Chapter 1

Introduction

Modern sequencing technologies are catalysing a revolution in our understanding of cancer genetics, developmental disorders, and ageing (Behjati et al., 2018; Martincorena and Campbell, 2015; Stratton, 2011; Yates and Campbell, 2012). Over the past decade, genomic scrutiny of over a million cancers has revealed the oncogenic mutations responsible for causing most human malignancies (Tate et al., 2019). These discoveries have enabled development of novel targeted cancer therapies and sequencing-based cancer diagnostic methods (Chang et al., 2016; Gerstung et al., 2017; Zahn, 2016). In parallel, sequencing of normal tissues has demonstrated that somatic mutations accumulate in all cells with age due to a host of extrinsic and endogenous exposures (Alexandrov et al., 2013; Hoang et al., 2016; Ju et al., 2017; Martincorena and Campbell, 2015; Yizhak et al., 2018). Somatic genetic diversity in ageing tissues provides a substrate for natural selection at the cellular level. Most somatic mutations have no discernible impact on cell function (Martincorena et al., 2017). However, recent studies have demonstrated that canonical cancer driver mutations are remarkably common in morphologically and functionally normal tissues and frequently fuel clonal expansion (Bowman et al., 2018; Martincorena et al., 2018; Martincorena et al., 2015; Moore et al., 2018; Salk et al., 2018; Yizhak et al., 2018; Yokoyama et al., 2019). The ubiquity of subclonal cancer evolutionary processes represents a daunting challenge to sequencing-based early cancer detection efforts and may also increase the toxicity of novel precision oncology drugs targeting cancer driver mutations present in a significant fraction of normal cells (Busque et al., 2018; Cohen et al., 2018; Martincorena et al., 2015). The landscape of somatic genetic diversity is currently best understood in the haematopoietic system, largely due to ease of representative sampling. Clonal haematopoiesis (CH) becomes increasingly common with age and is associated with an increased risk of haematological malignancies,

though only a small minority of individuals with CH ever develop a blood cancer (Busque et al., 2018). The main aim of this dissertation has been to explore the premalignant mutational landscape of haematological cancers and the extent to which indolent clones can be distinguished from CH at high risk of malignant transformation. The general introduction to this thesis provides an overview of somatic evolution in cancer and normal tissues, with an emphasis on the haematopoietic system.

1. Somatic evolution in cancer

"At last gleams of light have come, & I am almost convinced (quite contrary to opinion I started with) that species are not (it is like confessing a murder) immutable."

- Charles Darwin to Joseph Hooker, 11 January 1844

"*One general law, leading to the advancement of all organic beings, namely, multiply, vary, let the strongest live and the weakest die…. Natural Selection, as we shall hereafter see, is a power incessantly ready for action*"

- Charles Darwin, *The Origin of Species*, 1959

"If, as I believe that my theory is true & if it be accepted even by one competent judge, it will be a considerable step in science."

- Charles Darwin to Emma Darwin 5 July 1844

As presciently anticipated by Darwin, natural selection is relevant to much more than the evolution of free-living species. The cells that make up multicellular organisms possess the requisite features for natural selection according to Darwin: heritable variation that impacts fitness. Cells, like species, are mutable, inevitably accumulating changes in their genomes due to extrinsic factors (e.g., radiation) and endogenous processes (e.g., errors in DNA replication and repair) (Alexandrov et al., 2013; Martincorena and Campbell, 2015). According to current estimates, most cells accumulate one to two mutations per cell division (Yizhak et al., 2018), though this rate may vary considerably (Hoang et al., 2016). Somatic mutations generate variety and starting from early embryogenesis, multicellular organisms become mosaics of

genetically distinct cells (Behjati et al., 2014; Blokzijl et al., 2016; Ju et al., 2017). This variety creates a substrate for natural selection. Although few somatic mutations impact cell function (Martincorena et al., 2017), occasionally a mutation confers a fitness advantage, favouring clonal expansion of the cell harbouring it (Martincorena and Campbell, 2015; Yates and Campbell, 2012). The competitive advantage conferred by a given mutation may be context-dependent, varying with environmental exposures (Bondar and Medzhitov, 2010; Wong et al., 2015b; Yates and Campbell, 2012; Yokoyama et al., 2019). Cell competition has been most extensively studied in simpler model organisms, where it is often a beneficial physiological process that helps ensure that tissues are made up of the healthiest cellular constituents (Amoyel and Bach, 2014; Baker and Li, 2008). In humans, somatic evolution has primarily been studied in the context of cancer, where the process produces a cell with a complement of mutations enabling it to escape normal constraints on proliferation and to invade other tissues (Hanahan and Weinberg, 2000, 2011). However, recent studies of somatic mutation in the context of human development, ageing, pre-cancer, cancer and non-malignant disease have indicated that the border between normal age-related somatic evolution and malignancy can be indistinct (Martincorena et al., 2018; Martincorena et al., 2015; Moore et al., 2018; Salk et al., 2018; Yizhak et al., 2018; Yokoyama et al., 2019). This introduction will provide an overview of somatic evolution in cancer and ageing with a focus on the haematopoietic system, which has been particularly well characterised due to ease of representative tissue sampling.

1.1 Cancer is a genetic disease

"…a malignant cell is a cell with an irreparable defect, located in the nucleus. There is a permanent change in the condition of the chromatin which forces the cell to divide."

- Theodore Boveri, '*The Origin of Malignant Tumours'*, 1914 (Manchester, 1995)

"I got sort of amused tolerance at the beginning."

- Janet Rowley recalling the response of the scientific community to her 1972 discovery that chromosomal translocations could cause cancer. (Fox, 2013)

The history of the mutational theory of cancer is a reminder of the power of simple experiments interpreted well and of the amount of time it can take for pivotal discoveries to elicit follow-up work and acceptance. Theodore Boveri is generally credited with being the first biologist to recognise that abnormal genetic content is responsible for malignant transformation (Rowley, 2001). His observations stemmed from meticulous light microscope scrutiny of sea urchin embryo divisions and the observation that aberrant mitoses seemed to trigger developmental defects.

"Experiments on sea urchin embryos have led to the result that most chromosome combinations that vary from the normal lead to the death of the cell; however, other combinations occur, in which the cell, while it remains viable, does not function in a typical way."

- Theodore Boveri, '*The Origin of Malignant Tumours'*, 1914 (Manchester, 1995)

Boveri concluded that chromosomal content guides embryogenesis and further speculated that the entities responsible for Mendelian traits must reside within chromosomes:

"I feel beyond any doubt that the individual chromosomes must be endowed with different qualities and that only certain combinations permit normal development."

- Boveri, 1901 (Hardy and Zacharias, 2005)

''The probability is extraordinarily high that the traits examined in the Mendelian experiments are linked to individual chromosomes''

- Boveri, 1914 (Hardy and Zacharias, 2005)

These conclusions led Boveri to revisit observations made over twenty years previously by David Hansemann (1858–1920), a German pathologist who had documented asymmetrical nuclear segregation in a host of human cancers (Hardy and Zacharias, 2005). Hansemann maintained that nuclear abnormalities were most likely to represent characteristic sequelae of the malignant process (Hardy and Zacharias, 2005). Boveri, reinterpreting Hansemann's findings in the context of the sea urchin experiments, posited that cancers are the progeny of

a single cell that acquired uncontrolled growth potential due to abnormal chromosomal content (Hardy and Zacharias, 2005; Manchester, 1995). Boveri's hypothesis that chromosomes contained the material of inheritance was confirmed by the experiments of Avery, MacLeod and McCarty in 1944 (Avery et al., 1944). Further evidence that tumours often contain wildly bizarre chromosomes accumulated over the ensuing decades as cytogenetic methods improved. In the 1950s, Hauschka, Levan, Makino and others documented that most cancer cell lines contain aberrant chromosome numbers, as well as dicentric and ring chromosomes (Rowley, 2001). However, there was no apparent trend between particular abnormalities and cancer type, leading to further scepticism of any role in carcinogenesis (Rowley, 2001).

In the 1960s and 1970s, a clear association emerged between specific chromosomal abnormalities and particular leukaemias. In 1960, Nowell and Hungerford reported the Philadelphia (Ph) chromosome in almost all cases of chronic myeloid leukaemia (CML) (Nowell and Hungerford, 1960). Aided by improved chromosome banding techniques, Janet Rowley was able to establish that the Ph chromosome represented an interchange between chromosomes 9 and 22 (Rowley, 1973). Several other recurrent translocations were discovered in the 1970s by Rowley, Zech and others, notably the AML-associated t(8;21), t(8;14) in Burkitt lymphoma and t(15;17) in acute promyelocytic leukaemia (Rowley, 2001; Zech et al., 1976). It took until the early 1980s for the diagnostic and prognostic utility of these findings to be incorporated into clinical guidance (Rowley, 2001).

The advent of clinical cytogenetics coincided with further definitive proof that somatic mutations in DNA cause cancer. Weinberg, Cooper and colleagues demonstrated that human tumour DNA introduced into a mouse fibroblast cell caused malignant transformation (Krontiris and Cooper, 1981; Shih et al., 1981). Retrieval of the human sequence from the murine malignant cells ruled out spontaneous in vitro transformation, as can occur in many putatively normal cell lines (Krontiris and Cooper, 1981; Shih et al., 1981). Isolation of the oncogenic DNA fragment led to the discovery of an activating substitution mutation in *HRAS*, thus demonstrating for the first time that simple missense mutations, in addition to chromosomal rearrangements, can cause cancer (Reddy et al., 1982; Tabin et al., 1982). This discovery stimulated widespread concerted efforts to systematically identify genetic mutations capable of causing cancer.

Cancer gene discovery efforts further accelerated following the release of the first draft human genome sequence in 2000 (Lander et al., 2001; Venter et al., 2001) and the advent of massively parallel sequencing a few years later (Stratton, 2011; Stratton et al., 2009). The ensuing revolution in genomics has yielded unprecedented insights into the pathogenesis of cancer, as well as the inextricably related processes of human development and ageing. The next section will give an overview of some important concepts that have emerged from the study of the cancer genome.

1.1.1 Classifying mutations according to selection: 'driver' and 'passenger' mutations

To date, over 1.4 million tumour samples have been sequenced, including tens of thousands of whole genomes (Sondka et al., 2018). The ability to scrutinise whole genomes from diverse cancer types has revealed dramatic variation in somatic mutation burden, ranging from over 100 per megabase (Mb) in some melanomas and mismatch-repair deficient tumours to fewer than 0.01 mutations/Mb in some childhood cancers and leukaemias (Alexandrov et al., 2013; Shlien et al., 2015; Stratton, 2011).

A key focus of cancer genomics has been to classify somatic mutations according to whether or not they are under positive, neutral or negative selective pressure. Identifying the minority of mutations that are under positive selection and playing a causative role in oncogenesis (hereafter referred to as 'driver mutations') from mutations that do not confer a fitness advantage ('passenger mutations') is an ongoing and complex task (Lawrence et al., 2013; Martincorena et al., 2017; Stratton et al., 2009). The phenotypic features under positive selection in cancers have been conceptualised as the "hallmarks" of cancer and all, in essence, promote survival and/or growth (Hanahan and Weinberg, 2000, 2011). The most recent release of the Cancer Gene Census included 719 genes implicated in driving human cancers (Tate et al., 2019), although this list is constantly being amended and expanded to accommodate new genomic and functional evidence. The extent to which negative selection shapes somatic evolution in cancers and normal tissues is contentious, though at present most evidence suggests that positive selection plays a much more important role in governing clonal dynamics (Martincorena et al., 2017; Zapata et al., 2018).

1.1.2 Classifying cancer genes: tumour suppressors and oncogenes

Although often an oversimplification, it has proven conceptually useful to broadly classify cancer genes as either tumour suppressor genes or oncogenes. Tumour suppressor genes are implicated in oncogenesis through loss-of-function mutations (Stratton et al., 2009). Tumour suppressor genes frequently encode negative regulators of cell cycle progression (e.g., *RB1, PTEN*), suppressors of cell growth (e.g., *NF1*), pro-apoptotic signalling molecules (e.g., *DAXX*), proteins linking the DNA damage response to the cell cycle (e.g., *ATM*, *TP53*), cell-adhesion mediators (e.g., *APC*), DNA damage repair proteins (e.g., *BRCA1*) and epigenetic regulators (e.g., *KDM6A, SETD2, DNMT3A, TET2*) (Martincorena et al., 2017; Stratton, 2011). Many tumour suppressors, like the prototypical *RB1* that gave rise to Knudson's 'two-hit' hypothesis (Knudson, 1971), function in a recessive manner (Stratton, 2011). However, for many tumour suppressors, haploinsufficiency alone promotes cancer development (e.g., *TP53, RUNX1, PTEN, TET2, DNMT3A*)(Döhner et al., 2015; Inoue and Fry, 2017). Many types of mutations can inactivate tumour suppressor genes, including truncating mutations (e.g., nonsense, frameshift, disruptive rearrangements, essential splice site mutations, gene deletions) as well as variants that disrupt key functional domains (Inoue and Fry, 2017).

Oncogenes are implicated in cancer through activating mutations and often encode growth factors or cytokine receptors (e.g., *EGFR, JAK2, KIT, PDGFRA*), their downstream signalling mediators (e.g., *PIK3CA, BRAF, NRAS, KRAS*) or negative regulators of tumour suppressors (e.g., *PPM1D*) (Nangalia et al., 2016; Ruark et al., 2013; Stratton, 2011). The types of mutations that result in activation or upregulation of oncogenes are diverse and include canonical hotspot missense mutations (e.g. *JAK2* V617F, *BRAF* V600E), chromosomal translocations or gene amplifications as well as deletions or truncating mutations that disrupt inhibitory regulatory domains (e.g., truncating mutations in *PPM1D* exon 6, intragenic *BRAF* deletions)(Forbes et al., 2011; Ruark et al., 2013; Stratton, 2011; Wegert et al., 2018).

It is increasingly recognised that many cancer genes, particularly those implicated in epigenetic regulation, do not fit tidily into this classification scheme. Many function as either tumour suppressors or oncogenes in different cancer types or even at different stages of the same cancer type (e.g., *EZH2*), reflecting the influence of cell-type, developmental context

and epistasis on the functional significance of many cancer driver mutations (Feinberg et al., 2016; Kim and Roberts, 2016; Shen et al., 2018; Van Vlierberghe and Ferrando, 2012).

Haematological cancers, and acute myeloid leukaemia in particular, are among the most extensively sequenced and genomically well-characterised of all cancer types (Medinger and Passweg, 2017; TCGA et al., 2013). Hence, the landscape of tumour suppressor and oncogenes relevant to these conditions has been well charted and the types of mutations that appear to be under positive selection in these genes is reasonably well defined, with concordance between many large studies (Bahr et al., 2018; Chen et al., 2018; Medinger and Passweg, 2017; Petti et al., 2018; TCGA et al., 2013; Tyner et al., 2018). The experiments described in this dissertation have taken a conservative approach to driver curation based on the criteria described in the largest relevant cancer genomics to date (Chapter 2).

1.1.3 Germline contributions to cancer risk

Studies of familial cancer predisposition and rare childhood cancer syndromes identified some of the first known cancer genes (Knudson, 1971; Maris, 2015). Germline variation plays an increasingly recognised role in cancer development, though its impact likely remains underestimated (Frick et al., 2018; Hermouet and Vilaine, 2011; Hinds et al., 2016; Huang et al., 2018; Loh et al., 2018; Parsons et al., 2016; Zhang et al., 2015). According to current estimates, overall approximately 1-2.7% of individuals without cancer have a putatively deleterious germline mutation in a cancer-associated gene, compared with 8.5 – 12.6% of cancer patients (Pritchard et al., 2016; Schrader et al., 2016; Zhang et al., 2015), though this rate appears considerably higher for some rare cancer types (Ballinger et al., 2016; Lu et al., 2015). Germline variants can influence cancer development by diverse mechanisms, including by directly driving clonal growth (Loh et al., 2018; Lu et al., 2015), increasing global mutation rate (Nik-Zainal, 2014; Shlien et al., 2015), increasing the likelihood of acquiring particular somatic driver events (Hermouet and Vilaine, 2011; Hinds et al., 2016; Loh et al., 2018) or altering carcinogen metabolism (Ding et al., 2010).

Studies of cancer predisposition syndromes have also demonstrated that the biological and clinical significance of germline and somatic variants in a given gene are often dramatically different (Maris, 2015; Maris and Knudson, 2015). For example, childhood myeloproliferative disease with germline mutations in *PTPN11* may follow an indolent, self-

resolving course, whereas somatic *PTPN11* mutations presage rapid progression and warrant prompt haematopoietic stem cell transplantation (HSCT)(Hasle, 2016). Furthermore, germline and somatic mutations in several cancer genes, notably *TP53* and *RB1,* drive a distinct spectrums of cancer types with predilections for different tissues and age groups (Maris and Knudson, 2015). The distinction between germline and somatic drivers is particularly relevant when interpreting the results of unmatched sequencing experiments such as those described in this thesis, and will be discussed further later on.

1.1.4 Mutational signatures

The entire complement of somatic mutations in a genome constitutes a record of the types of mutational processes operative during the lifetime of the organism. Certain patterns of mutation are characteristic of particular mutagenic exposures. For example, ultraviolet light-induced pyrimidine dimers are typically repaired by transcription-coupled nucleotide excision repair, which tends to result in C>T mutations on the untranscribed strand (Alexandrov et al., 2013). Substitutions, small insertions and deletions (indels) and complex structural events can be classified according to sequence context, thus allowing formal mathematical extraction of mutational signatures (Alexandrov et al., 2013; Li et al., 2017; Petljak et al., 2014).

Substitution mutational signatures have been most extensively studied. The six types of substitution mutation (C>A, C>G, C>T, T>A, T>C and T>G) can be classified into 96 subtypes based on their trinucleotide context. Various statistical approaches, predominantly based on non-negative matrix factorisation, can discern distinct patterns of co-occurrence of substitution types (Alexandrov et al., 2018; Alexandrov et al., 2013). At present, only a minority of putative mutational signatures have a known cause (Alexandrov et al., 2018; Alexandrov et al., 2013). Nevertheless, mutational signature analysis has yielded compelling insights into the causes and epidemiology of several cancer types, and are increasingly being used clinically to guide diagnosis, prognostication and therapeutic strategy (Behjati et al., 2016; Hoang et al., 2013; Ma et al., 2018; Petljak and Alexandrov, 2016; Poon et al., 2015).

All cancers harbour a significant number of mutations attributed to ageing-associated single base substitution signatures 1 (SBS1) and 5 (SBS5) (Alexandrov et al., 2018; Alexandrov et al., 2013). SBS1 is dominated by C>T mutations attributed to spontaneous deamination of

5-methylcytosine, whilst SBS5 is of unknown aetiology (Alexandrov et al., 2018; Alexandrov et al., 2013). Myeloid malignancies are characterised by very low mutation burdens, similar to those observed in normal haematopoietic stem cells from age-matched individuals (Welch et al., 2012). Consistent with this finding, most of these mutations are attributable to SBS1 and SBS5 (Alexandrov et al., 2018; Alexandrov et al., 2013). A significant proportion of AML demonstrate evidence of SBS18, attributed to reactive oxygen species-mediated DNA damage (Alexandrov et al., 2018). A small proportion of myelodysplasia and myeloproliferative disease specimens harbour mutations attributable to SBS32, a signature thought to be caused by azathioprine treatment (Alexandrov et al., 2018). Although lymphoid neoplasms are also generally dominated by age-related SBS1 and SBS5 (Alexandrov et al., 2018; Alexandrov et al., 2013), they tend to have higher mutation burdens than myeloid cancers and a more complex mutational signature complement, with some specimens harbouring evidence of defective DNA repair mechanisms or APOBEC activity (Alexandrov et al., 2018; Alexandrov et al., 2013).

1.2 Cancer is an evolutionary process

The notion that cancer development is a clonal (originating from a single ancestral cell) evolutionary process can be traced back to Boveri and was further advanced in the 1950s based on histological observation of the natural history of precancerous lesions and their response to extrinsic irritants (Denoix, 1954; Foulds, 1958). Following the acceptance of the mutational theory of cancer, Peter Nowell and John Cairns conceptualised the modern understanding of cancer evolution in their seminal 1970s reviews (Cairns, 1975; Nowell, 1976).

"The acquired genetic instability and associated selection process, most readily recognized cytogenetically, results in advanced human malignancies being highly individual karyotypically and biologically. Hence, each patient's cancer may require individual specific therapy, and even this may be thwarted by emergence of a genetically variant subline resistant to the treatment. More research should be directed toward understanding and controlling the evolutionary process in tumors before it reaches the late stage usually seen in clinical cancer."

- Peter Nowell, 1976 (Nowell, 1976)

Cairns spoke more explicitly in terms of natural selection acting on inevitable mutations arising in stem cells throughout the lifespan of an organism:

"Survival of the rapidly renewing tissues of long-lived animals like man requires that they be protected against the natural selection of fitter variant cells (that is, the spontaneous appearance of cancer)."

- Cairns 1975 (Cairns, 1975)

The ability to sequence many specimens of the same tumour type demonstrated remarkable genetic diversity within the same histopathological diagnosis (Yates and Campbell, 2012). Phylogenetic inference, multi-region tumour sequencing and single cell methods revealed striking intra-tumour heterogeneity (Anderson et al., 2011; Gerlinger et al., 2012; Greaves, 2015; Navin et al., 2011). These observations established that the evolutionary routes to cancer are diverse and that malignant clones continue to acquire mutations, compete and evolve (Ding et al., 2012; Greaves and Maley, 2012; Nik-Zainal et al., 2012). It became possible to construct phylogenetic trees at unprecedented resolution. Consistent features of these trees illustrate key principles of cancer pathogenesis. At their base, all cancer phylogenetic trees have the ancestral cell with the initial complement of driver mutations, along with all other mutations previously acquired by that cell and captured as the clone expanded (Yates and Campbell, 2012). Each cell within the expanded clone continues to acquire mutations, which are subclonal. With a few exceptions (e.g., chromothripsis causing multiple simultaneous driver mutations (Stephens et al., 2011)), in almost all cases cancer phylogenies support the gradual, multi-step model of carcinogenesis (Greaves, 2015; Yates and Campbell, 2012). Tumour cells continually diversify through acquisition of additional mutations and clonal architecture may follow branching, parallel or convergent evolutionary trajectories (Greaves, 2015; Yates and Campbell, 2012). The relative influence of mutation-induced cell-intrinsic growth advantage, selective pressures and genetic drift in cancer evolution remains contentious (Martincorena and Campbell, 2015; Martincorena et al., 2017; Sun et al., 2017; Zink et al., 2017). Phylogenetic trees constructed from multi-region or serial sampling have yielded insights into some of the selection pressures implicated in cancer clonal competition, discussed briefly in the next section.

1.2.1 Selection pressures shaping cancer evolution

1.2.1.1 The tumour microenvironment

The idea that the tumour microenvironment influences cancer development was first put forward in the late 19th century by Ernst Fuchs and Stephen Paget based on detailed anatomical studies of tumour metastases (Fuchs, 1882; Paget, 1889). Paget likened tumour cells to 'seeds' that required a favourable microenvironment, or 'soil' to survive and grow (Paget, 1889). The factors underpinning the predilection of metastases for certain organs are still incompletely understood (Hunter et al., 2018). However, several studies that used multi-region sampling or tumour organoids have elucidated the phylogenetic relationships between primary tumour lesions and metastases and provided insight into the interplay between genetic diversification and organ-specific selection pressures (Altorki et al., 2019; Campbell et al., 2010; Gundem et al., 2015; Hunter et al., 2018; Makohon-Moore and Iacobuzio-Donahue, 2016; Roerink et al., 2018; Yachida et al., 2010). It is now clear that interactions between cancer cells and tissue microenvironment are relevant far beyond metastasis, exerting selective pressures important at all stages of solid and haematological cancer development (Medyouf, 2017; Scott and Gascoyne, 2014; Yates and Campbell, 2012; Yokoyama et al., 2019).

1.2.1.2 Cancer therapies

Anticancer therapy is often one of the most potent selective pressures governing cancer evolution (Yates and Campbell, 2012). Resistance mechanisms are diverse (Holohan et al., 2013), however, as sequencing technologies become more sensitive, it is increasingly clear that resistance mutations to both conventional cytotoxic agents and targeted therapies frequently predate treatment at extremely low subclonal levels (Karoulia et al., 2017; Kennedy et al., 2014; Schmitt et al., 2016; Wong et al., 2015a; Wong et al., 2015b). As presciently anticipated by Nowell (Nowell, 1976), the extensive genetic diversity present in fully fledged cancers represents a formidable arsenal of potential adaptive strategies and has greatly undermined targeted therapy efforts (Holohan et al., 2013).

Scrutiny of cancer genomes has yielded profound insight into the genetic drivers and evolutionary dynamics of most human cancer types. However, it is now evident that this work

did not adequately capture the somatic genetic diversity and selective pressures shaping the pre-cancerous phases of oncogenesis. Recent studies of somatic evolution in morphologically normal tissues have yielded compelling biological insights into normal ageing and its relationship with cancer development. The next section will give a broad overview of these advances with a focus on the haematopoietic system.

2. Somatic evolution in normal ageing tissues and its relationship to cancer

"*Cancer is a chronic disease with a long history extending back for many years before clinical signs are evident*."

- Leslie Foulds, 1958 (Foulds, 1958)

"…the whole body is seeded with tumor cells whose evolutionary potential is revealed at unpredictable times thereafter."

- Foulds's summary of a hypothesis proposed by Pierre Denoix in his 1954 paper 'De la diversité de certains cancers' (Denoix, 1954; Foulds, 1958)

The molecular basis of multi-step carcinogenesis was meticulously dissected in childhood leukaemia and colon cancer in the 1980s and 1990s and gave preliminary insights into the ambiguous boundary between normal tissue, pre-cancer and fully-fledged malignancy (Fearon and Vogelstein, 1990; Greaves et al., 2003). Studies of monozygotic twins concordant for leukaemia demonstrated that the initiating event, typically a fusion gene, arises in a single cell *in utero*, which transfers to the second twin via a monochorionic placenta (Greaves and Wiemels, 2003). For most childhood leukaemia, the latency to disease onset suggested that the initiating translocation (most commonly the *TEL–AML1* fusion gene), requires a second hit to trigger malignant transformation (Greaves and Wiemels, 2003). In support of this hypothesis, several studies screened healthy newborns for leukaemogenic fusions and found their prevalence to be considerably higher than the cumulative incidence of childhood leukaemia (Greaves et al., 2011; Lausten-Thomsen et al., 2011; Mori et al., 2002; Zuna et al., 2011). Furthermore, not all twins concordant for the initiating event are concordant for

leukaemia (Bateman et al., 2015). Collectively, these findings provided genetic evidence to support Foulds and Denoix's hypothesis that pre-cancer is considerably more common than cancer and that malignant progression is not readily predictable. These conclusions were also supported by the natural history and molecular features of the adenoma-carcinoma sequence in the colon (Fearon and Vogelstein, 1990).

The advent of sensitive sequencing methods has recently revealed that potentially pre-malignant clonal expansions are remarkably common in many normal ageing tissues (Bowman et al., 2018; Martincorena et al., 2018; Martincorena et al., 2015; Moore et al., 2018; Salk et al., 2018; Suda et al., 2018; Yizhak et al., 2018; Yokoyama et al., 2019). This phenomenon has been most extensively explored in skin (Martincorena et al., 2015), oesophagus (Martincorena et al., 2018; Yokoyama et al., 2019), endometrium (Moore et al., 2018; Salk et al., 2018; Suda et al., 2018) and blood (Bowman et al., 2018), though preliminary evidence from bulk RNA sequencing of diverse normal tissues suggests that clonal expansions harbouring canonical cancer driver mutations may be ubiquitous in most organs (Yizhak et al., 2018).

Several common themes are beginning to emerge from these findings. Firstly, there is generally a clear association between age and prevalence of readily detectable clonal expansions, with that latter apparently trending towards inevitability by midlife in many tissues (Martincorena et al., 2018; Martincorena et al., 2015; Suda et al., 2018; Young et al., 2016). However, it is not yet clear to what extent age-related mutation acquisition is a rate-limiting step in clonal expansion. Potent cancer driver mutations, including hotspot *TP53* mutations, may be dated to early infancy or childhood in several tissues and may never contribute to cancer even in high risk individuals (Greaves et al., 2011; Moore et al., 2018; Yokoyama et al., 2019). It is increasingly apparent that selective pressures, some correlated with ageing, impact the fitness advantage of particular mutations and hence modulate clonal dynamics (Hsu et al., 2018; McKerrell and Vassiliou, 2015; Murai et al., 2018; Wong et al., 2015b; Yokoyama et al., 2019). For example, exposure to smoking and alcohol accelerates clonal growth in normal oesophagus (Yokoyama et al., 2019) and ultraviolet radiation exposure influences the fitness advantage of epidermal *TP53* mutations (Murai et al., 2018). The proliferation of clonal expansions with age may reflect both mutation accrual and ageing-associated changes in tissue microenvironments that confer increasing fitness advantage on oncogenic mutations (Armitage and Doll, 1954; Nordling, 1953; Rozhok and DeGregori, 2015).

A second observation that has been made in several tissue types is that the mutational spectrum of age-associated clonal expansions may differ from that seen in cancer (Busque et al., 2018; Martincorena and Campbell, 2015; Martincorena et al., 2018; Xie et al., 2014; Yokoyama et al., 2019). For example, putative driver mutations in *NOTCH1* are more frequently seen in clonal expansions in histologically normal skin and oesophagus than in cancers arising from these tissues (Martincorena et al., 2018; Martincorena et al., 2015; Yokoyama et al., 2019). Similarly, activating mutations in *PPM1D*, which encodes a negative regulator of TP53, are more frequent in normal blood and oesophagus than in malignancy (Bowman et al., 2018; Xie et al., 2014; Yokoyama et al., 2019). Most relevant experiments have employed targeted sequencing of known cancer-associated genes, thus hindering an unbiased comparison between the mutational landscape of cancer and normal ageing. Equally, the ubiquity of certain mutations in normal tissues, and by extension their recurrence in the trunks of tumour phylogenetic trees, could lead to overestimates of their importance in cancer pathogenesis (Ciccarelli, 2019).

How mutations and selective pressures interact to determine the likelihood of malignant transformation is an important biological question with compelling clinical implications. As predicted by Cairns (Cairns, 1975), emerging evidence suggests that some epithelial tissues have evolved mechanisms for restraining growth of clones harbouring oncogenic mutations (Murai et al., 2018; Ying et al., 2018). Senescence and immune surveillance are also involved in policing mutated clones (Collado et al., 2005; Schreiber et al., 2011). However, understanding of the factors governing physiological cell competition and tissue homeostasis in humans and their relationship with carcinogenesis remains very limited. A significant obstacle to studying these questions in most organs is the inability to obtain representative tissue samples. The haematopoietic system has proven a privileged setting in which to explore somatic evolution and its relationship with ageing and ageing-associated pathologies (Bowman et al., 2018; Geiger et al., 2013; Latchney and Calvi, 2017; Lee-Six et al., 2018). The next section will summarise current understanding of clonal haematopoiesis and its clinical relevance.

3. Clonal haematopoiesis

3.1 Prevalence and mutational landscape of clonal haematopoiesis

Blood has one of the highest turn-over rates of any tissue, necessitating the production of trillions of cells per day by a population of haematopoietic stem cells (HSCs) estimated to number between 50,000 and 200,000 (Carrelha et al., 2018; Doulatov et al., 2012; Lee-Six et al., 2018). Replicative mutagenesis and other sources of genotoxic stress cause HSCs to accumulate DNA damage with age, with an estimated 14 mutations accumulating per cell per year (Flach et al., 2014; Osorio et al., 2018; Rossi et al., 2007; Welch et al., 2012; Yahata et al., 2011). Clonal haematopoiesis (CH) refers to the disproportionate expansion of one somatically mutated HSC clone relative to others. Many reports have now identified this phenomenon in a significant proportion of individuals without a haematological cancer (Acuna-Hidalgo et al., 2017; Akbari et al., 2014; Artomov et al., 2017; Bonnefond et al., 2013; Buscarlet et al., 2017; Busque et al., 1996; Busque et al., 2012; Coombs et al., 2017; Forsberg et al., 2012; Frick et al., 2018; Genovese et al., 2014; Gibson et al., 2017; Gillis et al., 2017; Jacobs et al., 2012; Jaiswal et al., 2014; Jaiswal et al., 2017; Laurie et al., 2012; Loftfield et al., 2018b; Loh et al., 2018; Machiela et al., 2015; McKerrell et al., 2015; Rodriguez-Santiago et al., 2010; Savola et al., 2017; Schick et al., 2013; Takahashi et al., 2017; Thompson et al., 2019; Vattathil and Scheet, 2016; Xie et al., 2014; Young et al., 2016; Zhou et al., 2016; Zink et al., 2017). Clonal haematopoiesis was first recognised in the 1990s when Busque and colleagues demonstrated that ageing was associated with increasingly skewed X-inactivation in blood cells (Busque et al., 1996). Busque et al. applied a PCR-based X-inactivation clonality assay to peripheral blood samples from a cohort of 295 healthy females spanning a broad age range (Busque et al., 1996). Using stringent criteria for skewing (allele ratios >= 10:1), this approach identified imbalanced X-inactivation in 22.7%, 4.5% and 1.9% of women aged >=60 years, 28-32 years and <1 month, respectively (Busque et al., 1996).

The advent of molecular karyotyping using SNP arrays demonstrated that a significant proportion of the general population harbours clonal, somatic chromosomal abnormalities in blood cells (Artomov et al., 2017; Bonnefond et al., 2013; Forsberg et al., 2012; Jacobs et al., 2012; Laurie et al., 2012; Loftfield et al., 2018a; Loh et al., 2018; Machiela et al., 2015;

Rodriguez-Santiago et al., 2010; Schick et al., 2013; Vattathil and Scheet, 2016; Zhou et al., 2016). These studies identified a clear correlation between age and frequency of clonal mosaic aneuploidy or copy-neutral loss of heterozygosity (LOH) events, with prevalence varying from <0.5% in individuals under age 50 years to 1.9-3.4% in persons aged >60 (Forsberg et al., 2012; Jacobs et al., 2012; Laurie et al., 2012). The most recurrent abnormalities included del(13q), trisomy 8, del(20q), del(5q) and del(7q), chromosomal changes characteristic of haematological malignancies (Forsberg et al., 2012; Jacobs et al., 2012; Laurie et al., 2012). Mosaic chromosomal changes were associated with a five- to ten-fold higher risk of subsequently developing haematological cancers (Jacobs et al., 2012; Laurie et al., 2012; Schick et al., 2013). Longitudinal tracking of clonal chromosomal abnormalities has yielded variable results, with one study suggesting that aberrant clones may become undetectable over time (Forsberg et al., 2017), while another series of 47 individuals sampled several years apart found that most clones expanded with age (Machiela et al., 2015).

Next-generation sequencing technologies enabled higher resolution scrutiny of the genetic changes driving clonal haematopoiesis. Sequencing of healthy women with skewed X-inactivation identified mutations in the epigenetic regulator *TET2* in 5.5% (10/182 individuals) (Busque et al., 2012). In 2014, three large exome sequencing studies identified leukaemia-associated point mutations in the blood of >2% of individuals unselected for haematological phenotypes (Genovese et al., 2014; Jaiswal et al., 2014; Xie et al., 2014). All three studies reported a steep rise in CH prevalence with age, ranging from <1% under age 50 years to around 10% in individuals over age 70 (Genovese et al., 2014; Jaiswal et al., 2014; Xie et al., 2014). The majority of candidate driver mutations occurred in *TET2*, *DNMT3A* and *ASXL1*, epigenetic regulators commonly mutated in myeloid malignancies (Arber et al., 2016; Genovese et al., 2014; Jaiswal et al., 2014; Xie et al., 2014). Jaiswal et al. interrogated a predefined set of 160 cancer-associated genes, whereas Genovese et al. and Xie et al. screened for CH in an unbiased manner on the basis of unusual allele frequencies (Genovese et al., 2014; Xie et al., 2014). The latter approach identified a broader spectrum of putative CH drivers, most notably a remarkably high frequency of mutations in *PPM1D*, a negative regulator of TP53 that is infrequently mutated in haematological or solid cancers (Genovese et al., 2014; Ruark et al., 2013; Xie et al., 2014). Other recurrently mutated genes included *JAK2, TP53*, spliceosome genes (*SF3B1, SRSF2* and *U2AF1*), *CBL, BCORL1, ATM, MYD88* and *GNAS* (Genovese et al., 2014; Jaiswal et al., 2014; Xie et al., 2014).

Later studies of CH in the general population used more sensitive targeted sequencing approaches and demonstrated that CH prevalence increases dramatically with assay sensitivity (Acuna-Hidalgo et al., 2017; Buscarlet et al., 2017; McKerrell et al., 2015; Young et al., 2016; Zink et al., 2017). Young et al. used molecular barcoding to enable detection of mutations at a variant allele frequency (VAF) as low as 0.0003 and found CH to be ubiquitous in otherwise healthy individuals aged >50 years (Young et al., 2016). The genes recurrently implicated in CH were broadly consistent across these studies. However, whilst the prevalence of mutations in all genes increased with age, certain mutations were found to be particularly enriched in older individuals (McKerrell et al., 2015). In particular, spliceosome gene mutations were seen almost exclusively in individuals aged >70 (Acuna-Hidalgo et al., 2017; McKerrell et al., 2015), whereas the frequency of mutations in *DNMT3A* and *JAK2* increased more linearly with age (Acuna-Hidalgo et al., 2017; Buscarlet et al., 2017; McKerrell et al., 2015). A less dramatic age-dependence has been observed for *TET2* mutations (Buscarlet et al., 2017).

Ageing is just one example of how the mutational landscape of CH varies according to clinical context. CH is extremely common in aplastic anaemia patients and displays a distinct spectrum of somatic mutations (Stanley et al., 2017; Yoshizato et al., 2015). Similarly, CH enriched in *TP53* and *PPM1D* mutations is prevalent in individuals who have been exposed to chemo- and/or radiotherapy (Coombs et al., 2017; Gibson et al., 2017; Gillis et al., 2017; Takahashi et al., 2017). Further discussion of the interplay between somatic mutations and dynamic selection pressures is discussed in section 3.4.

Zink et al. conducted a broader, though less sensitive, screen for CH by interrogating 11,262 whole genomes (median coverage 35x) for unusual SNV allele frequency distribution, similar to the variant calling strategies applied by Xie et al. and Genovese et al. (Genovese et al., 2014; Xie et al., 2014; Zink et al., 2017). Consistent with previous data and predictions, CH was almost universally detectable in individuals >85 years of age (McKerrell et al., 2015; Young et al., 2016; Zink et al., 2017). The overall prevalence of CH (identified on the basis of having > 20 putative mosaic point mutations) was 12.5%, higher than that observed in previous studies (Zink et al., 2017). Presumptive driver mutations were most frequent in *DNMT3A*, *TET2*, *ASXL1* and *PPM1D* (Zink et al., 2017). However, candidate driver mutations were only identified in a minority of individuals with CH (Zink et al., 2017). The authors suggest genetic drift as a likely explanation for this result. However, numerical and structural

chromosomal changes were not systematically identified and may account for a significant proportion of the CH cases without an apparent point mutation driver (Artomov et al., 2017; Bonnefond et al., 2013; Forsberg et al., 2012; Jacobs et al., 2012; Laurie et al., 2012; Loftfield et al., 2018a; Loftfield et al., 2018b; Loh et al., 2018; Machiela et al., 2015; Rodriguez-Santiago et al., 2010; Schick et al., 2013; Vattathil and Scheet, 2016; Zhou et al., 2016). Contiguous gene deletions and rearrangements are common initiating driver events in many haematological cancers. It is possible that structural variants under positive selection underpinned a significant proportion of the CH cases attributed to drift. It is also conceivable that there is only partial overlap between cancer drivers and the mutations that are under positive selection in somatic evolution in normal ageing blood. The preponderance of *PPM1D* and *NOTCH1* mutations in clonal expansions in normal tissues compared to cancers may support this hypothesis (Bowman et al., 2018; Martincorena et al., 2018; Martincorena et al., 2015; Yokoyama et al., 2019). Zink et al. did perform an unbiased search for novel driver genes, but did not identify many candidates (Zink et al., 2017).

Mutations in certain common myeloid cancer genes, notably *FLT3* and *NPM1*, were consistently absent in even the most sensitive CH screens, supporting their role as late cooperating/transforming mutations rather than initiating events (Acuna-Hidalgo et al., 2017; Genovese et al., 2014; Jaiswal et al., 2014; McKerrell et al., 2015; Xie et al., 2014).

3.2 Germline influences on CH

Extensive evidence demonstrates that germline variation is an important determinant of clonal haematopoiesis risk and clinical outcome (Buscarlet et al., 2017; Frick et al., 2018; Hinds et al., 2016; Jones et al., 2009; Kilpivaara et al., 2009; Koren et al., 2014; Loftfield et al., 2018a; Loh et al., 2018; Olcaydu et al., 2009; Thompson et al., 2019; Wright et al., 2017; Zhou et al., 2016; Zink et al., 2017). Heritable polymorphisms can influence CH development by increasing susceptibility to somatic mutagenesis (Hinds et al., 2016; Jones et al., 2009; Kilpivaara et al., 2009; Koren et al., 2014; Loh et al., 2018; Olcaydu et al., 2009; Zhou et al., 2016) or by modulating positive or negative clonal selection (Hinds et al., 2016; Loh et al., 2018). For example, the *JAK2* 46/1 haplotype is a well-recognised risk factor for acquiring *JAK2* V617F-positive CH and progressing to a myeloid neoplasm (Jones et al., 2009; Kilpivaara et al., 2009; Olcaydu et al., 2009). Polymorphisms in several other genes, including *TERT, TET2, ATM* and *CHEK2*, are also associated with *JAK2* V617-driven myeloproliferative neoplasms and

hence perhaps also antecedent clonal haematopoiesis (Hinds et al., 2016). Over 150 loci have now been strongly linked to overall CH risk, or risk of particular chromosomal losses or likelihood of specific LOH events amplifying the selective advantage conferred by inherited or somatic driver events (Loh et al., 2018; Thompson et al., 2019; Wright et al., 2017; Zink et al., 2017). Additionally, several germline polymorphisms have been shown to impact leucocyte DNA replication timing, and by consequence, the susceptibility of nearby sequence to somatic mutagenesis (Koren et al., 2014). In a recent large survey of mosaic chromosomal changes in peripheral blood, Loh et al. identified several highly penetrant heritable variants associated with increasing mutability of nearby DNA sequence, including in the myeloid oncogene *MPL* (Loh et al., 2018). Several of the variants were also subject to clonal selection and impacted risk of progression to haematological cancer (Loh et al., 2018).

A main emerging message from these studies is the increasingly blurry distinction between heritable and somatically acquired determinants of clonal haematopoiesis development and natural history. Furthermore, the influence of germline variation on CH incidence and outcome probably remains underestimated. Several studies report familial or ethnic clustering of CH suggesting yet to be discovered heritable risk factors (Buscarlet et al., 2017; Frick et al., 2018; Loftfield et al., 2018a). Moreover, a large number of uncommon germline variants have emerged as important determinants of haematological phenotypes in the general population, and it is plausible that these exert epistatic, lineage biased effects on CH evolution (Astle et al., 2016).

3.3 Clinical significance of clonal haematopoiesis

3.3.1 Impact of clonal haematopoiesis on blood indices

Mutations common in CH are implicated in ineffective haematopoiesis, impaired differentiation and cytopenias when they occur in individuals with MDS or AML (Papaemmanuil et al., 2016; Steensma et al., 2015). However, CH harbouring putative driver mutations (CH-PD) is not generally associated with any abnormalities in blood cell counts (Buscarlet et al., 2017; Jaiswal et al., 2014; McKerrell et al., 2015). Jaiswal et al. analysed blood indices data available for 3107 individuals, 4.5% of whom had CH-PD and found no significant differences in haemoglobin levels, platelet counts or white-cell differential counts (Jaiswal et

al., 2014). The only blood index that differed significantly was red cell distribution width (RDW), which was higher in individuals with CH-PD and correlated with mutation VAF (Jaiswal et al., 2014). Moreover, although the prevalence of a single cytopenia was not influenced by CH status, individuals with multiple cytopenias were more likely to have CH (odds ratio 3.0)(Jaiswal et al., 2014).

While CH may rarely cause haematological indices to deviate to a clinically significant degree, Loh et al. recently demonstrated that some acquired mutations correlate with trends in blood counts, though generally within the reference range (Loh et al., 2018). Their findings suggest lineage-specific clonal selection pressures mirroring those observed in blood cancers (Loh et al., 2018). For example, chromosome 9p LOH (encompassing *JAK2*) and trisomy 12 (highly recurrent in CLL) were associated with higher granulocyte and lymphocyte counts, respectively (Loh et al., 2018).

3.3.2 Clonal haematopoiesis and haematological malignancy

Numerous studies have reported a clear association between CH in haematologically normal individuals and risk of developing a haematological malignancy (Coombs et al., 2017; Genovese et al., 2014; Gibson et al., 2017; Gillis et al., 2017; Greaves and Wiemels, 2003; Jacobs et al., 2012; Jaiswal et al., 2014; Laurie et al., 2012; Loh et al., 2018; Schick et al., 2013; Takahashi et al., 2017; Zink et al., 2017). This is perhaps unsurprising given that the multi-step model of cancer implies a premalignant phase in cancer evolution (Yates and Campbell, 2012). Furthermore, several studies of haematological cancer evolution have demonstrated that myeloid malignancies evolve from a population of preleukaemic stem cells harbouring initiating driver mutations, and that such preleukaemic HSCs can persist during long-term remission and serve as a reservoir for relapse (Greaves et al., 2003; Jan et al., 2012; Shlush et al., 2017; Shlush et al., 2014). Similar observations hold true for the commonest lymphoid malignancies (Landgren et al., 2009; Ojha et al., 2014; Rawstron et al., 2008). However, the prevalence of preleukaemic HSC clones and the rate and determinants of progression to leukaemia remain unknown. The studies cited above demonstrate that the rate of CH in the general population, and in particular CH harbouring putative driver mutations (CH-PD), vastly exceeds the cumulative incidence of blood cancers (Bowman et al., 2018). Given the variation in cohort characteristics, follow-up time and CH detection sensitivity, it is unsurprising that

the strength of the association reported between CH and haematological cancer risk has varied between studies (Coombs et al., 2017; Genovese et al., 2014; Gibson et al., 2017; Gillis et al., 2017; Greaves and Wiemels, 2003; Jacobs et al., 2012; Jaiswal et al., 2014; Laurie et al., 2012; Loh et al., 2018; Schick et al., 2013; Takahashi et al., 2017; Zink et al., 2017). Notably, Zink et al. and Genovese et al. found that the risk of malignant progression was the same regardless of whether a point mutation driver (versus no driver) was identified (Genovese et al., 2014; Zink et al., 2017). However, as discussed previously, it is possible that CH without such mutations may reflect unsought structural driver events.

Most studies of cohorts unselected for cancer or haematological phenotype have reported an approximately ten-fold increased risk of blood cancer among individuals with CH (Genovese et al., 2014; Jaiswal et al., 2014). However, this still reflects a low absolute risk for malignant progression. Jaiswal et al. found that individuals with CH-PD (assay sensitivity limit 3.5% and 7.0% for SNVs and indels, respectively) had a 4% risk of blood cancer diagnoses over a median follow-up period of 7.9 years (Jaiswal et al., 2014). This translates into an overall annual progression rate of 0.5%, rising to 1% per year among individuals with driver mutations present at VAF > 0.1 (Jaiswal et al., 2014). Similarly, Genovese et al. reported similar findings, and in addition were able to demonstrate a clonal relationship between CH and blood cancer in the two individuals for whom diagnostic bone marrow specimens were available (Genovese et al., 2014). In both of these cases, the interval between blood sampling and cancer diagnosis was modest (2 and 34 months) (Genovese et al., 2014). Both Jaiswal et al. and Genovese et al. found that only a minority of the blood cancers arising during follow-up were diagnosed in individuals with antecedent CH: 5/16 (31%) and 13/31 (42%), respectively (Genovese et al., 2014; Jaiswal et al., 2014). This finding, in conjunction with the ubiquity of CH relative to blood cancer incidence, raises clinically and biologically compelling questions about the natural history of haematological cancers and the pathophysiological relevance of CH.

From a clinical perspective, it is sobering that the main cause of mortality from many of the commonest adult haematological cancers remains treatment resistance, despite a growing arsenal of novel targeted therapies (Abdi et al., 2013; Döhner et al., 2015; Woyach and Johnson, 2015). There is hence a compelling rationale for identifying and treating a genomically simpler antecedent of the disease. In this context, reduction of clonal size rather than complete clonal extinction may be sufficient to significantly reduce the risk of malignant progression. Such an approach has proven very effective in CML, which has been transformed

into a chronic condition by targeted therapy, whereas CML blast crisis remains very challenging to treat (Gore et al., 2018; Hunger, 2017; O'Brien et al., 2003). The eventual feasibility of earlier detection and intervention for nascent blood cancers will invariably be hampered by the high prevalence of benign CH, given the relative rarity of the former. However, CH is associated with and may play a causal role in several much commoner conditions, which may broaden indications for its use as a clinical biomarker or a therapeutic target for non-haematological pathologies. The broader clinical significance of CH is summarised in the following sections.

3.3.3 Clonal haematopoiesis and non-haematological cancers

Clonal haematopoiesis has been associated with both a higher risk of solid cancers (Akbari et al., 2014; Artomov et al., 2017; Bowman et al., 2018; Ruark et al., 2013; Thompson et al., 2019) and with higher mortality among solid tumour and lymphoma patients (Coombs et al., 2017; Gibson et al., 2017). However, it is challenging to study the relationship between CH and solid cancer risk given that cancer treatments dramatically increase CH incidence and many study participants were not chemotherapy/radiotherapy naïve (Akbari et al., 2014; Artomov et al., 2017; Ruark et al., 2013). It is also possible that germline cancer predisposition is a confounding risk factor for both CH and overall cancer risk.

The association between CH and mortality among cancer patients has been consistently observed across diverse cohorts (Coombs et al., 2017; Gibson et al., 2017; Gillis et al., 2017), though may also be subject to some confounding factors, e.g., germline cancer predisposition. Furthermore, cancer treatment intensity correlates with CH risk (Coombs et al., 2017; Gibson et al., 2017; Gillis et al., 2017; Takahashi et al., 2017) and toxicity-related mortality, and may be higher in individuals with more advanced malignancies. These potential confounders are hard to control for across retrospective cohorts comprising individuals with diverse solid cancer types.

Any mechanistic link between CH and solid tumour pathogenesis remains speculative. It is possible that clonal haematopoiesis may promote solid tumour growth by fostering hospitable tissue microenvironments (Bowman et al., 2018). The term ‘tumour-associated macrophage’ (TAM) encompasses phenotypically diverse cells that can play both oncogenic or tumour-suppressive roles (Mantovani et al., 2017). It is intriguing that the cytokine profile

of the *TET2*-mutated macrophages implicated in atherosclerosis (Jaiswal et al., 2017) shares key features with that seen in oncogenic TAMs (Storr et al., 2017; Wang et al., 2018).

3.3.4 Clonal haematopoiesis and non-malignant conditions

Several studies have found that clonal haematopoiesis is associated with a higher overall mortality rate that is only partially due to cancer deaths (Coombs et al., 2017; Genovese et al., 2014; Gibson et al., 2017; Jaiswal et al., 2014; Loftfield et al., 2018a; Loh et al., 2018; Zink et al., 2017). The majority of excess mortality has been attributed to cardiovascular disease (CVD), ischaemic stroke and diabetes (Bonnefond et al., 2013; Coombs et al., 2017; Fuster et al., 2017; Genovese et al., 2014; Gibson et al., 2017; Jaiswal et al., 2014; Jaiswal et al., 2017; Loftfield et al., 2018a; Sano et al., 2018a; Sano et al., 2018b). Preliminary evidence also links CH with rarer inflammatory conditions, such as rheumatoid arthritis (Savola et al., 2017).

It has long been recognised that known cardiovascular risk factors - namely hypertension, lipid profile, smoking and obesity – only partially account for atherosclerotic diseases burden and that other poorly characterised pro-inflammatory processes likely contribute (Ross, 1999). A large prospective case-control study recently confirmed the association between CH and risk of coronary heart disease, independent of age and other known risk factors (Jaiswal et al., 2017). This association held regardless of whether CH harboured mutations in *DNMT3A, TET2, JAK2* or *ASXL1* (Jaiswal et al., 2017). Individuals with CH had significantly more coronary artery calcification, a surrogate marker of atherosclerosis severity (Jaiswal et al., 2017). Moreover, compelling evidence now supports a causal role for CH in atherosclerosis and cardiometabolic disease (Fuster et al., 2017; Jaiswal et al., 2017; Sano et al., 2018a). Jaiswal et al. engrafted *TET2*-mutated cells into hypercholesterolaemia-prone mice and found that the TET2-deficient animals developed accelerated atherosclerotic disease (Jaiswal et al., 2017). Transcriptional profiling of *TET2*-mutant macrophages from arterial plaques revealed increased expression of pro-inflammatory mediators implicated in atherosclerosis, including CXCL1, CXCL2, IL-1b and IL-6 (Jaiswal et al., 2017). These findings were corroborated by a similar mouse model study by Fuster et al., which further demonstrated that inhibition of IL-1b secretion was more effective in slowing atherosclerosis in mice engrafted with TET2-deficient bone marrow than in controls (Fuster et al., 2017). Sano

et al. found that *TET2*-mutant CH increases IL-1b levels, accelerates cardiac failure in mice, and can be mitigated with anti-inflammatory therapy targeting IL-1b production (Sano et al., 2018a). A recent randomised, double blind trial of canakinumab, a therapeutic monoclonal antibody targeting IL-1b, reduced cardiovascular morbidity and mortality in humans independent of lipid profile (Ridker et al., 2017). Trial participants were not screened for CH, so it remains to be investigated whether CH could serve as a useful human biomarker or therapeutic target in its own right.

Myeloproliferative diseases are associated with increased cardiovascular morbidity and mortality mediated by multiple mechanisms (Deininger et al., 2017). In a retrospective nested case-control study including 10,000 individuals without a known myeloid neoplasm, *JAK2*-mutant CH was associated with an increased thrombosis risk (Wolach et al., 2018). This association appears at least partially attributable to a mutant *JAK2*-mediated increase in pro-thrombotic neutrophil extracellular trap (NET) formation (Wolach et al., 2018). In a mouse model of *JAK2*-mutant CH, NET formation and thrombosis was reduced upon administration of ruxolitinib, a *JAK2* inhibitor (Wolach et al., 2018).

It is not yet known whether CH with mutations in other genes plays a causative role in atherosclerosis, though the strong association between *DNMT3A*- and *ASXL1*-mutant CH and CVD (Jaiswal et al., 2017) warrants further investigation. It is intriguing that atherogenic haemodynamic stress appears to reprogram endothelial gene expression via a DNA methyl-transferase (DNMT)-dependent mechanism and that DNMT inhibition with siRNA or decitabine can reduce vascular endothelial inflammation and atherosclerosis formation in multiple mouse models (Dunn et al., 2014; Zhou et al., 2014). It is therefore possible that *DNMT3A*-mutant CH promotes endothelial dysfunction by epigenetic mechanisms, and might conceivably be amenable to nucleoside analogue treatment.

The hypothesis that CH can contribute to inflammatory conditions is further substantiated by a recent study investigating the impact of donor CH on allogeneic haematopoietic stem cell transplantation (HSCT) outcomes (Frick et al., 2018). Frick et al. found that recipients of CH-positive transplants had a significantly higher rate of chronic graft versus host disease and lower rate of relapse (Frick et al., 2018).

Collectively, these studies suggest a causal link between CH and non-malignant conditions, including leading causes of morbidity and mortality in the general population. It

is therefore possible that CH may prove to be a useful biomarker and/or modifiable risk factor in a range of clinical contexts.

3.4 Selection pressures influencing clonal haematopoiesis

Which selective pressures influence somatic evolution in the haematopoietic system? Do certain driver events confer strong enough cell-intrinsic growth advantage that they render clonal expansion inevitable? To what extent do environmental selection pressures determine the fitness advantage conferred by mutations and the pathophysiological outcome of CH? Are any of these selective pressures clinically modifiable? Although these questions remain largely unanswered, it is clear that the incidence and natural history of CH is influenced by clinical context.

3.4.1 Ageing

CH prevalence consistently rises with age, which is itself the dominant risk factor for most haematological malignancies (Busque et al., 2018). Haematopoietic ageing is characterised by HSC functional decline and myeloid bias reflected in a tendency towards anaemia and innate and adaptive immune senescence (Pang et al., 2011; Rossi et al., 2007; Rossi et al., 2005). Although HSCs accumulate mutations throughout life, ageing is associated with accelerated accrual of DNA damage (Flach et al., 2014; Osorio et al., 2018; Rossi et al., 2007; Welch et al., 2012). Age-associated genotoxic stress can induce apoptosis or differentiation, thus potentially depleting the functional HSC pool (Adams et al., 2015; Flach et al., 2014; Geiger et al., 2013; Rossi et al., 2007; Yahata et al., 2011). These factors may create an environment where HSCs with greater proliferative capacity or resistance to DNA-damage induced apoptosis and/or terminal differentiation contribute disproportionately to haematopoiesis (Latchney and Calvi, 2017; Pang et al., 2017). Mutations in many recurrent CH drivers, notably *DNMT3A, ASXL1* and *TET2*, may confer a competitive advantage through their ability to increase HSC self-renewal and inhibit differentiation (Abdel-Wahab et al., 2012; Challen et al., 2011; Dominguez et al., 2018; Jeong et al., 2018; Ko et al., 2011; Moran-Crusio, 2011). Similarly, HSC harbouring mutations in *TP53* or *PPM1D* are likely to have a particular competitive advantage in the context of genotoxic stress (Bondar and Medzhitov, 2010; Hsu et al., 2018; Kahn et al., 2018; Wong et al., 2015b).

3.4.2 Cytotoxic therapies

Studies of CH in cohorts of cancer patients who have received intensive chemo and/or radiotherapy have demonstrated an elevated prevalence of CH with marked enrichment for *PPM1D* and *TP53* mutated clones (Akbari et al., 2014; Coombs et al., 2017; Gibson et al., 2017; Gillis et al., 2017; Ruark et al., 2013; Takahashi et al., 2017). These findings suggest that exogenous genotoxic stress confers a strong competitive advantage on HSCs harbouring mutations that interfere with the DNA-damage response and apoptosis. In vivo studies of murine HSC competition have demonstrated that cells with *TP53* or *PPM1D* mutations outcompete their wild-type peers in the context of ionising radiation and chemotherapy, respectively (Bondar and Medzhitov, 2010; Hsu et al., 2018; Kahn et al., 2018). CH arising in the context of cancer treatment and its relationship with therapy-related myeloid neoplasms is further discussed in the introduction to chapter 5.

3.4.3 Immune-mediated selection

CH is particularly common in the context of bone marrow failure syndromes (Mehta et al., 2010; Reina-Castillon et al., 2017; Stanley et al., 2017; Yoshizato et al., 2015), corroborating the notion that HSC functional decline and depletion promotes cell competition. CH arising in the context of autoimmune-mediated acquired aplastic anaemia (AA) is another example of environmental context influencing HSC somatic evolution (McKerrell and Vassiliou, 2015; Yoshizato et al., 2015). CH is present in the majority of AA patients, and the mutational spectrum reflects the selective pressure exerted by immune attack on HSCs (Stanley et al., 2017; Yoshizato et al., 2015). For example, mutations in *PIGA* are highly recurrent and result in reduced cell surface expression of glycophosphotidylinositol-anchored autoantigens (McKerrell and Vassiliou, 2015; Yoshizato et al., 2015). Deletion of chromosome 6p, which encompasses human leucocyte antigen alleles, is likely to further aid immune escape (Stanley et al., 2017).

4. Sequencing strategies for studying somatic evolution

High resolution insight into somatic evolution in normal ageing tissues requires detection of rare mutations and represents a considerable technical challenge. The Illumina sequencing platform currently has the lowest error rate, though this varies considerably across different genomic regions according to the GC content and other base composition features (Hoang et al., 2016; Ross et al., 2013). With sophisticated post-sequencing analysis techniques, mutations in less error-prone genomic regions can be detected with a sensitivity >0.1%, though this is still inadequate for detecting rare mutations in cells that have not undergone appreciable clonal expansion (Gerstung et al., 2014; Hoang et al., 2016; Martincorena et al., 2015; Ross et al., 2013).

Strategies for overcoming this challenge include growing single-cell derived colonies (Lee-Six et al., 2018) or organoids (Blokzijl et al., 2016; Roerink et al., 2018), laser capture microdissection of clonal units from tissue sections (Moore et al., 2018), single cell sequencing (Navin et al., 2011; Potter et al., 2013; Zong et al., 2012) and error-corrected sequencing using molecular barcodes (Kennedy et al., 2014; Kinde et al., 2011; Mattox et al., 2017). The latter method involves using barcoded adaptors to label both strands from a single DNA molecule. This manoeuvre greatly helps distinguish artefacts (which will almost always be called on one strand only) from real mutations (apparent in both strands from the same DNA molecule)(Kennedy et al., 2014; Schmitt et al., 2012). However, error-corrected sequencing is tractable only for very limited target regions and can be insensitive, in part due to inefficient pull-down of target regions (Kennedy et al., 2014; Schmitt et al., 2012). It is also more labour-intensive and expensive due to the need to sequence each individual molecule sufficiently deeply to generate consensus sequences (Kennedy and Ebert, 2017; Kennedy et al., 2014). A main emphasis of the work in this thesis is to better define pathophysiologically significant clonal haematopoiesis, ideally using clinically tractable sampling and sequencing approaches that might eventually be applied in a 'real world' setting. The experiments described here have primarily used bulk peripheral blood and bone marrow samples. For a subset of this work (Chapter 3), we compared the performance of consensus sequencing with molecular barcodes and ultradeep targeted sequencing, which is now routinely available in clinical diagnostic laboratories.

5. Thesis Aims

In summary, cancer is a clonal genetic disease adept at evolving resistance to both conventional and targeted therapies (Stratton, 2011; Stratton et al., 2009; Yates and Campbell, 2012). Knowledge of the genetic basis of cancers has galvanised research into early detection using increasingly sensitive sequencing technologies (Cohen et al., 2018; Etzioni et al., 2003; Newman et al., 2016). It is conceivable that earlier detection of asymptomatic, genetically simpler pre-cancerous lesions might enable therapeutic intervention, including targeted therapies for single oncogene addictions, analogous to treatment of chronic phase CML (O'Brien et al., 2003) or therapies to mitigate selection pressures that favour clonal expansion. The success of early cancer detection efforts will hinge upon the ability to distinguish pre-cancer from ubiquitous benign clonal expansions in normal ageing tissues. In the blood system, CH harbouring canonical leukaemia-associated mutations is a risk factor for haematological malignancy (Bowman et al., 2018). However, only a small minority of affected individuals progress, and determinants of evolutionary trajectories remain poorly understood (Figure 1.1). This dissertation investigates the pre-malignant landscape of several common haematological neoplasms and the feasibility of identifying individuals with CH at high risk of developing a blood cancer. The main aims of this project are as follows.

1. Describe the premalignant mutational landscape of the commonest haematological neoplasms and compare this with age-related CH in the general population.
2. Investigate the extent to which benign clonal haematopoiesis can be distinguished from clones at high risk of malignant transformation.
3. Investigate the prevalence of CH in childhood cancer survivors and the natural history of childhood therapy-related myeloid neoplasms.

Figure 1.1

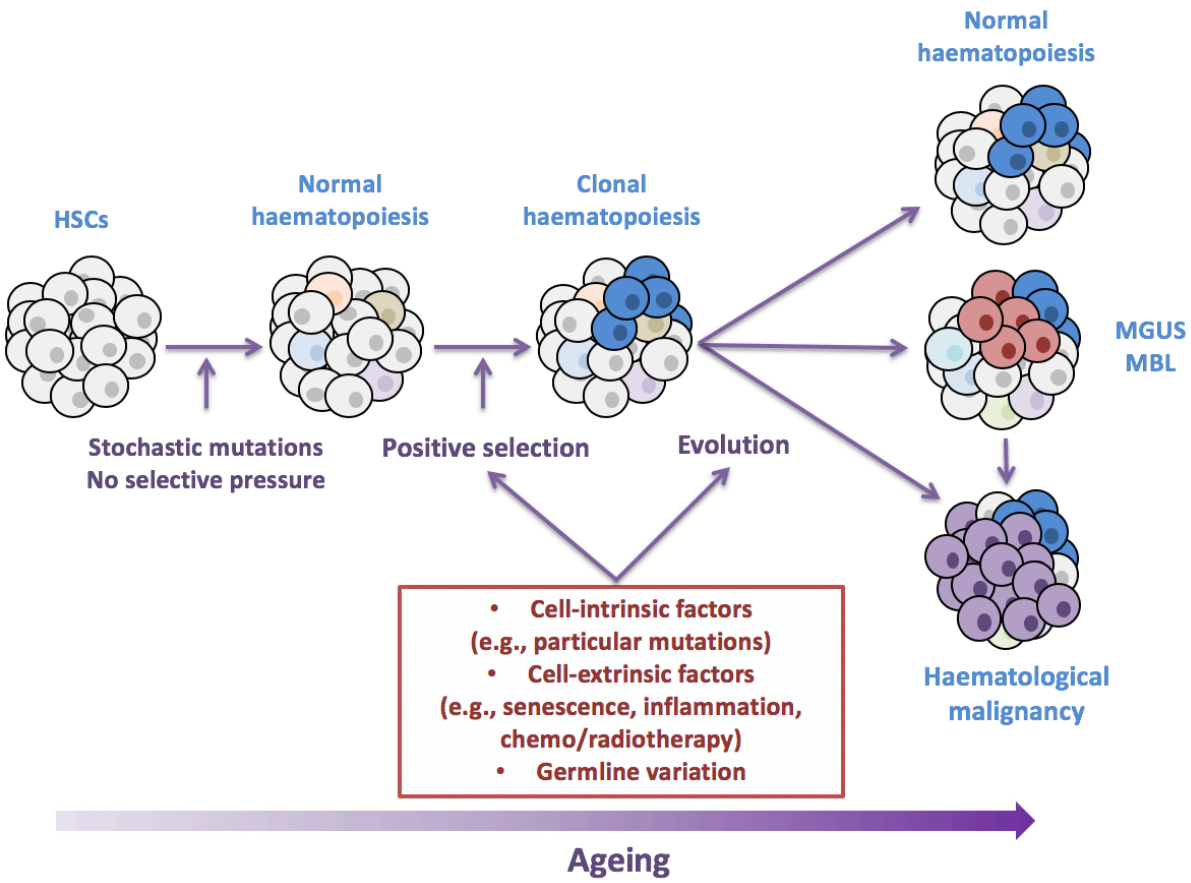

Figure 1.1 | Initiation and evolution of clonal haematopoiesis
Shown is a model illustrating the process of somatic mutation accumulation in HSCs and different clonal trajectories, with known and hypothetical influences on mutation acquisition and/or positive selection highlighted in red. As yet poorly-defined mutational processes acting on HSCs generate somatic genetic diversity in the HSC pool with time, represented here as a mosaic of distinctly coloured cells. Cells with a relative fitness advantage under the selective pressures prevailing in the haematopoietic microenvironment undergo clonal expansion. Clonal haematopoiesis is a nearly inevitable consequence of ageing, and may play a role in maintaining adequate haematopoiesis in a senescing haemopoietic niche. A minority of individuals may progress to a neoplastic disorder. MGUS, monoclonal gammopathy of unknown significance; MBL, monoclonal B-cell lymphocytosis.