 18(1):1–11, 2018.

Hohenstein, P., Pritchard-Jones, K., and Charlton, J. The yin and yang of kidney development and Wilms' tumors. *Genes & development*, 29(5):467–482, 2015.

Holland, E. C. and Varmus, H. E. Basic fibroblast growth factor induces cell migration and proliferation after glia-specific gene transfer in mice. *Proceedings of the National Academy of Sciences*, 95(3):1218–1223, 1998.

Hurst, L. D. and McVean, G. T. Growth effects of uniparental disomies and the conflict theory of genomic imprinting. *Trends in Genetics*, 13(11):436–443, 1997.

ICGC TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82–93, 2020.

Iqbal, K., Jin, S.-G., Pfeifer, G. P., and Szabó, P. E. Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. *Proceedings of the National Academy of Sciences*, 108(9):3642–3647, 2011.

Iwata, K., Yumoto, K., Sugishima, M., et al. Analysis of compaction initiation in human embryos by using time-lapse cinematography. *Journal of assisted reproduction and genetics*, 31(4):421–426, 2014.

Jacobs, P. A., Brunton, M., Court Brown, W. M., Doll, R., and Goldstein, H. Change of human chromosome count distribution with age: evidence for a sex differences. *Nature*, 197:1080–1, 1963.

Jaiswal, S., Fontanillas, P., Flannick, J., et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N Engl J Med*, 371(26):2488–98, 2014.

Jauniaux, E., Watson, A. L., Hempstock, J., et al. Onset of maternal arterial blood flow and placental oxidative stress. a possible factor in human early pregnancy failure. *Am J Pathol*, 157(6):2111–22, 2000.

Jones, D., Raine, K. M., Davies, H., et al. cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. *Curr Protoc Bioinformatics*, 56:15 10 1–15 10 18, 2016.

Ju, Y. S., Martincorena, I., Gerstung, M., et al. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature*, 543(7647):714–718, 2017.

Jukam, D., Shariati, S. A. M., and Skotheim, J. M. Zygotic genome activation in vertebrates. *Developmental cell*, 42(4):316–332, 2017.

Kalousek, D. K. and Dill, F. J. Chromosomal mosaicism confined to the placenta in human conceptions. *Science*, 221(4611):665–7, 1983.

Kalousek, D. K., Howard-Peebles, P. N., Olson, S. B., et al. Confirmation of CVS mosaicism in term placentae and high frequency of intrauterine growth retardation association with confined placental mosaicism. *Prenat Diagn*, 11(10):743–50, 1991.

Keller, P. J., Schmidt, A. D., Wittbrodt, J., and Stelzer, E. H. Reconstruction of zebrafish early embryonic development by scanned light sheet microscopy. *science*, 322(5904): 1065–1069, 2008.

Kimmel, C. B., Warga, R. M., and Schilling, T. F. Origin and organization of the zebrafish fate map. *Development*, 108(4):581–594, 1990.

Knudson, J., A. G. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A*, 68(4):820–3, 1971.

Kobayashi, T. and Surani, M. A. On the origin of the human germline. *Development*, 145 (16), 2018.

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., and Teichmann, S. A. The technology and biology of single-cell RNA sequencing. *Molecular cell*, 58(4):610–620, 2015.

Kovacs, G., Akhtar, M., Beckwith, B. J., et al. The Heidelberg classification of renal cell tumours. *J Pathol*, 183(2):131–3, 1997.

Kretzschmar, K. and Watt, F. M. Lineage tracing. *Cell*, 148(1-2):33–45, 2012.

Kuijk, E., Blokzijl, F., Jager, M., et al. Early divergence of mutational processes in human fetal tissues. *Science advances*, 5(5):eaaw1271, 2019.

Kwon, G. S., Viotti, M., and Hadjantonakis, A.-K. The endoderm of the mouse embryo arises by dynamic widespread intercalation of embryonic and extraembryonic lineages. *Developmental cell*, 15(4):509–520, 2008.

Laks, E., McPherson, A., Zahn, H., et al. Clonal decomposition and DNA replication states defined by scaled single-cell genome sequencing. *Cell*, 179(5):1207–1221, 2019.

Larsen, E. C., Christiansen, O. B., Kolte, A. M., and Macklon, N. New insights into mechanisms behind miscarriage. *BMC medicine*, 11(1):154, 2013.

Lawson, A. R. J., Abascal, F., Coorens, T. H. H., et al. Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science*, 370(6512):75–82, 2020.

Lee, H. J., Hore, T. A., and Reik, W. Reprogramming the methylome: erasing memory and creating diversity. *Cell stem cell*, 14(6):710–719, 2014.

Lee-Six, H., Obro, N. F., Shepherd, M. S., et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature*, 561(7724):473–478, 2018.

Lee-Six, H., Olafsson, S., Ellis, P., et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature*, 574(7779):532–537, 2019.

Lemischka, I. R., Raulet, D. H., and Mulligan, R. C. Developmental potential and dynamic behavior of hematopoietic stem cells. *Cell*, 45(6):917–927, 1986.

Li, H. and Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–60, 2009.

Lodato, M. A., Woodworth, M. B., Lee, S., et al. Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science*, 350(6256):94–98, 2015.

Lodato, M. A., Rodin, R. E., Bohrson, C. L., et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science*, 359(6375):555–559, 2018.

Loftfield, E., Zhou, W., Graubard, B. I., et al. Predictors of mosaic chromosome Y loss and associations with mortality in the UK Biobank. *Sci Rep*, 8(1):12316, 2018.

Los, F. J., Van Opstal, D., Van Den Berg, C., et al. Uniparental disomy with and without confined placental mosaicism: a model for trisomic zygote rescue. *Prenatal Diagnosis: Published in Affiliation With the International Society for Prenatal Diagnosis*, 18(7): 659–668, 1998.

Love, M. I., Huber, W., and Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, 15(12):550, 2014.

Ludwig, L. S., Lareau, C. A., Ulirsch, J. C., et al. Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell*, 176(6):1325–1339, 2019.

Maher, E. R., Reik, W., et al. Beckwith-Wiedemann syndrome: imprinting in clusters revisited. *The Journal of clinical investigation*, 105(3):247–252, 2000.

Margol, A. S. and Judkins, A. R. Pathology and diagnosis of SMARCB1-deficient tumors. *Cancer genetics*, 207(9):358–364, 2014.

Martincorena, I., Roshan, A., Gerstung, M., et al. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*, 348(6237):880–6, 2015.

Martincorena, I., Raine, K. M., Gerstung, M., et al. Universal patterns of selection in cancer and somatic tissues. *Cell*, 171(5):1029–1041 e21, 2017.

Martincorena, I., Fowler, J. C., Wabik, A., et al. Somatic mutant clones colonize the human esophagus with age. *Science*, 362(6417):911–917, 2018.

Mazzarello, P. A unifying concept: the history of cell theory. *Nature cell biology*, 1(1): E13–E15, 1999.

McKenna, A., Findlay, G. M., Gagnon, J. A., et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*, 353(6298):aaf7907, 2016.

McLaren, W., Gil, L., Hunt, S. E., et al. The Ensembl Variant Effect Predictor. *Genome Biol*, 17(1):122, 2016.

Milewski, R. and Ajduk, A. Time-lapse imaging of cleavage divisions in embryo quality assessment. *Reproduction*, 154(2):R37–R53, 2017.

Molè, M. A., Coorens, T. H. H., Shahbazi, M. N., et al. A single cell characterisation of human embryogenesis identifies pluripotency transitions and putative anterior hypoblast centre. *Nature Communications*, 12(1):1–12, 2021.

Moore, L., Leongamornlert, D., Coorens, T. H. H., et al. The mutational landscape of normal human endometrial epithelium. *Nature*, 580(7805):640–646, 2020.

Morris, R. J., Liu, Y., Marles, L., et al. Capturing and profiling adult hair follicle stem cells. *Nat Biotechnol*, 22(4):411–7, 2004.

Morris, S. A., Teo, R. T., Li, H., et al. Origin and formation of the first two distinct cell types of the inner cell mass in the mouse embryo. *Proceedings of the National Academy of Sciences*, 107(14):6364–6369, 2010.

Mottla, G., Adelman, M., Hall, J., et al. Lineage tracing demonstrates that blastomeres of early cleavage-stage human pre-embryos contribute to both trophectoderm and inner cell mass. *Human reproduction (Oxford, England)*, 10(2):384–391, 1995.

Moulton, T., Crenshaw, T., Hao, Y., et al. Epigenetic lesions at the H19 locus in Wilms' tumour patients. *Nat Genet*, 7(3):440–7, 1994.

Nakamura, T., Okamoto, I., Sasaki, K., et al. A developmental coordinate of pluripotency among mice, monkeys and humans. *Nature*, 537(7618):57–62, 2016.

Narod, S. A. and Lenoir, G. M. Are bilateral tumours hereditary? *Int J Epidemiol*, 20(2): 346–8, 1991.

Needham, J. and Hughes, A. *A history of embryology*. Cambridge University Press, 2015.

Nicholson, A. M., Olpe, C., Hoyle, A., et al. Fixation and spread of somatic mutations in adult human colonic epithelium. *Cell Stem Cell*, 22(6):909–918 e8, 2018.

Nik-Zainal, S., Van Loo, P., Wedge, D. C., et al. The life history of 21 breast cancers. *Cell*, 149(5):994–1007, 2012.

Nik-Zainal, S., Davies, H., Staaf, J., et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47–54, 2016.

Novelli, M., Cossu, A., Oukrif, D., et al. X-inactivation patch size in human female tissue confounds the assessment of tumor clonality. *Proceedings of the National Academy of Sciences*, 100(6):3311–3314, 2003.

Nowell, P. C. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976.

Nowotschin, S. and Hadjantonakis, A. K. Guts and gastrulation: Emergence and convergence of endoderm in the mouse embryo. *Curr Top Dev Biol*, 136:429–454, 2020.

Okamoto, K., Morison, I. M., Taniguchi, T., and Reeve, A. E. Epigenetic changes at the insulin-like growth factor II/H19 locus in developing kidney is an early event in Wilms tumorigenesis. *Proc Natl Acad Sci U S A*, 94(10):5367–71, 1997.

Olafsson, S., McIntyre, R. E., Coorens, T., et al. Somatic Evolution in Non-neoplastic IBD-Affected Colon. *Cell*, 182(3):672–684 e11, 2020.

Osorio, F. G., Rosendahl Huber, A., Oka, R., et al. Somatic mutations reveal lineage relationships and age-related mutagenesis in human hematopoiesis. *Cell Rep*, 25(9): 2308–2316 e4, 2018.

Paradis, E., Claude, J., and Strimmer, K. Ape: analyses of phylogenetics and evolution in r language. *Bioinformatics*, 20(2):289–290, 2004.

Pasupathy, D., Dacey, A., Cook, E., et al. Study protocol. a prospective cohort study of unselected primiparous women: the pregnancy outcome prediction study. *BMC Pregnancy Childbirth*, 8:51, 2008.

Pei, W., Feyerabend, T. B., Rössler, J., et al. Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. *Nature*, 548(7668):456, 2017.

Petljak, M., Alexandrov, L. B., Brammeld, J. S., et al. Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. *Cell*, 176(6):1282–1294, 2019.

Pich, O., Muiños, F., Lolkema, M. P., et al. The mutational footprints of cancer therapies. *Nature genetics*, 51(12):1732–1740, 2019.

Pijuan-Sala, B., Griffiths, J. A., Guibentif, C., et al. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*, 566(7745):490–495, 2019.

Pleguezuelos-Manzano, C., Puschhof, J., Rosendahl Huber, A., et al. Mutational signature in colorectal cancer caused by genotoxic pks(+) E. coli. *Nature*, 580(7802):269–273, 2020.

Poetsch, A. R. The genomics of oxidative DNA damage, repair, and resulting mutagenesis. *Comput Struct Biotechnol J*, 18:207–219, 2020.

Popescu, D.-M., Botting, R. A., Stephenson, E., et al. Decoding human fetal liver haematopoiesis. *Nature*, 574(7778):365–371, 2019.

Rahbari, R., Wuster, A., Lindsay, S. J., et al. Timing, rates and spectra of human germline mutation. *Nat Genet*, 48(2):126–133, 2016.

Raj, B., Wagner, D. E., McKenna, A., et al. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat Biotechnol*, 36(5):442–450, 2018.

Rheinbay, E., Nielsen, M. M., Abascal, F., et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature*, 578(7793):102–111, 2020.

Robinson, P. S., Coorens, T. H. H., Palles, C., et al. Elevated somatic mutation burdens in normal human cells due to defective DNA polymerases. *bioRxiv*, 2020.

Roe, S. A. *Matter, life, and generation: Eighteenth-century embryology and the Haller-Wolff debate*. Cambridge University Press, 2003.

Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S., and Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol*, 17:31, 2016.

Rouhani, F. J., Nik-Zainal, S., Wuster, A., et al. Mutational history of a human cell lineage from somatic to induced pluripotent stem cells. *PLoS genetics*, 12(4):e1005932, 2016.

Schleiden, M. J. *Beiträge zur phytogenesis*. 1838.

Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics*, 27(4):592–593, 2011.

Schwann, T. Mikroskopische Untersuchungen über die Uebereinstimmung in der Struktur und dem Wachsthum der Thiere und Pflanzen. *Sander, Berlin*, 268, 1839.

Serbedzija, G. N., Bronner-Fraser, M., and Fraser, S. E. A vital dye analysis of the timing and pathways of avian trunk neural crest cell migration. *Development*, 106(4):809–816, 1989.

Shahbazi, M. N. Mechanisms of human embryo development: from cell fate to tissue shape and back. *Development*, 147(14), 2020.

Shahbazi, M. N., Wang, T., Tao, X., et al. Developmental potential of aneuploid human embryos cultured beyond implantation. *Nature communications*, 11(1):1–15, 2020.

Short, K. M. and Smyth, I. M. The contribution of branching morphogenesis to kidney development and disease. *Nat Rev Nephrol*, 12(12):754–767, 2016.

Sidorov, G., Gelbukh, A., Gómez-Adorno, H., and Pinto, D. Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas*, 18(3): 491–504, 2014.

Sirchia, S., Garagiola, I., Colucci, G., et al. Trisomic zygote rescue revealed by DNA polymorphism analysis in confined placental mosaicism. *Prenatal Diagnosis: Published in Affiliation With the International Society for Prenatal Diagnosis*, 18(3):201–206, 1998.

Smith, H. O., Qualls, C. R., Prairie, B. A., et al. Trends in gestational choriocarcinoma: a 27-year perspective. *Obstet Gynecol*, 102(5 Pt 1):978–87, 2003.

Snippert, H. J., van der Flier, L. G., Sato, T., et al. Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell*, 143(1):134–44, 2010.

Sovio, U., Gaccioli, F., Cook, E., et al. Prediction of preeclampsia using the soluble fms-like tyrosine kinase 1 to placental growth factor ratio: A prospective cohort study of unselected nulliparous women. *Hypertension*, 69(4):731–738, 2017.

Sparks, A. B., Morin, P. J., Vogelstein, B., and Kinzler, K. W. Mutational analysis of the APC/β-catenin/Tcf pathway in colorectal cancer. *Cancer research*, 58(6):1130–1134, 1998.

Spencer Chapman, M., Ranzoni, A. M., Myers, B., et al. Lineage tracing of human development through somatic mutations. *Nature*, 595:85–90, 2020.

Stamatakis, A. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.

Stratton, M. R., Campbell, P. J., and Futreal, P. A. The cancer genome. *Nature*, 458(7239): 719–24, 2009.

Strnad, P., Gunther, S., Reichmann, J., et al. Inverted light-sheet microscope for imaging mouse pre-implantation development. *Nat Methods*, 13(2):139–42, 2016.

Sulston, J. E. and Horvitz, H. R. Post-embryonic cell lineages of the nematode, caenorhabditis elegans. *Dev Biol*, 56(1):110–56, 1977.

Suzuki, K., Tonien, D., Kurosawa, K., and Toyota, K. Birthday paradox for multi-collisions. In *International Conference on Information Security and Cryptology*, pages 29–40. Springer, 2006.

Swofford, D. L. Paup*: Phylogenetic analysis using parsimony (and other methods) 4.0. B5. 2001.

Swofford, D. L., Waddell, P. J., Huelsenbeck, J. P., et al. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Systematic biology*, 50(4):525–539, 2001.

Tavaré, S. Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on mathematics in the life sciences*, 17(2):57–86, 1986.

Thompson, D. J., Genovese, G., Halvardson, J., et al. Genetic predisposition to mosaic Y chromosome loss in blood. *Nature*, 575(7784):652–657, 2019.

Toutain, J., Goutte-Gattat, D., Horovitz, J., and Saura, R. Confined placental mosaicism revisited: Impact on pregnancy characteristics and outcome. *PLoS One*, 13(4):e0195905, 2018.

Treger, T. D., Chowdhury, T., Pritchard-Jones, K., and Behjati, S. The genetic changes of Wilms tumour. *Nat Rev Nephrol*, 15(4):240–251, 2019.

Van de Velde, H., Cauffman, G., Tournaye, H., Devroey, P., and Liebaers, I. The four blastomeres of a 4-cell stage human embryo are able to develop individually into blastocysts with inner cell mass and trophectoderm. *Human reproduction*, 23(8):1742–1747, 2008.

Van Loo, P., Nordgard, S. H., Lingjærde, O. C., et al. Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences*, 107(39):16910–16915, 2010.

Viotti, M., Nowotschin, S., and Hadjantonakis, A.-K. SOX17 links gut endoderm morphogenesis and germ layer segregation. *Nature cell biology*, 16(12):1146–1156, 2014.

Virchow, R. L. K. *Die Cellularpathologie in ihrer Begründung auf physiologische und pathologische Gewebelehre*. A. Hirschwald, 1859.

Voet, T., Vanneste, E., and Vermeesch, J. The human cleavage stage embryo is a cradle of chromosomal rearrangements. *Cytogenetic and genome research*, 133(2-4):160–168, 2011.

Vogt, W. Gestaltungsanalyse am Amphibienkeim mit örtlicher Vitalfärbung. II. Teil. Gastrulation und Mesodermbildung bei Urodelen und Anuren. *Arch Entw Mech*, 120:384–706, 1929.

von Mohl, H. *Ueber die Vermehrung der Pflanzen-Zellen durch Theilung*. Fues, 1835.

Watson, C. J., Papula, A., Poon, G. Y., et al. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science*, 367(6485):1449–1454, 2020.

Weaver, J. R., Susiarjo, M., and Bartolomei, M. S. Imprinting and epigenetic changes in the early embryo. *Mammalian Genome*, 20(9-10):532–543, 2009.

Weksberg, R., Shuman, C., and Beckwith, J. B. Beckwith-Wiedemann syndrome. *Eur J Hum Genet*, 18(1):8–14, 2010.

Whitman, C. O. *The embryology of Clepsine*. JE Adlard, 1878.

Wilcox, A. J., Harmon, Q., Doody, K., Wolf, D. P., and Adashi, E. Y. Preimplantation loss of fertilized human ova: estimating the unobservable. *Human Reproduction*, 35(4):743–750, 2020.

Xiang, L., Yin, Y., Zheng, Y., et al. A developmental landscape of 3d-cultured human pre-gastrulation embryos. *Nature*, 577(7791):537–542, 2020.

Xue, Z., Huang, K., Cai, C., et al. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature*, 500(7464):593–597, 2013.

Yamamoto, M., Shook, N. A., Kanisicak, O., et al. A multifunctional reporter mouse line for cre- and flp-dependent lineage analysis. *Genesis*, 47(2):107–14, 2009.

Yan, L., Yang, M., Guo, H., et al. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nature structural & molecular biology*, 20(9):1131, 2013.

Yang, Z. and Rannala, B. Molecular phylogenetics: principles and practice. *Nature Reviews Genetics*, 13(5):303, 2012.

Ye, A. Y., Dou, Y., Yang, X., et al. A model for postzygotic mosaicisms quantifies the allele fraction drift, mutation rate, and contribution to de novo mutations. *Genome Res*, 28(7): 943–951, 2018a.

Ye, K., Guo, L., Yang, X., et al. Split-read indel and structural variant calling using pindel. *Methods Mol Biol*, 1833:95–105, 2018b.

Yoshida, K., Gowers, K. H. C., Lee-Six, H., et al. Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature*, 578(7794):266–272, 2020.

Young, M. D., Mitchell, T. J., Vieira Braga, F. A., et al. Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *Science*, 361(6402):594–599, 2018.

Yu, G., Smith, D. K., Zhu, H., Guan, Y., and Lam, T. T.-Y. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, 8(1):28–36, 2017.

Zernicka-Goetz, M. Cleavage pattern and emerging asymmetry of the mouse embryo. *Nature reviews Molecular cell biology*, 6(12):919–928, 2005.

Zhang, Y. and Tycko, B. Monoallelic expression of the human H19 gene. *Nat Genet*, 1(1): 40–4, 1992.

Zhou, F., Wang, R., Yuan, P., et al. Reconstituting the transcriptome and DNA methylome landscapes of human implantation. *Nature*, 572(7771):660–664, 2019.

Zink, F., Stacey, S. N., Norddahl, G. L., et al. Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood*, 130(6):742–752, 2017.