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.

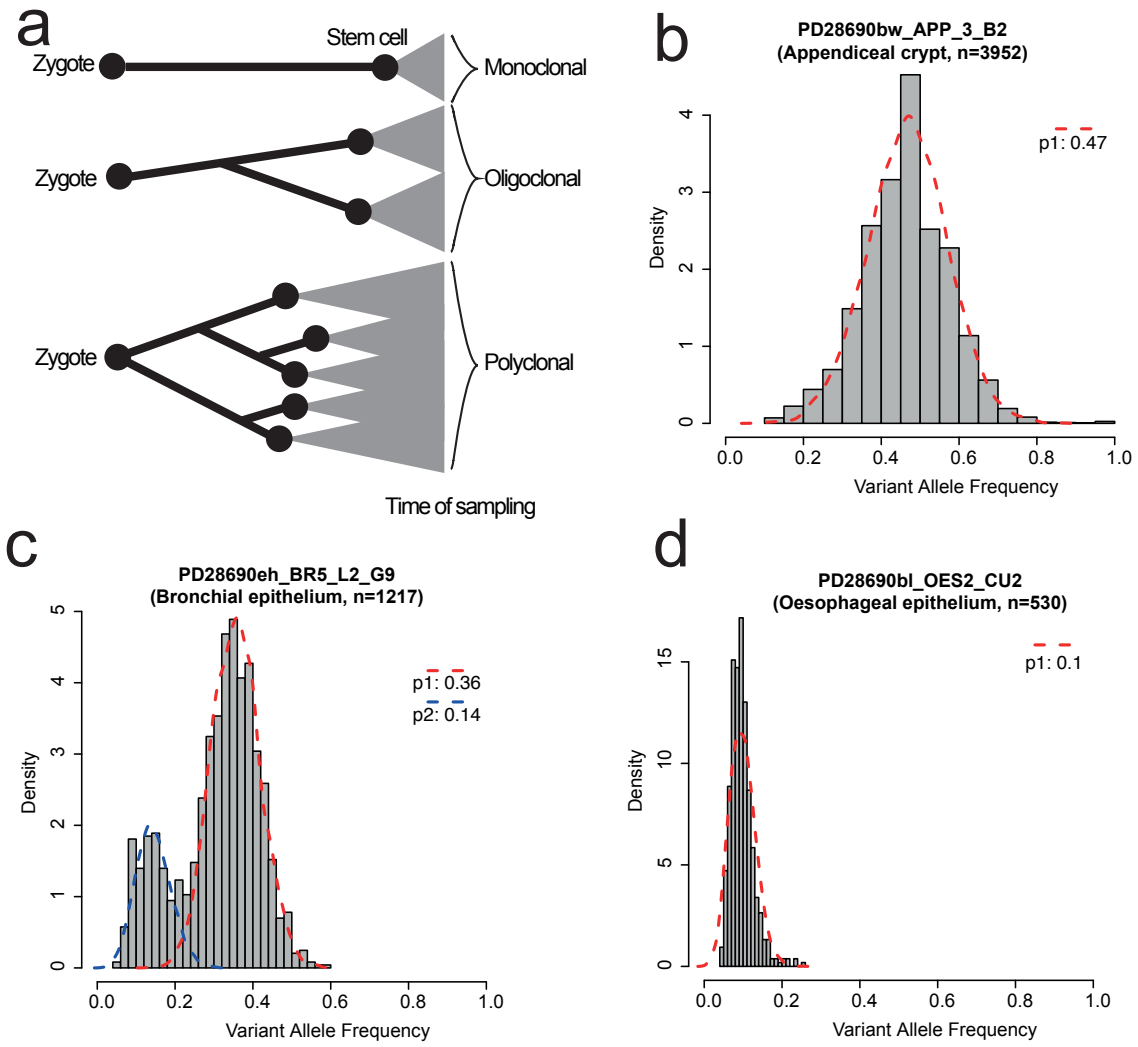

Figure 3.1 (a) Schematic of three different progenitor or stem cell contributions to the eventual sample. Monoclonal samples consist of the progeny of one cell, while oligo- and polyclonal are derived from a few and many progenitors, respectively. VAF histograms and binomial decompositions for a monoclonal (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.

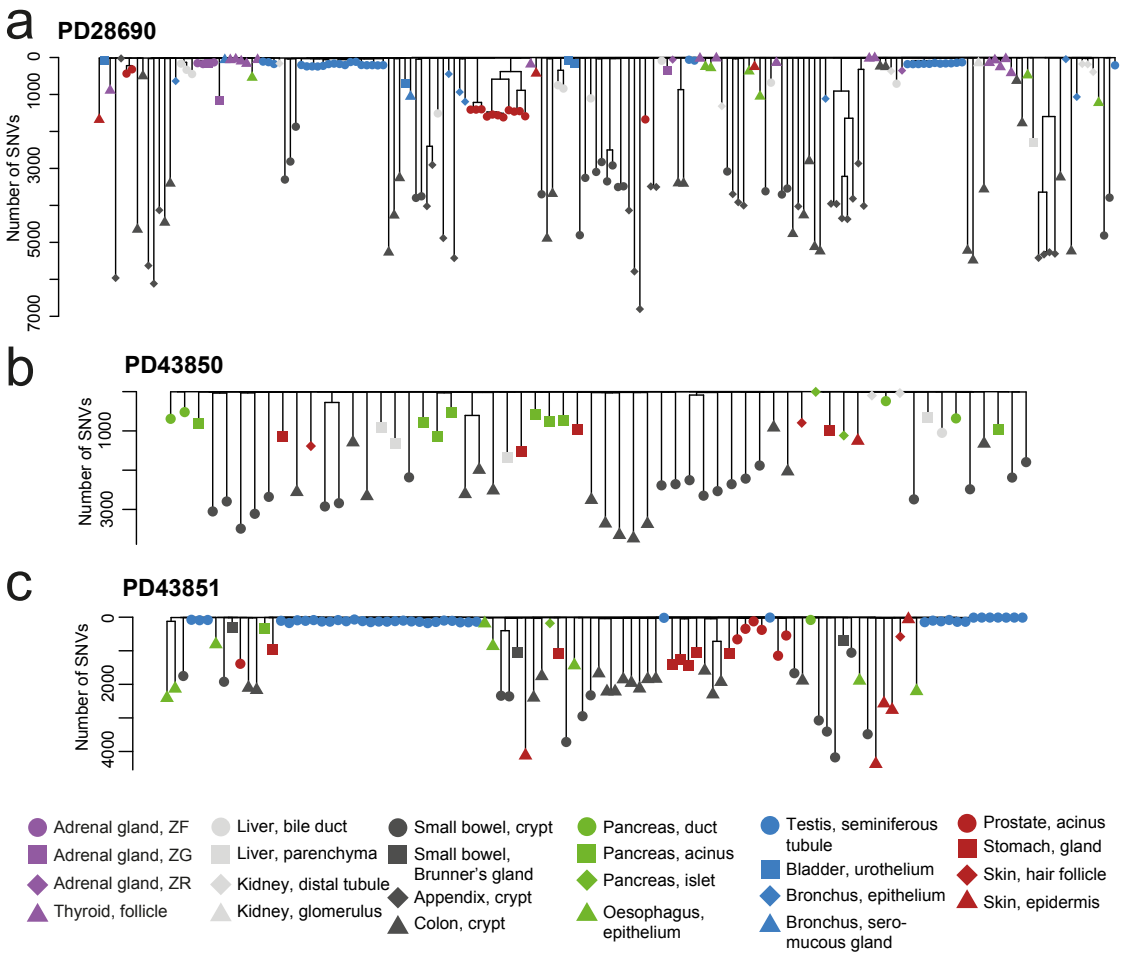

Figure 3.2 Phylogenetic trees of clonal LCM cuts of PD28690 (a), PD43850 (b) and PD43851 (c). The branch length conveys the number of SNVs acquired in that lineage. The shape and colour of the dots reflect the tissue type of the sample as per the legend.

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.

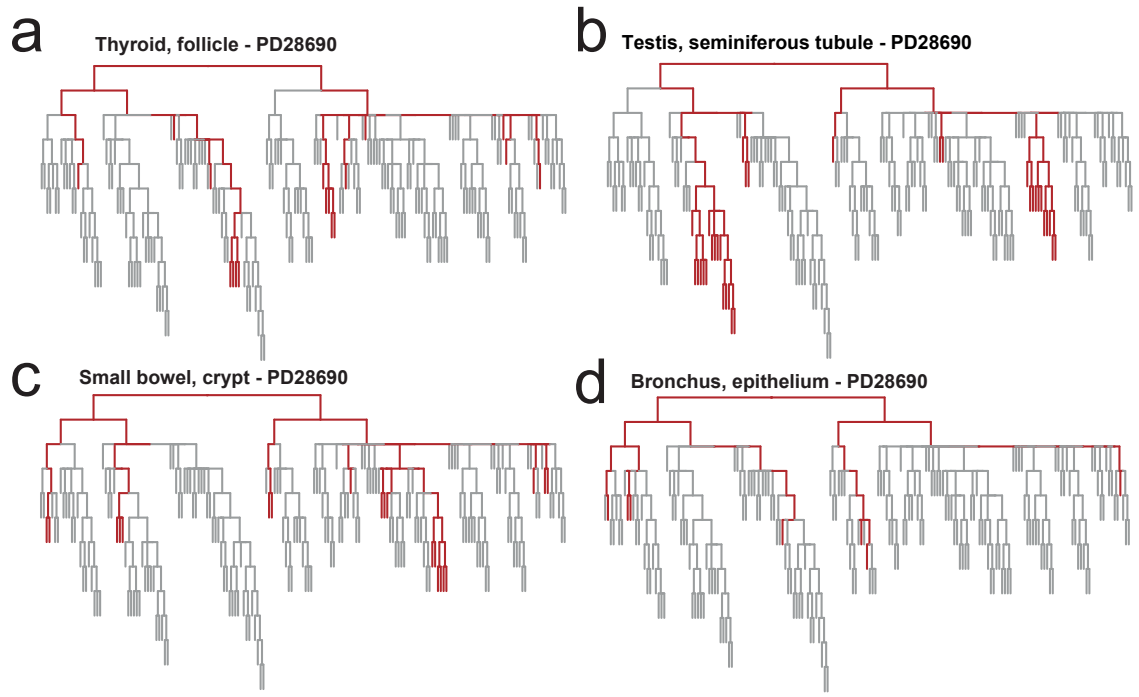

Figure 3.3 Phylogenetic trees with unit branch lengths for PD28690, showing the coalescence (red) of all samples from four tissues types: thyroid follicles (a), seminiferous tubules (b), small bowel crypts (c) and bronchal epithelium (d). The most recent common ancestor for all these tissues is the root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

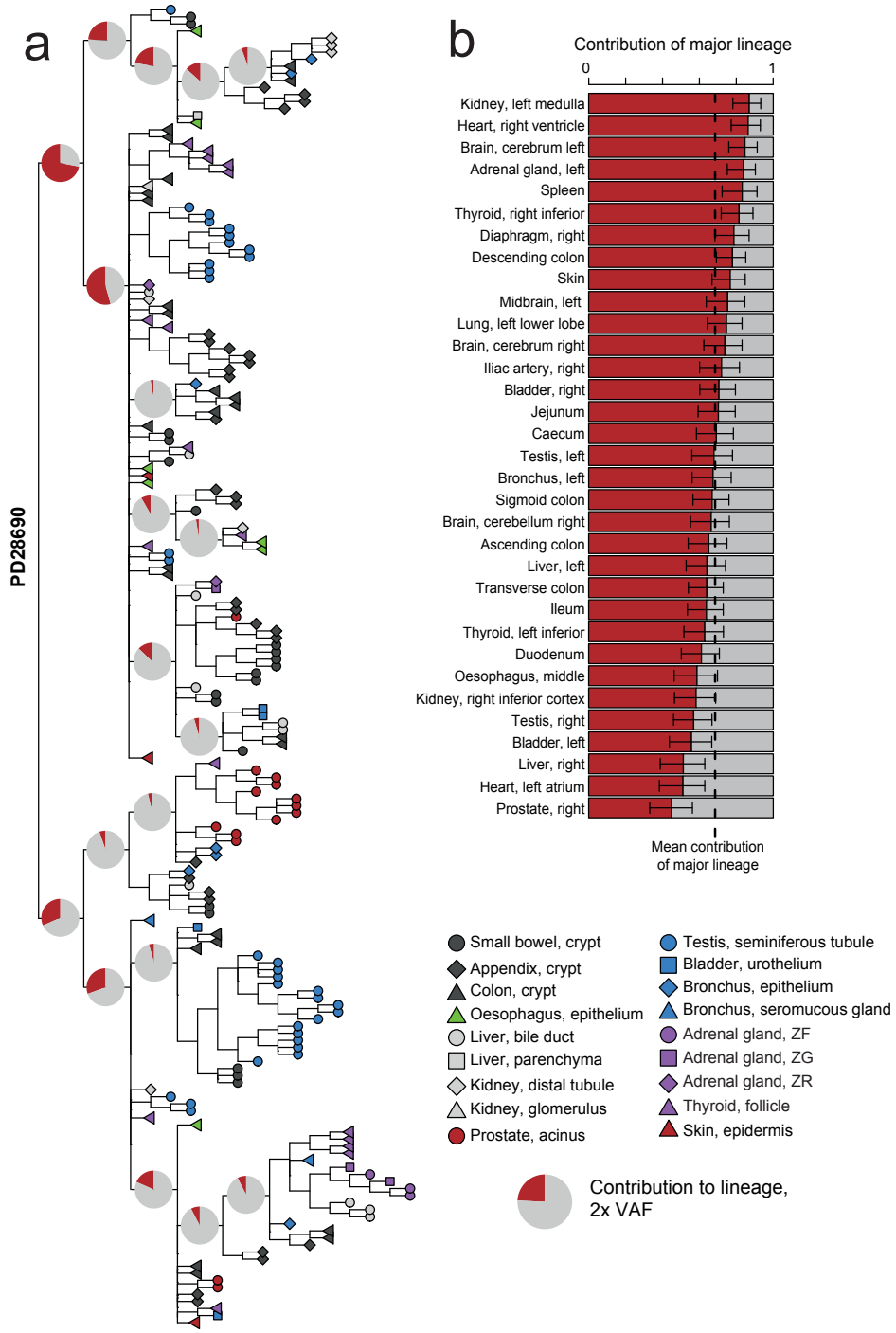

Figure 3.4 (a) Phylogenetic tree for PD28690. The shape and colour of the labels indicate tissue type. Pie charts show the mean contribution (twice the VAF) of that lineage to the 33 bulk biopsies of this patient. (b) The contribution of the major lineage to each bulk biopsy.

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).

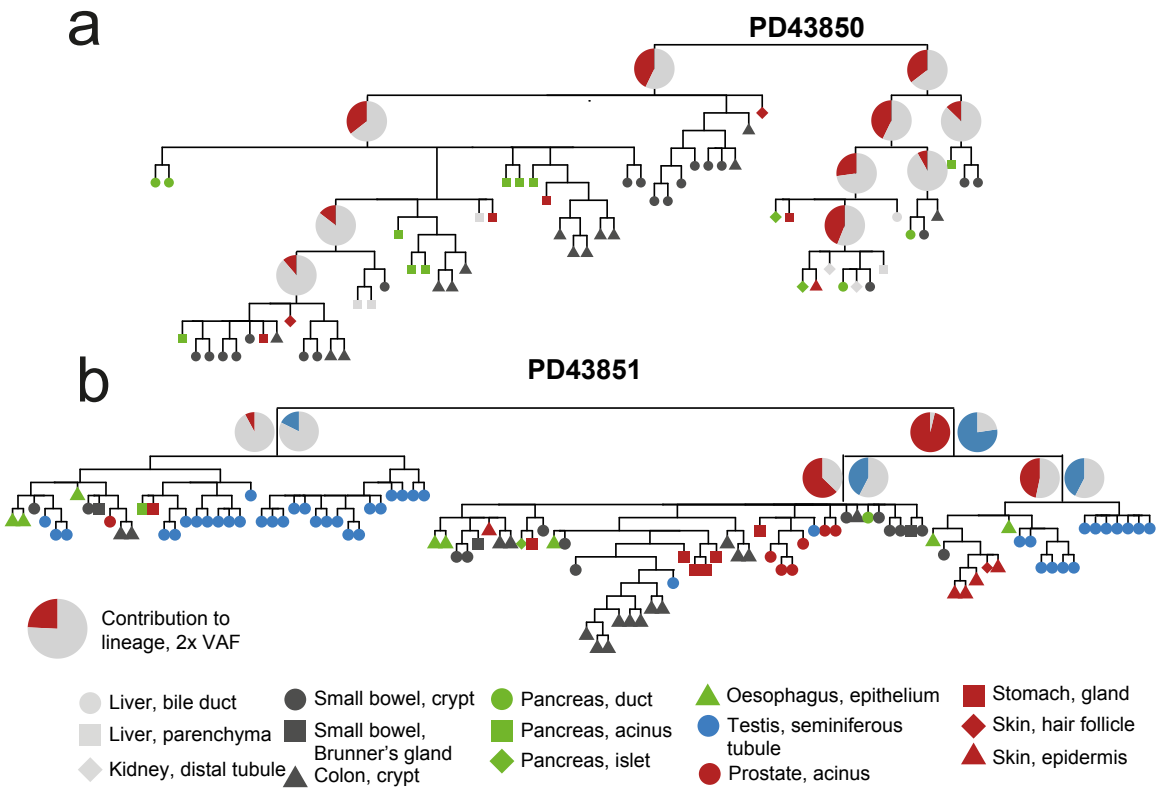

Figure 3.5 Phylogenetic tree topology for PD43850 (a) and PD43851 (b). The shape and colour of the tip labels reflect the tissue type of the sample as per the legend. Pie charts show the mean contribution (twice the VAF) to a bulk brain sample of PD43850, or a bulk brain (red) or colon (blue) sample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

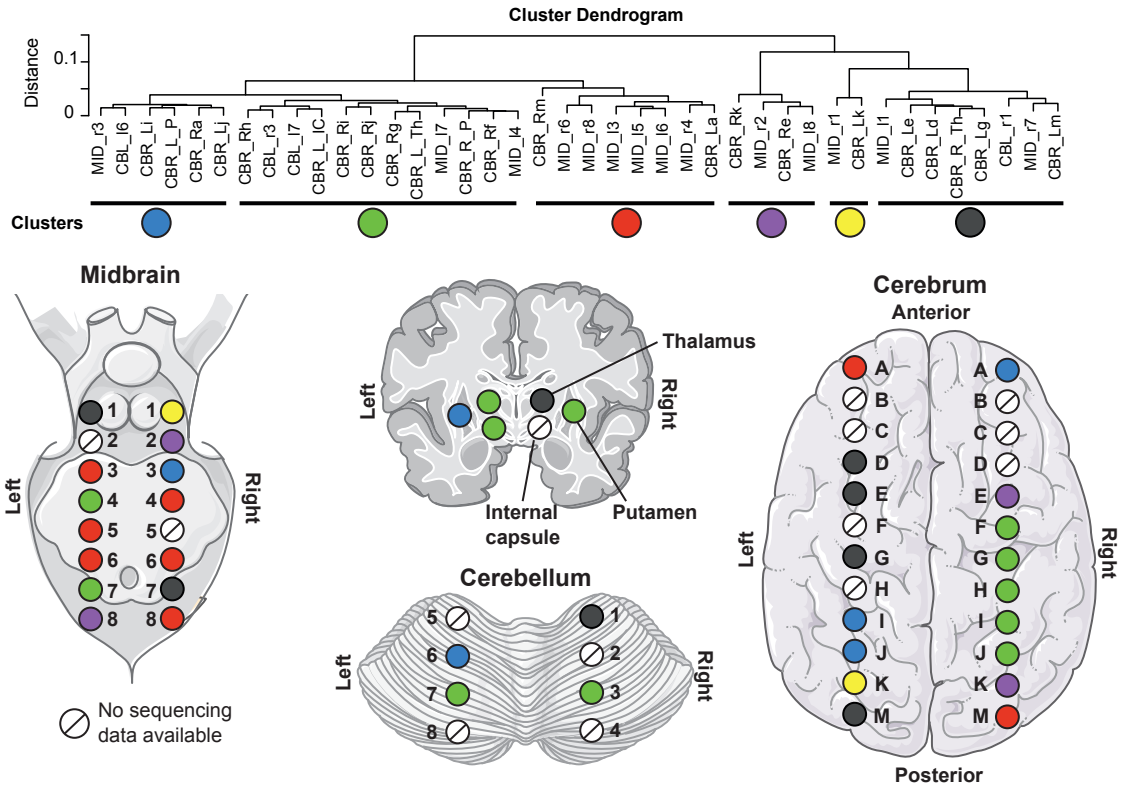

Figure 3.6 Cluster dendrogram of soft cosine similarity of VAFs of early embryonic mutations in bulk brain samples of PD28690. The data is split into six clusters, which are spatially displayed in the various brain regions. CBL=cerebellum, CBR=cerebrum, IC=internal capsule, L=left, MID=midbrain, P=putamen, 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).

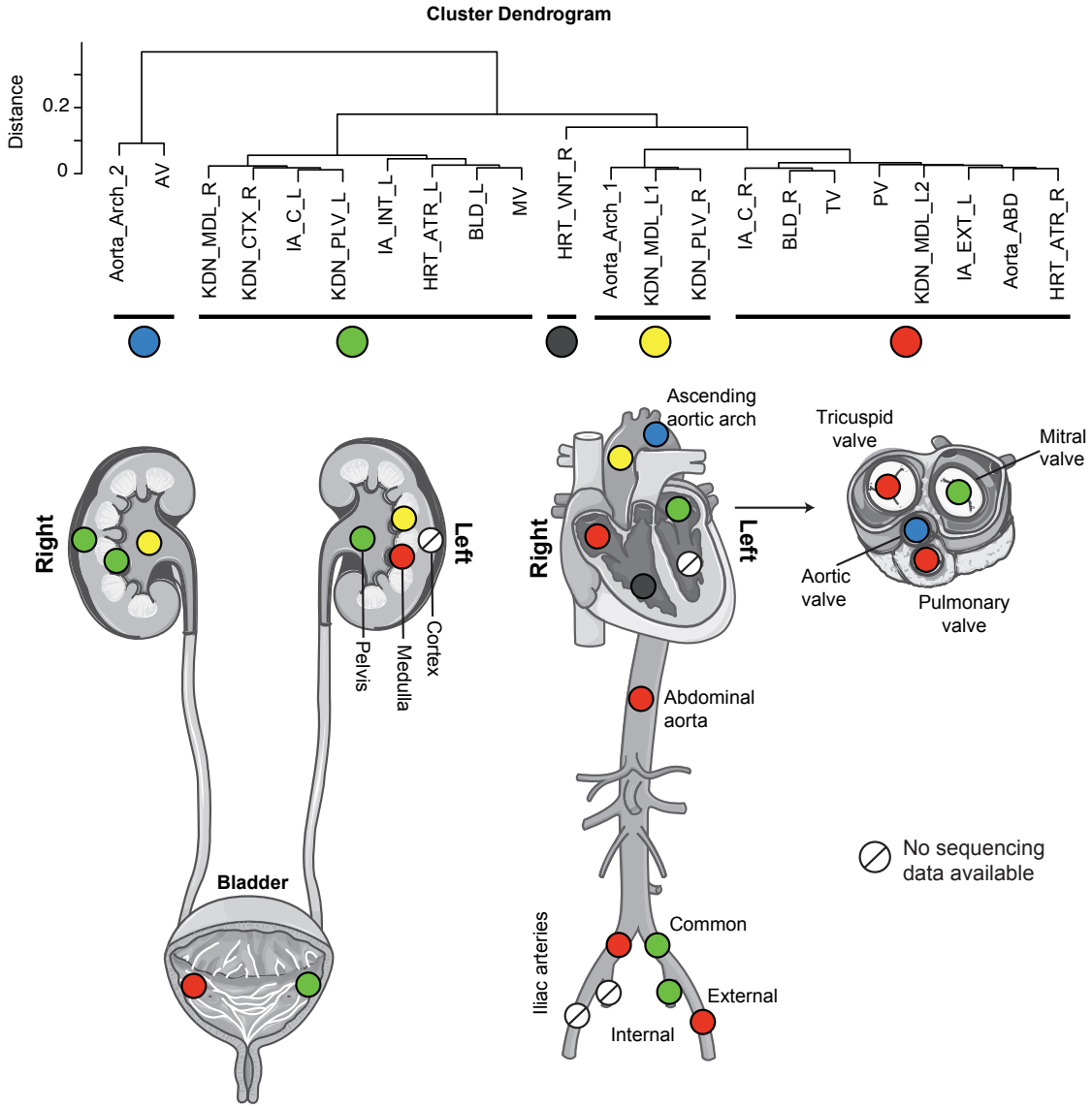

Figure 3.7 Cluster dendrogram of soft cosine similarity of VAFs of early embryonic mutations in mesodermal samples of PD28690. The data is split into five clusters, which are spatially displayed in the various organs and tissues derived from the mesoderm. ABD=abdominal, ATR=atrium, AV=aortic valve, BLD=bladder, 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.

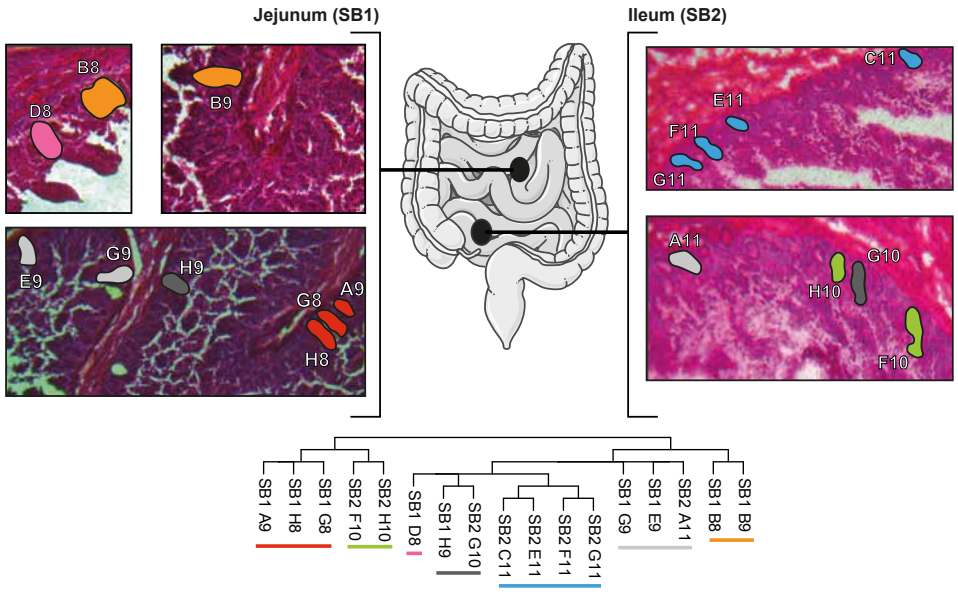

Figure 3.8 Histology sections of ileum and jejunum of PD28690 coloured by location on the phylogeny. Parts of the figure are composed of pictures from Servier Medical Art. Servier Medical Art by Servier is licensed under a Creative Commons Attribution 3.0 Unported License (https://creativecommons.org/licenses/by/3.0/).

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).

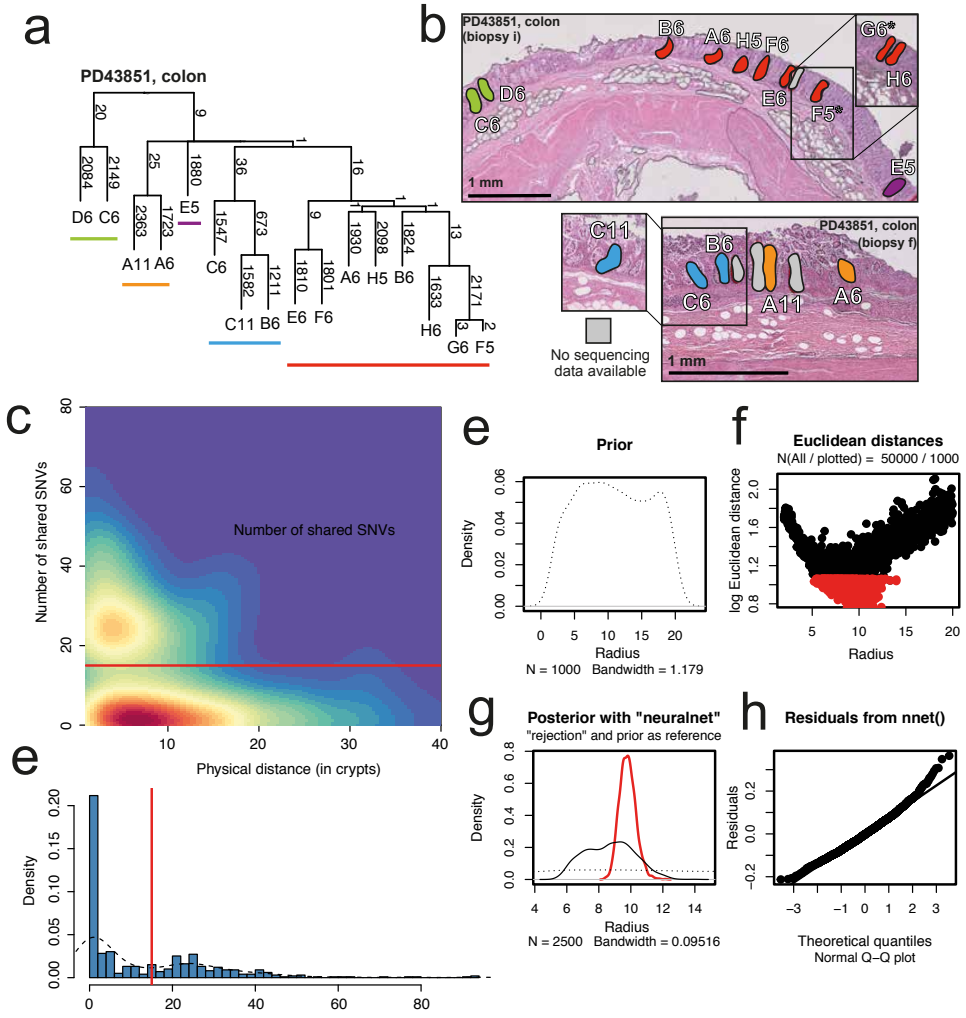

Figure 3.9 (a) Phylogeny of colonic crypts in PD43851, with coloured lines indicating distinct embryonic patches, each diverged after 20-46 SNVs. (b) Histology sections of samples shown in (a). Connected lines and boxes indicate z-stacks, with the asterisk indicating these samples were taken from different 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).

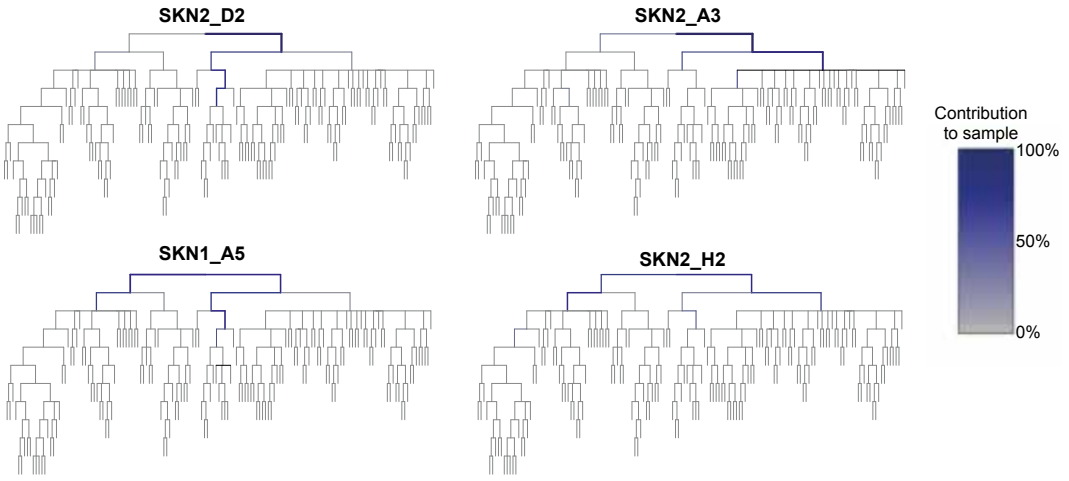

Figure 3.10 Phylogenetic trees with unit branch lengths for four polyclonal samples of epidermis of PD28690, showing the contribution (blue) of early embryonic progenitors in the phylogeny to the sample.

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.

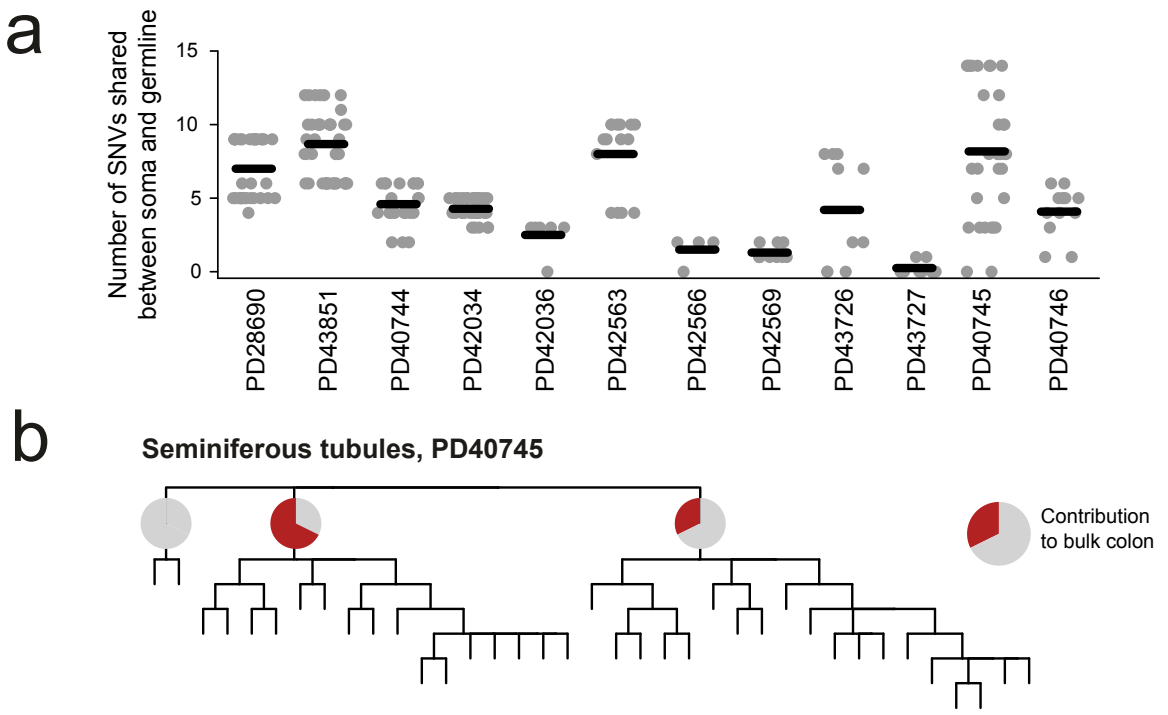

Figure 3.11 (a) Number of SNVs shared between any seminiferous tubule microdissection and another microdissection in PD28690 and PD43851. In other patients, number of SNVs shared between seminiferous tubule microdissections and matched bulk sample. (b) Phylogeny of seminiferous tubules of PD40735, with 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.

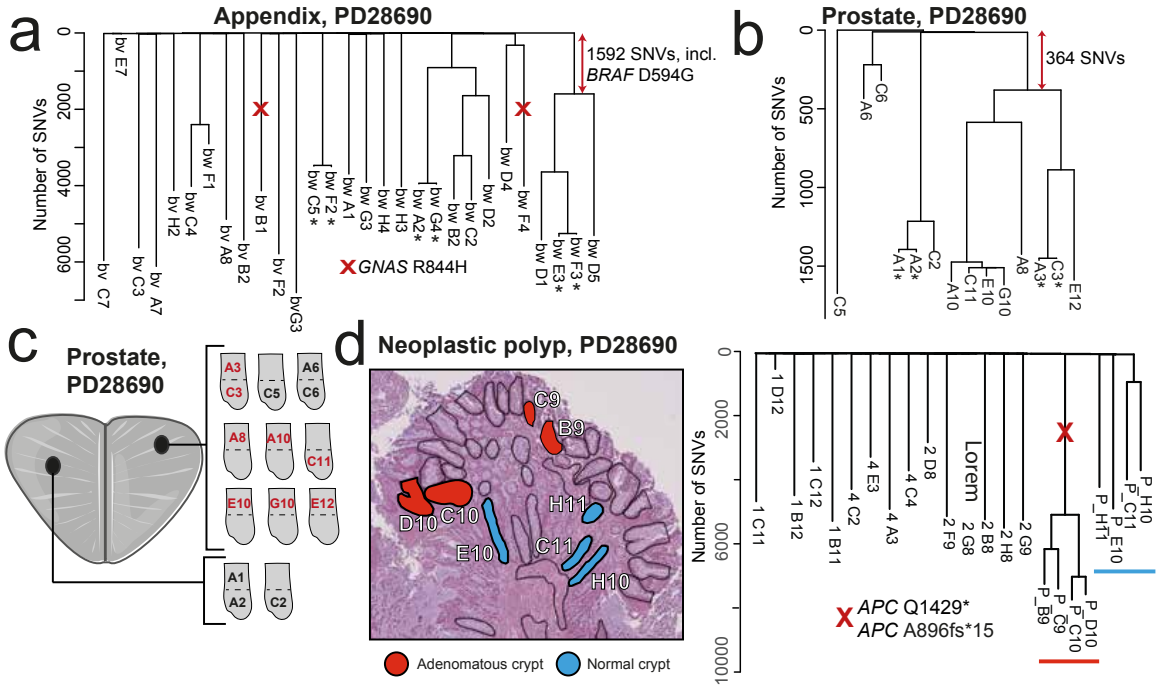

Figure 3.12 (a) Phylogenetic tree for appendiceal crypts in PD28690, with annotated cancer driver mutations. An asterisk indicates the two neighbouring crypts were taken as biological replicates of one another. (b) Phylogeny and sampling overview for prostatic acini in PD28690, showing widespread benign 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.

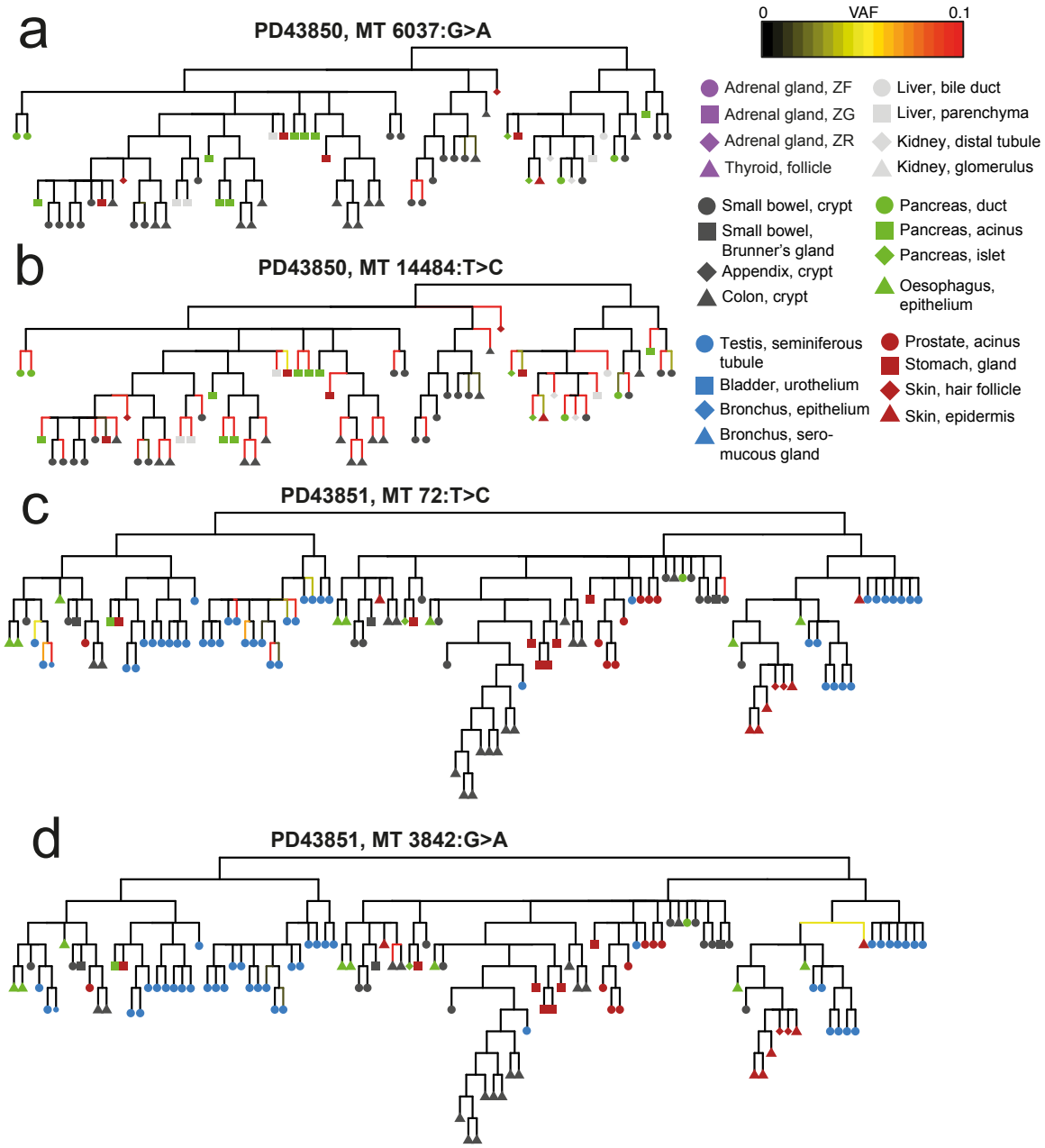

Figure 3.13 Phylogenies of nuclear SNVs with the VAF of mitochondrial mutations overlaid on them, showing a late shared SNVs (a), an SNV that was heteroplasmic in the zygote (b), an SNV that is consistent with a shared subclone or stromal contamination (c) and an SNV recurrently acquired in samples from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

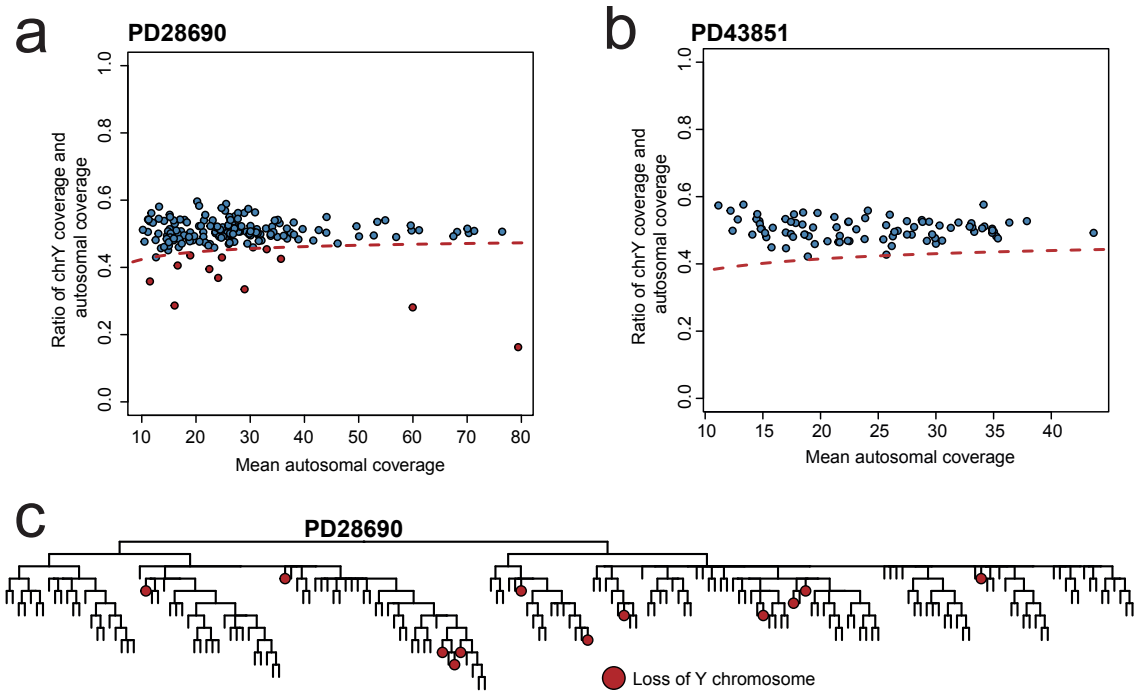

Figure 3.14 Scatterplot showing the ratio between the mean Y-chromosomal coverage and autosomal coverage against the mean autosomal coverage for all samples of PD28690 (a) and PD43851 (b). The dashed red lines indicates the 95% confidence interval around an expected ratio of 0.5. Red-coloured dots indicate 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.