INTRODUCTION

1. Somatic evolution

> *One general law, leading to the advancement of all organic beings, namely, multiply, vary, let the strongest live and the weakest die.*
>
> (Charles Darwin, *On the Origin of Species*).

Darwin's theory applies not only to whole organisms but also to the cells that constitute them; this is what is meant by 'somatic evolution'. Variation and selection occur within the body and so evolution must follow. Cancer represents the archetype of this process, although the principles apply beyond the disease, and it is through the study of cancer genetics that our and other laboratories have approached the concept of somatic evolution more broadly. Much of this dissertation is framed in terms of cancer, therefore, and, for these reasons, I begin with a discussion of cancer genomics.

2. Cancer genomics

2.a. Cancer is a disease of the genome

Cancer has been viewed as a disorder of cells since the nineteenth century. In 1840, Langenbeck stated 'every single carcinoma cell must now appear as an organism endowed with life-force and developmental ability' (cited in Bignold 2006). A series of discoveries over the following 150 years demonstrated that mutations were the molecular basis of cancer. Chromosomes were discovered in 1870, and the observation that the sum of the numbers of chromosomes in sperm and egg equalled that in the zygote suggested that they contributed the individuality of the tissue. In 1890, von Hansemann described abnormal mitoses in cancer cells and hypothesised that an imbalance in chromosome numbers might be important (von Hansemann 1890). A decade later, Boveri observed that sea urchins with multipolar mitoses (due to two spermatozoa fertilising an egg) had developmental defects and went on to propose that cancer

might be due to an 'abnormal chromosome complex' (Boveri 1914). He further postulated the existence of tumour suppressor genes and oncogenes:

> [...] in every normal cell there is a specific arrangement for inhibiting, which allows the process of division to begin only when the inhibition has been overcome by a special stimulus. To assume the presence of definite chromosomes which inhibit division, would harmonize best with my fundamental idea [...] Cells of tumours with unlimited growth would arise if those 'inhibiting chromosomes' were eliminated [...] On the other hand, the assumption of the existence of chromosomes which promote division, might satisfy this postulate [...] If three or four such chromosomes meet, the whole number of chromosomes being otherwise normal, then the tendency to rapid proliferation would arise.
>
> (Boveri 1914, cited in Knudson 2001)

It was observed that mutagenic exposures (first ionising radiation (Muller 1927) and then chemicals such as those in coal tar (reviewed in Loeb and Harris 2008)) were frequently carcinogenic. The discovery of the Philadelphia chromosome (Nowell and Hungerford 1960) showed that a specific chromosomal alteration could be associated with cancer, providing evidence for the concept of driver mutations. In 1971, Knudson fitted Poisson models to the number of tumours found in hereditary and sporadic cases of retinoblastoma and noted that the data were consistent with the need for two mutations occurring at the same rate to develop the disease (of which one was already inherited in familial cases). Identification of the *RB1* gene suggested that these were two inactivating hits in the same gene rather than mutations in two separate genes (Friend et al. 1986). This was the first tumour suppressor gene. In the same period, revival of work from the beginning of the century on tumour viruses indicated that specific genes carried by viruses, homologous to – but different from – host genes, were capable of causing cancer (Martin 1970; Stehelin et al. 1976). Similarly, it was noted that moving DNA from cancer cells into a phenotypically normal cell line could transform it, indicating the existence of dominantly-acting oncogenes (Krontiris and Cooper 1981; Shih et al. 1981). Isolation of the causal gene (*HRAS*) from a bladder carcinoma cell line and the identification of the particular base change with oncogenic activity resulted in the description of the first oncogenic point mutation (Reddy et al. 1982). In the decades since, through first targeted and later systematic analyses of thousands of genes across many different cancer histologies the number of cancer driver genes has risen to a few hundred

(and thus 1-3% of all genes), although this is still a matter of debate (Stratton et al. 2009; Vogelstein et al. 2013; Forbes et al. 2017). The number of driver mutations per tumour is also contentious. This is discussed in the context of colorectal adenocarcinoma in Results Chapter 2.

2.b. Mutational processes

Genome sequencing has revealed that only a small fraction of mutations in a tumour are driver mutations. The vast majority are passenger mutations, exerting no effect (or, at least, one that is too weak to be selected over the lifespan of a cancer) on the fitness of the cells that carry them. There is substantial variability between and within different cancer types in mutation burden, but most adult cancers have thousands to tens of thousands of single base substitutions, tens to hundreds of dinucleotide substitutions (where two adjacent bases both mutate), hundreds to thousands of small insertions and deletions (indels), and up to a few hundred structural variants (Alexandrov et al. 2018; Li et al 2017). In addition, there may be insertion of viral sequences or reactivation of endogenous mobile elements, and most cancers display changes in their epigenome relative to their cells of origin, but this is beyond the scope of the present dissertation and will not be discussed here.

Although most are innocuous, these scattered mutations nonetheless represent the substrate for evolution. Studying their properties tells us about the processes that cause the mutations that do exert a functional effect and provides a window into the basis of the 'variation' component of somatic evolution. The vast numbers of somatic mutations yielded by large scale sequencing efforts permits a systematic analysis of the mutational processes that cause cancer. To date, approximately 85 million somatic mutations of all classes have been described across thousands of cancers (Alexandrov et al. 2018; Li et al. 2017), from which signatures of mutational processes (or mutational signatures) can be extracted. Here, I explain mutational signatures in the context of single base substitutions, but the same methodology can be applied to all mutation types.

The analysis of mutational signatures rests upon the idea that different mutational processes mutate the genome in distinct ways, such that their activity leaves a characteristic 'signature' in the genome. To begin to describe signatures, we must first classify mutations. For example, substitutions can be grouped according the identity of the base change, presented with the

pyrimidine of the mutated Watson-Crick base pair as the original base, into six classes: C>A, C>G, C>T, T>A, T>C, and T>G substitutions. It has long been recognised that particular mutational processes have a predilection for some base changes over others. For example, *TP53* mutations in skin cancers are often C>T, while *TP53* in lung cancers from smokers – but not non-smokers – has a large number of C>A mutations (Alexandrov and Stratton 2014). These observations have a molecular basis. C>T mutations in skin cancers are thought to be due to a combination of, firstly, ultraviolet irradiation inducing cyclobutane-pyrimidine dimers in which the C is unstable and spontaneously deaminates to a uracil, opposite which an A is added in replication, and secondly, error-prone translesion polymerase bypass of ultraviolet-induced photolesions (Ikehata and Ono 2011). C>A mutations associated with smoking are thought to be due to bulky adducts such as benzo(a)pyrene metabolites binding guanines. The damaged guanine is frequently complemented by an adenine when replicated by the translesion polymerase POLH (Christmann et al. 2016).

Mutational processes with a reasonably well-characterised origin and molecular explanation, such as ultraviolet irradiation and smoking, were the exception rather than the rule before the advent of mutational signature analysis, and most somatic mutations in cancer genomes were of unknown cause. Even now, the aetiology of approximately half of mutational signatures is unknown (Petljak and Alexandrov 2016). It is not the case that the genome of one cancer represents one mutational process: most cancer genomes seem to have been sculpted by a number of co-occurring processes, which will have overlapping features (for example, there may be multiple processes that all cause C>T mutations). We do not know *a priori* the pattern of mutations associated with each process, nor how much each contributes to a cancer genome. The only information we have is the counts of mutations in every category that we have defined in each cancer. Remarkably, though, by comparing the mutational patterns across thousands of cancer samples one can deconvolute these mutations into their constituent signatures.

Humans can deconvolute simple cases intuitively. Imagine that three tumours have been sequenced, and the mutations in each tabulated. If the mutations in tumour 1 are 50% C>A, 30% C>G, and 20% C>T, tumour 2 only has C>A mutations, and tumour 3 has 60% C>G and 40% C>T mutations, we would naturally explain this with two mutational processes: process 1 is C>A and process 2 is C>G and C>T, and each process accounts for half of the mutations in tumour 1. For cancer genomes, which are far more complicated, computational methods are required. The most widely-applied method to date is non-negative matrix factorisation (NNMF), used to perform

the first comprehensive and field-defining analysis of mutational signatures (Alexandrov et al. 2013), but many others along similar lines have arisen since (EMu (Fischer et al. 2013), signeR (Rosales et al. 2016), SignatureAnalyzer (Kim et al. 2016), HDP (Roberts 2018)). In this thesis, NNMF and a hierarchical Dirichlet process (HDP) are both used, and a brief explanation of their inner workings is supplied in the methods section (Methods, section 12).

For greater resolution of the distinction between different mutational processes, mutations can be divided into a larger number of biologically meaningful classes. For the analysis of substitutions, the base change is further subcategorised by the bases 3' and 5' of the mutated base, considering, for example, C>T in an A<u>C</u>G (the mutated base is underlined) context differently from C>T in A<u>C</u>C. This allows, for example, discrimination between the C>T mutations in an N<u>C</u>G context associated with the hydrolytic deamination of 5-methylcytosine and the C>T mutations in an N<u>C</u>C context associated with ultraviolet irradiation and thus the separation of these two processes. Depending on the richness of the dataset and the behaviour of mutational processes (it may be, for example, that widening the context beyond a certain point adds no extra information), finer categories may be defined by integrating topological features of the genome such as transcription strand bias or positioning relative to histones.

So far, using a trinucleotide categorisation of substitutions, an analysis of 4,645 whole genomes and 19,184 exomes has resulted in the description of 49 single base substitution signatures, as well as 11 doublet base substitution signatures and 17 indel signatures (Alexandrov et al. 2018). Similarly, nine signatures of structural variation have been characterised (Li et al. 2017). Some signatures are of extrinsic origin (such as the one due to smoking) and some are of intrinsic origin (such as the spontaneous deamination of 5-methylcytosine). Particular signatures will be discussed at relevant points in this thesis.

It should be noted that mutational signatures represent sets of correlated features that need not be a result of a single molecular process. If two separate carcinogens that left different imprints on the genome were always present at exactly the same proportion in tobacco, their combined effect would always be extracted as one signature. They could only be separated if samples were found in which one process had predominated over the other. We may, therefore, expect some signatures to split into two as more samples are sequenced, which has indeed occurred when updating the TCGA analysis (Alexandrov et al. 2013) to the PCAWG analysis (Alexandrov et al.

2018). In due course, signatures should stabilise and their experimental validation (by, for example, exposing cell lines to putative carcinogens) will provide a final and robust set.

The exciting feature of mutational signatures is that they are a link between epidemiological observations and our molecular understanding of cancer. Their systematic exploration over time and across different tissues, people, and states of health, will be central to improving models of cancer and other diseases. Hopefully, mutational signatures will allow us to discover preventable exposures and ways to intervene in diseases of somatic evolution. For example, the mutational signature of aristolochic acid has been found in 47% of liver cancers in China, indicating that this exposure is likely to be a cause of significant morbidity (Ng et al. 2017).

The cancer genome has – in the decade since the first cancer genomes were sequenced (Ley et al. 2008, Pleasance et al. 2010a, Pleasance et al. 2010b) – largely been characterised (Campbell et al. 2017, Alexandrov et al. 2018, Li et al. 2017, Sabarinathan et al. 2017, Dentro et al. 2018, Gerstung et al. 2017), and bulk sequencing of additional common tumours is unlikely to change our understanding of it radically. This is not to say that all questions have been resolved – far from it – but perhaps more is to be gained by probing deeply into cancer biology through other means. This involves extending our understanding of the life history of cancer and the forces that have shaped its evolution.

2.c. Cancer as a multi-step branching evolutionary process

Foulds made the case in 1958 that cancer was the end result of a series of qualitatively different changes, based on the concepts of ‘initiation’ and ‘promotion’ that came out of rabbit skin carcinogenesis and the fact that mammary hyperplasia often preceded frank carcinomas in mouse models of breast cancer (Foulds 1958). By the 1970s, the clonal origin of tumours was widely recognised (Nowell 1976) based on three observations: all cells from tumours had the same karyotype (Sandberg and Hossfeld 1970); G6PD heterozygote women only expressed one allele per tumour (the gene is X-linked, and so in a given cell one copy is randomly inactivated) (Fialkow 1974); and plasma cell cancers only produced one immunoglobulin (discussed in Nowell 1976). Cairns related the cellular turnover of normal tissues to their acquisition of mutations, and discussed cancers as a product of natural selection, with sequential clonal expansions each driven

by a mutation (Cairns 1975). Nowell made the same argument, but explicitly added the component of selection, removing the least fit cells (Nowell 1976). Importantly, Nowell stated the capability of 'continued variation so long as the tumour persists', making the point that evolution has no endpoint. In this section, I attempt to cover succinctly the evidence for the multi-step nature and branching evolution of cancer, a brief discussion of mutation rates, and an acknowledgement of the importance of microenvironmental selection pressures. I will focus on the earliest stages of cancer and will not discuss metastasis or the response to therapy.

2.c.i. Multi-step tumorigenesis

Early evidence of cancer's multi-step nature came from the lag between carcinogenic exposures and the development of cancers, which suggested that other events must occur in the meantime, and modelling of age-incidence curves which indicated a number of rate-limiting events (Armitage and Doll 1954, Nordling 1953). Incidence modelling of colorectal cancer is discussed in Results Chapter 2. In the 1980s and 1990s, these steps were tied to a series of histopathological and molecular changes: the adenoma-carcinoma sequence. It was known that most adenomas arose from a single crypt (Ponder and Wilkinson 1986). Histopathology had shown that most carcinomas grew out of adenomas (Sugarbaker 1985), and their shared origin was confirmed by finding the same *RAS* mutation in both the adenoma and its linked cancer. The progression was further strengthened by epidemiological observations: for example, people whose adenomas were not removed were at greater risk of developing cancer. Once the histological progression was established it was possible to investigate its molecular basis. Fearon and Vogelstein's review of this work is the bedrock of our understanding of multi-step carcinogenesis (Fearon and Vogelstein 1990). Early adenomas, intermediate adenomas, late adenomas, and carcinomas were all sequenced, and the number of drivers in each tabulated. More drivers were found in more progressed lesions, with four to five (of the drivers known at the time and assayed) found in most frank carcinomas. Particular drivers tended to be associated with particular stages. For example, *RAS* mutations were found in the same proportion of intermediate adenomas as carcinomas, which suggested that *RAS* mutations' main effect was to reach the intermediate adenoma stage without driving further progression, while loss of 17p (containing *TP53)* was enriched in carcinomas. This

led Fearon and Vogelstein to propose the famous sequence of: 5q (containing *APC*) loss driving the formation of an early adenoma; *KRAS* mutations driving that of an intermediate adenoma; 18q (containing *SMAD4*) loss driving that of a late adenoma; and 17p (*TP53*) loss driving that of a carcinoma. The authors stress that it is the total sum of alterations that matters rather than their precise order, since this ordering of mutations is common but far from universal. Additionally, that certain mutations are found in more progressed lesions does not necessarily imply that they occurred late: *TP53* would also appear mostly in carcinomas if it were, in fact, the first mutation but encouraged extremely rapid progression to cancer such that very little time was spent as an intermediate lesion. This model has been extrapolated to most other cancer types and a similar multi-step progression has been observed. For example, in the progression to oesophageal adenocarcinoma, *TP53* mutations tend not to occur before high grade dysplasia and *SMAD4* mutations are not commonly found before full-blown cancer (Weaver et al. 2014).

Pre-malignancy in blood warrants some discussion, since the clonal dynamics of blood make up half of the present dissertation. In childhood acute lymphoblastic leukaemia, the observation that monozygotic twins can both develop the disease years after birth, but their leukaemias share an initiating clonotypic gene fusion that must have occurred prenatally, indicates that extra events are required to trigger the conversion to malignancy (Greaves et al. 2003, Ford et al. 1993, Greaves 2018). In addition, many children harbour the initiating fusion but never develop leukaemia (Greaves 2018). In adult blood, the myeloproliferative neoplasms typically only bear a small number of known driver mutations per neoplasm (two thirds only have one known driver), and acquisition of certain additional driver mutations is associated with conversion to an acute leukaemia (Nangalia and Green 2018). Intriguingly, the order in which these first few mutations occur influences both the clonal dynamics of the neoplasm and its clinical manifestation (Ortmann et al. 2015). By assaying single cell-derived colonies from patients with myeloproliferative neoplasms for common driver mutations it has been shown that if *JAK2* mutations occur before *TET2* mutations, most sampled haematopoietic stem and progenitor cells had both mutations, whereas if *TET2* occurred before *JAK2*, large numbers of cells with only the *TET2* mutation were found (Ortmann et al. 2015). This indicates that *TET2* mutation promotes a more rapid clonal expansion than *JAK2* mutation does.

More recently, much excitement has been generated by the discovery of known myeloid driver mutations in the blood of adults with normal blood counts and the fact that the bearers of

these clones are at greater risk of cancer. For such mutations to be detectable they need to be in a reasonable proportion of blood cells, indicating a clonal expansion, and so this phenomenon has been termed 'clonal haematopoiesis'. Skewing of the proportion of cells with each X chromosome inactivated in women without cancer had been observed for a long time and ascribed to various causes, but in 2012 it was found that some of these women had *TET2* mutations (Busque et al. 2012). Of seven women followed over five years, one developed essential thrombocythaemia. A series of articles in 2014 based on the targeted sequencing of thousands of normal blood samples demonstrated that clonal haematopoiesis was widespread, with mutations in multiple leukaemia and lymphoma-related genes reported (Xie et al. 2014, Jaiswal et al. 2014, Genovese et al. 2014). The proportion of people in whom clones were detected increased with age. For example, by ultra-deep sequencing of 15 hotspots, clonal haematopoiesis was detected in 1% of patients under 60 and 20% of patients aged over 90 (Mckerrell et al. 2015). Importantly, the presence of mutations in driver genes was associated with a hazard ratio of 11.1 (95% CI 3.9-32.6) of developing a leukaemia, indicating that the mutant clones might be pre-malignant (Jaiswal 2014). The definition of clonal haematopoiesis is somewhat in flux, however. Firstly, the number of clones detected depends on the sensitivity of the detection methodology: when mutations could be accurately called down to a frequency of one in 10,000 cells, clonal haematopoiesis could be detected in 19 out of 20 healthy people aged between 50 and 60 (Young et al. 2016). Secondly, clones have been found in which no known driver mutations were detected (Zink et al. 2017). It is therefore unclear whether clonal haematopoiesis represents a qualitatively distinct stage between normal haematopoiesis and cancer, or if the clones detected are merely a result of normal neutral drift that results occasionally in clones of a detectable size. We currently have a very limited understanding of the range of blood clone sizes in healthy humans. It is to be hoped that this will be resolved by studying further the clonal dynamics of normal human haematopoiesis.

2.c.ii. Branching evolution

Branching evolution implies that complete clonal sweeps through a tumour are infrequent. Experimentally, this means that if different samples are taken from a tumour, each will have a large number of private mutations. This can be depicted as a phylogeny with long private branches

leading from the samples to their coalescence into the trunk of shared, clonal mutations. This is frequently observed in cancers (Yates and Campbell 2012). The length of private branches is determined by a number of factors. Branches are lengthened by restricted movement within the founding clone preventing clonal sweeps (the most extreme example being that of monozygotic twins with concordant clonal acute lymphoblastic leukaemia), by an increased mutation rate, and by decreased competition for resources (such as space) between tumour cells.

In the last few years, much has been made of a "Big Bang" model of colorectal cancer (Sottoriva et al. 2015). The same sub-clonal copy number changes were found in different parts of colorectal carcinomas. In the absence of cell migration (which is perhaps a strong assumption given that the ability to migrate is a hallmark of a carcinoma), this would be consistent with the emergence of multiple subclones and a lack of complete selective sweeps since. Subsequent work showed that the distribution of clone sizes across many tumours, determined based on the allele fraction of mutations, was not significantly different from what one would expect in the absence of selection (Williams et al. 2016). Conversely, it has been shown that a similar distribution of allele fractions can also be observed in the presence of weak subclonal selection (Tarabichi et al. 2017). The authors of the latter study conclude that allele fractions alone are not sufficient to distinguish between neutrality and weak or occasional subclonal selection. Indeed, incomplete clonal sweeps within a tumour need not equate with neutrality: carcinomas have long been observed to grow out of adenomas and this has never been interpreted as neutral evolution. Orthogonal genetic approaches, such as dNdS (a method that compares the proportion of nonsynonymous and synonymous mutations in order to detect selection), have shown the presence of weak subclonal positive selection (Tarabichi et al. 2017). Furthermore, there is no reason that selection should stop operating during tumour evolution. Nonetheless, the similarity between observed allele fractions and simulations under a neutral model indicates that neutral drift is likely to play a large part in determining clone sizes in many tumours.

2.d. Selective pressures

The interactions between the tumour and its microenvironment are multifarious (Greaves and Maley 2012). I cannot hope to do them justice in this short section, but aim merely to

acknowledge their importance. The role of tissue architecture in regulating clonal competition is discussed in more detail below. Here, I provide a few brief vignettes on the effect of selection pressures in the earliest stages of cancer development.

Immune surveillance provides a well-established selection pressure. In aplastic anaemia, CD8 cytotoxic T lymphocytes attack one's own haematopoietic stem cells, reacting against antigens on their surface. In 50% of people with this condition, a cell that acquires a mutation in the X-linked *PIGA* gene that results in the loss of expression of glyosylphosphatidylinositol anchors (such that surface antigens can no longer be tethered to the cell surface) escapes immune attack and proliferates to form a clone. This is paroxysmal nocturnal haemoglobinuria, which is associated with a low risk of transformation to leukaemia (Hoffbrand et al. 2011). Indeed, even in the absence of a well-described pathology such as paroxysmal nocturnal haemoglobinuria, 15% of people with aplastic anaemia will develop myelodysplastic syndrome or acute myeloid leukaemia (Yoshizato et al. 2015), indicating that these microenvironmental changes create the conditions for a mutant cell to outcompete its wild-type neighbours. Even when we do not understand the selection pressure, there is evidence that a change in environment affects clonal expansions. For example, in clonal haematopoiesis, mutations in spliceosome components suddenly increase in frequency in the elderly (Mckerrell et al. 2015). This may be due to a change in selection pressures associated with old age (Mckerrell and Vassiliou, 2015).

Inflammation may provide a selective pressure that favours the outgrowth of certain clones. Dominant negative *P53* mutations in mouse colonic crypts conferred no advantage in a physiological setting but did in a mouse model of colitis (Vermeulen et al. 2013). This might be – at least in part – because of a loss of sensitivity to cellular damage which allows the mutant cells to continue to proliferate while their wild-type counterparts arrest for DNA repair (Breivik 2001). *TP53* mutations are enriched in colitis-associated neoplasia, including in non-dysplastic crypts that neighbour lesions (Leedham et al. 2009), and indeed, ulcerative colitis is associated with an increased risk of colorectal cancer. Similarly, in a model of paediatric acute lymphoblastic leukaemia, B cell progenitors with the *ETV6-RUNX1* fusion (a common initiating lesion) proliferate more slowly than their wild-type counterparts in homeostasis but gain a competitive advantage in the presence of the cytokine TGFβ (Ford et al. 2009).

One experiment produced such a peculiar result that it deserves mention. *Ctnnb1* mutations induced in small populations of cells in mouse skin were enveloped by wild-type skin, and after a

time, expelled from the tissue (Brown et al. 2017). When the proliferation of wild-type cells was inhibited, mutant cells were not expelled and were able to form large cysts. This indicates negative selection that is driven by wild-type cells. In a study of sequencing human skin, no negative selection was detected (Martincorena et al. 2015). If microenvironmental selection pressures arrested the growth of mutant cells, rather than eliminating them, however, no negative selection would be observed. The authors of the latter study did note that the range of mutant clone sizes detected in human skin was unexpectedly small, indicating that there may be some constraint on clonal expansion.

Clearly, there is a world of complexities of selection pressures in normal tissues that remains to be explored. Some have gone so far as to claim that the age incidence curves of cancer are due to changes in the microenvironment with age rather than the need to acquire multiple driver mutations (DeGregori 2017).

2.e. Mutation rates in cancer

There has been much debate over whether the mutation rates in normal tissues are sufficient to accumulate the number of driver mutations needed for cancer. A higher mutation rate allows for more rapid evolution, but it may be unnecessary and could even cause deleterious mutations. Loeb proposed the 'mutator phenotype' in 1974, based on the observation that polymerases isolated from acute lymphoblastic leukaemia cells replicated DNA less faithfully than those from normal lymphocytes. Various modelling approaches were used to determine whether an increase in mutation rate was necessary, some finding that it was (Loeb 1991) and others that it was not (Tomlinson et al. 1996). These models seem to be relatively sensitive to the number of hits that are necessary before the first clonal expansion and the extent of clonal expansions. Estimates of mutation rates were based on single gene reporter assays and so not representative of the whole genome.

The data, too, are mixed. By the 1970s, a number of studies had indicated that tumours made more mitotic errors than normal cells, and that the numbers increased as the tumour progressed (discussed in Nowell 1976). The discovery of Lynch syndrome and other genetic predisposition syndromes that result in an increased mutation rate certainly provides strong

evidence that in these circumstances an increased mutation rate is beneficial, but it is unclear whether this also applies to sporadic tumours where more hits would be necessary in order to increase the mutation rate. In colorectal cancers, it was observed that tumours that were mismatch repair deficient rarely exhibited chromosomal instability, and conversely tumours that were chromosomally unstable rarely lost mismatch repair, suggesting that there might be an advantage to increasing the mutation rate by whichever means (Fearon 2011).

With the advent of cancer genome sequencing, it has become clear that most tumours have tens of thousands of mutations due to a number of mutational signatures (Alexandrov et al. 2013, Alexandrov et al. 2018). Without sequencing normal tissues, however, one cannot know if this is abnormal. In one of the few tissues in which the comparison of tumour and normal has been performed, it was found that the mutation burden of acute myeloid leukaemia cells was not increased relative to healthy haematopoietic stem and progenitor cells (Welch et al. 2012). Nonetheless, blood cancers may represent a special case, since they are on the whole remarkably un-mutated and require few driver mutations. More recently, sequencing of organoids derived from normal colonic stem cells (discussed in Results Chapter 2 section I.4.) has shown that normal colonic cells have a lower mutation burden than colorectal cancers (Blokzijl et al. 2016, Alexandrov et al. 2018). Finally, a recent analysis shows that – with the exception of a few tens of genes in haploid regions which are very rarely mutated – negative selection is virtually absent in cancers, indicating that the mutation rates observed in cancers are not deleterious through the inactivation of essential genes (Martincorena et al. 2017).

Taken together, it seems that there is more evidence for an increased mutation rate in cancer than against it, in some tissues at least. While elevated mutation rates may not be necessary in order to acquire the sufficient number of driver mutations, an increased mutation rate may still increase the probability of this occurring. Definitive answers, however, can only come through the comparison of normal and cancerous cells.

In summary, cancer is a multi-step branching evolutionary process, with ongoing clonal competition and complex selection pressures. Despite these great advances in our conception of cancer, many questions remain unanswered. Two lacunae that are of relevance to this dissertation stand out. First, most descriptions of tumour progression are not quantitative (with a few exceptions, e.g. Bozic et al. (2010), Mitchell et al. (2018)). To understand cancer evolution fully,

one would need to measure the selective benefit conferred by each change in the context of its particular microenvironment. Second, we have a limited understanding of somatic evolution before the advent of histological changes.

3. The mutation rates and stem cell numbers of normal tissues affect cancer risk

Under a simple model of cancer, where a cell needs to acquire a defined number of driver mutations to become cancerous and driver mutations are independent events that occur at a constant rate over time, the incidence rates of cancer are closely related to the number of driver mutations required. If a single driver were required, the incidence rate of cancer at all ages would be the same, since the probability of acquiring the driver is the same in every decade. If two driver mutations were required, the incidence rate would increase linearly, since the probability of a cell having already acquired one driver increases linearly with time and the probability of acquiring a second driver is constant. If three drivers were required, the cancer incidence rate would increase proportionally to the square of age, and so on. For seven drivers:

$$\text{Incidence rate} = k * p1 * p2 * p3 * p4 * p5 * p6 * p7 * t^6$$

where k is a constant and t is age. p1 to p7 are the rate of acquisition of each of the driver mutations, which in this model will be entirely dependent on the mutation rate *in normal tissues*. Plotting the logarithm of cancer incidence *versus* age, the slope of the line should be one fewer than the number of driver mutations required to cause a cancer. This is the reasoning famously applied by Nordling (1953) and Armitage and Doll (1954) to infer the number of rate-limiting events necessary to cause cancer (commonly interpreted as driver mutations) from age incidence data (Nordling used mortality as a proxy for incidence). Driver mutations are unlikely to be independent, since they may cause clonal expansions and/or increase the mutation rate. Nonetheless, at the very least, cancer incidence will be related to the probability of acquiring the first driver mutation (p1), which in turn is likely to depend approximately linearly on both the mutation rate in normal tissues and the number of cells at risk of transformation.

A higher mutation rate in normal tissues, therefore, should increase the risk of cancer. In support of this, many risk factors for cancer increase the mutation rate, whether genetic (biallelic mismatch repair deficiency, xeroderma pigmentosum, Fanconi anaemia, or polymerase proof-reading polyposis syndrome, to name but a few examples) or environmental (such as smoking, ionising radiation, or aristolochic acid). Although many of these might act in other ways as well (such as by changing the clonal dynamics of the tissue through inducing inflammation and proliferation) in a study of cancers whose relative risk is elevated by smoking, the mutation burden of cancers from smokers was higher relative to that of cancers from non-smokers, which is consistent with smoking exerting its effect at least in part through elevated mutation rates (Alexandov et al 2016).

3.a. Stem cells are the cell-of-origin of many adult cancers

The more cells at risk of transformation, the greater the risk of cancer. Determining which are the cells at risk is therefore of primary importance. Given that the capability for self-renewal is shared by both cancers and adult tissue stem cells, the latter have been proposed as the cell-of-origin for a number of malignancies. Here, I will discuss the evidence for this in colon and blood.

Barker and colleagues inactivated *Apc* in cells that express *Lgr5* in colonic crypts (Barker et al. 2009). Lgr5 is a marker expressed at the base of the crypt, and a subset of Lgr5-expressing cells behave as functional stem cells (to be discussed in more detail in Results Chapter 2) (Barker et al. 2007, Kozar et al. 2013). Loss of Apc in the stem cell compartment resulted in the rapid formation of large adenomas, whereas loss of Apc in the more differentiated transit-amplifying cells produced microadenomas that were only very rarely observed to progress into macroscopic lesions, and these rare cases were explained by occasional unintended Apc inactivation in stem cells (Barker et al. 2009). Nonetheless, in inflammation or under the influence of certain driver mutations, more differentiated cells within the crypt have been observed to re-express stem cell markers and reacquire the ability to form adenomas (Schwitalla et al. 2013). Similarly, quiescent Paneth cell-precursors have been shown to produce multilineage output following injury (Buczacki et al. 2013), and if Lgr5+ stem cells are deleted by inducing the expression of diphtheria toxin only in those cells, more differentiated cells seem to be able to take their place (Tetteh et al. 2016). It is

possible, therefore, that in some circumstances initiating lesions in more differentiated cells could result in cancer.

The evidence that haematopoietic stem cells are the cell-of-origin of certain adult blood cancers is less direct. The leukaemic blasts from G6PD heterozygote women with chronic myeloid leukaemia only expressed one isoenzyme, and B lymphocyte populations also only expressed the same isoenzyme (Fialkow et al. 1978, Martin et al. 1980), indicating that the cell-of-origin of the leukaemia had multilineage potential. It is, of course, also possible that the cell-of-origin could have restricted lineage potential and the driver mutations themselves alter the cell's lineage output. In a model of the disease, self-renewal properties were not conferred by transduction of the pathognomonic *Bcr-Abl* fusion into progenitor cells, but they were by certain other gene fusions, indicating that differentiating cells could, in some cases, also act as the cell-of-origin (Huntly et al. 2004). In acute myeloid leukaemia, it was observed that the cells capable of causing leukaemia when transplanted into immunocompromised mice bore the same surface markers as haematopoietic stem cells (Bonnet and Dick 1997). The authors argued that for a stem cell to transform was more parsimonious than for a more differentiated cell to transform and reacquire stem-like properties and markers. Childhood acute lymphoblastic leukaemias, in contrast, may mostly arise from more differentiated cells, as myeloid cells are not usually found to share markers with leukaemic cells (Greaves 1993).

3.b. Stem cell divisions and cancer risk

It was recently and controversially asserted that two thirds of the variation in cancer risk is a result of the "bad luck" of the random replication errors of normal cells (Tomasetti and Vogelstein 2015). For 31 cancer types, the authors correlated the lifetime cancer risk in the US against estimates of the number of stem cell divisions in the tissue of origin. A correlation of 0.8 meant that 65% (95% CI 39% to 81%) of cancer risk was associated with stem cell divisions. Nothing proves that the risk associated with stem cell divisions is due to somatic mutations, but this is the most likely mechanism. Such a strong correlation is perhaps surprising given all the different factors that affect cancer development and vary across tissues: carcinogenic exposures,

microenvironmental and stem cell architecture differences (discussed below), the number of driver mutations needed by different cancers, and the proliferative advantage provided by each driver.

A number of criticisms can be levelled at this analysis. First, there is inaccuracy in the estimates of the number of stem cell divisions. The estimate of haematopoietic stem cell numbers is likely to be a significant overestimate (as will be discussed in Results Chapter 1), and other errors have been noted (Rozhok et al. 2015). Second, little attempt is made to explain some of the largest departures from this correlation. Most strikingly, the number of stem cell divisions in small intestine and colon is similar, and yet the incidence of colorectal cancer is two orders of magnitude higher. Third, it has been pointed out that this correlation cannot distinguish between extrinsic and intrinsic effects both because the effect of extrinsic mutational processes can be linked to the stem cell division rate if rapidly cycling cells are more likely to fix mutations, and because some extrinsic mutational processes might act upon multiple cell types to similar extents, thus preserving the correlation (Wu et al. 2016). An analysis of mutational signatures across normal tissues and exposures would clarify this. Even if valid, the conclusions of the Tomasetti and Vogelstein study were presented in an unhelpful way in terms of public health. The relative risk of cancer across *tissues* is quite different to that across *people* and the implications of this were not made clear enough in the mass media (e.g. Galagher 2015). Whatever my behaviours I will always be more likely to develop cancer of the colon than an osteosarcoma, but I can significantly reduce my risk of cancer relative to my identical twin who drinks like a fish and smokes like a chimney. The "bad luck" applies to my colon relative to my femur rather than to me relative to my debauched twin. Despite these criticisms, this study identifies those tissues that have more cancer than one would anticipate from the number of stem cell divisions. Similar approaches might represent an interesting way to explore cancer aetiology across different tissues and potentially identify preventable exposures, although perhaps better estimates of stem cell numbers and division rates are needed first.

A related point was raised by Tomasetti and Vogelstein's follow-up article (Tomasetti and Vogelstein 2017), in which a distinction is drawn between the amount of cancer that is due to an exposure and the proportion of driver mutations in a tumour that are caused by it. They calculate that even for an exposure such as smoking, which is responsible for 90% of lung adenocarcinomas, still 35% of total driver mutations are likely to be due to random errors of DNA replication. This is perhaps unsurprising from a perspective of species evolution. Natural selection has endowed us

with tumour suppressive mechanisms such that cancer is unlikely during a reproductive lifespan with the number of drivers that are likely to occur by replication errors alone across a tissue. Mutagens need only tip us over the threshold of the number of driver mutations that we can tolerate for them to have a pronounced carcinogenic effect.

3.c. Tissue architecture limits the selective advantage of mutant cells

Cairns, after noting the potential of natural selection to promote malignancy, wrote:

> We may therefore expect to find fast-multiplying tissues arranged in such a way that neighbouring stem cells (or sets of stem cells) are restricted to limited territories so that they cannot easily compete with each other.
>
> (Cairns 1975).

Going on to discuss intestinal crypts, he noted that both restricting the number of stem cells and restraining the competition between them would decrease the accumulation of stem cells with driver mutations (Cairns 1975). Here, I focus on the colon; the same mechanisms of restricting stem cell competition do not seem to be operative to the same extent in normal blood since, here, stem cells are able to recirculate (although many aspects of the haematopoietic stem cell niche are incompletely understood).

The organisation of the colonic epithelium into crypts means that even if a stem cell with a driver mutation manages to sweep through an entire crypt, its clone size is still limited. Clonal expansion cannot occur through proliferation alone: crypt fission is needed. Limiting clone sizes at this early stage would presumably significantly reduce the probability of another driver mutation occurring in the same clone. The role of driver mutations in promoting crypt fission is discussed in Results Chapter 2.

Less intuitively, the small number of competing stem cells in the intestinal stem cell niche strengthens the hand of chance in stem cell competition (Calabrese and Shibata 2010, Rozhok and DeGregori 2015). In population genetics terms, population structure decreases the effective population size, and the probability of fixation of a beneficial allele is related to the effective

population size divided by the census population size (Whitlock 2003). Stochastic factors are stronger in smaller niches, and so most driver mutations, even with a strong selective advantage, are likely to be lost from the crypt by chance. The other side of this coin, of course, is that there will be less negative selection as well, and colonic crypts are more likely to accumulate deleterious mutations than a system like blood. One can speculate that the need to balance positive and negative selection in tissues contributed to the evolution of different stem cell niche structures (in addition to obvious physiological reasons). Thus, the number of stem cells and their mutation rates are only one part of the story.

4. Somatic evolution in ageing and diseases beyond cancer.

Little is known about the role of somatic mutations beyond cancer, and so this section will necessarily be brief. The somatic mutation theory of ageing (Szilard 1959, Morley 1995) posits that the stochastic and progressive accumulation of mutations over life results in a loss of function of tissues that is responsible for ageing. It was based on observations that large scale irradiation of mammals seems to result in premature ageing (Henshaw et al. 1947). Several features of Szilard's model are now known to be wrong, such as the assumption that one mutation might inactivate a whole chromosome, but in the absence of a better explanation, the somatic mutation theory of ageing – with some alterations to Szilard's model – remains attractive.

Progeria syndromes, such as Cockayne syndrome, are one source of evidence that somatic mutations play a role in ageing (Garinis et al. 2008). In these, features of the ageing process are accelerated, and the underlying genetic defect is frequently in a component of a DNA damage repair pathway (for example, transcription-coupled nucleotide excision repair is impaired in Cockayne syndrome). Some syndromes, such as xeroderma pigmentosum, are associated with both progeria and an increased risk of cancer. Furthermore, long-term survivors of chemotherapy and radiotherapy show signs of premature ageing (Garinis et al. 2008). A recent study of single cell sequencing of individual neurones (the method is discussed below) from individuals with Cockayne syndrome, xeroderma pigmentosum, and controls, showed that the mutation rate was indeed increased in somatic tissues from individuals with both diseases (Lodato et al. 2018).

Nonetheless, it is questionable whether the mutation rate in normal tissues would be sufficient to bring about a functional defect. For example, only ~10 exonic single nucleotide variants are observed in a blood cell from a person in their 70s (Welch et al. 2012). Assuming a constant mutation rate, by age 100 a person is unlikely to have more than 20 protein-coding mutations per cell, which, scattered among the thousands of genes in the genome, most of which are diploid and non-essential, seems unlikely to have much functional effect in most cells. The mutation rates in other tissues are likely to be higher, but probably not high enough to cause a functional decline.

If the loss of function were associated with a selective advantage, however, an unlikely event affecting a small number of cells by chance could, after a period of clonal expansion, affect the whole tissue. Evidence of this has come from the recent discovery of clonal haematopoiesis, which increases the risk of all-cause mortality even after excluding malignancy (Jaiswal et al. 2014). Much of this risk seems to come from increased cardiovascular events, seemingly through the aberrant expression of atherogenic cytokines by mutant cells (Jaiswal et al. 2017). Other diseases may be clonal in origin too. There is evidence, for example, that somatic mutations may play a role in autoimmune diseases such as Sjögren's syndrome (Nocturne et al. 2013). While these are examples of specific diseases, the same may be true of ageing.

Thus, although the evidence is not yet fully convincing, the exploration of somatic mutations in normal tissues may reveal a somatic mutation basis for ageing and a number of diseases. Establishing a baseline of somatic mutation burden and processes for normal tissues from healthy people will provide a reference against which different disease states may be compared.

5. Massively parallel sequencing

All of the work in this thesis relies upon massively parallel sequencing in order to detect somatic mutations, and so a brief overview of the method is in order. Only Illumina/Solexa sequencing will be discussed, since it is the dominant technology and the only one used here.

After extraction of DNA from cells, the DNA is sheared (typically by sonication, but in some experiments by enzymatic fragmentation (see Methods)) into fragments that are less than 1,000 bases long. The ends of the fragments are repaired to remove overhangs, and adaptors are

ligated. The molecules are washed over a flow cell, where they anneal by a short sequence of the adaptor tail to oligonucleotides covalently bound to the surface of a glass slide. Each molecule of DNA is replicated by bridge amplification, resulting in the formation of DNA colonies that can be sequenced.

Illumina/Solexa sequencing uses the method of sequencing by synthesis. Reversible chain-terminating nucleotides, each coupled to a fluorophore, are washed over the flow cell. A free-floating nucleotide binds to extend a strand of DNA along the template if it is complementary to the next base of the template to be copied. This terminates the reaction. On excitation by a light source, the fluorophore is cleaved, emitting a light signal that is detected by a camera. The colour of the fluorophore gives the identity of the base that has just been added. Cleavage of the fluorophore also unblocks the end of the molecule such that the next base can be added. The process continues until a read of the desired length has been achieved (150 base pairs for the XTEN platform which was mostly used in this study). Both ends of the DNA molecule are sequenced, resulting in paired-end reads. The informatic representation of the DNA sequence is then aligned to the human reference genome and mutations can be called (Methods).

6. Methods for studying mutations in normal tissues

The polyclonal nature of normal tissues makes them difficult to study. Standard sequencing protocols require hundreds of nanograms of DNA and so tens of thousands of cells. Mutations can only be called accurately if they are present in a substantial proportion of these cells. Unlike a cancer in which large numbers of mutations are shared by most cells in a biopsy sample, a biopsy of tens of thousands of cells from most normal tissues will reveal very few shared (and so detectable) somatic mutations. A number of methods have been used in recent years to tackle this problem: single cell sequencing; heavily error-corrected deep sequencing; *in vitro* clone formation; and microbiopsy sequencing. The latter two are used in this thesis.

In single cell sequencing, the DNA from single nuclei is amplified using a method such as multiple displacement amplification (MDA) and sequenced. Clean single cell whole genome sequencing would be the perfect solution to the study of normal tissues, and all the methods discussed below would largely be obsolete. Unfortunately, the process of whole genome

amplification from such a small amount of input DNA both introduces large numbers of false positive errors and amplifies the genome unevenly, resulting in large portions of the genome that are not covered (Wang and Navin 2015). Despite these caveats, it has been used to determine the mutation rates in single neurones (Lodato et al. 2018). In order to reduce false positives, mutations were only called if they were correctly phased with a germline polymorphism; the mutation burden was then extrapolated to the whole genome. A correlation with age was observed, indicating that real mutations were being called. This method, cannot, however, be used to catalogue every mutation in the genome of a single cell, as mutations in a high proportion of the genome cannot be called.

Heavily error-corrected deep sequencing provides a way to study a polyclonal population. The most successful methods have been based on the principles of 'duplex sequencing' (Schmitt et al 2012, Kennedy et al 2014), where the sequences of either strand of a DNA duplex are identified separately and can be compared: artefacts introduced in library preparation or sequencing errors are not likely to affect both strands of a duplex in the same way. One difficulty of this sort of approach is that recapturing both strands of a DNA duplex is unlikely. The probability of this was improved by a strong dilution step before the polymerase chain reaction (Hoang et al. 2017), and the method could then be used to call mutations in polyclonal tissues. Identifying and reproducing the dilution sweet-spot is technically challenging, and, even if achieved, the full complement of mutations in a genome cannot be identified, since the information of which cell each DNA molecule comes from is lost. This makes analyses of clonal dynamics (such as phylogenetic reconstruction) more challenging.

In vitro clone formation involves the isolation of single cells and their expansion in culture, in the presence of growth factors, into colonies that are sufficiently large to be whole genome sequenced. The cell's own replicative machinery is less error prone than whole genome amplification methods, and the genome is naturally amplified evenly. Furthermore, the precise cell type of interest can be isolated prior to amplification (for example, by flow cytometry). This method has been used successfully for intestinal and liver cells (Blokzijl et al. 2016) and for blood stem and progenitor cells (Ortmann et al. 2015) amongst others. While the range of tissues for which this is possible is being broadened, not all cells can be expanded in this way and it seems unlikely that this approach will be possible for post-mitotic tissues without extensive manipulation. There are two other caveats of this approach: mutations can occur *in vitro* (although this can be

estimated and largely corrected: see Methods), and there is potential for positive or negative selection of mutations during growth (van de Wetering et al. 2015).

Microbiopsies exploit *in vivo* clonal expansions to identify mutations present in small numbers of adjacent cells. Microbiopsies may be blind, in which a random biopsy of a tissue is taken and one hopes to encounter a clone, or they may be targeted at histologically defined clonal units, such as a colonic crypt (discussed in greater detail in Results Chapter 2). Blind microbiopsies have been used successfully to identify mutations in human skin (Martincorena et al. 2015). Using the 'pigeonhole principle', which states that clones whose sum is greater than half of the biopsy size must be nested, it may, in some cases, be possible to infer clonal relationships, and the size of clonal patches can inform on selection. The technique only works for tissues in which clonal patches are of a detectable size, however, which may limit its use to squamous epithelia or tissues with a defined and large clonal unit. The mutations called are those that were present in the most recent common ancestor of the clone rather than those in extant cells.

7. Blood is well suited to studying the clonal dynamics, and colon the mutational processes, of normal tissues

Blood and colon are both well-studied tissues, with properties that lend themselves to the investigation of different aspects of somatic evolution. Blood can be sampled randomly and longitudinally with relatively little discomfort to the patient, making the investigation of clonal dynamics tractable. Although haematopoietic stem cells are perhaps the best-studied adult stem cell, we lack the answers to basic questions in humans, such as the number of stem cells and range of clone sizes in healthy humans. These are important parameters to model leukaemia and other diseases, such as those of bone marrow failure. Furthermore, the recent discovery of widespread clonal haematopoiesis in the elderly associated with an increased risk of malignancy and death from other causes adds urgency to the need to establish a baseline for the normal clonal dynamics of blood in humans.

The colon is well-suited to the investigation of mutational processes and driver events. Colonic crypts form visible clonal units, permitting a microdissection approach to identifying somatic mutations that have occurred in colonic crypts, without the caveats of selection for or

against certain mutations or the acquisition of *in vitro* mutations. Colonic mucosa is frequently sampled, allowing us to investigate the range of somatic mutational processes across a range of people. The archetypal model of the progression from normality to cancer was defined in the colon (Fearon and Vogelstein 1990), and it has historically been the battleground for debates on the mutator phenotype (Loeb 1991, Tomlinson et al. 1996). The investigation of normal colon complements these studies. Finally, colorectal cancer is a common and lethal disease, with 42,000 new cases diagnosed in the United Kingdom in 2015, and 16,000 deaths from it in 2016 (Cancer Research UK, accessed August 2018). Epidemiological observations (discussed in Results Chapter 2 section I.3.) indicate that some risk is preventable. The elucidation of the mutational processes that cause normal cells to become cancerous can hopefully contribute to the identification of the causes of excess risk and indicate opportunities for early intervention, as well as contribute to our understanding of the pathogenesis of the disease.

8. Thesis aims

This dissertation is divided into two chapters. The first investigates the clonal dynamics of normal blood, and the second the mutational landscape of normal colonic epithelium.

The principle aim of the study of normal haematopoiesis in Results Chapter One is to estimate the number of haematopoietic stem cells that actively contribute to blood in one healthy human, and the rate at which haematopoietic stem cells replace one another. Second, it aims to investigate clonal relationships between the myeloid and lymphoid lineages. Third, it seeks to establish the mutation burden per genome and mutational processes of normal haematopoietic stem and progenitor cells.

The main aim of the study of the colon in Results Chapter Two is to describe the repertoire of mutational processes and their mutational burden operative across different sites on the colon of a number of different people. Second, it seeks to establish the frequency of driver mutations in normal colon. Third, it investigates how the mutation rate and mutational processes change in the progression from normal to cancer.

Background specific to each tissue is presented at the beginning of the relevant results chapter.

METHODS

1. Samples

1.a. Primary bone marrow and peripheral blood samples

A bone marrow (BM) aspirate, peripheral blood (PB) sample, and a buccal swab were obtained in collaboration with the laboratories of David Kent and Tony Green from a 59 year-old man with normal blood counts and no history of blood disorders. Follow-up PB samples were taken every two months for the following six months, and thereafter once every six months by me, Carlos Gonzalez-Arias or Jacob Grinfeld.

1.b. Colon samples for laser capture microdissection

We obtained healthy colonic biopsies from four cohorts of patients. The first represents seven organ donors ranging in age from 36 to 67, from whom colonic biopsies were taken at the time of transplantation. These samples were provided by Kourosh Saeb-Parsy. The second represents patients aged 60 to 72 who were having a colonoscopy following a positive faecal occult blood test as part of the Bowel Cancer Screening Programme; we selected 16 patients who were not found to have either an adenoma or a carcinoma on colonoscopy, and 15 patients who were found to have a colorectal carcinoma (the normal biopsies that we use were distant from these lesions). These samples were provided by Nick Coleman. The third represents three paediatric patients who were having a colonoscopy for inflammatory bowel disease following suggestive symptoms, but in whom no diagnosis of inflammatory bowel disease was made. These samples were provided by Matthias Zilbauer. Finally, access to samples from one 78 year-old gentleman with oesophageal cancer who underwent a warm autopsy was provided by Rebecca Fitzgerald. A table of which samples come from which patients is provided in Appendix C.

1.c. Colon samples for organoid derivation

Tissue material was obtained in collaboration with the Clevers lab from The Diakonessen Hospital, Utrecht. From the resected colon segment, both normal and tumour tissues were isolated. The isolated tumour tissue was subdivided into 4-5 segments. Normal tissue was taken at least 5 cm away from the tumour.

2. Sample preparation

2.a. Isolation of haematopoietic stem and progenitor cells and peripheral blood fractions by the Kent lab

The following work was performed by the Kent lab. Mononuclear cells (MNCs) from BM and PB were isolated by density gradient centrifugation (Lymphoprep; Axis-Shield, Oslo, Norway), and enriched for CD34 positive cells (EasySep Human CD34+ enrichment kit, STEMCELL Technologies, Vancouver, Canada (STEMCELL)) as per the manufacturer's guidelines except that only one round of depletion in the magnet was performed. Cells were then stained with the following antibodies; CD38-FITC (Clone HIT2, BD Biosciences, San Jose, CA, USA (BD)), CD34-PerCPCy5.5 (Clone 581, Biolegend, San Diego, USA (Biolegend)), CD10-APC-Cy7 (Clone HI10a, Biolegend), CD90-APC (Clone 5E 10, Biolegend or BD), CD45RA-Violet450 (Clone HI100, BD) and CD135-PE (Clone BV10A4H2, Biolegend). Single $CD34^{+}CD38^{-}CD90^{+}CD45RA^{-}$ cells (hereafter "HSCs") from BM and PB were isolated for liquid culture using an Influx sorter (BD), equipped with the following lasers; 405nm, 488nm, 561nm, and 640nm, and filter sets; 530/40 (for FITC), 710/50 (for PerCPcy5.5), 750LP (for APC-Cy7), 670/30 (for APC), 460/50 (for Violet450), and 585/29 (for PE). Bulk HSCs and $CD34^{+}CD38^{+}CD90^{-}CD135^{+}CD45RA^{-}$ (common myeloid progenitors, CMPs), $CD34^{+}CD38^{+}CD90^{-}FLK2^{+}CD45RA^{+}$ (granulocyte-macrophage progenitors, GMPs) and $CD34^{+}CD38^{+}CD90^{-}FLK2^{-}CD45RA^{-}$ (megakaryocyte-erythroid progenitors, MEPs) from BM and PB were isolated for colony forming cell (CFC) assays using the same setup.

For liquid culture, single HSCs were sorted into 96-well plates containing StemSpan (STEMCELL), cc100 cytokine cocktail (STEMCELL), Penicillin/Streptomycin (Sigma-Aldrich, St. Louis, MO, USA (Sigma-Aldrich)), L-Glutamine (Sigma-Aldrich), 2-Mercaptoethanol (Life Technologies, Carlsbad, CA, USA (Life Technologies)). Ten days post-sort the medium was supplemented with additional recombinant growth factors (GM-CSF (Miltenyi Biotec, Bergisch Gladbach, Germany (Miltenyi)), G-CSF (Miltenyi), M-CSF (Miltenyi), Epo (Janssen Pharmaceutica, Beerse, Belgium), IL-3(Miltenyi), TPO (Miltenyi), and IL-6 (Miltenyi)). After 4-6 weeks in culture, single-cell derived clones were harvested into PBS for subsequent DNA extraction.

For colony assays, bulk HSCs, CMPs, GMPs and MEPs were sorted into 1.5ml Eppendorf tubes containing StemSpan before distribution and plating in MethoCult H4435 (STEMCELL). After 2-4 weeks in culture individual colonies were picked into PBS for subsequent DNA extraction.

For longitudinal time points, PB samples were collected (40-60 ml, Lithium-Heparin tubes) and MNCs and granulocytes were isolated using Lymphoprep. The granulocyte fraction was subject to two rounds of red blood cell lysis (Ammonium chloride, STEMCELL) and defined numbers of cells (1,000, 10,000 and 100,000 cells/well) were plated into 96-well plates for later DNA extraction (stored in PBS). Bulk $CD4^+$/$CD8^+$ T-cells and $CD19^+$ B-cells were sorted and plated in 96-well plated for later DNA extraction (1,000 cells/well).

DNA was extracted by the Cancer Genome Project Lab core facility from cell pellets using DNeasy kit (Qiagen 69504) for whole genome sequencing of colonies, and using the BioRobot according to the protocol for dried blood samples (Qiagen 965942) for targeted sequencing of bulk granulocytes and lymphocytes.

2.b. Laser capture microdissection of colonic crypts

Fresh frozen biopsies were embedded in optimal cutting temperature (OCT) compound. 30 micrometre sections were fixed in methanol for five minutes, washed three times with phosphate-buffered saline, and then stained with Gill's haematoxylin for 20 seconds. Crypts were isolated by laser capture microdissection, with every crypt falling into a separate well. They were lysed using

the Arcturus PicoPure Kit (Applied Biosystems) according to the manufacturer's instructions. DNA library prep then proceeded without clean-up or quantification.

In an attempt to avoid batch effects, most plates of crypts included samples from multiple patients. For example, rather than dissecting all the crypts from one patient into one plate, in most cases the samples from a given patient were split into three groups, and crypts were cut into separate plates, which they shared with samples from other patients. Almost all patients had some samples dissected on a different day and into a different well to other samples from the same patient. Batch effects that affect coverage should be accounted for by statistical analysis that explicitly takes coverage into account. Batch contamination with DNA of another species should be detectable at quality control, as a low proportion of reads mapping to the genome; in other work, for example, Mycoplasma contamination of sequenced cell lines was detected. Batch effects cannot be solely responsible for the presence of any novel signature, as all novel signatures have been detected in crypts from multiple plates and sequencing runs. Nonetheless, we cannot exclude batch effects that could affect crypts from the same section of tissue, as normally these are processed together.

2.c. Isolation and growth of colonic organoids by the Clevers lab

The following work was performed by the Clevers lab. Human normal and tumour colon organoids were established and maintained from isolated colonic epithelium. In brief, long term normal colonic organoid culture required Human Intestinal Stem Cell Medium (HISM) composed of Advanced DMEM/F12 (AdMEM) with penicillin/ streptomycin, 10 mM HEPES, 1xGlutaMAX,1xB27 (Invitrogen) and 1 μM N-acetylcysteine (SIGMA), supplemented with 50 ng mL^{-1} human recombinant EGF (Peprotech), 0.5 μM A83-01 (Tocris), 3 μM SB202190 (SIGMA), 1 μM Nicotinamide (SIGMA), 10 nM Prostaglandin E2,Wnt3A-condition medium (CM) (50% final concentration), Noggin-CM (10% final concentration), and R-Spondin1-CM (10% final concentration). Tumour organoids were cultured in medium containing only EGF, Noggin-CM, R-Spondin1-CM and A83-01.

For clonal organoids from normal crypts, single crypts were embedded in 10 μl Matrigel and cultured in HISM medium. For clonal tumour organoid cultures, tumour cell suspensions were

cultured for 7-14 days in HISM without Wnt3A-CM. 10-15 individual organoids were picked and separately dissociated into single cells by TryPLE express (Thermo Fisher), washed and suspended in AdMEM containing Propidium iodide (PI). 48 single cells were sorted into tumour organoid medium (HISM plus 10 μM ROCK inhibitor Y-27632 (Tocris BioScience); no Wnt-CM) from each tumour organoid. Sorting was based on FCS Area/FCS peak and PIneg/FCS Area using a Moflo machine (Beckman Coulter). Sorted cells were spun down at 1000 g and 4°C for 5 minutes, after which single cells were each embedded into 10 μl of Basement Membrane Extract (BME, Amsbio) and seeded into 96 well-plates at a ratio of 1 cell/well. The gel was left to solidify in a 37°C incubator after HISM (no Wnt3A-CM) was added. Y-27632 was added to the medium for the first a week after sorting. For each original tumor organoid, a single clonal organoid was selected and expanded for further study.

DNA was extracted from frozen tissue samples or organoid cultures using AllPrep DNA/RNA minikit (Qiagen 80204).

3. Library preparation and sequencing

3.a. Library preparation

Libraries for the blood study and for the colonic organoids were prepared by the Sanger Institute Core Pipelines. Library preparation of low-input microdissected crypts was performed by Peter Ellis and Chris Alder. Two library preparation methods were used for low-input material: initially, sonication was used to fragment DNA, and later, an enzymatic fragmentation method was implemented as it could make libraries from even lower input. Comparison of the two methods showed no difference in mutation calls once post-processing filters (described below) had been implemented. For sonication libraries, 20ul of LCM lysate was sheared with focused acoustics and DNA isolated with magnetic beads. End preparation and A tailing was performed and duplexed adaptors were ligated. After another bead-cleanup and 12 PCR cycles, a final bead-cleanup was performed and sequenced on the HiSeq X platform (Illumina). For enzymatic libraries a magnetic bead separation was performed up-front. DNA was not washed off beads, but a bead/DNA slurry

was incubated with Ultra II FS buffer and Ultra II FS enzyme (New England BioLabs). DNA that was fragmented and A-tailed was ligated with duplex enhancers and processed as for sonication.

3.b. Sequencing

For whole genome sequencing, paired end sequencing reads (150bp) were generated using Illumina XTEN® machines. For targeted sequencing, sequencing reads (75bp) were generated using Illumina HiSeq2000® machines. Sequences were aligned to the human reference genome (NCBI build37) using BWA-MEM for 150bp reads and BWA-align for 75bp reads by the Sanger Institute Core Informatics team.

4. Bait-set design for targeted sequencing of peripheral blood

Very deep sequencing of every single substitution detected in the clones would be cost-prohibitive. We therefore selected mutations to be included in the bait-set according to the following criteria:

- All mutations called by the Shearwater algorithm (see below) in two or more samples
- Mutations called in at least one clonal sample and also in at least one polyclonal sample
- We aimed to sample approximately 20 mutations private to each colony (for both clonal and polyclonal). These were selected in the following way:
    - Mutations were ranked according to their sequencing error rate from whole genome sequencing of a large panel (approximately 1000 tumours and normal samples). The Shearwater algorithm was used to calculate the error rate.
    - We then picked the 10 mutations on the X chromosome with the lowest error rate, since mutations on the X chromosome should provide greater sensitivity as our patient is male.
    - We then picked the 10 mutations from elsewhere in the genome with the lowest error rate.
    - When selecting mutations private to polyclonal colonies, we prioritised mutations with the highest VAF, that were likely to be in the dominant clone.

- We also included 100 mutations that are known hotspot mutations found in clonal haematopoiesis of indeterminate potential (CHIP) and leukaemias in order to detect CHIP should it arise.

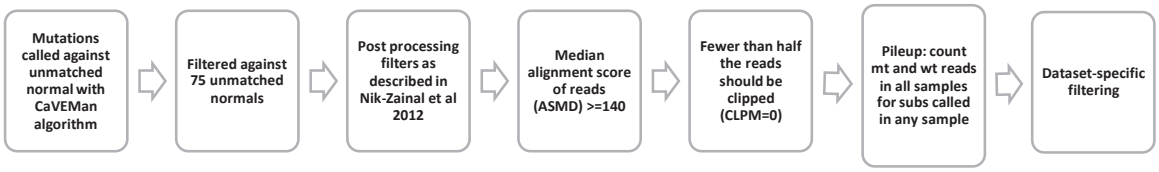
We performed custom RNA bait design following manufacturer's guidelines (SureSelect®, Agilent®, UK). Of the 19336 mutations selected, only 9175 could have a high-quality bait designed for them (some mutations are not suitable for bait design because, for example, they fall in a repetitive region of the genome). Driver mutations that had been included to allow us to detect clonal haematopoiesis of indeterminate potential could not, of course, be assigned to the tree, and neither could mutations that were present in polyclonal samples and not in any clonal samples. After final data curation and re-assessment of mutation assignment to the tree, 7116 mutations could be assigned to the tree of 140 clonal samples unambiguously and used for analysis.

5. Mutation calling

5.a. Substitutions

For all studies, substitution calling was broken down into three steps: mutation discovery; post-hoc filtering to produce a list of clean sites; and, finally, genotyping, where the presence or absence of every mutation in every sample is evaluated. Mutations were initially discovered using the Cancer Variants through Expectation Maximisation (CaVEMan) algorithm (Jones et al. 2016). CaVEMan uses a naïve Bayesian classifier to derive the probability of all possible genotypes at each nucleotide. CaVEMan copy number options were set to major copy number 5 and minor copy number 2 for normal clones, as in our experience this maximises sensitivity, but for analysing tumour organoids the tumour copy number profile from ASCAT (see below) was used. For all studies, the algorithm was run using an unmatched normal in order to be able to derive phylogenies: had another sample from the same individual been treated as a matched normal, early embryonic mutations would have been treated as germline and discarded, resulting in incorrect trees. Some post-processing filters were used in all studies. These included filtering against a panel of 75 unmatched normal samples to remove common single nucleotide polymorphisms, and two

filters (only applied to whole genome sequencing data) designed to remove mapping artefacts associated with BWA-MEM: the median alignment score of reads supporting a mutation should be greater than or equal to 140, and fewer than half of these reads should be clipped. After these filters, a pileup of all the samples from a given patient was constructed, counting the number of mutant and wild type reads in every sample over every site that had been called in any sample from that patient. Only reads with a mapping quality of 30 or above and bases with a base quality of 30 or above were counted.



Different post-processing filters were applied depending on the dataset.

5.a.i. Substitution calling in blood colonies

To build a phylogeny from 140 colonies sequenced at low coverage requires a stringent set of mutation calls. To minimise the number of false positive mutations, the following post-processing filters were included in addition to those detailed above:

- To remove germline:
    - We knew from initial manual tree building that there was an early split in the tree into two sample groups with more than 20 clonal colonies in either group. Therefore, if a mutation had mutant reads in greater than 120 out of 140 clonal colonies we considered it as germline.
- To remove false positives, or mutations that would be uninformative for tree building:
    - Mutations that fell within 10 base pairs of each other or within 10 base pairs of indels were removed.

- Mutations that had a sequencing coverage in a given sample of less than 6 reads for autosomes and less than 3 reads for sex chromosomes were given a status of NA in that sample. Mutations with this NA status in more than 5 samples were removed.
- We observed that mapping artefacts could cause a mutation to be seen at low VAF in multiple samples. We therefore applied the following filters
    - the mean VAF for a mutation, across all samples in which there are any mutant reads at all, should be greater than 0.3.
    - exclude sites where more than 10% of the samples in which there are any mutant reads have a VAF of less than 0.1

This resulted in 134,411 sites that were carried forward for further evaluation using the Shearwater algorithm (Gerstung et al. 2014), which was run by Sebastian Grossmann. This algorithm compares allele frequencies of variants to a background error model derived from sequencing thousands of samples from unrelated studies on the same platform. Sequencing errors are known to occur at different frequencies across different sites of the genome. By obtaining a comprehensive view of the number of variant calls at each position in unrelated normal genomes, we built an error model for each nucleotide change for each position. This method has previously been employed to find variants supported by small numbers of reads (Gerstung et al. 2014; Martincorena et al. 2015). For this study, the position-specific error probabilities were inferred from 234 whole genome sequences from unrelated normal samples from various donors and tissues. The ShearwaterML model to compare observed variant frequencies with the background error model is described in the R-package deepSNV-1.21.4 (Gerstung et al. 2014). After correcting for multiple testing with the Benjamini-Hochberg procedure, only variants were kept that were significantly mutated over the error model, using a corrected p value cutoff of 0.05. Although most true variants largely exceed this threshold, this procedure maximizes the chance of retaining variants on a relatively low number of reads but that occur in parts of the genome that suffer from few false positive calls. As a further stringent filter to minimise false positive calls, variants had to be supported by at least 3 mutant reads to be considered by the algorithm. In this way, every mutation called in an individual was genotyped as being present or absent in each sample.

5.a.ii. Substitution calling in colon microbiopsies

In addition to the filters described in section 5.a., additional filters were designed by Mathijs Sanders to remove artefacts associated with low-input library construction. The library preparation protocol for microbiopsies produced shorter library insert sizes than standard methods. It was, therefore, common for paired-end reads from microbiopsy libraries to overlap, resulting in double counting of mutant reads. Mathijs Sanders therefore generated fragment-based statistics to prevent the calling of variants supported by a low number of fragments. Variants were annotated by ANNOVAR (Wang et al. 2010) and fragment-based statistics (fragment coverage, number of fragments supporting the variant, fragment-based VAF) were calculated for each variant after the exclusion of marked PCR duplicates. In the rare event of a disagreement in the called base at the variant position between overlapping paired-end reads, the base with the highest quality score was selected. Fragment-based statistics were calculated separately for all fragments, only counting those with alignment score ≥ 40 and base scores ≥ 30. Variants supported by at least 3 high quality fragments were retained and used for the next stage of variant filtering.

Examination of variants called in microbiopsies demonstrated that an excess was present within inverted repeats capable of forming hairpin structures, that they were supported by reads with very similar alignment start positions (i.e., not marked as PCR duplicates), and were primarily located close to the alignment start within the supporting reads. These variants were frequently within 1-30 base pairs of another variant. Filtering based on variant proximity alone would also remove actual kataegis events, and so could not be used. *In silico* modelling of the potential hairpin revealed that the variants called in the same read were aligning to each other in the stem of the structure, but could not form a base pair (i.e., mismatched), while all other bases could. Careful consideration indicated that the artefacts were the consequence of erroneous processing of cruciform DNA (existing either prior to DNA isolation or formed during library preparation) by the enzymatic digestion protocol applied. Mathijs Sanders considered modelling the hairpin structures to filter these variants, but given the fact that read clustering (i.e. similar alignment position) serves as a strong hallmark for these artefactual variants, he opted to use the proximity of the variant to the alignment start, and the standard deviation (SD) and median absolute deviation (MAD) of the variant position within the supporting reads, as features for filtering. These statistics were calculated separately for positive and negative strand aligned reads. In case the variant was

supported by a low number of reads (i.e., 0-1 reads) for one of the strands, the filtering was based only on the statistics generated for the other strand. Per variant, if one of the strands was supported by too few reads, it was required for the other strand that either: (I) there should be ≤ 90% of supporting reads to report the variant within the first 15% of the read starting from the alignment start, or (II) the statistics MAD > 0 and SD > 4. Per variant, if both strands were supported by a sufficient number of reads it was required for both strands separately that either: (I) there should be ≤ 90% of supporting reads to report the variant within the first 15% of the read, (II) the statistics MAD > 2 and SD > 2, or (III) that the other strand should have the statistics MAD > 1 and SD > 10 (i.e., the variant is retained if the other strand demonstrates strong measures of variance). The proposed strategy reduced the number of artefactual variants while retaining all other variants, as assessed by running the last filtering step on WGS data from non-LCM experiments.

After applying these filters, mutations were genotyped based on the number of mutant and wild type reads at each locus. Mutations were called based on a variant allele fraction (VAF) > 0.2, a depth > 7, and at least 4 mutant reads. If the depth over a locus was less than seven in a given sample, or if there was more than one mutant read but the other criteria were not met, the genotype was set to NA for tree construction purposes. Loci that were set to NA in more than one third of the samples were removed for construction of the phylogeny. Positions were called as germline if they were either called as present or NA in all of the samples from a given patient.

5.a.ii. Substitution calling in colonic organoids

As organoids were sequenced with normal library preparation methods at higher coverage than other samples, they were fully clonal, and there were fewer samples from each patient (which makes phylogeny construction easier), less stringent filtering was required. However, a calling method had to be developed that allowed the detection of a pre-malignant clone that could contain both the normal organoids (which were derived from >5cm away from the tumour) and the tumour (a field effect). For each patient, the only germline reference available was healthy colorectal tissue >5cm distant from the tumour, consisting of epithelial and connective tissue. We reasoned that if there were a field effect the somatic mutations in this tissue should not be fully clonal. We therefore deducted germline mutations on the basis that they were fully clonal in the bulk normal, while

mutations that were subclonal in the bulk were not removed from the analysis. To define a mutation as subclonal in the bulk, the probability of finding the observed number of mutant reads or fewer given the sequencing coverage had to be less than 0.005, based on the binomial distribution with a probability of 0.5 for autosomes. Mutations that failed to meet this criterion were considered to be germline and were removed. Mutations were then genotyped using the Shearwater algorithm, as for blood colonies. The Shearwater algorithm was run by Sebastian Grossmann.

5.b. Small insertions and deletions (indels)

As for substitutions, calling of indels was broken down into mutation discovery, filtering, and genotyping. Mutations were called with the Pindel algorithm (Raine et al. 2015) using an unmatched normal. Post processing filters were applied as in Nik-Zainal et al. (Nik-Zainal et al. 2012), and the number of mutant and wild-type reads was tabulated as above. The same dataset-specific filters were applied as for substitutions. Indels were then genotyped based on a VAF>0.2, a depth of at least 10, and support of at least 5 mutant reads.

5.c. Short tandem repeats (STRS)

Short tandem repeats were called only for the blood colonies in order to test the robustness of the tree. The HipSTR algorithm to call them was run by Sebastian Grossmann (Willems et al. 2017). HipSTR was run using *de novo* stutter estimation and STR calling with *de novo* allele generation as per the author's recommendation. The calls were filtered on a calling quality of at least 0.95 and at least six reads per sample spanning STR loci on autosomes and at least three reads on sex chromosomes. Calls were also filtered if more than 15% of the reads were affected by PCR stutter or featured indels in the STR flanking regions. If STR calls were removed in more than five samples, the locus was flagged for all samples.

5.d. Structural variants

Genomic rearrangements were called using the BRASS algorithm (Li et al. 2017, https://github.com/cancerit/BRASS). Abnormally paired read pairs from WGS were grouped and filtered by read remapping. Read pair clusters with ≥50% of the reads mapping to microbial sequences were removed, as were rearrangements where the breakpoint could not be reassembled. Candidate breakpoints were matched to copy number breakpoints defined by ASCAT within 10kb. Only structural variants where the two breakpoints were more than 1000 base pairs apart were considered. Blood colony structural variants were called against another colony, colonic microbiopsies against a matched normal when available and against another crypt when not, and colonic organoids against bulk normal epithelium.

5.e. Copy number changes

Copy number changes were called using the Allele-Specific Copy number Analysis of Tumours (ASCAT) algorithm (Van Loo et al. 2010, Nik-Zainal et al. 2012, https://www.crick.ac.uk/peter-van-loo/software/ASCAT). The same matched normal sample was used as for calling structural variants. For additional validation of copy number changes in normal colon, the QDNAseq algorithm (Sheinin et al. 2014) was run. ASCAT uses both the read depth and ratios of heterozygous single nucleotide polymorphisms to determine an allele-specific copy number, while the QDNAseq relies solely on variations in sequencing depth. To call amplifications and deletions in the colonic microbiopsy cohort, only those that were called by ASCAT and showed a clear departure from the background on QDNAseq were retained. To call copy neutral loss of heterozygosity in this cohort, all such events called by ASCAT were checked visually on Jbrowse (Buels et al. 2016) to verify an imbalance of parental snps.

6. Construction of phylogenies

Phylogenies were constructed using the matrix of genotypes (with samples as columns and mutations as rows) derived for every patient.

6.a. Derivation of the phylogeny of blood

The phylogeny of blood was built using the maximum likelihood implementation of SCITE (Jahn et al. 2016), a tree-inference algorithm that uses Markov chain Monte Carlo sampling with an error model that takes potential false positives and false negatives into account for tree scoring. The SCITE algorithm was run by Sebastian Grossmann. The false positive error rate was known as the probability of every mutation being real is provided by Shearwater. False negative error rates were estimated by visual inspection using jBrowse of over 1,000 sites that had been provided to Shearwater but not called as present. We ran five independent Markov chain Monte Carlo simulations with one million iterations to test for robustness of the obtained phylogeny that all resulted in the same tree. Furthermore, the results were consistent with manual tree building attempts on a subset of clonal colonies.

To test the robustness of phylogeny, we bootstrapped the substitution input matrix for SCITE 1,000 times and ran SCITE 1,000 times to generate bootstrapping p values for each node in the tree (figure 1.4a). Furthermore, we took the following steps to rebuild our tree using different data and/or different tree building methodology:

1. We repeated tree building from the same matrix of substitution calls by maximum parsimony, bootstrapping 100 times.
2. We repeated tree building from a combined matrix of substitution and indel calls by maximum parsimony, bootstrapping 100 times.
3. We used the HipSTR algorithm to call STRs and built a tree with neighbour joining, bootstrapping the input matrix of mutation calls 100 times.
4. We built the tree using all variants combined (substitutions + indels + STRs) using neighbour joining, bootstrapping 100 times.
5. We built the tree using substitutions and non-STR indels using maximum parsimony, bootstrapping 100 times.

These are detailed below, and the results of each tree-building approach are shown in figure 1.4. SCITE and HipSTR were run by Sebastian Grossmann, and all other work performed by me. It should be noted that the shape of our tree, with multiple short branches that are ancestral to many

cells, makes bootstrapping conservative. As some of the earliest splits in the tree are supported by a single mutation which may be omitted in a bootstrap replicate (a given mutation will be present on average in 630 out of 1000 replicates), some of the earliest splits may not be supported by a particularly high proportion of replicates, even though the mutation calls are highly confident (based on multiple colonies carrying 5-10 reads reporting the variant, while being completely absent from other colonies). Indeed, shorter branches have lower p values.

When comparing trees, the quantity of interest is whether high-confidence nodes from other datasets and tree-building methods agreed or disagreed with our tree. For the purpose of this analysis, we define a high confidence node as one that created the same split of samples into two groups in >=70% of bootstrap replicate trees built with a given method.

1. Substitutions using maximum parsimony and bootstrapping.

The input matrix of substitution calls was bootstrapped 100 times and trees were built with each bootstrap replicate using maximum parsimony. We used the mix programme from the phylip (Felsenstein 1989) suite of tools, using the Wagner method. Ignoring terminal nodes (which are uninformative), 58 out of 139 nodes had >70% support. The fact that such a low proportion of nodes was well-supported is most likely due to the fact that many of the branches at the top of the tree are supported by no mutations (hence the polytomies in our original tree) or a small number of mutations, which need not be present in every bootstrap replicate. Of the 58 well-supported nodes, 56 were present in our tree, and two were absent. The differences only involve a minor reordering of branches within a clade. The discrepancies are not supported by many mutations in the SCITE tree.

2. (Non-STR) Indels using maximum parsimony and bootstrapping

There are only 228 shared indels outside of short tandem repeats (STRs), and so it is not surprising that a tree built only from these should have a high degree of uncertainty and several polytomies. Only 22 nodes (out of 126 this time because of polytomies in the parsimony tree) are supported by >70% confidence. Two of these 22 disagree with our original SCITE tree. As with the substitutions, these two differences are minor departures from the original SCITE tree.

3. STRs with neighbour joining bootstrapping results

Short tandem repeat indels were called with HipSTR (see above). The phasing quality was low and so the sum of alleles at a given locus was used, ignoring the parental origin. Trees were built with neighbour joining using the absolute distance between cells. We bootstrapped the matrix of calls 100 times, recalculating the distance matrix each time. The resultant consensus tree was uncertain, with multiple polytomies and only 2 nodes supported by 70% of the trees. Of these two high confidence nodes, one agreed with our original SCITE tree and one disagreed. This discordant node sits at the top of one clade with 10 descendants in the STRs NJ tree. In our original SCITE tree, however, these cells are scattered around the tree. Some of these cells share large numbers of substitutions with cells not contained in the STRs NJ clade (such as BMH97 (inside the STR NJ clade), which shares ~500 mutations with BMH73 (outside the STR NJ clade)), which leads us to trust the substitution tree over the STR tree.

4. Combined substitutions, indels, and STRs using neighbour joining

We tried neighbour joining on the combined dataset of STRs, substitutions, and indels, bootstrapping the matrix of mutation calls 100 times. As there were 92,661 shared STRs, 9,982 substitutions, and 228 indels, the STRs dominate each bootstrap replicate. As with the tree built only from STRs using neighbour joining, few nodes are well-supported. Of 27 nodes that were present in 70% or more of the trees, 23 were present in our original SCITE tree and 4 were discordant. Three of these 4 nodes were nested inside each other, and so really represent only two discordant clades.

5. Combined substitutions and indels analysis using maximum parsimony

Because of the lack of consistency among trees built by neighbour joining from bootstrap replicates of STRs, we performed an analysis on combined substitutions and indels, excluding STRs. We used maximum parsimony, as different false positive and false negative rates for substitutions and indels meant that SCITE would not have been appropriate. Of 59 nodes that were present in 70% or more of the trees, 56 were present in our original SCITE tree and 3 were discordant. The arrangements of the nodes in these three discordant clades is similar to their arrangement in the original SCITE tree.

In summary, the vast majority of high confidence nodes found using other data types and tree building methods support the tree that we built originally with SCITE. Furthermore, with the exception of a couple of overlaps between maximum parsimony substitution and indel trees compared to maximum parsimony on combined subs and indels (i.e. the same tree building method on nested datasets), the disagreements between our tree and those built with different datasets or tree-building methods were not recurrent: *different disagreements were found with each alternative tree*. We are therefore confident that the tree that we have provided is the best tree that we could have built.

The accuracy of the reconstruction of the tree matters principally for four findings:

1. The uneven contribution to the embryo of the first division that we can reconstruct.
2. The relationship between different cell types (stem cell vs progenitor, and peripheral blood-derived stem cell vs bone marrow-derived stem cell).
3. The timing of branch points throughout life for phylodynamic inference (both for the stem cell population size trajectory and the estimation of the number of stem cells contributing actively to granulopoiesis).
4. Timing when in life mutations occur, such that we can assign our mutations to the tree. This is important both in terms of estimating stem cell pool size and for the analysis of stem cell clone contribution to different mature blood cell types.

For the embryology analysis, since mutations were checked manually we are confident that the uneven split into two groups at the top of the tree is correct. For the latter three points, the most important features of the tree are later branchpoints. The precise arrangement of short branches right at the top of the tree is less informative than the branches that we observe later in molecular time. These branches are necessarily supported by more mutations shared by a small number of colonies, and so we are very likely to reconstruct them correctly. Indeed, none of these later branch-points were called into question by our different tree-building analyses.

6.b. Derivation of the phylogeny of colonic microbiopsies

Derivation of a phylogeny of colonic crypts was less challenging than that of blood because there were fewer samples per patient, and accuracy in the precise arrangement of short early branches was less important because we perform no analyses of the relatedness of different crypts. All that phylogenies are used for in this analysis is for timing mutations. The most informative branches in this case are the long branches shared by a small number of crypts, which are very robust to all tree construction methods. For this reason, trees were built using maximum parsimony, as for the validation of the blood phylogeny. The tree was bootstrapped 100 times and the consensus taken.

6.c. Derivation of the phylogeny of colonic organoids

Phylogenies were constructed by maximum parsimony, as detailed in the validation of the phylogeny of blood. The SCITE (Jahn et al. 2016) and Sifit (Zafar et al. 2017) algorithms were run for validation and produced the same results (data not shown).

7. Assignment of mutations to the phylogeny

Phylogeny inference programmes provide the topology of the tree but not the assignment of mutations. Mutations from the input matrix of genotypes therefore have to be re-assigned to branches. In order to assign a set of mutation calls with no false negative and no false positives to a tree, each branch of the tree would be considered in turn. If a mutation was called in all the descendants of a given branch, and in no samples that were not descendants of the branch, mutations would be assigned to that branch. However, the complexity of the phylogeny and the number of false positive and false negative calls in the mutation matrix affect how many mutations fit the tree perfectly. Because of differences in the datasets, slightly different approaches were taken to assign mutations to each tree.

7.i. Assignment of mutations to the phylogeny of blood

Given the size of the phylogeny of blood, its shape, and non-zero false negative and false positive rates of mutation calls, we expected that a proportion of mutations would not fit our tree perfectly. Consider a mutation that is truly present in 50 colonies. With 15x coverage over the site in every sample, if we simulate resampling of mutant reads from the binomial distribution (with size of 15 and probability of 0.5), 17% of the time the mutation will have fewer than the 3 reads required to call it in at least one sample. Variation in sequencing depth and sequencing errors would further decrease the probability of calling the mutation perfectly in every sample. Mutations were therefore assigned in the following way:

- Private mutations, only called in a single sample, could be assigned perfectly to the branch ancestral just to that sample.
- 6,819 out of 9,982 mutations called in more than one clonal sample fit the tree perfectly. These were assigned to the branch of the tree that was ancestral to all the colonies that bore the mutation to none of the colonies that did not bear the mutation.
- The 3,163 mutations that did not fit the tree perfectly were assigned in a probabilistic manner.
    - True positive and true negative rates of 0.99 were chosen.
    - For every node, the probability that a mutation that was truly present in that node given the number of positive and negative calls in the descendants of the node and in all of the colonies that do not descend from that node, was calculated.
    - The mutation was assigned to the node which had the highest probability of the mutation being there. 1,306 mutations that had ambiguous placements (they fit two or more nodes in the tree equally well) were not assigned.

7.ii. Assignment of mutations to the phylogenies of colonic microbiopsies

Colonic microbiopsies also suffered from low coverage, and, what is more, some stromal contamination. For this reason we did not expect mutations to fit the tree perfectly. Unlike in the phylogeny of blood, where assignment of mutations to branches mattered for the recapture phase, here mutations were only assigned to the tree in order to determine the mutational processes active at a particular time. We reasoned, therefore, that it was preferable to assign only mutations that fit the tree perfectly, and rather adjust the branch lengths based on the power to call mutations at a given branch. Using the clonality and coverage of all descendants of a branch, the proportion of true substitutions or indels on the branch that would be first discovered (whether by CaVEMan or Pindel) and then genotyped as present according to the criteria described above was calculated. The observed branch length was then adjusted by dividing by this proportion. This was done for both substitutions and indels, but not for structural variants and for larger copy number changes due to a lack of data: most branches have no large variants and so could not be extended appropriately. Rearrangements and copy number changes were assigned to phylogenies manually.

7.iii. Assignment of mutations to the phylogenies of colonic organoids

The colonic organoid data was high coverage (~30X) and fully clonal, and with relatively small numbers of samples per patient. Ignoring private mutations, which necessarily fit any tree, 97.7% of shared mutations fitted the tree structure from patient 1 perfectly, 89.7% fitted the tree from patient 2 perfectly, and 88.1% fitted the tree from patient 3 perfectly. The lower concordance with the tree for patients 2 and 3 reflects the increased copy number changes that have occurred in these phylogenies. Examination of the copy number state at loci where there were discordant mutations showed that the majority could be explained by deletions of those mutations in a subclone. Substitutions that did not fit the tree perfectly were therefore assigned to the most recent common ancestor of the samples in which they were called.

The approach for assigning rearrangements was slightly different, as the same rearrangement may be called in related samples with slightly different breakpoints. To identify rearrangements that had been sequenced in related clones as the same, both the upstream and the

downstream breakpoints had to fall within 500bp of each other. The majority of rearrangements fitted the tree. Visualisation of discordant rearrangements using IGV (http://www.broadinstitute.org/igv) showed that often an overlapping rearrangement meant that the rearrangement was lost in a clone.

8. Analysis of relationships in phylogeny for blood

Testing for differences in the relatedness of cell types on the tree was carried out by analysis of molecular variance (AMOVA (Bird et al. 2011; Excoffier et al. 1992)), comparing stem cells derived from the bone marrow with those from peripheral blood, comparing stem cells relative to progenitors, and different progenitor types relative to one another. The phylogeny was first made ultrametric by extending private branches to the maximum branch length of 1210 mutations. Then the mutational distance between two samples (i.e. the number of mutations over which you would have to walk to go from one cell to another on the tree) was calculated for all sample pairs. Within-population and between population sum of squared distances were calculated, divided by their degrees of freedom, and used to estimate the fraction of total variation explained by differences between cell populations (the F statistic). To calculate p values, the population labels of cells were randomly re-assigned 30,000 times, and the same statistic was calculated. The p value is the proportion of random re-assignments that has an F statistic more extreme than the observed value.

9. Analysis of population size trajectory for tree of blood

Analysis of population size trajectory was performed using the Phylodyn package (Karcher et al. 2017, Lan et al. 2015), using a phylogeny made ultrametric by extending private branches to the maximum branch length of 1210 mutations, and with 70 grid points.

10. Targeted sequencing mutation analysis of peripheral blood

A stringent error-correction method of counting the number of mutant and wild type reads over a locus was devised and executed by Robert Osborne. BAM files were annotated with read coordinate (rc), mate coordinate (mc) and optical (od) auxiliary tags using biobambam2 (Tischler and Leonard 2014). Reads were included for analysis if they were marked as proper-pairs, had a minimum mapping quality ≥ 30, < 3 mismatches and were not marked as optical duplicates, supplementary, QC fail, unmapped or secondary alignments. We also restricted analyses to bases with base quality ≥ 30. Overlapping reads can result in double-counting of a base. If the base calls on read 1 and read 2 were not identical then the call was discarded. If the base calls on reads 1 and 2 matched then the call was assigned to the read with the highest base quality score. If the call on reads 1 and 2 matched and the base quality scores on both reads was identical then the call was randomly assigned to one or other reads. Instead of discarding PCR duplicate molecules, we generated consensus calls by grouping calls from reads with the same fragmentation breakpoints, read orientation and read number. A consensus base was called if >90% of the reads shared the same base. There were no minimum number of reads (if the group contained only one read then its call was retained).

For plotting purposes, signal was then separated from noise by using data from control cord blood from two individuals using a Bayesian generalised mixed effects Poisson model which was written by Peter Campbell, with minor amendments by me. The input data comprise the observed number of reads reporting each variant in the control sample and the corresponding total depth across that base; and the same for each test sample from our subject. A Poisson model was fitted in which the dependent variable was the number of reads reporting the variant, including an offset for log(total depth) (so that we are really estimating the fraction of reads). Random effects were included for the branch in the test sample (allowing for the VAF of mutations on the same branch to be correlated); the mutation (allowing for variable error rates among different mutations); and the interaction term of specific mutation by test sample (allowing for the test sample to have excess reads reporting the variant than the control sample). Parameter estimates and residuals are assumed to be drawn from a multivariate normal distribution with uninformative conjugate priors on the covariance. The model was fitted with the R package 'MCMCglmm', with 300,000 iterations, a burn-in of 10,000 iterations and thinning to 1/1000 iterations. Mutations were called as 'detected'

if more than 90% of iterations of the MCMC chain after burn-in estimated a positive term for the test sample.

11. Detection of driver mutations and positive selection

11.a. Detection of driver mutations and positive selection in blood

We searched both for particular mutations that could confer a selective advantage (driver mutations), and for global signals of positive selection in colonies that had undergone whole genome sequencing.

First, we first took all coding mutations called by CaVEMan or Pindel, but without applying post processing filters beyond those detailed in Nik-Zainal et al (2012) in order to maximise our sensitivity. We also searched for copy number changes and rearrangements that might affect known myeloid cancer genes. We then intersected this with a list of 348 genes that are found to be mutated in at least 10 haematopoietic and lymphoid neoplasms from the COSMIC database, found to be mutated in CHIP (Jaiswal et al. 2014; Genovese et al. 2014, Xie et al. 2014, Mckerrell et al. 2015), in AML (Cancer Genome Atlas Research Network et al. 2013), or in the COSMIC cancer gene census (Forbes et al. 2017). Any mutations in genes in our list were visually inspected on jBrowse as a form of verification, and we used prior knowledge of the gene's function and the ways in which it is mutated to decide whether the mutation that we observed was a likely driver mutation.

Second, we tested for positive and negative selection of mutations using dNdScv (Martincorena et al. 2015, Martincorena et al. 2017). In brief, this algorithm compares the number of synonymous and non-synonymous mutations in coding regions, adjusting for the pattern of mutational processes in the sample and local mutation rates, to detect if there is an excess or paucity of non-synonymous mutations relative to what is expected by chance under a neutral model. An excess of non-synonymous mutations indicates positive selection, while a paucity indicates negative selection. This method of driver detection is therefore independent of prior knowledge. This allowed us to search both for genes that might be under positive selection, and to see the

global dNdS of all the mutations, which tells us whether there is any excess of non-synonymous over synonymous mutations in the whole dataset.

Finally, 100 of the most frequent mutation hotspots in clonal haematopoiesis of indeterminate potential were included in the bait-set for targeted sequencing of peripheral blood, allowing us to detect driver mutations that may not have been present in the clones that were originally whole genome-sequenced (Appendix A).

11.a. Detection of driver mutations and positive selection in colonic microbiopsies

Again, driver mutations were detected both through an unbiased dNdS method and through manual annotation. For these analyses, the CaVEMan and Pindel calls were used without post-processing filters in order to maximise our sensitivity. All putative driver variants were visually inspected using Jbrowse (Buels et al. 2016), and so we could afford a higher false positive rate in the mutation discovery phase.

dNdScv (Martincorena et al. 2015, Martincorena et al. 2017) was used to conduct three tests: first, using only the whole genome sequencing data, an analysis of selection over all genes; second, using combined whole genome and targeted sequencing data, over all the genes covered by the bait-set; and finally, using again this combined dataset, over 90 selected cancer genes (Appendix B).

Manual annotation of driver variants based on prior knowledge complemented this. A list of 90 colorectal cancer genes (Appendix B) curated from the literature, that were also covered by the bait-set were intersected with the list of substitutions and indels from combined whole genome and targeted sequencing. Mutations were annotated as putative drivers if they were either missense mutations that fell in an oncogene hotspot (based on visualisation of the distribution of mutations in the gene on COSMIC (Forbest et al. 2017)), or if they were nonsense mutations that fell in a tumour suppressor gene.

Finally, structural variants that might act as drivers were assessed by intersection of genes involved in each structural variant with the twelve genes involved in gene fusions that have been reported in colorectal cancer in COSMIC (*VTI1A, TCF7L2, TPM3, NTRK1, PTPRK, RSPO3, ETV6, NTRK3, EIF3E, RSPO2, C2orf44*, and *ALK*). No fusion genes were found. The genes

involved in structural variants in our data did not overlap with the list of 90 cancer genes used for assessing substitutions and indels, and nor were there any genes that were affected by more than one structural variant.

11.a. Detection of driver mutations and positive selection in colonic organoids

Driver analysis in colonic organoids was performed by Sophie Roerink. To classify driver events in substitutions, indels and rearrangements the following criteria were used: 1) deleterious mutations in genes identified in CRC by TCGA (Cancer Genome Atlas Research Network et al. 2012) 2) all other known oncogenes carrying a canonical activating mutation 3) tumour suppressor genes with loss of function, and/or carrying two deleterious mutations.

12. Adjusting crypt mutation burden by the callable proportion of the genome

Whole genome sequences from crypts have variable coverage and clonality, which affect the proportion of the genome that can be called. In order to compare crypts, crypt mutation burdens throughout the text are adjusted for their coverage and clonality. The read depths for 1,000 sites were sampled from each crypt in order to capture the variability in its coverage. For each site, a mutation was simulated. Numbers of mutant reads over a given simulated position were drawn from a binomial distribution with probability equal to the clonality of the crypt, and the number of trials equal to the coverage that has been sampled for that site. The proportion of simulated sites for which mutations are callable serves as an estimate of the callable proportion of the genome. Raw mutation burdens per crypt are divided by this proportion to estimate the total mutation burden per crypt.

13. Extraction of mutational signatures

13.a. Categorisation of mutations

Mutations were categorised following the method used by the Mutational Signatures working group of the Pan Cancer Analysis of Whole Genomes (Alexandrov et al. 2018). Single base substitutions were categorised into 96 classes according the identity of the pyrimidine mutated base pair, and the base 5' and 3' to it. Doublet base substitutions were categorised into 78 classes according to the identity of the reference and alternative bases. Indels were classified according to whether they were an insertion or a deletion, the identity of the inserted/deleted base, the length of the mononucleotide tract in which they occurred, or the degree of homology with the surrounding sequence into 83 classes (figure 2.1).

13.b. Non negative matrix factorisation (NNMF)

Consider a matrix, with the name of every sample along one side, the name of every mutation category that we have defined (e.g. C>A in A<u>C</u>A) on the other. The matrix is populated by the counts of mutations in each sample of each category. NNMF considers this observed matrix to be the product of a matrix of signatures and a matrix of exposures, plus some noise (Alexandrov et al. 2013b). The matrix of signatures is made up of the categories on one side, and the names of the signatures on the other, and populated by the proportion of a signature that is in a particular context. The matrix of exposures is made up of the samples on one side, the signatures on the other, and the amount that each signature contributes to each sample populates the matrix. The numbers that populate the matrices of signatures and of exposures are unknown. They are learnt by trying different combinations of numbers in these two matrices, and measuring the similarity of their product with the observed set of counts. In the classical version of the algorithm written by Ludmil Alexandrov, the distance metric used is the Frobenius norm (Alexandrov et al. 2013, Alexandrov et al. 2013b). The number of signatures that are used is not learnt from the data. Different numbers of signatures are tried, and a measure of the stability of the solution (calculated by bootstrapping the input data) and how well it approximates the observed data is used to select

the optimal number of signatures. With more signatures, the observed counts are approximated more accurately but the reconstruction is less stable.

13.c. Hierarchical Dirichlet Process

The use of a Hierarchical Dirichlet Process to extract mutational signatures was developed by Nicola Roberts (Roberts 2015, Roberts 2018). A Dirichlet Process is a non-parametric clustering method that takes as input a probability distribution and produces as output a more discretised probability distribution over the same domain. Signatures are treated as a multinomial probability distribution of the set of mutation classes (e.g. the trinucleotide context). The mutation counts per category in a sample are considered to be the result of draws from a sample-specific mixture of a shared set of signature probability distributions. The hierarchical nature comes from the fact the samples are put in groups (for example, in my analysis all samples from one patient were put in a group), and the amount that each signature contributes to a sample is drawn from parent node of the group; this has the effect that samples within a group are considered to be more similar to each other than samples from different groups.

Signatures are learnt by Gibbs sampling (Teh et al. 2006). A set of clusters is initialised with random counts in each category, all contributing randomly to every sample. In every iteration of the chain, mutations are shuffled between clusters. They are more likely to be drawn into clusters that resemble them (i.e. that have a high proportion of mutations in the same class). With some low probability, mutations can also form their own clusters, which means that the number of signatures in the dataset can be learnt rather than having to be specified. Averaging over posterior samples over the chain produces the number of signatures, their identity, and their contribution to each sample.

The algorithm can be conditioned on known signatures by including nodes of "fake data" that containing pseudocounts in the category distribution of a known signature. Mutations from the real datat that are similar to these known signatures are likely to be drawn into their clusters. Mutations that are not similar to any of the preconditioned clusters, however, can still form their own clusters, allowing simultaneous matching to known signatures and discovery of new ones.

14. Timing copy number changes in colonic microbiopsies

If substitutions are assumed to occur at a constant rate, the copy number of mutations over a particular copy number segment allows us to time when the copy number change occurred. For example, if one chromosome copy is amplified (i.e. a change from 1+1 to 2+1, where the number indicates the number of copies of each parental chromosome), then mutations that occurred on the amplified chromosome prior to its amplification will be on two copies, whereas those that occurred after the amplification will only be on one copy. Mutations were timed using the MutationR package (Gerstung et al. 2017, https://github.com/gerstung-lab/MutationTimeR), which accounts for lower detection sensitivity of mutations at lower copy number. To maximise reliability only copy number changes with that contained 100 substitutions at good coverage were timed. Trinucleotide profiles of substitutions estimated to occur before and after the copy number change were inspected and found to be very similar, suggesting that there has been no change in the mutational processes operative since the copy number change, and so supporting our assumption of a constant mutation rate.

15. Timing mutations relative to a whole genome duplication

A whole genome duplication (WGD) was observed in the trunk of the tumour for patient 2 in the organoid study. Timing as many mutations as possible relative to this allowed the investigation of the evolution of mutational processes.

15.a. Timing substitutions relative to the whole genome duplication

For every truncal substitution in every tumour clone from patient 2, the copy number segment (as called by ASCAT) in which that mutation fell was defined. Mutations could only be timed in samples in which there was a minor copy number of 0 and a major copy number greater than 1. Fortunately, because of the extensive copy number changes in this tumour, all mutations

fell in a region that met these criteria in at least one sample. For a given mutation that fell in such a copy number segment in a given sample, the VAF in that sample of known germline single nucleotide polymorphisms that fell in that segment (that necessarily occurred before the WGD) and the VAF of somatic mutations assigned to branches further down the tree (that necessarily occurred after the WGD) was examined. If, in a given sample, a mutation had a VAF greater than 90% of the VAFs of the mutations that were known to occur further down tree it was considered to have occurred before the WGD, whereas if it had a VAF less than 90% of the VAFs of the SNPs it was considered to have occurred after the WGD. If there was any overlap between the 90th percentiles of the SNPs and the later mutations, or if the mutation fitted neither of these criteria, it was considered uninformative and was not used in the signature analysis. This accounted for 9,094 mutations (out of a total of 12,623 assigned to the trunk), that were not used in signature analysis. There is no reason to believe that mutations that were excluded for these reasons should be attributable to different mutational signatures than those that could be included, and indeed their trinucleotide mutation contexts are similar (data not shown). For each mutation, then, the number of samples in which it had been counted before and after the WGD was tallied. If a mutation was called as occurring before the WGD in some samples and after the WGD in others, the mutation was designated as conflicting and excluded from the analysis. 82 mutations fell into this category, and the remaining 3,447 could be timed unambiguously relative to the WGD and used in the signature analysis. In figure 1 we extrapolated the preWGD and postWGD fractions and their relative signature components to all mutations identified in the clonal trunk of P2.

15.b. Timing driver mutations relative to the whole genome duplication

Driver mutations in *TP53* and *APC* were timed relative to the WGD in patient 2. The *TP53* mutation was at VAF 1 in a region that was 2+0 in all samples, indicating that it occurred before the WGD. There were mutations in both alleles (which we will call mutation 1 and mutation 2) of *APC*. P2.T4.2 and P2.T5.1 both had the *APC* locus called as 2+2, and both mutations were at VAF 0.5. P2.T1.1, P2.T1.3, and P2.T6.2 were 2+1 in the *APC* region. Mutation 1 was at VAF 0.67 and mutation 2 at 0.33. In P2.T2.5 the region was also called as 2+1, but mutation 1 was at VAF 0.33 and mutation 2 at VAF 0.67. This shows biallelic inactivation of *APC* prior to the WGD.

15. Placing a lower bound on the age at which cells acquired aberrant mutational processes

The onset of signature 18 was timed in each patient relative to the rate of signature 1 (see Results Chapter 2, section R.7.b.). Calculations are found in table M.1.

	patient 1	patient 2	patient 3
branch to calculate ratio of signature 1 : signature 18	Ancestor of all clones except for PD21928c6	Segment after WGD but before the most recent common ancestor	Average of the two branches after the trunk
signature 1 mutations in branch	2088	975	2043
signature 18 mutations in branch	2926	410	2571
signature 1:signature 18 in branch	0.7	2.4	0.8
signature 18 mutations in trunk	1054	934	1617
signature 1 mutations in trunk	2508	3509	3050
estimate of signature 1 mutations after signature 18 onset	752	2221	1285
estimate of signature 1 mutations before signature 18 onset	1756	1288	1765
estimated years since signature 18 onset	24	20	22

Table M.1. Calculations to estimate the onset of signature 18.

RESULTS CHAPTER 1
CLONAL DYNAMICS OF NORMAL BLOOD

Introduction to this chapter

I.1. Haematopoietic stem cells

In adult homeostasis, hundreds of billions of blood cells must be produced every day, in the correct proportion of many specialised cell types, over the whole of life. Mature blood cell types include erythrocytes that transport oxygen, platelets that control haemostasis, and a battalion of cells with complementary immune functions: neutrophils, basophils, eosinophils, dendritic cells, macrophages, natural killer cells, B lymphocytes, T lymphocytes, and innate lymphocytes. Many of these have multiple subtypes. All mature blood cells ultimately derive from haematopoietic stem cells (HSCs). The conceptual definition of an HSC, therefore, is that it should have the potential to produce all blood cell types and to self-renew, producing daughters that have equal potential for the lifespan of the organism. In between stem cells and their mature progeny in the hierarchy of differentiation are progenitor cells, with a weaker capability for self-renewal and a more restricted set of possible differentiation fates. A progenitor will typically undergo multiple cell doubling and differentiating divisions to produce large numbers of functional mature blood cells.

The term *Stamzelle* was coined by Ernst Haeckel in 1868 as the first organism from which all life stemmed, in line with his Darwinian views (Haeckel 1868, Laurenti and Gottgens 2018). He later used it to describe the fertilised egg, from which all cells in the organism derive. Thus, as explained by Laurenti and Gottgens, the term *stem cell* has always had the connotation of being at the root of a phylogeny, whether germline or somatic (Laurenti and Gottgens 2018). Bone marrow cells that could generate mature blood cells were described independently by Bizzozero and Neumann (Bizzozero 1868, Neumann 1868) and a cell type that might give rise to all of blood was proposed in the late nineteenth century, when it was suggested that red and white blood cells might have a common ancestor (Pappenheim 1896). This idea remained controversial (Cooper 2011)

until 1945, when Owen noted that fraternal twin cattle that had shared a placenta also shared the blood cell types of both calves for life, thus providing evidence of a long-lived cell type that could be the source of blood (reviewed in Weissman and Shizuru 2008). Owen wrote:

> Since many of the twins in this study were adults when they were tested, and since the interchange of formed erythrocytes alone between embryos could be expected to result in only a transient modification of the variety of circulating cells, it is further indicated that the critical interchange is of embryonal cells ancestral to the erythrocytes of the adult animal. These cells are apparently capable of becoming established in the hemopoietic tissues of their co-twin hosts and continuing to provide a source of blood cells distinct from those of the host, presumably throughout his life.
>
> (Owen 1945)

In the 1950s, transplantation studies provided the means to begin to characterize these cells. Bone marrow was shown to be the site of residence of HSCs, as lethally irradiated mice could be saved by injecting them with bone marrow from other animals (Lorenz, 1951). Although regrowth of haematological tissues was observed in the salvaged mice, it was initially believed that this was due to a humoral factor in the marrow stimulating autochthonous regeneration. In 1956, however, a chromosomal (Ford et al. 1956) or enzymatic (Nowell et al. 1956) marker was shown to be shared by mature blood cells from both the donor and the recipient, indicating that cells themselves were transferred and that the restoration of haematopoiesis might be cellular. The approach described in this results chapter of using mutational markers to track stem cell clones traces its origin to these original experiments.

In a series of seminal experiments in the 1960s, Till and McCulloch showed the existence of multipotent stem cells. They induced chromosomal markers in one mouse by irradiation and transplanted them into a second mouse. Spleen colonies in the recipient were clonal for a given marker, indicating that a single cell could make all the cells in a colony (Becker et al. 1963). Re-transplanting cells from suspensions of spleen colonies into secondary recipients showed the capacity of some cells to self-renew through multiple rounds of transplantation and produce multilineage output (Siminovitch et al. 1963). Some chromosomal markers were found in both myeloid and lymphoid tissues (Wu et al. 1968), from which the authors concluded:

> These findings […] indicate that hematopoietic colony-forming cells, erythroblasts, granulocytes, thymic cells, and the cells of lymph nodes may all belong to the same clone. Our studies, however, do not allow us to determine precise parent-progeny relationships within such a clone. For example, we do not know whether or not some or all hematopoietic colony-forming cells can differentiate to give rise to both myeloid and lymphoid descendants or, alternatively, if colony-forming stem cells and lymphoid cells have a common, as yet unidentified, precursor. The resolution of this problem will require further detailed analysis of patterns of differentiation within large hematopoietic clones.
>
> (Wu et al. 1968)

These questions are still not fully answered.

The invention of multi-parameter flow cytometry in the late 1970s provided new tools to study HSC behaviour. In the 1980s, the field moved to the prospective isolation of HSCs by flow cytometry on cell surface markers, assaying the potential of different populations to make the various mature cell types by *in vitro* colony assays or transplantation experiments (Visser et al. 1981, Visser et al. 1984). This work led to the hierarchy of stem cell differentiation found in most textbooks today (reviewed in Weissman and Shizuru 2008). Crucially, in 1988 it was shown that functional stem cells could be isolated (Spangrude et al. 1988).

The current working experimental definition of a stem cell has not moved on very far: a cell is generally accepted to be a stem cell if it can give rise to cells of both the myeloid and lymphoid lineages for at least 16 weeks within an irradiated host (Bryder et al. 2006), although a more rigorous test is whether its progeny can do the same when re-transplanted into a secondary recipient (Eaves 2015). In the intervening decades, multiple properties of stem cells have been characterised: for most of the time they should be quiescent (Wilson 2008, Foudi 2009, Cabezas-Wallscheid 2017), glycolytic (Simsek et al. 2010, Ito and Suda 2014), rely on autophagy (Warr et al. 2013, Ho et al. 2017), and not synthesise as many proteins as their descendants (Signer et al. 2014). None of these is absolute, and most of the time functional assays are required to show stemness.

I.2. The embryological origins of blood

Our study permits us to examine retrospectively the ancestry of adult blood cells, and so a very brief discussion of the embryology of blood aims to put this in context. The first blood cells in the developing embryo appear in the yolk sac (Sabin 1920). Primitive erythrocytes derived from the yolk sac assure oxygen transport in the developing fetus. They derive from a progenitor that is shared with the endothelial cells of the yolk sac's blood islands, and consequently called a haemangioblast (Choi et al. 1998). Based on shared surface markers, haemangioblasts are thought to originate in the posterior region of the primitive streak (Huber et al. 2004). Yolk sac-derived cells, however, are not the ancestors of most adult haematopoietic tissue. This was elegantly demonstrated by grafting quail embryo bodies onto chicken yolk sacs before development of vascular connections between the chicken body and yolk sac (Dieterlen-Lievre 1975). Quail cells have a prominent nucleolus, which could be used as a lineage marker. Blood from these chimaeras was shown to derive exclusively from quail cells in 16 out of 17 animals. The site of development of the cells that go on to provide life-long haematopoiesis was established by transplanting various parts of the mouse conceptus into adult irradiated recipients (reviewed in Dzierzak and Speck 2008). The aorta-gonad-mesonephros (AGM) region and the vitelline and umbilical arteries contained cells with the capability of long-term multilineage reconstitution. Haematopoietic clusters were seen to emerge from the ventral aspect of the dorsal aorta only three days after the beginning of yolk sac haematopoiesis, which indicates that the AGM and yolk sac haematopoietic precursors are unlikely to be particularly closely related. Haematopoietic cells made in the AGM, the yolk sac, and the placenta, all migrate to and reside in the fetal liver for the duration of embryogenesis, where they proliferate (reviewed in Mikkola and Orkin 2006). They colonise the bone marrow around the time of birth.

I.3. Estimates of the numbers of active HSCs

One of the aims of this thesis is to estimate the number of active HSCs in a healthy human. Estimates have varied widely over the past few decades. Here, I review important contributions to the field.

I.3.a. Animal models

A feature of studies that aim to determine the identity and output of stem cells is the need to label cells such that their progeny can be traced. The induction of labels generally involves genetic manipulation, which is not feasible in humans. Most studies, therefore, have relied upon animal models.

I.3.a.i. Transplantation studies in animal models

Transplantation studies carry a triple benefit for tracking cells: first, the cell type of interest can be isolated by flow cytometry; second, it is easier to label these cells outside the body than inside it; and third, transplantation into a recipient allows an assay of stem cell function, based on the concept that stem cells are able to reconstitute long term multilineage haematopoiesis. These studies, however, bear two major and linked disadvantages: first, they provide an assay of stem cell *potential* rather than *fate* (cells that behave as stem cells physiologically may not do so in transplant and *vice versa*); second, the stress of cell labelling and transplantation affects stem cell dynamics.

Initial studies of transplanting retrovirally-barcoded cells into genetically anaemic or irradiated mice indicated that one to three haematopoietic stem and progenitor cells (HSPCs) were responsible for the majority of the blood produced in the salvaged animal (Dick et al. 1985, Keller et al. 1985). Barcodes could be found in all lineages and after secondary transplantation, proving that the cells capable of reconstitution were from high up the differentiation hierarchy (Lemischka et al. 1986). Remarkably, transplantation of even a single cell was sufficient to reconstitute long-term multilineage haematopoiesis (Osawa et al.1996).

More quantitative approaches have relied upon limiting dilution transplantation experiments in mice to determine the frequency of four to eight HSCs per 100,000 nucleated marrow cells (Abkowitz et al. 2000). As mentioned in the General Introduction in the context of determining the clonality of tumours, measuring the expression of an allele on the X chromosome

(X chromosome inactivation or XCI skewing) of female heterozygote mammals can be used to detect an imbalance in clone sizes: because of random X chromosome inactivation early in embryogenesis, under a stable polyclonal model of haematopoiesis half of cells should express the paternal allele and the other half the maternal one. This approach was used to estimate stem cell numbers in cats. Following autologous transplantation with different numbers of stem cells, the degree of XCI skewing was used to estimate the frequency of feline stem cells in the bone marrow (Abkowitz et al. 1996). Combining frequency estimates with counts of nucleated marrow cells per animal (determined from the distribution of radioactive transferrin) resulted in the estimate of 11,000-22,000 HSCs per mouse and 6,000-16,800 HSCs per cat (Abkowitz et al. 2002).

Transplantation studies of an animal model closer to humans physiologically and genetically have also suggested the existence of thousands of stem cells. Autologous transplantation of irradiated macaques with large numbers of retrovirally-marked HSPCs suggested that at least 1,000 HSPCs contributed to multilineage haematopoiesis in the first year following transplantation (Kim et al. 2000). Longer-term follow-up of macaques has suggested that the first year post-transplantation is one of clonal instability (Kim et al. 2014). Nevertheless, follow-up for years still supports highly polyclonal long term multilineage haematopoiesis, with hundreds to thousands of unique markers detected (Kim et al. 2014, Koelle et al. 2017). It should be noted that long-term clonal marking assays will all underestimate the total number of *cells* contributing to haematopoiesis due to sequencing detection thresholds (a 0.05% detection threshold is reported in Koelle et al. 2017): small clones are missed, and a detectable one is counted as a single clone rather than the hundreds of stem cells that may descend from it.

I.3.a.ii. Studies of native haematopoiesis in animal models

Due to the disadvantages of transplantation studies outlined above, recent work has exploited advances in inducible genetic labelling to mark cells *in vivo*, allowing assays of actual stem cell fate rather than stem cell potential.

In 2014, the Camargo group used an inducible sleeping beauty transposon system to label HSPCs indiscriminately at one time-point (Sun et al. 2014). Four months later (allowing the shortest-lived cells to exhaust), they searched for tags among different blood fractions. They found

that over 90% of granulocyte tags were only ever found at one timepoint for up to 12 months of chase. The authors ascribe this result to a model of clonal succession, where stem cell clones are recruited successively to drive haematopoiesis and then exhaust (Kay 1965). For a label to be detectable by their semi-quantitative method, it probably needs to be present in a number of stem cells, given that they estimate the sensitivity of their assay to be ~0.1% and there are likely to be thousands to tens of thousands of active stem cells. If this is indeed the case, then the model of clonal succession would require a number of closely related stem cells to enter synchronously into a phase of doubling and differentiating cell divisions to fuel blood production, until they all exhaust at the same time. The coordination of multiple stem cell members of a clone seems implausible, and such a conceptualisation of haematopoiesis is perhaps a consequence of the experimental approach of labelling cells at one particular time-point. This separates a large branching phylogeny into distinct clades (or clones) by drawing a horizontal line at the time of labelling and suggests that a clone is a meaningful biological entity. If labelling had occurred earlier, clones would have contained more stem cells, and if later, fewer; there is no reason for all members of a clone to share a pattern of behaviour when the level of grouping is so arbitrary. The original hypothesis of clonal succession (Kay 1965), proposed that coordination of clones might be spatial or involve a positive feedback loop. Little evidence of these feedback loops has emerged in the fifty years since the hypothesis was formulated. In the absence of this evidence it is hard to reconcile clonal succession with polyclonal haematopoiesis.

Sun and colleagues found that only 7% of bone marrow granulocyte tags were shared with nascent pro/pre-B cells, which was interpreted to mean that haematopoiesis was largely driven by unipotent progenitors. There are, however, other plausible interpretations of this finding. Firstly, analysis of their results is complicated by the fact that both stem and progenitor cells are labelled: tags from unipotent progenitors will necessarily be short-lived and lineage-restricted, and will far outnumber tags in stem cells and multipotent progenitors. Putting that aside, as a long chase period was allowed, these results can be explained by the polyclonality of blood without the need to invoke clonal succession or consign the bulk of steady-state haematopoiesis to lineage-restricted populations. Let us suppose that 15,000 stem cells are present in a mouse (Abkowitz et al. 2002). Allowing for some neutral drift of the population over the chase period, these might retrace their ancestry to perhaps 12,000 stem cells at the time of labelling. The labelling strategy tags ~30% of HSCs, so ~4,000 clones are tagged. Sun et al. generally recapture a few hundred tags, so let us say

that 400 tags from granulocytes and another 400 from B lymphocytes are captured. Assuming that these are derived from HSCs that make exactly equal proportions of B lymphocytes and granulocytes, these each represent independent samples of 10% of the labelled stem cell clones, and we would only expect a 1% overlap between the two of them. This 1% of the total pool of labelled clones represents 10% of the labelled granulocyte clones that were captured, and so the authors would present this finding as showing that only 10% of tags from granulocytes were recaptured in B lymphocytes – a number not dissimilar from the 7% that they found. The same argument can be used for why granulocyte tags are infrequently recaptured. The overlap could be further reduced if the number of stem cells were larger (which it could quite conceivably be), some stem cells were even slightly biased towards one lineage or another, there were an imbalance in the number of tags recaptured from different samples, or if there were a degree of transposon reactivation during the chase period. Indeed, the transposon was found to have remobilized in the absence of induction in one of 24 secondary colonies derived to test the leakiness of their assay. Nonetheless, this study remains an elegant demonstration of the vast polyclonality of blood in unperturbed haematopoiesis.

Busch et al. took the alternative lineage-tracing approach of permanently inducing the expression of yellow fluorescent protein (YFP) in 1% of immunophenotypic HSCs and all their descendants (Busch et al. 2015). The population is studied as a whole: individual cells are not marked and consequently any stem cell heterogeneity is missed. By limiting dilution, they estimated that ~30% of immunophenotypic HSCs contributed to haematopoiesis over most of a mouse's adult life. As most strategies for isolating HSCs also mark some progenitors, this proportion could well be higher: it is conceivable that all stem cells contribute. No attempt was made to estimate the number of HSCs, but rather this proportion was combined with estimates from limiting dilution transplantation experiments of ~17,000 stem cells to derive a lower bound of ~5,000 HSCs actively contributing to haematopoiesis in the investigated time period. Slow equilibration between the proportion of the label in the stem cell and progenitor compartments was taken to mean that the bulk of haematopoiesis must be sustained by cells downstream of the most undifferentiated HSCs. Finally, YFP+ cells were transplanted into recipients in order to investigate the differences between *in vivo* and transplant haematopoiesis. They found very variable engraftment, with a mean of one in 33 engrafting. This helps to explain some results of oligoclonal haematopoiesis from transplantation experiments and strengthens the case for *in vivo* studies.

A similar study was performed by Sawai et al. (2016), where a different system was used to lineage-trace a high proportion (80-90%) of HSCs with the least differentiated phenotype and slowest proliferation rate. Longitudinal analysis allowed a precise timing of when the label appeared in different compartments and the kinetics of differentiation. Broadly, their results agree with those of Busch et al. and Sun et al. on the large numbers of stem cell clones contributing to haematopoiesis, although they find that the rate of transition from stem cell to mature blood cell types is faster, estimating that 3-8% of the labelled HSCs enter differentiation every day. This supports a more traditional model in which progenitors are less long-lived.

Finally, a polylox recombination system which marks cells uniquely by randomly excising and inverting ten cassettes was used by Pei et al. (2017) to mark embryonic cells in mice at the emergence of HSCs. 10 months after birth, barcodes were detected, showing that the adult HSC compartment is a mosaic of hundreds of embryonic clones, ranging in size from 0.2-3.8%, mostly with multilineage output. Since adult clones are nested within embryonic clones, the number of embryonic clones forms an extreme lower bound for the number of active stem cells.

In summary, all studies of *in vivo* unperturbed haematopoiesis in mice support a polyclonal model of haematopoiesis where large numbers of adult stem cells, likely in the thousands, contribute to haematopoiesis. It could be argued that they have not changed our understanding of haematopoiesis dramatically relative to transplantation studies, but they do at least demonstrate similar findings in a more physiological model. It should be noted that even in animal models there are no direct estimates of the number of active stem cells at steady state.

I.3.b. Estimates of human HSC numbers

65 million years of evolutionary divergence, coupled with the long lifespan and large size of humans indicate that – even if we had good estimates from mice – it may not be sufficient to extrapolate from mouse studies to estimate human stem cell numbers and dynamics. Mice are typically studied under pathogen-free conditions, the proportion of peripheral blood cells that are myeloid in humans is larger, and there are many known differences between HSCs from the two species, including immunophenotypic definition, cytokine requirements, and differences in the HSC niche (Doulatov et al, 2012, Abkowitz et al.1996, Catlin et al. 2011, Larochelle 1996). Below,

I discuss experiments that aim to estimate stem cell numbers in humans, first by transplantation into mice or other humans, and then in unperturbed haematopoiesis by necessarily indirect means.

I.3.b.i. Xenotransplantation of human cells into animals

Xenotransplantation studies face all the caveats of autologous and allogeneic transplantation, with the added complication of differences between the human niche and that of the recipient immunosuppressed animal. Furthermore, they have produced varying results, with the estimate of the number of stem cells varying from 250,000 (Wang et al. 1997) to ~7,000 (discussed in Bystrykh et al. 2012). It is unclear how to reconcile these experiments, although differences in the degree of immunocompromise of the animal, the transplantation regimen, and the amount of *ex vivo* manipulation of stem cells may all be important.

I.3.b.ii. Transplantation of human cells into humans

The advent of gene therapy over the last few decades has allowed analysis of unique gene insertion sites as a side-product of clinical trials. Although initial attempts were leukaemogenic through insertional mutagenesis (Hacein-Bey-Abina et al. 2003), more recently it seems that insertion of the genes acts as a neutral tag. In addition to the caveat that transplantation measures potential rather than fate, these patients are by definition unwell. Despite the fact that analysis of viral insertion sites is only semi-quantitative, counting the number of unique insertion sites detected years after transplantation provides a lower bound on the number of active stem cells in this non-physiological setting. At least hundreds (Cartier et al. 2009) or thousands (Biasco et al. 2016) of stem cells contributed to long term multilineage haematopoiesis in these patients.

I.3.b.iii. Indirect methods of studying native human haematopoiesis

Methods for studying native human haematopoiesis have been based on the detection of markers that vary naturally between different somatic human cells. Firstly, in women heterozygous for an X-linked marker gene, detection of the proportion of cells that have inactivated either X chromosome (X chromosome inactivation or XCI skewing) provides insights into population dynamics. Early work inferred from the small degree of XCI skewing in most women that the number of active stem cells must be at least 400, based on binomial statistics (Buescher et al. 1985). A more recent study (Catlin et al. 2011), used the proportion of women with XCI skewing as a proxy for the rate of clonal drift. The rate of clonal drift is determined jointly by the population size and the rate of symmetric cell divisions (see below); knowing any two of these three terms allows the third to be inferred, but one is insufficient. Catlin et al. assumed that humans had the same number of stem cells as had been inferred in mice and cats by limiting dilution transplantation experiments (discussed above) and used that to infer the rate of symmetrical stem cell divisions. Furthermore, this study predates the description of clonal haematopoiesis (General Introduction), which may be responsible for a proportion of the skewing observed, particularly in the older of the two cohorts used. Finally, the model of how XCI skewing occurred is unusual: it was assumed to be due to hemizygous selection of one X allele over the other, with the parameter for the strength of the selective advantage drawn from a distribution based on the strength of hemizygous selection in Safari cats (cats that are the offspring of the distantly-separated South American and Eurasian breeds). There is little evidence of hemizygous selection in humans beyond the observation of skewed XCI. This Catlin et al. article is often cited in relation to the number of human stem cells despite the fact that no attempt is made to estimate it there (for example, Welch et al. 2012, Laurenti and Gottgens 2018, Young et al. 2016).

Secondly, analysis of telomere length distributions over the life of patients (Werner et al. 2015) provides a window into tracking unperturbed haematopoiesis. Telomerase is expressed in HSCs, but at insufficient levels to stop telomere attrition completely over the course of life (Lansdorp 2008). Werner et al. compared simulations to observed telomere data to infer the proportion of symmetric to asymmetric stem cell divisions over life. The underlying principle is that, under a model of purely asymmetric stem cell divisions, the telomeres shorten in all stem cells at the same rate, resulting in a Poisson distribution of telomere lengths over the population of cells, where the

lambda of the distribution decreases linearly with age. In contrast, in a model with mostly symmetric stem cell divisions, some stem cells have been through many more cell divisions than others. More stem cells are held in reserve, without cycling, and the decay in telomere length is logarithmic. Analysis of telomere data from a range of patients of different ages showed that the decrease in telomere length in the first decade of life is logarithmic and becomes linear later, indicating many symmetrical stem cell divisions and an increase in population size in childhood which then plateaus in adolescence. This, incidentally, argues against a model of clonal succession in adulthood.

Taken as a whole, this large body of work across humans and animal models suggests that the more physiological the model, the greater the evidence for a large number of multilineage clones driving blood production at steady state. Nonetheless, no studies have estimated *ab initio* the number of stem cells in humans, due to the ethical contraints on inducing clonal markers in healthy, unperturbed humans.

I.4. HSC clone lineage biases

Over the past fifteen years or so a number of findings have challenged the textbook model of haematopoiesis, in which a homogeneous pool of stem cells – through a rigid series of progenitors with defined lineage restrictions – produces the right balance of all cell types. In particular, the introduction of assays of single cells, whether colony-forming assays (Doulatov et al. 2010), transcriptomic analyses (Paul et al. 2015), or single cell barcoding experiments (Perie et al. 2015) have suggested alternative patterns of progressive lineage restriction and generally pointed towards lineage priming occurring higher up the hierarchy than anticipated by the original model, which was derived from observations of bulk populations that shared a surface marker (Laurenti and Gottgens 2018). However, this work is not the focus of the present chapter: the discussion here is restricted to assays of the cell types that descend from *clones* of stem cells (rather than the precise series of progenitors that lie between an individual stem cell and its mature progeny) as that is of greater relevance to the results presented below.

I.4.a. Transplantation studies in animals

Transplantation experiments in mice have shown that cells that meet the criteria for stemness can be further stratified based on their surface markers into those with greater or lesser self-renewal potential (Kent et al. 2009), indicating intrinsically-determined HSC heterogeneity. Furthermore, immunophenotypic HSCs can be biased towards producing one mature cell type or another (reviewed in Copley et al. 2012, and in Schroeder 2010). Some HSCs have been found to be biased towards the myeloid or lymphoid lineage (Muller Sieburg et al. 2004, Dykstra et al. 2007, Challen et al. 2010, Muller-Sieburg et al. 2012), and some towards the megakaryocyte-erythrocyte lineage (Sanjuan-Pla et al. 2013). Interestingly, HSC biases are recapitulated upon their transplantation, suggesting that this is a somatically heritable phenotype, although it may additionally be modulated by extrinsic signals such as growth factors (Muller-Sieburg et al. 2004, Dykstra et al. 2007, Challen et al. 2010).

I.4.b. *In vivo* studies in animals

As previously discussed, transplantation experiments and colony assays determine a cell's potential rather than its fate had it remained in its host under physiological conditions. HSCs *in vivo* could therefore be more or less biased than is suggested by *ex vivo* experiments, including the experiments that built the classical roadmap of haematopoiesis in the 1980s and 1990s. As described in the section on estimating stem cell number above, it has recently become possible to induce unique and heritable tags in haematopoietic stem and progenitor cells such that their clonal output can be determined in homeostasis. In a study that used the same sleeping beauty transposon system as Sun et al. (2014), after eight weeks of chase many tags were shared across all lineages except for megakaryocyte precursors, suggesting most stem cells are ancestral to multiple cell types, with the exception of megakaryocytes (Rodriguez-Fraticelli et al. 2018). Conversely, a large number of tags were found only in megakaryocytes, indicating that in unperturbed haematopoiesis a reasonable proportion of stem cells make megakaryocytes and negligible numbers of other cell

types. Importantly, some of these megakaryocyte-restricted HSCs produced multilineage output when transplanted into a recipient, highlighting the suggestibility of these biased HSCs.

4.c. Observations and transplantation studies in humans

In humans, it has been observed that with increasing age a myeloid bias develops (Pang et al. 2011). Two explanations have been put forward: stem cells represent one congruent population without bias in youth, but a bias towards the myeloid lineage develops with increasing age; or the stem cell pool represents at least two groups, one of which has a myeloid bias and one a lymphoid bias, and over the course of life the myeloid-biased group gradually predominates over the other (Schroeder 2010). Which of these is the case remains to be determined. Gene therapy studies have provided the opportunity to track transplanted cells in humans using the site of vector integration as a unique clonal mark. In a study of four patients with Wiskott-Aldrich syndrome who were transplanted with autologous genetically-modified HSPCs, integration sites were frequently shared between myeloid and lymphoid cells (unfortunately, no separate analysis of B and T lymphocytes was reported) showing the existence of multipotent stem cell clones (Biasco et al. 2016). In contrast, in one B-thalassaemia patient who received a genetically modified HSPC transplant, a long-lived (33 months) B and T lymphocyte-deficient clone was observed, suggestive of a myeloid biased stem cell (Cavazzana-Calvo et al. 2010). Intriguingly, the lentiviral vector in this clone had integrated into and increased the expression of *HMGA2*, a known regulator of HSC self-renewal durability (Copley et al. 2013). Its behaviour may therefore reflect normal biology or be due to insertional mutagenesis.

In summary, there is compelling evidence for some degree of HSC lineage bias, although most stem cell clones in humans and unperturbed mice still seem to have multipotent output. Very little is known about biases in unperturbed human haematopoiesis. What bias there is in animal models seems to be both cell-intrinsic and flexible in the case of stress. One can conjecture that the ability of HSCs to modulate output based on environmental stressors might be advantageous in evolutionary terms, given that the haematopoietic system is faced with the challenge of massively amplifying the production of one cell type – such as granulocytes in acute infection or platelets in haemorrhage – without necessarily scaling up the production of other cell types.

I.5. Spontaneous somatic mutations as natural barcodes

To track the clonal dynamics of stem cells in a healthy unperturbed human requires a label that occurs naturally in cells at a high rate, persists over life, and can be reliably detected. Somatic mutations provide such a marker. Induced mutations were famously used by Till and McCulloch to study normal haematopoiesis, and spontaneous mutations have been used to track the clonal dynamics of tumours (Ding et al. 2010, Ding et al. 2012, Tsao et al. 1999). We build on this work to use spontaneous somatic mutations to study normal blood.

In blood, all available evidence – limited though it is – indicates that the mutation rate over life is relatively constant. Welch and colleagues cultured and whole exome sequenced three single HSPCs from each of seven healthy individuals spanning in age from birth to the eighth decade of life (Welch et al. 2012). Although the total number of mutations per exome is low, these data show a linear accumulation of mutations. Furthermore, in acute myeloid leukaemia the mutation burden increases linearly with the age of diagnosis (Welch et al. 2012, Alexandrov et al. 2015). More noise is present in analyses of blood cancers, most likely due to a combination of occasional aberrant mutational processes in some cancers which inflates the mutation burden and a variable time to the most recent common ancestor (MRCA) of the tumour (a longer time to the MRCA decreases the number of mutations called since many subclonal mutations will not be detected). Not only does the mutation rate seem constant, but the mutation burden is sufficient for us to detect large numbers of potential clonal markers per cell. Welch et al.'s normal HSPC exomes acquired mutations at a mean of 0.13 exonic mutations per year of life. Assuming that approximately 1% of mutations occur in coding regions, this translates to 13 mutations per genome per year, or 780 in a 60 year-old. Most AMLs have approximately 400-500 mutations per genome (Cancer Genome Atlas Research Network et al. 2013, Welch et al. 2012). Thus, by sequencing 100-200 normal blood genomes in a 60 year-old, one could hope to detect over 100,000 nested clonal barcodes. Such an approach has become feasible in recent years because of the falling costs of sequencing and reliable methods of culturing single HSPCs into large colonies amenable to whole genome sequencing.

I.6. Population genetics models for studying stem cells

In this study, we use methods from the field of population genetics to study the clonal dynamics of blood. We conceive of blood stem cells as a population of asexually-reproducing individuals. It is therefore necessary to introduce briefly the population genetics models that we apply: the neutral Moran model and the neutral Wright-Fisher model, and the associated coalescent process.

The population of stem cells can be thought of as behaving like the neutral Moran model of population genetics (Moran 1958, Gladstein 1978). In this model, in every generation two different individuals (or stem cells, in our case) are chosen from the population at random; one dies (or differentiates), and the other reproduces (a symmetrical doubling stem cell division), leaving two daughter individuals which carry the set of mutations carried by the parent individual. The waiting time from the birth of an individual until it next gives birth or dies, in the Moran model, is equivalent to the waiting time from the origin of a stem cell (in a symmetrical cell division) until the first time its stem cell descendant undergoes a new symmetrical cell division. We are blind to asymmetric cell divisions, as they do not change the genotype composition of the stem cell population.

Associated with the Moran model, which evolves forwards in time, is a genealogical process, where a sample of individuals from the population retrace their ancestry backwards through time. Here, in every generation back in time, a coalescent event may or may not occur among the ancestors of the sample. The occurrence of coalescent events is independent in each generation, and therefore the number of Moran model generations (going backwards) until a coalescent event has a geometric distribution, just as the number of coin tosses until a head is seen follows a geometric distribution. Kingman's coalescent is a stochastic process that generates genealogies, for any specified sample size, exactly as if the forward time process is a neutral Moran model (Kingman 1982). The timings of the coalescent events and the random pairing of lines of descent together determine the topology of the phylogenetic tree and the branch lengths.

When the population size N is large, N/2 generations of the Moran model closely approximates to one generation of the Wright-Fisher model (Fisher 1922, Wright 1931), which is more convenient to simulate forwards in time. In every generation of the Wright-Fisher model, the genotypes of individuals are obtained by sampling with replacement from the parent generation.

As with the Moran model, a reverse time genealogical process is associated with the forward time Wright-Fisher model, where every cell in one generation is assigned to a parent in the previous generation, independently of other cells; thus, a coalescence is generated when two cells happen to be assigned to the same parent.

For a sample of cells from a given population, the probability of two cells picking the same parent in the preceding generation is dependent on the population size. If we only have two cells in our sample, and a population size of N, the probability that both will pick the same common ancestor in the preceding generation is 1/N. When the population size is large relative to the sample size, either zero or one coalescent event will occur in a given generation; that more than two cells should happen to pick the same common ancestor out of a large pool becomes very unlikely, making the model more similar to the Moran model. As with the Moran model, the number of generations until a coalescent event follows a geometric distribution. The probability of two cells not coalescing in a Wright-Fisher generation is (1-(1/N)). Therefore, the probability of coalescing G generations ago is the probability of not coalescing for (G-1) generations, and then of coalescing in the G^{th} generation: $P(G) = (1-(1/N))^{G}*(1/N)$.

It follows that if the population size is 10 times larger, the probability of two cells picking the same parent is 10 times smaller, and 10 times more generations are required to achieve the same probability of a coalescent event. This means that the pattern of branch points in a phylogeny derived from a certain population size having gone through a certain number of generations will be the same as that from a population 10 times as large having gone through 10 times as many generations.

It can be shown mathematically that the mean waiting time from the birth of an individual until it next gives birth or dies is the same as the generation time of the Wright-Fisher model (as shown by Kevin Dawson in the technical appendix of Lee-Six et al. (2018)). This in turn is equivalent to the mean time between symmetric cell divisions along a given line-of-descent. Because of this equivalence, we can simulate stem cell population drift using the Wright-Fisher model, which is computationally more convenient because fewer generations are required and each is of a fixed chronological duration.

Results

R.1. Experimental design

Here follows a brief overview of the experimental design. In order to exploit the somatic mutations of normal stem cells as clonal markers, our experiment was designed in two phases: a 'mutation discovery' phase, in which somatic mutations were discovered by whole genome sequencing of colonies derived from single HSPCs from one individual; and a 'mutation quantification' phase, in which the mutations that we had discovered by whole genome sequencing were detected and accurately quantified in mature blood fractions from the same individual, telling us the output of the clones that we had discovered. These data were then analysed with methods adapted from population genetics and ecology, effectively treating each stem cells as an individual within a population. An outline of the experimental workflow is provided in figure 1.1a.

R.2. Mutation discovery phase: HSPC isolation, clonal amplification, and sequencing

A bone marrow aspirate was performed (by Brian Huntly) on one healthy 59 year-old male volunteer with normal blood counts and no past history of blood disorders. Peripheral blood was taken on the same day. We also obtained a buccal swab as a germline control. An individual of this age was picked because he would be sufficiently old to have acquired a workable number of mutations (assuming the mutation rate reported by Welch and colleagues (2012)) but was young and healthy enough to allow, potentially, decades of longitudinal follow-up. Studying a male was advantageous because he would have only one copy of the X chromosome such that the allele fraction of X chromosome mutations could be evaluated more accurately. On a diploid chromosome, even a perfectly clonal mutation can still fluctuate in allele fraction because of binomial sampling of mutant and wild type reads, complicating the estimation of the true allele fraction, whereas a homozygous clonal mutation should be present on every single sequencing read.

Cells obtained from the bone marrow and peripheral blood harvest were sorted by flow cytometry. This served the dual purpose of isolating the immunophenotypic cell populations of

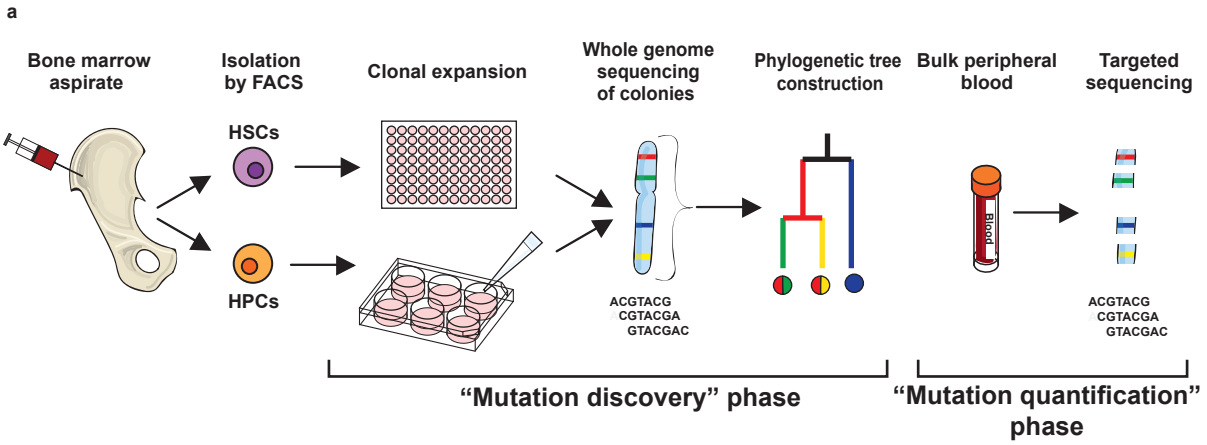
interest based on their cell surface markers and of obtaining single cells. Cells were cultured in methylcellulose or liquid culture for 2-6 weeks until colonies were derived that were sufficiently large to be whole genome sequenced. Overall, we harvested 285 stem cell-derived colonies and 390 progenitor-derived colonies, split across common myeloid progenitors (CMPs), granulocyte macrophage progenitors (GMPs), and megakaryocyte erythroid progenitors (MEPs). Progenitor cells that produced a colony with a different morphology to their immunophenotypic definition were not picked. All of the colony derivation work was performed by David Kent, Mairi Shepherd, Nina Friesgaard-Oebro, and Miriam Belmonte, with advice on culturing conditions from Elisa Laurenti.

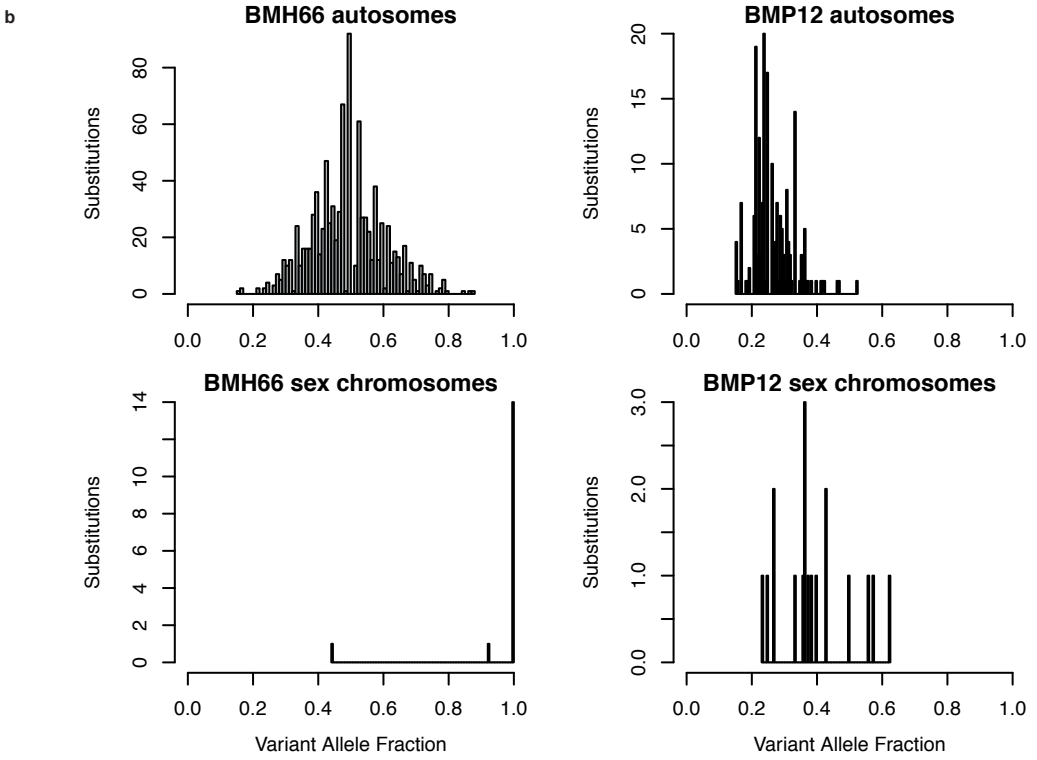
When we began this study, whole genome sequencing required large amounts of DNA (~100ng), and not all colonies could be expanded to such a large size. This imposed a selection bias on the colonies that we sequenced towards colonies from less differentiated cell types that have greater expansion potential. It is also possible that colonies with certain mutations may not amplify well under normal culturing conditions relative to their wild-type counterparts, as has been observed for cancer-derived organoids from a number of tissues (van de Wetering et al. 2015). We chose to whole genome sequence at ~15X the 100 stem cell-derived colonies with the largest amount of DNA, and the 98 progenitor-derived colonies with the largest amount of DNA.

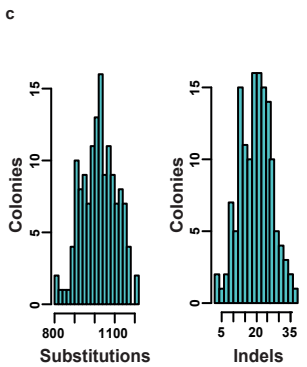
R.2.a. Quality control of colony whole genomes

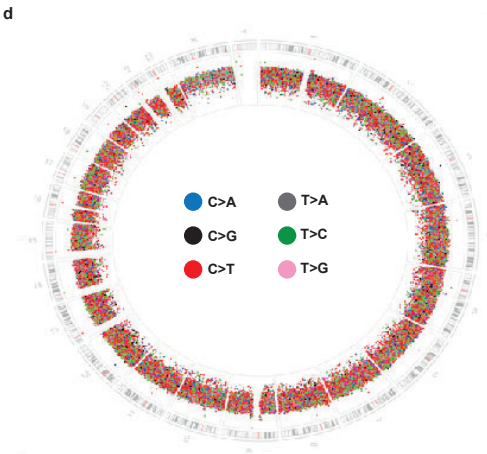
The clonal origin of each HSPC colony was determined by visual examination of histograms of their mutant allele fractions. In a colony that is truly single cell-derived, somatic mutations should only be present on one of two copies of an autosome, and so their allele fraction should be binomially distributed around 0.5. For the X chromosome of our male individual, allele fractions should be exactly 1. For 140 colonies, this was indeed the case. However, for the remaining 58, the allele fractions were lower, indicating that these colonies were derived from one or more cells (figure 1.1b). This is probably a result of colonies growing into one another in methylcellulose. Because some of the stem cells were grown in liquid culture, while all progenitors

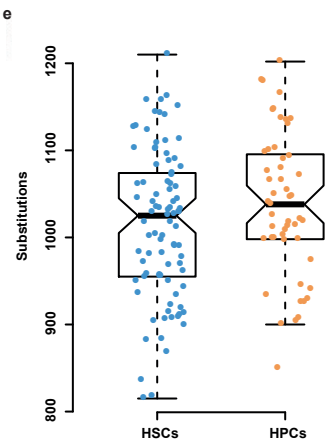
Figure 1.1. Experimental design and quality control. a, The experimental workflow. Following bone marrow aspiration, single haematopoietic stem cells (HSCs) and haematopoietic progenitor cells (HPCs) were isolated by fluorescence-activated cell sorting (FACS) and expanded in liquid culture or methylcellulose into colonies that could be whole genome sequenced. Somatic mutations discovered by sequencing were used to construct the phylogeny. This is the mutation discovery phase. In the mutation quantification phase, peripheral blood samples underwent deep targeted sequencing to quantify mutations that had been found in the discovery phase. **b**, histograms of the variant allele fraction (VAF) of mutations were used to identify whether colonies were derived from one cell or more. An example clonal sample (BMH66) and polyclonal sample (BMP12) are shown. In a clonal sample, the VAF of autosomal mutations should be distributed around 0.5, and that of sex chromosomes should be 1. For a subclonal sample it should be lower. Occasionally, even in a clonal sample lower VAFs are seen due to the failure to detect a mutation on a read, or a read from another locus being aberrantly mapped to the locus in question and lowering the apparent coverage, or a mutation acquired *in vitro*. **c**, histograms of the burden of substitutions and indels across all 140 clonal samples. **d**, the distribution of mutations from all 140 clonal samples around the genome, shown as a circos plot. The outermost ring of the circos plot depicts the karyotypic ideogram. Moving inwards, base substitutions are shown as rainfall plots where the height of the dot in the substitution ring is proportional to log10 of the distance to the next mutation and with the colour of the dot illustrating the base change, as shown in the key. (c) a comparison of the substitution burden between stem cells and progenitor cells. There were not significantly more mutations in progenitors than stem cells (p=0.14, Wilcoxon Rank Sum test).


a
Bone marrow aspirate
Isolation by FACS
HSCs
HPCs
Clonal expansion
Whole genome sequencing of colonies
ACGTACG
CGTACGA
GTACGAC
Phylogenetic tree construction
Bulk peripheral blood
Blood
Targeted sequencing
ACGTACG
CGTACGA
GTACGAC
"Mutation discovery" phase
"Mutation quantification" phase


b
BMH66 autosomes
BMP12 autosomes
BMH66 sex chromosomes
BMP12 sex chromosomes
Substitutions
Variant Allele Fraction


c
Colonies
Substitutions
Indels


d
C>A
C>G
C>T
T>A
T>C
T>G


e
Substitutions
HSCs
HPCs

were grown in methylcellulose, the stem cells were less affected than the progenitors. Furthermore, MEPs grow as more compact balls than CMPs and GMPs, and so were relatively spared. For these reasons, after removal of 58 polyclonal colonies we were left with 63 bone marrow-derived HSCs, 16 peripheral blood-derived HSCs, 38 MEPs, five CMPs, and eight GMPs.

We could also use allele fractions to assess what proportion of mutations might have occurred *in vitro* during the process of colony derivation. Unless a selective sweep occurred *in vitro* (which seems unlikely), only mutations that were present in the sorted founder cell will be fully clonal in a colony. The proportion of mutations that are subclonal within a colony with a VAF peak at 0.5 informs on the mutation rate. Because of binomial sampling of mutant and wild type reads combined with the low sequencing depth, it can be difficult to tell if a mutation on an autosome is truly subclonal. This is further complicated by sequencing false negative artefacts, and more frequently mismapping of reads to this location in the genome and appearing as a wild type read. Mismapping events are relatively frequent with Illumina HiSeqX-sequenced and BWA-mem-mapped data. The bottom left-hand panel of figure 1.1d demonstrates two X chromosome mutations with a VAF below 1. Visual inspection of these mutations shows that one is present on 4 out of 8 reads. This mutation is in a relatively low complexity part of the genome and could represent either a sequencing artefact or a mutation acquired *in vitro*. The other mutation is present on 12 out of 13 reads. The read that does not support the mutation maps uniquely to that position in the genome, and so probably represents a sequencing error. The variant, however, is likely to be a true clonal mutation. Thus, even true clonal mutations on the X chromosome may not have a VAF of 1, but the mutation is likely still to be present on a large proportion of the reads. Variants acquired *in vitro*, however, should be present in half of the cells in a clone or fewer. Therefore, a more informative metric of the proportion of variants that might have occurred *in vitro* is the proportion of X chromosome variants at variant allele fraction of 0.5 or less (assuming equal amplification of both daughters of the first cell division). Considering only positions with depth 10 or above, at which the allele fraction can be estimated accurately, and considering only colonies with more than five high coverage X chromosome mutations called, the mean percentage of mutations per colony at a variant allele fraction of <0.5 is 5.6%.

An additional source of evidence that *in vitro* mutations do not contribute substantially to our mutation catalogue is that two colonies share 1,115 substitutions and have only 51 and 87

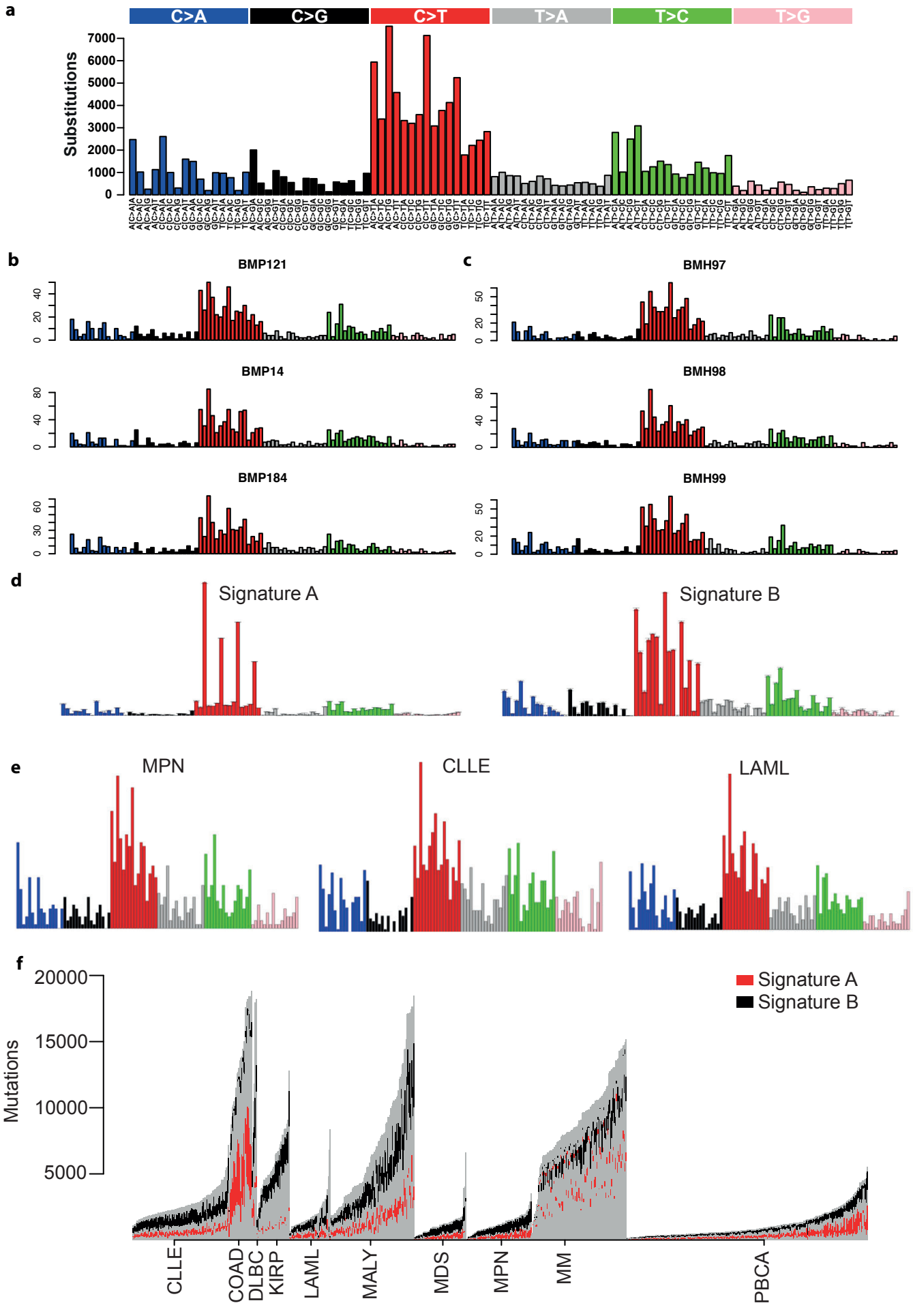
private mutations respectively. *In vitro* mutations will typically be private, so fewer than 7% (87/(87+1115)) of the total mutations called in these colonies could have occurred *in vitro*.

R.2.b. Genomics of normal blood cells

A mean of 1,023 (range 815-1,210) single base substitutions and 20 small insertions and deletions (indels) (range 2-37) were observed across the 140 colonies (figure 1.1c). No copy number changes or structural variants were detected. Progenitor cells had slightly more substitutions than stem cells (mean of 1,038 for progenitors and 1,024 for stem cells), but the difference was not statistically significant ($p=0.14$ by Wilcoxon rank sum test) (figure 1.1e). This result does not necessarily mean that progenitors do not have an increased mutation rate over stem cells, as the line-of-descent to any given progenitor will have spent nearly all its time as an HSC. Imagine that a progenitor that we sampled exited the stem cell compartment one month before sampling, and over that last month acquired mutations at a 10-fold greater rate than in stem cells. In this scenario, the progenitor would only have 10-20 more mutations than its stem cell cousins.

Comparison of the mutation burden that we detect in normal HSPCs and that reported in acute myeloid leukaemia (AML) supports at the level of the whole genome the findings by Welch et al. that AML exomes do not have more mutations than normal HSPC exomes. Indeed, many AML genomes have a lower mutation burden than that reported here: the mean number of mutations per AML genome in a cohort with a mean age of 55 was just over 400 (Cancer Genome Atlas Research Network et al. 2013). Some of this difference may be due to different sequencing platforms and mutation calling algorithms. In addition, in some patients, contamination of the matched normal tissue with leukaemic blasts (such that somatic mutations may appear to be germline) could result in a significant underestimation of the mutation burden. Another – and biologically interesting – component is that the most recent common ancestor of the leukaemia may have existed a long time ago; if a single leukaemic cell were sequenced, it might have the same mutation burden as a single normal cell, but bulk sequencing misses a large proportion of the subclonal mutations in the tumour. Depending on the strength of the effect of the first three factors, comparison with our data would indicate that the MRCA of many AMLs occurred years to decades before diagnosis.

Figure 1.2. Trinucleotide context of mutations in normal blood. a, the counts of substitutions in each of 96 trinucleotide contexts from all 140 clonal samples combined. **b**, **c,** the same figure for representative individual progenitor colonies (**b**) and stem cell colonies (**c**). **d,** Non-negative matrix factorisation of the blood trinucleotide context along with a panel of blood cancers deconvoluted the substitutions in the our samples into two signatures. Signature A accounted for 30% of mutations and bears a strong resemblance to signature 1 (see text). Signature B accounted for 70% of mutations and was novel. **e**, trinucleotide context of mutations from three blood cancers, chosen for their similarity to the trinucleotide context of normal blood colonies. **f**, results from mutational signature decomposition of a panel of cancers. The mutations burden attributable to each signature from each sample of a given cancer type is shown. Signatures that are not signature A or B are all coloured grey. CLLE, chronic lymphocytic leukaemia; COAD, colorectal adenocarcinoma; DLBC, diffuse large B cell lymphoma; KIRP, kidney renal papillary cell carcinoma; LAML, acute myeloid leukaemia; MALY, malignant lymphoma; MDS, myelodysplastic syndrome; MPN, myeloproliferative neoplasms; MM, multiple myeloma; PBCA, paediatric brain cancer.


a
C>A
C>G
C>T
T>A
T>C
T>G
Substitutions
b
BMP121
BMP14
BMP184
c
BMH97
BMH98
BMH99
d
Signature A
Signature B
e
MPN
CLLE
LAML
f
Mutations
Signature A
Signature B
CLLE
COAD
DLBC
KIRP
LAML
MALY
MDS
MPN
MM
PBCA

The somatic mutations in blood stem cell genomes that we observed were mostly C>T and T>C transitions, particularly in a context of N<u>C</u>T (the mutated base is underlined) (figure 1.2a). The trinucleotide spectrum of mutations in all of our samples was similar (figure 1.2b and 1.2c). A similar spectrum has been observed in a study of clonal haematopoiesis (Zink et al. 2017). Visual inspection of all blood cancer genomes included in the Pancancer Analysis of Whole Genomes study revealed that it was present in some cancers (figure 1.2e). A mutational signature (General Introduction) extraction was run with the combination of our samples and all of these cancers. The trinucleotide context of our sample was deconvoluted and found to be ~30% a signature that similar to signature 1, a known signature present in all tissues that is discussed in Results Chapter 2, and ~70% a novel mutational process (signature B in figure 1.2d**)**. Quantification of the contribution of this mutational process to all blood cancer genomes analysed indicates that it is present at some level in most blood cancers. It is responsible for a large proportion of mutations in myeloproliferative neoplasms, myelodysplastic syndrome, and AML, which have relatively quiet genomes and, as discussed above, for the most part have not departed significantly from the somatic mutational landscape of normal blood (figure 1.2f). In more mutated cancers, such as certain lymphomas, the signature is obscured by a vast excess of additional process, which may explain why it has not been reported previously. The signature contributed proportionately less to solid tumour genomes included as a negative control, although it was detected in kidney cancers and paediatric brain cancers. The cause of this signature, as with many mutational signatures, remains unknown.

The mutations were scattered across the genome (figure 1.1d), with <1% causing non-synonymous changes in protein-coding genes. We searched our 140 genomes for driver mutations, (which would affect our interpretation of clonal dynamics in this experiment) in two ways: first, by manual annotation of known cancer genes, and second, using a dNdS approach which detects the enrichment of non-synonymous mutations in a gene above what would be expected by chance, indicating positive selection (Martincorena et al. 2017). The dNdS approach therefore does not rely on prior knowledge of which mutations are drivers. dNdS also allows us to estimate a global measure of selection across all the mutations in the dataset. The global dNdS across all 140 blood genomes is 1.0010 (95% CI 0.889-1.127), with values greater than one indicating positive selection and values less than one negative selection. No genes were found to be under positive selection by dNdS, and no known driver mutations were found by manual curation of known drivers of myeloid

neoplasms. Finally, in the second phase of the experiment in which we performed targeted sequencing, we included 100 common hotspot mutations frequently observed in clonal haematopoiesis and myeloid malignancies, and no drivers were detected despite deep sequencing. We cannot exclude the presence of as-yet undiscovered drivers that are rare in our dataset, but all the evidence that we have gathered indicates that this individual's blood is evolving neutrally.

R.2.c. Construction of a phylogeny of HSPCs

A phylogeny of HSPCs was constructed based on the sharing of somatic mutations across the 140 colonies. Our confidence in each of the branch-points was quantified by Felsenstein's non-parametric bootstrapping method (Felsenstein 1989), with 1,000 bootstrapping replicates. The tree was constructed independently based on indels and on short tandem repeats and using all types of data combined, using bootstrapping in all cases. We applied all three of a maximum likelihood, maximum parsimony, and neighbour joining methods (although not all on every dataset, as some are not appropriate for certain data types (Methods)). In general, the structure of the tree was supported with high confidence: few branch-points that were supported with high confidence using different methods disagree with the tree that we present (figure 1.4) (Methods). Tree-building was performed with substantial help from Sebastian Grossman (distribution of the work-load described in Methods).

The phylogeny that was reconstructed had a deep branching structure, with most mutations private to a given cell: of 129,582 substitutions on the tree, 8,676 were shared amongst different colonies (figure 1.3a). The construction of a phylogeny allowed us to perform multiple analyses: we could time mutations relative to one another (mutations on a branch shared by two colonies necessarily occurred before those on a branch private to one of those colonies); determine the relationship of all cells to one another; and apply phylodynamic models that were developed in the fields of population genetics and epidemiology to our data. These are explored below.

Figure 1.3. Inferences from the phylogeny. a. phylogeny of 140 single haematopoietic stem and progenitor cells showing the relationship between cell types. At each tip of the tree is a colony. Branches connect colonies to each other to form a family tree. Branch lengths are proportional to the number of somatic mutations: thus, a branch that is ancestral to two colonies and 100 substitutions long reflects the fact that these two colonies share 100 substitutions that are not present in any of the others. Symbols at the tips of the branches represent the phenotype of the cell. **b**, the same phylogeny as in **a**, but showing only the first 10 mutations of molecular time. **c**, the number of descendants of each node for the first 10 mutations of molecular time, used to estimate the embryonic mutation rate (Methods). **d**, phylodynamic inference of the population size trajectory of the stem cell pool reveals changes in the effective population size of stem cells over life based on the timing of coalescences (branch-points) in our observed phylogeny. Shading illustrates different credibility intervals. The y axis is shown in units of 'population size multiplied by generation time' (cell-years) because the same distribution of coalescences can be generated from a population of 10 times the size with 10 times as many generations.

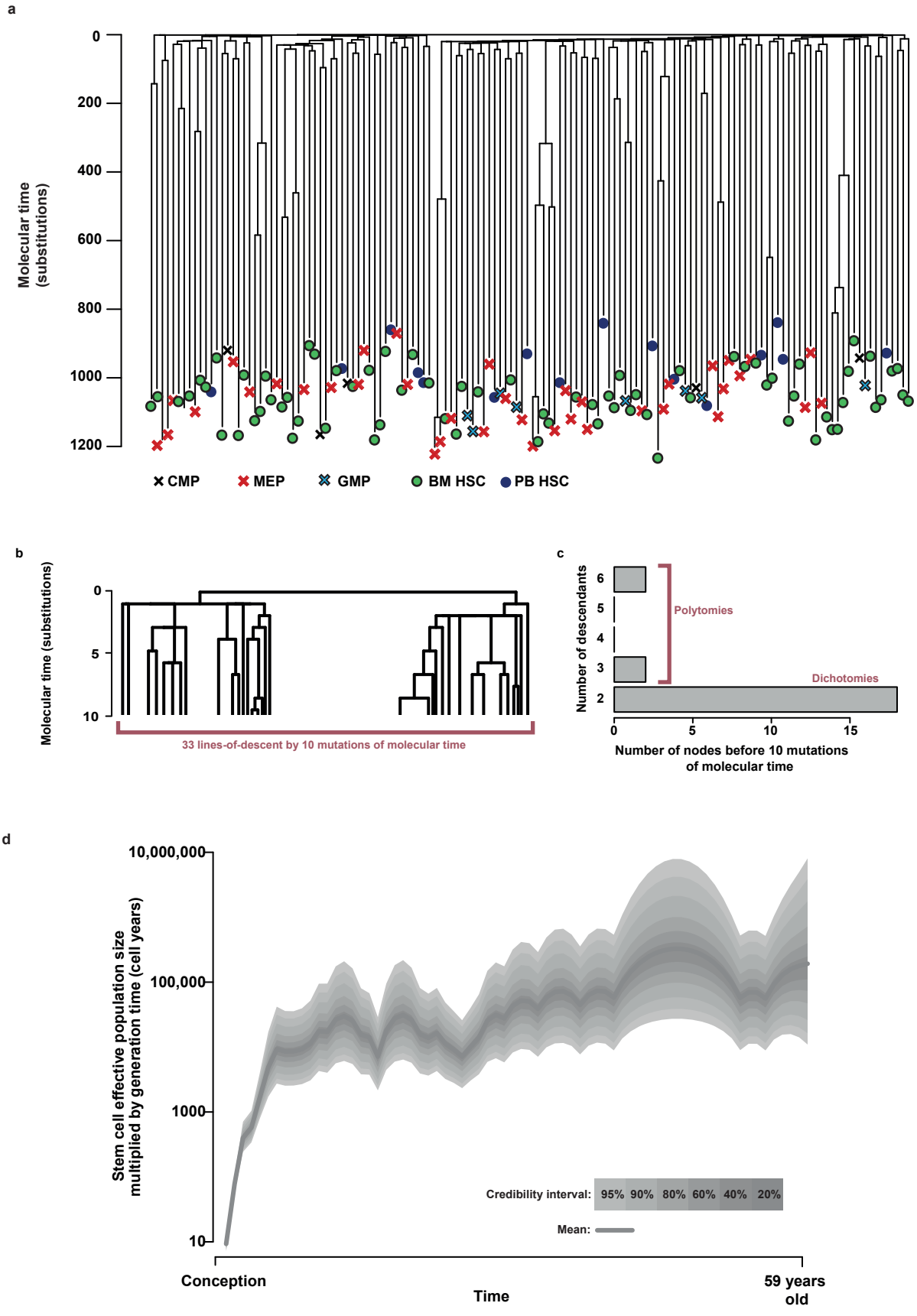
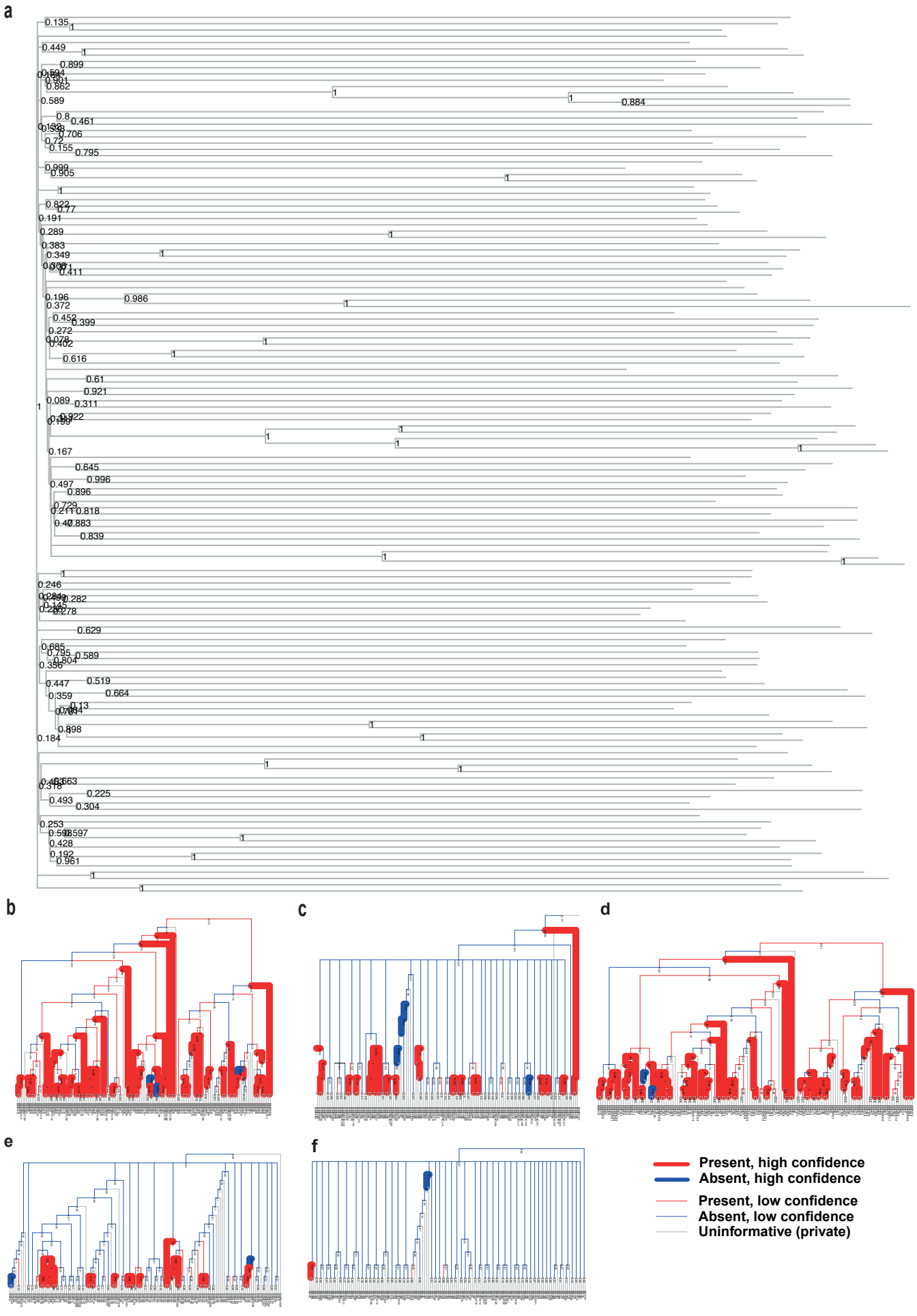

a
Molecular time (substitutions)
0
200
400
600
800
1000
1200
CMP
MEP
GMP
BM HSC
PB HSC
b
Molecular time (substitutions)
0
5
10
33 lines-of-descent by 10 mutations of molecular time
c
Number of descendants
6
5
4
3
2
Polytomies
Dichotomies
0
5
10
15
Number of nodes before 10 mutations of molecular time
d
Stem cell effective population size multiplied by generation time (cell years)
10,000,000
100,000
1000
10
Credibility interval:
95%
90%
80%
60%
40%
20%
Mean:
Conception
Time
59 years old

Figure 1.4. Construction of the phylogeny using different methods. a, the phylogeny of cells, presented with p values next to every node, derived by bootstrapping the substitution matrix 1000 times, building a tree using SCITE for each replicate, and counting the proportion of the bootstrapped trees that support each node. **b-f,** phylogenies constructed using different datasets and methods. In each case the phylogeny was constructed using 100 bootstraps of the data, and the p value for each node shown underneath it. Branches are coloured by whether a branch ancestral to exactly the same descendants is also present in the SCITE tree, and are drawn with a thicker line if the branch is recovered in >=70% of bootstrap replicates. **b**, substitution and indel datasets combined, building the tree by maximum parsimony. **c,** substitution, indel, and neighbour joining datasets combined, building the tree by neighbour joining. **d,** substitutions, tree build by maximum parsimony. **e,** indels, tree built by maximum parsimony. **f**, Short tandem repeats, tree built by neighbour joining.


a
b
c
d
e
f
Present, high confidence
Absent, high confidence
Present, low confidence
Absent, low confidence
Uninformative (private)

R.2.d. Embryological insights

R.2.d.i. The ontogeny of blood

As one looks up the phylogeny, one is looking back in time into the somatic history of the haematopoietic system. The probability of capturing the most recent common ancestor of a population with a sample of size n for a randomly bifurcating tree is equal to (n-1)/(n+1); in our case, assuming that ours is a random sample of blood, 197/199=0.99. The earliest branch-point on our phylogeny therefore in all likelihood represents the cell division of the most recent common ancestor of blood. It partitions the tree into two uneven groups, one with one third of the cells (52/140) referred to as 'clade A' and the other with two thirds (88/140) referred to as 'clade B'. All cells in clade A all share one mutation ('mutation A') which is absent from clade B, and, conversely, those in clade B share a mutation which is absent from clade A ('mutation B'). We searched for these early embryonic somatic mutations in whole genome sequencing of the buccal swab sample. Strikingly, mutation A was found in the buccal epithelium at a clonal fraction of 1/3 (on 6 out of 30 reads) and mutation B at a clonal fraction of 2/3 (on 11 out of 33 reads). Mutation B was sequenced more deeply using a bait-set (discussed below), and found to occur on 124 out of 414 read (0.30%). It is possible that the buccal swab has some blood contamination, but it seems very unlikely that it should be made up exclusively of blood with no epithelial contribution at all. These data therefore indicate that the most recent common ancestor of blood is shared with the most recent common ancestor of buccal epithelium. Thus, at no point in embryogenesis is there a cell that is fated to produce all of blood and only blood. Furthermore, since blood is mesoderm-derived and buccal epithelium ectoderm-derived (Rothova et al. 2012), this common ancestor must precede gastrulation. Indeed, it seems likely that it existed very early, perhaps as one of the first cells in the embryo.

R.2.d.ii. Early embryonic mutation rates

Beyond this first branch-point, we observed a number of short branches, supported by 0-4 mutations (figure 1.3b). We were able to use the structure of the phylogeny to estimate the mutation

rate for the first 10 mutations of molecular time, presumably reflecting the first days or weeks of life. Firstly, the rate of branching in our tree informs on the upper bound of mutation rate per cell division. By 10 mutations of molecular time, we observe 33 lines-of-descent in our tree. Therefore, at least 33 cells must have existed in the embryo by this stage of molecular time. With doubling divisions, in the absence of cell death, it would take at least five generations to grow to this number of cells ($2^5 = 32$). Under this simplified scheme, the mutation rate is likely to be under 2 mutations per generation (10 mutations divided by 5 generations). Two principal caveats complicate this model: first, cell loss (either death or contribution to extraembryonic tissues exclusively, such that no descendants contribute to the adult); and, second, we may not have sampled every embryonic clone. Both cell loss and incomplete sampling of embryonic clones would mean that more generations need to have elapsed for us to obtain 33 lines-of-descent in our tree. Therefore, both would lower our estimate of the mutation rate per cell division. Thus, we can use an upper bound of 2 mutations per cell division.

The number of polytomies (or multifurcations) in this early embryonic tree also informs on the mutation rate per cell division. Polytomies in the tree can only occur if no mutations were acquired in a cell division. The lower the mutation rate, the more cell divisions are associated with no mutations and the more polytomies appear in the tree. Cell loss can make a true polytomy appear as a dichotomy, but not the other way around, and so increasing cell loss decreases the number of polytomies seen. The number of polytomies, therefore, also provides an upper bound: a combination of increased cell loss and decreased mutation rate would result in the same number of polytomies.

In the absence of cell loss, the following calculations allow us to estimate the mutation rate per cell division. In the first 10 mutations of molecular time of our phylogeny we observe 18 dichotomies, two trichotomies, and two hexachotomies (figure 1.3b,c). In the absence of cell loss, a trichotomy means that two cell divisions occurred, creating four branches, and one out of the four was associated with no mutations. Similarly, a hexachotomy means that five cell divisions occurred, creating 10 branches, and four of them were associated with no mutations. In total, the 32 cell divisions required to make 33 lines-of-descent created 64 branches, 10 of which were associated with no mutations, one for each of two trichotomies and four for each of two hexachotomies. From the Poisson distribution, the maximum likelihood of 10 branches with no mutations is 1.86 mutations per division (-log(10.64)), with a 99% confidence interval of 1.19 to

2.80. We can therefore conclude that in the absence of cell loss, mutation rate per cell division in the early embryo is likely to fall between 1.2 and 2 mutations per cell division, a rate that overlaps with estimates from human neural progenitor clones (Bae et al. 2018) and *de novo* germline mutations (Rahbari et al. 2016). With more cell loss, the mutation rate could be lower.

R.2.e. Phylogenetic relationships between different cell types

The mapping of phenotypes (as defined by surface markers and colony morphology) onto our phylogeny provides us with the opportunity to study clonal relationships. A statistical test for clustering on the phylogeny (Analysis of Molecular Variance, or AMOVA (Methods)) allows us to perform three comparisons: bone marrow-derived *versus* peripheral blood-derived stem cells; stem cells *versus* progenitors; and different kinds of progenitors *versus* one another.

First, no clustering of bone marrow-derived stem cells was observed relative to their circulating counterparts ($p=0.14$). This shows that there is no detectable geographic clustering of related cells in the marrow and, as far as we can tell, our sample from iliac crest is representative of the bone marrow throughout the body of this individual. This is consistent with two explanations: 1) a large polyclonal group of stem cells seeds the iliac crest marrow (and probably other marrow locations) around the time of birth, and no local clonal sweep occurs thereafter; or 2) stem cells must regularly recirculate and redistribute sufficiently often that the population in the iliac crest is a reasonably random sample of the whole stem cell pool. If neither of these scenarios were occurring, we would expect to see marrow-specific clades of the phylogenetic tree.

Second, the degree of clustering of progenitors relative to stem cells informs on the proportion of stem cells contributing to blood. If only a small proportion of stem cells were contributing to blood, and we had sequenced the entire haematopoietic phylogeny, we would see clades of progenitors clustering around their parent stem cells, and large numbers of stem cells with no nearby progenitor. Even if the parent stem cell were not sequenced, clustering of its progenitor daughters would be observed. In contrast, if all stem cells were contributing equally, stem and progenitor cells should be distributed around the tree. We were not able to detect clustering of stem cells relative to progenitors ($p=0.12$), indicating that more than a few clones are generating progenitors. The test that we use to show that progenitors are not clustered relative to

stem cells does not have an intuitive statistic, and it is hard to encapsulate the relationships on the tree with a few numbers. We can say, for example, that of the 51 progenitors, 37 are more closely related to another progenitor than to a stem cell. This or a more extreme value occurs 8% of the time in a randomisation test. However, this does not tell us when the common ancestor of the pair of cells existed, which is important for establishing whether currently active progenitors derive from a small number of stem cell ancestors. Perhaps a more useful statement is that the 51 progenitor colonies distribute across 47 different lines-of-descent extant at 100 mutations of molecular time, compared to 89 stem cells across 76 lines-of-descent. Thus, the progenitors we sequenced are not drawn from a substantially more restricted set of historic lines-of-descent than the stem cells.

Third, we were not able to demonstrate any clustering of different types of stem or progenitor cells on our tree. For example, MEPs were no more likely to be closely related to one another than they were to GMPs or CMPs ($p=0.10$), although the small numbers of GMPs and CMPs reduce the power of this analysis. These data are consistent with either a model whereby stem cell clones provide multilineage replenishment of progenitor cells with individual commitments, or a model with unilineage priming but such a large pool of ancestors that the cells of the same type that we have sampled were unlikely to descend from the same clone.

All of the above analyses have limited power. It is possible that by sequencing more stem and progenitor cells we would have been able to detect clustering of these various cell types. It is difficult to quantify the effect size that we would have been powered to detect, as this will depend on the size of the stem cell pool and the clonal dynamics of the tissue, both of which we are exploring here largely for the first time. These results should be revised by future studies.

Nevertheless, the observation that stem and progenitor cells are interspersed on the tree is useful to our analyses. Taken in conjunction with the relative paucity of recent branch-points, it implies that the phylogeny is dominated by events that occurred in stem cells. Progenitors can be thought of as random samples of stem cell clones and can therefore be treated in the same way as stem cells. Furthermore, since progenitors have short life-spans, if one were to retrace their life history they would rapidly coalesce with a stem cell. Branch-points that occur more than a few tens of mutations up the tree, and so more than a few years ago, necessarily represent symmetrical cell divisions where one stem cell has divided to give rise to two daughter stem cells. We know that both daughters are stem cells since we have sampled their progeny years to decades later. It

does not matter that some of the branch-points that we observe in our phylogeny we might have discovered through sequencing two progenitors, since our data indicate that we are sampling the descendants of two different stem cells. In analogy to this, in studies of mitochondrial inheritance, male samples can be treated as proxies for their mothers (Griffiths and Tavaré 1994). Therefore, for the analyses that follow, which rely upon the pattern of branch-points in our phylogeny, we can treat progenitors as stem cells, and, indeed, hereafter we refer to them as such.

R.2.f. Population size trajectory over life

While most branch-points are observed at the top of the phylogeny, a sizeable number also occur later. As explained above, these represent symmetrical cell divisions of adult stem cells. The pattern of branch-points throughout the tree is remarkably informative on how the size of the population of stem cells has changed over life. This has been extensively studied in population genetics and epidemiology, where, for example, the seasonal rises and falls of influenza virus populations can be reconstructed through an analysis of the branching pattern of the phylogeny (Lan et al. 2015; Karcher et al. 2017). These approaches are based upon Kingman's coalescent (Kingman 1982), which is discussed briefly above. Essentially, under a constant population size, the time intervals between coalescent events on a phylogeny should follow independent geometric distributions (with means determined by the number of lines of descent which traverse each inter-coalescent interval). Deviations from this statistical pattern of coalescences indicate changes in the size of the population. By comparing the pattern of coalescences before and after each time point, fluctuations in the population size can inferred.

The population size trajectory was computed using an established Bayesian method (Lan et al. 2015, Karcher et al. 2017) (figure 1.3d). We observe almost logarithmic growth early in life, for the first 100 mutations of molecular time, as one might expect given that the population must expand from one cell in the early embryo to an adult stem cell pool size. At this point, however, the population size stabilises. Assuming a constant mutation rate over life, 100 mutations of molecular time represents approximately 6 years of age. We cannot be quite this precise with great certainty, since the mutation rate may be higher in the embryo and perhaps in early childhood. If this were the case, the inflexion point would occur at a younger age. It would be possible to draw

a straight horizontal line through the 95% confidence interval on the population size trajectory from mutation 100 right until the end of life, such that we cannot reject the null hypothesis of a constant population size in adulthood. In humans, the number of immunophenotypic HSCs increases in life (Pang et al. 2011). Here, we are assaying the effective population size – that is to say, stem cells that are still able to reproduce – and so our results are a readout of stem cell *function*. It should be noted too that our individual is only 59 years old, whereas most studies of stem cell ageing look at older people.

The stability of the stem cell pool indicates a degree of homeostasis: the symmetric stem cell divisions – where one stem cell begets two others – that we observe as branch-points in our tree must be counterbalanced by death and/or differentiation of a number of cells equal to that created. It is unknown at which level this control is exercised. It could be provided by a niche of limited size beyond which stem cells are forced to die/differentiate (Zhang et al. 2003), by stochastic fate decisions that are balanced at the population level, as has been observed in epithelium (Alcolea et al. 2014), or even through a deterministic model.

The absolute size of the stem cell pool cannot be estimated through this approach without knowing the generation time of the population of stem cells, which is equivalent to the average time between symmetrical stem cell divisions. As explained in the introduction to this chapter, the same expected pattern of coalescences in a phylogeny is generated by a population that is 10 times as large as another going through 10 times as many generations: they are confounded. Very little is known about the generation time of stem cells in humans, and so we could not simply use an estimate from the literature. Instead, the next phase of our experiment allowed us to learn both at the same time.

R.3. Mutation quantification phase: estimating the number of human stem cells

R.3.a. Targeted sequencing of peripheral blood

In this phase of the experiment, we aimed to quantify the frequency in peripheral blood of the clonal markers that we had discovered by whole genome sequencing. All 129,582 mutations assigned to branches of the tree were potential clonal markers. To design a bait-set for and perform

high depth-targeted sequencing of all of these mutations would have been extremely costly. We therefore designed a bait to as many of the 8,676 mutations present in more than one colony as possible and aimed to target approximately 10 mutations per private branch of the tree. When selecting private mutations, we picked sites in the genome with the lowest sequencing error rate possible, based on a panel of ~1000 genomes (Methods). Not all mutations are suitable for bait design (because, for example, they fall in a repetitive region of the genome), resulting in a bait-set of 7,116 mutations, of which 6,317 were shared and 799 were private.

With this bait-set, we performed targeted sequencing of granulocytes taken at 4-month (mean coverage 776X), 9-month (mean coverage 4,669X), and 14-month (mean coverage 268X) time-points after the bone marrow aspirate. Different samples were sequenced with different levels of coverage as a range-finding exercise. In addition, a negative control of two cord bloods from unrelated individuals (kindly provided by Grace Collord) was sequenced (mean coverage 5,305X). An algorithm for counting mutant and wild type reads with stringent error correction, written and run by Robert Osborne, was used to generate mutation counts over every bait. Further error correction was performed statistically, using a Markov chain Monte Carlo generalised linear mixed model that used the site-specific error rate from the control DNA to estimate the true VAF in test samples, written by Peter Campbell (with minor amendments by me). Reassuringly, mutations on higher branches were at higher VAF than mutations on lower branches. 96.5% of mutations had no mutations on lower branches with a higher allele fraction. The VAFs of mutations were stable across the three time-points (figure 1.7). This is reassuring from a quality-control point of view, but also of biological interest, demonstrating clonal stability (at this level of the phylogeny, at least) over a 10-month period.

Most mutations further down the phylogeny were not detectable (figure 1.5). In contrast, mutations on the majority of embryonic and early childhood branches were detectable, as indicated by multiple black or red branches at the top of the phylogeny in figure 1.5. Every clade with detectable mutations represents a clone whose descendants are making blood nowadays. The picture we observe, therefore, is one of polyclonal haematopoiesis, with contributions from stem cells dispersed on the phylogeny of blood. A total of 47 non-nested clones have mutations that are detectable in the 9-month granulocytes (not all of these can be seen in figure 1.5, as a VAF cut-off is used to allow comparisons between different timepoints, but they can be seen in figure 1.8 where no cut-off is used). As with the Rodewald group's embryonic clonal marking strategy (Pei et al.

2017, section I.3.a.ii. of this chapter), the number of early clones that we observe actively contributing represents an extreme lower bound on the total number of actively contributing stem cells. Furthermore, if we make the assumption that stem cells make equal numbers of granulocytes, the distribution of VAFs itself suggests that thousands of stem cells are ancestral to today's granulocytes. If only 500 stem cells made blood, we should hardly ever see mutations at a VAF much below one in 1,000. Over some sites, the coverage is sufficiently high and the error rate in controls is sufficiently low that we can detect mutations at VAFs of the order of 1 in 3,000-4,000. We observe VAFs right down to our detection limit, suggesting that at least 2,000 stem cells are contributing to the granulocyte pool.

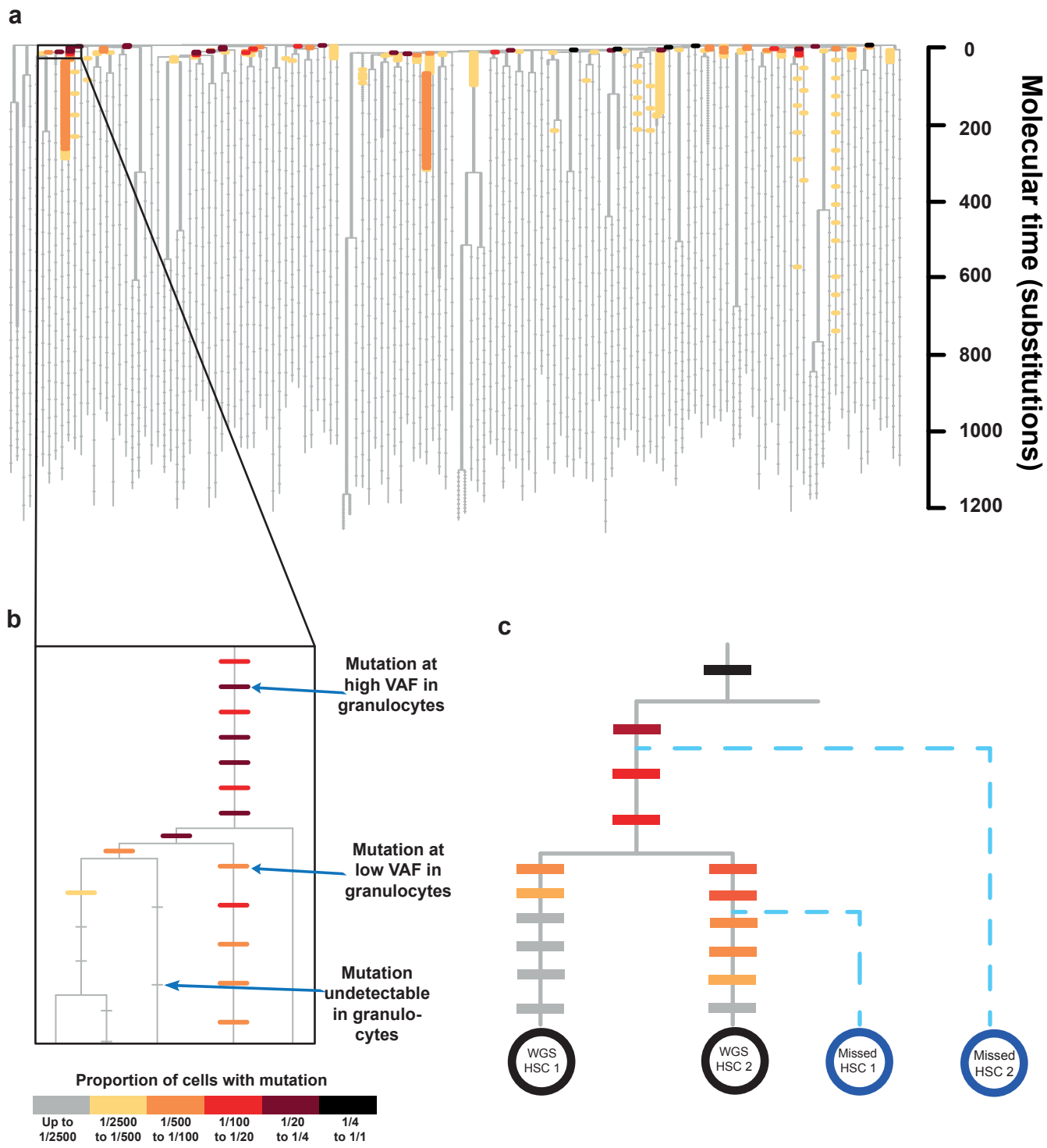
R.3.b. Capture-recapture approach for estimating stem cell number

We hypothesised that the actual number of stem cells might be estimated through a capture-recapture (also known as 'mark and recapture') approach that is typically used in ecology. The total number of animals in a population, N, can be estimated by marking n animals on one visit and recapturing K animals on a second visit. The number of recaptured animals that have been marked, k, tells us the size of the population through the formula: N=(K*n)/k. In our study, 'stem cells' can be substituted for 'animals', and 'mutations' for 'markers': by taking multiple blood samples from one time-point, lysing and sequencing them separately, and comparing the overlap between the mutant alleles found in one sample and those in another sample, one could begin to estimate the size of the population.

In our study, however, this is complicated by several factors. First, mutations are not discovered by sequencing of peripheral blood, but have rather been discovered previously by whole genome sequencing; the question is if they are recaptured in different peripheral blood samples. Second, we separated our blood sample into six subsamples, rather than two, for greater resolution. To adapt the analysis for these first two factors only requires minor amendments.

More challenging, however, is the fact that a mutation does not mark a single stem cell, but rather a clone of stem cells. For this to be most informative we need to know when the MRCA of the clone existed (i.e. when the clonal mutation occurred). Although we know on which branch a

Figure 1.5. Quantification of mutations by targeted sequencing. **a,** the phylogenetic tree of cells is shown as in Figure 1.3a, but information from targeted sequencing of peripheral blood granulocytes from the 9 month time-point is overlaid. This is shown more clearly in the inset (**b**), which zooms in on a portion of the tree. The underlying structure of the tree is shown in grey. On top are placed horizontal bars, one for every mutation in the bait-set for targeted sequencing. The bars are coloured according the proportion of cells in the sample that carry the mutations (obtained by multiplying the variant allele fraction for autosomal mutations by two), indicated in the colour scale. Undetectable mutations are coloured grey and shown as smaller bars. Mutations are assigned to a branch based on which colonies they are present in. We cannot know the ordering of mutations along a branch other than from the targeted sequencing of peripheral blood; mutations have therefore been spaced evenly along a branch according to their mean VAF from targeted sequencing of all granulocyte and lymphocyte time-points combined. Small fluctuations in the estimated VAF due to random sampling of mutant reads mean that sometimes a mutation might be at a higher allele fraction in one particular sample than the mutation placed above it; this explains why sometimes a mutation at a low allele fraction in a particular sample is placed higher up the tree than a mutation which is at lower allele fraction in that sample. A higher density of baits was designed for branches shared by more than one colony. On these branches the mutations are so close together that they can appear as one continuous bar. **c**, This schematic explains that the VAFs of mutations decline down branches because of undetected coalescences with stem cells that were not whole genome sequenced but that are producing granulocytes.


a
0
200
400
600
800
1000
1200
Molecular time (substitutions)
b
Mutation at high VAF in granulocytes
Mutation at low VAF in granulocytes
Mutation undetectable in granulo-cytes
Proportion of cells with mutation
Up to 1/2500
1/2500 to 1/500
1/500 to 1/100
1/100 to 1/20
1/20 to 1/4
1/4 to 1/1
c
WGS HSC 1
WGS HSC 2
Missed HSC 1
Missed HSC 2

given mutation falls, we do not, from the whole genome sequencing alone, know the ordering of mutations along that branch. Targeted sequencing of peripheral blood informs on this: the VAF of mutations should decline down a branch, since they represent nested clones. To obtain a point estimate of their molecular time, therefore, we can spread mutations evenly down a branch in order of decreasing VAF. By performing a capture-recapture experiment on mutations from a given slice of molecular time, one could estimate the number of cells that existed at that time with extant descendants (i.e. the number of clones).

The rate of decline of VAFs down the tree is governed by the same coalescent process that determines the pattern of branchpoints in the phylogeny. Indeed, the VAFs decline because of undetected coalescences (figure 1.5c), and so this rate largely measures the same quantity as the rate of observed branching, albeit more sensitively. Recall that this defines the relationship between stem cell number and generation time but does not allow one to be calculated without knowledge of the other.

Extra information, however, comes from the mutations found at the very bottom of the tree, which occurred in the last few generations of stem cells: those that are ancestral to our granulocyte sample with few intermediate stem cells. The number of 'clones' at this level of the tree is essentially the same as the number of active stem cells. We had lysed and sequenced separately six different granulocyte samples from the same time point. Based on flow-sorting, each sample contained ~90,000 cells, and each position in the bait-set was sequenced in each sample at a mean depth of 800X.

Imagine the extreme case where only 10 stem cells make granulocytes at any one time. In this scenario, each of the six granulocyte samples will contain descendants from all of the 10 stem cells, and the VAFs of mutations in all six peripheral blood samples will be very similar to each other. In contrast, imagine a scenario in which one billion stem cells are making granulocytes at any one time. In this case, different samples of granulocytes will contain cells that descend from different ancestral stem cells: a sample of 90,000 granulocytes will not contain descendants from all stem cells. The VAFs of mutations will therefore differ more between different granulocyte samples than in the scenario of a small number of contributing stem cells. This is the information, in conjunction with the pattern of branchpoints in the tree and the rate of VAF decline, that we set out to use to estimate the total number of stem cells in the population.

R.3.c. Approximate Bayesian Computation to estimate stem cell numbers

Because of the complexity of using these multiple sources of information at the same time, it was not possible to solve this problem analytically. In future, it would be possible to apply a fully Bayesian approach, but because computation of the likelihood function involves integration of a large number of possible phylogenies, such an approach is impractically slow. In contrast, an Approximate Bayesian Computation (ABC) can be parallelised and run within the timescale available.

The concept of an ABC is simple: a model is simulated many times, varying the parameters of interest; for each simulation, summary statistics that capture the features of the data of interest are computed; simulated summary statistics are compared to the same statistics calculated on the observed data; and the values of the parameters of interest that produce summary statistics similar to the observed are deemed to be more likely than those that produce very different summary statistics (Tavaré et al. 1997; Beaumont et al. 2002; Bertorelle et al. 2010). The b simulations (where b is a number to be determined, but typically around 1,000) that are most similar to the observed are chosen and the values of the parameters in these simulations can be treated as a credibility interval. Finally, a regression step on a given parameter may be performed (Blum and Francois 2010, Csillery et al. 2012), regressing the value of the parameter against a measure of the dissimilarity between observed and expected summary statistics, thus finding the parameter value that would most closely approximate the observed data.

In our case, a simple model of haematopoiesis (explained below) was simulated thousands of times, varying the number of active stem cells and the generation time, and the whole experiment was repeated *in silico* on every simulation, from whole genome sequencing of the colonies, to design of the bait-set, to estimating the VAFs of mutations in peripheral blood. Such an approach allows us to recreate the uncertainties in our data: for example, we can pretend in the simulation that we do not know the true ordering of mutations, and rather reconstruct it by ordering mutations down a branch according to their estimated VAF (which need not be equal to their true VAF, since it is calculated from the binomial resampling of mutant and wild-type reads), just as in the real experiment. Similarly, we can use the observed sequencing depths from our data to generate a slight unevenness in coverage, making the simulation more realistic. Summary statistics

were chosen to include a measure of the pattern of branch-points on the tree and to reflect a capture recapture approach, by looking at how many mutations were shared between different blood samples. A more detailed description of each step of the method follows.

Please note that while I wrote the ABC code, I received extensive guidance from Kevin Dawson and Peter Campbell.

R.3.c.i. The model of haematopoiesis

R.3.c.i.1. Overview of model

Haematopoiesis is complex, and to model it requires a number of assumptions. We therefore opted for a simple model of haematopoiesis: individual stem cells within the population of total stem cells replicate stochastically over life and their clonal dynamics approximate neutral drift. Stem cells were simulated with a forward-in-time Wright-Fisher model, varying the number of stem cells and generation time. Because of uncertainty in the clonal dynamics of growth to adulthood, we commence our simulations at the point where the adult stem cell pool has reached a constant size, at 100 mutations of molecular time (figure 1.4d). In each cell division, each stem cell acquired a number of somatic mutations drawn from the Poisson distribution. The mean of the Poisson distribution is chosen such that the mean number of mutations per stem cell after the requisite number of generations in that simulation would equal the observed mean of 1,023.

It should be noted that we are using the term ‘stem cell’ to refer to cells that have self-renewed over the 59 years of life of our subject and are ancestral to circulating granulocytes. This population of cells may not translate directly to the standard definition of haematopoietic stem cells as those capable of long-term multi-lineage reconstitution when transplanted into immunocompromised mice. Nevertheless, our ‘stem cells’ meet the conceptual definition of a stem cell since they have self-renewed for so long, and in addition to producing granulocytes are likely also to have produced B lymphocytes (figure 1.8, discussed in section R.4.). We only concern ourselves with the stem cells that have given rise to cells actively making blood. In this experimental set-up, we are blind to long-term dormant stem cells (i.e., years of not making any

sort of long-lived multi-lineage progeny) and to stem cells that exclusively produce other mature blood cell types but no granulocytes.

R.3.c.i.2. Assumptions

First, we model stem cells with active contributions from their progeny as belonging to one population. There is evidence that there may, in fact, be different pools of stem cells cycling at different rates (Wilson et al. 2008), but it is currently unclear whether this applies to humans, how big each pool is, whether cells move from one pool to the other during homeostasis and if so at what rate, and what the relative contribution of each pool to circulating granulocytes is. We therefore chose to model the pool of stem cells actively contributing to granulopoiesis as one population. It is also important to note that while all stem cells share the same probability of dividing in a given window of time, our model in effect allows them to behave stochastically, as in previous models of haematopoiesis (Abkowitz et al. 1996, Catlin et al. 2011).

Second, we assume that the size of this stem cell pool is constant over the period of adulthood studied (figure 1.4s, Werner et al. 2015). We do not concern ourselves with the dynamics of how the population grew to be this size: all our summary statistics only use information from after 100 mutations of molecular time, when we see that the stem cell population size has stabilised. We thus have a population of stem cells of constant size that we allow to replicate stochastically over the course of adult life.

Third, we make the assumption that there is no selection in the stem cell pool. This is based upon our inability to detect selection in our experiment (discussed above) and the assumption that the vast majority of somatic mutations will not affect stem cell function. We thus have a population of stem cells that is replicating stochastically, resulting in a process of neutral drift.

Fourth, we assume that those stem cells that are making granulocytes are all making an approximately equal number of them. This assumption is likely to matter less with larger stem cell population sizes.

The final important assumption that we make is that stem cells accumulate somatic mutations at the same rate over life, drawing the number of mutations that each stem cell acquires at every generation from a Poisson distribution. This is justified by the linear accumulation of

somatic mutations over time reported by others (Welch et al. 2012) and the relatively narrow range of mutation burden that we observe from our 140 clonal whole genomes (figure 1.1c).

With this model, we simulate haematopoiesis hundreds of thousands of times with different values of the stem cell pool size and time between symmetrical stem cell divisions, drawing both from a uniform prior on the log scale (which minimal knowledge of the distribution of the data). We replicate the experiment on each simulation: we sample cells for whole genome sequencing, design a bait-set, sample granulocytes from our stem cell population, and count the number of mutant reads found in the granulocytes over each position in the bait-set.

R.3.c.i.3. Choice of priors

R.3.c.i.3.a. Prior on stem cell population size

The adult stem cell population size was drawn from a uniform distribution on the log scale between 1,096 (exp(7)) and 3,269,017 (exp(15)) stem cells. A minimum number of 1,096 stem cells was chosen because we knew from preliminary simulations and from the reasoning above that a smaller number of stem cells than 1,000 could not produce the low VAF mutations that we observe. The maximum number of stem cells was chosen because, firstly, it was at the limits of what was computationally feasible with the resources available (each simulation at this upper limit requires approximately 150 GB of memory), and, secondly, because it was two orders of magnitude higher than the proposed number of stem cells (Abkowitz et al. 2002).

R.3.c.i.3.b. Prior on generation time

The Wright-Fisher generation time (which is equivalent to the mean time between symmetrical cell divisions for one line-of-descent) was drawn from a uniform distribution on a log scale between 20 days (exp(3)) and 8,103 days (exp(9) i.e. 22 years). The minimum time of 20 days was chosen because stem cells are reportedly relatively quiescent (Arai and Suda 2007; Orford and Scadden 2008). Not all of a stem cell's divisions need be symmetrical: a proportion is

likely to be asymmetrical, producing one daughter stem cell and one progenitor cell (Werner et al. 2015). We are blind to asymmetrical divisions. Therefore, if a cell is dividing asymmetrically in addition to symmetrical divisions on average every 20 days, it will be dividing significantly faster than every 20 days, which seemed unlikely given prior knowledge of stem cell quiescence. Furthermore, shorter times between cell divisions mean that more generations need to be simulated, which is computationally costly. The maximum time between symmetrical cell divisions of 22 years was chosen because it required a very small number of HSCs to create a phylogeny of the right shape, and such a small number was not compatible with the observed range of VAFs.

R.3.c.i.4. *In silico* recapitulation of our experiment

3.c.4.i.a. Mutation discovery phase

After simulating drift for the whole population of stem cells for the requisite number of generations, we choose 155 colonies for whole genome sequencing and construct a phylogeny from them. Of 198 colonies sequenced in our experiment, only 140 were clonal. These 140 were used to build the tree, and for all analyses except for the ABC. However, of the 58 polyclonal colonies, we could salvage 15 because they had a dominant clone and shared more than ten mutations with a clonal colony that was on our tree. We could therefore graft these 15 extra colonies onto the tree of 140 clones. This is helpful because it provides an extra time point on each branch onto which an extra colony is grafted. Mutations can then be classified as being shared with the polyclonal colony that has been grafted on, or absent from the polyclonal colony, thus providing additional information about the timing of the mutation. No mutations that were present only in polyclonal colonies (and not in clonal colonies) were used, as we could not be sure where to place them on the tree.

R.3.c.i.4.b. Mutation quantification phase

We design a bait-set for the simulated tree, using the same criteria as were used to design the real bait-set (Methods). We then simulate the sampling of peripheral blood granulocytes. We generate a sample of M = 540,000 granulocyte by sampling with replacement from the stem cell population. This simulated sample of 540,000 cells is then split into six sub-samples each of size 90,000 (to simulate the six sub-samples obtained from our volunteer at the nine-month time-point). We then simulate targeted sequencing of the mutations in the bait-set. In the observed data, there were 3,952 mutations in the bait-set that (after our duplicate removal and consensus calling step (methods)) were covered by at least 4,000 reads in the control cord blood DNA, but where no mutant reads were found in the cord blood. We therefore used these 3,952 mutations for analysis of the observed data, and also only use 3,952 mutations in the simulated bait-set. For every bait-set locus in every sub-sample, we randomly draw a sequencing depth from the empirical distribution of sequencing depths for the real targeted data from the nine month time-point. We sample the chosen number of reads over this locus from the granulocytes without replacement and count the number of reads that have the mutation.

Sequencing errors were included in the simulation as follows. The sequencing error rate was learnt from the control cord blood. For all 7,116 positions in the bait-set, the VAF in the cord blood was calculated. Where the VAF was zero (if there were no mutant reads), the VAF was set to 1/10,000, a value just below the rarest mutations that we could detect. For each of the 3,952 loci used in simulations, then, an error rate was drawn from these VAF distributions. The number of false positive reads was obtained by drawing from the binomial distribution, with parameter p equal to the randomly chosen error rate and parameter n equal to the sequencing depth. We also tested two other error models: one with no sequencing errors, and one with double the sequencing error rate observed in the cord blood controls. These made little difference to the median of the posterior distribution of the model but did affect its width. A separate false positive rate was included based on the estimated rate of homoplasy, assuming that every granulocyte has 2,000 mutations (double the number of mutations present in a stem cell, which seemed a reasonable upper bound for the number of additional mutations that a granulocyte could acquire) and that these are spread evenly across the genome.

R.3.c.ii. Extraction of summary statistics

We then extract summary statistics from the resulting simulated data set. There are two categories of summary statistics used. Only the first category of summary statistics was extracted for the first set of simulations, as explained below.

The first category of summary statistics is comprised of those designed to capture the shape of the phylogeny, referred to as lineages through time or LTT summary statistics. We divide the molecular time scale, from mutation 100 to mutation 800 (beyond which some branches in our observed phylogeny end), into bins of molecular time each 100 substitutions wide. For each mutation of molecular time in a given bin, we count the number of branches extant at that slice of time and take the mean across all hundred mutations in that bin.

The second category of summary statistics is made up of those designed to exploit the 'capture-recapture' aspect of our approach with targeted sequencing of multiple peripheral blood samples. We wanted statistics that would capture whether different granulocyte samples descended from the same stem cell population or not. As explained above, the less overlap in the granulocyte sub-samples, the larger the contributing stem cell population is likely to be. For a given mutant read threshold (from 1 to 6 mutant reads), and a given number of samples (from 0 to 6 samples) we count – out of the 3,952 mutations included – how many loci have this many or more mutant reads in precisely this many samples. For example, one summary statistic would be the number of mutations in the bait-set that are supported by at least three mutant reads in two samples. These summary statistics are recorded for every simulation, and the same summary statistics were calculated for the observed data.

R.3.c.iii. Details of ABC implementation

Two sets of simulations were run, referred to hereafter as ABC1 and ABC2, both using the same model, but with a different prior and extracting different summary statistics.

First, we generated 121,329 simulations drawing both the number of stem cells and the time between symmetrical stem cell divisions from a uniform prior on a log scale (figure 1.6a) and

extracted summary statistics that reflect the shape of the phylogeny (the LTT summary statistics explained above). This allowed us to identify a relationship between stem cell number and generation time, effectively defining a plausible diagonal band on the sample space of stem cell numbers plotted against generation time (figure 1.6b). Going backwards in time, the faster the rate of random drift, the more rapidly the number of lineages decreases. Thus, simulations that have too rapid a random drift rate (simulations with a small population size and short generation time) have LTT statistics that are too low for early bins of molecular time (figure 1.6o), and simulations that have too slow a random drift rate (with a large population size and long generation time) have LTT statistics that are too high for early bins of molecular time (figure 1.6p).

For ABC2, we ran another 80,762 simulations targeting this area of the sample space (figure 1.6c). For this set of simulations, both the LTT and peripheral blood-derived summary statistics were extracted.

3.c.iv. Comparison of simulations and observed data

Summary statistics were normalised, and for each simulation the Euclidean distance between the simulated and observed vector of summary statistics was calculated. For ABC1, we simply defined the region that contained the most similar 20% of simulations, which was then used as the prior for ABC2. Simulations outside this range can be shown to have a branching structure that does not resemble that of our phylogeny (figure 1.6o-p).

For ABC2, to maximise the accuracy of our model, we cross-validated both the number of simulations included in the acceptance region, and the weighting to give to the LTT statistics. For each of 1,000 cross validation samples, we drew one simulation to act as fake observed data and removed it from the pack of simulations. We then analysed the data as though our fake observed data were the true data. We took the n (where n is the number of accepted simulations) simulations that produced summary statistics that were most similar to our fake observed data, as determined by their Euclidean distance.

We then calculated a number of statistics (figure 1.6e-i). First, we plotted the accepted simulations on a graph of stem cell numbers *versus* generation time (both on a log scale) and drew an ellipse that contained 90% of the n points inside it. We then saw whether the true value of the

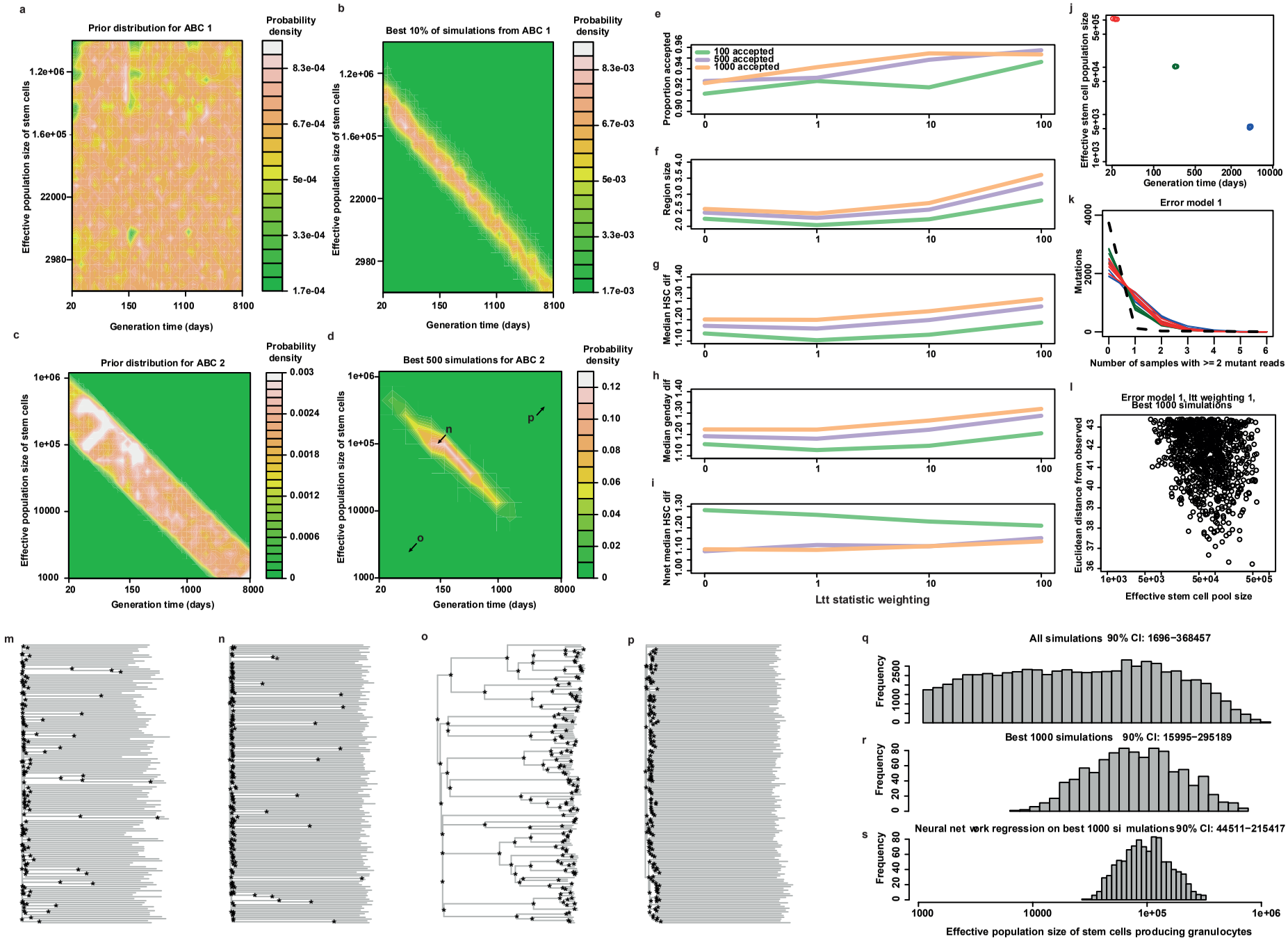
fake observed data fell inside this ellipse. The proportion of cross validation samples for which the fake observed data fell inside the ellipse is shown in figure 1.6e. We also measured: the mean area of the ellipse (figure 1.6f); the distance between the median number of stem cells of the accepted simulations and the fake observed data number of stem cells (figure 1.6g); the distance between the median generation time of the accepted simulations and the fake observed data generation time (figure 1.6h); and, finally, the distance between median of the posterior of a neural network regression run on the accepted simulations and the fake observed stem cell number (figure 1.6i). We chose an LTT weighting of 1 and an acceptance region of 1,000, since this resulted in an accurate prediction of the stem cell number from the neural network regression and a high proportion of the fake observed values falling in the ellipse, while keeping the size of the ellipse relatively small.

We then analysed the true observed data, using the error model that took VAFs from the observed control data, weighting the LTT summary statistics by 1 and choosing the 1,000 most similar simulations to fall in the accepted region. Of the best 1,000 simulations, 90% had more than 15,995 and less than 295,189 HSCs, and a generation time of more than two and less than 20 months (figure 1.6r).

For added precision on the number of stem cells, neural network regression was performed on these best 1,000 simulations using the R package *abc* (Csillery et al. 2012), to find the number of stem cells that minimised the distance between the observed and simulated summary statistics. The neural network regression was run using the default of one hidden layer with five units. As predictions from different neural networks can vary, thirty neural networks were run and the median provided. We found that the most likely number of stem cells ancestral to the sampled granulocytes was 97,000, with a 90% credibility interval of 45,000 to 215,000 (figure 1.6s).

To test the robustness of our analysis, we repeated it with the other two error models described and ignored summary statistics that used a mutant read cut-off of 1, since these would be most sensitive to incorrect modelling of the sequencing error rate. Both of these additional analyses resulted in a widening of the posterior distribution for the number of contributing stem cells but did not significantly change its location.

Figure 1.6. Approximate Bayesian Computation (ABC) to estimate HSC numbers. **a,** The joint prior distribution for HSC numbers and the generation time for the first ABC. **b**, The location in sample space of the 10% of simulations that produced ltt summary statistics most similar to the observed summary statistics. **c**, the joint prior distribution for the second ABC, in the area of sample space indicated to be plausible by the first set of simulations. **d**, The joint posterior distribution of the best 500 simulations from the second ABC. Letters n, o, and p on the plot indicate the position in sample space from which panels **n**, **o**, and **p** were drawn, respectively. **e-i,** cross-validation of the model to choose the number of accepted simulations and the weighting applied to the ltt summary statistics. **j**, for illustrative purposes, five simulations were sampled for each of three population sizes along the plausible diagonal of sample space indicated in panel **b**. One set of summary statistics is shown for these simulations in **k**. A red line indicates a simulation coming from the area of sample space indicated by a red point in **j**; *idem* for blue and green lines. The black dotted line indicates the observed values for these summary statistics. This set of summary statistics counts, for different numbers of samples (x axis), how many of the 3952 mutations considered (y axis) are in this many samples with two or more reads, using error model 1 (which simulates errors according to the error rate in control DNA). **l**, For each of the 1000 simulations that produce summary statistics most similar to the observed, the Euclidean distance from the observed (y axis) is plotted against the number of stem cells in that simulation (x axis). This information is used by the neural network regression to define the most likely value for the number of stem cells. The most similar values are seen at around 100,000 stem cells, which was the location of the median of the posterior distribution from neural network regression. **m**, the observed phylogeny, with branch points indicated by asterisks. **n-p,** phylogenies from simulations that occur at the points in sample space indicated in panel **d**. **n** represents a plausible simulation, since the pattern of branch points is not dissimilar from that seen in the observed phylogeny **m**. Simulations with smaller stem cell populations and faster stem cell turnover rates resulted in phylogenies where the stem cells are very closely related to each other (**o**), whereas those with larger populations and slower turnover result in phylogenies where the stem cells only share an embryonic common ancestor (**p**). **q**, the prior distribution for the number of stem cells contributing to granulocytes for the second ABC (i.e. the stem cell numbers for all 80,000 simulations). **r,** the distribution of stem cell numbers for the 1,000 simulations that produced summary statistics most similar to the observed summary statistics. **s,** the posterior distribution of a neural network regression run on these 1,000 simulations. The 90% credibility interval is quoted for the stem cell population in each of **q-s**.


a
Prior distribution for ABC 1
Probability density
8.3e-04
6.7e-04
5e-04
3.3e-04
1.7e-04
Effective population size of stem cells
1.2e+06
1.6e+05
22000
2980
20
150
1100
8100
Generation time (days)
b
Best 10% of simulations from ABC 1
Probability density
8.3e-03
6.7e-03
5e-03
3.3e-03
1.7e-03
Effective population size of stem cells
Generation time (days)
c
Prior distribution for ABC 2
Probability density
0.003
0.0024
0.0018
0.0012
0.0006
0
Effective population size of stem cells
1e+06
1e+05
10000
1000
20
150
1000
8000
Generation time (days)
d
Best 500 simulations for ABC 2
Probability density
0.12
0.10
0.08
0.06
0.04
0.02
0
n
o
p
Effective population size of stem cells
Generation time (days)
e
Proportion accepted
0.90 0.92 0.94 0.96
100 accepted
500 accepted
1000 accepted
0
1
10
100
f
Region size
2.0 2.5 3.0 3.5 4.0
g
Median HSC dif
1.10 1.20 1.30 1.40
h
Median genday dif
1.10 1.20 1.30 1.40
i
Nnet median HSC dif
1.00 1.10 1.20 1.30
Ltt statistic weighting
j
Effective stem cell population size
1e+03
5e+03
5e+04
5e+05
20
100
500
2000
10000
Generation time (days)
k
Error model 1
Mutations
4000
2000
0
0 1 2 3 4 5 6
Number of samples with >= 2 mutant reads
l
Error model 1, ltt weighting 1,
Best 1000 simulations
Euclidean distance from observed
36 37 38 39 40 41 42 43
1e+03
5e+03
5e+04
5e+05
Effective stem cell pool size
m
n
o
p
q
All simulations 90% CI: 1696−368457
Frequency
0 1000 2500
r
Best 1000 simulations 90% CI: 15995−295189
Frequency
0 20 40 60 80
s
Neural net work regression on best 1000 si mulations 90% CI: 44511−215417
Frequency
0 20 40 60 80
1000
10000
1e+05
1e+06
Effective population size of stem cells producing granulocytes

4. Clonal contributions to granulocytes and lymphocytes.

In order to examine lineage relationships at the clonal level, we performed targeted sequencing of B (mean coverage 2,830X) and T lymphocytes (mean coverage 3,298X), both from the 9-month time-point, using the same bait-set and error-correction method as for granulocytes (figures 1.6 and 1.8). As can be observed from the preponderance of black bars at the top of the phylogeny, most early embryonic branches contribute to all three cell types at the time of sampling. Thus, clones marked in the embryo, and perhaps early childhood, have multilineage potential, as has been observed in mice (Pei et al. 2017).

Beyond 100 mutations in molecular time, when the population size reached a plateau (figure 1.4d), 464 mutations distributed across 39 branches could be detected, of which 217 on 12 branches were detected in more than one cell type. Hardly any of these were detected in all three lineages. We do, however, observe a number of clones that make detectable amounts of both B lymphocytes and granulocytes. Sequencing coverage was equally good in T lymphocytes and in B lymphocytes, so the failure to detect these mutations in T lymphocytes is not a result of decreased sensitivity. At least five of these shared granulocyte-B cell branches stretch beyond the 100 mutations mark (figure 1.8) that we take to signify adult stem cell dynamics. Furthermore, the VAFs observed are similar between B cells and granulocytes, indicating a more recent common ancestor: the neutral drift to detectable clone sizes will have been gradual, with much of it occurring since the last detectable mutation. If the last detectable purple mutation in figure 1.8a represented the last common ancestor of granulocytes and B lymphocytes, both the B-restricted and granulocyte-restricted clones would have had to drift to similar sizes in the intervening time independently. It is more likely that they share a common ancestor well beyond the last detectable mutation found in both B cells and granulocytes. Thus, we demonstrate the presence of adult stem cell clones with both myeloid and lymphoid output in a human at homeostasis. Admittedly, the number of clones in which we have detected this multilineage output is small. However, there are no clones in which we would have been able to detect multilineage output in which we have not detected it (no branches beyond 100 mutations of molecular time that are exclusively dark red, blue, or green). This suggests that multilineage clonal output is not rare.

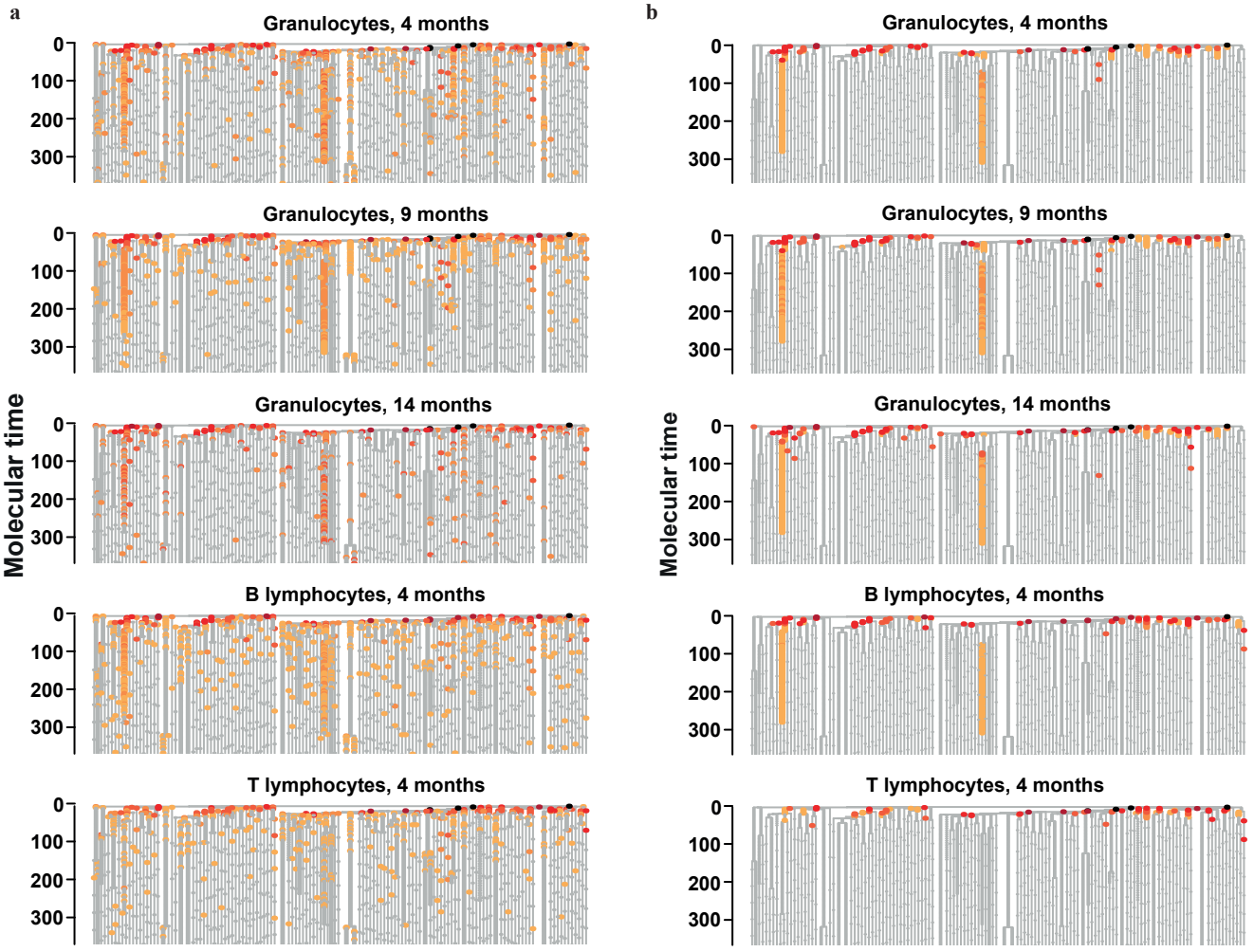
We cannot exclude smaller contributions of these granulocyte- and B cell-producing clones to the T lymphocyte pool on the order of less than one in three thousand T lymphocytes, which

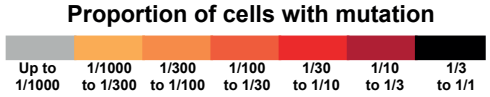
would fall below our detection limit for lymphocytes. Nonetheless, since this is less than half the proportion that these clones are contributing to granulocytes and B lymphocytes, we can unequivocally show an imbalance. In contrast, we observe some branches where we can detect very rare mutations (allele fractions of less than 1/2,000) in granulocytes (pink branches in figure 1.8), which we did not have the sequencing depth to detect reliably in lymphocyte populations. It is quite plausible that these branches contribute to lymphocytes, too; we cannot, therefore, infer the presence of exclusively myeloid clones.

Only a handful of mutations were found to be shared by B and T lymphocytes and absent from granulocytes. Even if lymphoid-biased clones were frequent, however, we would not expect to find many of them, as the 'capture' phase of our experiment was biased towards myeloid lineages, since isolation strategies for human stem cells inevitably also capture more differentiated cells, predominantly myeloid progenitors (despite possessing the same combination of surface markers as true HSCs).

The observation of granulocyte-B cell restricted clones in adults would be reinforced by studying more cells from more people and with the ability to resolve rarer mutations. Even without these, our data already indicate that adult stem cell clones with both myeloid and lymphoid output are not infrequent in humans. In mice, tags have been found to be shared between granulocytes and B lymphocytes (Rodriquez-Fraticelli et al. 2018), but, unfortunately, T lymphocytes were not assessed. The observation of clones that produce detectable granulocyte and B lymphoid but not T lymphoid output could be explained in three ways. First, some individual stem cells may have a propensity to produce more of one mature cell type than another. This would have to be a somatically heritable trait, since we observe this behaviour at the level of clones that are ancestral to hundreds of stem cells. Such a finding of heritable stem cell heterogeneity is anticipated by the animal and *in vitro* studies discussed at the beginning of this chapter. Second, unequal clonal contributions to present-day blood samples might reflect the long life-span of T lymphocytes relative to granulocytes and some B cells: T lymphocytes could reveal the stem cell clones that dominated decades ago, while granulocytes and B lymphocytes show us more recent clone distributions. A third and related possibility is of a strong bottleneck in the T lymphocyte population imposed by thymic selection or infections, which would reduce the overlap between these populations. This could be resolved by including lymphocyte progenitors in the whole genome sequencing phase of the study, and by performing this experiment in children (since we

Figure 1.7. Quantification of mutations in different fractions of peripheral blood. a, b, targeted sequencing data represented as in figure 1.5 for all fractions of peripheral blood that were sequenced, showing only the first 350 mutations of molecular time, beyond which no mutations were detectable. To allow a better comparison between samples sequenced at different depths, a different detection threshold is used relative to figure 1.5. Data are shown prior to sequencing error correction using cord blood controls in the Bayesian generalised poisson mixed effects model (**a**) and after applying the error correction (**b**). **c**, correlations between the VAFs of all sequenced samples, shown on a log scale. Note that samples that were sequenced to lower depth cannot have VAFs as small as samples sequenced to higher depths.


a
b
Granulocytes, 4 months
Granulocytes, 9 months
Granulocytes, 14 months
B lymphocytes, 4 months
T lymphocytes, 4 months
Molecular time
0
100
200
300


Proportion of cells with mutation
Up to 1/1000
1/1000 to 1/300
1/300 to 1/100
1/100 to 1/30
1/30 to 1/10
1/10 to 1/3
1/3 to 1/1

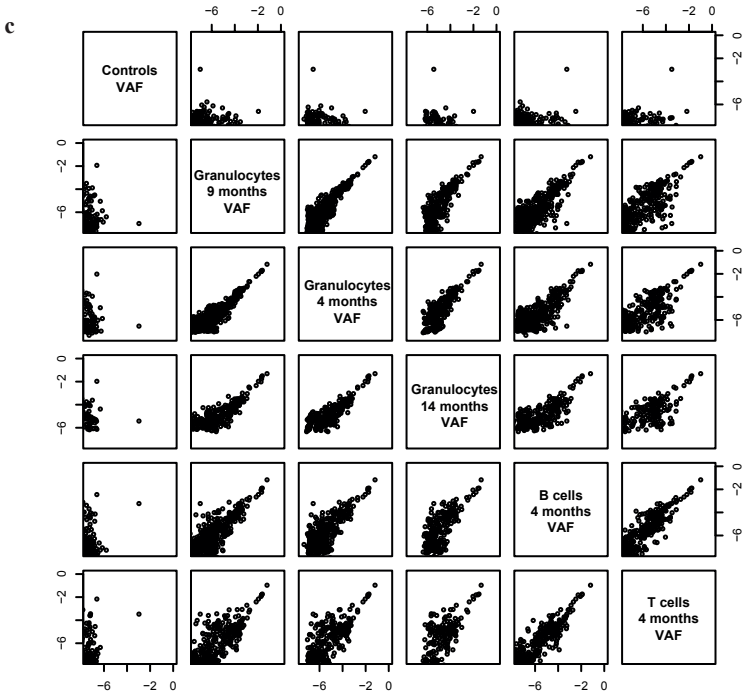
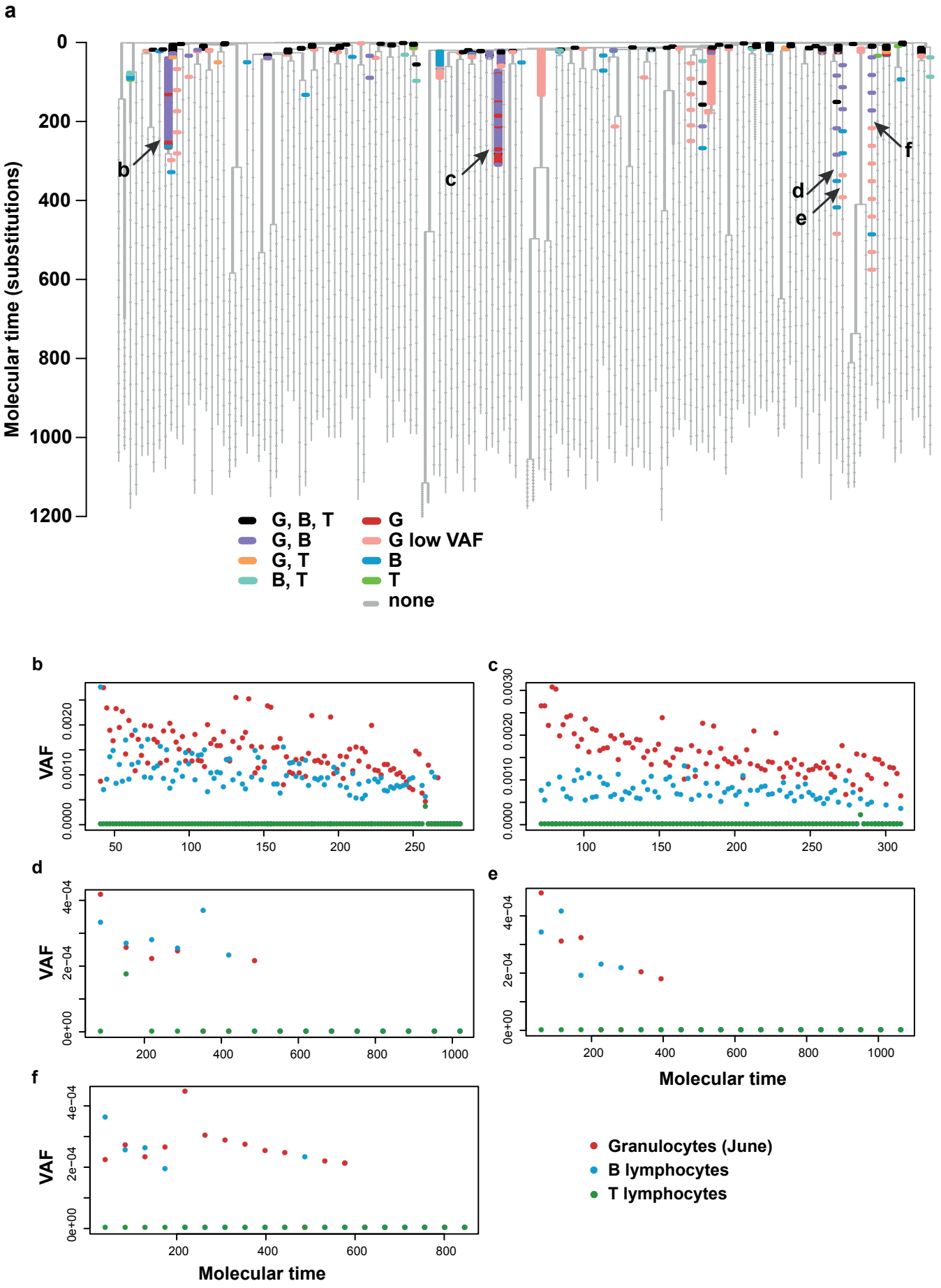

c
Controls VAF
Granulocytes 9 months VAF
Granulocytes 4 months VAF
Granulocytes 14 months VAF
B cells 4 months VAF
T cells 4 months VAF
−6
−2
0

Figure 1.8. Comparison of allele fractions in granulocytes and lymphocytes. a, the phylogeny is depicted as in figure 1.5, with the underlying structure of the tree shown in grey, and horizontal bars drawn to represent every mutation in the bait-set. Here the colouring of mutations reflects which peripheral blood cell fractions they could be detected in, as indicated by the colour key. Granulocytes were sequenced at greater depth than lymphocytes and so mutations could be detected at lower allele fractions in granulocytes than in lymphocytes. Therefore, two colours are used for granulocytes: red for mutations only detected in granulocytes that were are at a sufficiently high allele fraction to have been found in lymphocytes should they have been present at a similar allele fraction in lymphocytes, and pink for mutations that were only detected in granulocytes, but were at such a low allele fraction (<1/2000 reads) that if they had been present in lymphocytes at this allele fraction they would not have been detected. Arrows indicate adult clones with multilineage output. G, granulocytes; G low VAF, granulocytes, allele fraction too low to be detected in lymphocytes; B, B lymphocytes; T, T lymphocytes. **b-f,** VAFs of all mutations on branches (indicated by arrows in **a**) with mutations beyond molecular time 100 that are detectable in granulocytes and B lymphocytes but not T lymphocytes.


a
Molecular time (substitutions)
0
200
400
600
800
1000
1200
b
c
d
e
f
G, B, T
G, B
G, T
B, T
G
G low VAF
B
T
none
b
c
d
e
f
VAF
0.0000
0.0010
0.0020
0.0030
4e−04
2e−04
0e+00
50
100
150
200
250
300
400
600
800
1000
Molecular time
Granulocytes (June)
B lymphocytes
T lymphocytes

were able to culture a reasonable number of stem cells from peripheral blood, a bone marrow aspirate would not be necessary).

Summary of results in this chapter

The number of haematopoietic stem cells, their clonal dynamics, and the clonal relationships between different cell types are all poorly understood in humans because of the difficulty of tracking clones in unperturbed people. We used spontaneously acquired somatic mutations to reconstruct lineage relationships among stem cells in normal human haematopoiesis. We sequenced the whole genomes of 140 colonies derived from single HSPCs from one healthy 59 year-old man. We identified the somatic mutations, reconstructed the phylogenetic relationships of the cells to one another, and sequenced at high depth bulk populations of granulocytes and lymphocytes for mutations that had been discovered by whole genome sequencing.

We were able to reconstruct cell divisions in the early embryo. Comparisons with buccal epithelium indicated that the most recent common ancestor of blood existed prior to gastrulation. The mutation rate in the early embryo is likely to be less than two mutations per mitosis. Averaging over the whole of life, the mutation burden and mutational signatures were consistent across different HSPCs. No positive or negative selection of mutations could be identified.

The trajectory of the number of haematopoietic stem cells over life could be inferred from the branching patterns in the phylogeny. The stem pool size increased rapidly during development and childhood and reached a stable plateau in adulthood. Using deep sequencing data, we inferred that the number of stem cells contributing actively to granulocytes at any given time is in the range 45,000 – 215,000.

Finally, we found that T and B lymphocytes have different clonal dynamics. We observed adult clones that produced detectable fractions of both granulocytes and B lymphocytes, but not T lymphocytes. Our data indicate a contribution of multipotent HSCs to B lymphopoiesis throughout life.

These findings are discussed in the Discussion chapter.