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 $\geq$ 0.3, '0' (absent) for VAF $\leq$ 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).