Chapter 4

The pre-clinical evolution of lymphoid neoplasms

1. Introduction

As discussed in Chapter 1, the initial exome-based screens for CH in the general population established that most somatic mutations occur in a limited number of genes most frequently implicated in myeloid neoplasms (Genovese et al., 2014; Jaiswal et al., 2014; Xie et al., 2014). However, two of these studies screened broadly for candidate driver events and revealed a broader mutational spectrum, including rare oncogenic mutations in several genes closely associated with lymphoid malignancies, such as *ATM*, *CREBBP* and *MYD88* (Genovese et al., 2014; Xie et al., 2014). The majority of the sensitive, targeted surveys of CH-PD in the general population have since been biased towards detecting mutations in myeloid cancer genes (Acuna-Hidalgo et al., 2017; Coombs et al., 2017; McKerrell et al., 2015; Young et al., 2016). Collectively, these studies have yielded several important insights into CH that were inaccessible to the initial exome screens, for example the high prevalence of small clones harbouring spliceosome gene mutations in older individuals (discussed in Chapter 1, section 3.4.1)(McKerrell et al., 2015; McKerrell and Vassiliou, 2015). Although there is considerable overlap between the cancer genes involved in the commonest lymphoid and myeloid malignancies, the former are generally characterised by more diverse genetic landscapes, with a significant proportion of driver events occurring in infrequently mutated cancer genes (Bolli et al., 2014; Landau et al., 2015; Landau and Wu, 2013; Reddy et al., 2017; Sabarinathan et al., 2017). Given the current literature on CH, it is unclear whether or not a similar spectrum of mutations affecting these less recurrent cancer genes is mirrored in the general ageing population at very low VAF. This is relevant to understanding the selective pressures operative in the ageing haematopoietic niche and to understanding the relationship between CH-PD and lymphoid neoplasms.

As discussed in the introduction to Chapter 3, the studies reporting an association between CH and haematological malignancies were not powered to study distinct classes of blood cancer (Genovese et al., 2014; Jaiswal et al., 2014). The work described in Chapter 3 delineates notable differences in the prevalence and mutational landscape of CH-PD in individuals who later develop *de novo* AML versus that seen in controls, and demonstrates that these genetic features have predictive value for future AML development. However, the extent to which the same is true for other blood cancers remains poorly understood.

The work described in this chapter aims to explore this question by undertaking a broader survey of candidate CH-PD driver genes (Appendix 6) in a cohort of individuals later diagnosed with a lymphoid neoplasm and healthy controls, using a nested case-control experimental design similar to that described in Chapter 3 for AML.

Aims:

1) Compare the prevalence and mutational landscape of CH-PD in the general population with that observed in individuals who go on to develop a lymphoid neoplasm.
2) Correlate genetic features and routinely collected clinical variables with risk of progression to lymphoid malignancy
3) Investigate the combined predictive power of genetic, clinical and demographic features to identify individuals at high risk of developing a lymphoid neoplasm.

2. Results

2.1 Cohort overview

Our EPIC-Norfolk (Day et al., 1999) collaborators (Nick Wareham, Robert Luben, Shabina Hayat and Abigail Britten) identified a discovery cohort comprising 118 study participants diagnosed with a lymphoid neoplasm a mean of 8.0 years (IQR 4.3 - 11.1) after peripheral blood sampling and 118 age- and sex-matched controls with no record of any cancer or haematological disorder (Appendix 12). Individuals were excluded if they were sampled less than 6 months before diagnosis or had a lymphocyte count of 5 x 10^9/L or above, which might be high enough to trigger a clinical work-up for monoclonal B-cell lymphocytosis (MBL) according to current diagnostic criteria (Swerdlow et al., 2016). Given that MBL is a known risk factor for chronic lymphocytic leukaemia (Strati and Shanafelt, 2015), the commonest chronic leukaemia in adults (Dores et al., 2007), we focussed on individuals with lymphocyte counts that would not, in isolation, elicit clinical suspicion of an underlying neoplasm (Swerdlow et al., 2016). The mean age at blood sampling for discovery cohort cases was 64.6 years (IQR 57.0 - 71.8). A validation cohort was also sourced from EPIC-Norfolk and included 71 pre-lymphoid neoplasm (pre-LN) cases and 71 controls (Appendix 13). The mean interval between blood sampling and diagnosis for the validation cohort cases was 8.4 years (IQR 4.1 – 12.3) and mean age at sampling was 64.0 years (IQR 59.4 – 69.8). For the controls, the mean duration of follow-up was 15.4 and 16.4 years for the discovery and validation cohorts, respectively. Serial premalignant samples were available for a subset of the discovery cohort cases and controls. Clinical metadata including full blood count, lipid profile, blood pressure and anthropomorphic measurements were available for the majority of cases and controls. Moreover, out of the 262 controls with clinical metadata described in Chapter 3, 189 were adequately age-and sex-matched to the pre-LN cases, providing a case:control ratio of 1:2 for analysis of clinical factors associated with progression to lymphoid malignancy. These controls were also used to compare mutation frequency in genes that overlapped across the gene panels (Appendices 4 and 6).

The spectrum of future LN diagnoses was similar between the discovery and validation cohorts and is summarised in Table 4.1 with complete metadata for both cohorts detailed in Appendices 12 and 13. For many cases, particularly individuals later diagnosed with a non-

Hodgkin lymphoma, histopathological subtype is unknown. Furthermore, disease classification schemes have evolved dramatically over the course of the recruitment period (Campo et al., 2011; Chapuy et al., 2018; Swerdlow et al., 2016), which would complicate translating historical diagnoses into currently recognised disease entities, and is not essential for the aforementioned aims of this study.

Table 4.1 | Pre-LN cohort summary

Diagnosis	Diagnosis abbreviation	Number of individuals	Mean interval between sample and diagnosis (years)	Mean age at sampling (years)
Peripheral T-cell lymphoma NOS	PTCL NOS	6	8	65.0
Mycosis fungoides	MF	1	3.1	69.9
Non-Hodgkin lymphoma NOS	NHL NOS	37	6.5	65.2
Acute lymphoblastic leukemia	ALL	1	13.3	50.2
Lymphoblastic lymphoma	LL	1	18.5	60.0
Multiple myeloma	MM	43	7.9	63.7
B-cell non-Hodgkin lymphoma	B-NHL	26	7.2	63.2
Diffuse large B-cell lymphoma	DLBCL	25	10.9	64.5
Chronic lymphocytic leukemia	CLL	20	9.1	67.3
Monoclonal gammopathy of undetermined significance	MGUS	12	8.2	65.3
Hodgkin lymphoma	HL NOS	4	14.1	56.1
Small cell B-cell lymphoma	SLL	4	8.4	61.5
Waldenstrom macroglobulinaemia	WM	3	3.9	67.9
Hairy-cell leukemia	HCL	2	4.8	72.8
Nodular sclerosis Hodgkin lymphoma	NScHL	2	5.1	61.3
Extramedullary plasmacytoma	EP	2	6.4	64.2

2.2 Prevalence of CH-PD and driver mutation burden

Peripheral blood samples were deep sequenced with a custom panel comprising 95 genes implicated in haematological malignancies (Methods section 2.4 and Appendix 6). Average sequencing coverage was >5,000 (IQR 4,750 – 5,800). The prevalence of CH-PD was significantly higher in pre-LN cases than in controls (*P* = 0.0019, two-sided Fisher's exact test), though the difference was less dramatic than that observed for pre-AML (Figure 4.1a). Overall the prevalence of CH-PD in pre-LN cases and controls was 35.4% and 20.6%, respectively (Figure 4.1a,b). These proportions were similar across the discovery cohort (CH-PD prevalence of 33.9% in cases and 17.8% in controls) and validation cohort (38% and 25.4% for cases and controls, respectively). The average number of driver mutations identified in pre-LN cases was 0.43 compared to 0.25 for controls (*P*=0.0016, two-sided Wilcoxon rank-sum test), with a significant trend towards increasing driver mutation burden with age (Figure 4.1c). Moreover, as seen for pre-AMLs, the VAF of driver mutations was significantly higher in pre-LN cases versus controls (median VAF 6.9% and 2.8%, respectively; *P* = 0.00036, Wilcoxon rank-sum test; Figure 4.1d).

Figure 4.1

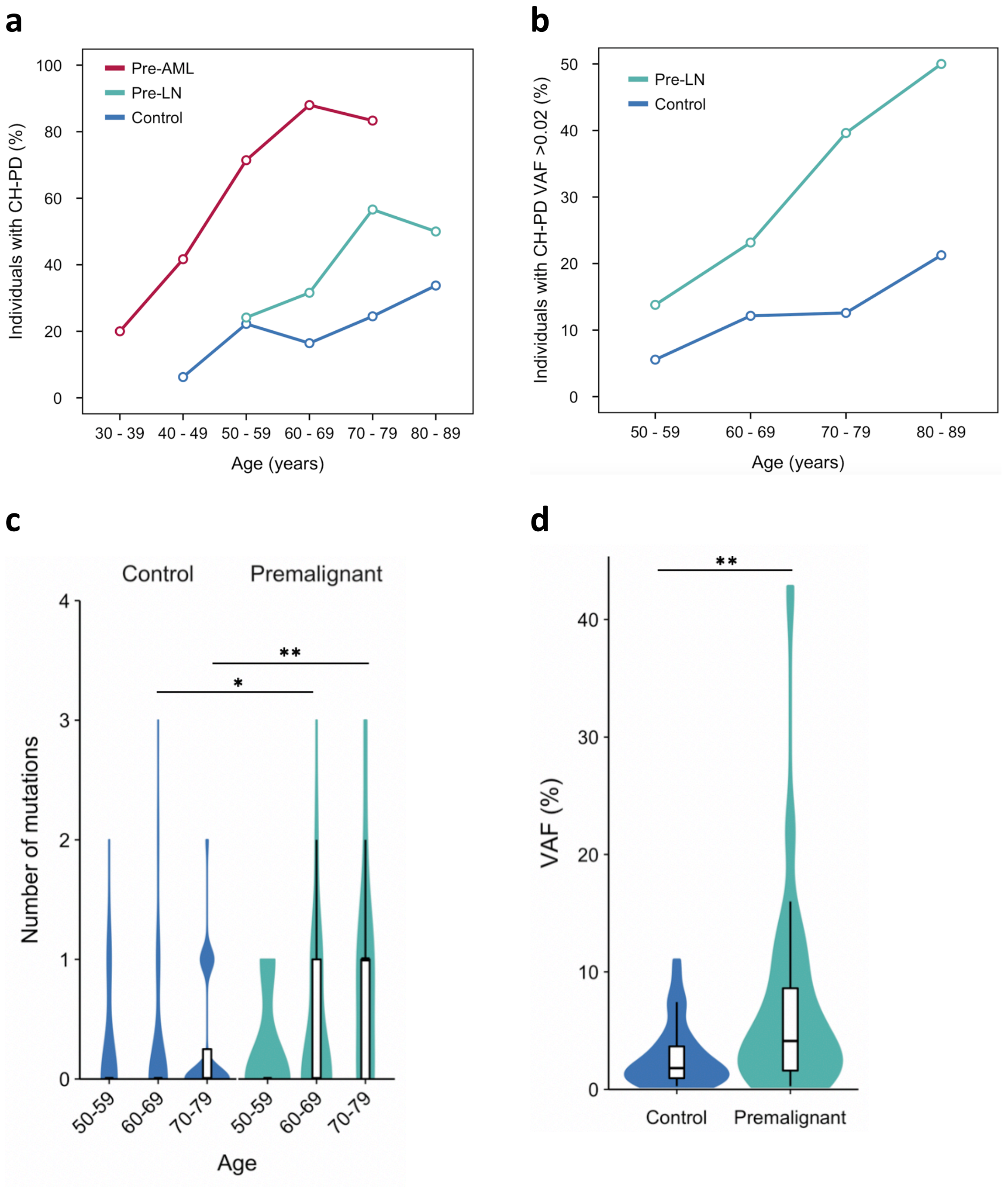


Figure 4.1 | Prevalence of CH-PD, number of mutations and clone size in pre-LN and control cohorts. a, Prevalence of CH-PD among pre-LN cases (green), controls (blue) and pre-AML (red; data from chapter 3). **b,** Prevalence of CH-PD clones with VAF > 2% among pre-LN cases (green) and controls (blue) is shown to put the data in the context of the historical definition of 'clonal haematopoiesis of indeterminate potential' (CHIP). **c**, The number of CH-PD mutations detected in pre-LN cases and controls according to age. Box plot centres, hinges and whiskers represent the median, first and third quartiles and 1.5× interquartile range, respectively. **d**, VAF of CH-PD mutations in pre-LN cases (green) and controls (blue). * indicates P < 0.1; ** indicates P<0.001, two-sided Wilcoxon rank-sum test with Benjamini-Hochberg multiple testing correction.

2.3 Mutational spectrum of CH-PD in individuals who later developed a lymphoid neoplasm

Among the 189 discovery and validation cohort controls, the top three most frequently mutated genes were *DNMT3A, TET2* and *ASXL1* (Figure 4.2a-c, Appendix 14), consistent with the findings of other studies of CH-PD in the general population (Bowman et al., 2018). By contrast, among individuals who later developed a lymphoid blood cancer, the most recurrently mutated genes were *DNMT3A* (16.4% of cases versus 14.4% of controls), *TET2* (6.9% of cases vs 2.7% of controls), *ATM* (2.7% of cases vs 0.53% of controls) and *TP53* (2.7% of cases and 1.1% of controls). Among the genes recurrently mutated in both cases and controls, the mean mutation VAF was consistently higher in cases, though this difference only reached statistical significance on an individual gene level for *DNMT3A* (mean VAF in cases and controls 5.9% and 2.8%, respectively; *P* = 0.029, two-sided Wilcoxon rank-sum test with BH multiple testing correction). Furthermore, CH-PD in the pre-LN cases demonstrated a remarkably diverse spectrum of mutations, with putative driver variants identified in a total of 24 genes, compared to 11 genes among the controls (Figure 4.2a,b). Although there is broad overlap between the cancer genes implicated in myeloid and lymphoid malignancies (Arber et al., 2016; Sabarinathan et al., 2017; Swerdlow et al., 2016), several of the genes mutated among the cases are predominantly implicated in the latter, including *POT1*, *XPO1*, *HIST1H1E*, *NOTCH1*, *NOTCH2*, *ATM* and *CCND3* (Arber et al., 2016; Hing et al., 2016; Lunning and Green, 2015; Sabarinathan et al., 2017; Swerdlow et al., 2016).

Although data were too sparse to discern significant changes in the mutational spectrum with age, it is noteworthy that mutations in spliceosome genes (*SF3B1*, *SRSF2* and *U2AF1*) were only observed in controls over the age of 70, consistent with previous studies strongly associating these mutations with CH-PD in older individuals (Figure 4.2d)(McKerrell et al., 2015). Among the cases, the splicing gene mutation with the highest VAF (*SF3B1* p. K700E, VAF 2.1%) occurred in a 54-year-old man (PD00315) sampled 8 years before diagnosis with chronic lymphocytic leukaemia (CLL).

Figure 4.2

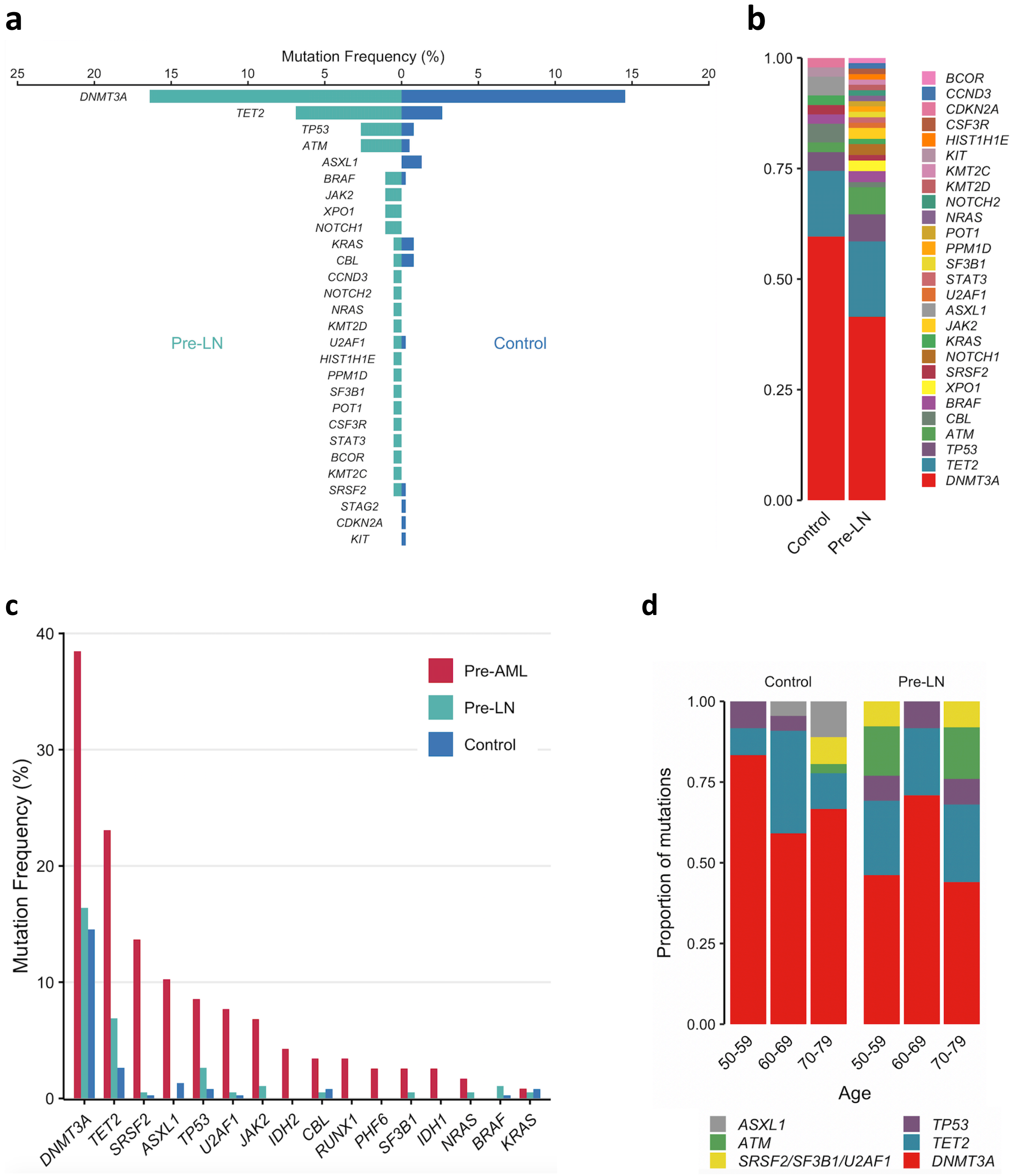
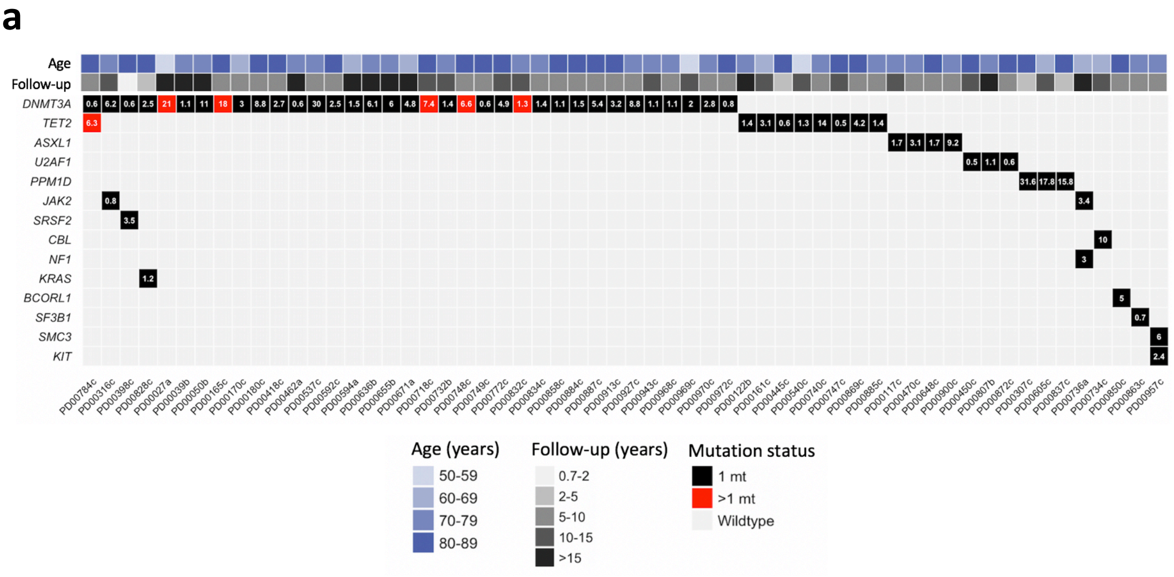


Figure 4.2 | The mutational spectrum of clonal haematopoiesis in individuals who developed a lymphoid neoplasm years later versus controls. a, Proportion of pre-LN cases (green) and controls (blue) with driver mutations each given gene. **b**, Relative frequency of mutations in the indicated genes according to age group for pre-LN cases and controls. **c**, Proportion of pre-AML (red), pre-LN (green) and control (blue) individuals with driver mutations in genes sequenced for both the pre-AML (chapter 3) and pre-LN cohorts. **d,** Relative frequency of mutations in the indicated genes according to age group for pre-LN cases and controls; only genes mutated at least 5 times included in panel, with spliceosome genes *SRSF2, SF3B1* and *U2AF1* aggregated.

2.4 Mutational spectrum in an extension cohort of older individuals with no record of cancer or a blood disorder

The more diverse genetic landscape of CH-PD in the pre-LN cases is intriguing, though the limited sample sizes and 1:1 case:control ratio warrant cautious interpretation. Although collectively a significant proportion of the mutations observed in the pre-LN cases occur in genes never or rarely reported in CH-PD in the general population, individual genes were infrequently mutated. Hence, despite the notable differences in mutational spectra between pre-LN cases and controls, considering all genes mutated more than 5 times across both cohorts on an individual basis, only *TET2* mutations approached significance for enrichment among the pre-LN cases (6.9% vs 2.7% mutated) (P = 0.05, one-sided Fisher's exact test with BH multiple testing correction). Is the absence of recurrent LN-drivers in the 189 age-and sex-matched controls included in the discovery and validation cohorts truly representative of the frequency of such mutations in the general ageing population? As mentioned in the introduction, most of the sensitive targeted surveys of CH-PD have used gene panels restricted to the most recurrent CH-PD driver genes and have not included the aforementioned LN-associated cancer genes (Acuna-Hidalgo et al., 2017; Coombs et al., 2017; Gibson et al., 2017; McKerrell et al., 2017; McKerrell et al., 2015; Young et al., 2016). The cumulative incidence of both common adult lymphoid malignancies and of CH-PD increases dramatically with age (Howlader et al., 2011), and it is conceivable that a more diverse CH-PD genetic landscape enriched for recurrent LN drivers emerges at higher rates in older age groups, analogous to the trend observed for spliceosome gene mutations (McKerrell et al., 2015). To investigate this possibility, we sequenced an extension cohort of 234 individuals (n=238 samples) with no record of any prior or subsequent cancer diagnosis or known blood disorder. The mean age at blood sampling was 74.4 years (IQR 67.5-81.6), more than ten years older on average than the control cohort. The mean follow-up was 11.9 years (IQR 8.0-16.4). Out of the 234 individuals, 58 (24.8%) had CH-PD (Appendix 14). Despite high coverage (median >5,000X) and sensitivity to detect small clones down to VAF 0.5%, the genetic landscape was consistent with that observed in previous studies of CH-PD in the general ageing population. In particular, no canonical drivers associated with lymphoid malignancies were identified (Figure 4.3a,b), in contrast to the pre-LN cohort.

Figure 4.3



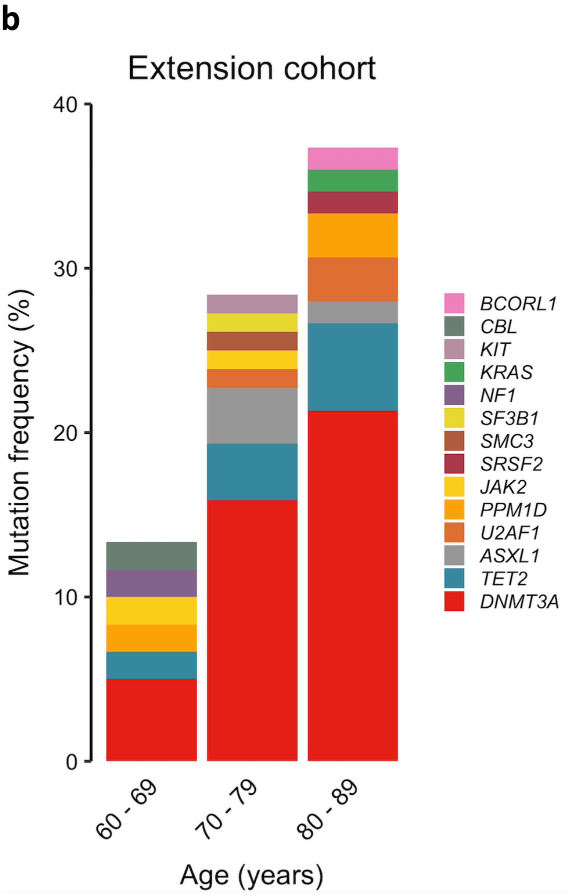


Figure 4.3 | The mutational spectrum of clonal haematopoiesis in an extension control cohort of older individuals with no history of cancer or haematological disorder. a, Co-mutation plot including only individuals with CH-PD (58 out of 234 individuals in the older extension cohort). The top two rows indicate age at sampling and follow-up period in years. Tiles are coloured according to mutation status for each given gene and number of drivers identified: pale grey, wild type; black, one driver mutation; red, two driver mutations. The mutation VAF (%) is indicated in white text within each tile. Where two mutations were identified in a given gene and sample (red tiles), the highest VAF is shown. **b,** Proportion of individuals with driver mutations in each given gene according to age group.

2.5 Clonal dynamics over time and relationship with future lymphoid neoplasm

Examining co-mutation patterns in those with a future LN diagnosis (Figure 4.4a-b) invites some initial speculation regarding the relationship between CH-PD and future LN. For many cases, the only CH-PD mutations detected occur in genes that are seldom implicated as drivers in the lymphoid cancer type diagnosed years later. The most notable example is *DNMT3A*, the most frequently mutated gene among both cases and controls (Figure 4.2a). Although *DNMT3A* does play a role in some lymphoid malignancies, particularly T-cell leukaemia/lymphoma (Couronne et al., 2012; Haney et al., 2016a; Haney et al., 2016b), it is not among the most recurrently mutated genes in these disorders (Brunetti et al., 2017; Sabarinathan et al., 2017). By contrast, the *BRAF* p.V600E, *POT1* p.K90E and *XPO1* p.E571 hotspot mutations preceding diagnoses of hairy cell leukaemia (HCL), small cell B-cell lymphoma (SLL) and CLL, respectively, are highly plausible drivers of the respective latent malignancies, but are rarely if ever associated with CH in the general population (Landau et al., 2015; Pinzaru et al., 2016; Tiacci et al., 2011).

In order to further investigate the relationship between CH-PD detected years before LN diagnosis and the future malignancy, serial peripheral blood DNA samples were sequenced from 104 individuals, including 69 pre-LN cases and 35 controls. The mean interval between earliest and latest sample was 7.3 years. No diagnostic specimens were available; however, for 16 of the pre-LN cases, at least one peripheral blood sample taken less than 6 months before diagnosis (n = 5 individuals) or after diagnosis (n = 11 individuals) was sequenced.

Of the 69 serially sampled pre-LN cases, 22 had at least one driver detected in an earlier time point sample. Out of the 26 distinct mutations identified, 25 persisted in the later sample and 1 became undetectable. The only non-persistent clone harboured a *KRAS* p.G13D mutation present at 1% VAF in PD00003 at age 62.4 and no longer detectable in a sample taken 8.5 years later. Among the 35 controls with serial samples, 7 had mutations detected in their earlier samples. Of the 10 distinct variants, 5 persisted and 5 were no longer detected in the subsequent sample. The latter group comprised low VAF mutations in *DNMT3A* (n=4) and *KRAS* (n=1). Consistent with the patterns seen in pre-AML cases and controls, examining the VAF trajectories of the persistent mutations over time demonstrated variable behaviour, including for clones with mutations in the same gene (Figure 4.5a,b). However, the numbers of cases and controls with mutations were insufficient to infer any significant overall difference in clonal growth rates between pre-LN cases and controls.

Examining the sequence of mutation acquisition and VAF trajectories among the pre-LN cases revealed several notable findings (Figure 4.6a-k). Among the 16 pre-LN cases with peri- or post-diagnosis samples available, 7 harboured antecedent CH-PD. All 7 individuals harboured at least one driver in *DNMT3A* (Figure 4.6a-g), all of which persisted across serial samples. In 4/7 cases, the size of the *DNMT3A* clone(s) diminished over time (Figure 4.6a,d,f,g), and in 2 of these cases this decline coincided with late acquisition of at least one driver mutation in a canonical lymphoid cancer gene, specifically *CCND3* and *CREBBP* in an NHL and *SF3B1* in a CLL case (Figure 4.6d,g)(Chapuy et al., 2018; Lunning and Green, 2015; Mullighan, 2014; Okosun et al., 2014; Sabarinathan et al., 2017). The same phenomenon is observed in two other cases, with the appearance of a relatively LN-specific driver mutation (e.g., in *NOTCH1, POT1* and *HIST1H1E*)(Sabarinathan et al., 2017; Swerdlow et al., 2016) years before diagnosis also coinciding with stable or falling VAF of mutations in the canonical CH/myeloid neoplasm drivers *DNMT3A* and *U2AF1*, respectively (Figure 4.6i,j). These observations strongly suggest the presence of distinct, potentially competing clones and supports the hypothesis that a significant proportion of the CH-PD in the pre-LN cases is not phylogenetically related to the future malignancy, despite large clone sizes in most instances. Four serially-sampled pre-LN cases harboured drivers in genes more frequently mutated in LN than in CH-PD, namely *CCND3, ATM, BRAF* and *TP53* (Figure 4.4a), and in each of these cases VAF increased over time. Hence, despite limited data, this time series experiment suggests that CH-PD in pre-LN cases represents a combination of pre-malignant clones and ‘bystander’ clones, analogous to the situation observed in pre-AML.

Figure 4.4

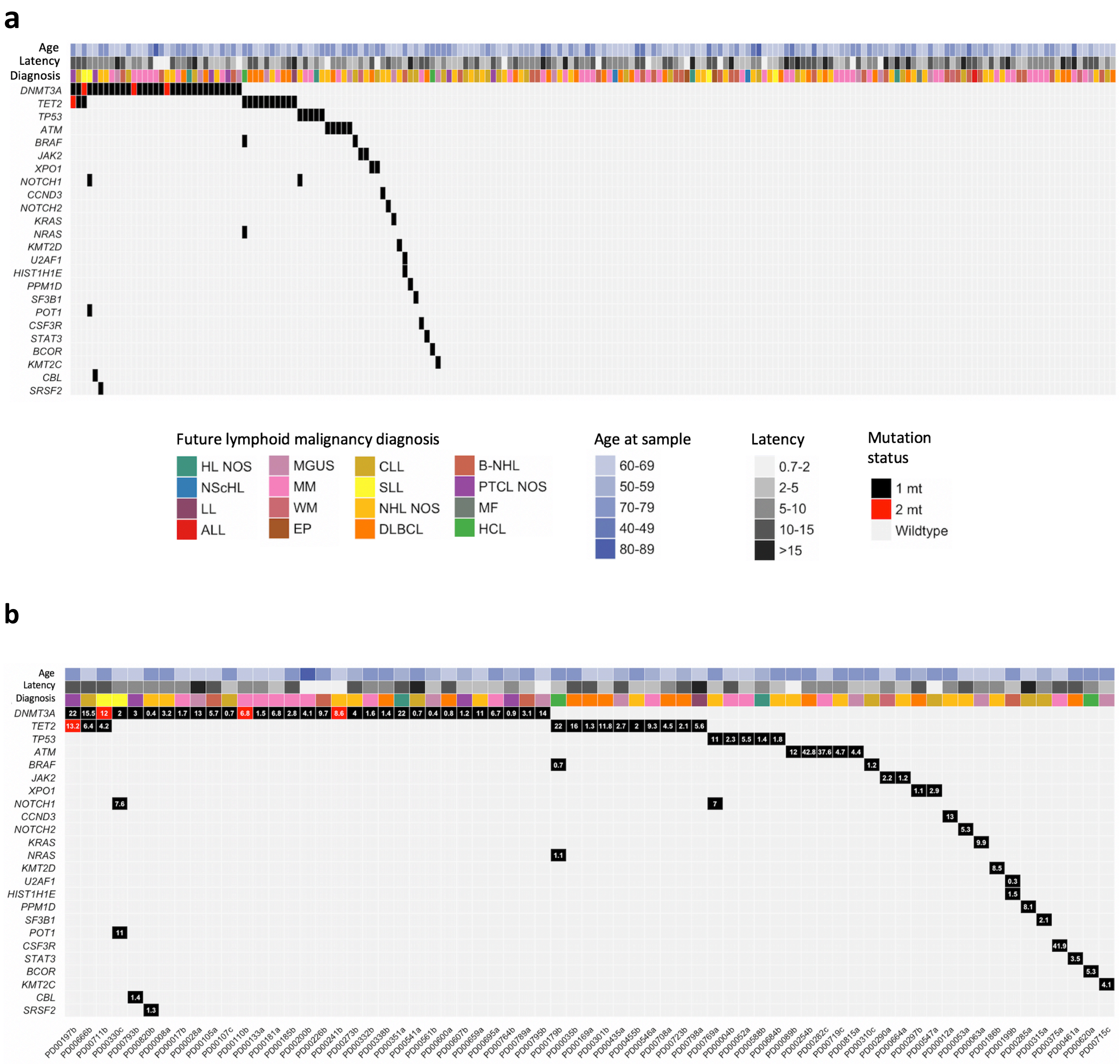


Figure 4.4 | Mutation co-occurrence in pre-LN cases according to diagnosis, latency and age at sampling. a, Co-mutation plot for all 189 pre-LN cases. Top three rows indicate age at sampling, latency and sample and future LN diagnosis. Tiles are coloured according to mutation status for each given gene and number of drivers identified: pale grey, wild type; black, one driver mutation; red, two driver mutations. **b**, Co-mutation plot including only cases with CH-PD. The mutation VAF percentage is indicated in white text within each tile. Where two mutations were identified in a given gene and sample (red tiles), the highest VAF is shown. MM, multiple myeloma; NHL NOS, non-Hodgkin lymphoma not otherwise specified; MGUS, monoclonal gammopathy of undetermined significance; DLBCL, diffuse large B-cell lymphoma; B-NHL, B-cell non-Hodgkin lymphoma; CLL, chronic lymphocytic leukemia; HCL, hairy-cell leukemia; PTCL NOS, peripheral T-cell lymphoma NOS; WM, Waldenstrom macroglobulinaemia; SLL, small cell B-cell lymphoma; HL, Hodgkin lymphoma; LL, lymphoblastic lymphoma; NScHL, nodular sclerosis Hodgkin lymphoma; EP, extramedullary plasmacytoma; MF, mycosis fungoides; ALL, acute lymphoblastic leukemia.

Figure 4.5

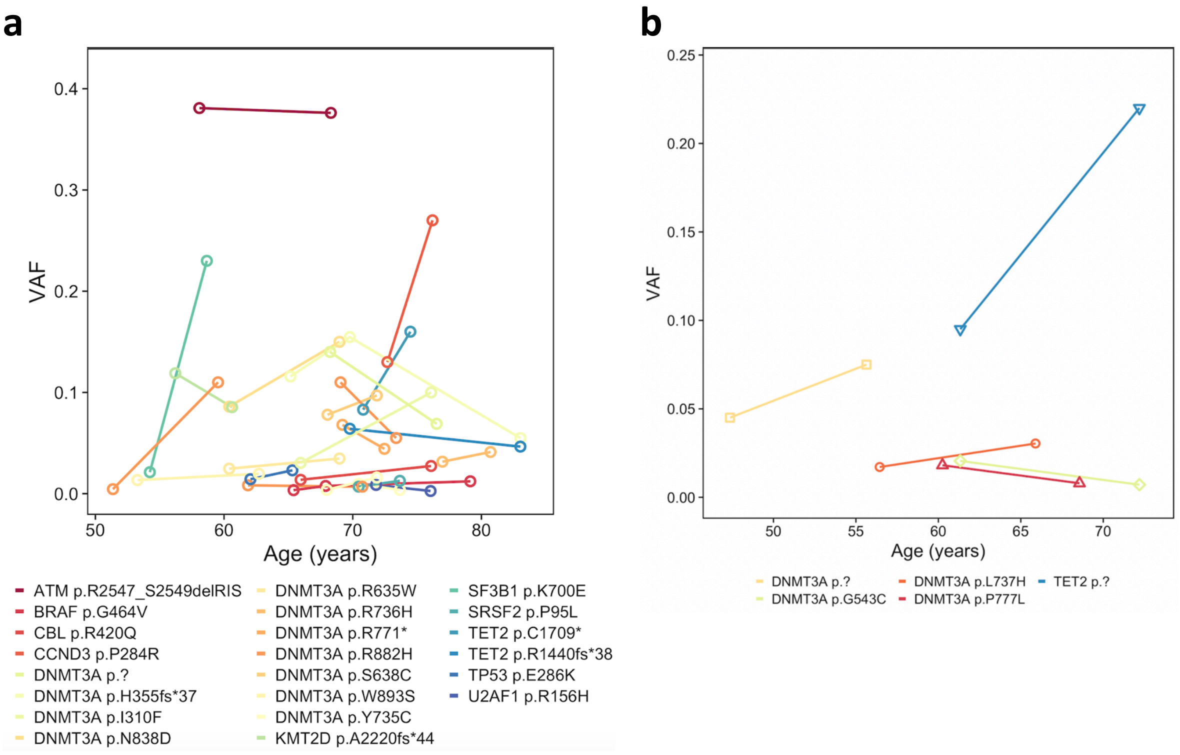


Figure 4.5 | VAF trajectories of persistent mutations in serially sampled pre-LN cases and controls. **a-b,** VAF trajectories of CH-PD driver mutations persisting across serial samples from cases sampled years before diagnosis of a lymphoid neoplasm **(a)** and controls **(b)**. X-axis denotes age at sampling and y-axis mutation VAF.

Figure 4.6

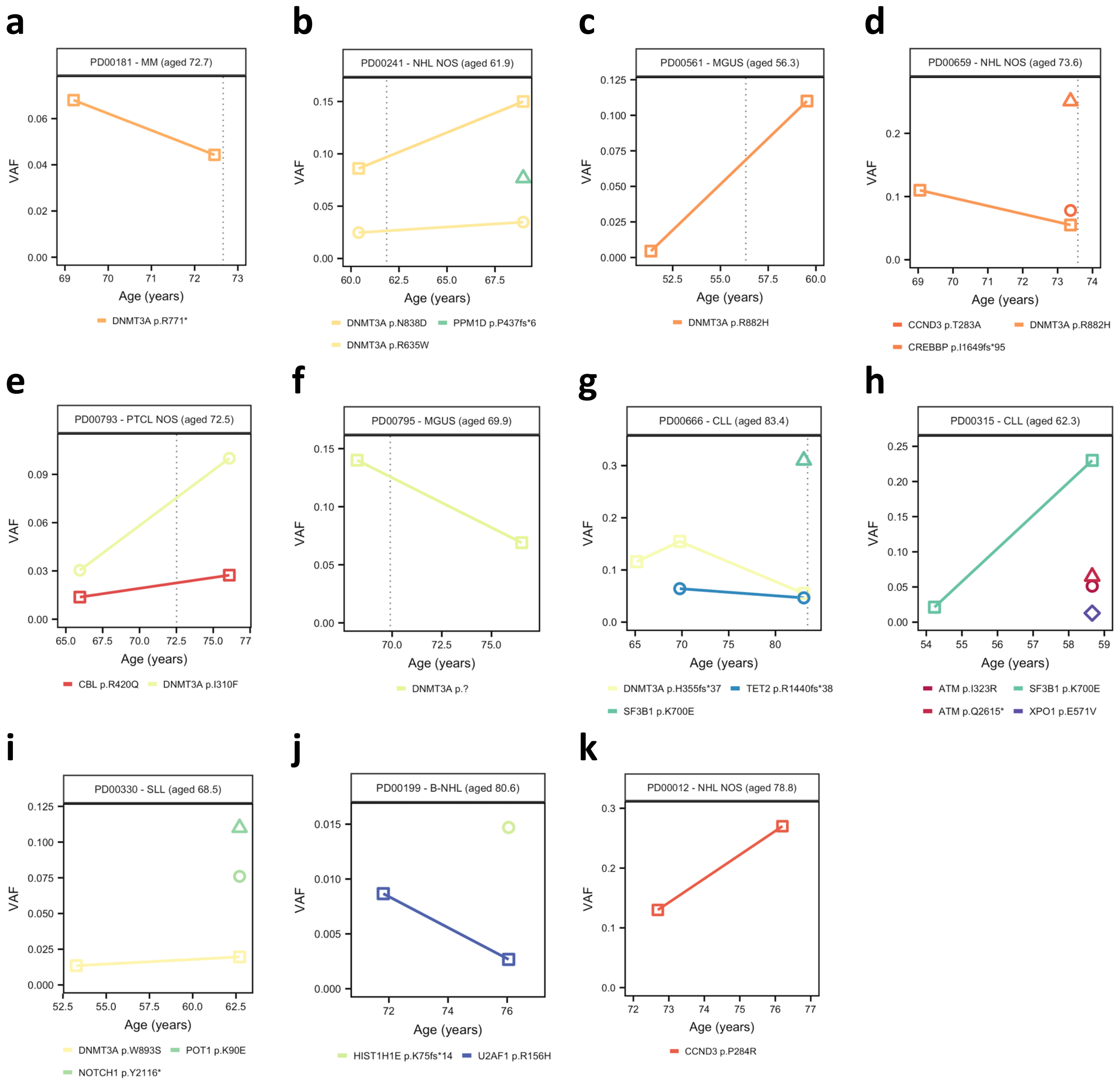


Figure 4.6 | Evolution of clonal haematopoiesis and relationship with future lymphoid neoplasm. a-h, VAF trajectories of putative driver mutations in 7 individuals for whom peripheral blood taken near or after cancer diagnosis was available for sequencing. Future LN diagnosis and age at diagnosis are indicated in parentheses above the plot. Vertical dotted lines demarcate pre- and post-diagnosis periods. **i-k,** VAF trajectories of putative driver mutations in an additional 5 cases sampled multiple times years before cancer diagnosis. Age at sampling and mutation VAF are shown on the x- and y-axis, respectively. LN, lymphoid neoplasm; VAF, variant allele fraction; MM, multiple myeloma; NHL NOS, non-Hodgkin lymphoma not otherwise specified; MGUS, monoclonal gammopathy of undetermined significance; B-NHL, B-cell non-Hodgkin lymphoma; CLL, chronic lymphocytic leukemia; PTCL NOS, peripheral T-cell lymphoma NOS; SLL, small cell B-cell lymphoma.

2.6 Clinical factors associated with future development of a lymphoid malignancy

Full blood count parameters, lipid profile, C-reactive protein, blood pressure and anthropomorphic measurements were available for most of the pre-LN cases and controls (Figure 4.7). The case:control ratio for this analysis was 1:2 due to inclusion of 189 age-and sex-matched controls from the validation cohort described in Chapter 3. Consistent with the observations in the pre-AML cases and controls and previous studies of CH-PD (Jaiswal et al., 2014; McKerrell and Vassiliou, 2015), blood counts did not differ significantly between pre-malignant cases and controls or between individuals with and without CH-PD (Figure 4.7). Assessing all clinical parameters available for the majority of pre-LN cases and controls revealed significantly lower levels of high-density lipoprotein (HDL) in pre-LN cases (*P*=0.048, two-sided Wilcoxon rank-sum test with BH multiple testing correction). No other trends in clinical variables remained significant after multiple testing correction. There were no significant differences in clinical parameters when only cases and controls with CH-PD were compared to each other or when all individuals (cases and controls) with CH-PD were compared to individuals with no detectable mutations. Kaplan-Meier analysis of the impact of clinical variables on LN-free survival showed trends towards shorter time to cancer progression with higher RDW, though this correlation did not reach significance (Figure 4.8).

Figure 4.7

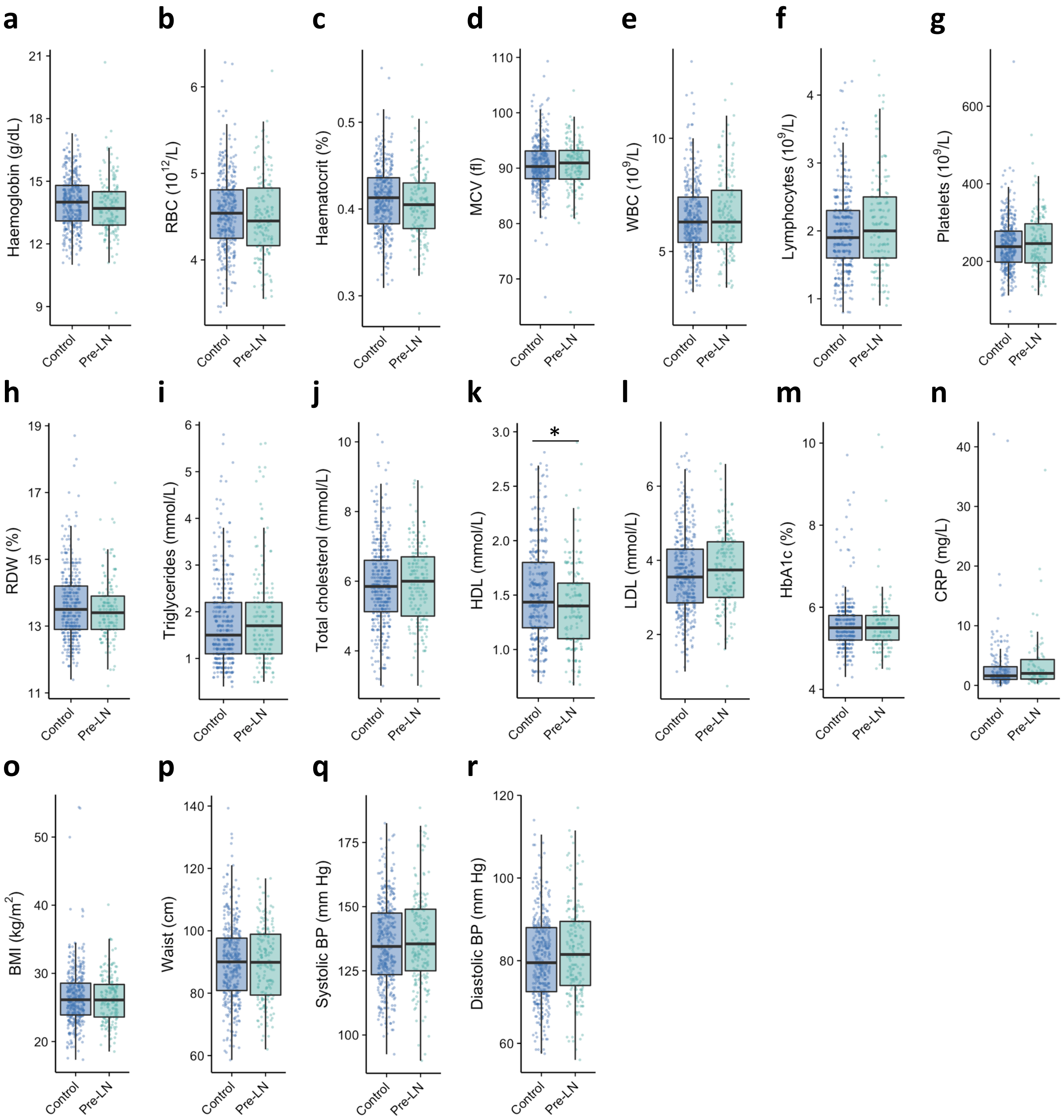


Figure 4.7 | Full blood count and metabolic parameters in pre-LN cases and controls. Box plots of full blood count parameters **(a-h)**, biochemistry measurements **(i-n)**, body mass index **(o)**, waist circumference **(p),** and blood pressure **(q-r)** available for a subset of cases and pre-LN controls. Boxplot centres, hinges and whiskers represent the median, first and third quartiles and 1.5× interquartile range, respectively. RBC, red blood cell; MCV, mean corpuscular volume; WBC, white blood cell; RDW, red cell distribution width; HDL, high density lipoprotein; LDL, low density lipoprotein; HbA1c, haemoglobin A1c; CRP, C-reactive protein; BMI, body mass index; BP, blood pressure. * *P*=0.048, two-sided Wilcoxon rank-sum test with BH multiple testing correction

Figure 4.8

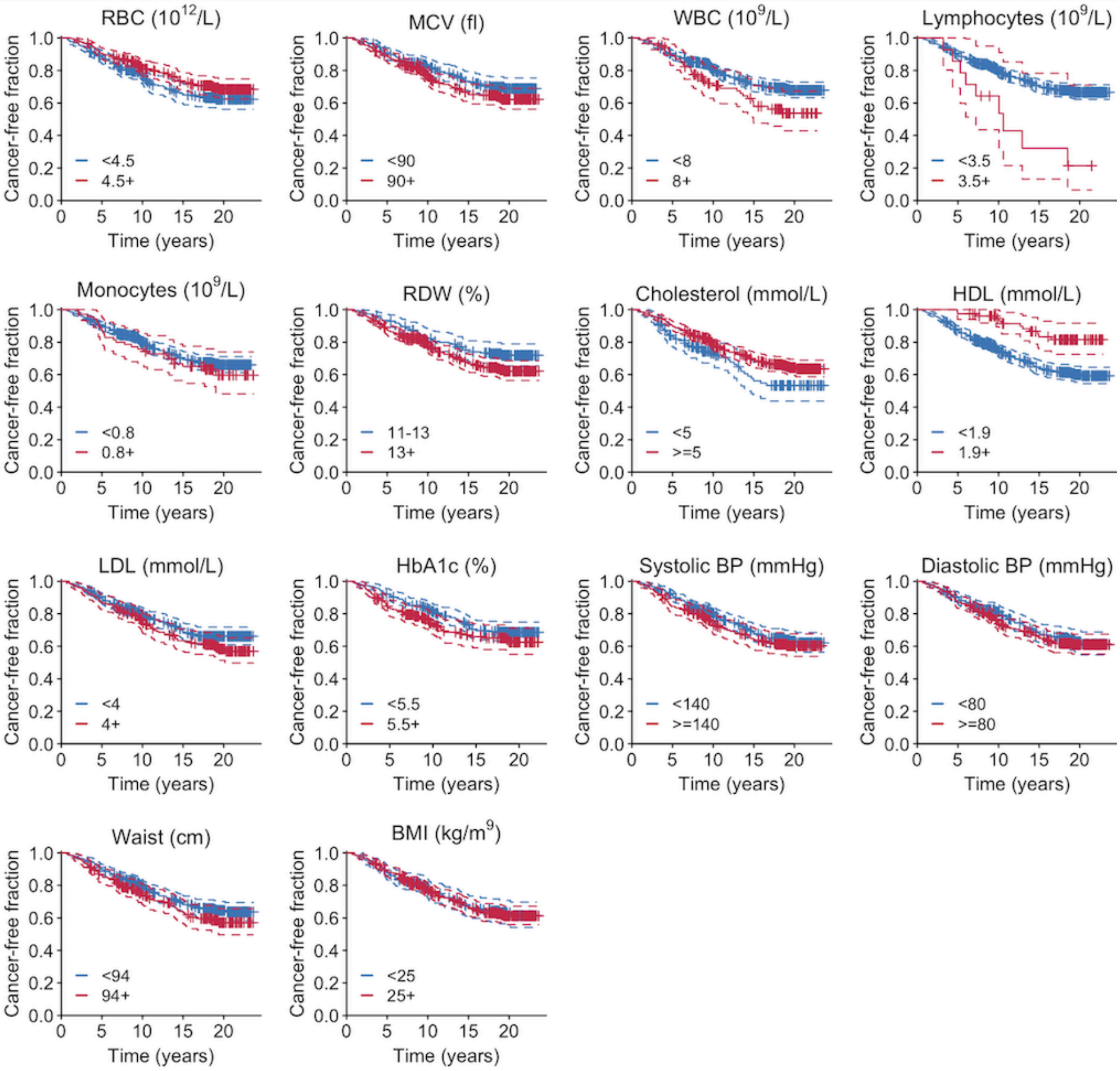


Figure 4.8 | Impact of clinical variables on lymphoid neoplasm-free survival. Kaplan–Meier curves of LN-free survival, defined as the time between sample collection and LN diagnosis, death or last follow-up. Survival curves are stratified according to cutoffs indicated in the lower left corner of each plot. *n* = 567 unique individuals, including 189 pre-LN cases and 378 age- and sex-matched controls. 95% confidence intervals indicated by dashed lines. RBC, red blood cell; MCV, mean corpuscular volume; WBC, white blood cell; RDW, red cell distribution width; HDL, high density lipoprotein; LDL, low density lipoprotein; HbA1c, haemoglobin A1c; BMI, body mass index; BP, blood pressure.

2.7 Predicting progression to lymphoid malignancy

On the basis of these findings, an approach similar to that described in Chapter 3 was developed to quantify the relative contributions of driver mutations, clone sizes and clinical factors to the risk of progressing to a lymphoid malignancy. In keeping with results from Chapter 3, Kaplan-Meier analysis of the impact of the number of drivers and mutation VAF demonstrated consistent correlation between mutation burden and progression-free survival, though these trends did not reach significance (Figure 4.9a). This correlation held even when the additional set of controls was incorporated and analysis was restricted to genes included in the myeloid panel used in Chapter 3 (Figure 4.9b). Although the relative infrequency of CH among pre-LN cases limited the power of KM analysis, a trend towards shorter LN-free survival was observed with larger *DNMT3A* clones (Figure 4.9c) or the presence of mutations in any of the LN-associated genes *XPO1, POT1, CCND3, HIST1H1E, NOTCH1* or *NOTCH2* (Figure 4.9d). KM curves for individual genes are shown in Figure 4.10.

Figure 4.9

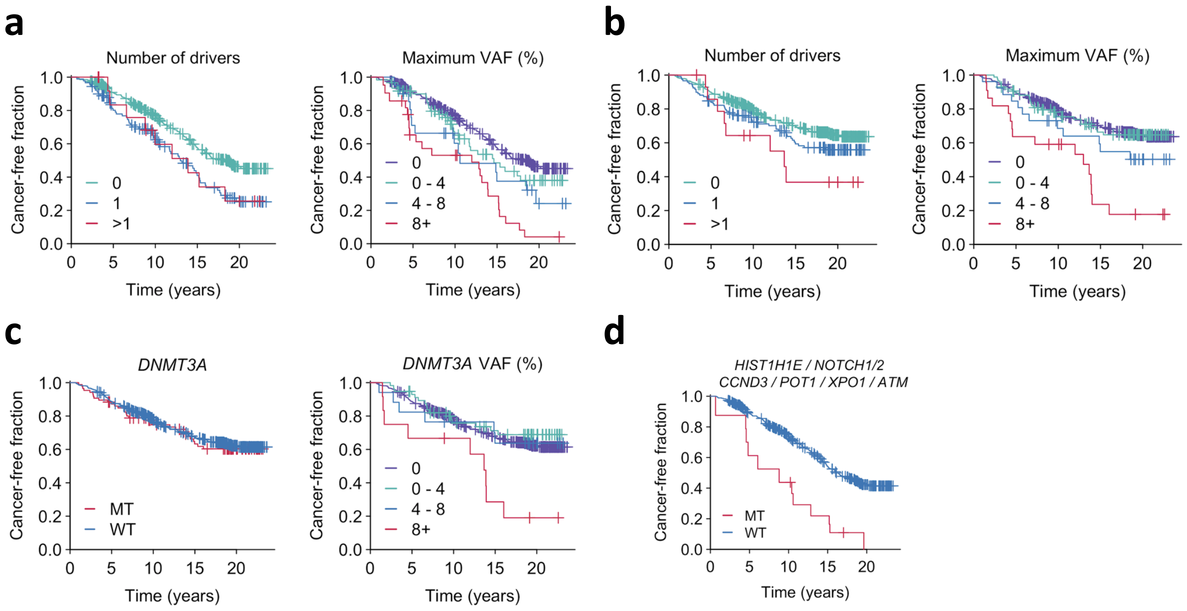


Figure 4.9 | Impact of mutation burden on lymphoid neoplasm-free survival. a,b Kaplan–Meier (KM) curves of LN-free survival, defined as the time between sample collection and LN diagnosis, death or last follow-up. Survival curves are stratified according to number of driver mutations per individual and largest clone detected. Panel **(a)** includes all genes sequenced across the 189 pre-LN cases and 189 age- and sex-matched controls. The same trends, albeit not reaching significance, persist when only mutations in genes sequenced by the myeloid panel are included in the analysis (189 pre-LN cases and 378 controls) **(b)**. **c,** KM curves of LN-free survival stratified by *DNMT3A* mutation status and VAF of *DNMT3A* mutations. **d,** KM curve of LN-free survival stratified according to mutation status in any of six infrequently mutated lymphoid neoplasm-associated driver genes. VAF, variant allele fraction.

Figure 4.10

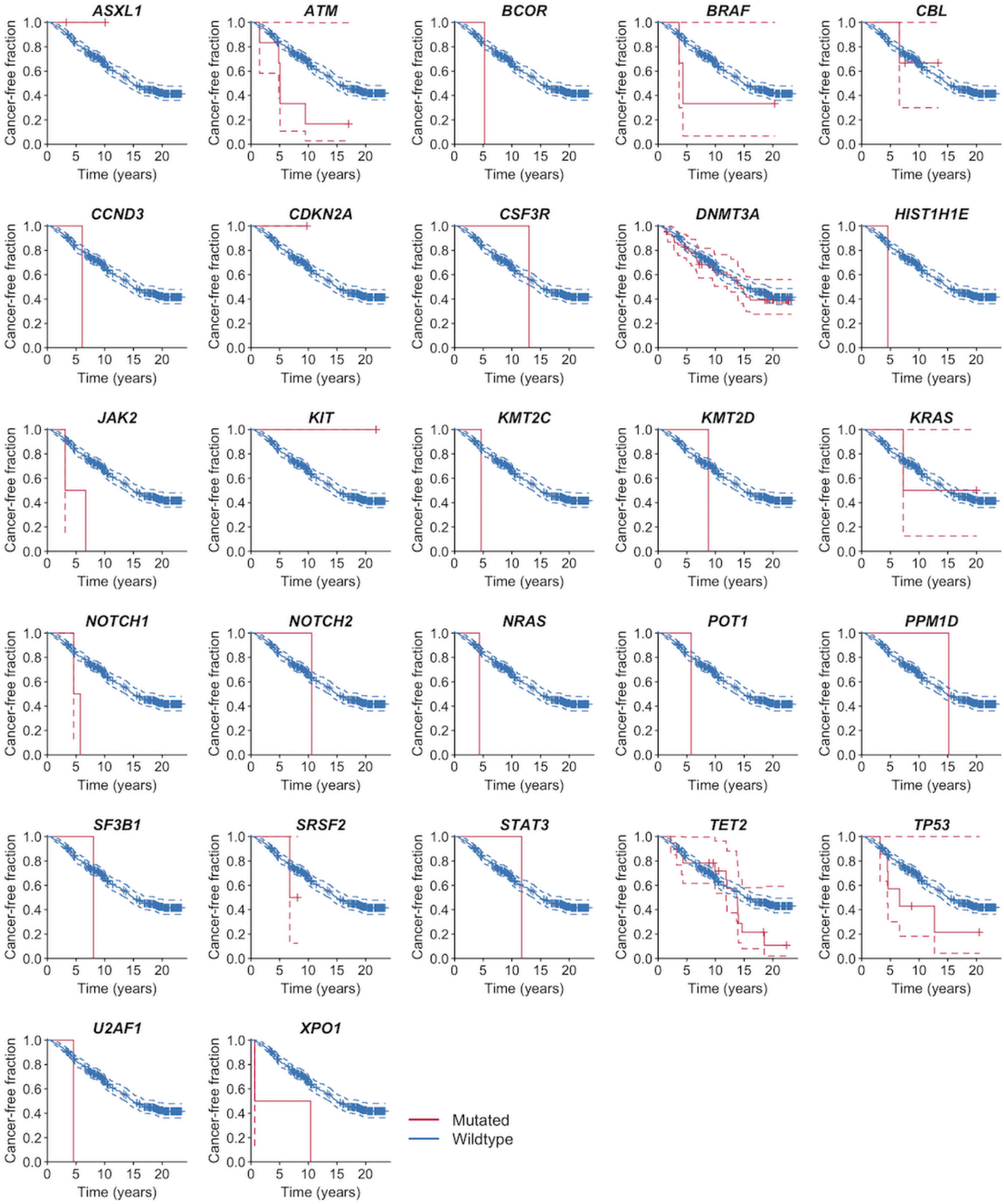


Figure 4.10 | Gene-level impact on LN-free survival. Kaplan–Meier (KM) curves of LN-free survival, defined as the time between sample collection and AML diagnosis, death or last follow-up. Survival curves are stratified according to mutation status. *n* = 378 unique individuals (189 pre-LN cases and 189 controls). LN, lymphoid neoplasm; VAF, variant allele fraction. Dashed lines indicate 95% confidence intervals.

However, the high proportion of infrequently mutated genes dominating the genetic landscape of CH-PD among pre-LN cases and lower prevalence of CH-PD among pre-LN relative to pre-AML hindered robust identification of gene-level risk factors for malignant progression. Regularised logistic and Cox proportional hazards regression approaches were applied as described in Chapter 3 (see Methods section 4). Excluding infrequently mutated genes from model training eliminated a significant proportion of CH-PD mutations from analysis and yielded fairly homogenous gene-level hazard ratios with wide confidence intervals for most genes (Figure 4.11). Notable exceptions were *DNMT3A* and *TET2*, which were the most recurrently mutated genes across both cohorts and were thus amenable to more accurate analysis of the mutation contribution to LN progression risk (Figure 4.11a). Quantitatively, driver mutations in *DNMT3A* and *TET2* conferred a 1.5 to twofold increased 10-year risk of LN per 5% increase in clone size (Figure 4.11a and Appendix 15). Remarkably, these hazard ratios are virtually identical to the effect sizes observed for these genes in the AML prediction model (Figure 3.5). In order to achieve more accurate estimates of HRs for clinical variables and the subset of genes sequenced across both gene panels, the model was retrained using an additional set of 189 controls sequenced with the myeloid panel used in Chapter 3 for a case:control ratio of 1:2. The genes analysed were restricted to those overlapping between both panels and mutated at least twice in either discovery or validation cohort. Hazard ratios for overlapping variables were concordant, albeit with narrower confidence intervals (Figure 4.11b).

Figure 4.11

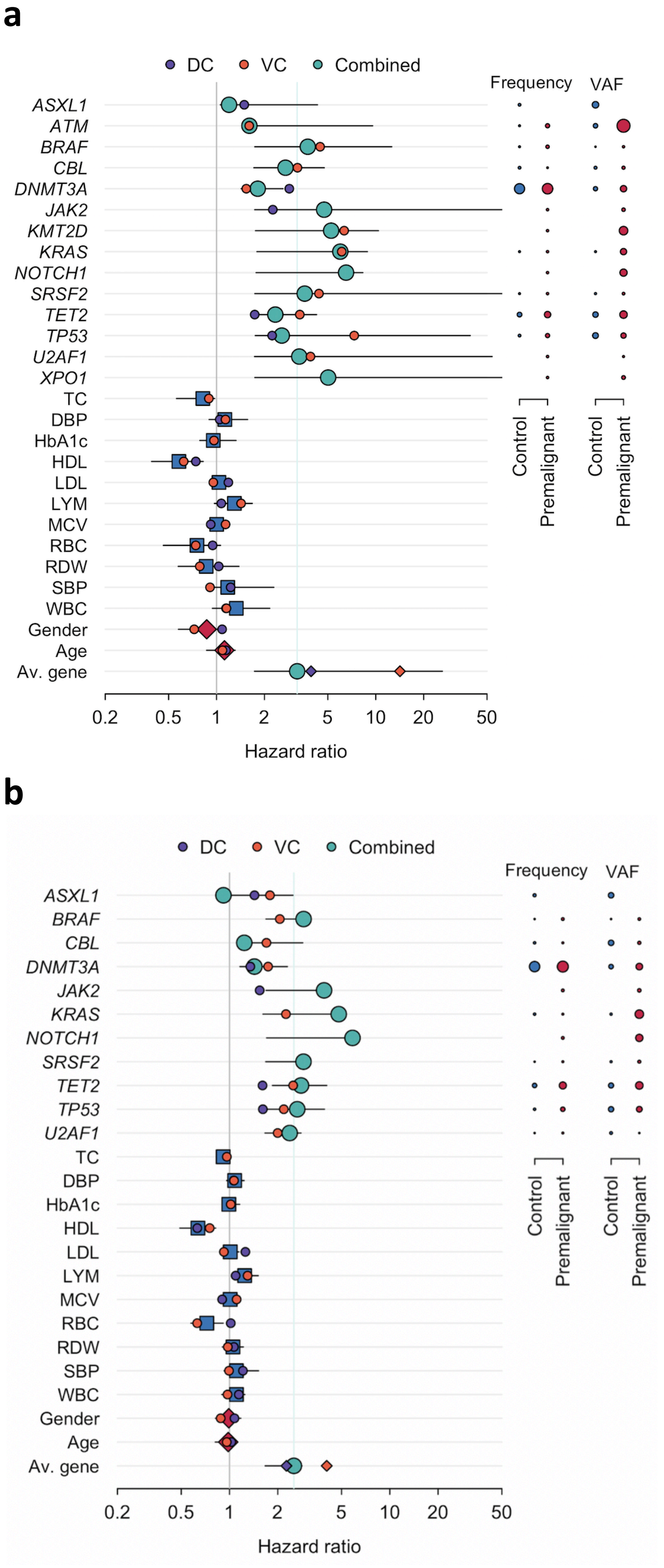


Figure 4.11 | Forest plots of hazard ratios for risk progression to lymphoid malignancy. a, Forest plot for Cox proportional hazards model using a 1:1 case control ratio and including all myeloid and lymphoid cancer genes. **b,** Model restricted to myeloid panel genes and incorporating an additional 189 age-and sex-matched controls for a 1:2 case:control ratio and hence more accurate estimates of risk associated with clinical factors and genes sequenced across both panels. Purple, orange and green circles indicate hazard ratios (HR) for the discovery (DC), validation (VC) and combined cohort, respectively. Horizontal lines denote 95% confidence intervals for the combined cohort. For each gene, the indicated HR applies to the 10-year risk of lymphoid blood cancer conferred by each 5% increase in mutation VAF. The green vertical line indicates the mean HR across all genes. Blue (controls) and red (pre-LN) circles to the right of the forest plot indicate the proportion of individuals with mutations in each gene and the average mutation VAF, which aids in the interpretation of hazard ratios. For example, *ATM*, a recurrent driver gene in several lymphoid malignancies, is almost exclusively mutated in pre-LN cases but at relatively high VAF, which translates into a modest HR for each 5% increase in clone size.

Overall, genetic and clinical parameters explained approximately 45% and 12% of the absolute variance in LN-free survival between individuals, respectively. Notably, clinical factors explained a comparable proportion of the variance. The coefficients for clinical variables were consistent between models trained on the discovery and validation cohorts (Figure 4.11a,b). Interestingly, lower HDL was associated with a modest but significant increase in risk of LN progression (Figure 4.11a,b). Consistent with this finding, lower total cholesterol was also associated with a smaller but still significantly increase in risk.

Unsurprisingly, models did not achieve anywhere near the predictive power observed for AML, with concordance and AUC both ≤0.7 for models trained on either cohort (Table 4.2). Nevertheless, this analysis yielded robust estimates of the risk conferred by lower HDL levels and mutations in *DNMT3A* and *TET2*, findings with compelling biological implications that warrant further investigation.

Table 4.2 Cox proportional hazard model performance

Cox proportional hazards model	Concordance	Standard error	Time-dependent AUC
VC data and fit	0.60	0.035	0.67
DC data and fit	0.70	0.029	0.64
VC fit DC data	0.58	0.035	0.60
DC fit VC data	0.60	0.027	0.67
Combined cohorts	0.67	0.022	0.67

*Derived from 100 bootstraps out-of-bag validation
DC, discovery cohort; VC, validation cohort

3. Discussion

The main aim of this experiment was to characterise the prevalence and genetic landscape of CH-PD in individuals who go on to develop a lymphoid neoplasm. To this end, I have deep sequenced peripheral blood specimens from 189 pre-LN cases and 189 age- and sex-matched controls using a much broader gene panel than has been applied in previous similarly sensitive assays for CH. To investigate potential enrichment for LN-associated mutations in older age, this study was extended to include samples from a further 234 healthy older individuals. Serial samples, including peri- and post diagnosis blood samples, provided insight into clonal dynamics and the relationship between CH-PD and future malignancy. Clinical metadata, including full blood count parameters and lipid profile, were analysed for any association with CH-PD or future LN risk. Genetic and clinical variables were then incorporated into predictive models to seek any significant risk factors for LN progression and assess their collective power to identify individuals at high risk of future LN development.

3.1 CH-PD frequently precedes LN diagnosis and is characterised by a diverse mutational spectrum

This work demonstrates that CH-PD becomes more prevalent among individuals who develop a lymphoid malignancy years before diagnosis and is characterised by a more diverse genetic landscape than that observed in pre-AML cases or in the general population. The experiment described in Chapter 3 demonstrated that pre-AML exhibits a mutational spectrum that closely overlaps with that seen in the general population but is enriched for mutations in particular genes. By contrast, the pre-LN cohort harboured rare events in a number of genes highly associated with LN pathogenesis and rarely if ever reported in the current CH literature, including *ATM, CCND3, POT1, HIST1H1E, XPO1, NOTCH1* and *NOTCH2* (Arber et al., 2016; Kandoth et al., 2013; Martincorena et al., 2017; Sabarinathan et al., 2017; Swerdlow et al., 2016). Among these, *ATM* was the most recurrently mutated in pre-LN cases, ranking third after *DNMT3A* and *TET2*. The genetic heterogeneity observed in the pre-LN cohort is reminiscent of the genomic landscapes of the most common lymphoid blood cancers in adults, which tend to be characterised by a large number of infrequently mutated putative cancer genes (Landau and Wu, 2013; Reddy et al., 2017; Sabarinathan et al., 2017; Swerdlow et al., 2016).

3.2 CH-PD as a biomarker for lymphoid blood cancer risk

Despite an overall more varied mutational spectrum in pre-LN CH-PD, the two top genes remained *DNMT3A* and *TET2*. Mutations in both of these genes, and in particular *TET2,* are implicated in both B- and T-cell lymphoid malignancies (Couronne et al., 2012; Dominguez et al., 2018; Haney et al., 2016a; Haney et al., 2016b; Mouly et al., 2018; Quivoron et al., 2011). TET2 deficiency in particular has been shown to increase HSC mutation rate and predispose to lymphoid and myeloid malignancies (Pan et al., 2017). However, the high frequency of *TET2/DNMT3A* mutations in in pre-LN CH-PD relative to lymphoid cancers, in conjunction with the results of the time series experiment, suggests that *DNMT3A*-mutated clones in particular often do not represent ancestors of the future cancer. Nevertheless, mutations in *DNMT3A* and *TET2* confer a significantly increased risk for progression to LN, with hazard ratios comparable to those observed in the AML prediction model (Figure 3.5 and Figure 4.11). Although speculative, there are several possible explanations for this observation. As alluded to in Chapters 1 and 3, it is possible that clones that are not phylogenetically related to the future malignancy are surrogate markers of selective pressures that impart a strong growth advantage on pre-malignant HSCs. There is increasing precedent for this hypothesis in the haematopoietic system and other tissues. For example, as discussed in depth in Chapter 5, activating mutations in *PPM1D*, a negative regulator of TP53, confer a selective advantage on HSCs in the context of cytotoxic therapy (Gibson et al., 2017; Hsu et al., 2018; Takahashi et al., 2017). *PPM1D*-mutated CH-PD is a biomarker of therapy-related AML risk, despite that the *PPM1D*-mutations often persist at low VAF alongside the evolving AML (Gibson et al., 2017; Gillis et al., 2017). Remarkably, a similar scenario has recently been described in oesophageal epithelium, which is increasingly populated by *PPM1D*-mutated clonal expansions with age (Yokoyama et al., 2019). Exposure to alcohol and smoking, strong risk factors for oesophageal cancer, were associated with expansion of *PPM1D*-mutated epithelial clones, though *PPM1D* is not a recurrent driver in oesophageal malignancies (Yokoyama et al., 2019).

Current understanding of the selective pressures influencing somatic evolution in the haematopoietic system remains limited. However, age-related increases in endogenous genotoxic stress and reduced HSC self-renewal capacity may be important factors (Pang et al., 2017; Yahata et al., 2011). It is plausible that inter-individual variation in the pace and nature of age-related processes may influence the spectrum of mutations that confer selective advantage on HSCs. In this context it is noteworthy that *TP53* and *ATM*, both critical mediators of DNA damage response and cell cycle checkpoint control (Roos et al., 2016), constituted the third and fourth most frequently mutated genes in this pre-LN cohort. Whilst this result warrants confirmation in larger studies, it is conceivable that some individuals experience more severe/earlier DNA-damage associated HSC senescence and that this favours expansion of clones with mutations that repress DNA-damage-induced apoptosis and cell cycle arrest. By extension, such individuals would likely be at higher risk of stochastic driver mutation acquisition and clonal evolution of any one of numerous pre-malignant clones. As mentioned in Chapter 3, Wong et al. recently reported a high prevalence of ‘bystander’ pre-leukaemic clones in AML patients at diagnosis, suggesting that their leukaemia arose from one of many candidate pre-malignant HSCs (Wong et al., 2015a).

3.3 RDW and lymphoid neoplasm risk

Notably, RDW was not significantly increased among pre-LN cases, in contrast to the scenario observed for pre-AML. As discussed in Chapter 3, higher RDW has previously been associated with CH in the general population (Jaiswal et al., 2014). However, we have shown that comparing pre-AML cases and controls with CH-PD revealed that RDW could help distinguish pre-AML (including cases without detectable CH-PD) from CH in individuals who did not develop a blood cancer during follow-up. The association between higher RDW and risk of developing AML was validated in a large electronic medical records dataset. It is possible that a weaker correlation does exists between pre-LN and RDW that this study was underpowered to detect, as hinted by the subtle trend discernible on KM analysis (Figure 4.8). However, this result nevertheless suggests that RDW is not a universally strong discriminator between indolent and pre-malignant CH-PD. This experiment may mask lymphoid cancer subtype-specific associations between RDW and warrants further investigation.

3.4 Lower high-density lipoprotein levels and lymphoid cancer risk

Among all clinical variables analysed, only HDL levels differed significantly between pre-LN cases and controls. The association between lower HDL and future LN was corroborated by Cox proportional hazards modelling, which identified a modestly increased risk of LN with lower HDL and total cholesterol (Figure 4.11a,b). Hypocholesterolaemia is a common finding in lymphoma and leukaemia patients, and has also been reported in association with some solid tumour types (Lim et al., 2007; Pirro et al., 2018). Lower HDL in particular has been previously identified as a preclinical feature of non-Hodgkin lymphoma discernible years before diagnosis (Lim et al., 2007). Low HDL at lymphoma diagnosis has also been correlated with poorer prognosis (Matsuo et al., 2017). The mechanisms underlying these observations are unclear with no compelling evidence of a causative link between low cholesterol and haematological malignancies (Pirro et al., 2018). However, numerous studies report that lymphoma cells and leukaemia blasts have higher HDL and/or LDL uptake receptor activity (Goncalves et al., 2005; Vitols et al., 1990; Vitols et al., 1985) and that cholesterol metabolism may represent a viable therapeutic target for several mature B-cell malignancies (McMahon et al., 2017). It is therefore possible that pre-malignant CH displays similar behaviour, leading to reductions in circulating levels of HDL even years prior to overt malignant transformation. This is a particularly intriguing hypothesis in view of the emerging causal role of CH-PD in atherosclerosis (Fuster et al., 2017; Jaiswal et al., 2017; Sano et al., 2018a). It is even conceivable that plaque-resident clonal haematopoietic cells may accelerate atheroma progression in part by increasing lipid accumulation at sites of inflamed endothelium.

3.5 Experiment limitations and future directions

This experiment has several important limitations. Firstly, the pre-LN cohort encompasses diverse diseases presenting over a long period during which histopathological classification schemes and diagnostic guidelines evolved considerably (Campo et al., 2011; Swerdlow et al., 2016). This limited the scope to investigate the natural history of or distinct genetic/clinical risk factors for individual cancer types. Furthermore, structural events, particularly translocations involving the immunoglobulin heavy chain (IGH) genes and numerical chromosomal aberrations, are frequent initiating events of lymphoid malignancies and their detection requires a much broader and more costly sequencing approach (Bolli et al., 2014; Landau et al., 2015). While the main aim of this experiment was to characterise the point mutation spectrum of CH-PD in pre-LN and investigate the predictive value of both putative ancestral and 'bystander' clones in assessing risk of progression, the power of predictive models would likely be increased by screening for subclonal large copy number changes and recurrent translocations.

Moreover, these results provide further evidence that malignant and cardiovascular adverse outcomes associated with CH might be linked. The association of lower HDL with LN progression risk, in conjunction with the clinical AML prediction model described in Chapter 3, hint that there may be unifying features of 'high risk' CH that could eventually help define a useful biomarker and/or therapeutic target. Hence this experiment reinforces the need for future studies of CH to correlate genetics with detailed clinical and phenotypic metadata and to try to move beyond investigating malignant and cardiometabolic disease associations in isolation.