Chapter 2

Materials and Methods

1. Patient samples

1.1 Pre-AML and control peripheral blood samples (Chapter 3)

For the study of the pre-clinical evolution of AML described in Chapter 3, samples from pre-AML cases and age- and sex-matched controls were collected by collaborators at the European Prospective Investigation into Cancer and nutrition (EPIC) study (Riboli et al., 2002). Samples were divided into discovery and validation cohorts and sequenced at the Wellcome Sanger Institute and the University of Toronto, respectively (see section Methods section 2.1 and 2.2).

Written informed consent was obtained from all participants in accordance with the Declaration of Helsinki and protocols approved by the relevant ethics committees (IARC Ethics Committee approval #14-31, the Weizmann institute of science Ethics board approval #60-1 and East of England - Cambridgeshire and Hertfordshire Research Ethics Committee reference number 98CN01). *De novo* AML patients were identified based on the following ICD9 codes: 9861/3 9860/3 9801/3 9866/3 9891/3 9867/3 9874/3 9840/3 9872/3 9895/3 9873/3. All patients provided peripheral blood samples from which the buffy coat fractions were separated and aliquoted for long-term storage in liquid nitrogen prior to DNA extraction.

1.1.1 Discovery cohort samples

A total of 509 DNA samples were collected from individuals upon enrolment into the EPIC study between 1993 and 1998 across 17 different centres (Riboli et al., 2002). The pre-AML group comprised 95 individuals who developed *de novo* AML an average of 6.37 years

(IQR=4.88 years) after the sample was collected. The control group included 414 age and gender matched individuals with no record of haematological disorders (mean follow-up period 11.9 years). The median age at recruitment was 56.75 years (range 36.08 to 74.42). In order to minimize any possible demographic biases, an approximate 1:4.5 pre-AML to control ratio was maintained across the different centres. Discovery cohort (DC) sample metadata is detailed in Appendix 1.

1.1.2 Validation cohort samples

Samples were ascertained from individuals enrolled in the EPIC-Norfolk longitudinal cohort study between 1994 and 2010 (Day et al., 1999). Samples and clinical metadata were available from 37 AML patients (of which 8 were already included in the discovery cohort) and 262 age- and gender-matched controls without a history of cancer or any haematological condition. The median time between the first blood sampling and AML diagnosis was 12.3 years (IQR 8.3 years). The median follow-up period for the control cohort was 17.5 years (IQR 3.8). For 12 individuals in the pre-AML cohort, 2-3 blood specimens were available, taken a median of 3.4 years apart. Of the 262 controls, 141 had multiple blood samples available, spanning a median of 10.5 years. Blood counts and other clinical parameters were available for all study participants (Appendix 2).

1.2 Childhood cancer survivor cohort samples (Chapter 5)

We obtained peripheral blood DNA samples from patients enrolled on long-term follow-up after treatment for a paediatric malignancy and from 3 age-matched controls with no cancer history. Written informed consent was obtained for sample collection and DNA sequencing from all patients or their guardian in accordance with the Declaration of Helsinki and protocols approved by the relevant institutional ethics committees (approval numbers 09REG2015, 1-09/12/2015). The median age at cancer diagnosis was 4.5 years, and the commonest malignancies were acute lymphoblastic leukaemia (n=21), neuroblastoma (n=17) and non-Hodgkin lymphoma (n=10). Nineteen patients had received a hematopoietic stem cell transplant (8 allogeneic and 11 autologous). The median interval between completion of

cancer treatment and blood sampling was 6 years (range 2 – 25). Patient characteristics are summarized in Table 4.1 and individual sample details are shown in Appendix 3.

1.3 Paediatric therapy-related myeloid neoplasm samples (Chapter 5)

We obtained bilateral bone marrow biopsies and serial peripheral blood DNA samples from a paediatric neuroblastoma patient who developed a therapy-related myeloid neoplasm. Written informed consent was obtained for sample collection and DNA sequencing from the guardian in accordance with the Declaration of Helsinki and protocols approved by the relevant institutional ethics committees (REC reference 16/EE/0394).

1.4 Pre-lymphoid neoplasm cohort and controls

For the study of the pre-clinical evolution of lymphoid neoplasms (LN) described in Chapter 4, samples from pre-LN cases and age- and sex-matched controls were collected by collaborators at the European Prospective Investigation into Cancer and nutrition (EPIC)-Norfolk study (Day et al., 1999; Riboli et al., 2002).

Written informed consent was obtained from all participants in accordance with the Declaration of Helsinki and protocols approved by the relevant ethics committees (IARC Ethics Committee approval #14-31, East of England - Cambridgeshire and Hertfordshire Research Ethics Committee reference number 98CN01). Pre-LN cases were identified based on the following ICD10 codes: C81*, C82*, C83*, C84*, C85*, C86*, C87*, C88*, C89*, C90*, C91*. All patients provided peripheral blood samples from which the buffy coat fractions were separated and aliquoted for long-term storage in liquid nitrogen prior to DNA extraction.

2. Library preparation and sequencing

2.1 Targeted sequencing of discovery cohort pre-AML and control samples (Chapter 3)

Library preparation and sequencing of discovery cohort samples was performed by Sagi Abelson and colleagues (Princess Margaret Cancer Centre, University Health Network, Toronto). Targeted deep sequencing was performed using error-corrected sequencing (ECS) as detailed below.

Ligation of sequencing adaptors. Shearing of genomic DNA, preparation of pre-capture sequencing libraries, hybridization-based enrichment, assessment of the libraries quality and enrichment following hybridization were performed as previously described (Newman et al., 2014). Briefly, 100ng of genomic DNA was sheared before library construction (KAPA Hyper Prep Kit #KK8504, Kapa Biosystems) with a Covaris E220 instrument using the recommended settings for 250bp fragments. Following end repair and A-tailing, adapter ligation was performed using 100-fold molar excess of Molecular Index Adapter. Library clean-up was performed with Agencourt AMPure XP beads (Beckman-Coulter) and the ligated fragments were then amplified for 8 cycles using 0.5µM Illumina universal and indexing primers.

Target capture. Targeted capture was carried out on pools containing 3 indexed libraries. Each pool of adaptor-ligated DNA was combined with 5µl of 1mg/ml Cot-I DNA (Invitrogen), and 1 nmol each of xGEN Universal Blocking Oligo – TS-p5, and xGen Universal Blocking Oligo – TS-p7 (8nt). The mixture was dried using a SpeedVac and then re-suspended in 1.1µl water, 8.5µl NimbleGen 2× hybridization buffer and 3.4µl NimbleGen hybridization component A. The mixture was heat denatured at 95°C for 10 minutes before addition of 4µL of xGen Lockdown Probes (xGen® AML Cancer Panel v1.0, 3pmol). The panel was designed to include all genes recurrently mutated in the 2013 TCGA study of AML (TCGA et al., 2013). Each pool was then hybridized at 47°C for 72 hr. Washing and recovery of the captured DNA was performed according to the manufacturer's specifications. Briefly, 100µl of clean streptavidin beads was added to each capture. Following separation and removal of supernatant on a magnet, 200µL 1X Stringent Wash Buffer was added and the reaction was incubated at 65°C for 5 min. Supernatant containing unbound DNA was removed before repeating the high stringency wash one additional time. Next, the bound DNA was washed as follows: 1) 200µl 1X Wash Buffer I and separation of the supernatants by magnetic separation, 2) 200µl 1X Wash Buffer II following magnetic separation, 3) 200µl 1X Wash Buffer III and removal of the supernatants using magnetic separation. The captured DNA on beads was resuspended in 40µl of Nuclease-Free water before dividing the total volume into 2 PCR tubes and subjecting the libraries to 10 cycles of post-capture amplification (manufacturer recommended conditions; Kapa Biosystems). Prior to sequencing, libraries were spiked in with 2% PhiX.

2.2 Targeted sequencing of validation cohort pre-AML and control samples and AML diagnostic specimens (Chapter 3)

This section describes the sequencing methods for the validation cohort (VC) pre-AML and control samples discussed in Chapter 3.

Targeted sequencing was performed using a custom cRNA bait set (SureSelect, Agilent, UK, ELID #0537771, contributed by Dr Elli Papaemmanuil and Dr Peter Campbell) designed complementary to all coding exons of 111 genes implicated in myeloid leukaemogenesis (Appendix 4). Genomic DNA was sheared using the Covaris M220. Equimolar pools of 10 libraries were prepared and sequenced on the Illumina HiSeq 2000 using 75 base paired-end sequencing as per Illumina and Agilent SureSelect protocols.

2.3 Multiplex PCR design and sequencing (Chapter 5)

This section describes the sequencing strategy used to screen peripheral blood samples from childhood cancer survivors for clonal haematopoiesis (Chapter 5). The multiplex PCR panel was designed by Dr Naomi Park and Dr George Vassiliou as detailed in a published protocol (Park and Vassiliou, 2017) and I performed PCR experiments with supervision from Dr Park. Primers were designed using mprimer software (Shen et al., 2010) and checked for specificity using MFEprimer (Qu and Zhang, 2015). In order to minimise primer dimer formation, primers were synthesised to include a single 2'-O-Methyl base substitution, one base from the 3'-end. The multiplex PCR amplifies 32 regions of 14 genes frequently mutated in CH or t-MN (Table 4.2) (Bowman et al., 2018; McNerney et al., 2017). This is an extension of a previously validated assay (McKerrell et al., 2015) to include all coding exons of *TP53* and *PPM1D*, genes implicated in t-MN pathogenesis (Gibson et al., 2017; Hsu et al., 2018; McNerney et al., 2017). Primer sequences are detailed in Appendix 5. Amplicon libraries were sequenced on the Illumina MiSeq platform as detailed in Park et al. (Park and Vassiliou, 2017).

2.4 Targeted sequencing using a custom pan-haematological cancer panel

This section describes the sequencing methods for the diagnostic AML bone marrow samples discussed in Chapter 3, the pre-lymphoid cancer specimens and controls discussed in Chapter 4 and the paediatric therapy-related myeloid neoplasm described in Chapter 5. Targeted sequencing was performed using a custom cRNA bait set (SureSelect, Agilent, UK, ELID ID: 3129591) designed complementary to all coding exons of 95 genes recurrently mutated in myeloid and lymphoid haematological cancers, including the genes most

frequently implicated in paediatric MPN/MDS (Appendix 6). Genes implicated in lymphoid neoplasms were selected with input from Dr Philip Beer. Genomic DNA was sheared using the Covaris M220. Equimolar pools of 10 libraries were prepared and sequenced on the Illumina HiSeq 2000 using 75 base paired-end sequencing as per Illumina and Agilent SureSelect protocols.

2.5 Whole genome sequencing

Whole genome sequencing of peripheral blood DNA (Chapter 5) was performed by 150-bp- paired-end sequencing on the Illumina Hiseq X10 platform. The Illumina no-PCR protocol was followed to construct short insert libraries, prepare flow cells and generate clusters (Kozarewa et al., 2009).

3. Variant calling

3.1 Variant calling in pre-AML and control samples

Variant filtering and annotation for the discovery cohort (section 3.1.1) and validation cohort (section 3.1.2) was performed by Dr Sagi Abelson and myself, respectively. After filtering and annotation, both datasets were combined and driver mutation calling and additional artefact filtering was performed by me as detailed in sections 3.1.3 and 3.1.4.

Figure 2.1

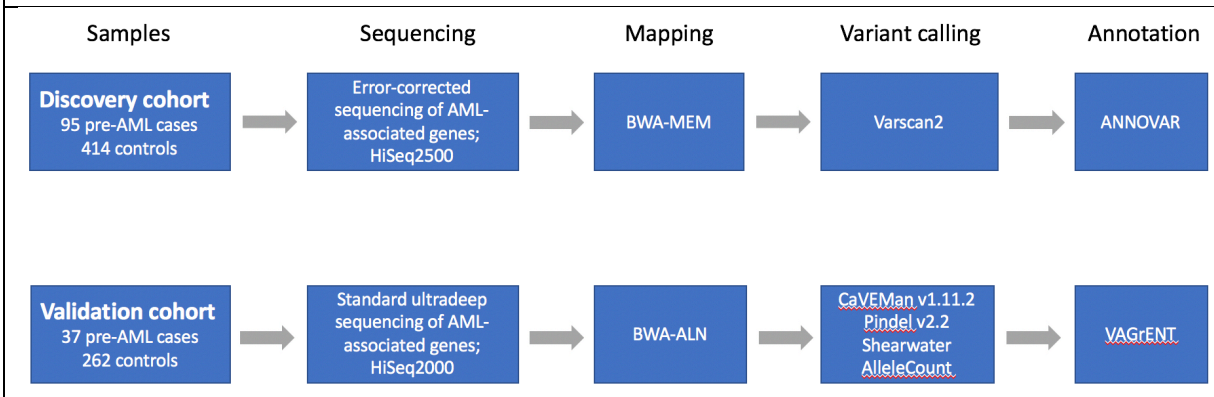

Figure 2.1 | Overview of Chapter 3 experimental design. Discovery and validation cohort pre-AML and control samples were processed using different sequencing and bioinformatic pipelines, summarised above.

3.1.1 Discover cohort variant calling and error correction

126bp paired-end read sequencing data from the Illumina HiSeq2500 platform was converted to fastq format. The 2bp molecular barcode information of each read was trimmed and incorporated into the read name. The thymine nucleotide required for ligation was removed from the sequences. The processed FASTQ files were then aligned to the hg19 reference genome using the Burroughs-Wheeler Aligner (BWA-MEM) (Li and Durbin, 2010). Indel-re-alignment was performed using GATK (McKenna et al., 2010). An in-house algorithm was written to collapse read families that share the same molecular barcode sequence, the left most genomic position of where each read of the pair maps to the reference and the CIGAR string. Families comprised of at least 2 reads were used to generate consensus reads (CR) and a consensus base was called when there was at least 70% agreement. When a consensus base was called, it was assigned with the maximum base quality score observed in its corresponding pre-collapsed reads. Furthermore, when possible, duplex reads (DR) were generated from two CR, from a singleton read (SR) and a CR, or from two SR (Kennedy et al., 2014). For each sequenced sample, we generated two BAM files, called bam1 and bam2. Bam1 consists of DR, CR and singleton reads, thereby including some error corrected and non-error corrected reads. Bam2 consists of DR and CR but not singleton reads. Both files were then analysed to detect single nucleotide variants (SNVs) and small insertions and deletions (indels) using Varscan2 (Koboldt et al., 2012). In order to further remove sequencing artefacts and improve sensitivity, we applied a two-step statistical polishing approach that models the error rate at each sequenced genomic position. For both steps, bam1 was used and all the samples except the sample being investigated were included for error rate modelling. At step one, as previously described (Newman et al., 2014), the error rates were modelled by fitting weibull distribution curves to the non-reference allele fractions. SNVs with allele fractions that were statistically distinguishable from the background error rates were further analysed. At Step 2, the coverage of the non-reference allele fractions was considered by using linear line fitting that describes the negative correlation that exist between the log (non-reference allele fraction) and the corresponding log(coverage) values. This allowed us to estimate different error rates at different coverage depths. Indel errors were filtered using barcode mediated error correction alone. At least 10 CR, 5 supporting reads on the forward strand, 5 supporting reads on the reverse strand, and 2 DR were required to call an indel. Variants were

annotated using Annovar (Yang and Wang, 2015). Additional post-processing steps applied to data from both the discovery and validation cohorts are detailed in section 3.1.3.

3.1.2 Validation cohort variant calling

Sequencing reads were aligned to the reference genome (GRCh37d5) using the Burrows-Wheeler aligner (BWA-ALN)(Li and Durbin, 2009). Unmapped reads, PCR duplicates and reads mapping to regions outside the target regions (merged exonic regions + 10bp either side of each exon) were excluded from analysis. Sequencing depth at each base was assessed using Bedtools coverage v2.24.0 (Quinlan and Hall, 2010).

Substitutions

Somatic single nucleotide variants (SNVs) were called using Shearwater, an algorithm developed for detecting subclonal mutations in deep sequencing experiments (https://github.com/gerstung-lab/deepSNV v1.21.5) (Gerstung et al., 2012; Gerstung et al., 2014; Martincorena et al., 2015) considering only reads with minimum nucleotide and mapping quality of 25 and 40, respectively. This algorithm models the error rate at individual loci using information from multiple unrelated samples. Additionally, allele counts at the recurrent AML mutation hotspots listed in section 3.1.4 were generated using an in-house script (https://github.com/cancerit/alleleCount) and manually inspected in the Jbrowse genome browser (Buels et al., 2016). To further complement our SNV calling approach, we applied an extensively validated in-house version of CaVEMan v1.11.2 (Cancer Variants through Expectation Maximization)(Stephens et al., 2012). CaVEMan compares sequencing reads between study and nominated normal samples and uses a naïve Bayesian model and expectation-maximization approach to calculate the probability of a somatic variant at each base (https://github.com/cancerit/CaVEMan). Post-processing filters required that the following criteria were met for CaVEMan to call a somatic substitution:

1) If coverage of the mutant allele was less than 8, at least one mutant allele was detected in the first 2/3 of the read.
2) Less than 3% of the mutant alleles with base quality ≥ 15 were found in the nominated normal sample.
3) Mean mapping quality of the mutant allele reads was ≥ 21.

4) Mutation does not fall in a simple repeat or centromeric region.
5) Fewer than 10% of the reads covering the position contained an indel according to mapping.
6) Less than 80% of the reads report the mutant allele at the same read position.
7) At least a third of the reads calling the variant had a base quality of 25 or higher.
8) Not all mutant alleles reported in the second half of the read.
9) Position does not fall within a germline insertion or deletion.

The following additional post-processing criteria were applied to all SNV calls:

1) Minimum VAF 0.5% with a minimum of 5 bidirectional reads reporting the mutant allele (with at least 2 reads in forward and reverse directions).
2) No indel called within a read length (75bp) of the putative substitution.

Small insertions and deletions

Small insertions and deletions were sought using two complementary approaches. Firstly, an in-house version of Pindel v2.2 (Raine et al., 2015) (https://github.com/cancerit/cgpPindel) was applied. We additionally used the aforementioned Shearwater algorithm (Gerstung et al., 2012; Gerstung et al., 2014; Martincorena et al., 2015) in order to increase sensitivity for indels present at low VAF. VAF correction was performed using an in-house script (https://github.com/cancerit/vafCorrect). Post-processing filters required that the following criteria were met for a variant to be called:

1) Minimum of 5 reads supporting the variant with minimum of 2 reads in each direction. For Pindel, the total read count was based on the union of BWA and Pindel reads reporting the mutant allele.
2) Minimum VAF 0.5%
3) Variant not present within an unmatched normal panel of approximately 400 samples.
4) No reads supporting the variant identified in the nominated normal sample.

Mutations were annotated according to ENSEMBL version 58 using VAGrENT (Menzies et al., 2002) for transcript and protein effects (https://github.com/cancerit/VAGrENT) and Annovar (Yang and Wang, 2015) for additional functional annotation.

3.1.3 Additional post-processing filters applied to all data

The following variants were flagged for additional inspection for potential artefacts, germline contamination or index-jumping event:

1) Any mutant allele reported within 75bp of another variant.
2) Any mutant allele with a population allele frequency > 1 in 1000 according to any of five large polymorphism databases: ExAC, 1000 Genomes Project, ESP6500, CG46, Kaviar that is not a canonical hotspot driver mutation with COSMIC recurrence > 100.
3) Mutations that were present in > 10% of the control cohort but not recurrent in COSMIC were flagged as potential germline variants or sequencing artefact.
4) As artefactual indels tend to be recurrent, any indels occurring in >2 samples were flagged for additional inspection.

3.1.4 Curation of oncogenic variants

Putative oncogenic variants were identified according to evidence for functional relevance in AML as previously described and used to define CH-PD (Gerstung et al., 2017; Papaemmanuil et al., 2016).

Variants were annotated as likely driver events if they fulfilled any of the following criteria:

1) Truncating mutations (nonsense, essential splice site or frameshift indel) in the following genes implicated in AML pathogenesis by loss-of-function: *NF1, DNMT3A, TET2, IKZF1, RAD21, WT1, KMT2D, SH2B3, TP53, CEBPA, ASXL1, RUNX1, BCOR, KDM6A, STAG2, PHF6, KMT2C*.
2) Truncating variants in *CALR* exon 9.
3) *JAK2* V617F
4) *FLT3* ITD

5) Non-synonymous variants at the following hotspot residues:
    a. *CBL* E366, L380, C384, C404, R420, C396
    b. *DNMT3A* R882
    c. *FLT3* D835
    d. *IDH1* R132
    e. *IDH2* R172, R140
    f. *KIT* W557, V559, D816
    g. *KRAS* A146, Q61, G13, G12
    h. *MPL* W515
    i. *NRAS* Q61, G12, G13
    j. *SF3B1* K700, K666
    k. *SRSF2* P95
    l. *U2AF1* Q157, R156, S34
6) Non-synonymous variants reported at least 10 times in COSMIC with VAF < 42% and population allele frequency < 0.003.
7) Non-synonymous variants clustering within a functionally validated domain or within 4 amino acids of a hotspot variant with population allele frequency < 0.003 and VAF < 42%.
8) Non-synonymous variants reported in COSMIC > 100 times with population allele frequency < 0.003 regardless of VAF.

This driver curation strategy inevitably runs a small risk of including germline variants in familial AML genes, e.g., *RUNX1*. However, in most settings, where a matched constitutional DNA sample is likely to be unavailable, this seems the best approach.

Of note, the entire validation cohort included 37 pre-AMLs, 8 of these were also included in the original discovery cohort and therefore were excluded from the validation cohort for downstream analysis. Both the discovery and the validation cohorts sourced samples from different centres participating in the EPIC study, hence the overlap. However, discovery and validation cohorts were sequenced by two independent research groups using different methods, as described above. Putative driver mutations detected for the duplicated samples by the two different methods were highly similar. All 9 driver mutations detected in

the discovery cohort with VAF>0.015 were detected in the validation cohort samples, while 8 other mutations (7 in TET2 or DNMT3A) with lower VAFs escaped validation. The latter is probably due to the higher VAF cut-off applied to the validation cohort sequencing method and the stochastic failure to sample a small clone in two independent experiments.

3.2 Variant calling from multiplex PCR sequencing

Reads were aligned to human genome build GRCh37d5 using the Burrows-Wheeler Aligner (Li and Durbin, 2010) and analysed for somatic single nucleotide variants and indels. Allele counts across target hotspots were generated using an in-house script (https://github.com/cancerit/alleleCount), considering only loci with ≥1000 reads and minimum base and mapping quality of 25 and 35, respectively. In order to identify SNV and indels in *TP53* and *PPM1D*, 3 variant callers were applied: Shearwater (https://github.com/gerstung-lab/deepSNV v1.21.5)(Gerstung et al., 2012; Gerstung et al., 2014; Martincorena et al., 2015), cgpPindel v2.2 (Raine et al., 2015) and CaVEMan v1.11.2 (Cancer Variants through Expectation Maximization, https://github.com/cancerit/CaVEMan)(Stephens et al., 2012) as describe in section 3.1.2 above.

3.3 Variant calling for non-AML pre-malignant samples and controls

SNV and indel calling was performed as described in 3.1.2 and 3.1.3. The strategy for curating putative driver variants was adjusted to account for the greater number of genes included in the larger bait panel (Appendix 6). Specifically, variants were flagged as candidate driver events if they fulfilled any of the following criteria:

1) Nonsense or frameshift mutations in the following genes: *ARID1A, ASXL1, ATM, B2M, BCOR, BCORL1, CALR, CDKN2A, CEBPA, CREBBP, CSF1R, CSF3R, CUX1, DNMT3A, EP300, FBXW7, KDM6A, KMT2C, KMT2D, NF1, NOTCH2, NPM1, PAX5, PHF6, POT1, PPM1D, PRDM1, PTEN, RAD21, SETD2, SOCS1, STAG2, TET2, TNFAIP3, TNFRSF14, TP53, WT1, ZRSR2*
2) Splice site mutations in the following genes: *ARID1A, ATM, BCOR, CBL, CD79B, CDKN2A, CUX1, DNMT3A, KDM6A, NF1, PAX5, PHF6, PRDM1, PTEN, SETD2, STAG2, WT1, ZRSR2*

3) Missense mutations in the following genes were considered if they passed SNP and artefact filters and had support as candidate drivers based on relevant literature (Tate et al., 2019): *ARID1A, ASXL1, ATM, B2M, BCL6, BCORL1, BRAF, CALR, CARD11, CBL, CD79B, CDKN2A, CEBPA, CREBBP, CSF1R, CSF3R, CUX1, DNMT3A, EP300, ETNK1, EZH2, FBXW7, FLT3, GATA2, GNAS, H3F3A, IDH1, IDH2, IL7R, JAK2, KIT, KMT2D, KRAS, MPL, MYD88, NF1, NOTCH1, NOTCH2, NRAS, PAX5, PDGFRA, PHF6, PIM1, POT1, PPM1D* (exon 6), *PRDM1, PTEN, PTPN11, RAD21, SETBP1, SETD2, SF3B1, SRSF2, STAG2, STAT3, TET2, TNFRSF14, TP53, U2AF1, WT1, XPO1, ZEB1, ZRSR2*

4) Non-synonymous variants reported at least 10 times in COSMIC with VAF < 35% and population allele frequency < 0.003.

5) Non-synonymous variants clustering within a functionally domain or within 4 amino acids of a hotspot variant with population allele frequency < 0.003 and VAF < 35%.

6) Non-synonymous variants reported in COSMIC > 150 times with population allele frequency < 0.003 regardless of VAF.

3.4 Screening for pathogenic germline variants

All mutations flagged by SNP filters (VAF > 0.42 and present in ExAC, 1000 Genomes Project, ESP6500, CG46 or Kaviar databases) were screened against the ClinVar database (Landrum et al., 2016) and Human Gene Mutation Database (HGMD) (Stenson et al., 2003) to identify potential cancer predisposition germline variants.

3.5 Variant calling from whole genome sequences (Chapter 5)

Whole genome sequences were mapped to the GRCh37d5 reference genome using the Burroughs-Wheeler Aligner (BWA-mem) (Li and Durbin, 2010). The Cancer Genome Project (Wellcome Trust Sanger Institute) variant calling pipeline was used to call somatic mutations which includes the following algorithms: CaVEMan (1.11.0)(Jones et al., 2016) for substitutions; an in-house version of Pindel (2.2.2; github.com/cancerit/cgpPindel)(Raine et al., 2015) for indels; BRASS (5.3.3; github.com/cancerit/BRASS) for rearrangements (Li et al., 2017), and ASCAT NGS (4.0.0) for copy number aberrations (Van Loo et al., 2010). In addition

to filters inherent to the CaVEMan algorithm, the following post-processing filtering criteria were applied for substitutions: a minimum two reads in each direction reporting the mutant allele; at least ten fold coverage at the mutant allele locus; minimum variant allele fraction 5%; no insertion or deletion called within a read length (150bp) of the putative substitution; no soft-clipped reads reporting the mutant allele; median BWA alignment score of the reads reporting the mutant allele ≥ 140. The following variants were flagged for additional inspection for potential artefacts, germline contamination or index-jumping event: any mutant allele reported within 150bp of another variant; any mutant allele with a population allele frequency > 1 in 1000 according to any of five large polymorphism databases: ExAC, 1000 Genomes Project, ESP6500, CG46, Kaviar.

To identify potential driver events in whole genome data, I considered variants presenting in established cancer genes (Tate et al., 2019). Tumour suppressor coding variants were considered if they were annotated as functionally deleterious by an in-house version of VAGrENT (http://cancerit.github.io/VAGrENT/) (Menzies et al., 2002), or alternatively if they were disruptive rearrangement breakpoints or homozygous deletions. Additionally, homozygous deletions were required to be focal (<1 Mb in size) or constitute a known contiguous gene syndrome implicated in t-MN (McNerney et al., 2017). Mutations in oncogenes were considered driver events if they were located at previously reported canonical hot spots (point mutations) or amplified the intact gene. Amplifications also had to be focal (<1 Mb) and increase the copy number of oncogenes to a minimum of 5 copies.

3.6 Copy number variation in targeted sequencing data

To detect copy number aberrations in the paediatric t-MN case discussed in Chapter 5, I applied FACETS (Fraction and Allele-Specific Copy Number Estimates from Tumor Sequencing), an allele-specific copy number analysis (ASCN) method (Shen and Seshan, 2016).

4. Predictive modelling

Regularised logistic and Cox proportional hazards regression approaches were tested in generating the predictive models described in Chapters 3 and 4.

Dr Moritz Gerstung wrote the initial version of the code for Chapter 3 and closely supervised all further iterations of the models described in Chapter 3. The code for the models described in Chapter 4 was written by me using a very similar analysis framework and methods as in Chapter 3.

4.1 Cox proportional hazards model with random effects

We used a Cox proportional hazards regression to model haematological malignancy-free survival as previously described (Gerstung et al., 2017). We used random effects for the Cox proportional hazards model in the CoxHD R package developed by Dr Gerstung (http://github.com/gerstung-lab/CoxHD). A key strength of this approach is the ability to include many variables in one model while shrinking estimated effects for parameters with weak support in the data, thus controlling for overfitting. We used weighting to minimise the biases introduced by the artificial case-control ratio (Antoniou et al., 2005) and calculated hazard ratios relative to the (approximate) true cumulative incidence of either AML (Chapter 3) or all lymphoid malignancies (Chapter 4) in the given age range over a follow up of 10-20 years. Full details of model derivation and comparisons with alternative methods are included in the accompanying code (Appendix 7). In brief, variables comprised age, gender, the variant allele fraction of putative driver mutations and selected clinical variables when available. We performed agnostic imputation of missing variables by mean and linear rescaling of gene variables by a power of 10 to a magnitude of 1.

All blood samples taken within 6 months of cancer diagnosis were excluded from model training. Among the pre-AML samples (Chapter 3), 4 individuals were thus removed from the discovery cohort. For one individual in the validation cohort who provided 3 pre-diagnostic samples, the 3rd sample was taken within this time frame and was also excluded (though their older samples allowed this individual to remain in the modelling analysis).

For each model, the following measures of predictive accuracy were evaluated before and after leave-one-out cross-validation (LOOCV): (i) concordance (C)(Harrell et al., 1996), (ii) time-dependent area under the receiver-operating characteristic curve (AUC)(O'Quigley et al., 2005) and (iii) Uno's estimator of cumulative/dynamic AUC (Uno et al., 2007). Coefficient confidence intervals were calculated using 100 bootstrap samples.

Concordance measures were obtained using the survConcordance() function implemented in the survival R package (Therneau and Grambsch, 2000). Dynamic AUC was calculated with AUC.uno() implemented in the survAUC package (Heagerty et al., 2000). Time-independent AUC was calculated by the performance function implemented in the ROCR package (Sing et al., 2005). The expected incidence of each haematological malignancy was calculated from the UK office of national statistics, available at http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/. All-cause mortality data was obtained from the office of national statistics (https://www.ons.gov.uk/).

4.2 Ridge regularised logistic regression

Using the same covariates as in the Cox proportional hazard models, we fitted a ridge regularised logistic regression model to dichotomised outcome data. While logistic regression is a common choice for case-control analyses, a downside of this approach is the inability to explicitly use time-dependent covariates. The penalty parameter was chosen using LOOCV on the full cohort; this value was then used on the discovery and validation cohorts to yield the same scaling of coefficients. Confidence intervals were calculated using 100 bootstrap samples. Fitting was performed using the glmnet R package (Simon et al., 2011). AUC as the primary performance metric was calculated using the ROCR R package (Sing et al., 2005).