University of Southampton

Faculty of Medicine

Human Development & Health

The Causes and Consequences of Clonal Haematopoiesis

by

Ahmed Abdelrazek Zaky Dawoud

ORCID ID: 0000-0003-0164-7773

Thesis for the degree of Doctor of Philosophy

January 2023

University of Southampton

Abstract

Faculty of Medicine

Human Development & Health

Doctor of Philosophy

The Causes and Consequences of Clonal Haematopoiesis

by

Ahmed Abdelrazek Zaky Dawoud

Introduction: Over the past decade, the availability of large population studies has allowed a detailed exploration of the relationship between genetics and clinical phenotypes. Clonal haematopoiesis (CH) is the expansion of blood cells with genetic features that are often observed in patients with haematological malignancies, particularly myeloid neoplasms. CH is a common finding in elderly individuals and associated with an elevated risk of developing haematological malignancies, cardiovascular diseases, and all-cause mortality. My study has four main aims. First, to characterise the inherited and environmental risk factors associated with myeloid CH. Second, to characterise the impact of myeloid CH on the risk of developing chronic inflammation-related diseases. Third, to investigate the utility of CH measures to predict the risk of myeloid malignancies. Fourth, to identify risk factors associated with age-related loss of the Y-chromosome (LOY) in men and its relationship to CH.

Methods: The UK Biobank represents a unique genetic and phenotypic dataset of about 500,000 individuals with 94.6% white ethnicity. CH was defined in this study by the presence of mosaic chromosomal alterations (mCA) and/or somatic driver mutations. I utilised B-allele frequencies, and genotypic intensities from single nucleotide polymorphism array data (n = 486,941) to identify mCA, and diagnostic data to classify mCA according to their association with myeloid, lymphoid or neither of these diseases. Furthermore, I utilised whole exome sequencing data (WES, 1st release, n = 49,956; 2nd release, n=150,685) and publicly available databases to identify putative somatic driver mutations. LOY calls in men were provided by the UK Biobank from published data.

Results: The frequency of myeloid CH increased per year of participant age and was associated with: two distinct germline predisposition signals within *TERT*, current smoking, and several blood features and clinical phenotypes indicative of chronic inflammation. Somatic loss-of-function mutations in *ASXL1* were found to be strongly associated with current and past smoking status.

Focusing on chronic kidney disease (CKD), myeloid CH was negatively associated with glomerular filtration rate (GFR) estimated from cystatin-C which is a marker of CKD but not with GFR estimated from creatinine which has previously been reported to be less informative. Furthermore, myeloid CH increased the risk of adverse outcomes, defined by a composite of all-cause mortality, myocardial infarction or stroke, in CKD cases compared to those without myeloid CH. Machine learning (ML) survival models which analysed high dimensional data including CH calls, blood counts and biochemistry markers were more predictive of myeloid malignancies in comparison to traditional regression-based models. Finally, LOY was significantly associated with CH and also with clonality inferred from non-CH somatic mutations. LOY was suggested to be causally associated with high levels of sex hormone binding globulin, and this relationship was linked to expression Quantitative Trait Locus (eQTL) associated with genes at the *DLK1*-*MEG3* locus.

Conclusion: This study demonstrates the wide scientific reach of CH and its broad impact on health outcomes. My results indicate that the type of CH, the identity of specific driver genes, inherited risk variants, and environmental factors are collectively determinants of the fitness of CH and influence the potential for development of myeloid neoplasms or non-malignant diseases. My findings also provide evidence that blood and serum measures hold additional information that helps to determine the clinical significance of CH.

Table of Contents

List of Figures

List of Tables

Research Thesis: Declaration of Authorship

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at the University of Southampton;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:

   Dawoud, A.A.Z., Tapper, W.J. & Cross, N.C.P. Clonal myelopoiesis in the UK Biobank cohort: ASXL1 mutations are strongly associated with smoking. *Leukemia* **34,** 2660–2672 (2020). https://doi.org/10.1038/s41375-020-0896-8

   Dawoud, A.A.Z., Gilbert, R.D., Tapper, W.J. & Cross, N.C.P. Clonal myelopoiesis promotes adverse outcomes in chronic kidney disease. *Leukemia* **36,** 507–515 (2022). https://doi.org/10.1038/s41375-021-01382-3

Signature: .... Date:

Published papers

- **Dawoud, A.A.Z.**, Tapper, W.J. & Cross, N.C.P. Clonal myelopoiesis in the UK Biobank cohort: ASXL1 mutations are strongly associated with smoking. *Leukemia* **34,** 2660–2672 (2020). https://doi.org/10.1038/s41375-020-0896-8
- **Dawoud, A.A.Z**., Gilbert, R.D., Tapper, W.J. & Cross, N.C.P. Clonal myelopoiesis promotes adverse outcomes in chronic kidney disease. *Leukemia* **36,** 507–515 (2022). https://doi.org/10.1038/s41375-021-01382-3
- Galatà G, García-Montero AC, Kristensen T, **Dawoud AAZ**, Muñoz-González JI, Meggendorfer M, Guglielmelli P, Hoade Y, Alvarez-Twose I, Gieger C, Strauch K, Ferrucci L, Tanaka T, Bandinelli S, Schnurr TM, Haferlach T, Broesby-Olsen S, Vestergaard H, Møller MB, Bindslev-Jensen C, Vannucchi AM, Orfao A, Radia D, Reiter A, Chase AJ, Cross NCP, Tapper WJ. Genome-wide association study identifies novel susceptibility loci for KIT D816V positive mastocytosis. *Am J Hum Genet* **108**:284-294 (2021). https://doi.org/10.1016/j.ajhg.2020.12.007

Submitted Papers

- **Dawoud, A.A.Z.**, Tapper, W.J. & Cross, N.C.P. Sex hormone binding globulin promotes the risk of age-related loss of the Y chromosome.

Ethical approval

The UK Biobank received ethical approval from the North West multi-centre Research Ethics Committee (REC reference 11/NW/0382). My study was undertaken as part of the UK Biobank approved project ID 35273: Myeloproliferative neoplasms and clonal haematopoiesis (Principal Investigator N.C.P. Cross) with local approval (ERGO II ID 61730: the clinical significance of clonal haematopoiesis).

Funders

This was supported by a Lady Tata International Award and the University of Southampton

Acknowledgements

I would mainly like to acknowledge Prof. Nicholas Cross and Dr. William Tapper for their valuable supervision. Prof. Cross shared his expertise and provided a solid academic and moral support throughout all the years spent in completing this work. Dr. Tapper provided an excellent guidance and close support that I enjoyed working under his supervision.

I would like to acknowledge Lady Tata Memorial Trust and University of Southampton for funding this work. Specifically, many opportunities for training, publishing, and attending conferences had been supported by my supervisors.

Of many friends and colleagues, I have met during my PhD, I must acknowledge the members of the genomic informatics group, and the myeloid research group at University of Southampton.

Last by no means least, I would like to thank my wife, Yasmin Elsery, for her unlimited encouragement and support until here. I wish to thank my parents, and other family members who have been back for everything.

Abbreviations

aCGH	Array comparative genomic hybridisation
aCML	Atypical chronic myeloid leukaemia
AFT	Accelerated failure time
AGM	Aorta-gonad mesonephros
AI	Allelic imbalance
AML	Acute myeloid leukaemia
AMPK	AMP-activated kinase
AUC	Area under the curve
aUPD	Acquired uniparental disomy
BAF	B Allele Frequency
BAP1	BRCA1-associated protein 1
BAT	Bioavailable testosterone
BBJ	Biobank Japan
BQSR	Recalibration of base quality score
BWA-mem	Burrows-wheeler aligner mem
CADD	Combined annotation dependent depletion
CBS	Circular binary segmentation
cDNA	Complementary DNA
CGC	Cancer gene census
CH	Clonal haematopoiesis
CHIP	Clonal haematopoiesis of indeterminant potential
CKD	Chronic kidney disease
CKDGen	Chronic Kidney Disease Genetics Consortium
CLL	Chronic lymphocytic leukaemia
CML	Chronic myeloid leukaemia
CMML	Chronic myelomonocytic leukaemia
CNG	Copy number gain
CNL	Copy number loss
CNN	Copy number neutral
CNV	Copy number variation
COPD	Chronic obstructive pulmonary disease
COSMIC	Catalogue of Somatic Mutations in Cancer
COX-PH	Cox proportional hazard
CRP	C-reactive Protein
CVD	Atherosclerotic cardiovascular disease
ddNTPs	Dideoxynucleotides
eGFR	Estimated glomerular filtration rate
EnWAS	Environmental WAS
eQTL	Expression quantitative trait locus
Erα	Oestrogen receptor α
ESE	Exon splicing enhancers
ESKD	End stage kidney disease

ET	Essential thrombocythaemia
FDR	False discovery rate
FE	Functional equivalence
FISH	Fluorescent in situ hybridization
FT	Free testosterone
GATK	Genomic analysis toolkit
GENEVA	Gene environment association studies consortium
G-proteins	Heterotrimeric guanine nucleotide-binding proteins
gVCF	Genome level variant call file
GWAS	Genome wide association study
H2AK119	Histone2A lysine-119
H3K27me3	Trimethylation of Histone3 lysine-27
HDL	High density lipoprotein
HSCs	Haematopoietic stem cells
HWE	Hardy-Weinberg equilibrium
ICD	International Classification of diseases
IGF-1	Insulin Growth Factor 1
IL6	Interleukin-6
InSIDE	Independent of Direct Effect
IPSS-R	International prognostic scoring system for myelodysplastic syndromes
ITD	Internal tandem duplication
JAK2	Janus kinase 2
JMML	Juvenile myelomonocytic leukaemia
LDLR	Low density lipoprotein receptor
LoD	Limit of Detection
LOH	Loss of heterozygosity
LPS	Lipopolysaccharide
LRR	log2 R ratio
MAF	Minor Allele Frequencies
MAPK	Mitogen activated protein kinase
mBAF	Mirrored BAF values
MBL	Monoclonal B-cell lymphocytosis
mCA	Mosaic chromosomal alteration
MCH	Mean corpuscular haemoglobin
MCV	Mean corpuscular volume
MDS	Myelodysplastic syndrome
MGUS	Monoclonal gammopathy of unknown significance
MI	Myocardial infarction
LOY	Mosaic loss of chromosome Y
MM	Multiple myeloma
MPN	Myeloproliferative neoplasm
MR	Mendelian randomisation
MRC	Medical Research Council
MR-RAPS	MR using robust adjusted profile score

NAHR	Non-allelic homologous recombination
NCCN	National comprehensive cancer network
NGS	Next generation sequencing
nORFs	Novel open-reading frames
OQFE	Original quality functional equivalence
PAR1	Pseudo-autosomal region 1
PCAWG	Pan-cancer analysis of whole genomes
PHESANT	PHEnome Scan ANalysis Tool
PheWAS	Phenome-wide association studies
PI3K	Phosphotidylinositol 3-kinase
PMF	Primary myelofibrosis
PON	Panel of normal
PRC2	Polycomb repressive complex 2
PR-DUB	Polycomb repressive deubiquitinase
PRS	Polygenic risk score
PTD	Partial tandem duplication
PV	Polycythaemia vera
RBD	RBCs distribution width
RNAseq	RNA sequencing
ROC	Receiver Operator Characteristic
RSF	Random survival forest
SAM	Sequence alignment map
SHAP	Shapley additive explanations
SHBG	Sex hormone binding globulin
SNP-A	Single nucleotide polymorphism arrays
snRNPs	Small nuclear ribonucleoproteins
SPB	Regeneron seal point balinese
SR	Splicing regulators
SS	Splice sites
STAT	Signal transducer and activator of transcription
TNFA	Tumour necrosis factor alpha
TOPMED	NHLBI trans-omics for precision medicine
TT	Total testosterone
UPD	Uniparental disomy
VAF	Variant allele frequency
VQSR	Variant quality score
WGS	Whole genome sequencing
WHO	World health organisation
XGB	Extreme gradient boosting
α-KG	α-ketoglutarate

Chapter 1 Introduction

In this Chapter, I will cover the background knowledge relevant to my studies. The first section discusses the concept of clonal haematopoiesis (CH) as a precursor to the development of haematological neoplasms. Next, I discuss two concepts relating to the pathogenesis of myeloid malignancies; the first is the genetic basis of these disorders, and the second is the role of inflammation. In the final section, I discuss next generation sequencing (NGS) and other genomic technologies used to characterise both CH and myeloid neoplasms.

1.1 Haematopoiesis and clonality

Haematopoiesis is the canonical differentiation process that generates cells of the blood and immune system. Haematopoietic stem cells (HSCs) are at the top of this hierarchical process, and are located at various sites, according to the developmental stage. In early gestation, HSCs are located in the yolk sac and the aorta-gonad mesonephros (AGM) region. Next, HSCs migrate and are active at the fetal liver and spleen until about 2 weeks after birth when they begin to migrate to the bone marrow. Finally, in human adults, bone marrow dominates the process of haematopoiesis [1]. HSCs flourish in a niche, which is cellular and extracellular matrix generated by stromal cells that provide a suitable environment for HSCs to survive, proliferate and differentiate. HSCs have two main characteristics. First, they can self-renew, and thus completely regenerate the haematopoietic system, e.g. after bone marrow transplantation. Second, they have the capacity to differentiate into haematopoietic progenitors that are committed to develop restricted cell types, such as myeloid or lymphoid cells. Long-lived progenitor cells are recruited to keep the steady-state blood production during adulthood rather than HSCs [2].

The regulation of haematopoiesis depends on internal and external factors. Externally, it is controlled by growth hormones and cytokines to keep the balance between the self-renewal capacity of the HSCs, and the promotion of differentiation. Within each cell, growth factor receptors, their signal transduction components, and downstream transcription factors control the fate of individual cells and thus the haematopoietic system.

1.1.1 Clonal haematopoiesis

Leukaemia and related conditions are clonal disorders whereby an initial somatic mutation in a single cell confers a selective advantage. Further mutations may be acquired that confer an additional selective advantage and give rise to further subclones that may initiate full blown disease or promote disease evolution [3]. It has become apparent, however, that clonality does not necessarily indicate malignancy. For example, skewed X-chromosome inactivation (a potential marker of clonality) in blood cells is known to be correlated with the age of healthy females who have no evidence of a haematological malignancy. This initial observation led to the identification of *TET2* mutations as the driver of clonality in some cases with skewed X-chromosome inactivation [4]. Other evidence for CH in the absence of a discernible haematological disorder came from a number of observations, including: (i) the prevalence of *JAK2* V617F was much greater than myeloproliferative neoplasms (MPN) in randomly selected subjects undergoing hospital-based clinical investigations [5]; (ii) as described in more detail below, mosaic chromosome alterations (mCA), many of which were characteristic of myeloid neoplasia, were seen in large cohorts ascertained for non-haematological conditions [6,7]; (iii) myeloid malignancy-associated driver mutations, most commonly involving *DNMT3A*, *TET2* or *ASXL1*, were found at a much higher frequency than expected in large cohorts also ascertained for non-haematological conditions [8-10] and (iv) exome sequencing of *de novo* acute myeloid leukaemia (AML), and matched remission samples identified some somatic mutations, particularly in *DNMT3A*, that persisted in remission suggesting reversion to a previously unsuspected pre-leukaemic clonal state [11]. These findings provided the foundation for our understanding of the relationship between age-associated clonality and myeloid leukaemia associated genes, as discussed in more detail below. The literature has other evidence that connects pre-malignant clonal states to haematological malignancies. For example, monoclonal gammopathy of undetermined significance (MGUS) is a premalignant clonal disease that is often succeeded by multiple myeloma (MM) [12], with a progression rate from MGUS to MM of 1% per year [13]. Similarly, monoclonal B-cell lymphocytosis (MBL) is a benign condition in many individuals but may progress to chronic lymphocytic leukaemia (CLL).

1.1.2 Mosaic chromosomal alterations

The carrying of two or more different karyotypes in different cells of the same individual is defined as chromosomal mosaicism. These aberrant events may arise very early in development or in somatic cells and include whole chromosome abnormalities (e.g., loss or gain of whole chromosomes), translocations, sub-chromosomal structural abnormalities (deletions or amplifications), and copy number neutral events known as uniparental disomy (UPD) or copy number neutral loss of heterozygosity (CNN-LOH). In cancer, acquired chromosomal alterations have been used for many years as a marker of clonality. In 2012, however, two consecutive studies identified chromosome mosaicism in the normal population, and an age-related pattern of incidence. In one of the two studies, they analysed 57,853 individuals (31,717 non-haematological cancer cases and 26,136 cancer-free controls) from 13 genome wide association studies (GWAS). Mosaic abnormalities were identified in peripheral blood-derived DNA from about 1% of the cohort. Although this finding was more frequent in participants with solid tumours (0.97%) compared to the cancer free group (0.74%), the most striking finding was a continuous increase in prevalence with age, ranging from 0.23% of samples under 50 years old to 1.91% of samples between 75 and 79 years [7]. In the second study, 50,000 individuals were analysed from 15 different studies in the Gene Environment Association Studies consortium (GENEVA) which identified mosaic alterations in <0.5% of the participants younger than 50 years old, but the frequency rapidly increased to 2-3% in participants older than 50 years. In addition, the mosaic events were associated with a 10-fold increase in the incidence of haematological malignancies [6]. These studies were the first comprehensive description of age-related clonality in peripheral blood cells of individuals that were not selected for a haematological abnormality.

1.1.3 Uniparental disomy (UPD)

UPD is defined as the finding that both homologues of a pair of chromosomes, or sub-chromosomal regions, are derived from the same parent. UPD can involve identical homologues in which case it is termed “isodisomy”, or non-identical homologues in which case it is called “heterodisomy”. The origin of UPD is related to its clinical consequences, and in particular whether it is germline or somatic.

Constitutional UPD: Germline UPD is associated with rare developmental disorders such as Angelman syndrome and Prader-Willi syndrome, conditions that arise by the aberrant expression of imprinted genes in the affected regions. Different hypotheses can explain how germline UPD arises. First, the gamete complementation hypothesis is based on the high frequency of aneuploidy in human gametes. Aneuploidy provides a chance for the fertilization of nullisomic gamete and a disomic gamete of the same chromosome, which generates heterodisomic UPD. Second, trisomic rescue is the loss of one of the extra-chromosomes in trisomic zygote cells, to save the conceptus. This mechanism may generate mosaicism between trisomic cells, and disomic cells, with a theoretical probability of 1/3 that the disomic cells are uniparental with respect to the originally trisomic chromosome. The trisomic event arises due to a segregation error in the meiosis 1, or meiosis 2. The first generates a heterodisomic UPD, whereas the second generates isodisomic UPD. Third, the compensation mechanism is the duplication of a normal chromosomal homologue to compensate for aneuploidy. This takes place at mitosis early in development and generates isodisomic UPD.

Acquired UPD: Non-allelic homologous recombination (NAHR) is the main mechanism for the generation of acquired UPD (aUPD), a phenomenon that is commonly associated with cancer, including myeloid malignancies. It arises through mitotic recombination between non-allelic homologous regions, which generates a region of isodisomic aUPD, usually between the region of the recombination and the telomere (Figure 1-1) or interstitially if there are two points of recombination [14,15]. Recurrent regions of aUPD are typically associated with somatic cancer driver mutations and confer a selective advantage to the cell, most commonly by conversion of a heterozygous mutation to homozygosity [16].

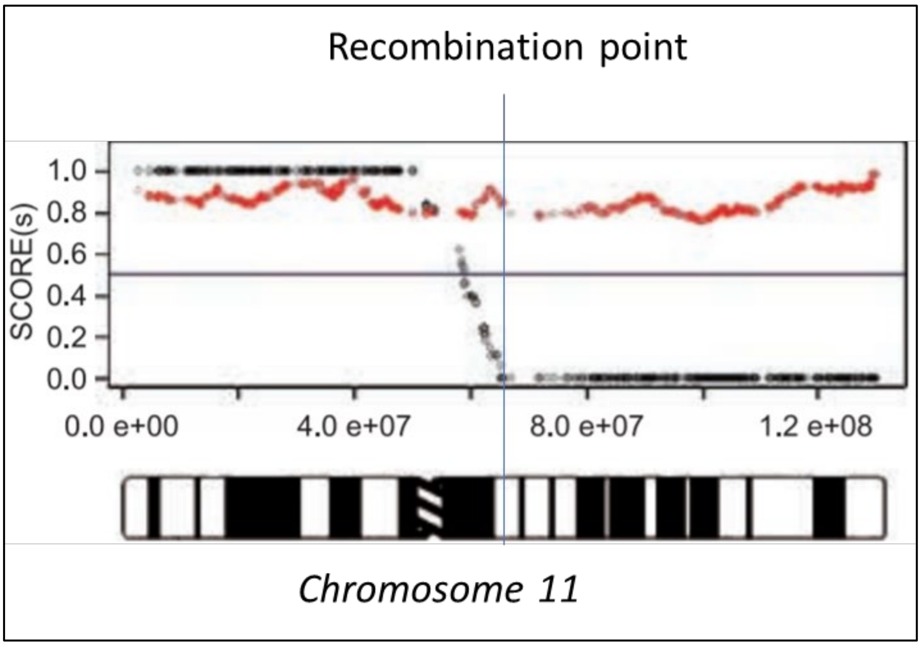


Figure 1-1: An example of aUPD at chr11q in an AML patient

Raghavan and colleagues identified aUPD with a breakpoint at chr11q by comparing calls ratio and signal intensities ratio between diagnosis and remission samples of an AML patient [17]. Black dots refer to ratio of heterozygous to homozygous calls in a window of 20Mb in the diagnosis sample divided by the similar ratio in the remission sample. Red dots refer to the ratio of the mean signal, and the blue line indicates the point of recombination.

1.1.4 The relationship between aUPD and cancer driver genes

Combining both single nucleotide polymorphism arrays (SNP-A) and NGS is a comprehensive strategy to identify pathogenic abnormalities in cancer. SNP-A provides a sufficient density of markers to identify most chromosomal abnormalities apart from balanced translocations. It provides quantitative data of the allelic frequency as well as the copy number, which can be used to distinguish between copy number variants (CNV) and copy number neutral changes (i.e. aUPD) [18]. On the other hand, NGS can also identify point mutations, *indel*s and, in the case of whole genome sequencing (WGS) or RNA sequencing (RNAseq), fusion genes arising from reciprocal translocations [19].

Despite the huge advances in our understanding of cancer genomes there are still opportunities to identify new cancer genes in recurrent regions of chromosomal alteration. The analysis of 3,131 microarrays (Affymetrix 250k) from 26 different cancers identified 76 focal amplifications, and 82 focal deletions, most of them with no known cancer genes [20]. Also, the analysis of 4,934 microarrays (Affymetrix SNP 6) from the Cancer Genome Atlas data set identified 140 recurrent regions, 102 of them with no known cancer related genes [21]. In general, focal somatic CNVs represent more than 80% of the copy number alterations in cancer, and they can be used to identify new driver genes [20]. Despite the huge progress in sequencing cancer genomes and exomes, an analysis based on signatures of positive selection indicates that 50% of cancer driver genes remain to be discovered [22]. In addition, non-coding regions are potential targets to discover new selection signals. The first evidence came from the discovery of two positions in *TERT* promoters that can be activated by somatic mutations [23,24]. Following this, the analysis of 2,658 genomes from the Pan-Cancer Analysis of Whole Genomes (PCAWG), identified more non-coding regions altered by point mutations such as 5'-end mutations in *TP53* and 3' UTRs of *TOB1* [25]. Novel open-reading frames (nORFs), that include both small ORFs (1-100 amino acids) and alternative ORFs, could represent a new dimension to discover new driver mutations [26]. Previous studies had identified some of the cancer associated nORFs, such as lncRNA HOXB-AS3 that encodes a 55 amino acid peptide downregulated in colon cancer [27].

In myeloid leukaemia, aUPD is associated with many known mutated genes. For example, *JAK2*, *MPL*, and *CALR* mutations, the main three diagnostic markers of the myeloproliferative neoplasms (MPN), are associated with aUPD of chromosomes 9p, 1p, and 19p respectively [28-30] although aUPD is most commonly seen in association with *JAK2* mutations. Indeed, refinement of the region targeted by 9p aUPD was one of the routes by which *JAK2* V617F, the most common mutation in MPN, was first identified [28]. This now classic pipeline of identifying mutated genes within regions of aUPD has been repeated, for example inactivating *EZH2* mutations in cases with 7q aUPD [31,32], and missense substitutions in *CBL* associated with 11q aUPD that abrogate CBL ubiquitin ligase activity [33,34]. Theoretically, the gene level mutation is the initial driver of clonal development and aUPD, or hemizygous deletion of the wild type allele, may represent a second hit that drives the development of a more aggressive clone with a homozygous or hemizygous driver mutation. For example, patients with the homozygous form of *JAK2* V617F tend to have a more symptomatic form of the disease, and are more likely to transform to primary myelofibrosis [35].

1.1.5 Clonal haematopoiesis of indeterminate potential

Clonal haematopoiesis of indeterminate potential (CHIP) is defined by the finding of clonality, most commonly the identification of somatic mutations in the myeloid malignancy-associated genes at ≥2% VAF, in the absence of any phenotypic characteristics of malignancy such as an abnormal blood cell counts. Following on from the studies of chromosomal mosaicism described above, CHIP was first identified as a widespread phenomenon by genomic analysis of large population cohorts under investigation for a variety of non-malignant conditions [8,10,36].

The prevalence of CHIP varies according to the sensitivity of the mutation assessment technique and the population studied. Notably, it is directly proportional with age, ranging from 1% of individuals in their 40s to >10% in their 80s, and is associated with a range of frequently mutated genes (*DNMT3A*, *TET2*, *JAK2*, *ASXL1*, *TP53*, *GNAS*, *PPM1D*, *BCORL1* and *SF3B1*) [10,37] (Figure 1-2). In a separate study, mutations in *DNMT3A*, *ASXL1* and *TET2* were identified in more than 10% of participants over the age of 70, and in 1% of participants below the age of 50 using whole exome sequencing with a limit of detection (LoD) of 5% variant allele fraction (VAF) [8]. However, using deep targeted error corrected sequencing methods with a LoD of 0.03%, CHIP were found in 95% of participants aged between 50-60 years old indicating that low level CHIP is very prevalent [38]. Briefly, error corrected sequencing is a technique that employs unique labels for each DNA molecule that directly allows technical artefacts to be distinguished from true variants by effectively improving the signal-to-noise ratio [38,39]. Individuals with CHIP had more than 10 times elevated risk of developing a haematological malignancy but the rate of progression was only about 1-2% per annum [8,10], similar to the rate of progression of MGUS to MM, and MBL to CLL [13,40]. Deep sequencing has provided valuable information about the incidence of CHIP mutations to the driver gene level among different age categories. Although *DNMT3A* mutations are the most common in all age categories, with a gradual increase in prevalence by age, the spliceosome genes *SF3B1* and *SRSF2* were exclusively mutated at age greater than 70 years old [41].

These findings have been used to generate a prediction model for the development of AML by using deep sequencing to compare pre-AML cases vs controls. The pre-AML cases had more somatic mutations, specifically an enrichment in *TP53* mutations, *SRSF2* and *U2AF1* mutations (spliceosome genes known to be associated with poor prognosis in AML), and mutations in *JAK2*, *IDH2*, and *ASXL1*. *DNMT3A* and *TET2* were frequently mutated in both groups. Surprisingly, *NPM1*, *CEBPA* and *FLT3*-ITD mutations were entirely absent, which suggests that these genes are later events that may more directly drive clinically manifest AML [42]. Another study identified somatic mutations in *IDH1*, *IDH2*, *TP53*, *DNMT3A* and *TET2* as predisposition factors for the risk of developing AML [43].

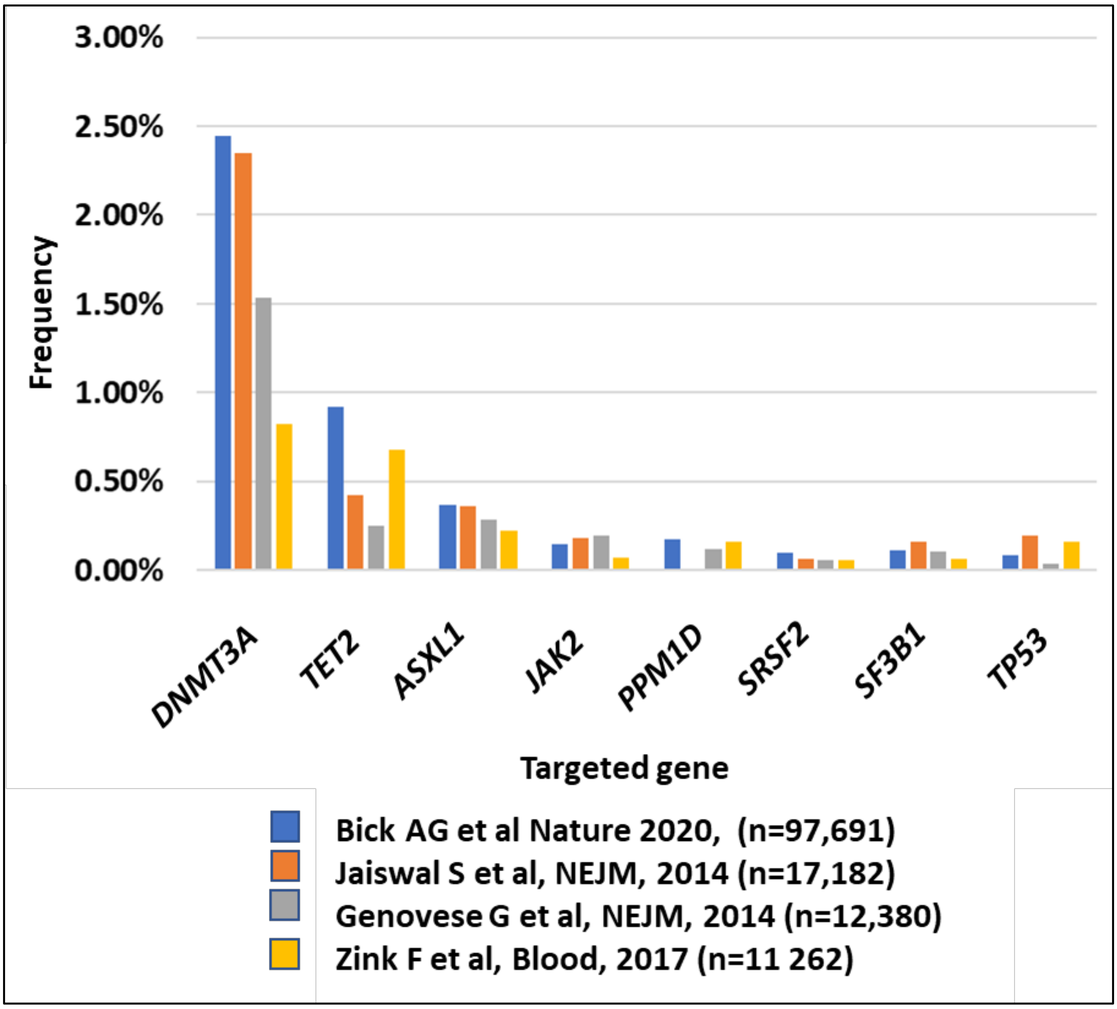


Figure 1-2: Mutations in the main genes associated with clonal haematopoiesis

Comparative data from four large cohort studies of CHIP. The three epigenetic regulator genes, *DNMT3A*, *TET2* and *ASXL1* contribute to more than 90% of somatic driver mutations. Although a long tail of mutated genes varies between studies, mutations in splicing genes (*SF3B1* and *SRSF2*), apoptotic regulators (*TP53*, *PPM1D*), and signal transduction (*JAK2*) were the most common targets after the epigenetic regulator genes.

1.1.6 The biological function of mutated genes in clonal haematopoiesis and myeloid neoplasms

Most driver mutations associated with CH target a small group of myeloid neoplasia-related genes that are implicated in epigenetic regulation, the splicing machinery, apoptosis, and signal transduction. I will cover these biological processes and the genes involved in the following sections.

1.1.6.1 Epigenetic regulator genes

DNA methyltransferase 3 alpha (DNMT3A): *DNMT3A* is located at 2p23.3. It encodes a methyltransferase enzyme, which adds a methyl group to C5 of cytosine to form 5-methylcytosine at CpG dinucleotides. In general, CpG are clustered in regions called CpG islands, and high methylation rates are associated with gene silencing. *DNMT3A* and *DNMT3B* have similar functions of carrying out *de novo* methylation [44] and differ from *DNMT1* which plays a role in maintaining pre-existing patterns of methylation. NGS as a game changing technology enabled the discovery of *DNMT3A* mutations in cytogenetically normal *de novo* AML [45]. The majority of *DNMT3A* mutations target the region encoding the methyltransferase domain, most frequently missense mutations at amino acid R882, or frameshift/stop mutations resulting in a truncated protein [46]. The functional effect of *DNMT3A* mutations can arise by different mechanisms; (i) haploinsufficiency, indicating a role as a tumour suppressor gene (ii) a dominant-negative effect, as mutated *DNMT3A* inhibits the activity of wild type *DNMT3A* and *DNMT3B* [47]. In clinical practice, *DNMT3A* mutations have been associated with a poor prognosis, reduced overall survival in AML [48], and co-occurrence with *NPM1*, *FLT3*, and *IDH1* [45]. However, the finding of *DNMT3A* mutations without *NPM1* present in AML blasts revealed the pre-leukaemic nature of *DNMT3A* mutations [11].

Ten Eleven Translocation (TET) methylcytosine dioxygenase 2 (*TET2*): The *TET2* gene, which is located at chromosomal position 4q24, belongs to the TET family of proteins. These proteins catalyse DNA demethylation by converting 5-methylcytosine into 5-hydroxymethylcytosine. Sequencing of the minimal overlap region in MPN patients with 4q mCA led to the finding of LOF mutations in the *TET2* gene [49]. In general, *TET2* mutations are highly associated with MDS and are frequently seen in AML, and MPNs. The prognostic impact of *TET2* is controversial, *TET2* mutations in MDS patients have been associated with a higher response rate to azacytidine [50]. Subsequently, *TET2* was found to interfere with other functional pathways; (i) *TET2* is regulated by α-ketoglutarate (α-KG) which is produced by isocitrate dehydrogenase (IDH1/2); mutations in *IDH1/2* produce 2-hydroxylglutarate (2-HG) that competitively inhibits α-KG, and alters the demethylation activity (Figure 1-3) [51]; (ii) *TET2* is a substrate for AMP-activated kinase (AMPK), which phosphorylates *TET2* serine 99 and stabilises *TET2* activity, a finding that connects *TET2* to glucose levels [52]; (iii) *TET2* is regulated by ascorbate which reduces the catalytic site Fe(III) to Fe(II) [53].

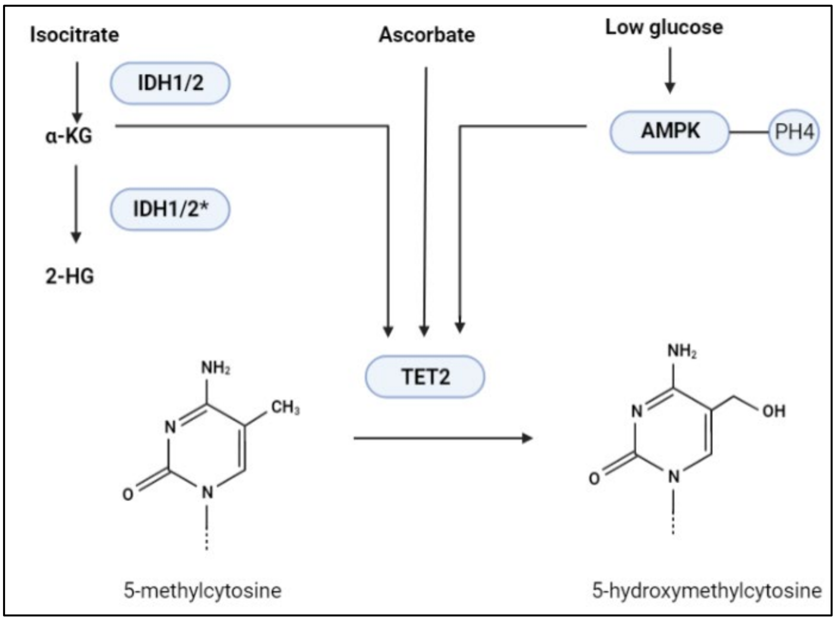


Figure 1-3: Pathways inhibiting the demethylation activity of *TET2*

TET2 catalyses DNA demethylation by converting 5-methylcytosine into 5-hydroxymethylcytosine Mutated *IDH1/2* induces the production of 2-hydroxylglutarate (2-HG) that competitively inhibit α-ketoglutarate required for the demethylation function of *TET2.* AMP-activated kinase (AMPK), regulated by glucose levels, phosphorylates *TET2* at serine 99. Ascorbate increases dioxygenase activity of TET2 by facilitating the Fe(III)/Fe(II) redox reaction.

Additional sex combs-like 1 (ASXL1): ASXL1 is located at chromosomal position 20q11. The sequencing of a recurrent region of interstitial deletion at 20q led to the finding of *ASXL1* mutations in myeloid malignancies [54,55]. Mutations in *ASXL1* alter chromatin conformation[56] by two mechanisms (Figure 1-4): (i) mutations in *ASXL1* target the polycomb repressive complex 2 (PRC2) which mediates trimethylation of Histone3 lysine-27(H3K27me3), an epigenetic mark associated with downregulation of nearby gene expression via the formation of heterochromatic regions [57]; (ii) ASXL1 and BRCA1-associated protein 1 (BAP1) form a protein complex called polycomb repressive deubiquitinase (PR-

DUB) that targets Histone2A lysine-119 (H2AK119) [58]. Histone H2AK119 mono-ubiquitination is essential to maintain PRC2-mediated transcriptional repression [59]. In general, *ASXL1* is frequently mutated in myeloid malignancies and associated with poor prognostic outcomes for AML, MDS and MPN patients [60].

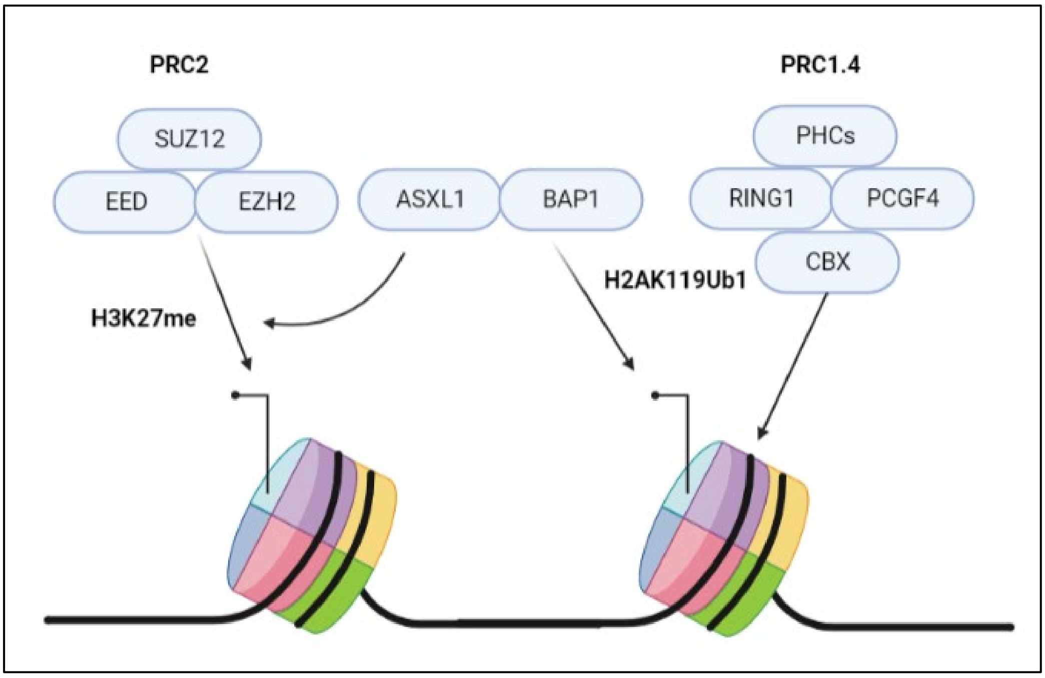


Figure 1-4: The role of *ASXL1* in chromatin modification

The ASXL1-BAP1 complex is recruited by the PRC2 complex that is composed of SUZ12, EED, and EZH2 proteins to mediate trimethylation of Histone3 lysine-27(H3K27me3). The ASXL1-BAP1 complex also functions by removing ubiquitin from histone H2A lysine 119 in regions targeted by the PRC1.4 complex. which is composed of PCGF4, RING1A, CBX, and PHCs.

1.1.6.2 Splicing machinery genes

Genes encoding components of the RNA splicing machinery are frequently mutated in MDS, include *U2AF1*, *ZRSR2*, *SRSF2* and *SF3B1* [61,62]. Small nuclear ribonucleoproteins (snRNPs) and other dependent proteins aggregate at the 3’ and 5’ splice sites (SS) of pre-mRNA to form the spliceosome. The 5’ SS binds to U1 small nuclear ribonucleoprotein particle (snRNP) which in turn binds to the 5’ splice site through base pairing (Figure 1-5). The AG bases of the 3’ SS bind to U2AF1. The branch point binds to SF1 and the polypyrimidine tract binds to U2AF2. SF3B1 and SF3A1 are components of U2 snRNP and it is thought that they bind pre-mRNA upstream of the branch site in a sequence-independent manner to anchor the U2 snRNP to pre-mRNA. SRSF2 is one of the splicing regulators (SR) proteins that binds to exon splicing enhancers (ESE) to direct splicing machinery components and to define exon/intron boundaries [63]. The majority of mutations in *SRSF2*, *U2AF1*, *SF3B1* are heterozygous missense point mutations that include SRSF2 P95, U2AF1 Q157, SF3B1 K700 [62].

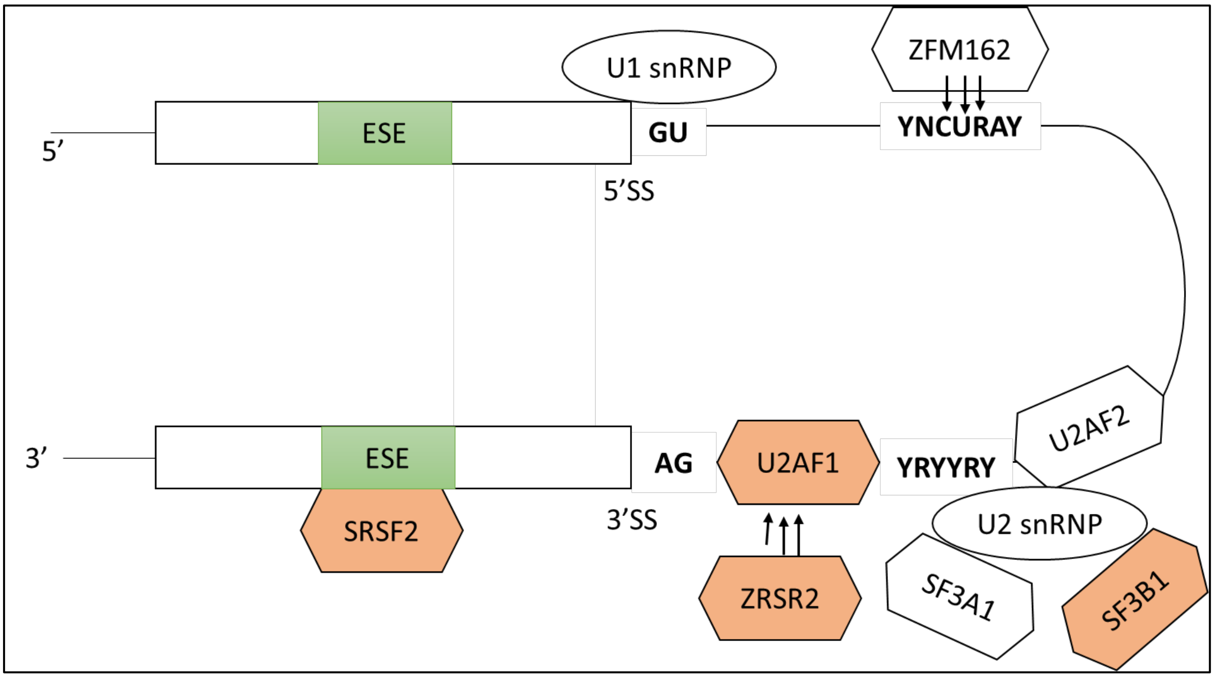


Figure 1-5: RNA splicing machinery

Orange coloured molecules indicate frequently targeted genes in myeloid malignancies. U1snRNP and U2AF35 (U2AF1) bind to the 5' SS, and 3' SS respectively, whereases the branching point binds to SF1 and U2AF65 (U2AF2). SF3A1, and SF3B1 guides U2snRNP. SRSF2 binds to the ESE [62].

1.1.6.3 Apoptosis-related genes

Protein Phosphatase, Mg2+/Mn2+ Dependent 1D (*PPM1D*): The serine-threonine phosphatase encoded by *PPM1D* is upregulated in response to DNA damage by a mechanism that depends on p53 [64]. Recurrent mutations in *PPM1D* have been identified by sequencing of blood samples from healthy individuals [37]. Pre-existing *PPM1D* mutated cells expand in patients treated with cytotoxic agents such as cisplatin [65], an effect mediated by elevated resistance to apoptosis [66]. Genetic alterations in *PPM1D* have been identified in a significant proportion of MPN patients which include truncating mutations (n=5/89 of MPN blast phase, n=4/135 of PV and ET), cytogenetic alterations in the *PPM1D* region at 17q23 (1.4%), and over-expression (42% of 31 MPN) [67].

1.1.6.4 Signal transduction genes

Janus kinase 2 (*JAK2*): *JAK2* is a non-receptor tyrosine kinase and member of the Janus-kinase family. *JAK2* function is mediated through its association with cytokine receptors that activates Signal Transducer and Activator of Transcription (STAT), mitogen activated protein kinase (MAPK) and phosphotidylinositol 3-kinase (PI3K) signalling pathways [68] (Figure 1-6). In 2005 several groups reported the acquisition of *JAK2* V617F in 95% of PV patients, and about 50% of ET and PMF patients [28,69-71]. V617F is a point mutation in the catalytically inactive pseudokinase domain (JH2) that results in constitutive activity of the JH1 kinase domain [72]. Recent structural modelling has localised V617 at the putative interface between JH1 and JH2, with mutation to phenylalanine predicted to destabilise the inactive conformation and stabilising the active conformation [73].

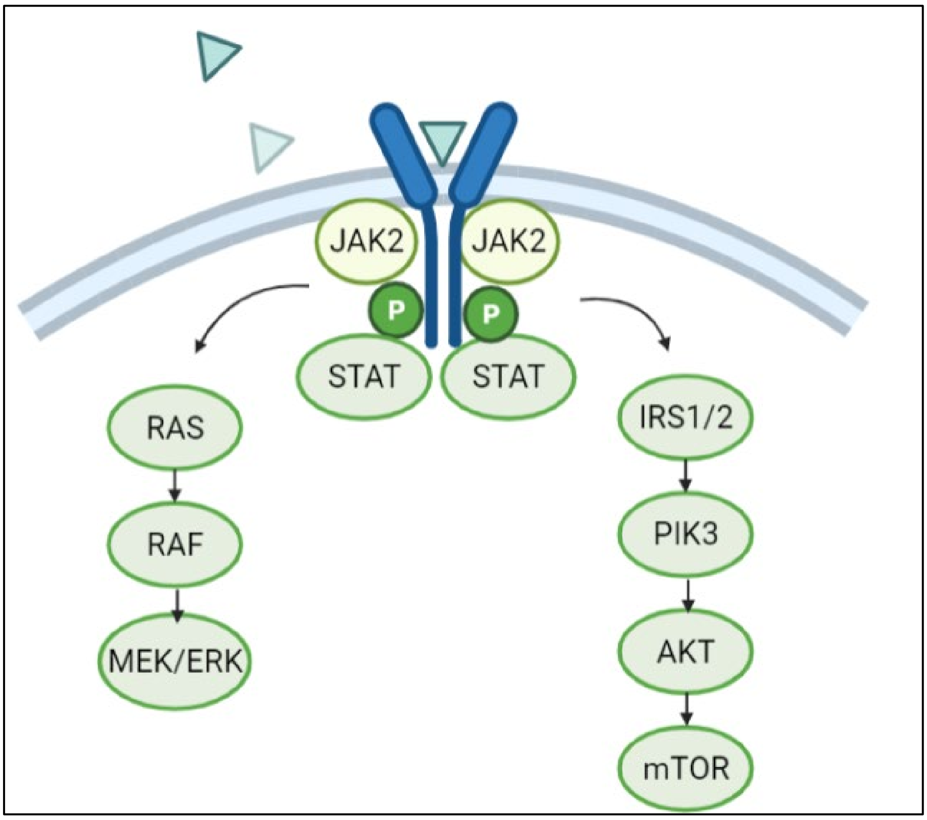


Figure 1-6: JAK2 signalling transduction pathways

Cytokine ligands (triangles) bind to cytokine receptors, resulting in JAK2 activation, phosphorylation and recruitment of STAT proteins. The activated signalling pathways include mitogen activated protein kinase (MAPK, RAS/RAF/MEK/ERK) signalling proteins and the activation of the phosphotidylinositol 3-kinase (PI3K)–AKT pathway via phosphorylation of Insulin Receptor Substrate 1/2 (IRS1/2) [74].

1.1.7 Mosaic loss of chromosome Y

Mosaic loss of chromosome Y (LOY) is the expansion of a 45,X karyotype in a subset of cells in the peripheral blood of males. The phenomenon was first identified nearly fifty years ago by karyotyping and described as common, occurring in 23% of males, and connected to ageing in generally healthy individuals as well as those with a haematological malignancy [75]. Large population studies have been performed using SNP-A data to detect LOY and to characterise its association with genetic and non-genetic risk factors. In a combined cohort of 1,153 men, LOY was associated with a 1.9 times higher risk of all-cause mortality, and 3.6 times higher risk for solid cancers [76]. LOY was significantly associated with current tobacco smoking in a dose-dependent manner, but not with previous smoking in 6014 males [77]. A genome-wide association study comparing 895 men with LOY against 11,474 controls detected a prominent association with rs2887399 (OR = 1.57, $P = 6.46 \times 10^{-11}$) which is located just upstream of *TCL1A* [78], a gene which encode a protein that co-activates AKT to enhance phosphorylation by signal transduction [79]. Subsequent single cell analysis identified over expression of *TCL1A* in cells with LOY and specifically in B-lymphocytes [80]. Further studies have discovered genetic risk variants for LOY associated with genes involved in cell proliferation and cell cycle regulation. A total of 19 genetic loci (18 + *TCL1A*) that predispose to LOY were identified following the analysis of 67,034 males in the UK Biobank [81]. An additional 137 loci (total=156) were discovered by assessing the SNP-A data from 205,011 participants from the UK Biobank [80].

1.1.8 The dynamics of clonal haematopoiesis

Clonal fitness is the proliferative advantage of mutated cells over normal cells [82]. It is a significant factor in determining the contribution of a clone in the pathogenesis of a malignant or benign phenotype. Recently, different methods were applied to mathematically model the fitness of a driver mutations:

(i) The first used aggregated VAF measurements and age at detection of a single mutation form different subjects [83] to model variant density as a function of VAF and to estimate fitness by considering age, mutation rate, number of haematopoietic stem cells (HSCs), and time between divisions as shown in the equation below. According to this formula the fitness of *DNMT3A* R882H was estimated to be 15% ± 1% per year

$$\rho(l)=\theta\exp(-e^{l}/\phi)$$

where l = log(VAF), $\theta = 2N\tau\mu$, and $\varphi=(e^{st}-1)/2N\tau s$

N=Number of HSCs, τ=time between divisions, μ=mutation rate, s=fitness effect

(ii) The second method used a longitudinal design that provided a direct estimation for fitness that tracked VAF changes over multiple-time points. Surprisingly, 46% of the identified CHIP with VAF > 0.02 were found to have VAF <0.02 in the following time points, which indicated the shrinkage over the study time [84]. Fluctuating VAFs of some mutations might be explained by natural drift in populations of cells and was supported by analysis of synonymous variations that are not expected to confer a fitness advantage. Both methods, VAF aggregation and the longitudinal design revealed canonical characteristics for CH; (i) the targeted gene was the main feature influencing fitness (ii) mutations in the splicing genes *SRSF2*, and *SF3B1* were identified as the most fit mutations in comparison to those targeting epigenetic regulator genes [83,84]. However, Watson and colleagues pointed to *GNB1* K57E as being one of the fittest mutations. *GNB1* encodes the subunit β, one of the three subunits compose heterotrimeric guanine nucleotide-binding proteins (G-proteins) that links signals from receptors to downstream proteins. On the other hand, exogenous factors such as radiation and chemotherapy play a fundamental role in influencing CH fitness. Mutations in *TP53*, *PPM1D*, and *CHEK2* confer a selective advantage in the context of treatment with radiation, platinum or topoisomerase II inhibitors [85]. In summary, the fitness of CH is controlled by internal factors that include the function of the gene concerned and the type of mutation, as well as external factors that provide an environment that confers a selective advantage for clones with specific mutated genes.

1.1.9 Clonal haematopoiesis and the risk of myeloid neoplasms

The relationship between the risk of developing myeloid malignancies and the finding of CH in healthy individuals has been well demonstrated in different cohorts [6,7]. The risk for myeloid malignancies varies according to the targeted gene, with mutations in *JAK2*, *SRSF2*, *U2AF1*, *IDH2*, and *RUNX1* predominant in individuals at higher risk of developing myeloid malignancies [42,86]. Regarding clone size, individuals with mutations of VAF > 0.01 are at higher risk of developing myeloid neoplasms, however smaller clones had unclear pathogenesis [87]. Individuals with multiple clones are at an even higher risk, that is independent of the correlation between point mutation and mCA [86,87]. Different studies modelled the risk of myeloid malignancies using features of CH (mutated genes and VAF) as independent variables. Abelson and colleagues encoded driver mutations as independent continuous variables and used a Cox proportional hazard (COX-PH) model to predict the progression-free survival of AML, the performance of the developed model achieved area under-the-curve Receiver Operator Characteristic (ROC) = 0.79 [42]. Saiki and colleagues used a combination of driver mutations and mCA to predict the risk of both myeloid and lymphoid malignancies, furthermore, CH was associated with higher risk of mortality in haematological malignancies (HR=2.8, for driver mutations and HR=2.6 for mCA [86].

1.1.10 The relationship between clonal haematopoiesis and non-malignant diseases

Theoretically, mutated HSCs may have a wider effect on non-haematopoietic organs as blood and immune cells are distributed to other tissues, and clonal changes may potentially alter the immune response and the inflammatory state. This hypothesis is supported by survival analysis of CHIP that identified a 40% increase in all-cause mortality from 2 different studies, that is not explained by haematological malignancies alone [8,10]. Indeed, the majority of participants with CHIP who died developed other diseases before developing any malignancy (Figure 1-7). Other factors also need to be considered for example smoking, the prevalence of which is double in CHIP cases in comparison to controls [8].

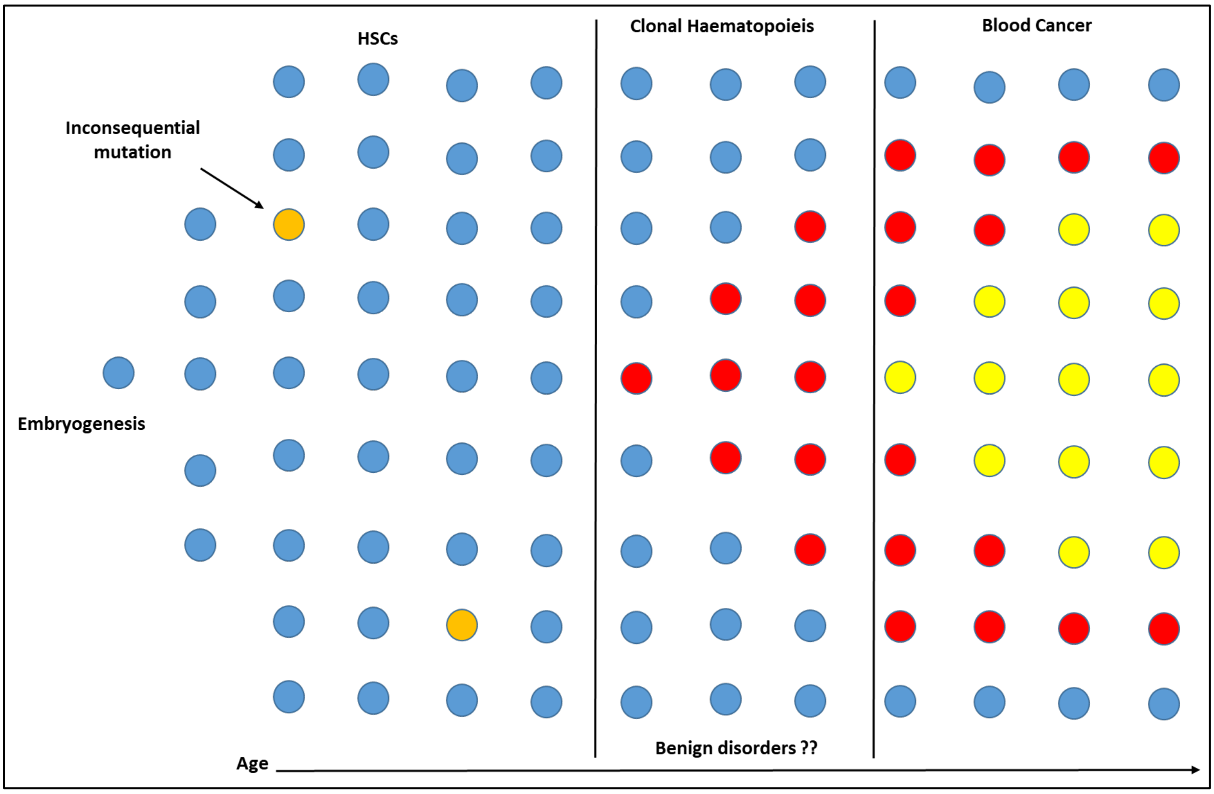


Figure 1-7: The relationship between age and clonal haematopoiesis

The colour change represents the acquisition of a new mutation. The orange mutation confers no selective advantage and is therefore inconsequential. The red mutation confers a selective advantage and results in CH. The yellow mutation confers a further selective advantage and drives progression to blood cancer. 10 to 20% of people aged above 70, develop a relatively large clone. They have more than 10 times higher risk of developing haematological malignancies but the majority of them will die or develop benign disorders such as cardiovascular diseases, before developing any malignancy [88].

1.1.10.1 Cardiovascular disorders

Importantly, the presence of CHIP was also associated with cardiovascular disease (CVD), indicating that clonality has wider health implications beyond haematological malignancies. Individuals with CHIP had four times higher risk of developing myocardial infarction, and 1.9 times of developing coronary heart disease [10]. There is functional evidence to support a role for mutations in two of the most commonly mutated CHIP genes in the pathogenesis of cardiovascular disease. The transplantation of bone marrow from *TET2* knock out mice to irradiated low density lipoprotein receptor (LDLR) knock out mice generated a double size lesion in aortic roots after 5 weeks of high cholesterol diet compared to controls [89]. Also, similar work that used bone marrow from *DNMT3A* knock out mice generated a 40% larger size lesion in aortic roots after 9 weeks of high cholesterol diet compared to controls, accompanied by upregulation of the chemokines *CXCL1*, *CXCL2*, *CCXL3*, and the cytokine interleukin 1B (*IL1B*). At the cellular level there was a reduction in T-lymphocytes, and increased macrophage count [90]. The effect was more intense when transplanting bone marrow from *JAK2* V617F mice: after 7 weeks the atherosclerotic lesions were 60% larger compared with controls and this was associated with neutrophilia, neutrophil adhesion to the lesion, and increased erythrophagocytosis that resulted in release of pro-inflammatory cytokines from macrophages [91].

1.1.10.2 Type 2 Diabetes

On the other hand, CH defined by mCA has been reported to have different clinical consequences in comparison to CHIP defined by mutations. A significant association was identified between mCA and type 2 diabetes with an odds ratio (OR) = 5.3, a relationship that was more pronounced in non-obese participants [92]. Differences in the specific targets of mCA and CHIP might be relevant to the different clinical consequences associated with them, e.g. mutations in CHIP most frequently target *DNMT3A*, *TET2* and *ASXL1* whereas mCA most commonly affects distinct regions e.g. chr9p, chr14q and chr20q. The general health condition of individuals with CH may also play a role in triggering some correlations. For example, functional analysis identified an overlap between *TET2* function and glucose levels. *TET2* protein is stabilised by targeted phosphorylation of serine 99, which is mediated by AMP-activated kinase (AMPK), which is regulated by glucose levels. *TET2* stability was restored by anti-diabetic metformin [52].

1.1.10.3 Chronic obstructive pulmonary disease:

A GWAS of pulmonary function identified a genome-wide significant signal in the *TET2* gene, the same signal was significant for individuals diagnosed with chronic obstructive pulmonary disease (COPD; [93]. Recently, a study of 2530 individuals free of haematological malignancies identified a significant association between CHIP defined by *TET2* somatic mutations and COPD [94].

1.1.10.4 Opportunistic infections:

A link between CH and inherited genomic variation might provide an explanation for some correlations. For example, *ASXL1* somatic mutations are enriched in myelodysplastic syndrome (MDS) patients with inherited *GATA2* variants. These inherited variants are also associated with human papilloma virus and non-tuberculous mycobacterial infections [95] as well as *ASXL1* somatic mutations [96].

1.2 The landscape of genetic abnormalities in myeloid malignancies

Classically, cancer driver genes are those that harbour a mutation which confers a selective growth advantage. Driver mutations are independently observed at a higher frequency than expected compared to normal background mutations across multiple malignancies. They can be distinguished from the random mutations, which may arise during the pre- or post-neoplastic stages, which are known as passenger mutations.

The average number of the exonic somatic driver mutations in leukaemia was estimated to be around 9.6 per tumour [97], although subsequent analyses have shown that the number is probably an overestimate and that the number depends on the type of leukaemia. Mathematical models of mutations in self-renewing tissues, such as haematopoietic stem cells, indicate that more than half of the somatic mutations associated with malignancy occur before the initiation of the tumour. In addition, a direct relationship was found between the number of somatic mutations in the tumour and the age of the proband [98]. This mathematical model supports the notion of the random accumulation of mutations in the genome of normal HSCs, until the acquisition of a driver mutation that confers a clonal advantage [99]. The linking of these studies suggests that most somatic mutations, which accumulate in relation to age, are non-pathogenic passenger mutations. Another feature of true driver mutations is that they are usually seen recurrently across different individuals whereas random passenger mutations are often unique.

The strong age relationship suggests that focusing analysis on young cancer patients might help to avoid the noisy background of mutations seen in the elderly. On the other hand, CHIP mutations accumulate with age, and increase the risk of developing haematological malignancies. This new knowledge suggests a division of driver genes into two groups, early mutations that initiate clonal expansion, and late mutations that promotes characteristic features of the disease including a pathological expansion in the size of the clone.

Although CHIP may develop into myeloid or lymphoid malignancy, development of myeloid malignancy is more common and the focus of my study [8,100]. Myeloid malignancies are classified into four main groups: AML, MPN, MDS and myelodysplastic/myeloproliferative neoplasms (MDS/MPN).

1.2.1 Acute myeloid leukaemia (AML)

AML is an aggressive myeloid disease defined by the presence of more than 20% myeloblasts (immature cells) in the bone marrow or peripheral blood, which is indicative of an increase in proliferation and a block in the differentiation of myeloid progenitors.

AML patients harbour a wide range of chromosomal structural abnormalities. Although some of these variants are rare, they are well established as diagnostic criteria by the World Health Organisation (WHO) classification [101]. Some translocations are associated with favourable prognostic outcomes, e.g. AML with either t(8;21), t(15;17) or inv16 [102]. Other karyotypic abnormalities and in particular a complex karyotype is associated with a worse outcome [103]. A summary of the prognostic value of the cytogenetic and the molecular abnormalities in AML are summarised in Table 1-1.

Table 1-1: Prognostic markers in AML according to European LeukemiaNet 2017 recommendations

Prognosis	Chromosomal alterations	Molecular mutations
Favourable	• Core binding factor fusions, inv(16), t(16;16), t(8;21) • t(15;17)	• *NPM1*, biallelic *CEBPA*
Intermediate	• Normal karyotype • Trisomy 8 • t(9;11)	• Core binding factor fusions + *KIT*
Unfavourable (poor)	• Any complex karyotype • Chr 5 monosomy • Chr 7 monosomy • 5q del • inv(3)* • t(6;9) • t(9;22)	• Chromatin (*ASXL1, STAG2, BCOR, MLL-PTD, EZH2, PHF6*) • Spliceosome (*SRSF2, SF3B1, U2AF1, ZRSR2*) • *TP53* • *FLT3-ITD*

* translocation (t), inversion (inv), deletion (del), Internal Tandem Duplication (ITD), Partial Tandem Duplication (PTD)[104]

Although chromosomal structural variations and defined diagnostic markers are well known, about 50% of AML patients have a normal karyotype. NGS and targeted sequencing has identified many somatically mutated genes that contribute to the clinical picture. *NPM1* mutations are identified in more than 25% of all AML patients, and this proportion increases to 50% of patients with a normal karyotype [105]. *FLT3* internal tandem duplication (ITD) in the juxta-membrane (JM) domain or mutations in the tyrosine kinase domain (TKD) occur in about 27% of the AML patients [106]. A large set of epigenetic modifier genes are recurrently mutated in myeloid leukaemia, they include *TET2, DNMT3A, EZH2, ASXL1*, and *IDH1/2*. Splicing factor gene mutations, *SRSF2, SF3B1, U2AF1 and ZRSR2,* are strongly associated with secondary AML, which evolved from clinically covert or overt MDS [107]. Prognosis is closely associated with recurrent genetic abnormalities and forms the basis for the classification of AML according to the National Comprehensive Cancer Network (NCCN) guidelines, and the European LeukemiaNet [104].

Like many cancers, large scale NGS analysis of AML has identified a long list of recurrently mutated genes, including many that are mutated infrequently. For example, in 2013, the sequencing of 200 patients with *de novo* AML identified 23 significantly mutated genes, but also 237 additional genes which were mutated in more than one sample [108]. Very large patient cohorts will be needed to understand the clinical significance of rare, recurrent abnormalities and to identify new driver mutations.

1.2.2 Myeloproliferative Neoplasms (MPN)

Myeloproliferative neoplasms (MPN) are a group of disorders characterised by the clonal proliferation of one or more myeloid cells lineage. The most common type is chronic myeloid leukaemia (CML), which is characterised by the Philadelphia-chromosome and *BCR-ABL1* fusion gene, arising from a reciprocal translocation between chromosomes nine and twenty two, t(9;22)(q34;q11). The disease is characterised by a slow progression and, in some cases, an ultimate block of the differentiation capability of the clonal cells resulting in transformation to acute leukaemia [109].

The other main three subtypes of MPN are polycythaemia vera (PV), primary myelofibrosis (PMF), and essential thrombocythaemia (ET). They are related to each other both clinically and in terms of their pathogenesis, and they can show transitional states between each other, as well as progression to AML. The landscape of driver mutations in MPN is well defined by somatic mutations in *JAK2, MPL,* and *CALR* [110]. 95% of PV patients have a clonal single substitution *JAK2* V617F [69] and some of the remaining cases have a somatic gain of function mutations in exon 12 of *JAK2* [111]. In addition to *JAK2* V617F, ET and PMF can be associated with *MPL* W515 point mutations and *CALR* indels [112,113]. As mentioned above, all three of these mutations are associated with aUPD, with 9p aUPD (associated with *JAK2* mutations) being by far the most frequent [29,114]. In addition, sub-clonal or pre-existing mutations in epigenetic modifier genes (*ASXL1, TET2, DNMT3A, EZH2,* and *IDH1/2*), splicing genes (*SF3B1, SRSF2* and *U2AF1*), and *TP53* are generally associated with worse clinical outcomes [115-118].

1.2.3 Myelodysplastic Syndrome (MDS)

MDS are a group of myeloid neoplasms characterised by bone marrow failure, peripheral blood cytopenia and high risk of evolution to AML. More than half of MDS patients have acquired chromosomal abnormalities, which are fundamental to the diagnosis and prognosis of these disorders. The most common single variants are del(5q), monosomy 7 or del(7q), trisomy 8, and del(20q) [119]. The most updated cytogenetic scoring system raised the number of the prognosis categories from three to five in the revised international prognostic scoring system for myelodysplastic syndromes (IPSS-R) [120,121]. A summary of the prognostic value of the cytogenetic abnormalities in MDS is summarised in Table 1-2.

Acquired somatic mutations are identified in about 90% of MDS patients, with molecular profiling now established in the diagnostic work up and prognostication of suspected MDS cases [122,123]. Like AML, a very wide range of genes are implicated in the disease, and, with the exception of *SF3B1*, mutated genes do not define specific MDS subtypes. It is interesting however that multiple components of the splicing machinery genes are strongly associated with MDS, altering the 3′-splice site recognition during pre-mRNA processing [61,62].

Table 1-2: Prognosis in MDS based on the IPSS-R

Prognosis	**Chromosomal abnormality**
Very Good	del(11q)
Good	Normal, del(5q), del(12), del(20q)
Intermediate	del(7q), del(17q), +8, +19
Poor	inv(3),t(3;3),del(3q),del(7),del(7q)
Very poor	Complex karyotyping > 3 events

1.2.4 Myelodysplastic/myeloproliferative neoplasms

MDS/MPN are a related group of myeloid clonal diseases which have both dysplastic and proliferative features. This group includes chronic myelomonocytic leukaemia (CMML), atypical chronic myeloid leukaemia (aCML), juvenile myelomonocytic leukaemia (JMML) and MDS/MPN-unclassified [124]. Of these, by far the most common is CMML, accounting for more than 80% of MDS/MPN cases.

Chromosomal alterations and somatic gene mutations are found in 30%, and 90% of CMML patients respectively. Trisomy 8, LOY, del(7), del(7q), trisomy 21 and del(20q) are the most common abnormalities, some of which are used for prognostication (Table 1-3) [125]. *TET2, SRSF2, ASXL1* and the oncogenic *RAS* pathway, are frequently mutated. *ASXL1, RUNX1, NRAS* and *SETBP1* are used for risk stratification according to CMML specific prognostic model (CPSS-Mol) [126]. In addition, 35% of patients have aUPD detected by SNP-A [127].

Table 1-3: Prognosis in CMML

Prognosis	**Chromosomal alterations**
Low risk	(normal karyotype or –Y)
Intermediate risk	Others (e.g., 20q-, der(3q), +21)
High risk	(trisomy 8, chromosome 7 abnormalities, or complex karyotype

JMML is characterised by mutations in the RAS pathway, including *PTPN11, NF1, NRAS, KRAS* and *CBL* [128,129]. In addition, secondary mutations in *SETBP1* and *JAK3* are seen [130]. Atypical CML is characterised by myeloid proliferation with low leukocyte alkaline phosphatase values, but with absence of *BCR::ABL1*. *SETBP1* and *CSF3R* mutations are associated with, but are not diagnostic of, this subtype [131,132].

1.2.5 Genetic predisposition to myeloid malignancies

The great majority of myeloid malignancies are sporadic, however in recent years it has become clear that genetic predisposition also plays an important role, with rare high penetrance predisposition genes leading to segregation of disease in families and common low penetrance variants playing a more subtle role.

Familial predisposition to MDS/AML is known to be associated with inactivating variants in several genes and is usually associated with presentation at an age of <40 years. Indeed, 'myeloid neoplasms with germline predisposition' is now recognised as a distinct entity within the WHO 2016 classification of myeloid malignancies [101]. The *GATA2* variant mutation p.T345M was identified in 3 families and segregated with aggressive MDS that transformed to AML [133]. Several other *GATA2* variants were subsequently identified and it was found that individuals who develop MDS/AML are enriched in *ASXL1* somatic mutations [95]. Heterozygous mutations in *CEBPA* gene were identified in AML cases [134], and in particular individuals with biallelic *CEBPA* mutations (one inherited and the second somatic). Germline mutations in *ETV6* and *ANKRD26* have been identified in families with dominant transmission of thrombocytopenia with progression to diverse haematological neoplasms [135], and *RUNX1* mutations in patients with familial platelet disorder with predisposition to develop AML [136]. Although germline mutations in these genes are usually associated with disease at a young age, predisposition variants may affect older patients. In particular, germline variants of *DDX41* are identified in 50% of MDS cases with somatic mutations in the same gene with most affected cases being >60 years old [137].

In 2009, three groups reported that a *JAK2* haplotype called 46/1 (also referred to GGCC) predisposes to the development of *JAK2* V617F-associated myeloid malignancies, and that the *JAK2* mutation generally arose specifically on the 46/1 allele [138-140] The haplotype spans a region of approximately 180kb and although the mechanism by which it predisposes to acquisition of *JAK2* V617F has not been defined, it has been suggested to involve either hypermutation of the 46/1 *JAK2* allele or an interaction that makes the outgrowth of a *JAK2* V617F mutant clone more likely if the mutation arises by chance on the 46/1 allele [141]. However, the predisposition model of *JAK2 V617F* may be more complicated, as haplotype 46/1 may predispose to early alterations of homologous recombination in the *JAK2* gene, before the development of *JAK2* V617F [142].

GWAS has identified a wider set of genes that predispose to the development of myeloid malignancies. A two stage GWAS of 3,437 MPN cases and 10,083 controls had identified a genome wide association signals at rs2201862 (*MECOM*) rs2736100 (*TERT*) and rs9376092 (*HBS1L/MYB*) [143]. Recently, a new GWAS of 2627 MPN cases, and 755,476 controls had identified 14 genome-wide significance loci near 11 genes (*JAK2*, *TERT*, *TET2*, *MECOM*, *KPNA4*, *HMG1*, *PINT*, *GFI1B*, *ATM*, *SH2B3*, and *RUNX1*) [144], also it identified a shared risk between MPN and longer leukocyte telomere length. In the post-GWAS analysis, the mapping of the identified loci to functional data identified *CHEK2* and *GFI1B* as altering the function of HSCs in relation to an increase in the risk of the disease.

1.3 The role of inflammation in myeloid malignancies

Several lines of evidence support the overlap between neoplasia and chronic inflammation. Pathological studies of pre-malignant lesions have identified inflammation mediated by innate immunity as a significant contributor to tumour progression [145]. An elevated risk of developing myeloid malignancies is reported in patients with autoimmune disorders: both AML and MDS are associated with rheumatoid arthritis, but AML is also associated with systemic lupus erythematosus, polymyalgia rheumatica, autoimmune haemolytic anaemia, systemic vasculitis, ulcerative colitis and pernicious anaemia [146,147]. No significant association, however, is reported with CML. Some of the key players in chronic inflammation are discussed below.

1.3.1 Sex hormones

Males and females differ in the development myeloid malignancies, that is supported by the predilection in the incidence of these neoplasms toward males [148]. Consequently, it has been hypothesised that oestrogen could play a protective role in the pathogenesis of malignancies. The main evidence supporting this hypothesis was the finding of over expression of oestrogen receptor α (ERα) in HSCs is associated with self-renewal and proliferation, furthermore, the activation of ERα by tamoxifen induces apoptosis in *JAK2* V617F positive HSCs [149]. On the other hand, lower levels of testosterone are associated with chronic inflammation in males, a high risk of CVD, and higher levels of IL-6 and C-reactive Protein (CRP) [150]. Testosterone has a simulating effect on erythropoiesis [151], and androgen medications increase platelets counts in MDS patients with thrombocytopenia [152].

1.3.2 Interleukin-6

Interleukin-6 (IL6) is a cytokine that plays an important role in inflammation. It binds to its cognate receptor encoded by the *IL6R* gene, and binding initiates an interaction with the gp130 component of the receptor to transduce the signal. Elevated expression of IL6 and other proinflammatory cytokines such as tumour necrosis factor alpha (*TNFA*) are a feature of myeloid neoplasia [153]. Inherited variation on exon 9 of *IL6R* gene (rs2228145; a nonsynonymous change D358A) impairs the function of IL6R, and supresses the interleukin 6 inflammation signal [154]. IL6R D358A has reduced expression at the cell surface in comparison to wildtype, as a consequence of elevated proteolytic ectodomain shedding mediated by ADAM17 [154]. The intronic variant rs4537545 is in linkage disequilibrium with rs2228145, a variant that has been used recently to determine that impaired IL6R function is associated with a reduced risk of development of MPN or its driver mutation *JAK2* V617F [155].

1.3.3 Red blood cell distribution width

Red blood cell (RBC) distribution width (RDW) is a measure for the range of RBC volume calculated by equation:

$$RDW = \frac{Sd\ (MCV)}{MCV}\ X\ 100$$

where MCV is mean cell volume and Sd is the standard deviation

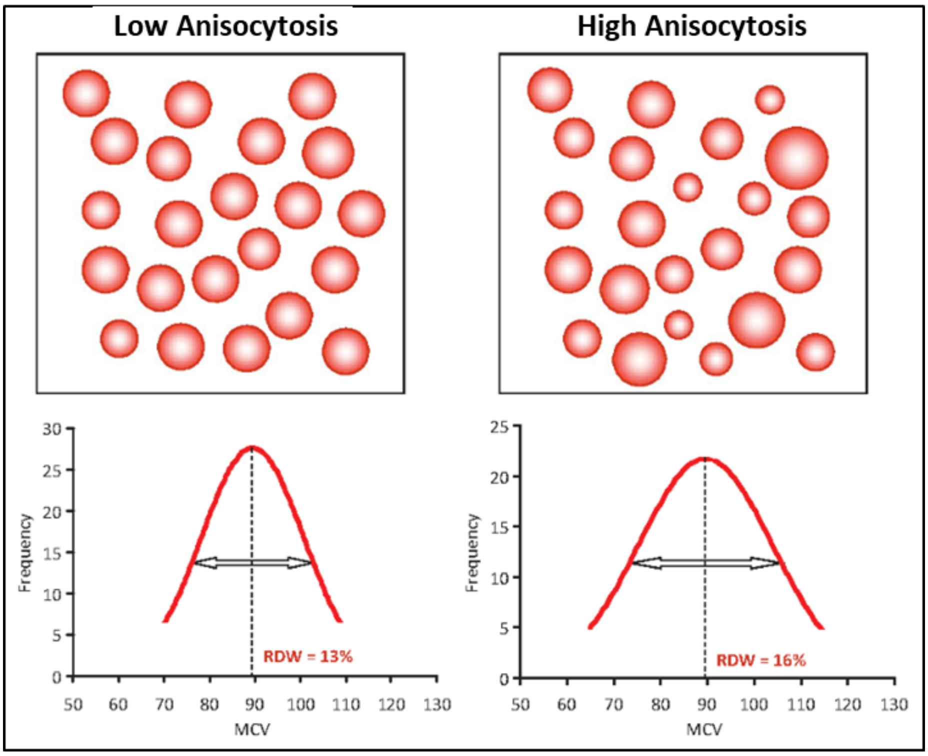


Figure 1-8: RDW is an indicator for the variation in erythrocyte volume

Low anisocytosis refers to RBCs of equal size, and high anisocytosis refers to RBCs of unequal size, that can be estimated from the ratio of the standard deviation of RBC volume to MCV

Elevated RDW is known as anisocytosis, which indicates unequal size of RBCs (Figure 1-8) [156]. This measure is known to be high in patients with elevated levels of inflammatory biomarkers [157]. The associated elevation of RDW with chronic inflammation can be explained by impairment of iron homeostasis, as the inflammatory cytokines such as interferon-γ (IFN-γ) and lipopolysaccharide (LPS) decrease iron uptake and increase its retention in monocytes. This effect is reversed by interleukin-10, an anti-inflammatory cytokine [158].

Elevation of RDW was associated with increased all-cause mortality (HR = 2.56) in a cohort of elderly diabetic patients with coronary artery diseases [159], and also (HR = 3.8) in a cohort of cases with myocardial infarction [160]. In the haematological malignancies, high levels of RDW is associated with Fanconi anaemia, one of the inherited bone marrow failure disorders [161]. Also, it is an independent predictor of development of MDS in patients of unexplained cytopenia [162]. Recently RDW has attracted more attention as it is significantly associated with CH defined by somatic mutations, and furthermore it synergises with clonality to increase the mortality rate [10], possibly as an indicator of disordered erythropoiesis [8].

1.3.4 C-reactive protein

C-reactive protein (CRP) is a circulating protein composed of five identical subunits, that was originally isolated in cases with pneumococcus infection [163]. Years later, CRP was detected in patients with myocarditis and rheumatic fever [164]. The majority of CRP is produced by the liver, and regulated by inflammatory cytokines such as IL-6, so it is considered a global marker for inflammation [165]. In myeloid malignancies, CRP is connected to negative prognosis as myelofibrosis patients with CRP ≥ 7 mg/L have lower leukaemia-free survival [166]. Furthermore, CRP was significantly associated with transformation, death, and thrombosis in patients diagnosed with ET (n=305), and PV (n=172), however, CRP values were not related to mutational profile [167]. Recently, CH was identified as a new risk factor for other chronic inflammatory diseases such as CVD, and COPD, but the relationship between CRP and CH is controversial. In a study of 1887 individuals, CHIP was identified in 427 subjects with 21% higher level of CRP in comparison to CHIP-free subjects [168]. However, in a much bigger study of CHIP consisting of 97,691 individuals, CRP was not significantly associated with CHIP [169].

1.3.5 Smoking as a risk factor

Smoking is a health hazard that drives chronic inflammation at mucosal surfaces and alters the immune response to external pathogens [170]. Early epidemiological studies reported a significant increase in the incidence of blood malignancies among smokers with a predilection toward acute features and myeloid phenotype [171]. Subsequent cytogenetic stratification of AML indicated a relationship between smoking and acquisition of the t(8,21)(q22,q22) [172]. For age related clonal haematopoiesis, CHIP defined by somatic mutations showed a significant association with smoking [8], however, mCA were not significantly associated with this risk factor [7].

1.4 Genetic and genomic screening

Over the last decades, a wide variety of genetic tests have been developed to satisfy clinical and scientific needs. The evolution of NGS provides a cheaper, faster and more accurate technology to assess the genome to a single nucleotide level with enormous depth suitable for cancer-related applications. I outline below the main genetic tests that have been used for genome screening.

1.4.1 Conventional methods and cytogenetics

1.4.1.1 Karyotyping

Karyotyping is a conventional technique of pairing and ordering all chromosomes captured at metaphase and stained by Giemsa to G-band DNA. Cytogenetics is still in wide use and was the first whole genome scan, albeit at low resolution. WBCs, bone marrow or other cells of interest are cultured and arrested at mitosis in metaphase by the use of colchicine. Next, cells are fixed, spread on slide, digested by trypsin and stained by Giemsa [173]. The distinct G-banding of each chromosome allows the identification of numerical changes in chromosomes (aneuploidy) and structural variations (deletions, inversions, translocations) at a resolution >5-10 Mb.

1.4.1.2 Fluorescent *In Situ* Hybridization (FISH)

FISH is the use of fluorescent probes to hybridise to and highlight target sequences. This method is used to detect or identify or confirm large deletions, duplications and translocations [174]. FISH is still in routine use and is particularly useful for quick highly targeted screens, e.g. for the PML-RARα fusion in AML and for many lymphoid disorders for which it is often difficult to obtain dividing cells in culture.

1.4.1.3 Array Comparative Genomic Hybridization (aCGH)

Arrays provides an efficient technique to scan large regions of the genome to identify CNVs such as deletions or duplications [175]. Technically, it is a quantitative comparison of the genomic DNA between a sample and a normal control. The enzymatically fragmented DNA from the test sample and a normal control are labelled with different fluorophores, and mixed together in equal proportions. The mixture is hybridised to unlabelled probes which represent complementary sequence of the targeted regions. Increased probe density on microarrays increased the resolution of the technique enabling the detection of CNVs in the order of a few kilobases (kb). The intensity of the fluorescence signal of each probe is measured and normalised to compare the case sample to the control. Genomic aberrations are detected if the ratio biased 2:2 ratio of sample to control, or biased from ($\log_2 R$ = zero), taking into consideration the diploid state of the human genome [176]. aCGH may need FISH or conventional chromosomal study to confirm its unbalanced translocation results and to overcome its weakness in identification of balanced translocations. aCGH is not generally used in the work of patients with haematological malignancies but is still widely performed for assessment of some rare diseases, e.g. childhood developmental disorders.

1.4.1.4 SNP microarray

SNP microarrays are panels of short length oligonucleotides are designed to hybridise to individual alleles of specific locus, scanning huge number of loci in the same time and comparing dosage of alleles tested to the equivalent value in healthy SNPs database. This technique can detect deletions and duplications with a similar resolution to aCGH but can also detect copy neutral changes (UPD or aUPD) as well as providing genome wide SNP profiles for genetic analysis, e.g. genome wide association studies (GWAS). The success of the SNP consortium in identifying 1.4 million SNPs [177], enabled high resolution DNA chips to be developed. This technique has been used in many international projects like; HAPMAP and 1000 genome projects, to build a haplotype of genetic variation and to assess genetic variations among populations. SNP arrays were used to profile the UK Biobank population cohort [178], and this dataset forms one of the core resources of my study.

1.4.1.5 Sanger sequencing

Sanger sequencing was until recently the most widely used technology to read the sequence of DNA. Developed by Sanger and his colleagues [179], the technique uses “chain termination” with labelled dideoxynucleotides (ddNTPs) which lack the 3′ hydroxyl group and thus cannot be extended further by DNA polymerase. In the original version, radiolabelled ddNTPs were used in four parallel reactions that were then run on four lanes of a polyacrylamide gel, with autoradiography used to identify the base type on the original template. The method was substantially improved by switching to fluorescently labelled ddNTPs and capillary electrophoresis [180,181]. Nowadays, Sanger sequencing is still used as a gold standard to confirm NGS results, and for targeted mutation analysis although its use is rapidly diminishing. It has a limit of detection of 15-20% for somatic mutations and can only generate individual sequences of up to 1000 bases, but this technology was massively scaled up to enable the initial sequencing of the human genome.

1.4.2 Next Generation Sequencing

NGS refers to a bundle of techniques that offer high throughput DNA sequencing at significantly low cost in comparison to Sanger-based sequencing technologies. Sequencing by synthesis is the core technique of NGS, a process that recruits DNA polymerase enzyme to read large numbers of shredded DNA pieces at the same time. The process includes the addition of one fluorescently tagged nucleotide at a time enabling visual signal detection, and repeated rounds of synthesis generates multi-read outputs that can be computationally assembled [182]. Different commercial platforms have been developed, but the technology used by Illumina is described in more detail below as it is by far the most widely used for large scale population level genome projects, and it was the first technology to achieve a $1000 sequencing cost for a full human genome [183].

1.4.2.1 Library preparation

This is the preparation of an indexed (barcoded) library, ready for targeted capture or sequencing. For example the lllumina library preparation [184] process is:

- Fragmentation: DNA of interest is shredded to small fragments
- Blunt end repair: T4 DNA polymerase fills in overhanging 5’ and 3’ ends and a phosphate group is added
- P5 and P7 adaptors are ligated, and filled in
- Indexes are added by PCR amplification
- Indexed libraries are pooled together ready for targeted capture

1.4.2.2 Target enrichment strategies

Target enrichment is a pre-sequencing step that aims to select and amplify the regions of interest [185]. Different targeting technologies have been developed to expand the applications of NGS; PCR based methods have an advantage of the high sensitivity of PCR primers to amplify a single, clearly defined DNA sequence. Multiplexing PCR is more problematic, as some amplicons and particularly those that are GC rich, do not amplify or sequence well, however solutions have been developed such as the Illumina Trusight Myeloid sequencing panel for haematological malignancies that consisting of 568 amplicons covering a region of interest 141Kb in 54 genes. Several other approaches have been developed, including molecular inversion probes that use a linear oligonucleotide that anneal to the targets and enable capture by circularisation [186], and fluidic platforms such at the Fluidigm system that enable multiple singleplex or small multiplex PCRs to be performed prior to pooling [187].

The main alternative to PCR-based approaches are hybrid capture techniques in solution which uses specific probes that define the region of interest and hybridise to a fragmented DNA library [188]. Next, magnetic beads are used for the clean-up, and the captured DNA is eluted for sequencing. This approach is commonly used for large panels and whole exome sequencing.

1.4.2.3 Illumina sequencing by synthesis

The technology uses a flow cell as a solid surface to bind the DNA templates, prepared as described above. To generate clusters of identical copies of each template, solid phase bridge amplification is applied that generates millions of clusters per centimetre [189]. In the sequencing cycles, a single labelled deoxynucleoside triphosphate (dNTP) is added per cycle that stops the polymerization. The dye is imaged to determine which sequence had that nucleotide at the next position (or a run of >1 instances of that nucleotide), and enzymatically cleaved to allow the addition of the next dNTP (Figure 1-9). The process is repeatedly cycled through the 4 nucleotides and the result will be base level sequencing with the coverage depending on the (i) number of the cycles, (ii) read length, and (iii) target length. The Illumina Novaseq platform was used for sequencing the UK Biobank samples.

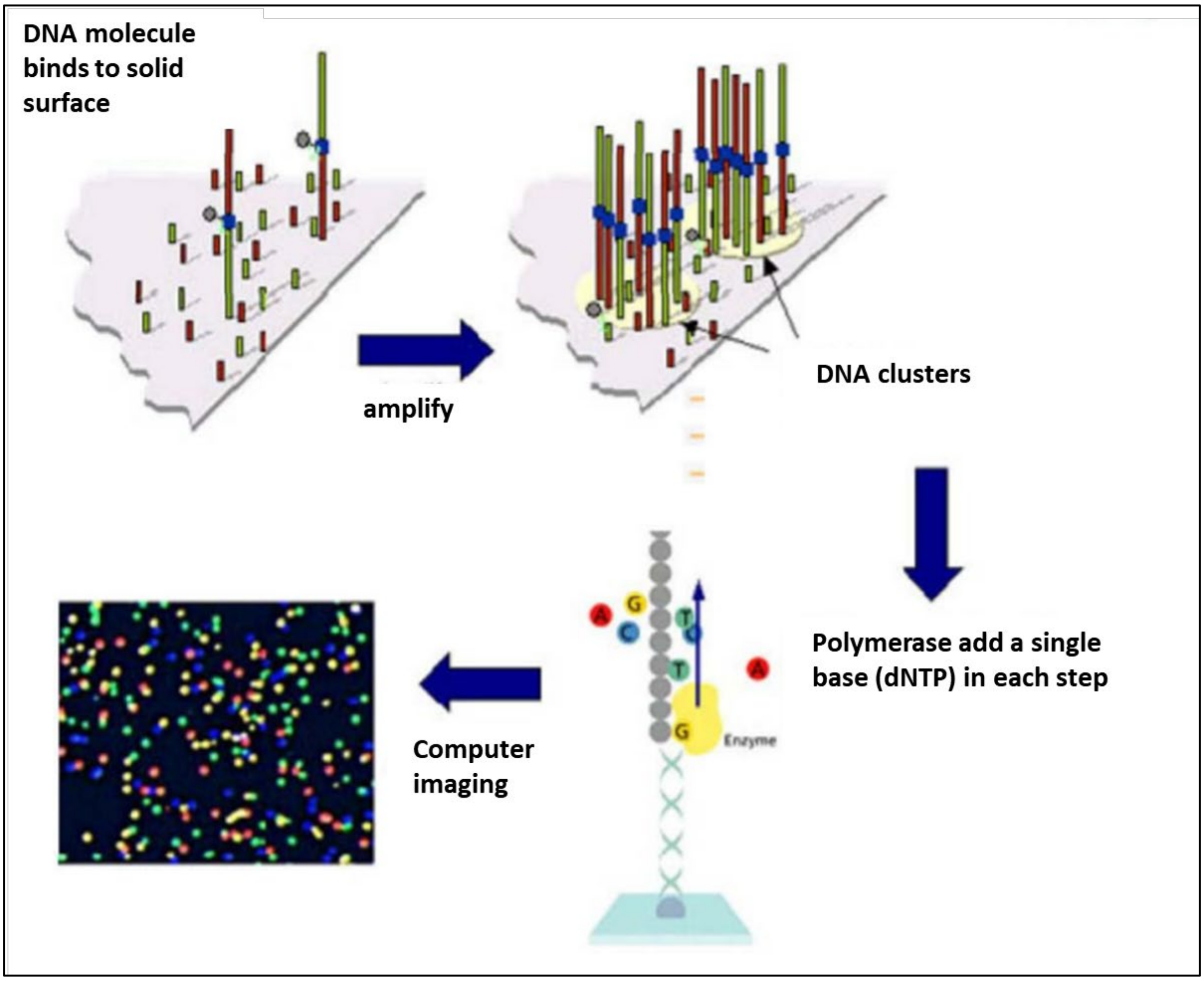

DNA molecule binds to solid surface
amplify
DNA clusters
Polymerase add a single base (dNTP) in each step
A
G
C
T
C
T
A
G
Enzyme
Computer imaging

Figure 1-9: Sequencing by synthesis

Sample libraries, consisting of similarly sized DNA fragments, are washed over the flow cell and bind to the complementary solid support via the appropriate adapter. DNA fragments are amplified for cluster generation by bridge amplification. DNA polymerase creates a complementary strand using the originally attached strand as a template. Next, the double strand molecule is denatured, and the original strand is washed away. The reverse strand bends and attaches to the oligo that is complementary to the top adapter on the flow cell forming a single strand bridge. DNA polymerase creates a complement forming a double strand bridge, that is denatured. The new single strand bends again to continue in the amplification process. The sequence of DNA fragments is determined by a process known as sequencing by synthesis whereby DNA polymerase is used to add chemically modified complimentary bases to the DNA template strand one nucleotide at a time. Each nucleotide contains a fluorescent tag and a reversible terminator that blocks incorporation of the next base. Computer imaging captures the fluorescent signal that indicates which nucleotide has been added. In the next step, the terminator is removed, and the next base is added.

1.4.2.4 NGS strategies

Whole Exome Sequencing: The human genome covers 3 × 10^9 base pairs, but only 1% of the genome (about 30Mb) represents coding sequences (the exome) of the 20,000 human genes. The exome harbours more than 85% of pathogenic mutations associated with genetic diseases and cancer [190] and thus targeting the exome by whole exome sequencing (WES) is efficient with regard to cost and data analysis. WES has a wide range of applications; (i) characterization of monogenic disorders (ii) identification of rare single nucleotide variants (SNVs) associated with complex diseases (iii) identification of somatically acquired mutations in cancer.

Whole Genome Sequencing: is predicted to be more cost efficient for the analysis of human genomes, as the capturing step is skipped. The technical advantage of WGS includes optimal exome coverage, more uniform coverage across the genome and the capability to identify large structural variations (Table 1-4), as well as other metrics of interest in cancer such as tumour mutation burden and mutational signatures. The development of data storage, data processing and faster algorithms are rapidly breaking down the barriers to widespread use of this technology.

RNA Sequencing (RNA-seq): is usually the sequencing of all mRNAs in a population of cells and thus the focus is on the expressed genes. RNA is isolated and converted into complementary DNA (cDNA) that is suitable for library preparation and sequencing [191]. The dynamic and complex nature of the transcriptome raises different scientific applications for RNA sequencing, with a particular focus on analysis of gene expression. For haematological malignancies, RNA-seq is particularly good at detecting fusion mRNAs arising from chromosomal translocations, and many point mutations and indels are also detected. RNA-seq combined with single cell sequencing technologies provides a powerful combination to understand heterogeneity in cell populations and the identity of specific cells.

Table 1-4: Variant types identified by SNP array, WES and WGS

	SNVs		Indel	Structural Variants	CNV
	Exonic	Intronic			
SNP array	Yes	Yes	No£	No*	Yes
WES	Yes	No	Yes	No	Yes
WGS	Yes	Yes	Yes	Yes	Yes

*Structural variations can be detected from SNParray data if associated with a genomic imbalance

£ Some markers on SNP array are binary indels

Technical features of NGS:

Sample Multiplexing is the addition of indexes (barcode) sequences to each DNA fragment library preparation which enables the simultaneous sequencing of different samples. The indexes are used to sort the sequencing reads before the data analysis.

Uniform library construction builds a sequencing library of uniform molar concentrations of different samples enables cost and time efficiency [192].

High sequencing depth is the sequencing of the same region multiple times. The ability to sequence regions hundreds of times independently has revolutionised the oncology field, and has enabled the detection of heterogeneous clones at low frequencies, e.g. a clone of size 2% defined by a heterozygous driver mutation may be detected on average by one read in 100, thus requiring high depth sequencing for its reliable detection.

Paired end sequencing is the sequencing of both ends of a fragment. It generates twice the amount of data for analysis and increases the accuracy of the alignment of the reads to the reference genome at regions of repeats. As the average distance between each pair of reads is known, this technique enables the detection of rearrangements and repetitive elements, and gene fusions.

Long sequencing reads on average standard NGS reads are 100-200 bp long whereas long sequencing using Nanopore (Oxford Genomic Technologies) or PacBio platforms can generate reads of 10-100kb or more. These long read sequences generate more sequence overlap, enable the construction of long range haplotypes resolving and are useful for *de novo* assembly of repetitive areas of the genome.

However the quality of individual base calls is lower than short read sequencing which makes the detection of single nucleotide variants (SNVs) challenging compared to short read sequencing.

Data analysis and bioinformatics pipelines: The ultimate goal of NGS applications is to identify the genomic or transcriptomic variations from the massive throughput of individual reads. In general, NGS has outstanding capability to identify SNVs and small *indels* in the targeted regions. Large *indels* e.g. >20bp-50bp and CNVs may be more difficult depending on the precise methodology employed and the depth of coverage. My focus, described in more detail in subsequent Chapters, is on the bioinformatics pipeline for processing WES data (Figure 1-10). In brief, raw NGS data are represented in FASTQ format files that include identifiers, sequence reads and phred-scaled quality scores for each base representing the estimated probability of an error (Figure 1-11).

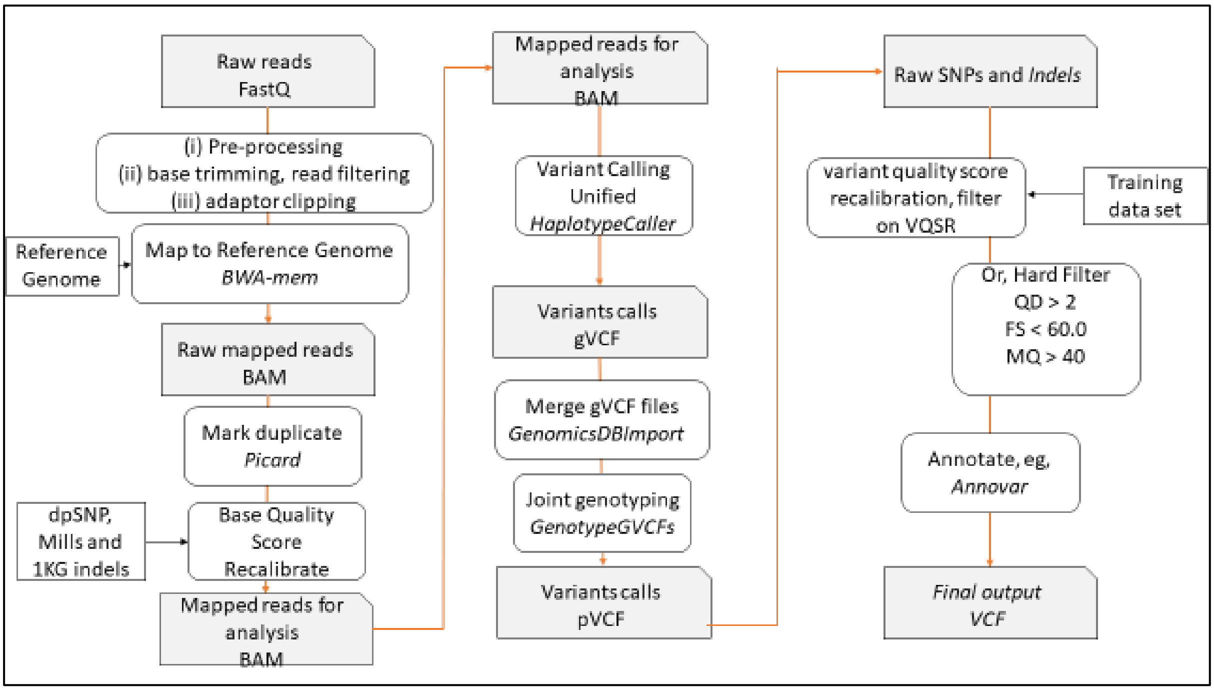
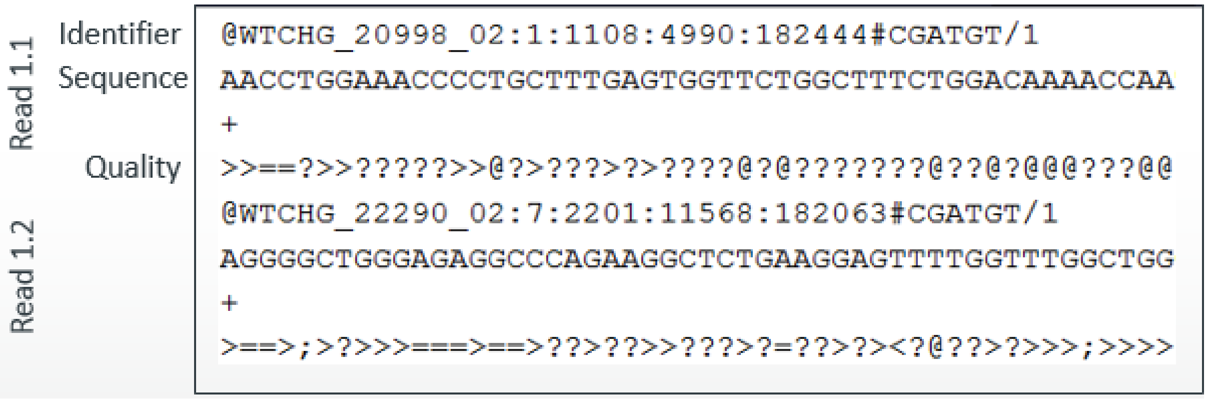

Raw reads FastQ
(i) Pre-processing
(ii) base trimming, read filtering
(iii) adaptor clipping
Reference Genome
Map to Reference Genome *BWA-mem*
Raw mapped reads BAM
Mark duplicate *Picard*
dpSNP, Mills and 1KG indels
Base Quality Score Recalibrate
Mapped reads for analysis BAM
Mapped reads for analysis BAM
Variant Calling Unified *HaplotypeCaller*
Variants calls gVCF
Merge gVCF files *GenomicsDBImport*
Joint genotyping *GenotypeGVCFs*
Variants calls pVCF
Raw SNPs and *Indels*
variant quality score recalibration, filter on VQSR
Training data set
Or, Hard Filter
QD > 2
FS < 60.0
MQ > 40
Annotate, eg, *Annovar*
Final output VCF

Figure 1-10: Data analysis pipeline in the light of GATK best practice guidelines.

For variant calling, each sample is processed independently, but the sample level calls are joined together to improve the genotyping quality of the whole project. Raw FASTQ files are pre-processed by read filtering, base trimming, and adaptor clipping. Next reads passing QC are aligned by Burrows-Wheeler Aligner mem (*BWA-mem*) to the reference genome. The mapped reads in BAM file are marked for duplicates by *Picard*, and Base Quality scores are recalibrated by reference variants against dbSNP and Mills 1000 genome [193] for structural variations. Mapped reads are processed for variant calling by the *GATK Unified Haplotype Caller*. The variant calls from multiple samples are merged by *GATK GenomicsDBimport*, and jointly genotyped by *GATK GenotypeGVCFs*. One of two methods can be used for refining the calls, (i) variant Quality score recalibration (ii) hard filters for quality indices. The variants calls passed quality filters are annotated by *Annovar*, and filtered to nominate the pathogenic variants. VCF refers to Variant Call File. gVCF refers to genomic VCF with additional data for each interval site. pVCF refers to project level VCF with data from multiple samples.



```
@WTCHG_20998_02:1:1108:4990:182444#CGATGT/1
AACCTGGAAACCCCTGCTTTGAGTGGTTCTGGCTTTCTGGACAAAACCAA
+
>>==?>>?????>>@?>???>?>????@?@???????@??@?@@@???@@
@WTCHG_22290_02:7:2201:11568:182063#CGATGT/1
AGGGGCTGGGAGAGGCCCAGAAGGCTCTGAAGGAGTTTTGGTTTGGCTGG
+
>==>;>?>>>===>==>??>??>>???>?=??>?><?@??>?>>>;>>>>
```

Figure 1-11: An example of a FASTQ file.

Each read is represented by 4 lines; (i) the sequence identifier provides data about: instrument name, flow cell lane, tile number, X and Y coordinate, index number and pair end read number (1 or 2), (ii) sequencing read, (iii) A separator, which is simply a plus (+) sign, (iv) base call quality in phred-scaled score using ASCII characters.

Initially, the quality of the raw reads is processed to assess the need for base trimming, read filtering, or adaptor clipping. Illumina sequencing reads are characterised by the presence of 3` end adaptor (Figure 1-12), and usually a drop in the quality at the end of the reads. The sequencing primer anneals to the adaptors, as the synthesis starts from 5' end, so the 5' adaptor is not sequenced, but the synthesis may exceed the targeted DNA fragment length and include the 3' adaptor.

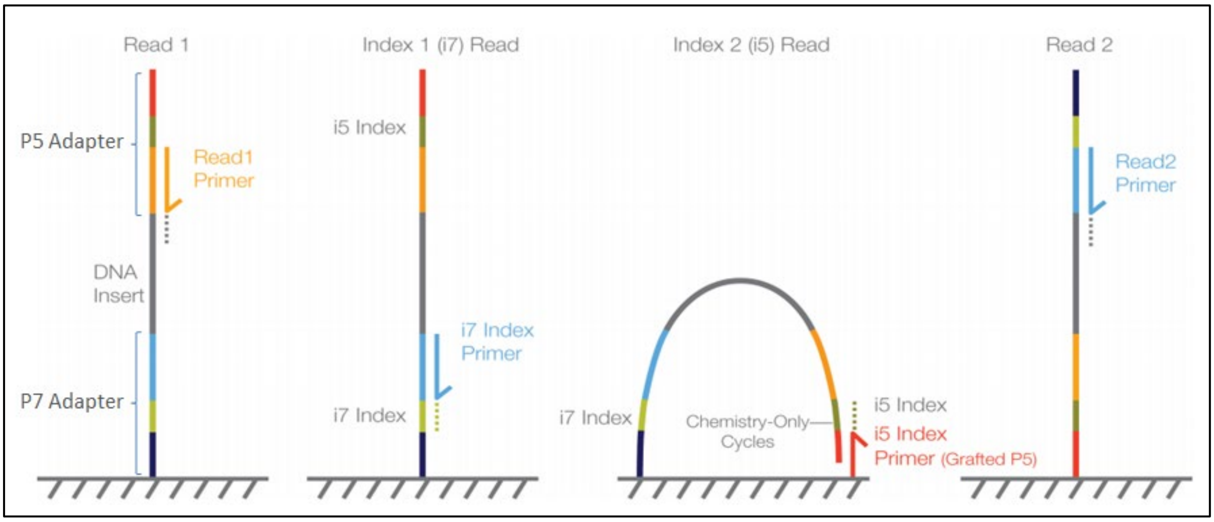


Figure 1-12: The positions of primer annealing used in Novaseq paired end flow cells

Four sequential steps of synthesis are used to generate read1, index1, index2 and read2, respectively. PE PCR primer 1.0 (P5), PE PCR primer 2.0 (P7), index 1 (i7), index 2 (i5)

Next, the short read sequences need to be accurately mapped to the reference genome in a way that preserves any relevant genomic variations. Choosing the proper aligner depends on the application of the experiment, for example Burrows–Wheeler Aligner (BWA) is a general purpose aligner that uses Burrows-Wheeler transform (BWT) algorithm [194] and is regularly applied to WES and WGS data. On the other hand, a splice aware aligner such as HISAT2 is needed to map reads derived from RNA-seq experiments [195].

The alignment process generates Sequence Alignment Map (SAM) files or their binary version (BAM) files. Both include an information header plus read name, read sequence, read quality and alignment information. To improve the data quality and accuracy, (i) read duplicates need to be identified and flagged. Picard MarkDuplicates [196] is one of the tools that uses 5′ mapping coordinates to identify duplicates, and ignores the 3` mapping coordinate which is typically of lower quality. GATK best practice guidelines for alignment include; (ii) *de novo* assembly at *indel* positions, to improve variant calling; (iii) per base quality recalibration considering variables such as the machine cycle, sequencing lane, and dinucleotide content of the current and previous base. GATK BaseRecalibrator [197] is an example, that excludes known variants e.g. those listed in the dbSNP database. Next, the aligned reads are assessed for genetic variations; SNVs, *indels* and CNVs. Different callers are available, but I have focused on two pipelines belong to Genomic Analysis Toolkit (GATK) used in my study:

GATK Germline short variant discovery: GATK standard pipeline recommends the use of GATK HaplotypeCaller [198] to call SNVs and *indel*s simultaneously for single samples. HaplotypeCaller reassembles haplotypes in active regions, that show signs of variations, and then align reads to each haplotype to generate a matrix of likelihood of haplotypes against reads in each active region. Bayes' rule is applied to calculate the likelihood of each genotype and to choose the most likely genotype, and a genome level Variant Call File (gVCF) is generated for each sample. To facilitate next steps, gVCF files from all the samples are assembled using a tool such as GATK4 CombineGVCFs [197-199]. Two main strategies are recommended by GATK best practice to filter the identified variants, the first is to use a machine learning method to recalibrate the variant quality score (VQSR) with the derived VQSR score being used for filtration. The second, hard filter method, is applying static cut-off of quality indices assigned to each variant. Mainly, (i) QualByDepth, QUAL divided by unfiltered read Depth, (QD) > 2.0; (ii) FisherStrand, phred scaled p-value of Fisher test of Strand bias, (FS) < 60.0; RMSMappingQuality, the Root Mean Square of the mapping quality of the reads across all project samples, (MQ) > 40.0.

GATK somatic short variant discovery: GATK best practice suggests the use of Mutect2 to call SNVs and indels from tumour-normal pairs or from tumour only using a Panel-Of-Normal to remove germline variants and artifacts [200]. Mutect2 uses the same method of reassembling haplotypes in regions with signs of variations and generating a matrix of likelihood of haplotypes mapped to reads. Mutect2 uses a Bayesian somatic likelihoods model to predict the odds of alleles to be somatic variants. Other features that distinguish GATK somatic pipeline include the capability to estimate the fraction of reads affected by cross-sample contamination and to filter orientation bias errors.

Annotation and filtration: the identified and genotyped variants are annotated with respect to genomic feature, gene symbols, exons and amino acid change. Annovar is one of the efficient tools that can combine data from a wide range of sources, such as Minor Allele Frequencies (MAF) from genomic databases such as the 1K genome and ESP6500 [201]. Also, it adds an empirical score from different tools to help evaluate the pathogenicity of the identified variants, such as the Combined Annotation Dependent Depletion (CADD) score. Attributes about the pathogenicity of the variants can be added from diverse sources such as the ClinVar and COSMIC databases.

The last step is variant filtration. Mainly, this involves removing variants of low quality, in particular strand bias (where a variant is seen in one direction only, suggesting of a sequencing artefact), and insufficient depth is essential. Other filters can be customised according to the aim of the experiment e.g. restricting the analysis to a set of genes of known significance. Also, selecting rare variants in the public databases is an efficient strategy. For trio data, i.e. a proband with both parents, the mode of inheritance (dominant or recessive) may play an important role in filtering variants.

1.5 Thesis hypothesis and aims

The main hypothesis behind my thesis is "Age-related clonal haematopoiesis is a common phenomenon that increases the risk for developing haematological malignancies and non-neoplastic disorders". Although this association had been established for some disorders prior to my study, the relationship between CH at the driver gene level, identification of other risk factors, and detailed health outcomes were not clear, and form the focus of my study. The introduction focused on providing the background knowledge on CH and myeloid malignancies. Also, I discussed the technological advances in genetic testing that are used to identify CH.

My study has four principal aims:

Aim 1: Assessing the causes and consequences of myeloid-related CH

Previous studies have focused on CH defined by mCA or somatic mutations but not both. I defined myeloid CH by both mCA and somatic driver mutations associated with myeloid malignancies. Chapter 3 describes the initial results using SNP array data from all the UK Biobank participants and WES data from 50,000 participants. Results were compared to age, germline variation, smoking and non-neoplastic disorders. As a result of this work, I was the first author to report the predominant association between *ASXL1* and smoking, confirmed the known association between *TET2* mutations and COPD, and found new associations such as the relationship between *TET2* and agranulocytosis [202]. WES data from an additional 150,000 UK Biobank participants was used as a replication cohort to confirm the findings.

Aim 2: Characterization of the inflammatory stress associated with clonal haematopoiesis

Chapter 4 describes the expansion of the definition of driver somatic variants from germline calls by utilising different filters such as COSMIC, and GnomAD databases. Driver mutations, and previously identified mCA were used to define myeloid CH and lymphoid CH in an expanded cohort of 200,631 participants. Next, I assessed the relationship between CH and chronic kidney diseases defined by eGFR scores. Finally, I assessed the impact of myeloid CH on the risk of developing adverse outcomes in CKD patients. This work produced first published study that define the relationship between myeloid CH, CKD, and CVD [203].

Aim 3: Prediction of the development of myeloid malignancies.

Chapter 5 describes the use of a somatic-specialised variant caller (Mutect2) to process the UK Biobank aligned reads in the absence of a matched normal sample. The method was used to define driver mutations in pre-myeloid cases (i.e. participants who developed a myeloid malignancy >1 year after study entry), and matched controls. CH, other blood measures, and health characteristics were used to model the development of myeloid malignancies by utilising different methods: (i) an Elastic-Net regularised COX-PH model, (ii) a random survival model and (iii) gradient boosted models. My work defined a small number of features that could predict the risk of developing myeloid malignancies with the strongest effect generated by number of lesions in myeloid genes. Different evaluation methods were used to compare the performance of different models. Machine learning models showed much better performance in comparison to the traditional COX-PH models.

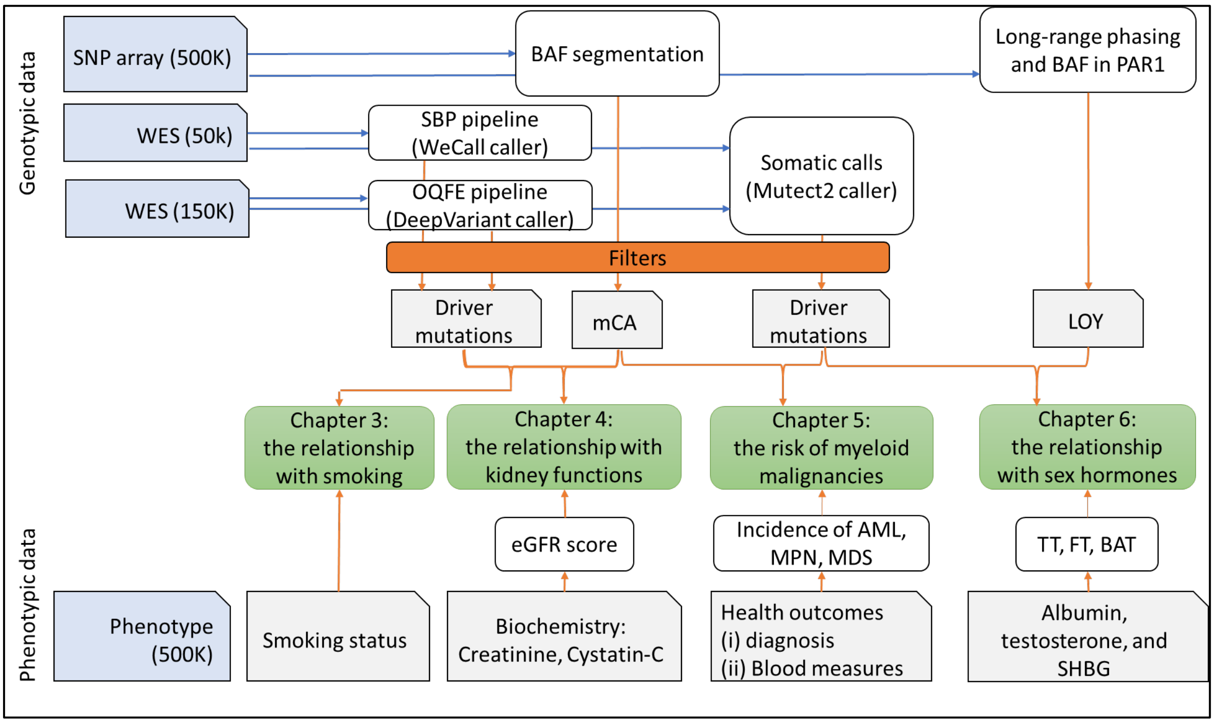
Aim 4: Define the relationship between sex hormones and mosaic loss of chromosome Y

Chapter 6 describes the use of regression and Mendelian randomisation methods to assess the relationship between LOY and sex hormones in the UK Biobank males. Next, eQTL data were utilised to assess molecular signals that potentially could be associated with LOY and its associated sex hormones. This work identified a new, likely causal relationship between sex hormone binding protein and LOY and also characterised the relationship between CH and LOY.

Chapter 2 Methods

2.1 The UK Biobank cohort

The UK Biobank is a major national resource for health research, established with the aim of improving the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses including cancer, heart disease, stroke, diabetes, arthritis, osteoporosis, eye disorders, depression and forms of dementia [178]. The Wellcome Trust and the Medical Research Council (MRC) agreed these goals by planning for a cohort of 500,000 UK participants, aged between 40 to 69 years, and to follow their medical records long-term. The recruitment strategy was based on targeted invitations to attend assessment centres across the country that aimed to enhance the generalisation of the project. Baseline information was collected during the participant visit to the assessment centre. He/she was asked for consent, undertook a series of questionnaires, physical measurements, and provided biological samples (blood, urine, and saliva). Next, the data, and the samples were transferred to the UK Biobank coordination centre. The samples were processed in a central laboratory, and the aliquots stored in an automated biological archive at – 80°C [178]. A summary of the genotypic and phenotypic data used in the thesis were presented in Figure 2-1.



Genotypic data
SNP array (500K)
WES (50k)
WES (150K)
BAF segmentation
Long-range phasing and BAF in PAR1
SBP pipeline (WeCall caller)
OQFE pipeline (DeepVariant caller)
Somatic calls (Mutect2 caller)
Filters
Driver mutations
mCA
Driver mutations
LOY
Chapter 3: the relationship with smoking
Chapter 4: the relationship with kidney functions
Chapter 5: the risk of myeloid malignancies
Chapter 6: the relationship with sex hormones
eGFR score
Incidence of AML, MPN, MDS
TT, FT, BAT
Phenotypic data
Phenotype (500K)
Smoking status
Biochemistry: Creatinine, Cystatin-C
Health outcomes (i) diagnosis (ii) Blood measures
Albumin, testosterone, and SHBG

Figure 2-1: Summary of the UK Biobank genotypic and phenotypic data

Genotypic data: both SNP array, and WES data from the UK Biobank were used to characterise CH across the thesis. **BAF segmentation:** aggregating the B-allele frequencies of consecutive SNPs was used to identify regions of allelic imbalance, and filters were applied to characterise mCA. **Long-range phasing in PAR1:** long-range haplotype phase information in pseudo autosomal region 1 were used to identify allelic imbalance and define the loss of chromosome Y (LOY). **WES**: Three pipelines based on different calling methods, WeCall, DeepVariant, and Mutect2, were used to identify variants and filters were applied to detect putative somatic driver mutations. **SBP pipeline**: Regeneron Seal Point Balinese (SPB). **OQFE pipeline**: Original Quality Functional Equivalence. **Phenotypic data:** the UK Biobank provided lifestyle, physical measurements, and health outcomes. **Chapter 3**: This chapter focuses on the relationship between smoking and CH defined by mCA in 500,000 (500K) participants and putative driver mutations in 6 frequently mutated genes in 50K participants with WES data. Findings were validated using variant calls from DeepVariant in a subset of 150K participants. **Chapter 4**: Chapter 4 investigates the association between kidney function defined by creatinine and/or cystatin-C estimated glomerular filtration rates (eGFR) and CH defined by mCA and somatic driver mutations in a wide range of cancer-related genes using DeepVariant calls in 200K participants. **Chapter 5**: This chapter looks at the incidence of myeloid malignancies (AML, MPN, and MDS) as estimated from hospital episodes, death registry, and cancer registry and their association with health outcomes and blood measures. Mutect2 calls were used to identify somatic driver mutations that were classified according to the name of the targeted gene and their functional impact, presented by the VAF value, and utilised as independent variable in different models to predict the risk of myeloid malignancies. **Chapter 6**: This chapter aimed to determine if any common biochemical measures, including sex hormone levels in men, are associated with LOY and to understand the interaction between genetic and biochemical factors. LOY was based on published findings returned to the UK Biobank.

2.2 The UK Biobank phenotypic data

A wide range of phenotypic data were collected for each participant [178] which can be classified as:

Recruitment data: Data collected by electronic questionnaire at the assessment centre, including socio-demographic and lifestyle data. This data covers all 500K participants.

Physical measurement data: includes blood pressure and weight. In addition, a subset of participants had an electrocardiogram (ECG), as well as hearing and sight tests.

Imaging data: includes magnetic resonance image (MRI) for the heart, brain and body.

Diagnosis and follow up data: supplied by the primary health records and national care records. This information covers 4 resources (i) cancer registries (ii) hospital in-patient data (iii) death registry (iv) primary care data. In addition, a subset of the participants was invited for repeat assessments every few years for calibration of the measurements, and for longitudinal assessment. The diagnostic data are encoded using the International Classification of Diseases, ninth revision (ICD-9) and tenth revision (ICD-10) coding system [204]. ICD-10 was introduced into the national cancer registry in 1995 and into the hospital admissions in 1996.

Biological samples measurements: includes complete blood features, and biochemistry measures for different biomarkers that relate to kidney and liver function as well as cancer.

Web based data: data collected by web questionnaire for a subset of the participants, for example mental health questionnaire for 150,000 participants in 2016.

2.3 Genotyping data, SNP array

DNA was extracted from the blood sample taken at recruitment and used for comprehensive genomic profiling. At the time of writing (August 2022) this includes genome wide SNP array on nearly all participants, and WES on 200,000 participants. WES data from a further 250,000 participants, and WGS for 200,000 participants were released in November 2021 and were not included in the analysis.

The UK Biobank genotypic (SNP array) data includes 488,377 participants. Of these, 85 withdrew consent as of January 2019, and reach, 132 by Feb 2021. These samples were genotyped using two similar microarrays: first a subset of 49,950 samples were analysed using the UK BiLEVE Axiom Array of 807,411 markers (cases selected based on lung function, smoking history and European ancestry), and the remaining samples were analysed using the UK Biobank Axiom Array consisting of 825,927 markers. The two methods share 95% of the markers and are thus highly comparable. The philosophy of marker choice in the UK Biobank Axiom Array was based on incorporating 95,490 known association markers, 111,904 rare markers with MAF <1%, and 629,368 markers to provide good coverage for participants of European ancestry population for imputation and downstream analysis [178].

DNA extraction was performed in Stockport, UK, in 96 well plates; each plate included 94 samples and 2 controls. The extraction procedure used 850µl buffy coat, which was generated from 9ml whole peripheral blood samples. The average DNA concentration was 37 ng/µl and the 260nm/280nm ratio was 1.91, indicative of good DNA quality. Only a fraction of the DNA was shipped on dry ice for genotyping, and the rest was stored for future analysis.

Genotyping was performed by Affymetrix laboratories, Santa Clara, CA, USA, in 106 batches of around 4,700 samples. Briefly, a cluster plot was made for each SNP based on the intensity of fluorescently labelled probes for the A and B allele in each sample. Genotype calls were then made by determining which genotype intensity, either AA, AB or BB, each sample was most likely to belong to.

2.3.1 Quality control by Affymetrix

As the genotyping was done in batches, marker quality was assessed on a batch by batch basis. In addition, Affymetrix checked the DNA concentration and the missingness rates. Overall, a total of 35,014 markers were excluded from the data due to either poor clustering of markers across multiple batches or evidence for more than two alleles (multi-allelic markers).

2.3.2 Marker-based quality control by the UK Biobank

Marker based quality control (QC) was performed using 463,844 participants, who represented the largest ancestral component (European), and the results were applied to the whole cohort. The marker-based QC involved six tests and a P-value threshold of 10^{-12} was used to reject the null hypothesis. This threshold is equivalent to the standard value of P=0.05 after adjusting for the number of tests, batches, plates, and markers (total $4.6\text{x}10^{9}$ tests).

Four tests [batch effect, plate effect, Hardy-Weinberg Equilibrium (HWE), and sex effect] were applied at the batch level whereby marker genotypes would be set to missing for the whole batch if the marker failed any of the four tests. If a marker failed in one of the four tests in all batches, it was excluded for all results. The two other tests (array effect and discordance across controls) were applied across all the batches, and if a marker failed in one of these tests it was excluded (Table 2-1).

Table 2-1: Quality Control tests for microarray markers

Test	**Null hypothesis**	**The probability test**
Batch effect	A batch has the same genotype frequency as all the other batches combined	Fisher's exact test (2*3 table)
Plate effect	A plate has the same genotype frequency as all the other plates within this batch	Fisher exact test
Hardy–Weinberg equilibrium		Exact test in *plink*
Sex effect	The gender has no effect on the genotype frequency of all markers except Y chromosome.	Fisher's exact test (2*3 table) for the autosomal markers
Array effect	The set of individuals typed on the UK Biobank Axiom array has the same genotype frequencies as those typed on the UK BiLEVE Axiom array.	Fisher's exact test (2*3 table) for the diploid markers
Discordance across control replicates	Two controls, HG00097 and HG00264, from the 1K Genomes Project were included in each plate	0.95 concordance is the minimum acceptable for a marker for each of the two controls.

2.3.3 Sample based QC by the UK Biobank

Although outlier heterozygosity or high rate of missingness may indicate poor sample quality, it may also be caused by biological phenomena. The UK Biobank used 605,876 high quality autosomal SNPs to calculate heterozygosity rates (the proportion of heterozygous non-missing SNPs). A total of 224 samples were flagged as having missingness above 0.05, and 744 samples were flagged as having outlying level of heterozygosity that could not be explained by admixture or consanguinity (high rates of heterozygosity may be caused by mixed ancestry, and low rates of heterozygosity may be explained by consanguinity). A further 652 samples were flagged as having sex chromosome karyotypes that did not match XY or XX, and 366 samples were flagged as the genotypic gender did not match the sex reported by the participants.

2.4 Identification of Allelic Imbalance and Loss of Heterozygosity

SNP arrays provide quantitative data for the probe intensity of both alleles and the copy number of each marker. The combination of these data from high density SNP arrays can be used to identify chromosomal abnormalities with the power to distinguish three states of allelic imbalance (AI) and loss of heterozygosity (LOH) associated with copy number loss (CNL), copy number gain (CNG) or copy number neutral changes associated with UPD [205].

2.4.1 BAF value in cancer samples

The B Allele Frequency (BAF) describes the ratio of intensity values for the A and B allele for each SNP in a single sample. In a normal genome, BAFs are expected to form three clusters: one close to the minimum value of 0 for the complete absence of the B allele (i.e. an AA genotype), one close to the maximum value of 1 for the complete absence of the A allele (BB genotype) and one around 0.5 for an equal presence of both alleles (AB genotype). Tumour genomes frequently have chromosomal alterations such as the gain of regions that harbour oncogenes or deletion of regions that harbour tumour suppressors. As a result, the BAF value for heterozygous SNPs in the affected region is shifted away from its expected value of 0.5. Theoretically, the expected BAF of one copy gain is either 0.33 (AAB) or 0.67 (BBA) depending on which allele is gained. However, normal cells in the cancer sample frequently have a diluting effect on the expected values, and in addition there may be multiple sub-clones present.

2.4.2 Log$_2$ R ratio in cancer samples

The relative copy number ratio is the logarithmic value of the observed intensity to the expected intensity of a marker. In the normal diploid genome, the log$_2$ R ratio (LRR) is expected to be zero. Positive shifting of this value is a marker for copy number gain, and negative shifting is a marker for copy number loss.

2.4.3 BAF segmentation of unpaired tumour sample

For detection of aUPD, comparative analysis of matched normal-tumour samples is desirable to remove constitutional homozygous SNPs which may interfere with the application of the segmentation algorithm. For solid tumours, normal constitutional DNA is typically obtained from a blood sample but obtaining matched normal DNA for individuals with haematological malignancies is challenging. For the UK Biobank only a single blood sample was taken for analysis.

For cases with an overt myeloid neoplasm and high white cell count we would expect the great majority of peripheral blood leucocytes to be part of the malignant clone and therefore in the absence of paired normal DNA it is generally not possible to distinguish aUPD from regions of autozygosity. For cases (either CH or myeloid neoplasm) with normal or modest blood counts, however, we expect samples to be a mixture of clonal and normal cells. Practically, this dilution effect of the normal DNA is valuable to distinguish between germline homozygous SNPs and acquired LOH SNPs due to allelic imbalance. This biological phenomenon suggests the use of a fixed threshold to remove non-informative homozygous SNPs. In individuals with normal blood counts, somatic clones are not expected to exceed 90% of the total DNA content, assuming that any clones are restricted to either lymphoid or myeloid cells. This dilution effect generates different relationships between BAF and the tumour content, according to the LOH state (i.e. CNL, CNG, aUPD). Regarding aUPD, the BAF is directly proportional to the tumour content (Figure 2-2).

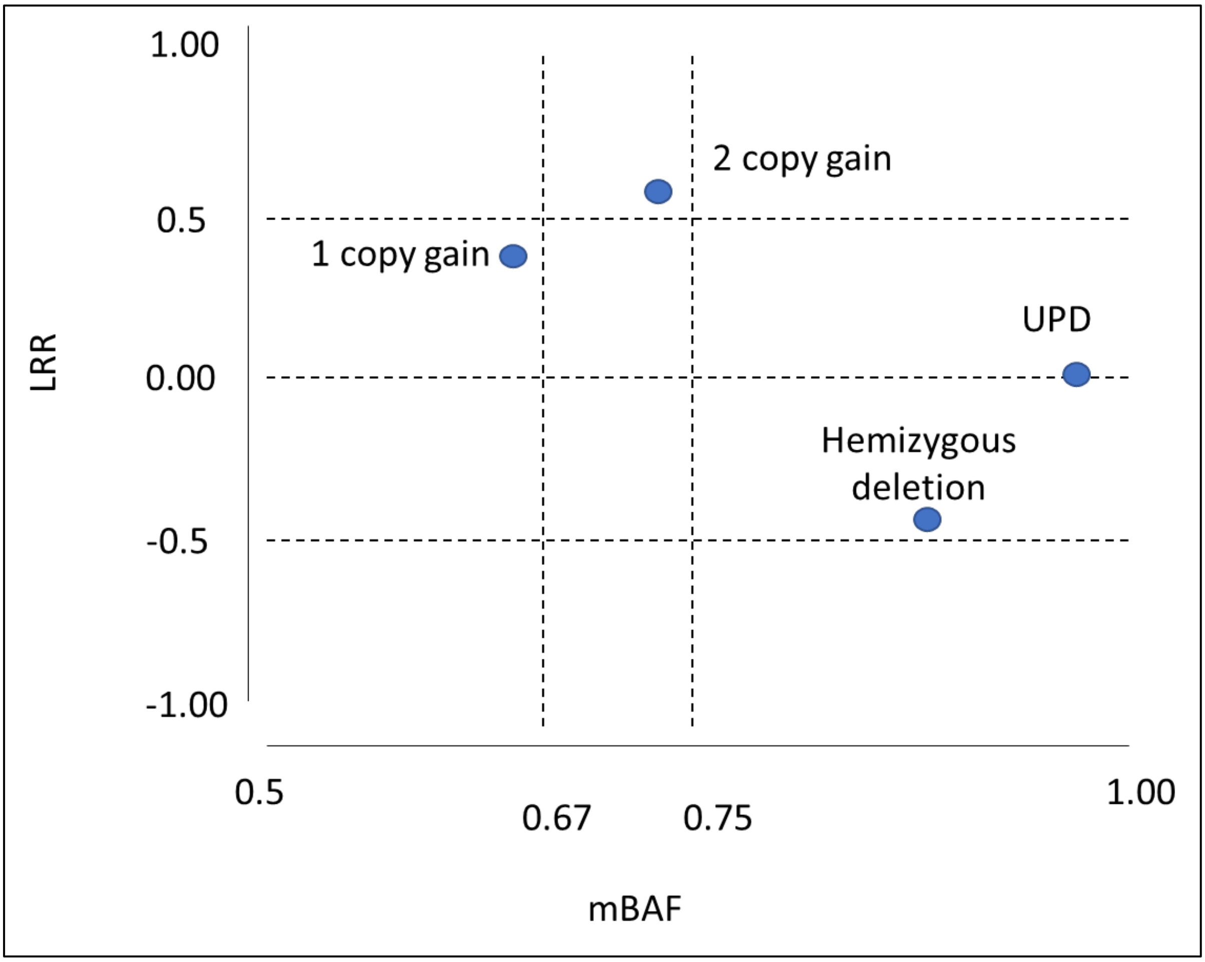


Figure 2-2: The relationship between mirrored mBAF and LRR

The relationship was used to differentiate between aUPD, CNL and CNG. The dilution effect of normal cells shifts the theoretical values of mBAF and LRR in the case of CNL or CNG, but for copy number neutral (CNN) events associated with UPD the LRR remains at zero, with a reduced mBAF that is proportional to the fraction of clonal cells in the sample [206].

Thus, for BAF segmentation of unpaired samples, non-informative homozygous SNPs with mirrored BAF values (mBAF) greater than 0.9 are removed. This threshold may not remove all non-informative SNPs. Any remaining non-informative SNPs are therefore removed by triplet filtering which calculates the absolute sum of the difference in mBAF between an investigated SNP and the SNP immediately before and after. SNPs with a triplet score exceeding a defined threshold are removed [207].

Equation 1: Triplet sum used to filter non-informative homozygous SNPs in BAF Segmentation tool.

$$triplet\ sum[i] = abs(mBAF[preceding\ SNP] - mBAF[i]) + abs(mBAF[succeeding\ SNP] - mBAF[i]) + mBAF[i] - 0.5$$

[i] Stands for the investigated SNP.

The validated cut-off of triplet filter is 0.6 and 0.8 for Affymetrix GeneChipArrays and Illumina Genotyping BeadChips, respectively [207].

2.5 Whole Exome Sequencing

2.5.1 Library preparation

In 2018, 50,000 samples were selected for WES by Regeneron Pharmaceuticals. Cases for sequencing were prioritised towards those with more complete phenotype data and also a primary diagnosis of asthma (16% among sequenced participants, compared to 13% amongst all participants). 100ng of blood-derived genomic DNA was enzymatically fragmented, end repaired, dA-tailed and a Y adaptor ligated to the fragments. The library was amplified by KAPA HiFi polymerase (KAPA Biosystems) in the presence of unique 10 bases barcode. A modified version of IDT's xGen probe library v1.0, was used to capture ~38Mb of the genome. Streptavidin-coupled Dynabeads were used to bind the captured fragments, and a stringent wash was used to remove the unbounded fragments. An amplification step with KAPA HiFi polymerase was applied before the sequencing using 75bp paired end reads on an Illumina Novaseq 6000 platform using S2 flow cells [208].

2.5.2 Alignment and variant calling

The sequencing process yields concatenated base call (CBCL) files of tiles from the same lane. Illumina bcl2fastq tool was used to convert CBCL into sample level FASTQ files based on the id barcodes that were attached during the library preparation. Three different analysis pipelines were applied by the UK Biobank, as shown in Figure 2-3;

(i) **Regeneron Seal Point Balinese (SPB)**: FASTQ files were aligned to the GRCh38 human reference genome using the BWA-mem tool to generate BAM files. Picard MarkDuplicates tool was used to flag the duplicated reads. The WeCall variant caller was then used to define the variants and generate a gVCF file [209]. gVCF files were jointly genotyped by GLnexus to generate a single pVCF file.

(ii) **Functional Equivalence (FE):** the published FE pipeline [210] was applied, characterised by recalibration of base quality score (BQSR). Firstly a model was built of covariation based on the input data from 2 resources: dbSNP138 [211], and the Mills_1000genome [193] databases from the GATK resource bundle (https://console.cloud.google.com/storage/browser/genomics-public-data/resources/broad/hg38/v0). Next the BQSR model was applied to adjust the score of each base to 4 score bins; 2-6, 10, 20, 30 rounded in probability space. The new BAM output was converted into CRAM files. GATK HaplotypeCaller was used to call SNPs and *indels* simultaneously [210].

(iii) **Original Quality Functional Equivalence (OQFE):** An updated version of FE protocol was applied by keeping the original quality score in CRAM files (OQFE). Next, small variants were called by DeepVariant Caller and generated gVCF files. gVCF files were jointly genotyped by GLnexus [212] .

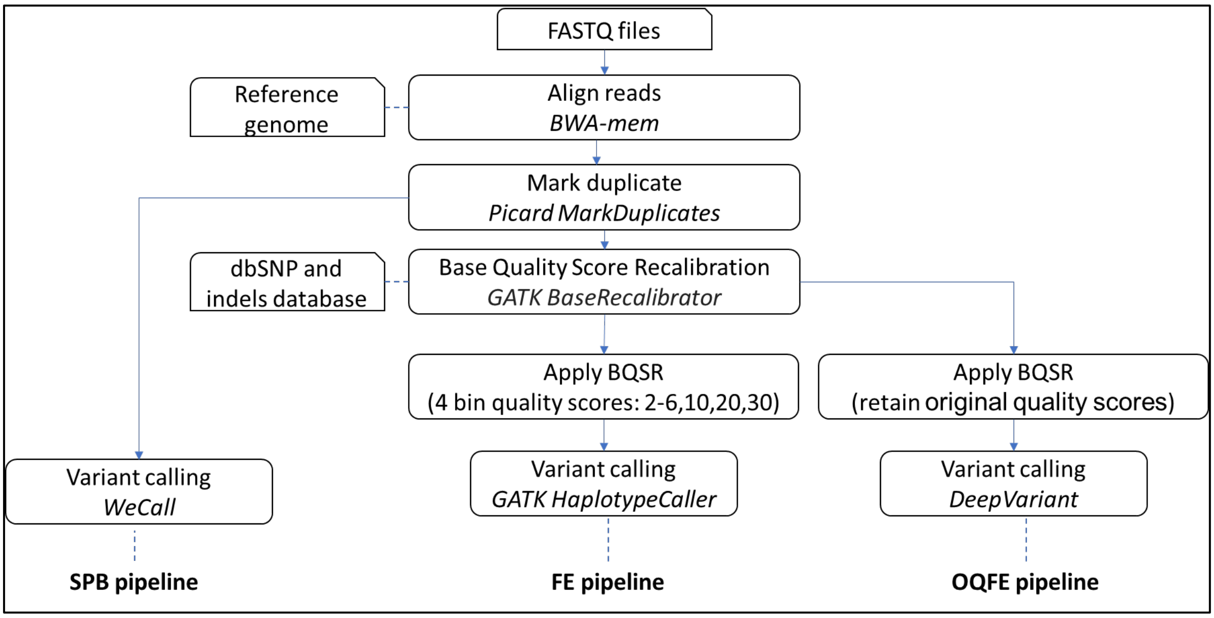


Figure 2-3: A summary of three pipelines used to call variants in the UK Biobank

FASTQ files were aligned to GRCH38 by BWA-mem with flagging split hits as 2048 "supplementary alignment", 100M as the minimum seed length. Picard v2.4.1 was used to mark duplicates. SBP: Regeneron Seal Point Balinese uses WeCall to call variants in all samples jointly. FE: Functional equivalent pipeline used GATK Base Quality Score Recalibration model built on dbSNP183, and Mills/1000 genome indels. The new score was based on 4 bins (2-6, 10, 20,30). OQFE pipeline: Original Quality Functional Equivalence is a modified version of FE that retained the original scores

2.6 Identification of candidate somatic variants

2.6.1 Using germline calls

WES is mainly used to genotype and identify germline variants across the genome with a relatively limited sequencing depth, but it can also identify somatic mutations as long as they have a relatively large allelic fraction. Different strategies have been applied to find somatic variants within germline calls, and all of these are used in this study:

(i) Driver somatic mutations are expected to be ultra-rare in genomic databases with MAF < 0.01.

(ii) Driver somatic mutation are likely to be harboured in a subset of cells, and consequently mutant allele would be present in less than 50% of the sequencing reads arising from that genomic site in many cases (Figure 2-4).

(iii) Recurrence in cancer data bases, mainly Catalogue of Somatic Mutations In Cancer (COSMIC) [213].

(iv) Informatics evidence such as disruptive mutations.

(v) High pathogenicity scores such as Combined Annotation Dependent Depletion (CADD)

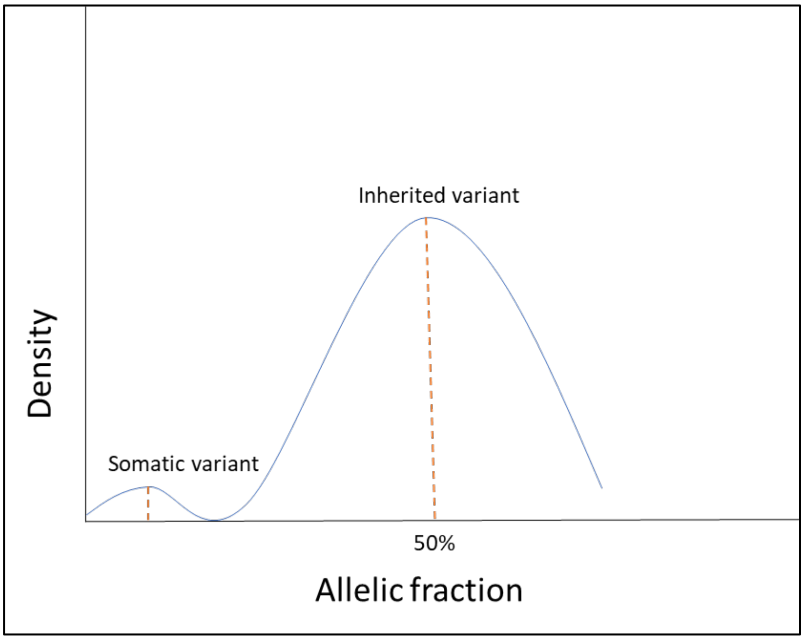


Figure 2-4: Clonal expansion of putative somatic mutations

Putative somatic mutations expand in a subset of cells that can be distinguished by an allelic fraction significantly less than the value of 50% seen for inherited variants

2.6.2 Somatic variant calling

The ideal strategy for calling somatic mutations is based on comparing tumour-normal pair and identifying loci that are present in the tumour sample and absent from the normal sample. Somatic calling is tuned to detect variants with a low fraction of mutant reads as tumour samples can be contaminated with germline cells, have intra-tumour heterogeneity, or copy number changes. In addition, somatic calling omits ploidy in the genotyping likelihood calculations and applies filtration strategies to flag common germline variations, multiallelic sites, and recurrent artifacts identified in a panel of normal samples. I chose GATK Mutect2 [197] to call somatic mutations as it can run in tumour-only mode that utilises a panel of normal and germline resources to identify somatic mutations in individual samples in absence of matched normal pairs.

2.7 Survival analysis

Survival analysis is the statistical method of predicting time to an event such as death or disease diagnosis. It mainly differs from other regression forms by the capability to deal with censored data, i.e. unobserved events during the study time. Two probabilities are estimated in survival analysis, the first is the survivor function *S(t)* that describes the probability of survival between 2 time points, the second is the hazard function *h(t)* that describes the probability of an event at specific time point.

2.7.1 Kaplan-Meier curves

The Kaplan-Meier method estimates survival probability $S(t_i)$ from observed survival $1- (d_i / n_i)$ at specific time points t_i. It is a univariate method that can be applied in a stepwise fashion at the time of each event [214] and is suitable for visualisation as a survival curve to show the relationship between time t_i and survival probability $S(t_i)$. The cumulative hazard is an estimation for the cumulative force of events at specific time, and it represents -log (survival function). A non-parametric log rank test is used to compare the probability of events between two groups at any time point.

2.7.2 Cox proportional hazards model

The Cox proportional hazards model was developed to adjust survival analysis for other variables that affect survival. Cox found that an effect parameter can be estimated for each covariate without consideration for the hazard function and the effect of a factor can be reported as a hazard ratio [215].

$h(t) = h_0(t) \times \exp(b_1x_1+b_2x_2+...+b_px_p) = h_0(t)\times\exp(b_1x_1)$

2.7.3 Machine learning methods

The UK Biobank and other large cohort studies acquire data from a wide range of sources that include questionnaires, laboratory assays, and hospital records. The vast amount of data sources shows high dimensionality, plentiful missingness, and heterogeneity in data types. Cox's model has problems in modelling survival with high dimensional data due to the correlations among factors. The extension of machine learning methods to handle censored data have allowed its use in survival analysis and previous studies have indicated the outperformance of machine learning over Cox in survival analysis [216].

2.7.3.1 Random Survival Forest

Random forest classifier is based on an ensemble of decision trees [217]. Each decision tree uses a randomly selected number of subjects and factors. Random forest achieves a reliable result in comparison to other classifiers regarding computational time, and the capability to rank variables. The capability of random forest to analyse right censored data, was introduced by developing new splitting rules for growing survival trees, and by using conservation-of-events rule that define ensemble mortality [218]

2.7.3.2 Gradient boosted models

Gradient boosting is a framework to combine the prediction of base learners to improve the overall survival model. The additive of each base model in a greedy fashion improves the overall model. Different loss functions and base learners had been used [219]. In my project, I tested each of (i) Cox's partial likelihood with regression tree to maximise the log partial likelihood function (ii) component-wise least squares base learners which minimises the residual sum of squares (iii) accelerated Failure Time (AFT) model with inverse-probability of censoring weighted least squares error.

2.8 Mendelian randomisation

Mendelian Randomisation (MR) is the use of genetic variation to assess causal relationships between risk factors and health outcomes [220]. The art of applying MR is based on selecting robust genetic variations associated with the study exposure and utilizing statistical methods to consistently estimate the effect on outcome. The selection of the genetic variants could be based on biological relevance or statistical significance such as the threshold used for GWAS ($P < 5 \times 10^{-8}$) [221]. The selection of genetic variants in the same cohort used for outcome investigation would exacerbate any biases, but this is overcome by using independent cohorts for genetic association and outcome evaluation. Three assumptions must be satisfied by each genetic variable, as shown in Figure 2-5, (i) the association with the exposure (ii) no association with cofounders of the relationship between exposure and outcome (iii) independent association with outcome [222].

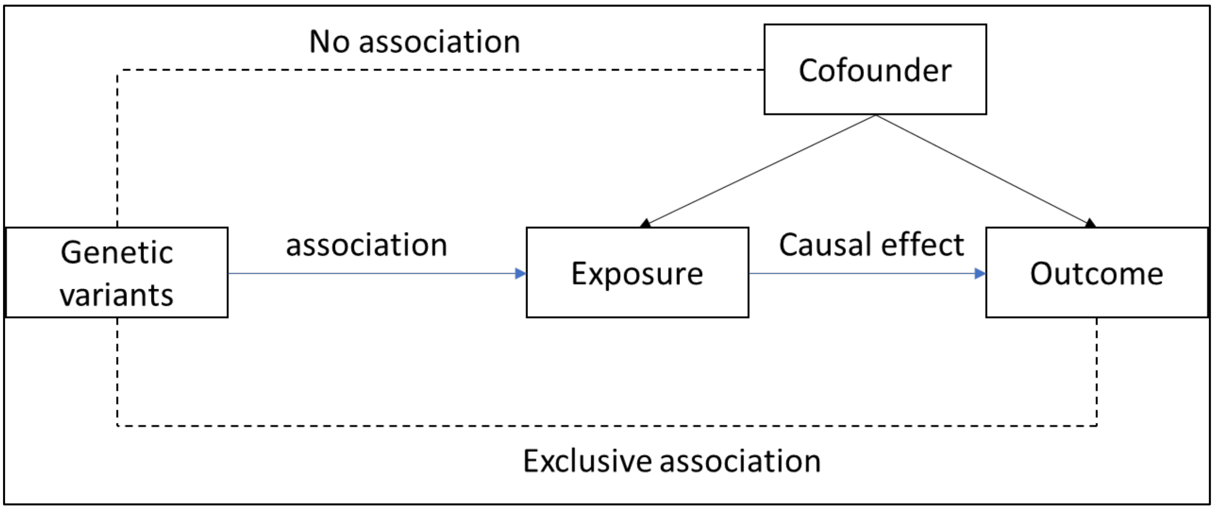


Figure 2-5: The assumptions of Mendelian randomisation

2.8.1 MR with single genetic variant

The simplest use of MR is the use of a single variant that is supported by biological knowledge. A single variant can be enough to describe gene expression or protein synthesis. For examples, *IL-6R* SNP rs7529229, a marker for *IL-6R* p.Asp358Ala, associated with increased IL-6, and decreased C reactive protein, was used to study IL6R blockade from infusions of tocilizumab as a potential therapeutic for to prevent coronary heart disease. rs7529229 was associated with a decreased odds of coronary heart disease events [223].

2.8.2 MR with multiple genetic variants

Multiple genetic variants can collectively explain more of the risk and have more statistical power than a single variant. They can be aggregated into a polygenic risk score (PRS), or be used as individual instruments, as I will discuss.

2.8.2.1 Allele score method

An allele score is the sum of multiple genetic variants that are associated with the risk factor. On the individual level, an unweighted score can be calculated as the total number of risk alleles. Weighted scores account for the estimate of the effect of each genetic variant as a weight reflection [224]. The calculated allelic score can be used a single instrumental variable to predict the risk of outcome in MR analysis.

2.8.2.2 Multiple instruments methods

Multiple genetic variants can be used as instrumental variables in a regression model to estimate the effect on outcome. Different statistical models have been developed to test different assumption`s, as I present in Table 2-2.

Table 2-2: some statistical methods used for mendelian randomisation

Statistical method	Description	Requirements
Inverse-variance Weighted [225]	The causal effect is estimated from the meta effect of the ratio estimates for individual variables.	All genetic variants are valid instruments
Weighted median [226]	It uses the average pleiotropic effect as the intercept which allows the use of instrumental variables with pleiotropic effects	More than half of the variants are valid instruments
MR-Egger [227]	The causal effect is estimated as the slope from the weighted regression of the ratio estimates for individual variables. It uses the average pleiotropic effect as the intercept which allows the use of instrumental variables with pleiotropic effects	Accept variants with pleiotropic effect; the pleiotropic effect should be independent of the exposure. The Instrument Strength Independent of Direct Effect (InSIDE) assumption
Robust Adjusted Profile Score (MR-RAPS)[228]	It uses random effects distribution to model the pleiotropic effects and provide estimates by using profile-likelihood of casual effect and the distribution effect.	Normal distributed pleiotropic effect

2.8.3 Populations used in MR studies

NHLBI Trans-Omics for Precision Medicine (TOPMED) [229] is part of precision medicine initiative, a wider framework to develop personalised medicine in USA [230]. WGS data of 97,691 individuals was used to characterise somatic driver mutations and identified 4,229 with CH [169]. Next, a GWAS was performed using individuals with high likelihood (>1%) of having CH (n=65,405) of them 3,831 CH cases. The study identified four independent signals (*TERT*: rs34002450, rs13167280; *TRIM59*: rs1210060191; *TET2*: rs144418061)

Chronic Kidney Disease Genetics Consortium (CKDGen) is an open consortium that aims to identify common genetic risk factors associated with kidney function, estimated by glomerular filtration rate, and albuminuria [231]. One of the biggest studies was a meta-analysis of 60 GWAS with total 625,219 individuals of them 64,164 cases which identified 23 genome wide significant loci [232].

Biobank Japan (BBJ) is a disease-based cohort of 260K patients representing 51 common diseases [233]. I used BBJ data to study the relationship between SHBG and LOY. Mean log-R ratio (LRR) in 95,380 men were used to estimate the degree of mosaicism in LOY [234]. GWAS had identified 50 independent signals associated with LOY in BBJ.

2.9 Programming tools and statistical tests

2.9.1 awk

Awk is freely available programming language for text processing and data extraction [235]. It provides a fast and efficient way to process big data in Unix/Linux operating systems with capability of building a complicated program using the simple conditioning and looping functions.

2.9.2 R programming

R is an open source programming language for statistics and graphics [236]. It has an environment capable of handling big data, operating calculations, and displaying graphics. It provides simple structured programming tools such as conditioning, and loops. R has a command line interface, but different graphical interfaces have been developed to support it such as R studio, and R notebook. Different packages are used to accomplish the study;

karyoploteR: It is a tool that combines many graphical sets in R to plot karyotypes on the genome [237]. It process input files that is encoded in GRanges format of GenomicRanges package (Chr, start, end, strand) [238].

Survival: Survival provides the tools to conduct a survival analysis that mainly includes survival object, and Kaplan-Meier and Cox model [239].

Lubridate: it is a tool to deal with dates and calculate the time intervals from different date format [240].

powerSurvEpi: it is a tool to calculate the power and sample size in the survival analysis [241].

TwoSampleMR: it is an R library for performing MR using GWAS summaries in two-sample strategy and provides a range of statistical methods to test multiple assumptions. The package allows the use of GWAS database of MRC integrative epidemiology unit [242].

2.9.3 Python

Scikit-learn: scikit-learn is a Python library for machine learning built on top of SciPy [243] and characterised by many algorithms for classification, clustering, and regression [244].

Scikit-Survival: scikit-Survival [245]is a Python library for survival analysis that utilise the pre-processing and cross-validation tools available by scikit-learn [244]. Scikit-Survival provides a variety of survival algorithms that include Cox proportional hazard models, ensemble-based methods, and survival support vector models.

matplotlib: matplotlib [246] is a visualization library for data and statistics presentation in Python .

seaborn: Seaborn [247] is a visualization library based on matplotlib for statistical graphics. Seaborn is used for the semantic mapping and statistical aggregation to produce informative graphs.

eli5: eli5 [248] is a package to explain the machine learning models. It is concordant with scikit-learn and explains the weights and predictions of its classifier and regressions. Eli5 uses the permutation importance rule to evaluate machine learning models by measuring the score decrease when the feature excluded.

2.9.4 BAF Segmentation

BAF Segmentation is a tool to detect regions of allelic imbalance from B Allele Frequencies of the markers on SNP array [207]. The tool can be used for paired tumour-normal samples and unpaired samples. Non informative homozygous SNPs are removed from the BAF profiles. Next, circular binary segmentation (CBS) is used to combine regions with similar allelic proportions which are called as allelic imbalance by comparison to a fixed threshold.

The tool can deal with a file of multi-samples or separated single file for each sample with BAF, and LRR data for each marker. The output is a list of the identified allelic imbalanced regions, and a set of 3 plots for each sample; a BAF plot, a mirrored BAF plot, and a log R ratio plot with all SNPs with average log R ratios within mBAF segments superimposed.

2.9.5 Annovar

Annovar is a functional annotation tool for genetic variants identified from different genomes [201]. It deals with simple tab delimited input of (chromosome, start position, end position, reference nucleotide and observed nucleotides). Information from a wide range of resources can be added, including: (i) annotated reference transcriptomes such as RefSeq to detect the targeted gene, exome, and transcript; (ii) annotated genomic intervals that include conserved regions, transcription binding sites, and DNAse I hypersensitivity sites (iii) annotated databases such as COSMIC, 1000 genome and GnomAd. Also, it can add pathogenicity scores such as SIFT, FATHMM, and polyphen.

2.9.6 PHESANT

PHEnome Scan ANalysis Tool (PHESANT) is a phenome scan tool that incorporates R scripts to scan the UK Biobank phenotype files and apply different association tests between the phenotypes and the trait of interests [249]. The tool can deal with different traits of interest according to the experiment design including single SNPs, genetic scores or different genetic features. This makes PHESANT suitable to conduct phenome-wide association studies (PheWAS) and Mendelian randomisation approaches. Also, it can test the association between different phenotypes, referred as Environmental WAS (EnWAS).

After the tool scan and categorisation of the phenotype file, PHESANT runs parallelised regressions for the trait of interest on the selected phenotype. The tool chooses the appropriate test for each phenotype based on rules documented in the variables file. In general, (i) linear regression is used for testing the association of continuous variables e.g., blood counts, after inverse normal transformation to counteract departures of continuous variables from normality; (ii) logistic regression is used for testing multiple categorical variables e.g., ICD-10 encoded diseases; (iii) ordinal logistic regression and multinomial logistic regression are applied for categorised variables if they have ordered categories or unordered categories, respectively.

The input data are (i) the UK Biobank phenotype set file (ii) the trait of interest file that varies according to the experiment (it may include single SNPs, genetic scores, or different genetic features); (iii) data coding information file (iv) a variable information file (v) a cofounder file [249].

2.9.7 Other statistical tests

Fisher's exact test: a test that assesses the null hypothesis of independence of the numbers in the cells of a 2x2 contingency table.

Mann-Whitney U test: a nonparametric test to assess the null hypothesis that it is equally likely that a randomly selected value from one population will be less than or greater than a randomly selected value from a second population.

Binomial test: compares the number of successes observed in a given number of trials with a hypothesised probability of success.

Chapter 3 Characterization of myeloid clonal haematopoiesis in the UK Biobank

3.1 Summary

In this chapter, I describe the investigation of CH in the UK Biobank (n = 502,524, median age = 58 years, range 40 to 70 years). Utilizing data from SNP arrays (n = 486,941), I identified 8,203 instances of mCA in 5,040 individuals, with the prevalence ranging from 0.85% at 40-45 years to 1.29% at 66-70 years, a significant age-related increase (OR = 1.017; 95% CI = 1.013 - 1.020; P = 1.80x10^{-19}, logistic regression test). Classifying these mCAs by chromosomal arm and copy number state identified 17 abnormalities involving 15 chromosomal arms that were significantly associated with myeloid disorders in 506 individuals. The risk of acquiring myeloid mCA (n=506) showed a sharper increase with age (OR = 1.10; 95% CI = 1.08 - 1.11; P = 1.57x10^{-38}, logistic regression test).

Within a subset of the cohort (n=49,956), WES data was used to identify likely somatic driver mutations in *DNMT3A, TET2, ASXL1, JAK2, SRSF2* or *PPM1D* that were rare (MAF ≤1%) in population databases and were either loss of function mutations or overlapped with known mutations. These criteria detected 721 candidate mutations in 678 individuals and, similar to myeloid mCA, were associated with age (OR = 1.10; 95% CI = 1.08 - 1.11; P = 5.89x10^{-47}; logistic regression test). In total, the analysis yielded 1,166 individuals with myeloid-related CH defined by one or more of myeloid associated mCA in 506 individuals (0.1% of subjects who had a SNP array) and/or likely somatic driver mutations in one or more of the six genes of interest in 678 individuals (1.4% of cases who underwent WES). A total of 18 subjects had both mCA and somatic mutations. Next, I investigated genetic features and exposure factors as causes for the development of myeloid CH. 30,892 individuals with WES data were selected as controls that were free of any mCA, had no putative somatic mutations in the six genes of interest and did not have any haematological malignancies during the study period. Using a genome-wide association analysis to compare these groups, I identified two distinct signals (rs2853677, OR = 1.32, P = 5.6x10^{-11}; rs7726159, OR = 1.33, P = 4.2x10^{-11}) within *TERT* that predisposed to myeloid CH, plus a weaker signal corresponding to the *JAK2* 46/1 haplotype. Smoking history was significantly associated with myeloid CH: 53% (n=622) of myeloid CH cases were past or current smokers compared to 44% (n=13,651) of controls ($OR_{previous}$=1.17; $OR_{current}$=1.76; P = 3.38x10^{-6};

multinomial logistic regression), a difference principally due to current (P_{FDR} = 6.14x10^{-6}, OR = 1.1; ordinal logistic regression) rather than past smoking (P_{FDR} = 0.085). Strikingly, breakdown of myeloid CH by specific mutation type revealed that *ASXL1* loss of function mutations were the most strongly associated with combined smoking status (OR_{past}=1.94, $OR_{current}$ = 4.68, P = 1.02x10^{-5}), that was more likely due to current smoking status (OR=1.07, P_{FDR} =1.92x10^{-5}), rather than past smoking (OR = 1.04, P_{FDR} = 2.60x10^{-3}). This finding was confirmed in a new release of WES data for 150,685 independent samples (OR_{past} = 1.34, $OR_{current}$ = 2.97, P = 3.43x10^{-6}), a finding that is largely attributable to current smoking status (OR = 1.04, P_{FDR} =2.01x10^{-7}) rather than past smoking status (OR = 1.01, P_{FDR} = 0.05). Indeed, 64% of participants with *ASXL1* mutations (n=327) were past or current smokers in both cohorts. Using meta-analysis to combine these results, the overall risk of carrying a somatic driver mutation in *ASXL1* was estimated to be 1.05 times higher per unit for current smokers (P_{FDR} = 8x10^{-13}).

Survival analysis revealed that individuals with myeloid CH and without any diagnosis of haematological malignancies (n = 911) have an increased risk of all-cause mortality (HR = 1.44, CI: 1.05 – 1.99, P = 0.02, Cox-hazard model). This suggests that myeloid CH is associated with other medical conditions and not just haematological malignancies. The correlation of myeloid CH with different clinical phenotypes, blood features and biomarkers highlighted a qualitative relationship between clonality and age-related inflammation. Mainly, myeloid CH was associated with red blood cell distribution width (RDW, OR = 1.02, P_{FDR} = 9.7x10^{-4}), and alterations in erythropoiesis and thrombopoiesis. However, each mutated gene has specific associations. Importantly, *TET2* mutations were associated with chronic obstructive pulmonary disease (COPD, OR = 1.16, P_{FDR} = 0.009) and significant agranulocytosis (OR = 1.23, P_{FDR} = 0.009), whereas *JAK2* V617F and chr9p mosaicism was associated with an elevation in platelet counts (OR = 1.04, P_{FDR} = 1.3x10^{-11}) and platelet crit (OR = 1.04, P_{FDR} = 7.5x10^{-12}). *ASXL1* mutations had an anaemia-like blood profile even after correction for smoking, mean corpuscular volume (MCV, OR = 0.98, P_{FDR} = 9.13x10^{-4}), and mean corpuscular haemoglobin (MCH, OR = 0.98, P_{FDR} = 2.86x10^{-3}). In general, myeloid CH has a heterogeneous genetic architecture that mirrors heterogeneity in age-related inflammation and plays a role in the pathogenesis of several diseases of aging.

3.2 Introduction

The combined prevalence of all myeloid malignancies, as defined by the 2016 WHO classification [101], ranges from between 0.3% for 3 year prevalence to 0.8% for 10 year prevalence with a median diagnostic age of 72.4 years [250]. These diseases are rare, tend to occur in elderly people, and are characterised by a wide genetic heterogeneity. The diagnosis of these disorders is very variable. Some cases are asymptomatic initially and are picked up by the finding of abnormal blood counts on routine assessment or investigation of another condition. Some cases present with non-specific symptoms such as easy bruising, lethargy and night sweats whereas others have an acute presentation with multiple abnormalities. The diagnosis of myeloid neoplasms has traditionally been made by morphological investigation of bone marrow cells along with bone marrow and peripheral blood counts, but the finding of blood cell clonality and specific characteristic somatic abnormalities are playing an increasingly important role to diagnose myeloid neoplasms, identify the subtype and predict prognosis. However, the identification of CH in healthy individuals as a result of pathogenic mutations in myeloid malignancy associated genes complicates the diagnosis of true myeloid malignancies, as well as raising an interest in assessing large prospective cohort to identify factors that promote the development of myeloid neoplasms from pre-existing CH.

Most CH studies have focused on mutations to define clonality [8,10], but clonality can also be defined by mCAs. Although the relationship between mCAs and underlying driver mutations is not straightforward, there is a notable overlap between some types of myeloid malignancy associated mCA and recurrently mutated genes in CH. For example, *TET2* is the second most recurrent CH gene in apparently healthy people, and *TET2* mutations are recurrently associated with LOH at 4q24 as a result of CNL or aUPD [251]. So, clonality can be confirmed by finding a putative driver somatic mutation that often represents the first hit, or by finding an acquired mCA event that often represents a second hit. However, the pathogenic value of any event can be assessed by testing its association with myeloid malignancies. Away from the malignant role of myeloid CH, granulocytes and monocytes play key roles in the inflammatory system across all tissues by accumulation at specific sites and by releasing inflammation mediators. It is possible that clonal granulocytes and monocytes have altered functions that impact or promote non-malignant conditions.

The UK Biobank provides a wealth of information to test different hypotheses relating to the causes of CH and its association with a wide range of benign phenotypes. Some key attributes of the UK Biobank are:

(i) Genotypic (SNP array) data is available for most of the 500,000 participants, with WES data from 50,000 at the time of initial analysis in 2019

(ii) Death registry data provides the date of death enabling survival analysis

(iii) The electronic questionnaire covered smoking and is thus suitable for assessing the relationship between CH and what is expected to be the most significant external factor

(iv) The data for individuals who developed health conditions are very detailed, covering about 10,000 different ICD-10 diagnoses

(v) A large number of blood and biochemical measures were performed at recruitment

In this Chapter, I present my results for detecting myeloid CH defined by (i) mCA and (ii) putative somatic mutations and their association with the development of myeloid malignancies in the UK Biobank. Next, I investigate genetic features and exposure factors as causes for the development of myeloid CH. Lastly, I sought to determine the role of myeloid CH in the pathogenesis of non-malignant diseases by assessing the relationship between myeloid CH and all-cause mortality, non-malignant diseases, blood features and other biomarkers.

3.3 Methods

3.3.1 Cohort structure

Participants from the UK Biobank were split into four phenotypic groups: myeloid malignancies, lymphoid malignancies, other cancers, and cancer-free based on the International Classification of Disease codes (ICD version 10) that were recorded by the national cancer registry (Data-Field 40006) and reason for admission to hospital (Data-Fields 41202, 41204, and 41270). ICD-10 codes used to define myeloid and lymphoid malignancies are listed in Table 3-1. Other cancers were defined by any other ICD-10 codes that were not used for haematological malignancies and were prefixed with a C or D0 to D48. The ICD-10 coding system was introduced into the national cancer registry in 1995 and to the hospital admissions in 1996. The cancer registry data was accessed as of 31st July 2018, and other clinical and phenotype data was accessed as of August 2019 (most recent record February 2018), thus providing data for a median of 9.1 years after recruitment and blood sampling, and a median age 58 at recruitment time. The four phenotypic groups were defined by events that occurred at any time between 1995 and 2018, i.e. at this stage of the analysis they included past, present and future malignancies with respect to the timing of the blood sample. Stricter definitions that focus on criteria for predicting the development of myeloid malignancies will be introduced later.

Table 3-1: ICD-10 used to define haematological malignancies

Group	ICD-10 Code
***(i)* Myeloid malignancies**	C92.0:Acute myeloid leukaemia, C92.1:Chronic myeloid leukaemia, C92.3:Myeloid sarcoma , C92.4:Acute promyelocytic leukaemia, C92.5:Acute myelomonocytic leukaemia, C92.7:Other myeloid leukaemia, C92.9:Myeloid leukaemia, unspecified, C93.0:Acute monocytic leukaemia, C93.1:Chronic monocytic leukaemia, C94.0:Acute erythraemia and, C94.4:Acute panmyelosis , C94.6:Myelodysplastic and myeloproliferative, C96.2:Malignant mast cell, D45:Polycythaemia vera , D46.0:Refractory anaemia without sideroblasts, so stated, D46.1:Refractory anaemia with sideroblasts, D46.2:Refractory anaemia with excess of blasts, D46.4:Refractory anaemia, unspecified, D46.7:Other myelodysplastic syndromes, D46.9:Myelodysplastic syndrome, unspecified, D47.0 :Histiocytic and mast cell tumours of uncertain and unknown behaviour, D47.1:Chronic myeloproliferative disease, D47.3:Essential (haemorrhagic) thrombocythaemia
***(ii)* Lymphoid malignancies**	C77.0 :Lymph nodes of head, face and neck, C77.1 :Intrathoracic lymph nodes, C77.2 :Intra-abdominal lymph nodes, C77.3 :Axillary and upper limb lymph nodes, C77.4 :Inguinal and lower limb lymph nodes, C77.5 :Intrapelvic lymph nodes, C77.8 :Lymph nodes of multiple regions, C77.9 :Lymph node, unspecified, C81.0 :Lymphocytic predominance, C81.1 :Nodular sclerosis, C81.2 :Mixed cellularity, C81.3 :Lymphocytic depletion, C81.7 :Other Hodgkin's disease, C81.9 :Hodgkin's disease, unspecified, C82.0 :Small cleaved cell, follicular, C82.1 :Mixed small cleaved and large cell, follicular, C82.2 :Large cell, follicular, C82.7 :Other types of follicular non-Hodgkin's lymphoma, C82.9 :Follicular non-Hodgkin's lymphoma, unspecified, C83.0 :Small cell (diffuse), C83.1 :Small cleaved cell (diffuse), C83.2 :Mixed small and large cell (diffuse), C83.3 :Large cell (diffuse), C83.4 :Immunoblastic (diffuse), C83.5 :Lymphoblastic (diffuse), C83.7 :Burkitt's tumour, C83.8 :Other types of diffuse non-Hodgkin's lymphoma, C83.9 :Diffuse non-Hodgkin's lymphoma, unspecified, C84.0 :Mycosis fungoides, C84.1 :Sezary's disease, C84.2 :T-zone lymphoma, C84.3 :Lymphoepithelioid lymphoma, C84.4 :Peripheral T-cell lymphoma, C84.5 :Other and unspecified T-cell lymphomas, C85.0 :Lymphosarcoma, C85.1 :B-cell lymphoma, unspecified, C85.7 :Other

	specified types of non-Hodgkin's lymphoma, C85.9 :Non-Hodgkin's lymphoma, unspecified type, C86.2 :Enteropathy-type (intestinal) T-cell lymphoma, C88.0 :Waldenstrom's macroglobulinaemia, C88.4 :Extranodal marginal zone B-cell lymphoma of mucosa-associated lymphoid tissue [MALT-lymphoma], C88.9 :Malignant immunoproliferative disease, unspecified, C90.0 :Multiple myeloma, C90.1 :Plasma cell leukaemia, C90.2 :Plasmacytoma, extramedullary, C90.3 :Solitary plasmacytoma, C91.0 :Acute lymphoblastic leukaemia, C91.1 :Chronic lymphocytic leukaemia, C91.3 :Prolymphocytic leukaemia, C91.4 :Hairy-cell leukaemia, C91.5 :Adult T-cell leukaemia, C91.9 :Lymphoid leukaemia, unspecified, C95.7 :Other leukaemia of unspecified cell type, C95.9 :Leukaemia, unspecified, C96.1 :Malignant histiocytosis, C96.3 :True histiocytic lymphoma, C96.8 :Histiocytic sarcoma, D47.2 :Monoclonal gammopathy
(iii) ICD-10 codes considered myeloid if accompanied with one of the codes in table (i) and no codes from table (ii), otherwise included under 'lymphoid'	C95.0 :Acute leukaemia of unspecified cell type, C95.1 :Chronic leukaemia of unspecified cell type, C96.7 :Other specified malignant neoplasms of lymphoid, haematopoietic and related tissue, C96.9 :Malignant neoplasms of lymphoid, haematopoietic and related tissue, unspecified, D47.7 :Other specified neoplasms of uncertain or unknown behaviour of lymphoid, haematopoietic and related tissue, D47.9 :Neoplasm of uncertain or unknown behaviour of lymphoid, haematopoietic and related tissue, unspecified

3.3.2 Calling mosaic chromosomal alterations from SNP array data

The UK Biobank provides comprehensive project level files for B-allele frequency (BAF: the ratio of intensity values for the A and B allele for each SNP in a single sample) and $\log_2$ R ratio (LRR: the logarithm value to the base 2 of the ratio of the observed intensity to the expected intensity for the diploid genome) for each SNP that passed QC (as detailed in Chapter 2). Raw input files were generated for each sample. Regions of allelic imbalance (AI) were then detected in in all participants for whom the array data passed QC (n=486,941), including X-chromosome imbalances for female participants (n=264,083) using BAF segmentation and the recommended parameters for Affymetrix array data [207]: minimum 4 SNPs, detection threshold mirrored BAF (mBAF) ≥0.56 (ΔBAF ≥ 12%), SNPs with mBAF > 0.9 were removed and triplet filter threshold was 0.6 (described in Chapter 2).

Next, a custom script was used for filtering, and merging events, removing likely constitutional events, and generating an empirical score for each event. First, bedtools was used to merge AI regions with a minimum density of 1 SNP per 20Kb that were separated by less than 2Mb [252], and removed merged events that cover <2Mb. Next, constitutional copy number gains were removed based on the following criteria which are similar to those used by other groups [253]. Constitutional CNG has a theoretical mBAF = 0.66. To exclude these non-informative events, large events (>10Mb) were removed if they had LRR > 0.35 or LRR > 0.2 and mBAF > 0.66. Small events (< 10Mb) were removed if LRR > 0.2 or LRR > 0.1 and mBAF >0.6. These thresholds were used by previous studies involving the UK Biobank SNP array data [253]. Next, merged AI regions were scored based on the product of (i) number of informative SNPs (ii) heterozygosity rate in the targeted region and (iii) coverage of AI regions for the merged event. Events that scored over an empirically defined threshold (≥9; described in detail below) were defined as mosaic chromosomal abnormalities (mCA), which were further broken down into CNL, CNG or aUPD using static LRR cut-offs (CNG > 0.07, CNL < -0.07). The parameters applied are estimated to identify clonal fractions larger than 0.1, 0.2, and 0.27 for aUPD, CNL, and CNG, respectively [207]. Since mCA may be derived from myeloid or lymphoid cells, we correlated mCA with clinical phenotype to specifically define myeloid mCA (detailed in results).

3.3.3 Identification of putative somatic mutations in WES data

The gVCF files from the UK Biobank were converted to VCF format and filtered to remove variants with low read depth (DP; <7 for SNVs, <10 for indels). SAMtools/Bcftools was used to merge the separate files into one multi-sample VCF, split multi-allelic positions into separate variants and to normalise the location of *indel*s using their left most position [254]. The multi-sample VCF was annotated in relation to genes (RefSeq), public databases of normal variation (1000 genome, https://www.internationalgenome.org/; ESP6500, https://evs.gs.washington.edu/EVS; GnomAD, https://gnomad.broadinstitute.org), and variant/protein pathogenicity scores using Annovar [201]. Putative somatic mutations (regardless of VAF) were identified in six genes known to be associated with myeloid neoplasia that were exonic, had an alternate allele frequency ≤1% in public databases of common variation (1000 Genome, ESP6500, GnomAD) and were either loss of function (LOF) mutations (*TET2*, *DNMT3A*, *ASXL1*, *PPM1D*) or known somatic mutations (*DNMT3A* R882, *JAK2* V617F, *SRSF2* P95). The data workflow is presented in Figure 3-1.

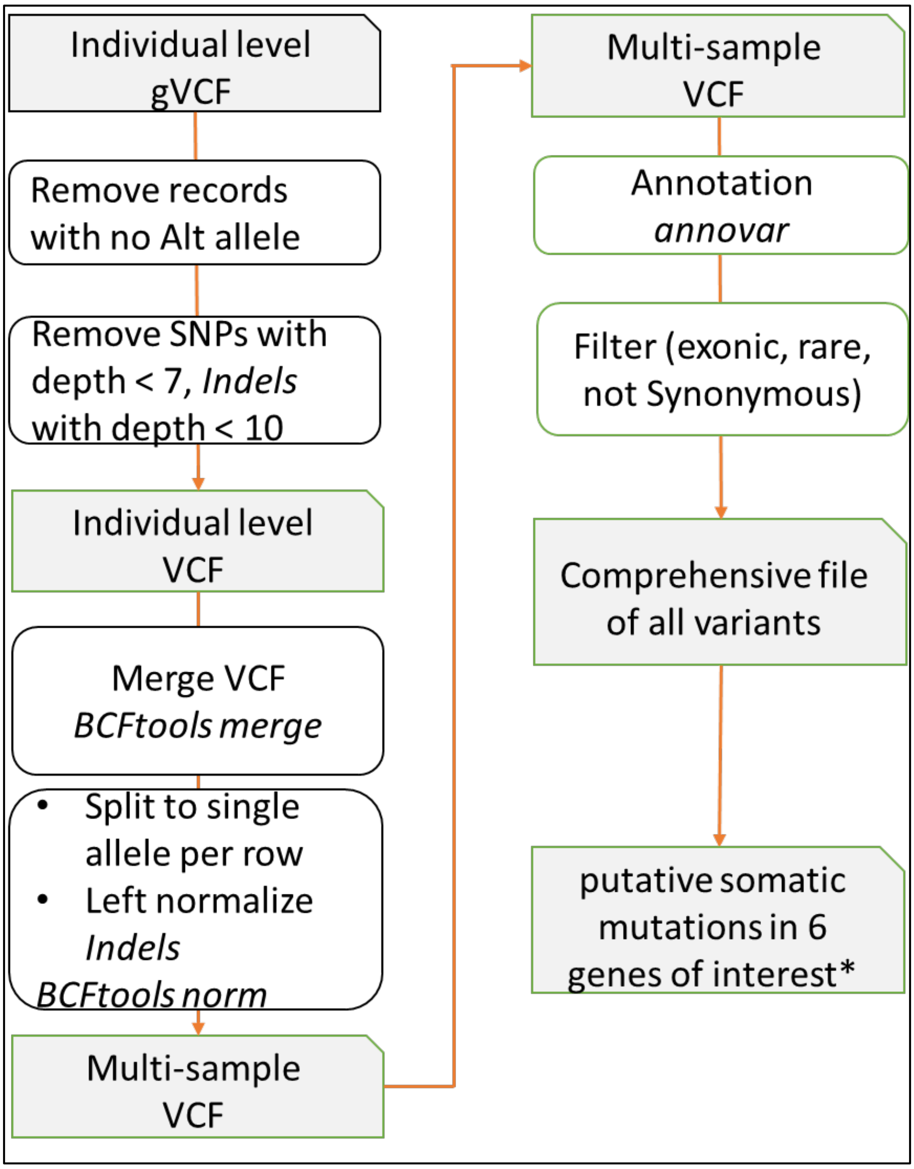

Individual level gVCF
Remove records with no Alt allele
Remove SNPs with depth < 7, *Indels* with depth < 10
Individual level VCF
Merge VCF *BCFtools merge*
• Split to single allele per row
• Left normalize *Indels*
BCFtools norm
Multi-sample VCF
Multi-sample VCF
Annotation *annovar*
Filter (exonic, rare, not Synonymous)
Comprehensive file of all variants
putative somatic mutations in 6 genes of interest*

Figure 3-1: Data processing of WES variant calls from the UK Biobank to identify putative somatic mutations

*The selected variants are disruptive mutations in *DNMT3A*, *TET2*, *ASXL1* and *PPM1D* plus 3 missense oncogenic variants *JAK2* V617F, *SRSF2* P95 and *DNMT3A* R882.

To validate the association between CH and smoking detected in 49,956 participants, I used the same pipeline to identify driver mutations in 150,685 newly released exomes from the UK Biobank. In addition to the previous filters, LOF mutations were considered if inferred as somatic by failing the hypothesis that the alternative allele is normally distributed with a mean of 0.45 and a false positive rate of P = 0.05 using a binomial test.

3.3.4 The association of common variants with clonality

Samples with SNP array data were split into cases (n = 1,166) and controls (n = 30,892). Cases were defined by the presence of one or more features associated with myeloid CH; either mCA in 15 genomic regions which are associated with myeloid disorders in this study, or at least one putative somatic mutation in six driver genes associated with myeloid disorders (*JAK2* V617F, *SRSF2* P95, *DNMT3A* R882 or frameshift/stopgain mutations in *DNMT3A*, *TET2*, *ASXL1* or *PPM1D)*. Controls were defined as samples without mCA, without likely somatic mutations in the genes of interest (including nonsynonymous variants) and without evidence of any haematological malignancy during the study period. A total of 265,112 common SNPs (MAF ≥0.1) without deviation from HWE (P>0.001) were assessed for association with myeloid CH using allelic chi square tests to compare allele frequencies between cases and controls. Association tests were performed using Plink V1.9. [255]. These results were visualised using the qqman, qqnorm and qqplot procedures in R to generate a Manhattan plot and quantile-quantile plot [256]. In regions with multiple SNPs reaching genome-wide significance, conditional logistic regression was used to determine the number of independent signals. All SNPs with P<5x10^{-8} and within 500kb of the index SNP were added to the regression model in order of significance. Linkage disequilibrium between SNPs was calculated by LDassoc; an interactive tool to visualise association P-value results and linkage disequilibrium patterns for a genomic region of interest [257].

3.3.5 Phenotype selection:

Cigarette smoking data were collected by electronic questionnaire at the first visit to the assessment centre and were captured in three data fields: (i) current smoking status “Data-Field 1239, question: Are you a current smoker?”; (ii) For participants who were not current smokers “Data-Field 1249”, question: Are you a previous smoker?”, and (iii) smoking status “Data-Field 20116” which combines data fields 1239 and 1249.

Phenotypic diagnoses were derived from primary/main diagnosis encoded in hospital inpatient records “Data-Field 41202”. These phenotypes were encoded using the international classification of disease, version 10 (ICD-10) and covered 7,920 different phenotypes with at least one incident across all participants. To minimise false positives due to small sample size, our investigation was limited to

395 phenotypes with a frequency >0.1% in the UK Biobank cohort. Irrelevant phenotypes, with ICD-10 prefixes “O”,”P”,”Q”, ”S” to “Z”, and any cancer related phenotypes, with ICD-10 prefixes “C” and “D00” to “D48”, were also excluded. Participants with any evidence of haematological malignancy were excluded from the analysis based on diagnoses from the national cancer registry (Data-Field 40006) and inpatient hospital records (Data-Fields 41202, 41204, and 41270). Blood counts and biochemistry were measured or estimated in the UK Biobank and encoded as continuous variables (Table 3-2, and Table 3-3).

Table 3-2: 29 blood counts measured, calculated or derived in the UK Biobank

	Calculated	**Calculated**	**Derived**
Red Blood Cells	Count (x 10^9 cells/L)	Haematocrit (%)	Mean Corpuscular Volume (fL) Distribution Width (%)
White Blood Cell	Count (x 10^9 cells/L)		
Haemoglobin	Concentration (g/dL)	Mean Corpuscular haemoglobin (pg) Mean Corpuscular (erythrocyte) haemoglobin concentration	
Platelet	Count (x 10^9 cells/L)	Platelet crit (%)	Mean platelet volume (fL) Platelet Distribution Width (%)
Lymphocyte	Percent (%)	Count (x 10^9 cells/L)	
Monocytes	Percent (%)	Count (x 10^9 cells/L)	
Neutrophil	Percent (%)	Count (x 10^9 cells/L)	
Eosinophil	Percent (%)	Count (x 10^9 cells/L)	
Basophil	Percent (%)	Count (x 10^9 cells/L)	
Reticulocyte	Percent (%)	Count (x 10^9 cells/L) High Light scatter Reticulocytes (x 10^9 cells/L) Immature Reticulocyte Fraction (Ratio) Mean Reticulocyte Volume (fL)	High Light scatter Reticulocytes (%) Mean Sphered Cell Volume (fL)

Table 3-3: 29 blood biomarkers measured in the UK Biobank

Biomarker group	Serum Assay
Liver	Albumin, Alkaline Phosphatase, Alanine Aminotransferase, Aspartate Aminotransferase, High Sensitivity C-Reactive Protein, Total Bilirubin, Direct Bilirubin, Gamma-Glutamyltransferase
Kidney	Uric acid, Urea, Cystatin-C, and Creatinine
Lipid	Cholesterol, Triglyceride, High Density Lipoprotein, Low Density Lipoprotein, Apolipoprotein A1, Apolipoprotein B, High Density Lipoprotein, Low Density Lipoprotein, Lipoprotein (a)
Sex	Testosterone, Oestradiol, and Sex Hormone Binding Protein
Bone	Calcium, Phosphate, Vitamin D
Others	Total Protein, Glucose, Insulin-like Growth Factor-1, Rheumatoid Factor

3.3.6 The association of myeloid CH with smoking, clinical phenotype, blood traits and biochemistry

The PHEnome Scan ANalysis Tool (PHESANT) was used to test the selected phenotypes from the UK Biobank for association [249] with myeloid CH defined by myeloid mCA and/or somatic mutations. PHESANT assigned the appropriate test for each phenotype; they include either ordinal logistic regression (polr R function for current smoking status and previous smoking status which are ordered categorical variables), multinomial logistic regression (multinom R function for combined smoking status which has three possible outcomes; never, previous and current), logistic regression (glm R function for binary clinical phenotypes; n = 395), or linear regression (lm R function for blood features n = 29 or biochemical markers n = 30). All regressions included covariates for age and sex with the addition of smoking status for the analysis of clinical phenotypes and blood features. Where appropriate, inverse normal transformation was applied to counteract departures from normality, and P-values were corrected for multiple testing using False Discovery Rate (FDR) method [258].

3.3.7 Survival analyses

To test the association between myeloid CH and either all-cause mortality, myocardial infarction (MI) or stroke, the ‘survival’ package [239] in R was used to perform Cox regression analyses with correction for age at study entry, sex and smoking status. Follow-up times were calculated using the ‘lubridate’ [240] package to determine the duration between study entry to last registration in either the date of death (Data-field 4000), date of MI (Data-field 42000) or date of stroke (Data-field 42006). Participants that had an event before the date of entry were excluded (left-truncated).

3.3.8 Statistical analysis

***Association of mCA categories and somatic driver mutations with phenotypic group*.** The frequency of mCA events, each of its subcategories (aUPD, CNG or CNL) and somatic driver mutations were tested for association with either myeloid, lymphoid, or other cancers compared to cancer free controls using Fisher’s exact tests in SPSS (Version 25). The average number of mCA events per sample in either myeloid, lymphoid, or other cancers were compared against cancer free controls using Mann-Whitney U tests [259].

***Association of specific mCAs with haematological phenotype*.** Autosomal mCAs were stratified by type (aUPD, CNL, CNG), chromosome arm (*p* or *q*), and position (telomeric or interstitial). Each type with at least one observation was tested for association with a haematological phenotype (myeloid or lymphoid) in comparison to cancer free controls using Fisher’s exact tests in SPSS (version 25). A total of 416 specific mCAs were tested after selecting mCA types with at least one observation. Interstitial events that were associated with a myeloid phenotype and not previously recognised as a recurrent abnormality were manually reviewed against known associations with myeloid malignancies (Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer; https://mitelmandatabase.isb-cgc.org/search_menu). After review, events associated with a myeloid phenotype were grouped and hereafter referred to as myeloid mCAs.

***Regression of mCA against age*.** The relationship between mCA and age was tested using multivariable logistic regression in SPSS where mCA status was treated as the dependent, age as a predictor and including gender as an independent covariate. This analysis was repeated for each subcategory of mCA (aUPD, CNG or CNL), myeloid mCAs and myeloid somatic mutations. The effect sizes were reported as odds ratios (OR) with 95% confidence intervals (CI).

Meta-analysis of the relationship between driver mutations and smoking. To validate the association of smoking with myeloid CH defined by driver mutations, the new release of WES data (n = 150,685) was screened for myeloid CH and tested for association using PHESANT as previously described. Results from both WES cohorts were combined using a fixed effect inverse variance weighted meta-analysis using STATA version 16 (StataCorp LLC, College Station, TX). Cochran's Q test was used to measure heterogeneity and results were presented as forest plot.

3.4 Results

3.4.1 Data summary

The phenotypic breakdown of the UK Biobank cohort is summarised in (Table 3-4), along with cases for whom SNP array and WES data were available. The classification shows the expected excess of myeloid malignancies in males (1.3 : 1) [250]

As described in Chapter 2, the SNP array data were provided by the UK Biobank for 488,377 SNPs following their own QC. I excluded 1,436 cases due to one or more of poor genotyping quality (missingness above 5%; n = 229), outlying levels of heterozygosity that could not be explained by admixture or consanguinity (principal component-adjusted heterozygosity above the mean 0.1903, n = 744), gender mismatch (n = 373), withdrawal of consent (n = 85), or absence of phenotypic data (n=10).

WES for 49,996 individuals were provided by the UK Biobank. I excluded 40 samples for QC reasons that were published by Regeneron [208] who performed the sequencing and initial QC. The reasons for exclusion were unmatched sex (n=15), high rates of heterozygosity/contamination (D-stat > 0.4) (n = 7), low sequence coverage (less than 85% of targeted bases achieving 20X coverage) (n=1), genetically duplicated samples (n = 14), and discordance between WES and SNP array (n = 9).

Table 3-4: Summary of the UK Biobank cohort

	Males n (%)	Females n (%)	Total
Phenotypic data [1]	229,129 (46)	273,395 (54)	502,524
Myeloid disorders [2]	1,157 (57)	873 (43)	2,030
Lymphoid disorders [3]	5,747 (44)	7,390 (56)	13,137
Other cancers	49,435 (41)	71,420 (59)	120,855
Cancer-free	172,790 (47)	193,712 (53)	366,502
SNP array data [4]	222,858 (46)	264,083 (54)	486,941
Myeloid disorders	1,097 (57)	816 (43)	1,913
Lymphoid disorders	5,566 (44)	6,980 (56)	12,546
Other cancers	48,101 (41)	68,820 (59)	116,921
Cancer-free	168,094 (47)	187,467 (53)	355,561
WES data [5]	22,714 (45)	27,242 (55)	49,956
Myeloid disorders	97 (53)	85 (47)	182
Lymphoid disorders	473 (46)	550 (54)	1,023
Other cancers	4,972 (41)	7,265 (59)	12,237
Cancer-free	17,172 (47)	19,342 (53)	36,514

1) Includes cases who had the specified disorder at any time during the study period.
2) Of 2030 participants with a myeloid disorder, 315 were also diagnosed with another non-myeloid haematological disorder during the study period.
3) 34 cases with unspecified haematological malignancy were included in the lymphoid group
4) Data available from 488,377 cases of which 1436 were excluded following QC.
5) Data available from 49,996 cases of which 40 were excluded following QC.

3.4.2 Mosaic chromosomal alterations in the UK Biobank

Genome wide SNP array data was analysed to identify autosomal regions of AI in all participants for whom the array data passed QC (n = 486,941), including X-chromosome imbalances for female participants (n = 264,083). The default parameters of the published tool “BAF Segmentation” [207] were applied, that define a region of AI by 4 consecutive SNPs of mBAF between 0.56 to 0.9. Initially, 6,546,768 AI signals were identified in 94% of the processed samples, which is vastly higher than the expected mCA prevalence of around 1% samples from published studies of other population cohorts s. Visualising a selection of the BAF plots suggested two major problems:

(i) **Likely germline events,** mainly small interstitial regions and constitutional CNG

(ii) **Artefacts,** including regions of poor coverage on the array, regions of low heterozygosity, or poor genotyping particularly near centromeres

A series of filters were applied to minimise signals related to poor marker coverage and the false discovery of constitutional copy-number variants, as detailed in methods and summarised in Table 3-5.

Table 3-5: The number of the identified allelic imbalance regions at each processing step

Filter	**Regions of AI (n)**	**Participants (%)**
The UK Biobank SNP array cohort		486,941 (100)
BAF Segmentation raw results	6,546,768	461,460 (94.8)
1 SNP/20,000 bp coverage	6,444,606	459,847 (94.5)
Merge with maximum 2Mb separation	5,838,835	459,847 (94.5)
≥2Mb coverage size	239,089	95,617 (19.6)
Remove constitutional CNG[1]	234,772	92,884 (19)
Score[2] ≥ 9	8,203	5,040 (1)

[1] remove large events (>10Mb) if they had LRR > 0.35 or LRR > 0.2 and mBAF > 0.66. Small events (< 10Mb) were removed if LRR > 0.2 or LRR > 0.1 and mBAF >0.6

[2] empirical score calculated as following (number of informative SNPs x mean Het rate x Coverage rate)

To further decrease the false discovery rate, the merged AI regions were scored according to the product of three parameters that are correlated with the calling accuracy, as indicated in Equation 2. Examples of plots with different empirical scores are provided in Figure 3-2.

Equation 2: The score assigned to each allelic imbalance event larger than or equal 2Mb size

$$Score = Number\ of\ informative\ SNPs\ x\ \ mean\ Het\ rate\ x\ Coverage\ rate$$

1) number of informative SNPs is the total number of SNPs that defines the event with mBAF between 0.56 and 0.9.

2) mean het rate is the mean heterozygosity rate at the targeted region of the event.

3) coverage rate is the size ratio (the sum of individual AI events / total size of the new covered event).

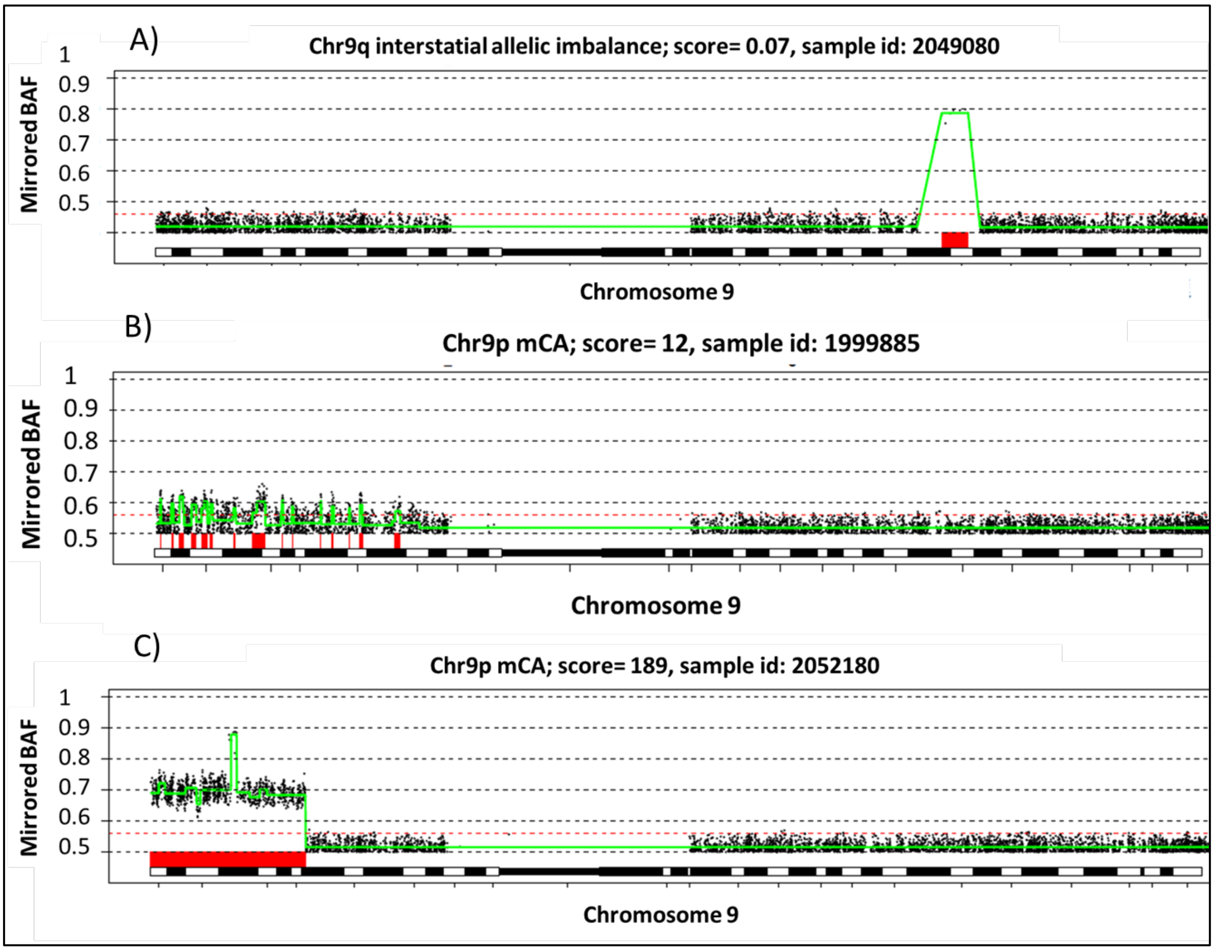


Figure 3-2: Plot of mirrored BAF on chromosome 9 in three samples in the UK Biobank.

Plot A) shows an interstitial AI event (score = 0.7; calculated as 7 SNPs x 0.1 het-rate x 1 coverage). This event was filtered out as its score was < 9. Plot B) shows a low level 9p mCA event (score =12.9; calculated 170 SNPs x 0.2 het-rate x 0.3 coverage). Plot C) shows a very clear 9p mCA event (score=189; calculated 1206 SNPs x 0.16 het-rate x 0.98 coverage).

Score cut-off selection: three methods were used to determine the empirical score threshold that was used to remove false positive AI regions; these 3 methods were used to select a score larger than or equal to 9.

The first method examined the frequency of AI events across chromosome 9 using non overlapping windows of 100Kb and a range of filtering scores (0 to 17). Our hypothesis was that the most frequent region of AI should include the *JAK2* gene due to the selective advantage of chromosome 9p aUPD in the presence of a somatically acquired *JAK2* V617F mutation [28], and the fact that there are no other known targets of 9p aUPD. At a filtering score of 9, the pattern of AI frequency across the chromosome stabilised and the most frequent region included *JAK2,* as hypothesised (Figure 3-3).

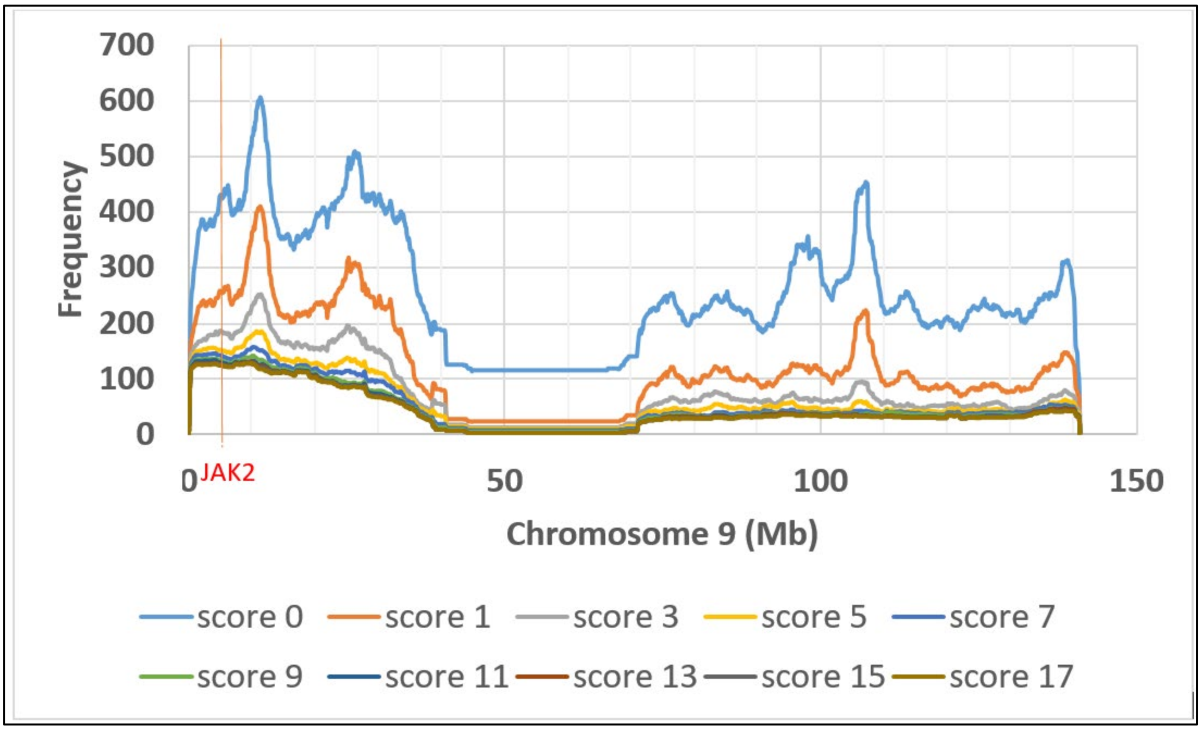


Figure 3-3: The frequency of mCA across chromosome 9 using different scores.

The plot was generated by calculating the number of samples with mosaicism within a sliding window of 100kb under different thresholds of the score filter. At a filtering score of 9 the pattern of AI frequency across the chromosome stabilised with the most frequent region including *JAK2*; an increase in score stringency provided no further improvement.

The second method examined the frequency of AI in the entire genome under a range of filtering scores (0 to 30) and six age categories that were defined by 5-year intervals from age 40 to 70 years old. Here the frequency of large AI regions (≥2Mb) was expected to be close to 1% over all participants given previous estimates of 0.89% [7] and 0.73% [260]. The frequency of large AI regions was also expected to increase with age [6]. At a filtering score of 9, the AI incidence aligned with those expected for all participants (1%) and increased with age to a frequency of 1.29% in participants aged 66 to 70 years old (Figure 3-4).

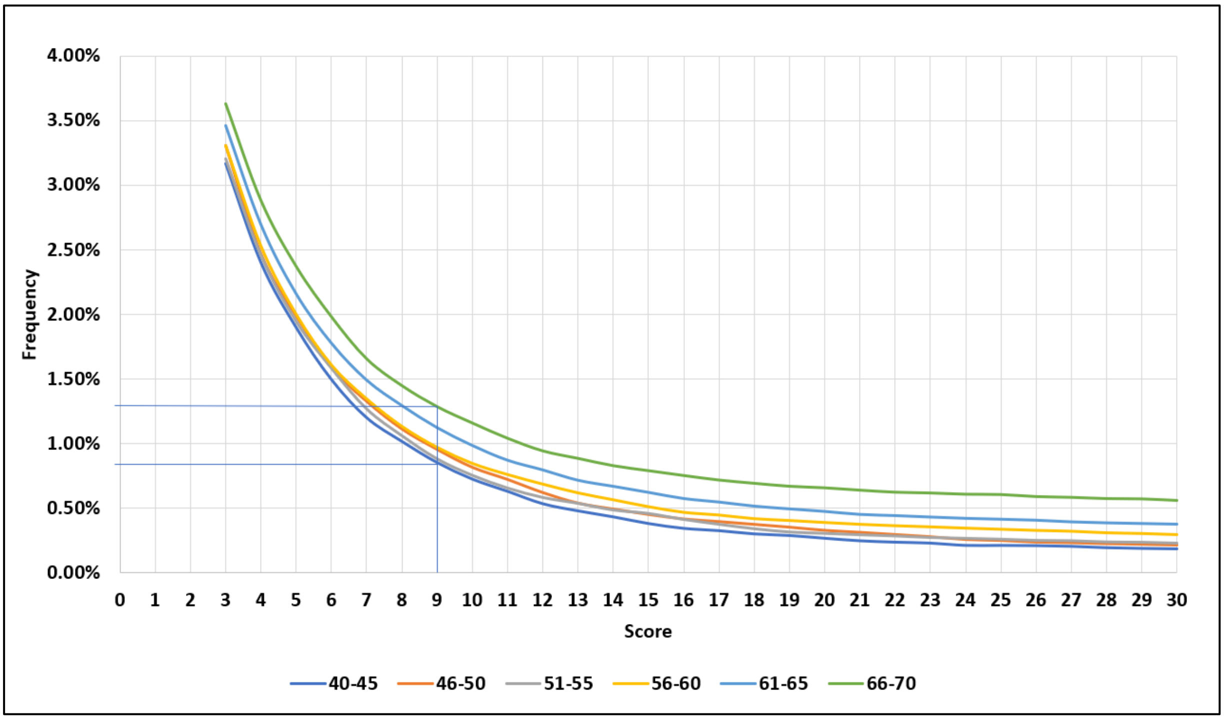


Figure 3-4: Age relationship of mCA using different scores.

Six age categories defined by 5 year intervals from age 40 to 70 years old were compared with a range of filtering scores. At a filtering score of 9, shown by the interpolation lines, the AI incidence aligned with that expected for all participants (1%) and increased with age to a frequency of 1.29% in participants aged 66 to 70 years old.

The third method was based on the relationship between 9p AI and *JAK2* V617F (n=40) in samples with WES data. In theory the great majority of samples with 9p AI should also harbour *JAK2* V617F. Assuming true positive 9p AI were also positive for *JAK2* V617F, the sensitivity and specificity were calculated under a range of AI filtering scores and shown to increase from 35% without filtering (AI score equal to zero) to a plateau of 86% at AI filtering scores of 8 and above (Table 3-6). Combining these three lines of evidence, an AI filtering score of ≥9 was chosen as the empirical threshold for selecting merged AI regions that were at least 2Mb in size. These regions were defined as mCA. The LRR was then used to classify each mCA region as either CNL (LRR ≤ -0.07), CNG (LRR ≥ 0.07) or aUPD (LRR ≥ -0.07 and ≤ 0.07) (Figure 3-5). Events were further classified by their genomic location as either telomeric if located within 2Mb of the p telomere (excluding acrocentric chromosomes) or 2Mb of the q telomere. Other mCA events were classified as interstitial (Figure 3-6).

Table 3-6: Specificity for calling mCA under different thresholds of the empirical score.

Empirical score	Number of chr9p mCA called that cover *JAK2*	True chr9p mCA (positive for *JAK2* V617F)	Specificity (%)
0	48	17	35.42
1	29	13	44.83
2	22	12	54.55
3	18	12	66.67
4	17	12	70.59
5	16	12	75
6	16	12	75
7	15	12	80
8	14	12	85.71
9	14	12	85.71
10	14	12	85.71
11	14	12	85.71
12	14	12	85.71

In total, the method identified 8,203 mCA >2Mb in size in 5,040 participants (1% of 486,941 analysed samples) which broke down into aUPD (n = 4,224), CNG (n = 659) or CNL (n = 3,320) as shown in Supplementary Table 3-1. The myeloid disorders group (n = 1,913) had the largest incidence of mCA with 11% of samples (n = 210) affected. Of these, more than 75% (n = 158) were affected by aUPD, a highly significant association (OR = 16.39; P = 8.78x10^{-124}; Fisher's exact test) that exceeded the relationship between all other mCA categories and phenotypes (Table 3-7). The frequency of mCA in lymphoid disorders was much lower than the myeloid group (363/12546; 2.9%) but the average number of events per positive sample (2.1) was significantly higher compared to cancer free controls (1.6, P = 4.1x10^{-6}) or myeloid samples (1.5, P = 0.015) according to the Mann-Whitney U tests (Table 3-7).

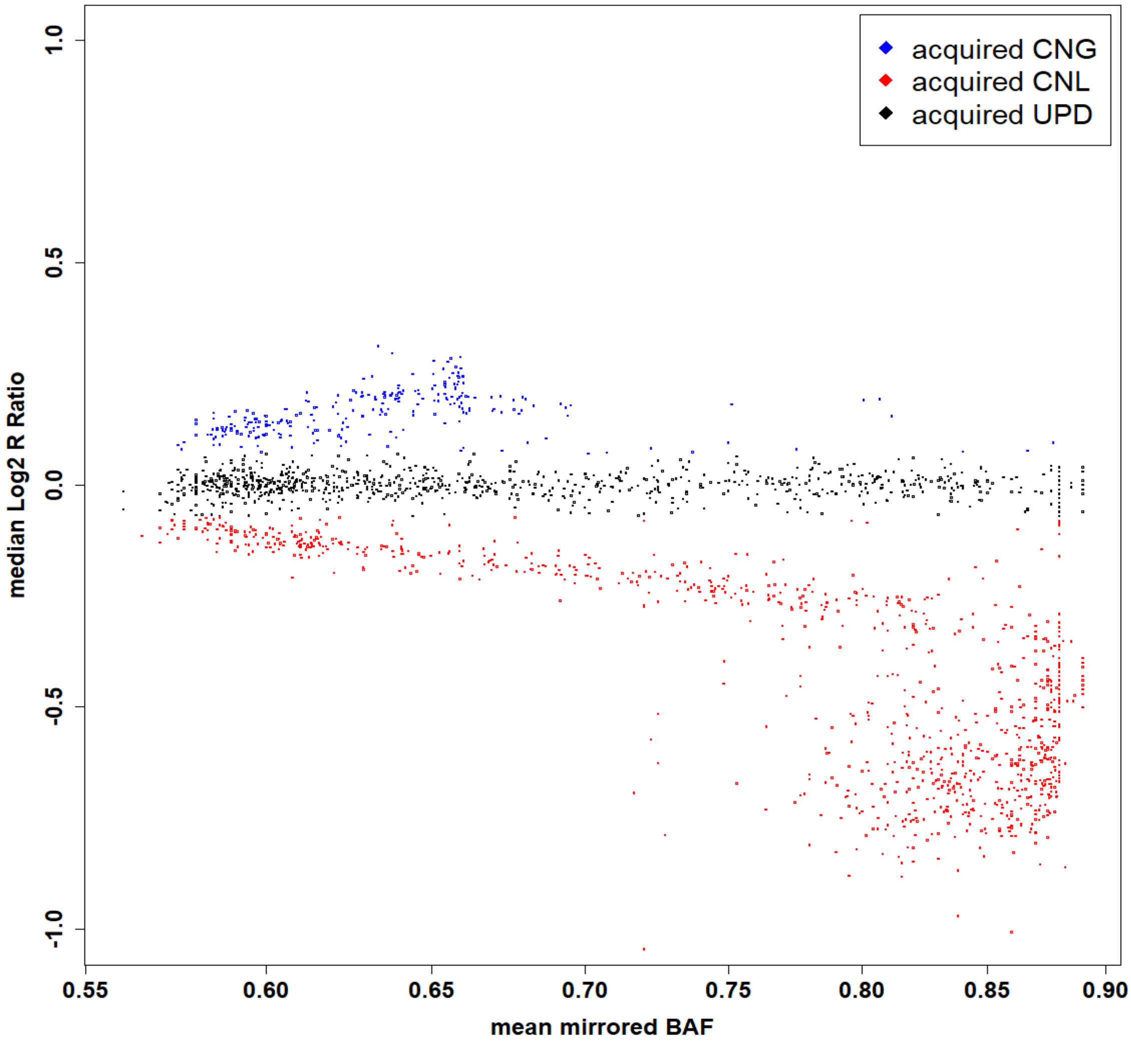


Figure 3-5: The relationship between mBAF and median LRR for telomeric mCA.

Table 3-7: Summary of mCA identified across the cohort

Group	SNP array samples	Total mCA events (per sample)	P*	Samples with at least one event											
				mCA			aUPD			CNG			CNL		
				n (%)	OR	P_{FDR}	n	OR	P_{FDR}	n	OR	P_{FDR}	n	OR	P_{FDR}
Myeloid disorders	1913	316 (1.5)	4.10×10^{-6}	210 (11)	13.21	3.74×10^{-145}	158	16.39	8.78×10^{-124}	37	40.88	3.10×10^{-44}	34	4.61	2.18×10^{-12}
lymphoid disorders	12546	768 (2.1)	0.54	363 (2.9)	3.19	1.41×10^{-105}	146	2.15	1.77×10^{-19}	81	14.00	3.09×10^{-55}	194	4.01	1.22×10^{-83}
Other cancers	116921	1854 (1.6)	0.89	1185 (1)	1.10	4.00×10^{-3}	657	1.03	0.26	67	1.24	0.09	527	1.16	3.6×10^{-3}
Cancer free	355561	5269 (1.6)		3282 (0.9)			1938			165			1386		

The number of mCA identified in each phenotypic group out of the total number of samples with SNP array data passing QC. The number of events for each mCA subcategory are also shown: aUPD, CNG and CNL. The mean number of mCA events in participants with either myeloid, lymphoid, or other cancers were compared with cancer free controls using Mann Whitney U tests (P*). Fisher's exact tests were used to compare the number of events which were corrected for 12 tests using the false discovery rate (P_{FDR}).

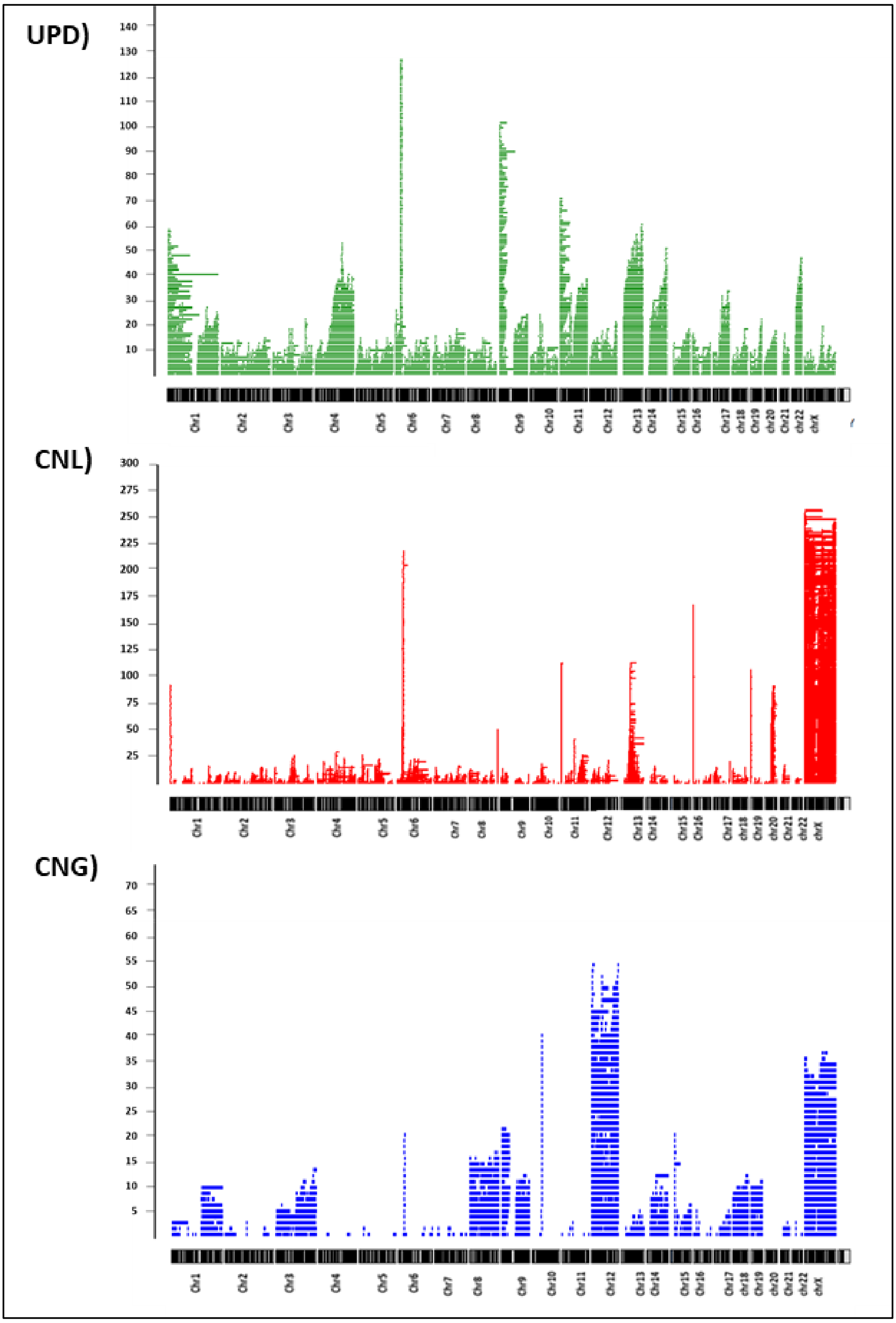


Figure 3-6: The chromosomal distribution of the identified mCA (n=8,203) in the UK Biobank

Relatively large mCA (>2Mb) with imperial score ≥ 9 were classified according to LLR into aUPD (LRR ≥ -0.07 and ≤ 0.07, n = 4,224, green panel), CNL (LRR ≤ -0.07, n = 3,320, red panel), or CNG (LRR ≥ 0.07, n = 659, blue panel).

3.4.3 The association of mCA with age

The association between mCA and age in years was assessed using logistic regression where mCA status was the dependent variable, age was the predictor and adjusting for gender in the model. The risk of mCA increased with age, showing a positive association with 1.02 fold annual increase (OR = 1.017; 95% CI = 1.013 - 1.020; P = 1.80x10^{-19}, logistic regression test). The frequency of mCA ranged between 0.85% at 40-45 years to 1.29% at 66-70 years (Figure 3-7). The regression was repeated for each category of mCA and the association with age was significant for aUPD (OR = 1.018; CI = 1.014-1.023; P = 3.14x10^{-14}), CNG (OR = 1.036; CI = 1.021-1.053; P = 1.09x10^{-6}) and CNL (OR = 1.01; CI = 1.05-1.015; P =3.66x10^{-4}).

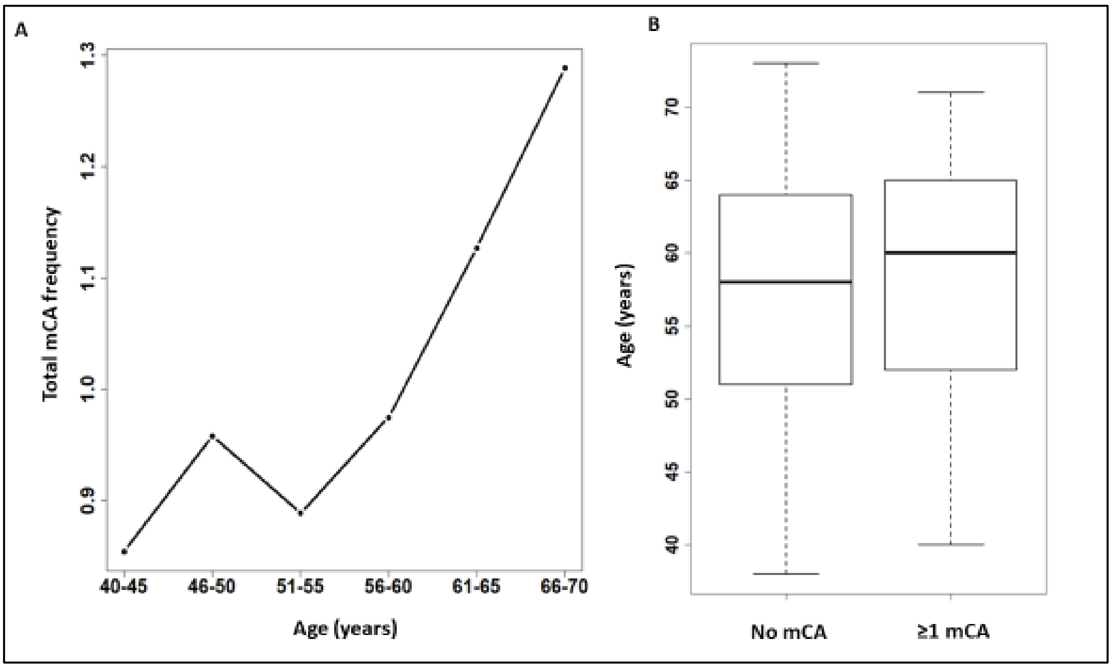


Figure 3-7: The relationship between mCA and age.

(A) Total mCA frequency across different age intervals. The risk of mCA was estimated to increase by 1.02 fold per year (P = 1.80x10^{-19}). (B) Box plot showing increased age in subjects with ≥1 mCA (median = 60 years; n=5,040) compared to those with no mCA (median = 58 years; n=481,901; P = 1.80x10^{-19}).

3.4.4 Myeloid malignancies-associated with mCA

Inspection of the mCAs shown on Figure 3-6. indicated several changes associated with both myeloid and lymphoid disorders. To examine the association with haematological malignancies in detail, I classified the autosomal chromosome alterations by type (aUPD, CNL, CNG), chromosome arm (p or q), and position (telomeric or interstitial). A total of 416 specific mCAs were tested after selecting mCA types with at least one observation in samples with myeloid or lymphoid disease.

As expected, distinct mCA were associated with myeloid and lymphoid disorders, with 9p aUPD most strongly associated with myeloid disorders (OR = 2858, $P_{FDR} = 6.28x10^{-191}$), and chr13q interstitial CNL most strongly associated with lymphoid disorders (OR = 23.16; $P_{FDR} = 1.24x10^{-63}$). For the downstream analysis, I selected all telomeric associated mCA with P ≤ 0.05 plus interstitial mCA of known cytogenetic relevance. This resulted in 17 abnormalities involving 15 chromosomal arms to define myeloid mCA: mCA for 16 telomeric abnormalities plus chr20q interstitial CNL (Table 3-8). A total of 25 abnormalities were associated with lymphoid disorders: 23 telomeric abnormalities and interstitial CNL targeting chr13q and chr11q (Table 3-9). Since the focus of my study is on myeloid clonality, I did not correlate the findings in lymphoid disorders with known cytogenetic aberrations.

Strikingly, the frequency of the myeloid associated mCA increased more sharply with age (OR = 1.1; 95% CI = 1.08 - 1.11; $P = 1.57x10^{-38}$, logistic regression test) compared to all mCA. It ranged between 0.04% at age 40 to 45 and 0.29% at age 66 to 70 (Figure 3-8). The frequency of the lymphoid associated mCA also increased with age (OR = 1.043; 95% CI = 1.033 - 1.053; $P = 1.75 \text{ x } 10^{-16}$, logistic regression test).

Table 3-8: Summary of mCA events significantly associated with myeloid disorders

Event	mCA type	Cancer free (n=355,561), No. positive†	Myeloid malignancies (n=1,614)*		
			No. positive	OR	P_{FDR}
chr9p	aUPD	7	86	2859	6.30 x10^{-191}
chr9p	CNG	1	15	3335	7.93 x10^{-33}
chr14q	aUPD	23	12	116	2.67 x10^{-18}
chr9q	CNG	1	8	1771	6.93 x10^{-17}
chr1p	aUPD	31	10	72	1.25 x10^{-13}
chr20iq	CNL	36	10	63	3.31 x10^{-13}
chr4q	aUPD	13	8	136	1.04 x10^{-12}
chr1q	CNG	1	4	833	4.26 x10^{-8}
chr8p	CNG	5	4	177	8.45 x10^{-7}
chr9p	CNL	1	3	662	5.82 x10^{-6}
chr8q	CNG	4	3	166	4.03 x10^{-5}
chr7q	aUPD	6	3	110	1.01 x10^{-4}
chr7q	CNL	0	2	-	2.31 x10^{-4}
chr17p	aUPD	4	2	110	2.85 x10^{-3}
chr19q	aUPD	12	2	37	0.01
chr11q	aUPD	24	2	18	0.04
chr22q	aUPD	27	2	16	0.05

†From a total of 355,561 cancer free sample

*From a total of 1,614 samples with a myeloid malignancy. 299/1913 myeloid cases were excluded from this analysis because they had both myeloid and lymphoid disorders

iq stands for interstitial events within the q arm

Table 3-9: Summary of mCA events significantly associated with lymphoid disorders

Event	mCA type	Cancer free (n=355,561)†	Lymphoid malignancies (n=12,546) *		
		No. positive	No. positive	OR	P_{FDR}
chr13iq	CNL	96	78	23	1.24x10^{-63}
chr12p	CNG	7	34	138	3.00x10^{-41}
chr12q	CNG	7	32	129	1.33x10^{-38}
chr13q	aUPD	25	25	28	7.63x10^{-22}
chr18q	CNG	1	9	255	1.80x10^{-11}
chr19q	CNG	0	8	-	5.02x10^{-11}
chr3q	CNG	1	8	226	4.11x10^{-10}
chr19p	CNG	0	7	-	1.30x10^{-09}
chr17p	CNL	1	7	198	9.54x10^{-09}
chr18p	CNG	1	6	170	2.00x10^{-07}
chr8p	CNL	2	6	85	7.47x10^{-07}
chr11iq	CNL	53	14	8	8.45x10^{-07}
chr11q	CNL	0	4	-	1.43x10^{-05}
chr3p	CNG	0	4	-	1.43x10^{-05}
chr13q	CNL	1	4	113	9.10x10^{-05}
chr1p	aUPD	31	8	7	4.71x10^{-04}
chr8q	CNG	4	4	28	8.84x10^{-04}
chr14q	CNL	1	3	85	1.50x10^{-03}
chr6q	CNL	1	3	85	1.50x10^{-03}
chr11q	aUPD	24	6	7	4.04x10^{-03}
chr9q	aUPD	15	5	9	4.04x10^{-03}
chr4q	aUPD	13	4	9	0.02
chr20q	aUPD	8	3	11	0.04
chr17q	CNG	2	2	28	0.05
chr7p	CNL	2	2	28	0.05

†From a total of 355,561 cancer free samples

*From a total of 12,546 lymphoid samples

iq stands for interstitial events within the q arm

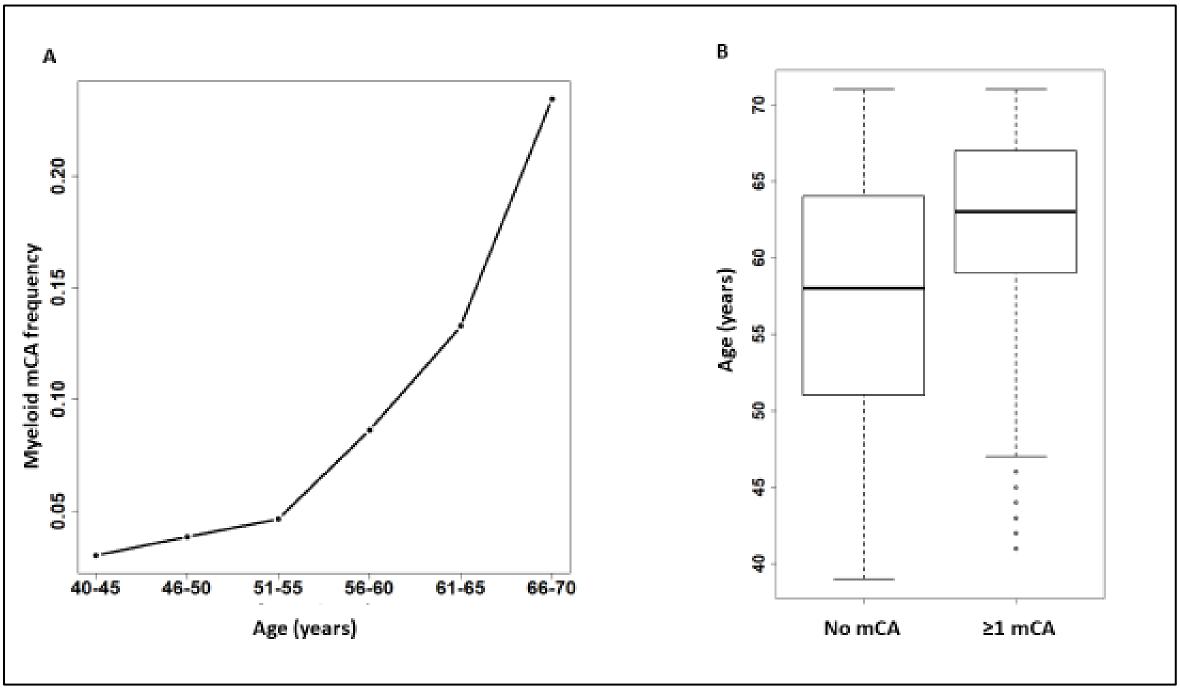


Figure 3-8: The relationship between myeloid mCA and age.

(A) Myeloid mCA frequency across different age intervals showing an annual 1.1-fold increase (P=1.57x10^{-38}). (B) box plot showing increased age in subjects with ≥1 myeloid mCA (median = 63 years; n=506) compared to those with no mCA (median = 58 years; n=481,901; P=1.57x10^{-38})

3.4.5 Clonality defined by somatic mutations in the UK Biobank

Although the WeCall pipeline (see Chapter 2) was not specifically designed to identify somatic mutations it did call variants with an allelic bias (lower number of reads supporting variant than expected for a germline variant). In my initial analysis I focused on identifying putative somatic mutations in 6 genes of known significance in myeloid malignancies (*TET2*, *DNMT3A*, *ASXL1*, *JAK2*, *SRSF2* and *PPM1D*) that are known to account for 95% of cases of CH in previous studies [9]. I focused on variants that had a high likelihood of being pathogenic driver mutations: with an alternate allele frequency of ≤1% in public databases of common variation (1000 genomes, ESP6500 and gnomAD) and were either loss of function mutations in *TET2*, *DNMT3A*, *ASXL1* or *PPM1D* or known somatic driver mutations in *DNMT3A* (R882), *JAK2* (V617F) or *SRSF2* (P95).

As summarised in Table 3-10 and detailed in Supplementary Table 3-2, I identified 721 candidate driver mutations in 678 subjects (1.4% of the 49,956 samples in the first release of WES data), with *DNMT3A* being the most commonly affected gene. Only 37 cases had more than one variant which had a higher frequency of myeloid disorders (11.8%) compared to participants with a single variant (5.8%). Of the 678 participants with CH defined by somatic mutations, 18 (*JAK2*, n=11; *DNMT3A*, n=4; *ASXL1*, *PPM1D*, *TET2*, n=1 of each) also had CH defined by mCA.

As expected, the prevalence of these putative somatic mutations was shown to be greatest in cases with myeloid disorders (22.5% versus 1.2% for cancer-free controls, P_{FDR} = 5.83x10^{-38}, OR = 23.7) compared to other groups (2.1% for lymphoid disorders versus cancer-free controls; P_{FDR} = 0.02; OR = 1.7) using Fisher's exact tests. Looking at individual genes, the only exception was *JAK2* V617F which was most commonly seen in myeloid disorders. There was a marginal increase in the prevalence of myeloid mutations in participants who had non-haematological cancers versus cancer-free controls (1.4% vs 1.2%; P_{FDR} = 0.04; OR = 1.2).

Table 3-10: Summary of putative somatic mutations by WES.

	Mutations		**Participants**				
	N	**VAF median (range)**	**Total**	**Myeloid n= 182**	**Lymphoid n=1,023**	**Other Cancer n=12,237**	**Cancer Free n=36,514**
***DNMT3A* LOF**	223	0.17 (0.07-0.50)	222	1	5	64	152
***DNMT3A* R882**	86	0.17 (0.11-0.40)	86	1	1	25	59
***TET2* LOF**	223	0.18 (0.06-0.68)	208	9	10	55	134
***ASXL1* LOF**	101	0.21 (0.08-0.49)	100	4	3	24	69
JAK2 V617F	40	0.27 (0.12-0.90)	40	25	0	4	11
***SRSF2* P95**	20	0.24 (0.11-0.47)	20	5	0	2	13
***PPM1D* LOF**	28	0.21 (0.10-0.51)	28	0	2	8	18
TOTAL	721	0.19 (0.08-0.90)	678	41	21	174	442

To demonstrate that these selected variants are indeed likely to be somatic, I plotted the variant density against VAF estimated by a Gaussian mixture model which showed that the mean VAF in each gene was less than the expected value of near 0.5 for heterozygous germline variants (Figure 3-9). In addition, the restricted criteria used to select the putative somatic variants identified C>T transitions as the most common type of single nucleotide substitution (n=245; 66%), as expected [261].

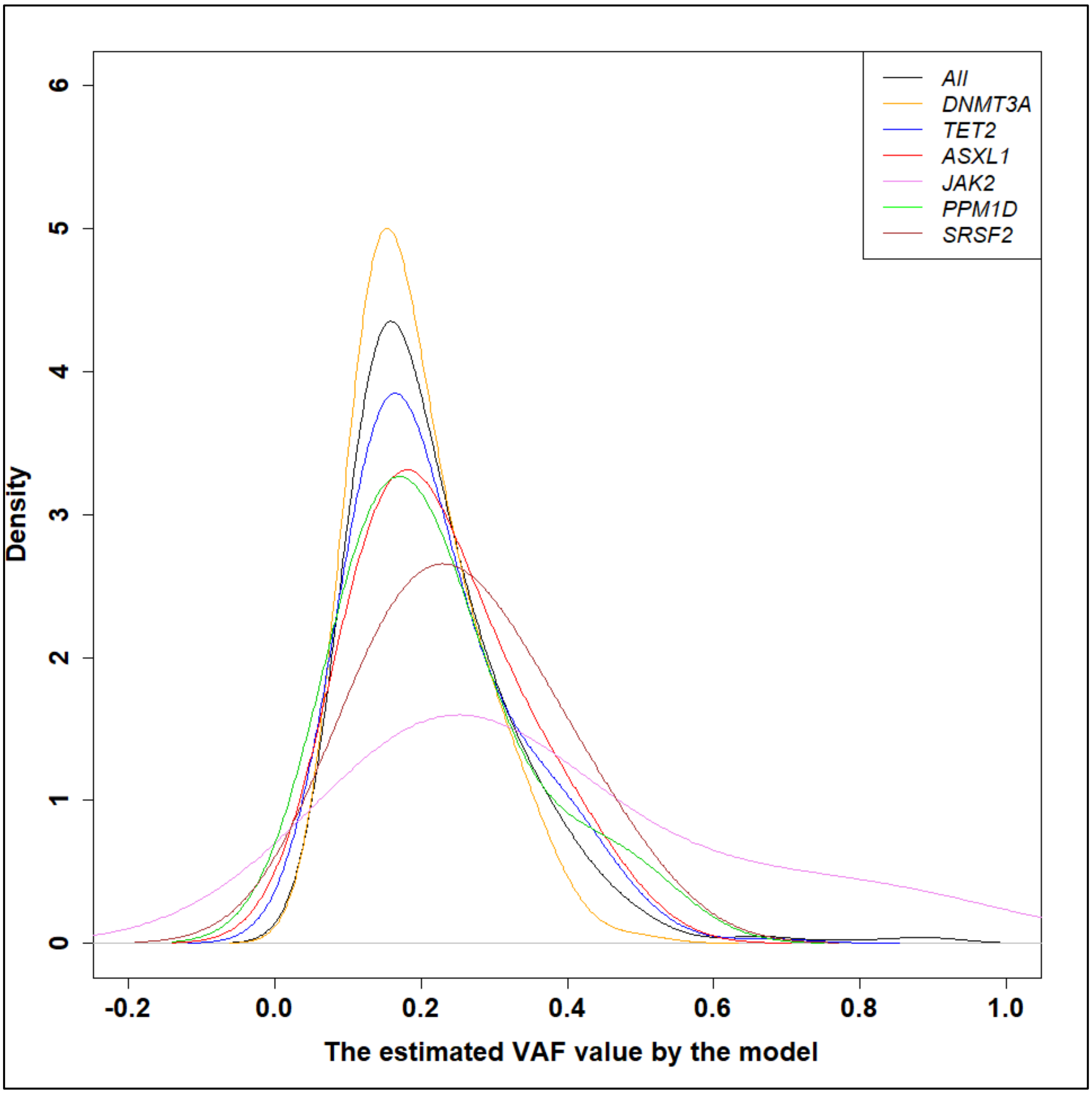


Figure 3-9: Density plot of estimated VAFs for variants in the genes of interest.

Variant density plotted against VAF estimated by a Gaussian mixture model which shows that the mean VAF in each gene was less than that expected for germline variants (near 0.5 for heterozygous variants).

3.4.6 The association of clonality defined by somatic mutations with age

The frequency of CH defined by somatic variants also increased with age and ranged between 0.4% at age <45 years to 2.8% at age >65 years. The overall age-related increase was similar to that seen for myeloid mCA (OR = 1.1; 95% CI = 1.08 - 1.11, P = 5.89x10^{-47}; logistic regression test) and was significant for all 6 genes (Figure 3-10). Also, clonality defined by each single gene was significantly associated with age as shown in Table 3-11. *SRSF2* had the strongest age dependency with an annual increased risk of 1.2 fold per year: *SRSF2* mutations were absent in participants <50 years old, but their prevalence was comparable to *PPM1D* and *JAK2* at age >60 years (Figure 3-10). These results are comparable to other studies, as somatic mutation in splicing genes such as *SRSF2* and *SF3B1* are only detected in individuals aged over 70 [41,262].

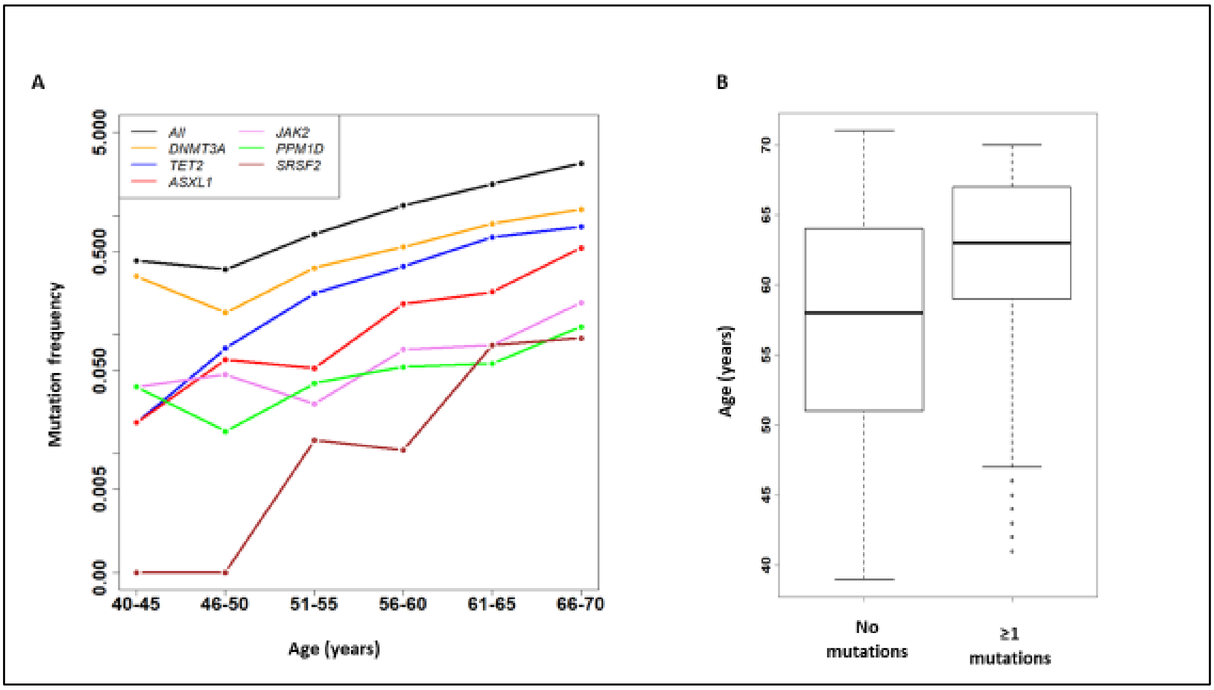


Figure 3-10: The relationship between putative somatic mutations and age.

(A) Frequency of individual mutations showing an age-related increase for all genes individually and combined. The risk of acquiring a mutation in at least one of these genes was estimated to increase by 1.1 fold per year (P=5.89x10^{-47}). (B) Box plot showing increased age in subjects with ≥1 mutation (median = 63 years; n=678) compared to those with no mutations (median=58 years; n=49,278; P=5.89x10^{-47}).

Table 3-11: Results of logistic regression between myeloid CH and age, corrected for sex.

Gene	OR	CI 2.5%	CI 97.5%	P_{FDR}	No. samples
SRSF2	1.21	1.11	1.34	$3.66x10^{-4}$	20
ASXL1	1.13	1.09	1.17	$2.33x10^{-11}$	101
TET2	1.11	1.09	1.14	$2.62x10^{-19}$	223
JAK2	1.09	1.04	1.14	$9.60x10^{-4}$	40
DNMT3A	1.08	1.06	1.10	$5.06x10^{-18}$	309
PPM1D	1.06	1.01	1.12	0.03	28
ALL somatic driver mutations	1.10	1.08	1.11	$5.89x10^{-47}$	678
CNG	1.04	1.02	1.05	$1.09x10^{-6}$	350
aUPD	1.02	1.01	1.02	$3.14x10^{-14}$	2899
CNL	1.01	1.01	1.02	$3.66x10^{-4}$	2141
All mCA	1.02	1.01	1.02	$1.80x10^{-19}$	5040
Myeloid mCA	1.09	1.08	1.11	$1.57x10^{-38}$	506

3.4.7 The association of myeloid CH with common genetic variation

In total, I identified 1,166 individuals with myeloid CH. Of these, 678 had somatic driver mutations, and 506 had myeloid mCA of which 18 subjects had both somatic driver mutations and myeloid mCA. A previous study has associated germline variation at the *TERT* locus (rs34002450) with CH in the Icelandic population [9]. To examine the influence of genetic variation on myeloid CH in the UK Biobank cohort, I performed a GWAS to assess the influence of common variants with MAF >0.1 in the 1,166 cases with at least one CH event defined by (i) mCA associated with myeloid malignancies (as defined in Table 3-8) and/or (ii) somatic mutations in the 6 genes of interest against 30,892 controls with WES data that were free of any mCA, had no putative somatic mutations in the six genes of interest and did not have any haematological malignancies during the study period. A total of 286,909 variants passed quality control (QC). The observed P values follow the expected distribution with lambda = 1.021 (Figure 3-11) indicating an absence of any systematic bias between cases and controls such as residual population stratification. Three SNPs with genome-wide significance were identified in the *TERT* gene

Table 3-12). Two of these were associated with an increased risk of developing CH (rs2853677 intron 2, OR = 1.32, P = 5.6x10^{-11}; rs7726159 intron 3, OR = 1.33, P = 4.2x10^{-11}) while the third and most significant single SNP was protective (rs2736100 intron 2, OR = 0.74, P = 3.1x10^{-12}) (Figure 3-12). Two of these SNPs (rs7726159, A allele, OR = 1.19, P = 0.003 and rs2853677, G allele, OR = 1.18, P = 0.004) were identified as independent association signals using stepwise logistic regression with an additive model and treating all three SNPs as covariates. LD analysis for the two primary signals (rs7726159 and rs2853677; Table 3-13) revealed (i) rs7726159 is in LD with rs7705526 (r^2=0.79); (ii) rs2853677 is not in LD with rs7705526, (r^2=0.19), (iii) rs2736100 is in modest LD with rs7705526 (r^2=0.51) (Figure 3-13). The association of these SNPs with myeloid malignancies was cross referenced with published GWAS results of self-reported PV, using the UK Biobank 150K V1 SNP array (http://big.stats.ox.ac.uk/), and with meta-analysis of MPN using three independent cohorts (the UK Biobank, 23andMe and FinnGen [144]. Only *TERT* SNPs in intron 2 were significantly associated with self-reported PV in the UK Biobank, as shown in Table 3-13 and Figure 3-13.

A second signal was seen just below the level of genome wide significance (Figure 3-12) and included rs3780381, rs17425819 and rs10974944. These SNPs are within *JAK2* and are in LD with the 46/1 haplotype, previously shown to be strongly associated with acquisition of *JAK2* V617F [263]. This association signal disappeared when cases with *JAK2* V617F (n=40) and mCA including *JAK2* (n=115) were removed from the analysis.

Table 3-12: Results of the allelic association between myeloid CH and common SNPs with MAF > 0.1.

CHR	SNP	BP	A1	Gene$	Risk allele frequency (A1)		A2	CHISQ	P*	OR
					Cases	Controls				
5	rs2736100	1286516	A	*TERT*	0.4249	0.4987	C	48.6	3.14 x 10^{-12}	0.743
5	rs7726159	1282319	A	*TERT*	0.3946	0.329	C	43.52	4.19 x 10^{-11}	1.329
5	rs2853677	1287194	G	*TERT*	0.4862	0.4179	A	42.94	5.64 x 10^{-11}	1.318
9	rs3780381	5114523	C	*JAK2*	0.3202	0.27	A	28.23	1.08 x 10^{-7}	1.274
9	rs17425819	5114773	T	*JAK2*	0.3184	0.269	C	27.36	1.69 x 10^{-7}	1.269
9	rs62554837	5266200	T	*JAK2*	0.1911	0.1517	C	26.67	2.41 x 10^{-7}	1.321
9	rs10974944	5070831	G	*JAK2*	0.301	0.2536	C	26.21	3.06 x 10^{-7}	1.267
9	rs10989523	104230977	C	*TMEM246 (PGAP4)*	0.1053	0.1392	T	21.65	3.27 x 10^{-6}	0.727
15	rs319889	35950483	C	*DPH6 (DAXX)*	0.4431	0.492	T	21.45	3.63 x 10^{-6}	0.821
1	rs80291200	57903916	G	*DAP1*	0.2415	0.2025	T	20.8	5.09 x 10^{-6}	1.254
9	rs10974900	4987958	T	*JAK2*	0.3666	0.413	C	19.87	8.27 x 10^{-6}	0.823

* Italic bold indicates genome wide significance. Chromosome (chr); odds Ratio (OR); allelic test chi-square (CHISQ), Allele 1 (A1), Allele 2 (A2), base pair (BP)

$ nearest gene

Table 3-13: LD analysis of *TERT* SNPs that predispose to CH and/or MPN

SNP ID	CH association (p value)	MPN association (p value)[1]	MPN association (P value)[2]	LD results (r^2/D')				
				rs34002450	rs7726159	rs7705526	rs2736100	rs2853677
rs34002450	N/A			1				
rs7726159	$4.19\mathrm{x}10^{-11}$	$2.5\mathrm{x}10^{-4}$	N/A	0.705/0.961	1			
rs7705526	N/A	$6.8\mathrm{x}10^{-6}$	5×10^{-54}	0.534/0.821	0.788/0.914	1		
rs2736100	$3.14\mathrm{x}10^{-12}$	$1\mathrm{x}10^{-5}$	N/A	0.311/0.695	0.516/0.977	0.51/1	1	
rs2853677	$5.64\mathrm{x}10^{-11}$	$2.9\mathrm{x}10^{-6}$	3×10^{-44}	0.092/0.319	0.181/0.487	0.185/0.507	0.435/0.784	1

N/A = not available

CH association is derived from 1,166 cases vs 30,892 controls in the UK Biobank.
[1] MPN association: is derived from the published GWAS results of self-reported PV, using the UK Biobank 150K V1 SNP array (http://big.stats.ox.ac.uk/).

[2] MPN association reported by meta-analysis of MPN in three cohorts the UK Biobank, 23andMe and FinnGen rs7705526 was not included in the final version of the UK Biobank v2 SNP array data used for my investigation. rs34002450 was significantly associated with CH defined by putative somatic variants [9].

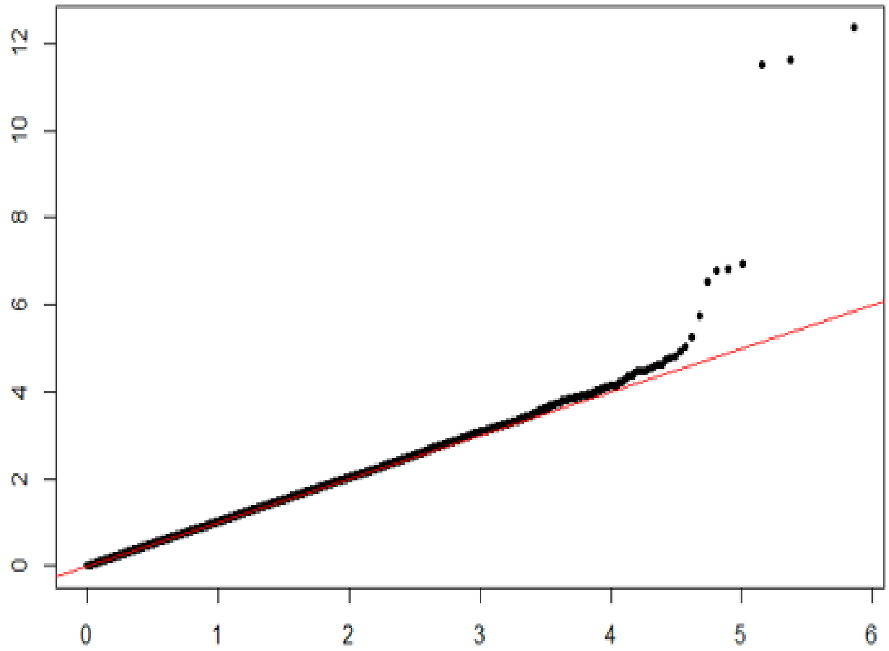


Figure 3-11: Quantile-quantile plot showing observed versus expected P values.

No evidence was seen for systematic bias between cases and controls, or population stratification (lambda=1.021).

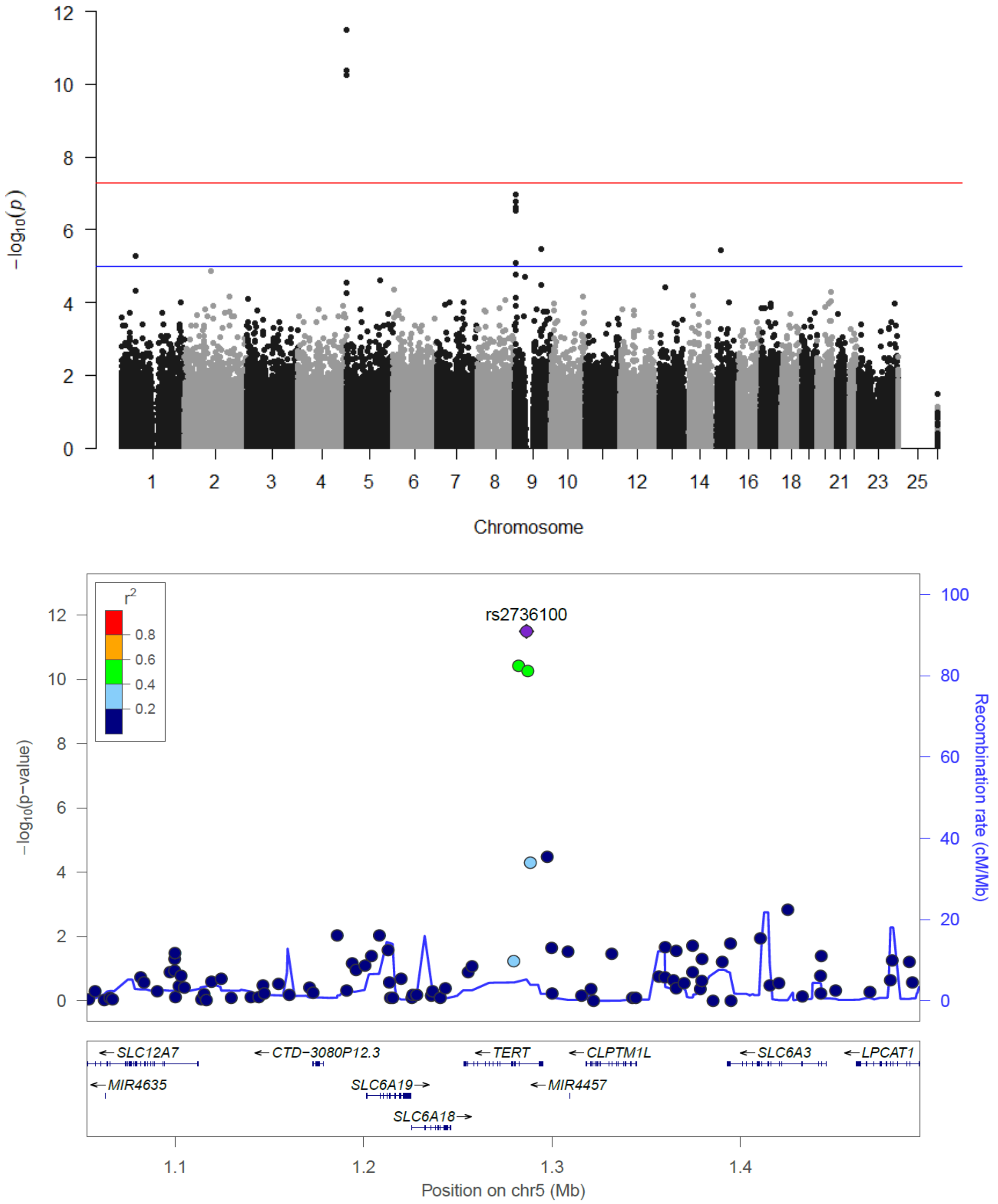


Figure 3-12: GWAS results for myeloid CH

Top panel: Manhattan plot summarising the significance of SNPs across the genome. The red line indicates genome wide significance ($P < 5x10^{-8}$) and the blue line indicates values that were of suggestive significance ($P < 10^{-5}$). Clusters related to *TERT* and *JAK2* are indicated. Lower panel: Locus zoom plot focusing on SNPs in the region of *TERT* at chromosome band 5p15. The lead rs2736100 variant is in purple.

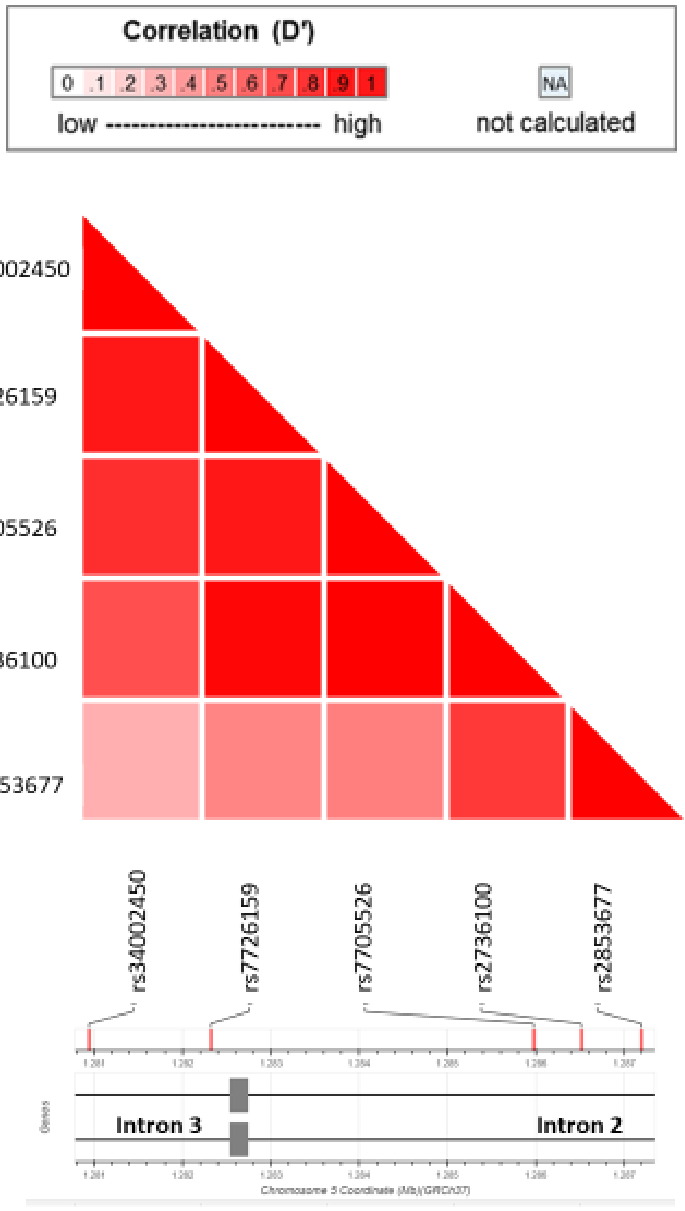


Figure 3-13: A visual plot of the coefficient of linkage disequilibrium (D) for 5 SNPs in *TERT*

rs2853677, rs7726159, and rs2736100 were assessed in the UK Biobank cohort. The other two SNPs, rs34002450 and rs7705526, were assessed in other studies [9,253].

3.4.8 The relationship between myeloid CH and smoking

To assess the relationship between smoking and CH, I used PHESANT to perform regression analyses of past, current and combined smoking status in 32,058 participants consisting of the 1166 cases (488 past smokers and 134 current smokers) with myeloid CH and 30,892 controls (10,952 past smokers and 2,699 current smokers). The odds of having ever smoked (combined status) were significantly higher in participants with myeloid CH (53% smokers, n = 622) than those without myeloid CH (44% smokers, n = 13,651; $P_{FDR} = 3.38\times10^{-6}$, Table 3-14). This effect was associated with current smoking status (OR = 1.10, $P_{FDR} = 6.14\times10^{-6}$) rather than past smoking status (OR = 1.02, $P_{FDR} = 0.08$).

Strikingly, breakdown of myeloid CH by specific mutation type revealed that variants in *ASXL1* were strongly associated with current smoking status in ordinal logistic regression analysis (OR = 1.07; P = $1.92\text{x}10^{-5}$), and the only abnormality associated with past smoking (OR = 1.04; P = $2.6\text{x}10^{-3}$). Indeed, 69% of participants with *ASXL1* mutations were past or current smokers. Myeloid CH cases without *ASXL1* mutations (n=1066) remained significantly associated with current smoking but the effect was weaker (P = $8.8\text{x}10^{-4}$). Both *TET2* and *DNMT3A* variants showed a significant, but relatively modest, association with current smoking status but there was no discernible association between smoking and variants in *JAK2*, *SRSF2*, *PPM1D* or for acquired myeloid mCA, Table 3-14 and Supplementary Table 3-3).

Table 3-14: The relationship between smoking and clonal haematopoiesis.

Marker[1]	No. myeloid	No. of smokers[2]		Previous smoking[3]		Current smoking[3]		Combined smoking[4]	
		Past	Current	OR	P_{FDR}	OR	P_{FDR}	OR^{a}; OR^{b}	P_{FDR}
mCA	506	218	48	1.01	0.39	1.03	0.18	1.19; 1.42	0.091
ASXL1 LOF	100	49	20	1.04	**2.60x10^{-3}**	1.07	**1.92x10^{-5}**	1.94; 4.68	**1.02x10^{-5}**
DNMT3A LOF or R882	308	117	35	1.00	1.00	1.05	**0.03**	1.03; 1.64	0.07
JAK2 V617F or chr9p mCA	155	64	7	0.99	0.68	0.95	0.18	0.95; 0.54	0.27
PPM1D LOF	28	15	3	1.02	0.16	1.01	0.68	2.07; 2.05	0.23
SRSF2 P95	20	10	1	1.01	0.55	1.00	1.00	0.88; 1.82	0.83
TET2 LOF	208	75	27	0.99	0.5	1.06	**6.40x10^{-3}**	0.88; 1.82	0.03
All myeloid CH	1,166	488	134	1.02	0.09	1.10	**6.14x10^{-6}**	1.17; 1.76	**3.38x10^{-6}**
Myeloid CH without *ASXL1*	1066	439	114	1.01	0.36	1.07	**8.8x10^{-4}**	1.12; 1.59	**5.81x10^{-4}**

Loss of function (LOF); clonal haematopoiesis (CH); mosaic chromosomal alterations (mCA)

Number of smokers encoded in the combined smoking status in the UK Biobank "Data-Field 20116"

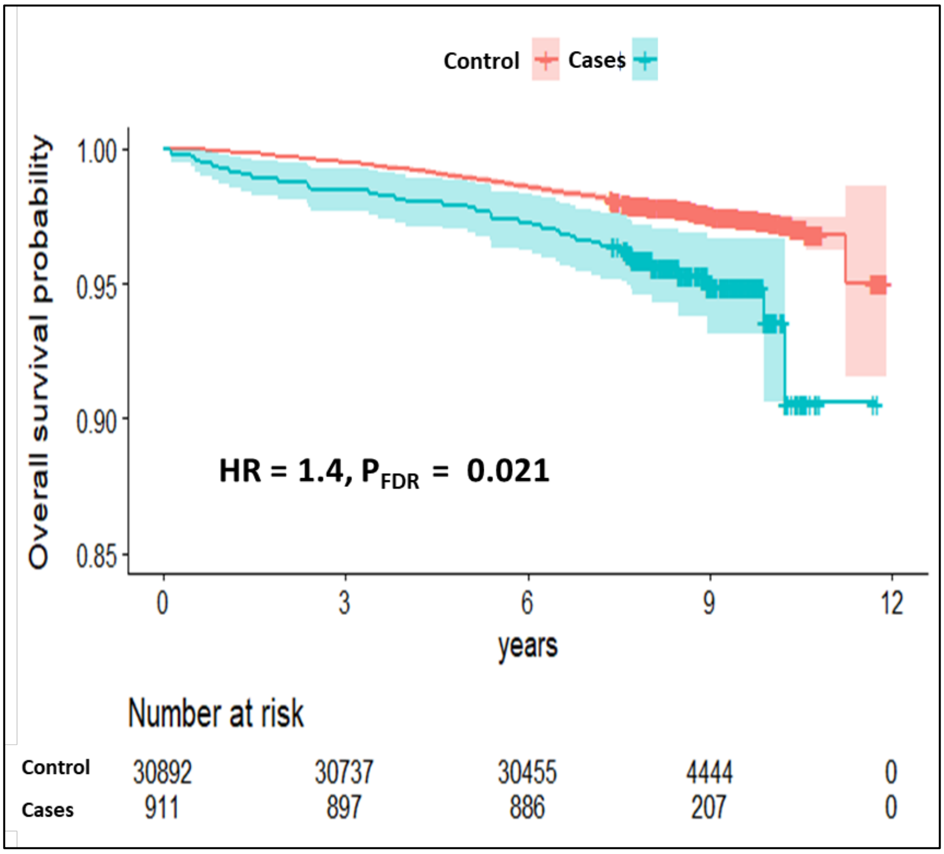
Results of ordinal logistic regression, total tests = 16, corrected for age, sex and FDR.

Results of multinomial logistic regression, total tests = 8, corrected for age, sex and FDR. Odds ratios are estimated for past smoking level (a), and current smoking (b)

3.4.9 The relationship with all cause of mortality

To identify any effect on mortality that was unrelated to blood cancer a subset of 911 myeloid CH cases were selected that were free of haematological malignancy at any time during the study period. These cases were compared with the 30,892 controls that were free of any haematological malignancy at any time point, and had no evidence of CH. Within the study period, 42 cases (4.6%) died compared to 674 controls (2.2%), a significant difference ($P < 2x10^{-6}$, log-rank test). Using multivariable Cox regression, myeloid CH was shown to be associated with an increased risk of all-cause mortality (HR = 1.44; P = 0.021; mean follow up = 8.1 years) after adjusting for age and sex. This analysis reveals that myeloid CH plays a pathogenetic role beyond predisposition to haematological malignancies (Figure 3-14).

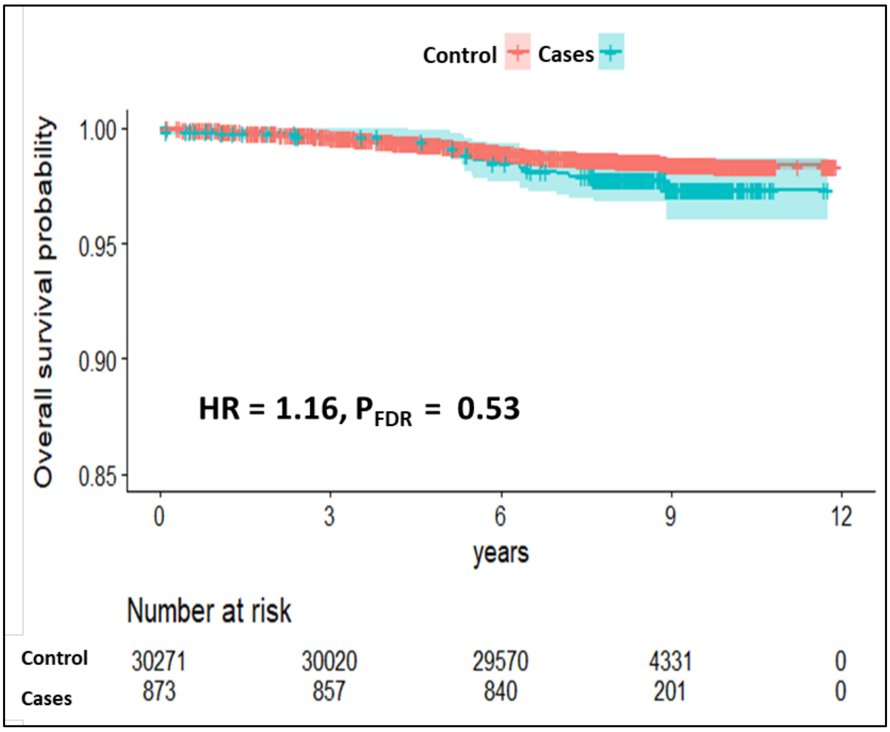
I examined the relationship between myeloid CH and the 2 main components of cardiovascular diseases, MI and stroke. For this analysis any participant who had an event (MI or stroke, as appropriate) prior to sampling was removed from the analysis. For MI, a difference was found in the number of events in cases (20/873; 2.1%) compared to controls (419/30,271; 1.4%; P = 0.03, long rank sum test), but this was not significant when a multivariate Cox hazard model was applied considering age, sex and smoking status as co-variates (HR = 1.16, P = 0.53) (Figure 3-15). Similar results were observed for stroke: 12/890 (1.35%) events were observed in cases with myeloid CH compared to 235/30,472 (0.78%) events in controls (P = 0.06, long rank sum test). On multivariate analysis considering age, sex and smoking status as co-variates, stroke was not significantly associated with myeloid CH (HR = 1.18; P = 0.58) (Figure 3-16).



Number at risk

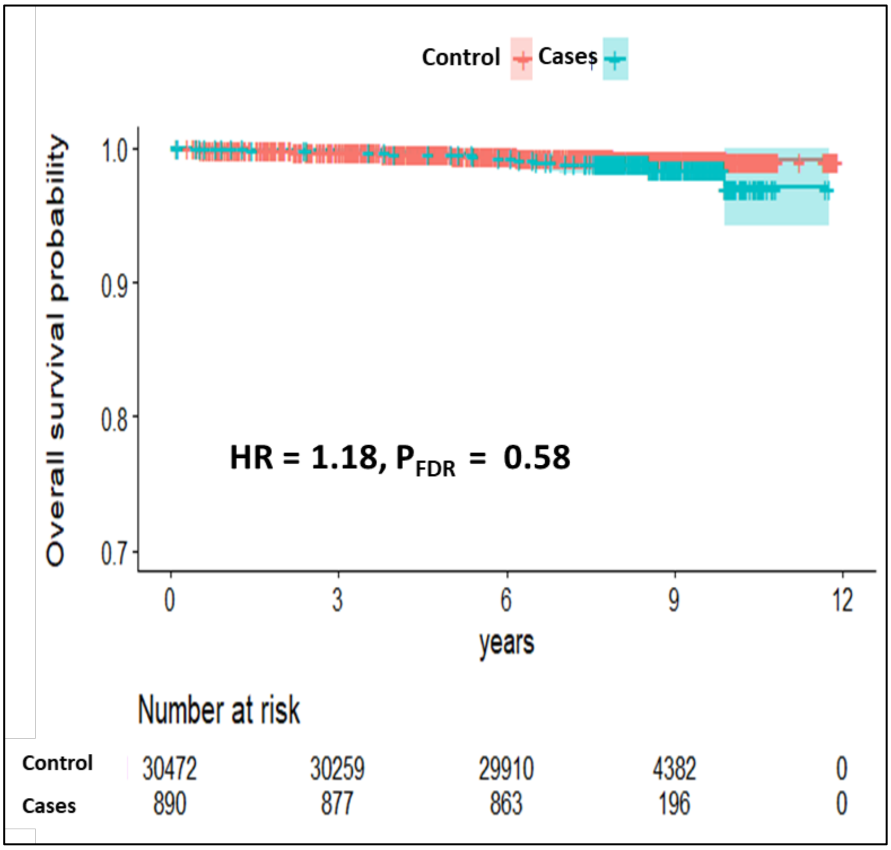
Control	30892	30737	30455	4444	0
Cases	911	897	886	207	0

Figure 3-14: Kaplan-Meier plot with overall survival probability and number at risk of all-cause mortality



Control	30271	30020	29570	4331	0
Cases	873	857	840	201	0

Figure 3-15: Kaplan-Meier plot with overall survival probability and number at risk for myocardial infarction.



Number at risk

Control	30472	30259	29910	4382	0
Cases	890	877	863	196	0

Figure 3-16: Kaplan-Meier plot with overall survival probability and number at risk for stroke

3.4.10 The relationship with blood features and clinical phenotype

To investigate the relationship with blood features and clinical phenotypes the myeloid CH cases without haematological malignancies (n=911) and controls without haematological malignancies or CH were used (n=30,892). All comparisons were corrected for sex, age, smoking status and multiple testing using FDR. Comparisons were performed for all cases as a single group, and for cases stratified by subtype (mCA, n = 301; *DNMT3A*, n = 300; *TET2,* n = 189*; SRSF2,* n = 15*; ASXL1,* n = 93*; PPM1D,* n = 26*; JAK2* V617F/chr9p mCA*;* n = 29).

Clinical phenotype:

This analysis confirmed an association between *TET2* mutations and chronic obstructive pulmonary disease (COPD) with acute lower respiratory infection "ICD10= J44.0" (logistic regression, OR = 1.16, P = 7.90x10^{-3}). In addition, a significant association was found with *TET2* mutations and agranulocytosis "ICD-10 = D70" (P = 7.9x10^{-3}, OR = 1.1) and ulcers of the lower limb "L97" (P = 1.99x10^{-3}, OR = 1.28). A significant association between myeloid mCA and urinary tract related disorders was seen, specifically "urethral stricture unspecified N35.9" (P = 8x10^{-3}, OR = 1.17), and "bladder-neck obstruction N32.0" P = 9x10^{-3}, OR = 1.9). No other significant associations were found; results are summarised in Table 3-15, and detailed in Supplementary Table 3-4.

Table 3-15: Clinical phenotypes significantly associated with myeloid CH

Marker	Phenotype	No. positive cases	No. positive controls	OR	CI 2.5%	CI 97.5%	P_{FDR}
mCA n=301	**(N35.9) Urethral stricture, unspecified**	10 (3.3%)	189 (0.6%)	1.169	1.085	1.244	0.008
	N32.0 Bladder-neck obstruction	7 (2.3%)	98 (0.03%)	1.192	1.093	1.280	0.009
***TET2* n=189**	**(L97) Ulcer of lower limb, not elsewhere classified**	4 (2.1%)	19 (0.06%)	1.276	1.137	1.394	0.004
	D70 Agranulocytosis	4 (2.1%)	51 (0.16%)	1.231	1.101	1.337	0.009
	(J44.0) Chronic obstructive pulmonary disease with acute lower respiratory infection	7 (3.7%)	137 (0.44%)	1.158	1.074	1.230	0.009

Blood counts and blood biochemistry:

Blood measurements (Data Category 100081, n = 29), and blood biochemistry markers (Data Category 17518, n=30) were also tested for association with myeloid CH using linear regression (Supplementary Table 3-5, and Supplementary Table 3-6). Nucleated red blood cell percentage and count were excluded because they have skewed binomial distribution. Regression models included sex, age, and smoking status as covariates and FDR to adjust P-values for multiple tests. As above, participants with any evidence of haematological malignancies were excluded. Overall, myeloid CH showed a significant association with elevated RBW, all the platelet related indices, low basophils and haematocrit percentage (Table 3-16, Table 3-17). The breakdown of myeloid CH to the driver gene level shows specific associations:

Myeloid-related mCA: the selected set of mCA showed a clear disruption in erythropoiesis, as this group is associated with decreases in red blood cellcounts, haemoglobin concentration and haematocrit percentage. On the other hand it was associated with high mean corpuscular haemoglobin (MCH) and RDW. The biochemistry measures of this group indicated a decrease in high density lipoprotein (HDL) cholesterol and apolipoprotein A, and also an association with low creatinine, phosphate and albumin levels.

JAK2 V617F/chr9p: JAK2 V617F and its related chr9p mCA were associated with an increase in platelet counts, percentage and distribution width. Given the established role of *JAK2* V617F in the pathogenesis of ET, this is likely to be a direct causal effect. *TET2* was significantly associated with decrease in eosinophils counts and percentage.

ASXL1 LOF: ASXL1 cases presented an anaemia-like profile as they were associated with low mean corpuscular volume (MCV), MCH and mean sphered cell volume (MSCV). *ASXL1* was also significantly associated with low Insulin Growth Factor 1 (IGF-1).

SRSF2 P95: SRSF2 cases were associated with a proliferative character of elevated reticulocytes indices, but also a decrease in HDL cholesterol.

Table 3-16: Significant blood features associated with myeloid markers

Group	Blood feature	Units	No.	Mean value		OR (CI 97.5%)	P_{FDR}
				Cases	Control		
mCA	Basophil count	10^9/L	30272	0.03	0.04	0.92 (0.90-0.94)	9.1×10^{-13}
	Platelet distribution width	%	30282	16.69	16.46	1.04 (1.03-1.05)	4.3×10^{-9}
	Haematocrit percentage	%	30282	41.05	41.61	0.98 (0.97-0.99)	1.0×10^{-4}
	Basophil percentage	%	30272	0.53	0.61	0.97 (0.96-0.99)	2.9×10^{-4}
	Mean corpuscular haemoglobin	g/dL	30282	34.43	34.27	1.02 (1.01-1.03)	4.2×10^{-3}
	Red blood cell count	10^9/L	30282	4.47	4.54	0.98 (0.97-0.99)	4.4×10^{-3}
	Red blood cell distribution width	%	30282	13.81	13.5	1.02 (1.01-1.03)	4.9×10^{-3}
	Haemoglobin concentration	g/dL	30282	14.13	14.26	0.98 (0.98-0.99)	6.8×10^{-3}
	Mean sphered cell volume	fL	28916	83.99	84.5	0.98 (0.97-0.99)	3.0×10^{-2}
ASXL1	Platelet distribution width	%	30082	16.72	16.46	1.03 (1.01-1.04)	5.9×10^{-4}
	Red blood cell distribution width	%	30082	13.93	13.5	1.03 (1.01-1.04)	7.0×10^{-4}
	Mean corpuscular volume	fL	30082	90.68	91.8	0.98 (0.97-0.99)	9.1×10^{-4}
	Mean corpuscular haemoglobin	Pg	30082	31.03	31.47	0.98 (0.97-0.99)	2.8×10^{-3}
	Mean sphered cell volume	fL	28712	83.36	84.5	0.98 (0.97-0.99)	1.0×10^{-2}
DNMT3A	Platelet count	10^9/L	30289	251.27	242.76	1.02 (1.01-1.03)	1.8×10^{-2}
JAK2	Platelet crit	%	30019	0.30	0.22	1.04 (1.03-1.06)	7.5×10^{-12}
	Platelet count	10^9/L	30019	341.46	242.76	1.04 (1.03-1.06)	1.3×10^{-11}
	Red blood cell distribution width	%	30019	15.31	13.5	1.04 (1.03-1.05)	3.6×10^{-9}
	Platelet distribution width	%	30019	17.14	16.46	1.03 (1.02-1.05)	1.0×10^{-6}
	High light scatter reticulocyte	10^{12}/L	28656	0.02	0.02	1.02 (1.01-1.03)	4.0×10^{-2}
PPM1D	Monocyte count	10^9/L	30007	0.59	0.48	1.02 (1.01-1.03)	3.5×10^{-2}
SRSF2	Reticulocyte percentage	%	28643	1.88	1.32	1.02 (1.01-1.03)	1.3×10^{-2}
	High light scatter reticulocyte	%	28643	0.62	0.4	1.02 (1.01-1.03)	2.4×10^{-2}
TET2	Eosinophil count	10^9/L	30169	0.15	0.17	0.98 (0.97-0.99)	1.1×10^{-3}
	Eosinophil percentage	%	30169	2.18	2.53	0.98 (0.97-0.99)	4.4×10^{-3}
	Monocyte percentage	%	30169	7.83	7.06	1.02 (1.01-1.03)	3.0×10^{-2}
All Myeloid CH	Platelet distribution width	%	30883	16.57	16.46	1.03 (1.02-1.04)	6.0×10^{-6}
	Basophil count	10^9/L	30873	0.04	0.04	0.95 (0.93-0.97)	5.9×10^{-4}
	Red blood cell distribution width	%	30883	13.71	13.5	1.02 (1.01-1.04)	9.7×10^{-4}
	Platelet crit	%	30883	0.23	0.22	1.02 (1.01-1.03)	1.1×10^{-3}
	Haematocrit percentage	%	30883	41.46	41.61	0.98 (0.97-0.99)	1.7×10^{-3}
	Platelet count	10^9/L	30883	248.32	242.76	1.02 (1.01-1.03)	3.3×10^{-3}
	Haemoglobin concentration	g/dL	30883	14.22	14.26	0.99 (0.98-0.99)	1.7×10^{-2}
	Basophil percentage	%	30883	0.58	0.61	0.98 (0.97-0.99)	4.4×10^{-2}

Table 3-17: Significant biochemical measures associated with myeloid markers

Group	Biochemistry measure	Units	N	Mean in		OR (CI 97.5%)	P_{FDR}
				Cases	Control		
mCA	Creatinine	µmol/L	29280	71.452	72.686	0.98 (0.97-0.99)	0.001
	Apolipoprotein A	g/L	27335	1.517	1.555	0.98 (0.97-0.99)	0.004
	Phosphate	mmol/L	27515	1.169	1.200	0.98 (0.97-0.99)	0.005
	HDL cholesterol	mmol/L	27546	1.420	1.474	0.98 (0.97-0.99)	0.010
	Albumin	g/L	27576	44.848	45.518	0.98 (0.97-0.99)	0.018
ASXL1	IGF-1	nmol/L	28977	19.043	21.697	0.98 (0.97-0.99)	0.033
SRSF2	HDL cholesterol	mmol/L	27294	1.237	1.474	0.98 (0.97-0.99)	0.027
All myeloid CH	Cholesterol	mmol/L	29874	5.619	5.697	0.98 (0.97-0.99)	0.033
	HDL cholesterol	mmol/L	28085	1.450	1.474	0.98 (0.97-0.99)	0.040
	Creatinine	µmol/L	29851	72.706	72.686	0.99 (0.98-0.99)	0.041

3.4.11 Validation of the association between *ASXL1* and smoking

In January 2021, the UK Biobank released a new set of variant calls from whole exome sequencing of 200,631 participants [264] including 49,946 individuals that were previously released and 150,685 additional individuals. These variants were identified using a new pipeline, DeepVariant version 0.10.0 [265], that employs a deep neural network for variant calling. Analysis of the 200,631 exomes will be reported in Chapter 4. To validate the findings in this chapter, I restricted the analysis to the new samples (n=150,685) and point mutations of *JAK2* V617F*, DNMT3A* R882*, SRSF2* P95, and LOF mutations in *TET2*, *DNMT3A*, *ASXL1* and *PPM1D*. LOF mutations were considered if inferred as somatic by failing the hypothesis that the alternative allele is normally distributed with a mean of 0.45 and a false positive rate of P = 0.05 using a binomial test. I identified 1,416 candidate driver mutations in 1,345 subjects, with *DNMT3A* being the most commonly affected gene. Only 67 cases had more than one variant. Of the 1345 participants with CH defined by somatic mutations, 56 (*JAK2*, n = 37; *DNMT3A*, n = 5; *ASXL1*, n = 6, *PPM1D*, n = 3 *TET2*, n = 10; *SRSF2*, n = 4) also had CH defined by myeloid mCA.

In keeping with the previous analysis, I used PHESANT to perform regression analyses of past, current and combined smoking status. *ASXL1* mutations were most strongly associated with smoking status

(OR_{past} = 1.34, $OR_{current}$ = 2.97, P = 3.43x10^{6}, multinomial regression), current smoking status (OR = 1.04, P_{FDR} = 2.01x10^{-7}), and past smoking status (OR = 1.01, P_{FDR} = 0.05), Indeed, 61% of participants with *ASXL1* mutations were past or current smokers. Using meta-analysis to combine evidence from the first (n = 49,946) and second (n = 150,685) releases of whole exome data, current smoking (OR = 1.05, P_{FDR} = 8 x 10^{-13}) past smoking (OR = 1.01, P_{FDR} = 2 x 10^{-7}) were both associated with *ASXL1* mutations and without evidence for heterogeneity (Cochran's Q test, P_{past} = 0.05 and $P_{current}$ = 0.1). *DNMT3A* mutations were associated with the combined smoking status (OR_{past} = 1.13, $OR_{current}$ = 1.81 P = 7.17x10^{-5}), that was due to current smoking (P_{FDR} = 6.12x10^{-5}). *TET2* mutations had a significant association with past smoking status (OR = 1.01, P_{FDR} = 0.03) Also, there was no discernible association between smoking and variants in *JAK2, SRSF2* or *PPM1D* (Table 3-18). In a meta-analysis, *DNMT3A* were significantly associated with both current and past smoking with no heterogeneity (P_{past} = 0.36, and $P_{current}$ = 0.59), but results for *TET2* mutations showed heterogenous results (P_{past}=0.04, and $P_{current}$ = 0.09) as illustrated in Figure 3-17.

Table 3-18: The relationship between smoking and clonal haematopoiesis in the validation cohort

Marker[1]	Count	No of smokers[2]		Previous smoking[3] "Data-field 1249"		Current smoking[3] "Data-field 1239"		Combined smoking[4] "Data-field 20116"		
		Past	Current	OR	P^{FDR}	OR	P^{FDR}	OR^a	OR^b	P^{FDR}
ASXL1 LOF	227	96	43	1.01	0.05	1.04	2.09×10^{-7}	1.34	2.97	5.71×10^{-7}
TET2 LOF	378	169	28	1.01	0.03	0.99	0.69	1.24	1.01	0.13
DNMT3A R882/LOF	581	219	79	1.01	0.07	1.04	6.12×10^{-5}	1.13	1.81	7.17×10^{-5}
JAK2 V617F	101	33	4	0.99	0.10	0.98	0.14	0.67	0.37	0.03
SRSF2 P95	61	28	7	1.00	0.69	1.01	0.53	1.15	1.48	0.65
PPM1D LOF	54	26	7	1.01	0.09	1.01	0.30	1.62	2.12	0.14

[1] *LOF* Loss of function

[2] Number of smokers encoded in the combined smoking status in the UK Biobank "Data-field 20116".

[3] Results of ordinal logistic regression, total tests = 12, corrected for age, sex, and FDR.

[4] Results of multinomial logistic regression, total tests = 6, corrected for age, sex, and FDR. Odds ratios are estimated for past smoking level (a), and current smoking (b).

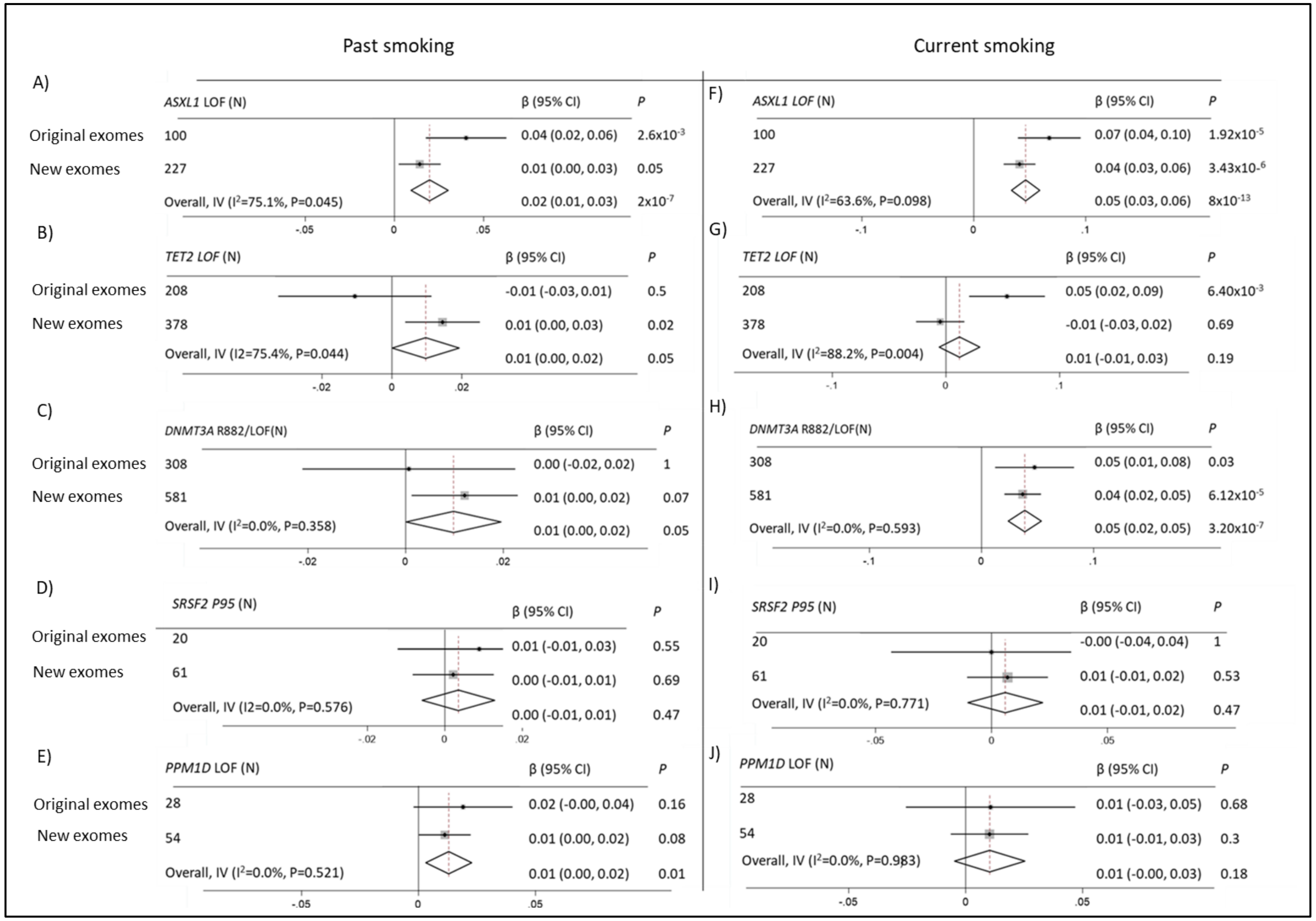


Figure 3-17: Meta analysis of smoking association results.

Forest plots of A, B, C, D, and E represent ordinal regression of past smoking on data in Data-field 1249. Forest plots of F, G, H, I, and J represent ordinal regression of current smoking on data in Data-field 1239.

3.5 Discussion

The prevalence and significance of CH has been reported in several cohorts, but my study has several distinctive features. The UK Biobank is a very large population-based cohort that includes an extensive repertoire of baseline phenotypic data as well as over 9 years of prospective clinical follow up information. Genome wide SNP data is available for the great majority of participants (n = 486,941), and WES data for 49,956 initially and 200,641 at the time of writing, which are all derived from a single baseline peripheral blood sample taken at study entry. Thus, I was able to assess CH associated with both mCA and somatic mutations, albeit with a modest limit of detection compared to some published studies. I focused on myeloid mCA, and genes known to be mutated in myeloid disorders with the specific aim of understanding the causes and consequences of myeloid CH.

To identify relevant mCAs, I processed the allelic frequencies and copy number calls from the UK Biobank SNP array data (n = 486,941) to identify regions of AI. Next, I developed an evaluation strategy to filter out artefacts and likely germline events. This method identified mCAs (n = 8,203) of relatively large size (≥ 2Mb) and relatively large clone size (> 10%) in 1% (n = 5,040) of the UK Biobank cohort. The incidence of these events increased with age from 0.85% at age 40 - 45 years to 1.29% at age 65 - 70 years. The age-related increase in risk of acquiring a mCA was greatest for those associated with myeloid disorders, with an estimated annual risk of 1.1 fold and an increase in frequency from 0.03 at age 40-45 to 0.23 at age 66-70, as shown in Figure 3-18. The relationship between age and CH reflects the fitness that depends on the mutation rate and/or the ability to form large clones [83].

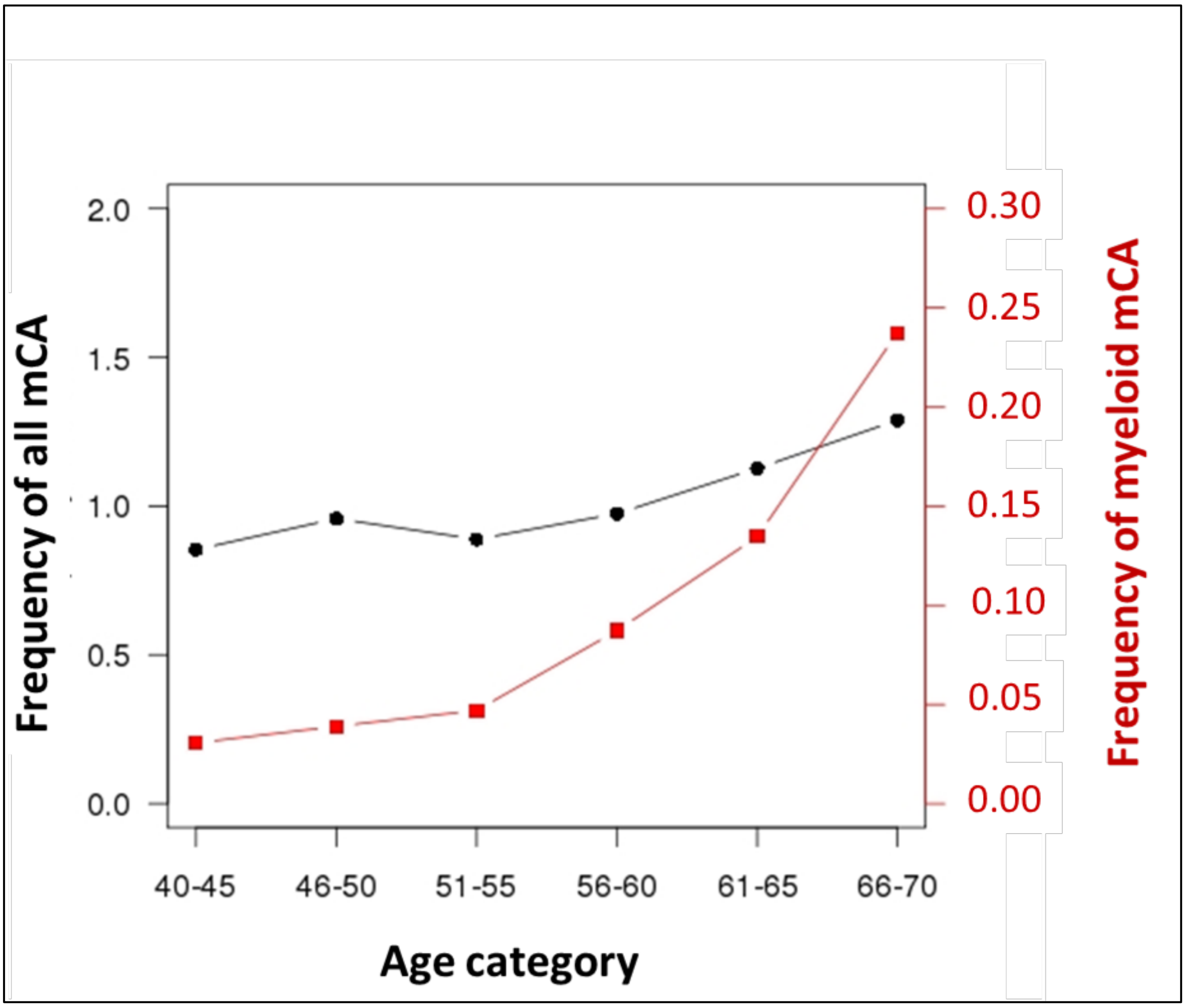


Figure 3-18: The age-related increase in myeloid mCA (red line/red squares) is steeper than for all mCA (black line/black circles)

3.5.1 Advantages and limitations of calling mCA

Specific features of our study compared to other studies include (i) the consistent genotyping of all samples on two very similar genotyping arrays (UK BiLEVE and UK Axiom arrays) whereas previous studies pooled SNP arrays from different projects and (ii) all the markers and samples were processed through the same quality control pipeline. On the other hand, the UK Biobank had a limited age for recruitment between 40 to 70 years that is much narrower in comparison to other studies. In general, though, our results are comparable to the previous publications. In a combined study of 31,717 cancer cases and 26,136 cancer-free controls, mCA of size >2Mb were identified in autosomes of 517 individuals (0.89%), ranging from 0.23% under 50 years to 1.91% between age 75 and 79 [7]. In another major study, mCA > 2Mb in 50,000 subjects from different GWAS was identified in less than 0.5% of subjects at age less than 50, rising up to 3% at age 80 years [6]. My study identified mCA in 5040 of 486,941 individuals that ranged from 0.85% at age 40 - 45 years to 1.29% at age 65 - 70 years.

Participants diagnosed with myeloid malignancies (n = 1,913) had the largest prevalence of mCA at 11% (n = 210) with at least one event. In addition, the majority of the myeloid malignancy cases (74%) acquired one or more region of aUPD. These results confirm the role of mCA in the pathogenesis of myeloid malignancies. I note that the prevalence of mCA in myeloid malignancies from the published literature varies according to the diagnosed disease, age, and the applied technique. In a study of 64 AML cases (median age = 55.5) 20% had aUPD [17]. This frequency is higher in CMML: in a detailed study of SNP array data (median age = 77) 55/70 (79%) cases tested had at least one AI event [266]. I explain the lower frequency of mCA in the myeloid malignancy group in my study (11%) by; (i) my definition of myeloid malignancy included all the diagnosed cases between 1995 and 2018, with no differentiation between prevalence or incidence; (ii) limitations of the data and methodology used (BAF segmentation) which I discuss below; (iii) the age of the UK Biobank cohort ranges between 40 and 69 years with median of 58 years

Regarding my method, the segmentation of allelic frequencies of SNPs to identify regions of AI is a proven method that has been used by similar studies [6,7]. But the method has some drawbacks:

(i) It cannot identify small clones at a frequency less than 10% (however in the clinical context, large clones are likely to be more significant with regard clinical phenotype).

(ii) Large clones at a frequency above 90% cannot be identified since they cannot be distinguished from germline events in the absence of a germline control.

(iii) Due to a high level of artefacts and germline events, I had to customise a scoring and filtration strategy to eliminate noise. Although this approach was validated by three approaches, it is unclear how many artefacts remained and how many true events were removed.

To explore these limitations in more detail I manually inspected the mBAF plots for all cases that tested positive for chr9p mCA with and without score filter, as they have high likelihood of *JAK2* V617F by WES. Five cases were identified that had been missed by my method. All had an mBAF close to 0.9, a VAF ranging between 0.77-0.91 but a low empirical score (Table 3-19, Figure 3-19). Low scores were due to low heterozygosity rate i.e. the method excluded AI regions with very few heterozygous SNPs. Importantly, all 5 cases were in the myeloid malignancy group suggesting that it would be useful to manually inspect the mBAF plots for all chromosomes from participants in this group to identify other regions of high level aUPD that might have been missed.

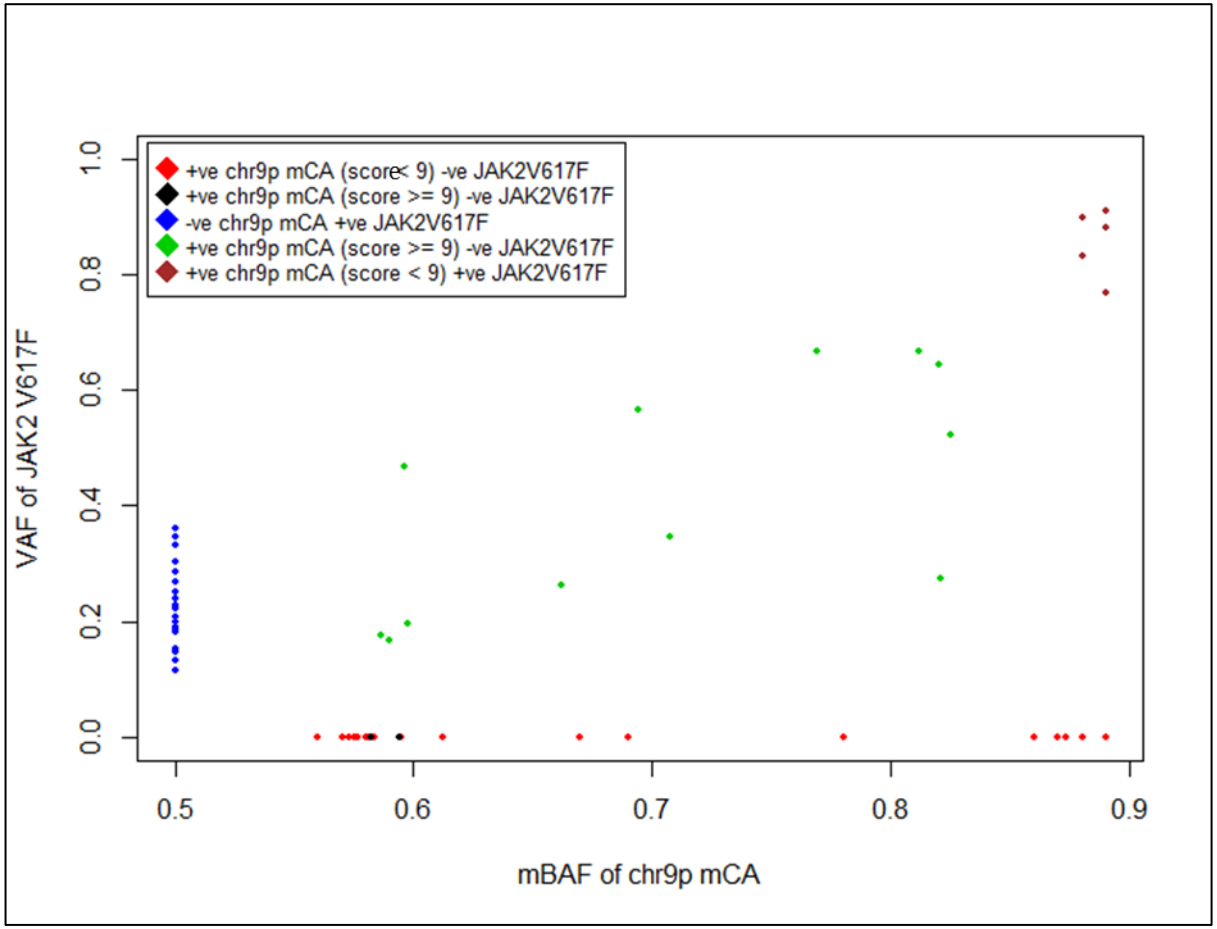


Figure 3-19: The relationship between mBAF of chr9p mCA and VAF of *JAK2* V617F

Five samples (brown), harboured *JAK2* V617F but failed the filtration criteria

Table 3-19: Five samples with chr9p aUPD and *JAK2* V617F that failed to satisfy the calling criteria

ID	Chr	Start	End	mBAF[1]	LRR[2]	Het rate*	Sum size	Total size	Coverage (sum size / total size)*	No of informative SNPs*	No of SNPs (Total)	Density	Score
3228229	9	1102943	22872228	0.88	0.02	0	21769286	21769286	1	33	7365	2955.78	0
3905325	9	1878831	33166348	0.89	0.05	0	31287518	31287518	1	34	9930	3150.81	0
2008888	9	342424	19335160	0.89	0.01	0	18992737	18992737	1	17	6795	2795.1	0
4324853	9	666119	14950193	0.89	0.01	0.01	14284075	14284075	1	32	5309	2690.54	0.32
4003272	9	334337	35415769	0.88	-0.05	0.01	35081433	35081433	1	133	11370	3085.44	1.33

* Columns used to calculate score

[1] mirrored B Allele Frequency

2 median Log R ratio

Statistical comparison between myeloid cases and cancer-free controls identified 17 regions of AI at 15 chromosomal positions. These regions can be classified into (i) highly recurrent regions of known aUPD at chr9p, chr14q, chr1p, and chr4q;(ii) less recurrent aUPD regions present in two to three samples with a myeloid malignancy involving 7q, 17p, 11q, and 22q. The highly recurrent events in myeloid malignancies are known to be associated with *JAK2* V617F [28], the imprinted *MEG3-DLK1* locus [267], mutations in *MPL* [29] and mutations in *TET2 [251],* respectively. Driver genes associated with the less recurrent events have also been identified; *EZH2* [32], *TP53* [268], *CBL* [33] and *PRR14L* [269]. I did not identify any instance of aUPD 13q, an event associated with mutations *FLT3*. This is consistent with the late role of *FLT3* in leukaemogenesis and the strong association with AML [41]. On the other hand, the most significant CNL targeted interstitial region on chr20q, a well-known abnormality in myeloid malignancies that is not fully understood but may also target imprinted genes [270]. Interestingly, CNG at chr1q and chr9q emerged as significantly associated with myeloid malignancies but the driver gene or genes is not known for these targets. Myeloid mCA were identified in 506 individuals with no past or present evidence of cancer. Importantly, 205 (40%) of them developed a haematological malignancy during the study period. The utility of the UK Biobank genetic data for predicting individuals who will develop a haematological malignancy will be explored in Chapter 5.

3.5.2 Limitations of calling driver somatic mutations:

Next, I focused on identifying myeloid clonality defined by somatic mutations by processing the variant calls in the subset of the UK Biobank with available WES data. I used strict criteria to call likely somatic mutations compared to published studies (Table 3-20), focusing on likely or known pathogenic variants in 6 genes associated with myeloid malignancies that are known to account for the great majority of instances of CH in the literature [9]. These genes play distinct roles and how they lead to clonal dominance is incompletely understood. However, the three most common genes, *DNMT3A*, *TET2* and *ASXL1,* all influence gene regulation at multiple loci by epigenetic mechanisms. *DNMT3A* is a member of DNA methyltransferase family that includes *DNMT1* and *DNMT3B*. *DNMT3A* is part of a complex that catalyses DNA methylation, which is in turn linked to downregulation of target gene expression [271]. Although most *DNMT3A* mutations in myeloid malignancies are loss of function, *DNMT3A* R882 missense variants are seen recurrently and are believed to exert a dominant negative effect, disturbing the transcriptional expression and cell-cycle regulation of haematopoietic cells [272,273]. Although *DNMT3A* R882 is a highly fit mutation, its high mutation rate may be explained by its CpG context [83]. *TET2* catalyses the oxidation of 5-methyl cytosine to 5-hydroxymethl cytosine, one of the steps in the removal of methylation marks from DNA. *In vivo*, *TET2* LOF increases self-renewal of stem cells, and promotes myeloproliferation with prominent monocytosis, and splenomegaly [274]. *ASXL1* mutations result in loss of polycomb repressive complex 2 (PRC2)-mediated histone H3 lysine 27 (H3K27) tri-methylation, which promote leukaemogenesis [57].

Table 3-20: Our study of CH in comparison to previous published studies

Study	Cohort	Variant caller	Average coverage (reads)	Age range	Frequency of CH	Definition of driver somatic mutations
[202] (This study)	WES n=49,956	WeCall	55	40-70	0.4% and 2.76% in age 40-45 and 66-79	Disruptive in *DNMT3A*, *ASXL1*, *TET2*, *PPM1D* Missense *JAK2* V617F, *DNMT3A* R882, *SRSF2* P95
[8]	WES n=12380	GATK	95	20-93	1% and 10% in age < 50, and > 65 years	Disruptive in *DNMT3A*, *ASXL1*, *TET2*, *PPM1D* Missense *JAK2*V617F, and *DNMT3A* exon 7 to 11 Other variants seen in COSMIC ≥7 times
[10]	WES n=17,182	Mutect	84	20-108	9.5% and 18.4% in age 70-79 and 90-108 years	156 genes cross referenced with COSMIC, excluding variants at the first or last 10% of the reading frame
[9]	WGS n=11,262	GATK	35.6	Median 55 Maximum 110	10% in age > 85	18 genes list from [100]; any variant seen in COSMIC ≥5 times
[169]	WGS n=97,691	Mutect2	40	Median 55 Maximum 98	4.3%	Variant in 74 genes found in COSMIC Any missense variant in *TET2* and *CBL* if pass binomial distribution test

Other CH-associated genes are more diverse in function. Missense mutations at *SRSF2* P95 are highly fit and alter the recognition of specific exonic splicing enhancer motifs that alter splicing of haematopoietic regulators [275]. PPM1D is a regulator of TP53, and its LOF mutations target the sixth exon, resulting in a C-terminal truncated protein that suppresses the DNA damage response checkpoint protein CHEK2 [66]. *JAK2* encode a non-receptor tyrosine kinase involved in cytokine and interferon signalling. The V617F mutation alter the pseudokinase domain and activates the kinase domain resulting in constitutive signalling [28].

Despite using strict criteria for selecting likely mutations, I identified a significant number of instances of clonality (n = 721) in 1.4% of the assessed samples. Thus far, my findings are characterised by (i) a restricted list of genes and (ii) a limited depth of sequencing in the UK Biobank (mean = 55x in the 6 genes resulting in the smallest VAF of 721 mutations being 0.06). Nevertheless, my findings are comparable to previous studies of similar sequencing depth, as shown in (Table 3-21).

Table 3-21: Mutations in the UK Biobank compared to a previous study of similar sequencing depth

	This study	**A previous study [8]**
Cohort	**49,956**	**12,380**
***DNMT3A* R882**	86 (0.17%)	23 (0.19%)
***DNMT3A* LOF**	223 (0.45%)	48 (0.39%)
***TET2* LOF**	223 (0.45%)	30 (0.24)
***ASXL1* LOF**	101 (0.2)	35 (0.28%)
***PPM1D* LOF**	28 (0.06%)	15 (0.12%)
***SRSF2* P95**	20 (0.04%)	5 (0.04%)
***JAK2* V617F**	40 (0.08%)	24 (0.19%)

The capture kit used in the Genovese et al study, covered the complete exons of five of the six genes of interest (*ASXL1*, *SRSF2*, *JAK2*, *PPM1D* and *DNMT3A*) but excluded exons 1 and 2 and part of exon 3 for *TET2*. Comprehensive gene level coverage was not provided by Regeneron, but assessment of aligned reads from 10 randomly chosen samples showed that 100% of the targeted regions in our genes of interest were covered at ≥20x with a mean coverage of 55x.

Initially, the UK Biobank used a pipeline called the Regeneron Seal Point Balinese (SPB) for variant calling [208] which has a duplicate read marking issue whereby all duplicates within each flow cell lane were correctly marked, but duplicates across lanes (maximum of one duplicate per unique read pair) were not marked which could lead to false variant calls (the UK Biobank communication to all users, August 2019). Of the 721 candidate driver mutations identified in my study, only 3 variants (one each in *JAK2*, *ASXL1* and *TET2*) had just 2 alternative reads (the minimum alternative allele depth used by WeCall) along with 7, 21 and 15 reference reads, respectively. I thus believe that the impact of the calling error on my study was minimal.

It is important to note the similarity in the relationship with age for CH defined by myeloid-related mCA and somatic mutations. Both showed a 1.1 fold annual increase in the incidence which clearly indicates they are similar age-related abnormalities associated with myeloid clonality. There was a small degree overlap between the two groups as shown in Table 3-22. In particular, (i) co-occurrence of mCA for JAK2/chr9p and other myeloid mCA (n=19), (ii) *TET2* is relatively often seen with other somatic mutations, most commonly with *SRSF2* (n=5, 25%). This combination has previously been noted as characteristic of CMML [276]. Indeed, monocyte percentage in all 5 cases were high, with 4 exceeding the normal range (normal range: 2% - 8%). However, there is a clear difference in terms of pathogenicity among the driver events, as 40% of the individuals with mCA had or developed haematological malignancy during the study period compared to less than 10% of participants with somatic mutations. The age relationship was seen for all mutations and, as has been noted by other investigators [41,42,262], *SRSF2* P95 mutations (n=20) were seen in participants who were relatively old, i.e. ≥60 years. This points to variation in the fitness and pathogenicity among the driver genes, and the possibility that the environment, e.g. in the bone marrow, provides different selective landscapes with ageing. Indeed, it is well known that the function of the bone marrow environment is influenced by the ageing process, and that this impacts on normal haemopoiesis [277].

Table 3-22: The co-occurrence of myeloid mCA and somatic mutations

	DNMT3A N=308	*TET2* N=208	*ASXL1* N=100	*JAK2*/chr9p N=155	*SRSF2* N=20	*PPM1D* N=28
Myeloid mCA (excluding chr9p) n=395	3	0	1	19	0	1
	DNMT3A	11	0	1	0	0
		TET2	4	3	5	0
			ASXL1	0	2	1
				JAK2/chr9p	0	1
					SRSF2	0
						PPM1D

3.5.3 Genetic risk factors of clonal haematopoiesis

By combining myeloid related mCA and putative somatic mutations, I identified 1,166 cases with myeloid related clonality cases for further analysis, however, only 21% of these developed any haematological malignancies during the study. The remaining 911 individuals are part of a focus to understand the role of myeloid CH in the pathogenesis of different diseases, as described below. Published GWAS have identified intronic variants in *TERT* be associated with CH defined by putative somatic mutations (rs34002450; intron 2; [9]; MPN and *JAK2* V617F associated CH (rs2736100; intron 3; [143,278] and (rs7705526; intron 3; [279]. Not all these SNPs were included on the array platform used by the UK Biobank, but I identified two distinct signals within *TERT* that achieved genome wide significance: rs7726159 and rs2853677. LD analysis for these signals revealed (i) rs7726159 is in LD with rs7705526 (r^2=0.79) but does not reach genome-wide significance for association with self-reported PV in the UK Biobank (P=2.5x10^{-4}; http://big.stats.ox.ac.uk); (ii) rs2853677 is not in LD with rs7705526, (r^2=0.19), but is associated with PV (P=2.9x10^{-6}; http://big.stats.ox.ac.uk); (iii) rs2736100 is in modest LD with rs7705526 (r^2=0.51) and is associated with PV (P=1x10^{-5}; http://big.stats.ox.ac.uk). Thus, it appears that variation in intron 2 (rs7726159) is associated with myeloid CH but does not predict development of MPN but variation in intron 3 (rs2736100, rs2853677 and rs7705526) does predict development of MPN. SNP rs2853677 is not in LD with any of the other variants and is thus a unique independent signal for both CH and MPN.

3.5.4 The association of *ASXL1* LOF mutations with smoking

Smokers have a known predisposition to develop CH defined by putative somatic mutations as well as AML [9,172], but no association has been reported between smoking status and acquired mCA [6]. I confirmed an association with myeloid CH and smoking and showed for the first time that this effect is predominantly, but not exclusively, associated with *ASXL1* mutations. *ASXL1* mutations have recently been associated with smoking in a large cohort of post-therapy cancer patients [85], providing support for my findings. CH has previously been associated with chemotherapy and radiotherapy [280,281] which, along with the association with ageing, suggest that a link between stress haematopoiesis and the development of clonality. Smoking is known to increase mutation rates in bronchial epithelial cells [282] and there is some evidence that smoking may increase the mutation rate in T-cells [283]. Consequently, it is conceivable that smoking preferentially induces *ASXL1* mutations, however it is perhaps more likely that smoking promotes chronic inflammation which in turn creates a suitable environment for the positive selection of *ASXL1* mutant clones, i.e. smoking alters the fitness landscape. In the context, a Mendelian randomisation study suggested smoking as a causal risk factor for CH [284]. This hypothesis is supported by finding a significant association of *ASXL1* mutations with gastritis, as 1.4% of reported cases of “gastritis of unspecified reason” had an *ASXL1* mutation, which is 5 times higher than controls, but this association is lost when smoking status is considered as a covariate. In addition, *ASXL1* is associated with an anaemia-like blood profile that is not expected in smokers. Indeed, the association remains after correction for smoking status and may relate to the fact that *ASXL1* mutations are commonly seen in MDS which in turn is characterised by anaemia. A previous study noted an increase in the incidence of C>A transversions in smokers [281] but we found the C>A transversion rate in *ASXL1* was similar in smokers (18%) compared to non-smokers (17%). Overall, C>T transitions (n=245; 66%) represented the most common single nucleotide substitution, as expected [261].

On the functional level, loss of *TET2* upregulates inflammatory mediators, including IL-6, independently from its established epigenetic role in relation to DNA methylation [285]. This may be relevant to the finding of a significant association between CH defined by *TET2* mutations and COPD [94]. We confirmed this relationship with the specific COPD class “(J44.0) Chronic obstructive pulmonary disease with acute lower respiratory infection”. Also, I found further suggestive relationships between *TET2* and chronic inflammation by its association with “agranulocytosis”, and the decrease in “basophils counts and percentage”. Also, I found a significant association between *TET2* mutations and “ulcers of lower limb”. Although the frequency of this ulcer of feet is low, it would be interesting to investigate the relationship between *TET2* and diabetes, and investigate the independent effect of *TET2* distributive mutations on wound healing in the absence of a diagnosis of diabetes [52].

3.5.5 Potential bias in the UK Biobank

A selection bias was evident toward healthy volunteers in the UK Biobank. It had a fast and efficient recruitment process, that achieved a relatively small response rate of 5.5% [286]. A large proportion (30.5%) of participants were found to have a third degree or closer relative among other cohort subjects [178], also the recruitment process was affected by the geographic distribution of Biobank assessment centres. Participants were less likely to be obese, to smoke, and to drink alcohol [287]. Furthermore, the first release of WES data incorporated 50K participants that were enriched in asthma diagnosis (ICD10= J45 or J46; 16% in comparison to 13% in all 500k participants, as well as individuals who had undergone assessments by magnetic resonance imaging [208]. It was estimated that ~ 25% of participants diagnosed with asthma had self-reported COPD defined by Global Initiative for Chronic Obstructive Lung Disease (GOLD), and it is possible that the selection bias may affect the results of studying the relationship between CH, smoking, and COPD. The newly released (2022) analysis of the UK Biobank sequence data from all participants has highlighted differences between the new exomes in comparison to the initial release of 50K exomes data. For instance, genetically predicted IL-6 showed different relationships with driver mutations, with the relationship being significant in the initial 50K exomes [288], but not in 450K exomes of all other participants [289].

3.5.6 Limitation in assessing the relationship between CH and specific CVD

Survival analysis confirmed the association between all-cause mortality in absence of haematological malignancies diagnosis during the study time. However, I did not find an association between myeloid CH and MI or stroke. The reason that the association between CH and cardiovascular disease is very prominent in some studies [10,89], but not others [9,94] is presumably explained by differences in cohort structure, follow up time, and definitions of CH. The UK Biobank had an upper recruitment age of 69 years and the follow up was only 9.1 years. My analysis is estimated to have 86% power (Figure 3-20) to detect an association between CH and MI based on a HR of 1.9, as previously reported [89] and an overall event rate of 1.4% (439/31144). My definition of CH included both chromosomal and mutational events, with a stringent definition of pathogenicity for mutations, and all abnormalities being present at a clonal fraction >10%. Clearly, as more the UK Biobank cases are sequenced and the median follow-up is extended, more associations are likely to emerge.

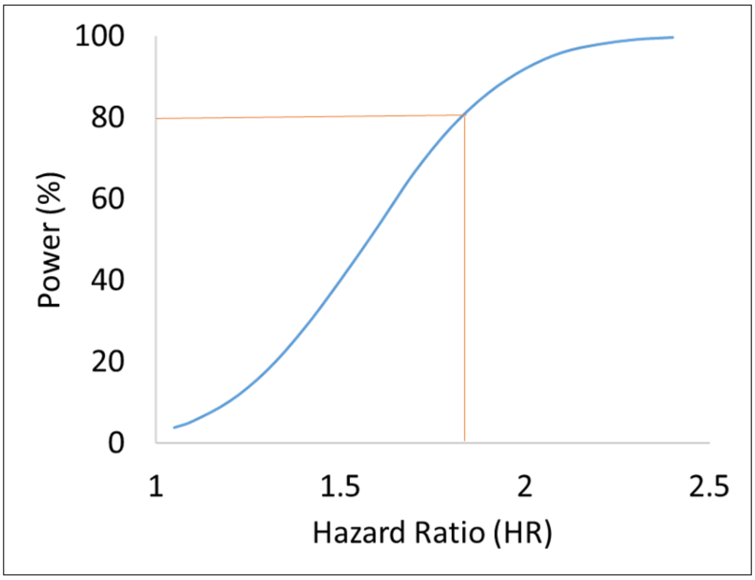


Figure 3-20: Power to detect an association between CH and MI

The plot shows the relationship between estimated power and hazard ratio (HR), under fixed type 1 error rate = 0.05; number of participants in the experimental group (cases with CH without evidence of haematological malignancy or MI prior to sampling; n=873); number of participants in control group (participants without CH and without evidence of haematological malignancy or MI prior to sampling; n = 30,271); probability of failure in experimental group (20 cases in the experimental group had a MI; P = 0.023); probability of failure in control group (419 participants in the Control group had a MI; P = 0.014). Analysis was performed using the R statistics package "powerSurvEpi". Interpolation line is added at 80% power, corresponding to HR = 1.83. My analysis is estimated to have 86% power to detect an association between CH and MI based on an HR of 1.9.

In summary, I investigated CH in the UK Biobank cohort and concluded that the risk of acquiring a myeloid associated lesion defined by mCA or driver mutation was estimated to increase by 1.1 fold per year. I found both genetic and environmental factors play an important role in the development of CH. Smoking history is strongly associated with *ASXL1* mutated CH and genetic variation at *TERT* may predispose to CH independently of predisposition to MPN. *TERT* encodes telomerase reverse transcriptase and is essential for telomere maintenance, but it also appears to function as a transcriptional co-activator [290] and impacts on the tumour microenvironment via diverse pathways, including inflammation. Chronic inflammation provides a link between genetic and environmental predisposition to CH. Myeloid CH was significantly associated with all-cause mortality, but haematological malignancies cannot explain the frequency of deaths consistent with the model that CH also impacts the pathogenesis of non-malignant diseases, mainly chronic diseases.

Chapter 4 The relationship between clonal haematopoiesis and chronic kidney disease

4.1 Summary

In the previous Chapter, I characterised myeloid-related CH and confirmed its contribution to age-related inflammation defined by smoking, serum RDW, and COPD. In this Chapter, I sought to determine the relationship between CH and chronic kidney disease (CKD). CH, defined as mCA and/or driver mutations was identified in 5,449 (2.9%) eligible the UK Biobank participants (n = 190,487 median age = 58 years). CH was negatively associated with glomerular filtration rate estimated from cystatin-C (eGFR.cys; β = −0.75, P = 2.37 × 10^{-4}), but not with eGFR estimated from creatinine, and was specifically associated with CKD defined by eGFR.cys < 60 (OR = 1.02, P = 8.44 × 10^{-8}). In participants without prevalent myeloid neoplasms, eGFR.cys was associated with myeloid mCA (n = 148, β = −3.36, P = 0.01) and somatic driver mutations (n = 3241, β = −1.08, P = 6.25 × 10^{-5}) associated with myeloid neoplasia (myeloid CH), specifically mutations in *CBL*, *TET2*, *JAK2*, *PPM1D* and *GNB1* but not *DNMT3A* or *ASXL1*. In participants with no history of cardiovascular disease or myeloid neoplasms, myeloid CH increased the risk of adverse outcomes in CKD (HR = 1.6, P = 0.002) compared to those without myeloid CH. Mendelian Randomisation (MR) analysis provided suggestive evidence for a causal relationship between CH and CKD (P = 0.03). I conclude that CH, and specifically myeloid CH, is associated with CKD defined by eGFR.cys. Myeloid CH promotes adverse outcomes in CKD, highlighting the importance of the interaction between intrinsic and extrinsic factors to define the health risk associated with CH.

4.2 Introduction

In Chapter 3, I provided lines of evidence to support the relationship between myeloid CH and chronic inflammation including smoking status [291], RDW [292], agranulocytosis, and COPD [94]. CH is associated with an elevated relative risk of developing haematological malignancies compared to age and sex matched controls without CH [293] and also an elevated risk of developing non-malignant, immune and inflammatory disorders [294,295] such as atherosclerotic cardiovascular disease (CVD)[10,89], COPD [94] and premature menopause [296]. Chronic kidney disease (CKD) is persistent kidney failure defined by low estimated glomerular filtration rate (eGFR) defined as < 60 mL/min/1.73m^2 and/or elevated urine albumin to creatinine ratio (uACR) defined as > 3mg/mmol [297]. It is a common disease but only a small minority of CKD cases progress to end stage kidney disease (ESKD), defined as eGFR<15 and/or uACR>30, and require kidney replacement therapy. Inflammation is a component of the pathogenesis of CKD and a marker of adverse outcomes that include CVD and mortality [298]. The majority of CKD cases are at an early stage of the disease process [299], which remains incompletely defined due to variation in eGFR and albuminuria measurements [300-303].

Like CH, CKD is associated with an elevated risk of CVD and mortality [304]. Atherosclerotic risk factors for CVD, such as diabetes, smoking, hypertension and dyslipidaemia, are prevalent in individuals with CKD, but there is an excess risk of CVD associated with CKD that is over and above that captured by atherosclerotic risk factors alone. In addition to sharing some risk factors, CH, CKD and CVD are characterised by persistent low-grade inflammation [305-308], however a specific relationship between CH and CKD has not been defined. In this study, I sought to assess the relationship between CH and CKD in the UK Biobank.

4.3 Methods

4.3.1 Cohort structure

I focused on participants with both genome-wide SNP array and WES data at the time of analysis (n=200,631; median age = 58y, median follow up = 11y). To investigate the relationship between CKD and either CH or myeloid neoplasia, the data were split randomly into equally sized discovery and validation cohorts. Results from the discovery and validation cohorts were combined using a fixed effects inverse variance weighted meta-analysis using STATA version 16 (StataCorp LLC, College Station, TX) and Cochran's Q test to measure heterogeneity.

4.3.2 Prevalent and incident myeloid neoplasia

Participants with myeloid malignancy were identified from the national cancer registry and hospital inpatient records using the ICD10 codes C920, C921, C923, C924, C925, C927, C929, C930, C931, C940, C944, C946, C962, D45, D460, D461, D462, D464, D467, D469, D470, D471 and D473. Myeloid malignancies were considered prevalent if diagnosed before or within one year of study (n=320) entry, or incident (n=419) if diagnosed a year or more after study entry. The relationship between CH and

ESKD in the absence of prevalent myeloid neoplasia was tested using multivariable logistic regression in R where ESKD diagnosed after the study entry was used as the dependant and CH as a binary predictor and adjusted for the same CKD risk factors.

4.3.3 Identification of CH

In the previous Chapter, I described the identification of myeloid, lymphoid or other mCA in the UK Biobank from SNP array data [202]. I expanded the definition of mutated genes, defined as myeloid-neoplasia related (‘myeloid’) according to previously published criteria [89], other genes were defined as ‘lymphoid’. The complete list of unique putative somatic driver variants (n=1,611) is shown in Supplementary Table 4-1. CH was defined as participants with any mCA and/or any somatic driver mutation; myeloid CH was defined as specific mCA events and/or a somatic driver mutation(s) that are associated with myeloid disease [89]. Lymphoid CH was defined by lymphoid mCA and/or lymphoid mutations, without myeloid mutations or myeloid mCA. To identify putative somatic driver mutations from WES data, individual gVCF files from the DeepVariant version 0.10.0 caller [265], were converted to VCF format and merged into one multi-sample VCF using SAMtools/Bcftools [309]. Multi-allelic variants were split into separate variants and the location of indels was normalised using their left most position [254]. The multisample VCF was annotated using Annovar and the RefSeq gene database [201]. Variants were defined as putative somatic driver mutations if they met the following criteria; (i) exonic or splice donor/acceptor site; (ii) the alternative allele had a minimum of 3 reads for point mutation and 6 reads for indels; (iii) alternate allele frequency ≤1% in GnomAD V2.1 [310]; (iv) predicted to be pathogenic CADD phred score >20 meaning that the variant is among the 1% most deleterious variants in the human genome [311]; (v) minor allele frequency (MAF) ≤ 0.01% in the UK Biobank; (vi) observed in COSMIC version 91 database at least 3 times in haematopoietic and lymphoid tissues [312]; (vii) inferred as somatic by failing the hypothesis that the alternative allele is normally distributed with a mean of 0.45 and a false positive rate of P=0.05 using a binomial test as described [8]. Several exceptions to these rules for defining putative somatic driver mutations were made in order to capture all relevant variants in known driver genes: (i) MAF >0.01 in the UK Biobank for *DNMT3A* R882 variants, *JAK2* V617F and *GNB1* K57E; (ii) *TP53*: all mutations seen at least once in COSMIC and validated in the International Agency for Research on Cancer database [313] (iii) *TET2*: all missense mutations in the catalytic domains (amino acids 1104-1481 and 1843- 2002); [314,315] (iv) any *DNMT3A* variant seen at least once in COSMIC; (v) all frameshift indels, stopgain, and splice site mutations in a list of known myeloid neoplasia related genes [89]. Exceptions to the binomial test were also made for established driver variants with high fitness, e.g. *U2AF1* Q157, *FLT3* Y842C, *JAK3* R657Q, *IDH2* R140L, *CBL* Y371H, and *KRAS* G12V [83]. Variants that were absent from COSMIC were only considered if they had a heterozygous “0/1” or homozygous “1/1” genotype and high genotype quality.

4.3.4 Kidney function

The eGFR in units of mL/min/1.73 m^2 was calculated in R using the Nephro package [316] and three different formulae as defined by the Chronic Kidney Disease Epidemiology Collaboration: creatinine (The UK Biobank field: 30700, eGFR.creat), cystatin-C (The UK Biobank field: 30720, eGFR.cys) or creatinine and cystatin-C (eGFR.creat.cys).[317] The creatinine based scores included ethnicity as recorded in the UK Biobank field: 21000. With respect to CKD, patients were considered as healthy (≥90), mild (≥60 and <90) moderate (≥15 <60) or end stage (<15) for each eGFR threshold.[317] In addition, uACR in mg/mmol was calculated as a further measure of kidney disease using albumin in urine (The UK Biobank field: 30500) and creatinine in urine (The UK Biobank field: 30510). Shrunken pore syndrome (SPS) is typically defined by an eGFR.cys/eGFR.creat ratio of ≤0.6 in the absence of factors that interfere with cystatin C or creatinine measurement, such as high muscle mass [318]. I used the recognised eGFR.cys/eGFR.creat ratio of ≤0.6 to define SPS.

4.3.5 The relationship between CH and CKD

To study the association between CH and CKD, I excluded 10,144 participants with (i) missing creatinine or cystatin-C data (n=9,913) or (ii) any form of ESKD (n=231) that was diagnosed before study entry according to relevant ICD10 codes (E85.3, N16.5, N18.0, N18.5, Q60.1, T82.4, T86.1, Y60.2, Y61.2, Y62.2, Y84.1, Z49.0, Z49.1, Z49.2, Z94.0, Z99.2) or interventions and procedures (OPCS4: L74.1, L74.2, L74.3, L74.4, L74.5, L74.6, L74.8, L74.9, M01.2, M01.3, M01.4, M01.5, M01.8, M01.9, M02.3, M08.4, M17.2, M17.4, M17.8, M17.9, X40.1, X40.2, X40.3, X40.4, X40.5, X40.6, X40.7, X40.8, X40.9, X41.1, X41.2, X41.8, X41.9, X42.1, X42.8, X42.9, X43.1),[319] or if any of the three eGFR scores was <15. Participants with ESKD were excluded due to the possibility of dialysis and/or erythropoietin treatment that would influence their eGFR scores and blood counts, and because the relationship between ESKD and CVD is well characterised. The relationship between CH and CKD was tested using multivariable logistic regression in R where CKD was used as the dependant and CH as a binary predictor. CKD was coded into cases (1) and controls (0) using the eGFR thresholds of <60 or ≥60 respectively and the analysis was repeated for each eGFR score (eGFR.creat, eGFR.cys, and eGFR.creat.cys). Logistic regressions were adjusted for potential confounding variables: age, sex, smoking status, systolic blood pressure, diastolic blood pressure, high density lipoprotein (HDL), low density lipoprotein (LDL), body mass index (BMI), and the first ten genetic principal components, as selected by using backward stepwise conditional ($P<0.05$) analysis. Effect sizes were reported as odds ratios (OR) with 95% confidence intervals (CI). The relationship between eGFR scores and CH were tested using multivariable linear regression in R where eGFR status was treated as the dependant and CH as a binary predictor and correcting for same confounding variables. The UK Biobank did not include follow up biochemical assessments for the great majority of participants and so incident ESKD was inferred from recorded hospital episodes as indicated above.

4.3.6 Mendelian Randomisation

MR was used to assess the possibility of a causal relationship between CH and CKD by using germline SNPs associated with the development of CH as instrumental variables. Following the STROBE guidelines [320], I investigated the use of two significance thresholds for selecting instrumental variables based on their association with CH defined by driver somatic mutations in a subset of the TOPMed cohort (n=65,405 total participants; n=3,831 CHIP cases) [169]. The first used a modest threshold ($P < 0.001$) to select 380 index SNPs after SNP clumping ($r^2 > 0.001$, within 10 Mb) with MAF ≥ 0.01 for a liberal analysis which aimed to investigate the evidence for a true null relationship. In the second, conservative, analysis I used a stricter threshold ($P < 1 \times 10^{-5}$) to select a subset of 28 index SNPs that were strongly associated with CH and would provide more robust evidence of causality. The effect sizes on CKD were obtained from a meta-analysis of 60 GWAS from the CKDgene consortium (n = 625,219, including 64,164 CKD cases [232]. I estimated that approximately ~2.4% of individuals from the TOPMed cohort are also included in the CKDgene consortium which could inflate false positive findings [321]. To mitigate against this, I performed a sensitivity analysis using the estimated effect sizes in a subset of patients from the CKDgene cohort with European ancestry (n = 480,698, including 41,395 cases). Detailed information for the SNPs used in both analyses is shown in Supplementary Table 4.2. MR was performed using the TwoSamplesMR package in R [242] to apply the Robust Adjusted Profile Score (MR-RAPS) methodology which enables the use of weak instrumental variables, is robust to pleiotropy and considers measurement error in the exposure estimate [228]. Additional sensitivity analyses were performed using methods that test the different assumptions of MR, specifically the inverse-variance weighted (IVW) method which performs a meta-analysis for the estimates of the instrumental variants [225], the MR-Egger method which uses the average pleiotropic effect as the intercept to allow the use of instrumental variables with pleiotropic effects [227], and the weighted median method which allows for a subset of instrumental variables to be invalid [226].

4.3.7 Prediction of adverse outcomes

A Cox proportional hazard model (survival package in R) [239] was used to determine if the risk of adverse outcome was associated with CH and/or CKD defined by each eGFR score or the uACR. Adverse outcomes were defined by a composite endpoint of either death (The UK Biobank data release April 2020), myocardial infarction (MI, field 40002, February 2018) or stroke (field 40006, February 2018). Participants who suffered MI or stroke before entering the UK Biobank were excluded. Follow-up times were calculated using the lubridate package [240] to determine the duration between study entry and the earliest of date of death (The UK Biobank field 40000), date of MI (The UK Biobank field 40002) or date of stroke (The UK Biobank field 40006). Patients without an adverse outcome were censored at the date of last follow-up for MI and stroke or the date they were lost to follow-up (The UK Biobank field 191). Univariate survival analyses were performed for all traditional risk factors (age, sex, smoking status, LDL, HDL, cholesterol, HbA1c, BMI, hs-CRP, systolic and diastolic blood pressure). Variables with $P < 0.2$ were entered into a multivariate survival analysis in a backward stepwise manner and retained if they reached nominal significance ($P < 0.05$).

To assess the potential for a non-linear relationship between eGFR scores and adverse outcomes, I used a restricted cubic spline function [322] to transform and segment the eGFR scores. Separate curves were fitted to each segment to generate a smooth fitted curve. The method was used to transform each eGFR score using the rms package in R [323] and default values for the number of knots ($n = 5$) and degrees of freedom ($n = 4$). The regression included the covariates described above. The adjusted spline values were plotted with 95% CI.

Receiver operating characteristic curves (ROC) and area under the curve (AUC) metrics [324] were used to evaluate the prediction accuracy of the multivariable survival models. AUCs were reported for three pairs of prediction models with and without CH: (i) traditional risk factors, (ii) traditional risk factors and eGFR.cys and (iii) traditional risk factors and uACR. Where relevant, P values for all tests were corrected for multiple testing using the false discovery rate (FDR).

4.4 Results

4.4.1 Definition and breakdown on CH in the UK Biobank

In the previous Chapter, I analysed SNP array data from the entire UK BIOBANK cohort and identified 8,203 mCA larger than 2 Mb in 5,040 participants.[202] In the subset of participants with available WES data (n= 200,631), 3,085 mCA were identified in 2,016 participants, of which 197 (185 participants) were associated with myeloid neoplasms and 278 (237 participants) were associated with lymphoid neoplasms. Analysis of the WES data identified 4,137 putative somatic driver mutations (1,611 unique variants) in 3,863 participants (Supplementary Table 4-3). In total, 5,718 (2.9%) participants had CH defined by one or more mCA and/or driver mutations and 194,913 participants were considered as CH-free controls. For further analysis, these data were split randomly into discovery and validation cohorts (Table 4-1).

Table 4-1: CH defined by both acquired mCA and/or driver somatic mutations

Participants	**Discovery cohort**					**Validation cohort**					**Total**
	Males		**Female**		**Total**	**Males**		**Females**		**Total**	
	N	**%**	**N**	**%**		**N**	**%**	**N**	**%**		
Total number	45,198	45%	55,118	55%	**100,316**	44,956	45%	55,359	55%	**100,315**	**200,631**
All CH	1286	45%	1582	55%	**2868**	1335	47%	1515	53%	**2850**	**5718**
Myeloid CH[a]	831	46%	960	54%	**1791**	885	49%	912	51%	**1797**	**3588**
Lymphoid CH[b]	135	48%	146	52%	**281**	140	46%	167	54%	**307**	**588**
All mCA	431	42%	585	58%	**1016**	439	44%	561	56%	**1000**	**2016**
Myeloid mCA	48	53%	42	47%	**90**	54	57%	41	43%	**95**	**185**
Lymphoid mCA	57	50%	56	50%	**113**	60	48%	64	52%	**124**	**237**
Other mCA	326	40%	487	60%	**813**	325	42%	456	58%	**781**	**1594**
All driver mutations	894	47%	1027	53%	**1921**	941	48%	1001	52%	**1942**	**3863**
Myeloid genes[c,d]	805	46%	933	54%	**1738**	854	49%	890	51%	**1744**	**3482**
DNMT3A	295	39%	470	61%	**765**	348	45%	419	55%	**767**	**1532**
TET2	193	46%	225	54%	**418**	188	46%	217	54%	**405**	**823**
ASXL1	103	64%	59	36%	**162**	92	64%	51	36%	**143**	**305**
JAK2	37	58%	27	42%	**64**	46	57%	35	43%	**81**	**145**
Other myeloid genes	220	54%	190	46%	**410**	225	52%	207	48%	**432**	**842**
Lymphoid genes	89	49%	94	51%	**183**	87	44%	111	56%	**198**	**381**
Control (CH-free)	43,912	45%	53,536	55%	**97,448**	43,621	45%	53,844	55%	**97,465**	**194,913**

a) 79 participants had both myeloid mutations and myeloid mCA; b) Lymphoid CH was defined by lymphoid mCA and/or lymphoid mutations, without myeloid mutations or myeloid mCA. 30 participants had both lymphoid mutations and lymphoid mCA; c) 14 participants had both myeloid and lymphoid mutations and were classed as myeloid; d) 218 participants had more than one myeloid gene mutation.

4.4.2 Assessment of the relationship between CH and CKD

I compared eGFR.cys, eGFR.creat and eGFR.creat.cys in participants with or without CH after excluding 10,144 ineligible participants with pre-existing ESKD or missing biochemistry measures. After excluding ineligible cases, the discovery cohort consisted of 2735 participants with CH and 92,457 CH-free controls, and the validation cohort compromised of 2714 participants with CH and 92,581 CH-free controls. As expected, the cystatin-C-derived eGFR score was lower than the scores that included creatine [319] and consequently fewer participants were determined to have moderate CKD, defined by an eGFR score between 15 and 60, according to eGFR.creat (n = 4194) and eGFR.creat.cys (n = 4433) compared with eGFR.cys (n = 8304). The median for all three eGFR scores was lower in participants with CH compared to those without CH (Figure 4-1) and the median uACR was higher (1.2 with CH versus 1.05 without CH; P < 0.001) indicating impairment of kidney function in association with CH. Participants with lower eGFR scores tended to be older, male, smokers, with low HDL, high LDL, high BMI, high systolic and diastolic blood pressure, and high albuminuria. (Table 4-2, Table 4-3, and Table 4-4).

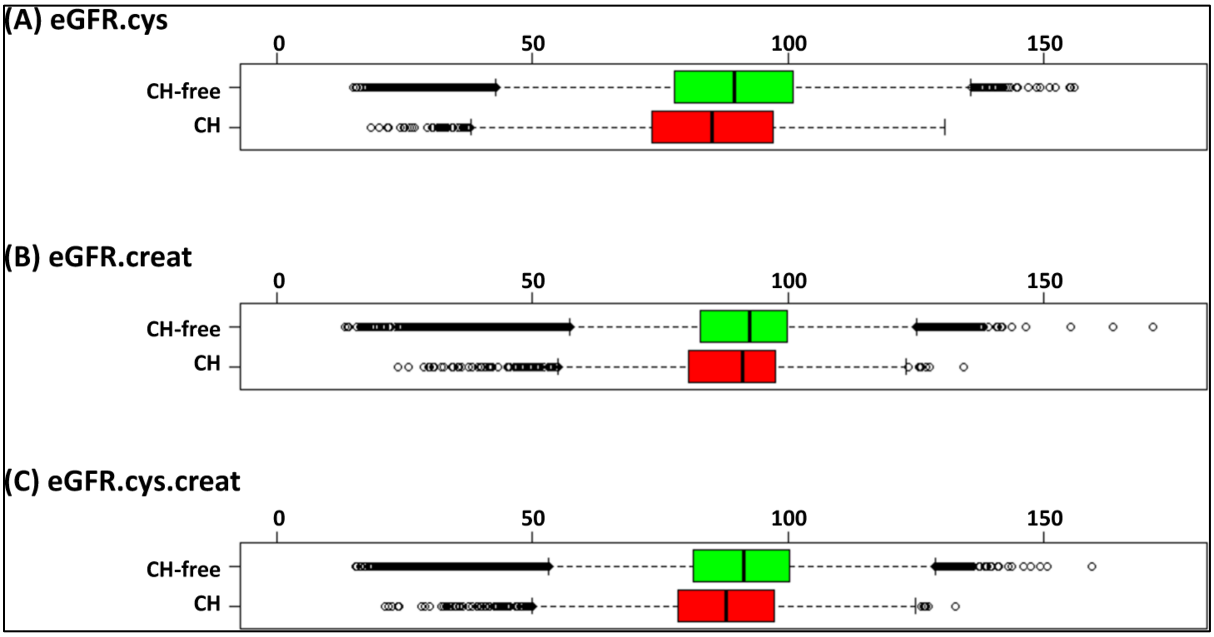


Figure 4-1: CH is associated with lower eGFR scores

Meta-analysis of discovery and validation cohorts (cases with CH, n=5,449; controls without CH, n=185,038). (A) eGFR.cys: CH, median = 84.4; CH-free, median = 88.6 (P <0.001; Mann-Whitney test), (B) eGFR.creat: CH median = 88.7; CH-free, median = 90.7 (P<0.001), (C) eGFR.creat.cys: CH, median=87.2; CH-free, median= 90.4 (P <0.001).

Table 4-2: Regression of CKD defined by eGFR.cys in the discovery cohort

		Discovery cohort						
Factor		**healthy >= 90**	**mild 60-90**	**Moderate 15-60**	**β**	**CI2.5%**	**CI97.5%**	***P***
N	**N**	46408	44638	4146	190487			
CH	**N (%)**	1060 (2.3)	1476 (3.2)	199 (4.8)	-4.01	-4.62	-3.40	<0.0001
Age	**Median**	53	61	65	-1.03	-1.04	-1.02	<0.0001
Sex	**male (%)**	19871	21131 (45.5)	1900 (45.8)	-0.79	-0.99	-0.58	<0.0001
Ethnicity*	**White (coded=1)**	43059 (92.8)	42386 (92.3)	3883 (93.7)				
	Mixed (coded=2)	423 (0.9)	211 (0.6)	17 (0.4)	6.56	5.33	7.79	<0.0001
	Asian (coded=3)	913 (2)	967 (2.1)	141 (3.4)	-1.65	-2.35	-0.94	<0.0001
	Black (coded=4)	1009 (2.17)	461 (1)	41 (0.988)	7.18	6.37	7.99	<0.0001
	Chinese (coded=5)	227 (0.5)	67 (0.1)	3 (0.1)	11.13	9.31	12.95	<0.0001
	other (coded=6)	559 (1.2)	320 (0.7)	31 (0.7)	5.13	4.09	6.17	<0.0001
	Unknown	154 (0.3)	166 (0.4)	19 (0.5)				
	No answer	14 (0.03)	20 (0.04)	2 (0.05)				
Smoking status $	**Never (coded=0)**	27308 (58.8)	23244 (50)	1848 (44.6)				
	Previous (coded=1)	15246 (32.9)	16293 (35.1)	1640 (39.6)	-2.32	-2.54	-2.11	<0.0001
	Current (coded=2)	3655 (7.9)	4856 (10.5)	623 (15)	-4.63	-4.99	-4.28	<0.0001

	No answer	149 (0.3)	205 (0.4)	26 (0.6)				
HbA1c	**Median**	34.40	35.80	37.60	-0.43	-0.45	-0.42	<0.0001
Cholesterol	**Median**	5.62	5.73	5.31	-0.02	-0.11	0.07	0.60
HDL	**Median**	1.47	1.36	1.23	7.14	6.86	7.41	<0.0001
LDL	**Median**	3.47	3.60	3.30	-0.55	-0.67	-0.43	<0.0001
uACR	**Median**	1.01	1.05	1.61	-0.10	-0.11	-0.09	<0.0001
Basophil count	**Median**	0.02	0.03	0.03	-11.77	-13.78	-9.77	<0.0001
BMI	**Median**	25.70	27.60	29.70	-0.94	-0.96	-0.92	<0.0001
systolic blood pressure	**Median**	134.00	141.00	142.00	-0.15	-0.16	-0.15	<0.0001
diastolic blood pressure	**Median**	81.00	83.00	81.00	-0.13	-0.14	-0.12	<0.0001
hs-CRP	**Median**	2.79	1.61	0.99	-0.69	-0.71	-0.67	<0.0001
Myocardial Infarction	**N (%)**	722 (1.6%)	1835 (4.1%)	411 (9.9%)	-10.80	-11.38	-10.22	<0.0001
Stroke	**N (%)**	706 (1.5%)	1271 (2.8%)	330 (8%)	-9.10	-9.76	-8.44	<0.0001
Death	**N (%)**	1365 (2.9%)	2926 (6.6%)	773 (18.6)	-11.07	-11.51	-10.62	<0.0001

* Ethnicity was encoded in integers from 1 to 6; 'white was used as a reference

$ Smoking was encoded in integers from 0 to 2; participants that never smoked were used as a reference

Table 4-3: Regression of CKD defined by eGFR.cys in the validation cohort

		Validation cohort						
		healthy >= 90	mild 60-90	Moderate 15-60	B	CI2.5%	CI97.5%	*P*
N	**N**	46575	44562	4158				
CH	**N (%)**	1060 (2.3)	1446 (3.2)	208 (5)	-4.38	-4.99	-3.77	<0.0001
Age	**median**	53	61	65	-1.02	-1.03	-1.01	<0.0001
Sex	**male (%)**	19612 (42.1)	21221 (47.6)	1885 (45.3)	-0.97	-1.17	-0.76	<0.0001
Ethnicity*	**White (coded=1)**	43276 (92.9)	42324 (95)	3875 (93.2)				
	Mixed (coded=2)	399 (0.9)	190 (0.4)	13 (0.3)	6.60	5.32	7.88	<0.0001
	Asian (coded=3)	904 (1.9)	940 (2.2)	161 (3.9)	-1.76	-2.46	-1.06	<0.0001
	Black (coded=4)	989 (2.1)	470 (1.1)	46 (1.1)	6.24	5.43	7.05	<0.0001
	Chinese (coded=5)	243 (0.5)	63 (0.1)	4 (0.1)	10.78	9.00	12.55	<0.0001
	other (coded=6)	554 (1.2)	338 (0.8)	26 (0.6)	4.57	3.53	5.60	<0.0001
	unknown	150 (0.3)	169(0.4)	21 (0.5)				
	No answer	15 (0.03)	18 (0.04)	4 (0.1)				
Smoking status $	**Never (coded=0)**	27608 (59.2)	23109 (51.9)	1878 (45.2)				
	Previous (coded=1)	15165 (32.6)	16299 (36.6)	1588 (38.2)	-2.47857	-2.69694	-2.26021	<0.0001
	Current (coded=2)	2613 (5.6)	4899 (11)	657 (15.8)	-5.09153	-5.4436	-4.73945	<0.0001

	No answer	145 (0.3)	205 (0.5)	27 (0.6)				
HbA1c	**median**	34.40	35.80	37.60	-0.41	-0.43	-0.40	<0.0001
Cholesterol	**median**	5.62	5.73	5.32	-0.07	-0.16	0.02	0.139
HDL	**median**	1.47	1.36	1.23	7.25	6.98	7.52	<0.0001
LDL	**median**	3.47	3.60	3.23	-0.62	-0.73	-0.50	<0.0001
uACR	**median**	1.02	1.04	1.57	-0.17	-0.19	-0.16	<0.0001
Basophil count	**median**	0.02	0.02	0.03	-10.89	-12.96	-8.82	<0.0001
BMI	**median**	25.70	27.60	29.90	-0.99	-1.01	-0.97	<0.0001
systolic blood pressure	**median**	134.00	141.00	143.00	-0.15	-0.16	-0.15	<0.0001
diastolic blood pressure	**median**	81.00	83.00	82.00	-0.13	-0.14	-0.12	<0.0001
hs-CRP	**median**	2.85	1.61	0.99	-0.67	-0.69	-0.65	<0.0001
Myocardial Infarction	**N (%)**	654 (1.4%)	1767 (4%)	375 (9%)	-10.92	-11.52	-10.33	<0.0001
Stroke	**N (%)**	652 (1.3%)	1328 (3%)	303 (7.3%)	-9.53	-10.20	-8.87	<0.0001
Death	**N (%)**	1358 (2.9%)	2937 (6.6%)	797 (19.2%)	-11.29	-11.74	-10.85	<0.0001

* Ethnicity was encoded in integers from 1 to 6; 'white was used as a reference

$ Smoking was encoded in integers from 0 to 2; never smokers were used as a reference

Table 4-4: Meta-analysis of the regression of CKD defined by eGFR.cys

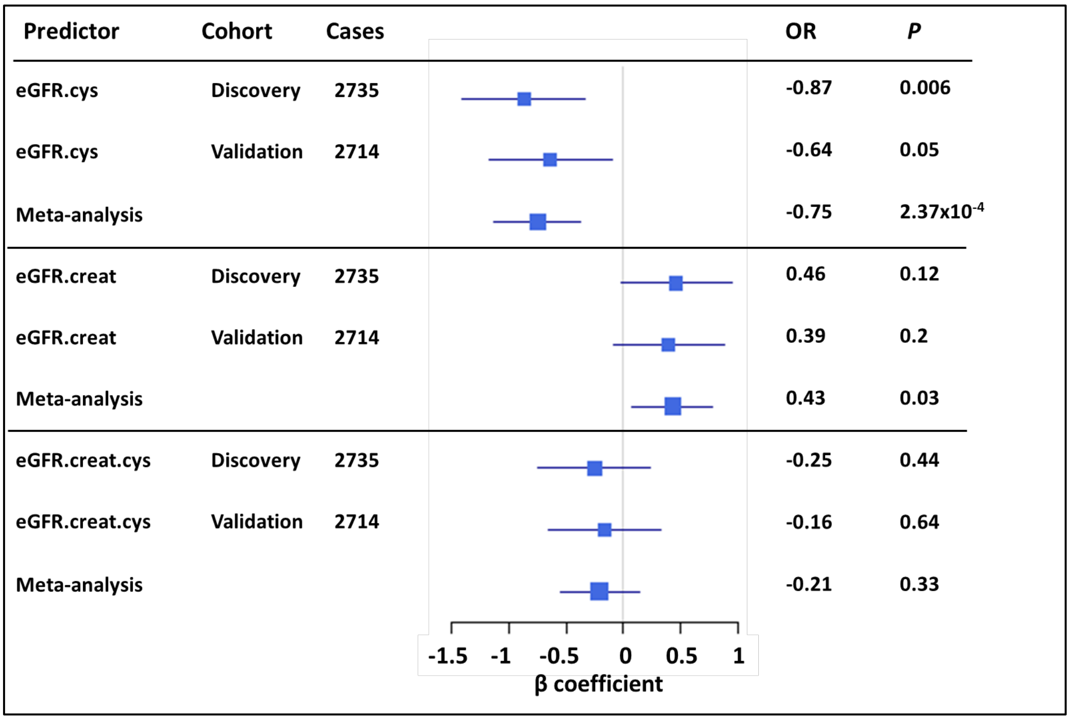
		Meta-analysis					
		Cochran's Q	***P Cochran's***	**B**	**CI2.5%**	**CI97.5%**	***P***
CH	**N (%)**	0.72	0.40	-4.19	-4.62	-3.76	3.04×10^{-81}
Age	**Median**	0.20	0.65	-1.03	-1.03	-1.02	$<1.0\times 10^{-300}$
Sex	**male (%)**	1.48	0.22	-0.88	-1.02	-0.73	1.00×10^{-32}
Ethnicity*	**White (coded=1)**						
	Mixed (coded=2)	0.00	0.96	6.58	5.69	7.46	5.31×10^{-48}
	Asian (coded=3)	0.05	0.83	-1.70	-2.20	-1.21	1.96×10^{-11}
	Black (coded=4)	2.58	0.11	6.71	6.14	7.29	2.11×10^{-116}
	Chinese (coded=5)	0.08	0.78	10.95	9.68	12.22	4.60×10^{-64}
	other (coded=6)	0.56	0.46	4.846	4.11	5.58	2.92×10^{-38}
Smoking status $	**Never (coded=0)**						
	Previous (coded=1)	0.95	0.33	-2.40	-2.56	-2.25	2.26×10^{-203}
	Current (coded=2)	3.26	0.07	-4.86	-5.11	-4.61	$<1.0\times 10^{-300}$
HbA1c	**Median**	3.40	0.07	-0.42	-0.43	-0.41	$<1.0\times 10^{-300}$
Cholesterol	**Median**	0.45	0.50	-0.05	-0.11	0.02	0.16
HDL	**Median**	0.34	0.56	7.20	7.00	7.39	$<1.0\times 10^{-300}$

LDL	**Median**	0.59	0.44	-0.58	-0.67	-0.50	4.48×10^{-43}
uACR	**Median**	66.50	0.00	-0.12	-0.13	-0.11	1.29×10^{-164}
Basophil count	**Median**	0.36	0.55	-11.35	-12.79	-9.90	9.77×10^{-54}
BMI	**Median**	8.09	$4.00x10^{-3}$	-0.97	-0.98	-0.95	$<1.0\times 10^{-300}$
systolic blood pressure	**Median**	0.01	0.92	-0.15	-0.16	-0.15	$<1.0\times 10^{-300}$
diastolic blood pressure	**Median**	0.00	0.95	-0.13	-0.14	-0.12	1.25×10^{-294}
hs-CRP	**Median**						
Myocardial Infarction	**N (%)**	0.08	0.77	-10.86	-11.28	-10.44	$<1.0\times 10^{-300}$
Stroke	**N (%)**	0.83	0.36	-9.32	-9.78	-8.85	$<1.0\times 10^{-300}$
Death	**N (%)**	0.49	0.48	-11.18	-11.50	-10.86	$<1.0\times 10^{-300}$

* Ethnicity was encoded in integers from 1 to 6; 'white was used as a reference

$ Smoking was encoded in integers from 0 to 2; never smokers were used as a reference

To determine the association between CH and CKD, I performed logistic and linear regression analyses where CKD was coded as either a binary (1 = moderate CKD eGFR > 15 and <60, 0 = eGFR ≥60) or as a continuous trait based on each eGFR score and adjusted for potential confounding variables (Table 4-5). In the logistic models, CH was associated with an increased risk of moderate CKD estimated from cystatin-C scores (eGFR.cys, OR = 1.02 [95% CI: 1.01–1.02], P = 8.44 × 10^{-8}). A weaker association was observed for eGFR.creat.cys (OR = 1.01 [95% CI: 1.00–1.01], P = 0.04) and there was no association with eGFR.creat (OR = 1.00 [95% CI: 0.995–1.004], P = 0.93), (Table 4-6). Similar results were obtained from linear regression analysis where eGFR scores estimated from cystatin-C were negatively associated with CH in the discovery, validation, and meta-analysis (eGFR.cys, β = −0.75, P = 2.37 × 10^{-4}) but not eGFR.creat.cys (β = −0.21, P = 0.33), or eGFR.creat (β = 0.43, P = 0.03, not significant in the discovery and validation cohorts) (Figure 4-2). For all tests there was no evidence for heterogeneity between the discovery and validation cohorts (P > 0.05, Cochran's Q test).



Predictor	Cohort	Cases	OR	*P*
eGFR.cys	Discovery	2735	-0.87	0.006
eGFR.cys	Validation	2714	-0.64	0.05
Meta-analysis			-0.75	2.37x10^{-4}
eGFR.creat	Discovery	2735	0.46	0.12
eGFR.creat	Validation	2714	0.39	0.2
Meta-analysis			0.43	0.03
eGFR.creat.cys	Discovery	2735	-0.25	0.44
eGFR.creat.cys	Validation	2714	-0.16	0.64
Meta-analysis			-0.21	0.33

Figure 4-2: CH is specifically and negatively associated with eGFR estimated from cystatin-C

eGFR.cys: eGFR estimated from cystatin-C, eGFR.creat: eGFR estimated from creatinine, eGFR.creat.cys: estimated from both creatinine and cystatin-C. Square sizes represent the precision of each eGFR score.

Table 4-5: Initial risk factors identified by linear regression model for eGFR.cys

	Discovery cohort				Validation cohort			
	β	**CI2.5%**	**CI97.5%**	***P***	**β**	**CI2.5%**	**CI97.5%**	***P***
age	-1.03	-1.04	-1.02	$< 2 \times 10^{-16}$	-1.02	-1.03	-1.01	$< 2 \times 10^{-16}$
sex	1.93	1.73	2.14	$< 2 \times 10^{-16}$	1.85	1.65	2.05	$< 2 \times 10^{-16}$
Smoking status	-1.36	-1.49	-1.23	$< 2 \times 10^{-16}$	-1.44	-1.57	-1.31	$< 2 \times 10^{-16}$
Diastolic blood pressure	-0.03	-0.04	-0.01	3.3×10^{-5}	-0.02	-0.03	-0.01	1.06×10^{-3}
Systolic blood pressure	0.03	0.02	0.04	$< 2 \times 10^{-16}$	0.03	0.02	0.03	2.88×10^{-15}
Cholesterol	-2.45	-2.88	-2.01	$< 2 \times 10^{-16}$	-2.51	-2.94	-2.08	$< 2 \times 10^{-16}$
HDL	7.48	7.05	7.91	$< 2 \times 10^{-16}$	7.54	7.11	7.97	$< 2 \times 10^{-16}$
LDL	2.71	2.18	3.25	$< 2 \times 10^{-16}$	2.74	2.21	3.28	$< 2 \times 10^{-16}$
HbA1c	0.02	0.00	0.03	0.03	0.01	0.00	0.03	0.04
BMI	-0.67	-0.69	-0.65	$< 2 \times 10^{-16}$	-0.69	-0.71	-0.67	$< 2 \times 10^{-16}$
hs-CRP	-0.30	-0.32	-0.28	$< 2 \times 10^{-16}$	-0.30	-0.32	-0.28	$< 2 \times 10^{-16}$

Table 4-6: Logistic regression between CH and CKD coded as a binary variable

	Discovery cohort					Validation cohort					Meta analysis					
Outcome defining score	**Cases**	**OR**	**CI 25%**	**CI 97.5%**	***P***	**Cases**	**OR**	**CI 25%**	**CI 97.5%**	***P***	**Cochrans' Q**	***P* Cochran's**	**OR**	**CI 25%**	**CI 97.5%**	***P***
eGFR creat	2735	1.00	0.99	1.00	0.66	2714	1.00	1.00	1.01	0.67	0.50	0.48	1.00	1.00	1.00	0.93
eGFR cys	2735	1.02	1.01	1.02	1.52×10^{-03}	2714	1.02	1.01	1.03	1.10×10^{-4}	0.30	0.58	1.02	1.01	1.02	8.44×10^{-8}
eGFR creat.cys	2735	1.00	1.00	1.01	0.39	2714	1.01	1.00	1.01	0.09	0.45	0.50	1.01	1.00	1.01	0.04

To investigate the relationship between CH and CKD in more detail, I tested the constituent components of CH for association with eGFR.cys as a continuous trait using linear regression. Low eGFR.cys scores were associated with myeloid mCA ($\beta = -4.44$, $P = 8.90 \times 10^{-5}$) but not lymphoid mCA ($\beta = -1.7$, $P = 0.12$) or 'other' mCA ($\beta = 0.61$, $P = 0.15$). Alterations involving chr9p were the most strongly associated subtype of myeloid mCA ($\beta = -8.06$, $P = 8.80 \times 10^{-5}$). For CH defined by somatic mutations, myeloid neoplasia-associated genes were strongly associated with lower levels of eGFR.cys ($\beta = -1.33$, $P = 5.52 \times 10^{-7}$), whereas lymphoid genes were not significant ($\beta = -1.31$, $P = 0.125$). At the gene level, the relationship was significant for CH defined by *JAK2* ($n = 139$, $\beta = -1.03$, $P < 1 \times 10^{-300}$) and *TET2* ($n = 788$, $\beta = -1.94$, $P = 4.50 \times 10^{-4}$) variants but not variants in *DNMT3A* or *ASXL1*. Again, for all tests there was no evidence for heterogeneity between the discovery and validation cohorts ($P > 0.05$, Cochran's Q test). Full results for the discovery and validation cohorts are presented in Table 4-7, Table 4-8, and Table 4-9.

The median VAF of CH defined by myeloid neoplasia associated genes was higher in participants with CKD (eGFR < 60) defined by eGFR.cys (median VAF = 0.24) compared to other participants (eGFR ≥ 60) (median VAF = 0.21, $P = 1.71 \times 10^{-7}$) but no difference was seen for CKD defined by eGFR.creat (median VAF = 0.23 vs. 0.21, $P = 0.12$) (Figure 4-3). At the level of individual genes, a significant difference was only seen for *JAK2* with a median VAF of 0.56 in cases with CKD defined by eGFR.cys compared to other participants (VAF = 0.20, $P = 4.70 \times 10^{-6}$).

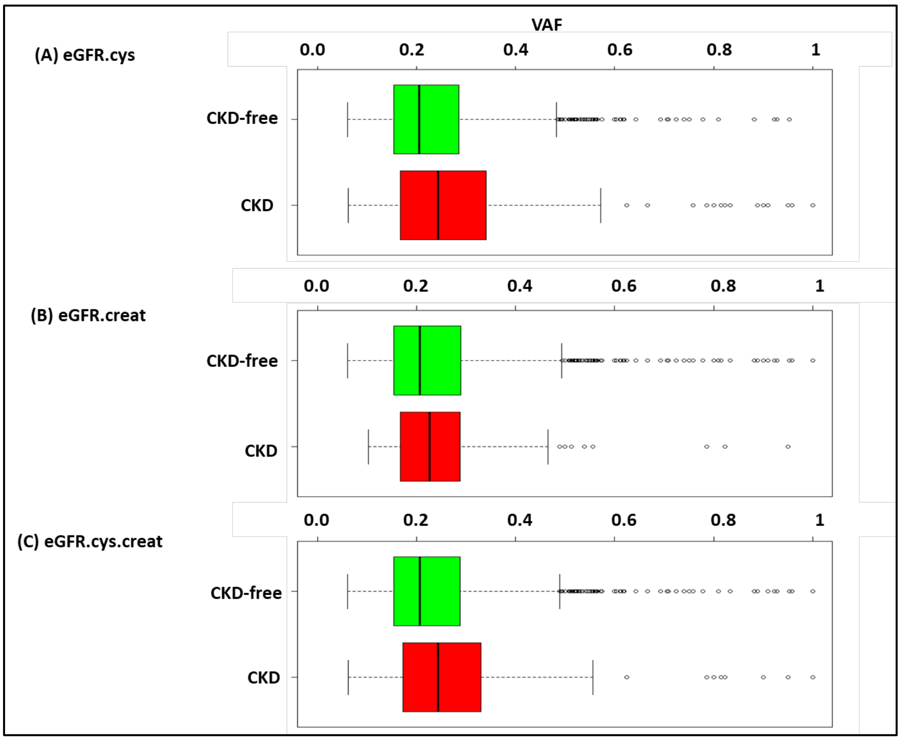


Figure 4-3: The relationship between eGFR scores and VAF

CKD, defined by eGFR <60, is associated with VAF of driver mutations in myeloid related genes in 3,328 participants. Meta-analysis of discovery and validation cohorts (A) eGFR.cys: CKD (n=293), median = 0.24; CKD-free (n=3,035), median = 0.21 (P = 1.71x10^{-7}; Mann-Whitney test), (B) eGFR.creat: CKD (n=111), median = 0.23; CKD-free (n= 3,217), median = 0.21 (P=0.12), (C) eGFR.creat.cys: CKD (n=144), median=0.23; CKD-free (n=3,184), median= 0.21 (P = 2x10^{-4}).

The link between myeloid neoplasms and reduced kidney function is well established and was replicated in a subset of the UK BIOBANK participants which included 320 participants with a prevalent myeloid neoplasm (diagnosed before or within a year of study entry) that was associated with lower eGFR.cys score (β = −5.22, P = 7.77 × 10^{-10}). Excluding these cases, eGFR.cys was still associated with myeloid CH (n = 3,330, β = −1.05, P = 8.80 × 10^{-5}), including both myeloid mCA (n = 148, β = −3.36, P = 0.01) and myeloid related-genes (n = 3241, β = −1.08, P = 6.25 × 10^{-5}). Stratification at the gene level identified associations between eGFR.cys and mutations in *CBL*, *TET2*, *JAK2*, *PPM1D* and, to a lesser degree, *GNB1* (Table 4-10) assesses the relationship between myeloid CH and the risk of developing ESKD in participants without prevalent myeloid neoplasms or prior ESKD. Myeloid CH (n = 3330) was weakly but significantly associated with ESKD incidence (n = 307, β = 0.002, P = 0.006).

Specifically, 0.33% (11 out of 3330) of participants with myeloid CH developed ESKD after study entry compared with 0.16% of controls (296 of 184,811).

Table 4-7: Association between CH and eGFR scores as continuous variables in the discovery cohort

		Discovery cohort				
Outcome defining score	**Predictor**	**Cases**	**β**	**CI 25%**	**CI 97.5%**	***P***
eGFR.creat	**CH**	2735	0.46	-0.02	0.95	0.12
eGFR.cys	**CH**	2735	-0.87	-1.41	-0.34	$6.22\text{x}10^{-3}$
eGFR.creat.cys	**CH**	2735	-0.25	-0.74	0.23	0.45
eGFR.cys	**Somatic mutations**	1829	-1.42	-2.08	-0.77	$2.22\text{x}10^{-4}$
eGFR.cys	**Somatic mutations (myeloid genes)**	1658	-1.41	-2.09	-0.72	$4.64\text{x}10^{-4}$
eGFR.cys	**Somatic mutations (lymphoid genes)**	171	-1.58	-3.71	0.55	0.26
eGFR.cys	***DNMT3A***	731	-0.40	-1.41	0.61	0.58
eGFR.cys	***TET2***	398	-2.21	-3.63	-0.78	$9.38\text{x}10^{-3}$
eGFR.cys	***ASXL1***	152	-0.93	-3.17	1.31	0.57
eGFR.cys	***JAK2***	62	-1.03	-1.04	-1.02	0.00
eGFR.cys	**somatic mutations other myeloid**	396	-3.11	-4.50	-1.73	$1.42\text{x}10^{-4}$
eGFR.cys	**any mCA >= 2Mb**	973	-0.51	-1.40	0.38	0.40
eGFR.cys	**mCA associated with myeloid and lymphoid**	198	-3.66	-5.64	-1.68	$1.90\text{x}10^{-3}$

eGFR.cys	**any mCA >= 2Mb (exclude myeloid and lymphoid)**	775	0.28	-0.72	1.27	0.70
eGFR.cys	**myeloid mCA**	86	-4.22	-7.14	-1.31	0.02
eGFR.cys	**mCA (lymphoid)**	112	-3.18	-5.88	-0.48	0.05
eGFR.cys	**chr9p mCA**	22	-6.94	-12.61	-1.26	0.05
eGFR.cys	**mCA (myeloid excluding chr9p)**	64	-3.25	-6.65	0.15	0.12
eGFR.cys	**myeloid CH**	1704	-1.35	-2.03	-0.68	6.68x10^{-4}
eGFR.cys	**Myeloid CH in prevalent myeloid malignancies**	37	-10.20	-15.05	-5.35	3.38x10^{-4}
eGFR.cys	**Myeloid CH in no prevalent myeloid malignancies**	1667	-1.18	-1.86	-0.50	3.10 x 10^{-3}
	CH free (Base line)	92457				
eGFR.cys	**Myeloid malignancies**	162	-5.63	-7.78	-3.47	7.17x10^{-6}
	Control	95030				0.12

Table 4-8: Association between CH and eGFR scores as continuous variables in the validation cohort

		Validation cohort				
Outcome defining score	**Predictor**	**Cases**	**β**	**CI 25%**	**CI 97.5%**	***P***
eGFR.creat	**CH**	2714	0.39	-0.09	0.88	0.21
eGFR.cys	**CH**	2714	-0.64	-1.17	-0.10	0.05
eGFR.creat.cys	**CH**	2714	-0.16	-0.65	0.33	0.64
eGFR.cys	**Somatic mutations**	1857	-1.22	-1.87	-0.58	$1.42\text{x}10^{-3}$
eGFR.cys	**Somatic mutations (myeloid genes)**	1670	-1.24	-1.92	-0.56	$1.97\text{x}10^{-3}$
eGFR.cys	**Somatic mutations (lymphoid genes)**	187	-1.08	-3.05	0.88	0.42
eGFR.cys	***DNMT3A***	729	0.71	-0.31	1.73	0.28
eGFR.cys	***TET2***	390	-1.67	-3.07	-0.27	0.05
eGFR.cys	***ASXL1***	138	-2.73	-5.03	-0.43	0.05
eGFR.cys	***JAK2***	77	-1.02	-1.03	-1.01	0.00
eGFR.cys	**somatic mutations other myeloid**	418	-3.30	-4.65	-1.96	$2.24\text{x}10^{-5}$
eGFR.cys	**any mCA >= 2Mb**	945	0.26	-0.64	1.16	0.69
eGFR.cys	**mCA associated with myeloid and lymphoid**	207	-2.20	-4.11	-0.30	0.06
eGFR.cys	**any mCA >= 2Mb (exclude myeloid and lymphoid)**	738	0.96	-0.05	1.98	0.12

eGFR.cys	**myeloid mCA**	90	-4.65	-7.57	-1.72	7.58x10^{-3}
eGFR.cys	**mCA (lymphoid)**	117	-0.42	-2.92	2.08	0.81
eGFR.cys	**chr9p mCA**	28	-8.91	-13.84	-3.98	1.97x10^{-3}
eGFR.cys	**mCA (myeloid excluding chr9p)**	62	-2.34	-5.97	1.30	0.32
eGFR.cys	**myeloid CH**	1709	-1.24	-1.91	-0.56	1.90x10^{-3}
eGFR.cys	**Myeloid CH in prevalent myeloid malignancies**	46	-12.27	-16.17	-8.37	2.06x10^{-8}
eGFR.cys	**Myeloid CH in no prevalent myeloid malignancies**	1663	-0.91	-1.59	-2.28× 10–01	0.03
	CH free (Base line)	92581				
eGFR.cys	**Myeloid malignancies**	158	-4.77	-7.02	-2.51	3.38x10^{-4}
	Control	95137				

Table 4-9: Meta analysis of the association between CH and eGFR scores as continuous variables

		Meta analysis					
Outcome defining score	**Predictor**	**Cochrans' Q**	***P*$_{Cochran's}$**	**β**	**CI2.5%**	**CI97.5%**	***P***
eGFR.creat	**CH**	0.04	0.85	0.43	0.08	0.78	0.03
eGFR.cys	**CH**	0.39	0.54	-0.75	-1.13	-0.38	2.37x10^{-4}
eGFR.creat.cys	**CH**	0.07	0.79	-0.21	-0.55	0.14	0.33
eGFR.cys	**Somatic mutations**	0.18	0.67	-1.32	-1.78	-0.86	1.35x10^{-7}
eGFR.cys	**Somatic mutations (myeloid genes)**	0.12	0.73	-1.33	-1.81	-0.84	5.52x10^{-7}
eGFR.cys	**Somatic mutations (lymphoid genes)**	0.11	0.74	-1.31	-2.76	0.13	0.13
eGFR.cys	***DNMT3A***	2.29	0.13	0.15	-0.57	0.88	0.73
eGFR.cys	***TET2***	0.27	0.60	-1.94	-2.94	-0.93	4.50x10^{-4}
eGFR.cys	***ASXL1***	1.21	0.27	-1.81	-3.41	-0.21	0.05
eGFR.cys	***JAK2***	0.50	0.48	-1.03	-1.04	-1.01	0.00
eGFR.cys	**somatic mutations other myeloid**	0.04	0.85	-3.21	-4.18	-2.24	1.01x10^{-9}
eGFR.cys	**any mCA >= 2Mb**	1.44	0.23	-0.13	-0.76	0.50	0.73
eGFR.cys	**mCA associated with myeloid and lymphoid**	1.08	0.30	-2.90	-4.28	-1.53	1.07x10^{-4}
eGFR.cys	**any mCA >= 2Mb (exclude myeloid and lymphoid)**	0.88	0.35	0.61	-0.10	1.33	0.15

eGFR.cys	**myeloid mCA**	0.04	0.84	-4.44	-6.50	-2.37	8.90x10^{-5}
eGFR.cys	**mCA (lymphoid)**	2.16	0.14	-1.70	-3.54	0.14	0.12
eGFR.cys	**chr9p mCA**	0.27	0.61	-8.06	-11.78	-4.35	8.80x10^{-5}
eGFR.cys	**mCA (myeloid excluding chr9p)**	0.13	0.72	-2.82	-5.30	-0.35	0.05
eGFR.cys	**myeloid CH**	0.05	0.82	-1.29	-1.77	-0.82	5.47x10^{-7}
eGFR.cys	**Myeloid CH in prevalent myeloid malignancies**	0.44	0.51	-11.47	-14.51	-8.44	2.97x10^{-12}
eGFR.cys	**Myeloid CH in no prevalent myeloid malignancies**	0.30	0.59	-1.05	-1.53	-0.56	8.90x10^{-5}
	CH free (Base line)						
eGFR.cys	**Myeloid malignancies**	0.29	0.59	-5.22	-6.78	-3.66	7.77x10^{-10}
	Control						

Table 4-10: Association between myeloid CH and eGFR.cys in the absence of prevalent myeloid neoplasia

		Discovery cohort					Validation cohort					Meta analysis					
Outcome defining score	**Predictor**	**Cases**	**β**	**CI2.5%**	**CI97.5%**	***P***	**Cases**	**β**	**CI2.5%**	**CI97.5%**	***P***	**Cochrans' Q**	***P* Cochran's**	**β**	**CI2.5%**	**CI97.5%**	***P***
eGFR.cys	**myeloid CH**	1667	-1.19	-1.87	-0.51	2.95×10^{-3}	1663	-0.91	-1.60	-0.23	0.03	0.31	0.58	-1.051	-1.54	-0.57	8.80x10^{-5}
eGFR.cys	**myeloid mCA**	76	-3.32	-6.46	-0.18	0.09	72	-3.40	-6.75	-0.05	0.10	0	0.972	-3.358	-5.647	-1.068	9.11x10^{-3}
eGFR.cys	**myeloid genes**	1619	-1.25	-1.94	-0.56	1.97×10^{-3}	1622	-0.91	-1.60	-0.22	0.03	0.48	0.489	-1.081	-1.566	-0.596	6.25x10^{-5}
eGFR.cys	***DNMT3A***	725	-0.39	-1.41	0.62	0.58	721	0.70	-0.33	1.73	0.30	2.16	0.14	0.14	-0.59	0.87	0.73
eGFR.cys	***TET2***	393	-1.91	-3.35	-0.48	0.03	385	-1.57	-2.98	-0.16	0.07	0.11	0.74	-1.74	-2.74	-0.73	1.84x10^{-3}
eGFR.cys	***ASXL1***	150	-0.83	-3.09	1.43	0.61	133	-2.41	-4.74	-0.08	0.09	0.92	0.338	-1.59	-3.21	0.03	0.09
eGFR.cys	***JAK2***	42	-4.23	-8.47	0.00	0.10	50	-5.07	-8.97	-1.18	0.03	0.08	0.78	-4.69	-7.56	-1.82	3.21x10^{-3}
eGFR.cys	***GNB1***	39	-4.22	-8.39	-0.04	0.10	47	-2.86	-6.85	1.14	0.27	0.21	0.65	-3.51	-6.40	-0.62	0.04
eGFR.cys	***SRSF2***	33	-3.65	-8.59	1.29	0.26	34	-1.14	-6.17	3.88	0.77	0.49	0.49	-2.415	-5.935	1.105	0.27
eGFR.cys	***TP53***	30	-0.18	-4.93	4.58	0.96	32	0.15	-4.42	4.72	0.96	0.01	0.92	-0.01	-3.30	3.29	1.00
eGFR.cys	***PPM1D***	33	-6.35	-11.29	-1.41	0.04	28	-5.35	-10.49	-0.22	0.09	0.07	0.79	-5.87	-9.43	-2.31	3.08x10^{-3}
eGFR.cys	***SF3B1***	22	-2.11	-8.29	4.07	0.63	30	-2.55	-7.57	2.48	0.46	0.01	0.91	-2.37	-6.27	1.52	0.32

eGFR.cys	***FLT3***	22	1.05	-5.33	7.43	0.81	14	-2.96	-10.39	4.46	0.58	0.65	0.42	-0.65	-5.49	4.19	0.81
eGFR.cys	***GNAS***	20	1.26	-4.57	7.08	0.77	13	2.84	-4.27	9.94	0.58	0.11	0.74	1.89	-2.61	6.40	0.49
eGFR.cys	***NF1***	12	-0.39	-8.20	7.42	0.96	16	-1.48	-8.06	5.10	0.77	0.04	0.83	-1.03	-6.07	4.01	0.73
eGFR.cys	***CBL***	10	-3.92	-12.65	4.81	0.53	18	-16.79	-23.36	-10.21	0.00	5.32	0.02	-12.14	-17.40	-6.88	3.44x10^{-5}
eGFR.cys	***STAG2***	14	6.95	0.10	13.80	0.10	13	-0.48	-8.68	7.73	0.96	1.85	0.17	3.90	-1.37	9.16	0.23
eGFR.cys	***PRPF40B***	12	0.32	-7.92	8.55	0.96	14	4.77	-2.05	11.60	0.28	0.67	0.41	2.96	-2.29	8.21	0.35
eGFR.cys	***CREBBP***	9	-0.52	-9.86	8.83	0.96	15	-4.38	-10.98	2.21	0.30	0.44	0.51	-3.10	-8.49	2.30	0.34
eGFR.cys	***KDM6A***	11	1.43	-7.31	10.16	0.81	11	-5.93	-13.71	1.86	0.24	1.52	0.22	-2.68	-8.49	3.14	0.45
eGFR.cys	***BRCC3***	15	-2.65	-10.10	4.80	0.61	6	-0.01	-10.06	10.04	1.00	0.17	0.68	-1.72	-7.70	4.27	0.66
eGFR.cys	***IDH2***	7	2.28	-8.76	13.33	0.77	7	-15.29	-27.59	-2.98	0.04	4.33	0.037	-5.56	-13.78	2.67	0.28
eGFR.cys	***KMT2D***	4	-6.75	-19.10	5.61	0.42	9	-1.78	-10.48	6.92	0.77	0.42	0.52	-3.43	-10.54	3.69	0.44
Control		92335					92476										
ESKD	**myeloid CH**	5	0.002	-0.001	0.004	0.210	6	0.002	0.000	0.005	0.06	0.33	0.567	0.002	0.001	0.003	6.00x10^{-3}
	control	158					140										

4.4.3 MR analysis to test causal effect of CH on kidney function

The possibility of a causal relationship between CH and kidney function was assessed using MR. In a liberal analysis, 380 independent SNPs associated with CH at (P < 0.001) [169] were used to estimate the effect of CH on CKD (Supplementary Table 4-2). To test the different assumptions and scenarios, several MR methods were used as recommended and the results corrected for multiple testing [321]. Only the MR-RAPS method, which is adapted to test weak instrumental variables as applicable to my study, identified a positive causal relationship [OR = 1.01; P = 0.029]. However, this relationship failed to reach significance (P = 0.81) in a more conservative analysis that applied stricter threshold ($P < 1 \times 10^{-5}$) to select 28 SNPs associated with CH (Figure 4-4). Due to the potential limited overlap between cohorts used to select instrumental variables, I performed a sensitivity analysis using a subset of samples with European American ancestry which yielded similar results for the causal association between CH and CKD [OR = 1.02; P = 0.029]. Detailed results are presented in Table 4-11.

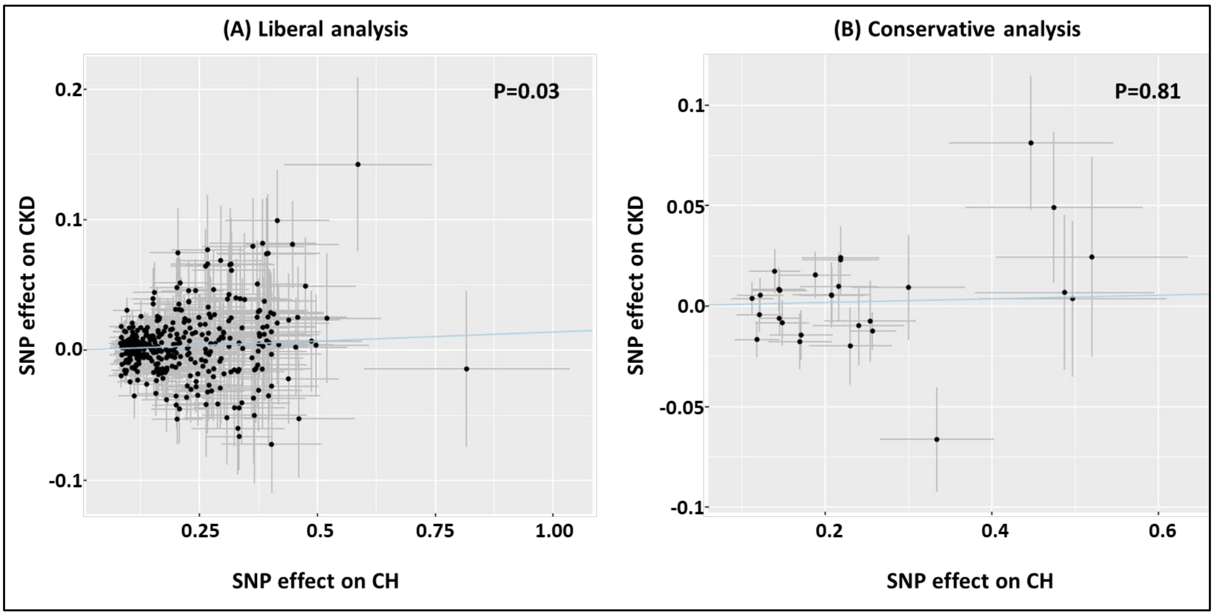


Figure 4-4: The relationship between CH and CKD using mendelian randomisation methods

MR using robust adjusted profile score (MR-RAPS) to estimate the effect of SNPs associated with CH against their effect in relation to CKD. (A) Liberal analysis using 380 independent SNPs associated with CH at P <0.001. The MR-RAPS test estimated a significant positive effect of CH on CKD (OR=1.014, CI 95%:1.003-1.024; P=0.03). B) Conservative analysis using 28 SNPs associated with CH at $P < 1x10^{-5}$. The line of regression is indicated in blue and the axes show β coefficients for SNP effects on CH and CKD.

Table 4-11: Mendelian randomisation results adjusted for multiple tests by false discovery rate

CKD Population	**p value cut-off**	**MR method**	**n SNPs**	**β**	**se**	***P***
All	1.00×10^{-3}	RAPS	380	0.01	0.01	**0.029**
European American	1.00×10^{-3}	RAPS	369	0.02	0.01	**0.029**
All	1.00×10^{-3}	IVW	380	0.01	0.00	0.063
European American	1.00×10^{-3}	IVW	369	0.01	0.01	0.079
All	1.00×10^{-3}	MR Egger	380	0.02	0.01	0.115
European American	1.00×10^{-3}	MR Egger	369	0.03	0.01	0.115
All	1.00×10^{-3}	Weighted median	380	0.01	0.01	0.131
European American	1.00×10^{-3}	Weighted median	369	0.01	0.01	0.292
All	1.00×10^{-5}	RAPS	28	0.01	0.02	0.806
European American	1.00×10^{-5}	RAPS	28	0.00	0.02	0.912
All	1.00×10^{-5}	IVW	28	0.03	0.04	0.624
European American	1.00×10^{-5}	IVW	28	-0.02	0.05	0.624
All	1.00×10^{-5}	MR Egger	28	0.01	0.02	0.744
European American	1.00×10^{-5}	MR Egger	28	0.01	0.02	0.810
All	1.00×10^{-5}	Weighted median	28	0.00	0.02	0.862
European American	1.00×10^{-5}	Weighted median	28	-0.01	0.02	0.862

4.4.4 Prediction of adverse outcomes by myeloid CH in CKD

As expected, established risk factors (myeloid CH, age, sex, ethnicity, smoking status, cholesterol, HbA1C, HDL, LDL, blood pressure, BMI, uACR, hs-CRP and eGFR scores) were associated on univariate analysis with an adverse outcome as defined by a composite endpoint of death, MI, or stroke (Table 4-12, Table 4-13, and Table 4-14).

Table 4-12: Regression of adverse outcomes defined by a composite end point of death, myocardial infarction, stroke (univariate analysis) in the discovery cohort

		Discovery cohort					
Factor		**adverse outcomes-free**	**adverse outcomes**	**HR**	**CI2.5%**	**CI97.5%**	***P***
N	**N**	85128	4707				
myeloid CH	**N (%)**	1386 (1.63%)	180 (3.82%)	2.32	2.00	2.69	<0.001
Age	**median**	57.00	63.00	1.10	1.09	1.10	<0.001
Sex	**male**	36821 (43.2%)	2850 (60.5%)	1.98	1.87	2.10	<0.001
Ethnicity*	**White**	79804 (93.7%)	4472 (95%)				
	Mixed	594 (0.7%)	28 (0.6%)	0.87	0.60	1.26	0.45
	Asian	1825 (2.1%)	77 (1.6%)	0.79	0.63	0.99	0.04
	Black	1373 (1.6%)	58 (1.2%)	0.81	0.63	1.05	0.11
	Chinese	281 (0.3%)	7 (.2%)	0.46	0.22	0.97	0.04
	other	837 (1%)	33 (0.7%)	0.73	0.52	1.03	0.08
	unknown	298 (0.3%)	22 (0.5%)				
	No answer	32 (0.03%)	3 (0.06%)				
Smoking status	**Never**	48165 (56.8%)	1942 (41.3%)				
	Previous	28914 (34%)	1927 (41%)	1.64	1.54	1.74	<0.001

	Current	7652 (9%)	800 (17%)	2.51	2.31	2.72	<0.001
	No answer	313 (0.4%)	31 (0.6%)				
Cholesterol	**median**	5.70	5.56	0.89	0.87	0.92	<0.001
HbA1c	**median**	35.10	36.40	1.03	1.03	1.03	<0.001
HDL	**median**	1.42	1.31	0.49	0.45	0.53	<0.001
LDL	**median**	3.55	3.49	0.91	0.88	0.94	<0.001
systolic blood pressure	**median**	138.00	145.00	1.02	1.01	1.02	<0.001
diastolic blood pressure	**median**	82.00	83.00	1.01	1.01	1.01	<0.001
hs-CRP	**median**	1.27	1.85	1.04	1.03	1.04	<0.001
BMI	**median**	26.60	27.45	1.03	1.03	1.04	<0.001
uACR	**median**	1.03	1.30	1.00	1.00	1.00	<0.001
eGFR.cys	**median**	90.42	79.75	0.97	0.96	0.97	<0.001
eGFR.creat	**median**	92.89	89.68	0.98	0.98	0.98	<0.001
eGFR.creat.cys	**median**	91.90	84.56	0.97	0.96	0.97	<0.001

* Ethnicity was encoded in integers from 1 to 6; 'white' was used as a reference

$ Smoking was encoded in integers from 0 to 2; never smoked was used as a reference

£ a composite end point includes death, MI and stroke

Table 4-13: Regression of adverse outcomes defined by a composite end point of death, myocardial infarction, stroke (univariate analysis) in the validation cohort

		Validation cohort					
Factor		**adverse outcomes-free**	**adverse outcomes**	**HR**	**CI2.5%**	**CI97.5%**	***P***
N	**N**	85420	4834				
myeloid CH	**N (%)**	1392 (1.6)	187 (3.9)	2.35	2.03	2.72	<0.001
Age	**median**	57	63 (1.3)	1.09	1.09	1.10	<0.001
Sex	**male**	36707 (43)	2916 (60.3)	1.98	1.87	2.10	<0.001
Ethnicity*	**White**	80153 (93.8)	4582 (94.8)				
	Mixed	562 (0.7)	19 (0.4)	0.61	0.39	0.96	0.03
	Asian	1795 (2.1)	95 (2)	0.97	0.79	1.19	0.78
	Black	1377 (1.6)	61 (1.3)	0.83	0.64	1.07	0.15
	Chinese	291 (0.3)	6 (0.1)	0.38	0.17	0.84	0.02
	other	832 (1)	34 (0.7)	0.75	0.53	1.05	0.09
	unknown	289 (0.3)	27 (0.6)				
	No answer	31 (0.04)	3 (0.1)				
Smoking status	**Never**	48425 (56.7)	2019 (41.8)				
	Previous	28921 (33.9)	1969 (40.7)	1.61	1.52	1.72	<0.001

	Current	7670 (9)	810 (16.8)	2.45	2.26	2.66	<0.001
	No answer	315 (0.4)	29 (0.6)				
Cholesterol	**median**	5.692	5.591	0.90	0.87	0.92	<0.001
HbA1c	**median**	35	36.5	1.03	1.03	1.04	<0.001
HDL	**median**	1.419	1.303	0.50	0.46	0.54	<0.001
LDL	**median**	3.544	3.514	0.92	0.89	0.95	<0.001
systolic blood pressure	**median**	137	144	1.02	1.01	1.02	<0.001
diastolic blood pressure	**median**	82	83	1.01	1.01	1.01	<0.001
hs-CRP	**median**	1.27	1.87	1.04	1.03	1.04	<0.001
BMI	**median**	26.6	27.5	1.04	1.04	1.05	<0.001
uACR	**median**	1.027156	1.269318	1.01	1.01	1.01	<0.001
eGFR.cys	**median**	90.51316	79.91814	0.96	0.96	0.97	<0.001
eGFR.creat	**median**	92.82222	89.46282	0.98	0.98	0.98	<0.001
eGFR.creat.cys	**median**	91.9188	84.45468	0.97	0.96	0.97	<0.001

* Ethnicity was encoded in integers from 1 to 6; 'white' was used as a reference

$ Smoking was encoded in integers from 0 to 2; never smoked was used as a reference

Table 4-14: Meta-analysis of regression of adverse outcomes defined by a composite end point of death, myocardial infarction, stroke (univariate analysis)

		Meta analysis					
Factor		**Cochran's Q**	***P* (Cochran's)**	**HR**	**CI2.5%**	**CI97.5%**	***P***
N	**N**						
myeloid CH	**N (%)**	0.02	0.90	2.34	2.10	2.59	3.19×10^{-57}
Age	**median**	0.47	0.49	1.10	1.09	1.10	$<1.0\times 10^{-300}$
Sex	**male**	0.00	0.99	1.98	1.90	2.06	7.92×10^{-234}
Ethnicity*	**White**						
	Mixed	1.34	0.25	0.75	0.57	1.01	0.05
	Asian	1.74	0.19	0.89	0.76	1.03	0.12
	Black	0.02	0.90	0.82	0.68	0.98	0.03
	Chinese	0.14	0.71	0.42	0.24	0.72	1.79×10^{-03}
	other	0.01	0.93	0.74	0.58	0.94	0.01
	unknown						
	No answer						
Smoking status	**Never**						
	Previous	0.10	0.75	1.63	1.56	1.70	6.99×10^{-103}
	Current	0.15	0.70	2.48	2.34	2.63	2.22×10^{-207}

	No answer						
Cholesterol	**median**	0.02	0.90	0.90	0.88	0.91	3.26×10^{-32}
HbA1c	**median**	5.69	0.02	1.03	1.03	1.03	$<1.0\times 10^{-300}$
HDL	**median**	0.11	0.74	0.49	0.46	0.52	1.68×10^{-117}
LDL	**median**	0.16	0.69	0.91	0.89	0.93	1.62×10^{-14}
systolic blood pressure	**median**	0.17	0.68	1.02	1.01	1.02	4.44×10^{-199}
diastolic blood pressure	**median**	0.61	0.44	1.01	1.01	1.01	5.92×10^{-27}
hs-CRP	**median**	0.34	0.56	1.04	1.03	1.04	2.97×10^{-164}
BMI	**median**	2.56	0.11	1.04	1.03	1.04	1.58×10^{-75}
uACR	**median**	49.84	0.00	1.00	1.00	1.00	1.63×10^{-20}
eGFR.cys	**median**	2.00	0.16	0.96	0.96	0.97	$<1.0\times 10^{-300}$
eGFR.creat	**median**	0.04	0.85	0.98	0.98	0.98	5.11×10^{-236}
eGFR.creat.cys	**median**	0.90	0.34	0.97	0.96	0.97	$<1.0\times 10^{-300}$

* Ethnicity was encoded in integers from 1 to 6; 'white' was used as a reference

$ Smoking was encoded in integers from 0 to 2; never smoking was used as a reference

To understand the influence of myeloid CH and CKD on adverse outcomes, I focused on participants without prevalent myeloid neoplasms (n = 320) or any prior history of CVD (n = 8459). Initially, Cox proportional-hazard analysis was used to identify risk factors unrelated to CH and CKD (Table 4-15, and Table 4-16) and then these factors were added into the model. To determine which of the three eGFR scores was most appropriate to use in the model, I tested the linearity of each score in relation to outcome using a restricted cubic spline test, as described previously [319]. Although all three scores were associated with adverse outcomes, eGFR.cys was more linear and negative compared to the scores that used creatinine in both the discovery and validation cohorts (Figure 4-5). Focusing on eGFR.cys, the risk of adverse outcomes was higher in subjects who had CKD (HR = 1.9, n = 1180/6970) compared to CKD free participants (n = 8295/172,857; P = 8.4 × 10^{-65}) (Table 4-17). The risk of adverse outcomes was estimated to be 1.56-fold higher (P = 1.4 × 10^{-11}) in cases with myeloid CH (n = 338/3078) compared to myeloid CH-free participants (n = 9137/176,749). Testing each component of adverse outcomes confirmed the previously reported features of the UK Biobank cohort [288] that CH was associated with all-cause mortality (HR = 1.91, P = 2.5 × 10^{-10}) but did not reach significance for MI (HR = 1.13, P = 0.38) or stroke (HR = 1.28, P = 0.15) considered independently, in accordance with previous findings [155,202] (Table 4-18). ROC analysis was used to assess the predictiveness of multivariable models that incorporated myeloid CH, eGFR.cys and uACR. The baseline model consisting of age, sex, smoking status, HDL, HbA1c, systolic blood pressure, hs-CRP, BMI (Table 4-18) and corrected for 10 genetic principal components had an AUC of 73.3% (72.8–73.9%). The addition of myeloid CH as a binary factor or eGFR.cys as a continuous trait improved the predictiveness of the model to an AUC of 73.4% and 74%, respectively, and including both further improved the AUC to 74.1% (73.5–74.6%), with very similar results achieved in both the discovery and validation cohorts (Figure 4-6, and Table 4-19).

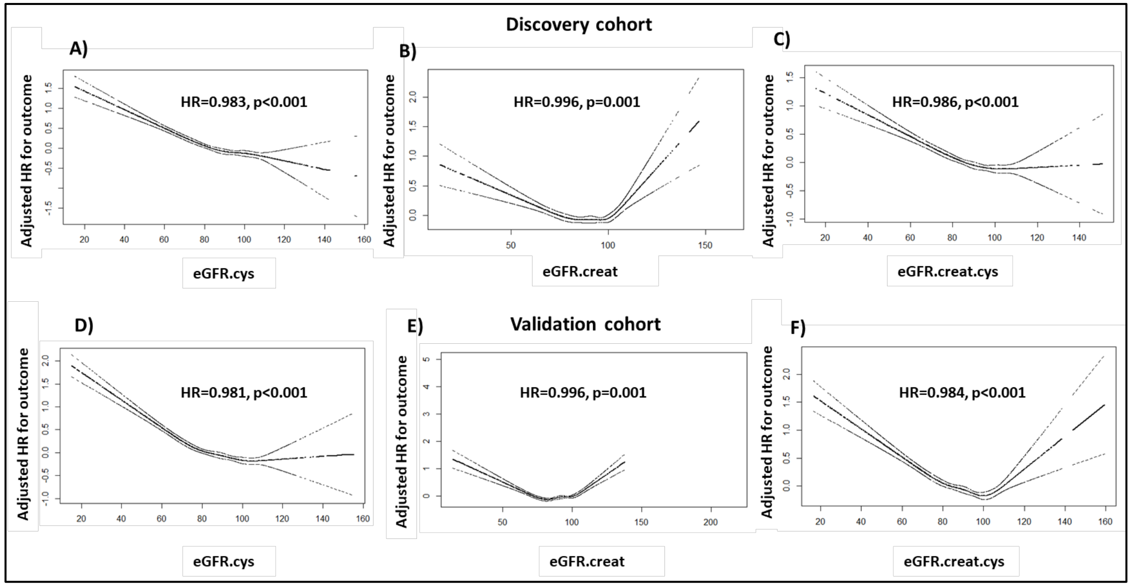


Figure 4-5: Restricted cubic spline to test the linearity of eGFR scores.

Adjusted spline of each eGFR score was plotted against HR for outcome with default values for the number of knots (n=5) and degrees of freedom (n=4). The upper and the lower dotted lines indicate 95% confidence intervals. A, B, and C refer to the discovery cohort. D, E, and F refer to the validation cohort for each eGFR score.

Table 4-15: The initial model of risk factors identified by Cox proportional-hazard analysis

	Discovery cohort				**Validation cohort**			
	OR	**CI2.5%**	**CI97.5%**	**P**	**OR**	**CI2.5%**	**CI97.5%**	**P**
age	1.09	1.08	1.09	**$< 2 \times 10^{-16}$**	1.09	1.08	1.09	**$< 2 \times 10^{-16}$**
sex	1.65	1.53	1.78	**$< 2 \times 10^{-16}$**	1.72	1.60	1.85	**$< 2 \times 10^{-16}$**
Smoking status	1.44	1.38	1.51	**$< 2 \times 10^{-16}$**	1.35	1.29	1.42	**$< 2 \times 10^{-16}$**
Diastolic blood pressure	1.00	0.99	1.00	0.06	1.00	0.99	1.00	0.1814
Systolic blood pressure	1.01	1.00	1.01	**1.57×10^{-6}**	1.01	1.00	1.01	**9.12×10^{-6}**
Cholesterol	1.09	0.94	1.26	0.27	0.95	0.82	1.10	0.50
HDL	0.68	0.59	0.79	**4.03×10^{-7}**	0.82	0.70	0.95	**8.00×10^{-3}**
LDL	0.90	0.75	1.08	0.25	1.07	0.89	1.28	0.49
HbA1c	1.02	1.01	1.02	**$< 2 \times 10^{-16}$**	1.02	1.02	1.02	**$< 2 \times 10^{-16}$**
BMI	1.01	1.00	1.02	**0.03**	1.02	1.01	1.02	**3.10×10^{-5}**
hs-CRP	1.03	1.02	1.03	**$< 2 \times 10^{-16}$**	1.03	1.02	1.03	**$< 2 \times 10^{-16}$**

Table 4-16: The final model of risk factors identified by Cox proportional-hazard analysis in the final model

	Discovery cohort				Validation cohort			
	OR	**CI2.5%**	**CI97.5%**	***P***	**OR**	**CI2.5%**	**CI97.5%**	***P***
age	1.09	1.08	1.09	**$< 2 \times 10^{-16}$**	1.09	1.08	1.09	**$< 2 \times 10^{-16}$**
sex	1.65	1.54	1.78	**$< 2 \times 10^{-16}$**	1.7	1.58	1.83	**$< 2 \times 10^{-16}$**
Smoking status	1.45	1.39	1.52	**$< 2 \times 10^{-16}$**	1.35	1.29	1.42	**$< 2 \times 10^{-16}$**
Systolic blood pressure	1.00	1.00	1.01	**2.55×10^{-6}**	1.00	1.00	1.01	**4.91×10^{-6}**
HDL	0.73	0.65	0.81	**2.97×10^{-9}**	0.78	0.71	0.87	**4.14×10^{-6}**
HbA1c	1.02	1.02	1.02	**$< 2 \times 10^{-16}$**	1.02	1.02	1.02	**$< 2 \times 10^{-16}$**
BMI	1.01	1.00	1.02	**0.05**	1.02	1.01	1.02	**7.39×10^{-5}**
hs-CRP	1.027	1.022	1.032	**$< 2 \times 10^{-16}$**	1.027	1.022	1.031	**$< 2 \times 10^{-16}$**

Model	AUC (95% CI)
Risk factors	73.3% (72.8%–73.9%)
Risk factors + CH	73.4% (72.9%–74.0%)
Risk factors and eGFR.cys	74.0% (73.4%–74.5%)
Risk factors and eGFR.cys +CH	74.1% (73.5%–74.6%)
Risk factors and uACR	72.3% (71.4%–73.2%)
Risk factors and uACR +CH	72.4% (71.5%–73.3%)

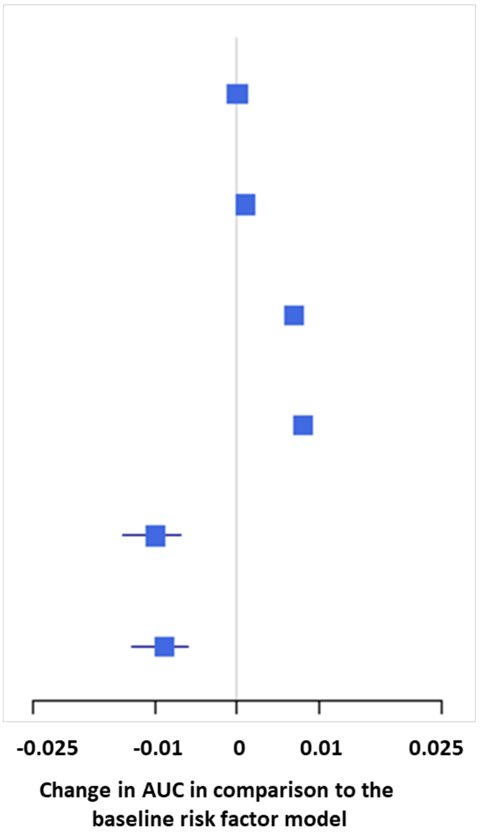


Figure 4-6: Risk factors for adverse outcome.

The baseline risk factors included age, sex, smoking status, HDL, HbA1c, systolic blood pressure, hs-CRP, BMI and was corrected for 10 genetic principal components. The effect on AUC of adding in CH, eGFR.cys, and uACR relative to the baseline model is shown (meta-analysis of discovery and validation cohorts).

Table 4-17: Prediction of adverse outcomes in participants with CKD (eGFR.cys<60) in the absence of prior myeloid malignancy or prior CVD

Group	OR	CI2.5%	CI97.5%	*P*	At risk	Incident event
	Discovery cohort					
Adverse outcomes (CKD)	1.83	1.65	2.04	$1.00\text{x}10^{-15}$	3439	560
Adverse outcomes (CKD-free)					86263	4111
	Validation cohort					
Adverse outcomes (CKD)	1.97	1.78	2.19	$1.00\text{x}10^{-15}$	3531	620
Adverse outcomes (CKD-free)					86594	4.18E+03
	Meta analysis					
Adverse outcomes (CKD)	1.90	1.77	2.05	$8.4\text{x}10^{-65}$	6970	1180
Adverse outcomes (CKD-free)					172857	8295

Table 4-18: Prediction of adverse outcomes associated with myeloid CH in the absence of prior myeloid neoplasia or prior CVD

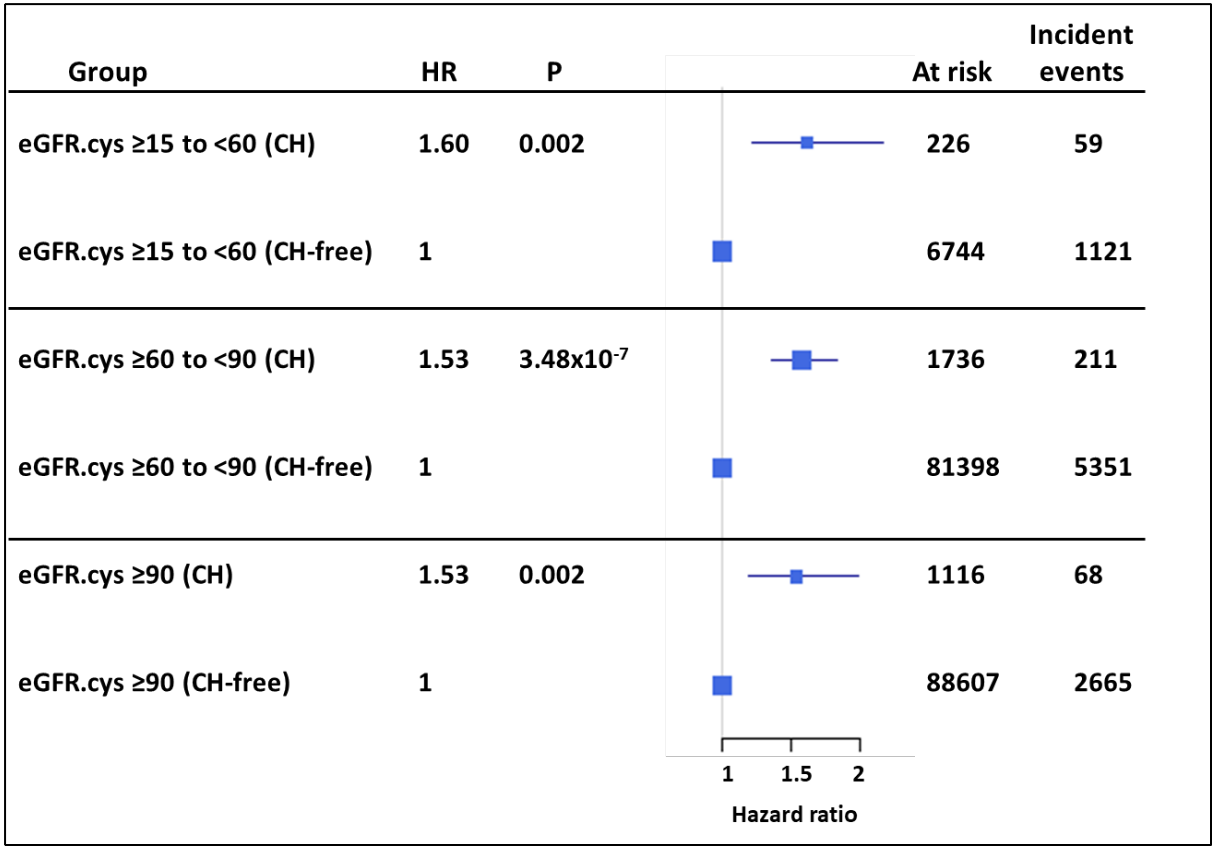
Group	**Discovery cohort**					
	OR	**CI2.5%**	**CI97.5%**	***P***	**At risk**	**Incident event**
Adverse outcomes (CH)	1.61	1.36	1.91	1.74×10^{-7}	1537	168
Adverse outcomes (CH-free)					88165	4503
Death (CH)	1.91	1.57	2.32	5.87×10^{-10}	1537	129
Death (CH-free)					88165	2878
Myocardial Infarction (CH)	1.07	0.72	1.58	0.74	1537	31
Myocardial Infarction (CH-free)					88165	1273
Stroke (CH)	1.1	0.66	1.84	0.74	1537	19
Stroke (CH-free)					88165	755
	Validation cohort					
Adverse outcomes (CH)	1.5	1.26	1.79	1.07×10^{-5}	1541	170
Adverse outcomes (CH-free)					88584	4634
Death (CH)	1.68	1.37	2.06	1.87×10^{-6}	1541	125
Death (CH-free)					88584	3001
Myocardial Infarction (CH)	1.19	0.82	1.71	0.41	1541	37

Myocardial Infarction (CH-free)					88584	1245		
Stroke (CH)	1.41	0.93	2.14	0.13	1541	29		
Stroke (CH-free)					88584	835		
	Meta-analysis						**Q**	***P* Cochran's**
Adverse outcomes (CH)	1.56	1.37	1.76	1.4×10^{-11}	3078	338	0.32	0.58
Adverse outcomes (CH-free)	1	1	1		176749	9137		
Death (CH)	1.91	1.57	2.321	2.5×10^{-10}	3078	254	0.01	0.94
Death (CH-free)	1	1	1		176749	5879		
Myocardial Infarction (CH)	1.13	0.86	1.48	0.38	3078	68	0.15	0.70
Myocardial Infarction (CH-free)	1	1	1		176749	2518		
Stroke (CH)	1.28	0.93	1.76	0.15	3078	48	0.54	0.46
Stroke (CH-free)					176749	1590		

Table 4-19: ROC analysis to compare the prediction accuracy of the models with and without CH and CKD measures

	Discovery cohort	
Predictors	**AUC (CI95%)**	**Addition of CH**
Risk factors	73.5% (72.7%–74.3%)	73.6% (72.8%–74.4%)
Risk factors + eGFR.cys	74.1% (73.3%–74.9%)	74.2% (73.4%–75.0%)
Risk factors + uACR	72.3% (71.1%–73.6%)	72.5% (71.2%–73.7%)
	Validation cohort	
Risk factors	73.3% (72.5%–74.1%)	73.3% (72.5%–74.1%)
Risk factors + eGFR.cys	73.9% (73.1%–74.7%)	74.0% (73.2%–74.8%)
Risk factors + uACR	72.7% (71.4%–74.0%)	72.7% (71.5%–74.0%)
	Combined cohort	
Risk factors	73.3% (72.8%–73.9%)	73.4% (72.9%–74.0%)
Risk factors + eGFR.cys	74% (73.4%–74.5%)	74.1% (73.5%–74.6%)
Risk factors + uACR	72.3% (71.4%–73.2%)	72.4% (71.5%–73.3%)

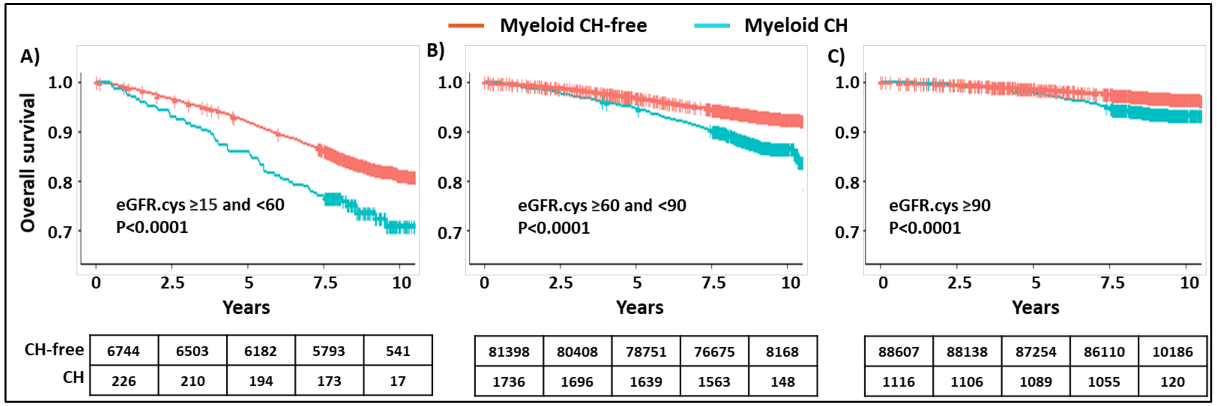
To further investigate the relationship between CH and adverse outcome in participants with CKD, I stratified the cohort (excluding prior CVD and prevalent myeloid malignancies), into participants with moderate renal impairment (eGFR.cys ≥15 to <60), mild impairment (eGFR.cys ≥60 to <90) and normal kidney function (eGFR.cys ≥90). I then tested the effect of CH in each subset using Kaplan–Meier survival analysis. CH increased the risk of adverse outcome in all groups but was particularly marked (HR = 1.6, 95% CI 1.2–2.14, P = 0.002) for participants with moderate CKD (n = 59/226 with myeloid CH compared to n = 1121/6744 without myeloid CH) (Figure 4-7, Figure 4-8, and Table 4-20). Much of the risk of adverse outcomes was related to incident myeloid neoplasms which were diagnosed in 19 participants at a median of 3.6 years after study entry. Of these, 11 (58%) had adverse outcomes in comparison to 48/207 (23%) who did not develop a myeloid neoplasm during the study period. Excluding the incident cases reduced but did not eliminate the risk of adverse outcomes (HR = 1.4, P = 0.05).



Group	HR	P	At risk	Incident events
eGFR.cys ≥15 to <60 (CH)	1.60	0.002	226	59
eGFR.cys ≥15 to <60 (CH-free)	1		6744	1121
eGFR.cys ≥60 to <90 (CH)	1.53	3.48x10^{-7}	1736	211
eGFR.cys ≥60 to <90 (CH-free)	1		81398	5351
eGFR.cys ≥90 (CH)	1.53	0.002	1116	68
eGFR.cys ≥90 (CH-free)	1		88607	2665

Figure 4-7: Myeloid CH predicts adverse outcomes in CKD

The forest plots show data stratified according to eGFR.cys as healthy (≥90), mild CKD (≥60 to <90) and moderate CKD (≥15 to <60). The risk of adverse outcomes was predicted by myeloid CH in all groups but was particularly marked (HR=1.6, P=0.002) for participants with moderate CKD.



A)

CH-free	6744	6503	6182	5793	541
CH	226	210	194	173	17

B)

CH-free	81398	80408	78751	76675	8168
CH	1736	1696	1639	1563	148

C)

CH-free	88607	88138	87254	86110	10186
CH	1116	1106	1089	1055	120

Figure 4-8: Kaplan–Meier survival estimates for the three CKD groups according to absence or presence of myeloid CH.

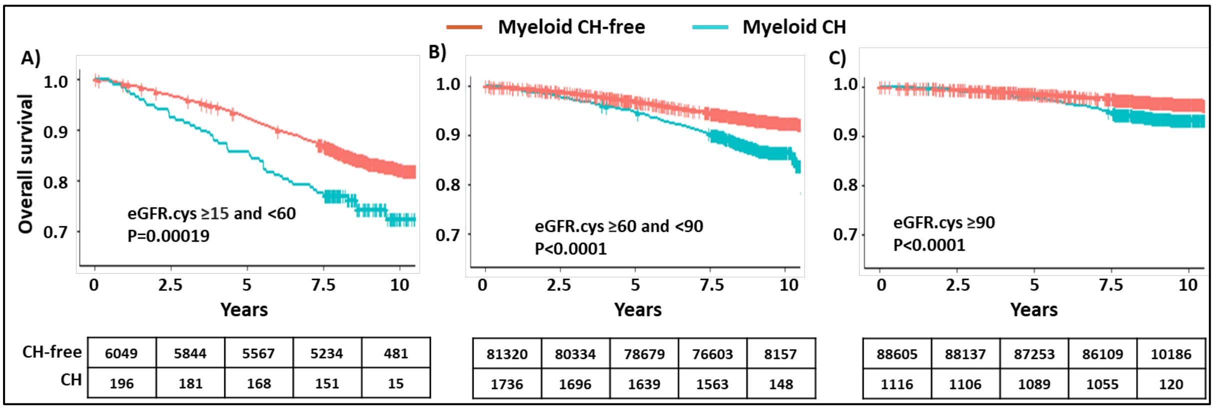
A) Moderate CKD B) mild CKD C) Normal. Log-rank test *P* values are reported for each group, and numbers at risk at 0, 2.5, 5, 7.5, and 10 years after study entry

Table 4-20: Adverse outcomes in relation to myeloid CH stratified by eGFR.cys score

	Discovery cohort							
Group	**OR**	**CI2.5%**	**CI97.5%**	***P***	**At risk**	**Incident event**		
eGFR.cys ≥15 to <60 (CH)	1.56	1.01	2.41	0.06	110	27		
eGFR.cys ≥15 to <60 (CH-free)					3329	533		
eGFR.cys ≥60 to <90 (CH)	1.59	1.28	1.97	6.65×10^{-5}	877	109		
eGFR.cys ≥60 to <90 (CH-free)					40683	2643		
eGFR.cys ≥90 (CH)	1.65	1.14	2.38	0.01	550	32		
eGFR.cys ≥90 (CH-free)					44153	1327		
	Validation cohort							
eGFR.cys ≥15 to <60 (CH)	1.64	1.11	2.42	0.02	116	32		
eGFR.cys ≥15 to <60 (CH-free)					3415	588		
eGFR.cys ≥60 to <90 (CH)	1.46	1.16	1.83	2.09×10^{-3}	859	102		
eGFR.cys ≥60 to <90 (CH-free)					40715	2708		
eGFR.cys ≥90 (CH)	1.42	0.98	2.05	0.09	566	36		
eGFR.cys ≥90 (CH-free)					44454	1338		
	Meta-analysis						**Cochran's Q**	***P* Cochran's**
eGFR.cys ≥15 to <60 (CH)	1.60	1.20	2.14	2.12×10^{-3}	226	59	0.03	0.87
eGFR.cys ≥15 to <60 (CH-free)	1	1	1		6744	1121		
eGFR.cys ≥60 to <90 (CH)	1.53	1.30	1.79	3.48×10^{-7}	1736	211	0.26	0.61
eGFR.cys ≥60 to <90 (CH-free)	1	1	1		81398	5351		
eGFR.cys ≥90 (CH)	1.53	1.17	1.99	2.12×10^{-3}	1116	68	0.31	0.58
eGFR.cys ≥90 (CH-free)	1	1	1		88607	2665		

4.4.5 Relationship between myeloid CH and shrunken pore syndrome

I identified 966 (0.5%) of the UK Biobank participants with potential SPS (eGFR.cys/eGFR.creat ratio ≤0.6). Of these, 6% (n=58) had myeloid CH. In comparison, 2.9% (n = 5391) participants had myeloid CH and eGFR.cys/eGFR.creat ratio > 0.6 (OR = 2.2, 95% CI = 1.6–2.9; P = 2.9 × 10^{-7} Fisher's exact test). After eliminating these SPS cases, myeloid CH remained associated with an adverse prognosis in CKD (HR = 1.61, 95% CI 1.17–2.21, P = 0.003) and remained most pronounced for participants with moderate renal impairment (Figure 4-9).



	A)					B)					C)				
CH-free	6049	5844	5567	5234	481	81320	80334	78679	76603	8157	88605	88137	87253	86109	10186
CH	196	181	168	151	15	1736	1696	1639	1563	148	1116	1106	1089	1055	120

Figure 4-9: Kaplan-Meier survival estimates for the three CKD groups according to absence or presence of myeloid CH and excluding participants with potential SPS.

The analysis excluded the 966 participants with potential SPS. A) Moderate CKD B) mild CKD C) Normal. Log-rank test P values are reported for each group, and numbers at risk at 0, 2.5, 5, 7.5, and 10 years after study entry.

4.5 Discussion

In this study I identified that CH, and specifically myeloid CH, is associated with CKD. The association was not seen with all markers of CH and, strikingly, not with mutations in *DNMT3A* or *ASXL1*, two of the most common drivers of clonality, although there was an overall association with clone size. These findings confirm previous observations that not all CH is equal [169,202,325], as well as the importance of having sufficiently large studies to understand the granularity of CH with respect to clinical outcomes.

I found that myeloid CH is specifically associated with eGFR.cys but not eGFR.creat and only marginally with eGFR.cys.creat. Similarly, recent studies have reported the superior utility of eGFR.cys in predicting the incidence of CVD and mortality in patients with CKD [319,326,327]. In the UK, the cost to measure cystatin C is 10-fold higher than that to measure serum creatinine, and consequently eGFR.creat is widely used for initial assessment of possible CKD. Although eGFR.cys is recommended to confirm CKD, this is not believed to be common practice, at least in the UK [319]. My findings provide further weight to the argument that eGFR.cys is more informative than eGFR.creat to define CKD.

The finding that myeloid CH is associated with eGFR.cys also provides further evidence for the importance of chronic inflammation in CH-related disorders. Levels of cystatin C correlate generally with oxidative stress and inflammation [319,328], a well-recognised feature of CKD [305] that is also associated with an elevated risk of development of CVD [298,329]. Other biomarkers of chronic inflammation have been associated with CH, e.g. C-reactive protein and IL-6 [169,306]. CH predisposes to haematological malignancies, particularly myeloid neoplasms [100], and both CKD and chronic inflammation have been described as features of myeloproliferative neoplasms [330,331]. My data show that myeloid CH increases the risk of adverse outcomes in the context of CKD and that this increase is only partly explained by incident myeloid neoplasms or SPS, a recently described phenomenon that may be observed in both children or adults with normal or reduced eGFR and is associated with increased mortality and morbidity in a variety of settings [318]. Although my analysis was corrected for hs-CRP, it is possible that part of the increase in adverse outcomes is due to chronic inflammation induced by CH.

MR uses genetic variation as a natural experiment to estimate causality in observational data [321] and has been used, for example, to detect a causal effect of cystatin C on risk of stroke [332]. My initial analysis of 380 SNPs that predispose to CH provided suggestive evidence for a causal relationship between CH and CKD (P = 0.03), but this link was not supported by a more conservative analysis of 28 SNPs that have lower P-value . Given that two of the most common CH genes (*DNMT3A* and *ASXL1*) were not associated with CKD, and that the 380 SNPs only explain 3.6% of the heritability of CH [169], the use of MR in this context is clearly challenging, and may be compounded by the possibility of other factors such as horizontal pleiotropy but these concerns are partly mitigated by the large sample size of the GWAS used for CH and CKD.

In summary, the role of CH in the pathogenesis of benign diseases varies widely and depends on intrinsic factors that define the clone as well as extrinsic factors that impact the inflammatory environment [88,202]. In this study, I have shown that CH is associated with CKD and confers an adverse prognosis over and above conventional risk factors for this common disorder. My findings suggest that screening for CH in CKD may be of clinical value to help predict outcomes.

Chapter 5 Prediction of myeloid malignancies in healthy individuals

5.1 Summary

In this Chapter, I describe the use of high dimensional data including CH metrics, blood counts, biochemistry measures, and healthcare data to predict the risk of developing myeloid malignancies in the UK Biobank cohort. The study base line was individuals with no reported myeloid neoplasm either before recruitment or up to one year after recruitment to exclude undiagnosed conditions. The analysis was conducted on 726 pre-myeloid cases, i.e. individuals who fulfilled the study base line criteria but developed a myeloid neoplasm during the study period (median follow-up from recruitment to diagnosis = 7.1 years, range = 1-13.4), and 7,260 controls (median follow-up = 11.7 years, range = 0.08-14.6) that were free from myeloid malignancies during the study period and were matched for age (mean age = 61.2) and sex (males = 53%). Participants were randomly split into a training set (80%) for model development and test set (20%) for evaluation. CH was defined by both mCA and driver mutations. mCA were classified according to their physical location, copy number state using log R ratio and estimated level of clonality using mBAF as described in Chapter 3. Driver mutations were identified by the Mutect2 somatic caller, classified according to gene name and mutation type (nonsynonymous, stopgain, frameshift-deletion, frameshift-insertion, or splicing mutation), and encoded by their VAF values. The model included four new features which represented the number of lesions targeting myeloid genes, myeloid mCA, lymphoid genes, and lymphoid mCA. As expected from previous studies, epigenetic regulators (*DNMT3A, TET2*, and *ASXL1*) were the most frequently mutated myeloid genes in both pre-myeloid cases and controls (49% and 73% of mutated genes, respectively). However, the pre-myeloid group was enriched in mutated splicing genes (*SRSF2*, *SF3B1*, and *U2AF1,* P= 3.18×10^{-50}), *JAK2* (P= 1.71×10^{-49}), *IDH1/2* (P= 5.33×10^{-16}*)*, in addition to epigenetic regulator genes (P= 4.68×10^{-30}). Interestingly, the number of lesions per individual were significantly higher in the pre-myeloid group compared with controls (P=3.02×10^{-119}, Mann–Whitney U test). This result is driven by the larger number of myeloid CH lesions in the pre-myeloid group versus controls (myeloid genes, P=7.17×10^{-119}; myeloid mCA, P=3.20×10^{-67}) and not lymphoid CH (lymphoid genes, P=0.04; lymphoid mCA, P=0.18). Several models were tested for

prediction of myeloid malignancies, the first was an Elastic-Net-regularised Cox proportional hazards (COX-PH) model. The best regularised COX-PH model for prediction of disease risk consisted of six features with non-zero coefficients. The number of lesions (myeloid genes) had the largest coefficient, and the other selected features were *JAK2* V617F, *SRSF2* P95, red cell distribution width (RDW), platelet count, and number of lesions (myeloid mCA). This model achieved a concordance index (C-index, defined as the proportion of concordant pairs divided by the total number of possible evaluation pairs) of 0.57 in the test data. Next, different machine learning models were tested which included all the features in a single model. These machine learning models out-performed the COX-PH model (C-index >0.57) when evaluated on the test data. Random Survival Forest (RSF) was the most predictive machine learning (ML) model, which achieved a C-index of 0.78 in the test data, a time-dependent AUC ranging between 0.9 and 0.74 at 2 years and 12.5 years from recruitment, respectively, and attributed the largest weights to platelet indices and number of lesions (myeloid genes). This research demonstrates that the number of mutated myeloid genes is a significant predictor for the risk of myeloid malignancies. In addition, ML survival models can deal effectively with large datasets combining both CH and other healthcare data in a single model with performance that exceeds the traditional COX-PH models.

5.2 Introduction

The finding of CH in healthy individuals years before the diagnosis of an overt myeloid malignancy [11,43] has raised the interesting prospect of utilising CH observations to predict the risk of these malignancies in healthy individuals from a single blood sample. CH is a measurable event with a dynamic nature that can be defined by (i) targeted gene (ii) clone size (iii) age at detection (iv) number of genetic lesions. Previous studies have demonstrated the potential of CH measures for prediction of myeloid malignancies [42,43,86]. On the gene level, mutations in epigenetic regulator genes were the most common in healthy individuals, however, mutations in the splicing genes *SRSF2*, *SFB31* and *U2AF1*, were notably enriched in pre-AML cases [42], and associated with high myeloid-related mortality compared to controls [86]. Mutant VAFs have been used to estimate clone size, which also help to predict subsequent AML [42,43]. In addition, individuals with a larger number of clones were at higher risk of developing myeloid malignancy, and this feature was independent of the correlation between point mutations and mCA, i.e. the number of clones remains significantly associated with the risk of developing myeloid malignancies after excluding samples with point mutations/mCA pairs such as *JAK2* V617F/9p UPD [86,87]. Although these studies which utilised CH measures were predictive of myeloid malignancies, the models had some limitations such as: (i) treating CH measures as independent variables without considering other risk factors such as blood counts [101]; (ii) a focus on specific subclasses of myeloid disorders such as AML despite the clear relationship between different types of myeloid neoplasms which have similar genetic profiles [101]; (iii) frequent use of the COX-PH model which cannot deal with model nonlinearity and interactions between variables. Much of these limitations are related to the high dimensionality of myeloid-neoplasm risk factors that traditional methods struggle to accommodate, resulting in reduced accuracy, overfitting, and longer time to train the model. Machine learning (ML) has been established as a modern method to improve cancer prediction, diagnosis and prognosis [333]. The expansion of ML methods to handle censored data allows its use in survival analysis [334].

In this Chapter, I aimed to predict the risk of developing myeloid malignancies in the UK Biobank subjects by considering healthcare data, CH metrics, blood counts, and biochemistry measures. Within the work (i) I applied a regularised COX-PH model to select small number of features that can predict the disease (ii) I tested different machine learning approaches and evaluated them against COX-PH model.

5.3 Methods

5.3.1 Study Cohort

The UK Biobank phenotype data from May 2021 was used that included updated follow-up data for hospital inpatient episodes, death registries and primary care for a subset of participants (45%). Focusing on participants with available WES data at the time of analysis, cases of myeloid neoplasms were identified based on hospital inpatient records (FID: 41202-41205, 41270, 41271), death registry (FID: 40001-40002), cancer registry (FID: 40006, 40013), self-reported medical conditions (FID: 20001, 20002), and primary-care data. The model development was conducted on cases that were diagnosed at least one year after recruitment (pre-myeloid group), according to the date of first occurrence from primary care, hospital inpatient data, and death registries. Participants with prevalent myeloid malignancy, i.e. those diagnosed before recruitment or up to a year after recruitment were excluded from the analysis. For the control group, participants were selected to be free from myeloid malignancies according to the registries listed above but were not selected to be free from other types of cancer. A propensity score matching method in R [335] was used to select ten controls per case and to match for age and sex in comparison to the myeloid incidence group.

5.3.2 Whole Exome Sequencing:

Processed CRAM files, released in February 2021, were obtained from the UK Biobank. These files were generated using an updated Functional Equivalence (FE) protocol [210] to map reads to the reference human genome sequence (version GRCh38) and retain Original Quality scores. The resulting CRAM files are referred to as OQFE.

5.3.3 Somatic variants calling

Somatic mutations were called using GATK (Version 4.1.9) and Mutect2 [336] to process individual CRAM files in the tumour-only mode. Following best practice guidelines (https://gatk.broadinstitute.org/hc/en-us/articles/360035531132), a Panel Of Normal (PON) and germline resources were used to remove artefacts and germline variants (Figure 5-1). First, Mutect2 was run using an option to output read count statistics (--f1r2-tar-gz) for subsequent orientation bias modelling, and the following input files (i) individual CRAM file, (ii) reference genome sequence used by the UK Biobank, (iii) PON from the Broad institute (1000g_pon.hg38.vcf.gz) that were generated using Mutect2 to process samples from the 1000 genomes project and identify recurrent artifacts, (iv) germline variants from GnomAD (af-only-gnomad_grch38.vcf.gz) and (v) a list of regions targeted by the WES experiment (xgen_plus_spikein.b38.bed). This step generated a list of raw variants in VCF format including number of reads in the F1R2 orientation that were used to learn the orientation bias model and exclude potential artefacts. Second, GetPileupSummaries was run to summarise reads for a set of germline variants from EXAC data (somatic-hg38_small_exac_common_3.hg38.vcf.gz). Third, contamination was estimated by CalculateContamination that determines the fraction of reads resulting from cross-sample contamination and estimates the allelic copy number segmentation. Fourth, the orientation bias model was fitted. Fifth, Mutect2 raw calls were filtered based on the previously generated contamination data and orientation model.

Publicly available resources were obtained from the Broad institute (gs://gatk-best-practices/somatic-hg38) and 1000 genomes project:

(ftp://ftp.ncbi.nlm.nih.gov/1000genomes/ftp/technical/reference/GRCh38_reference_genome/)

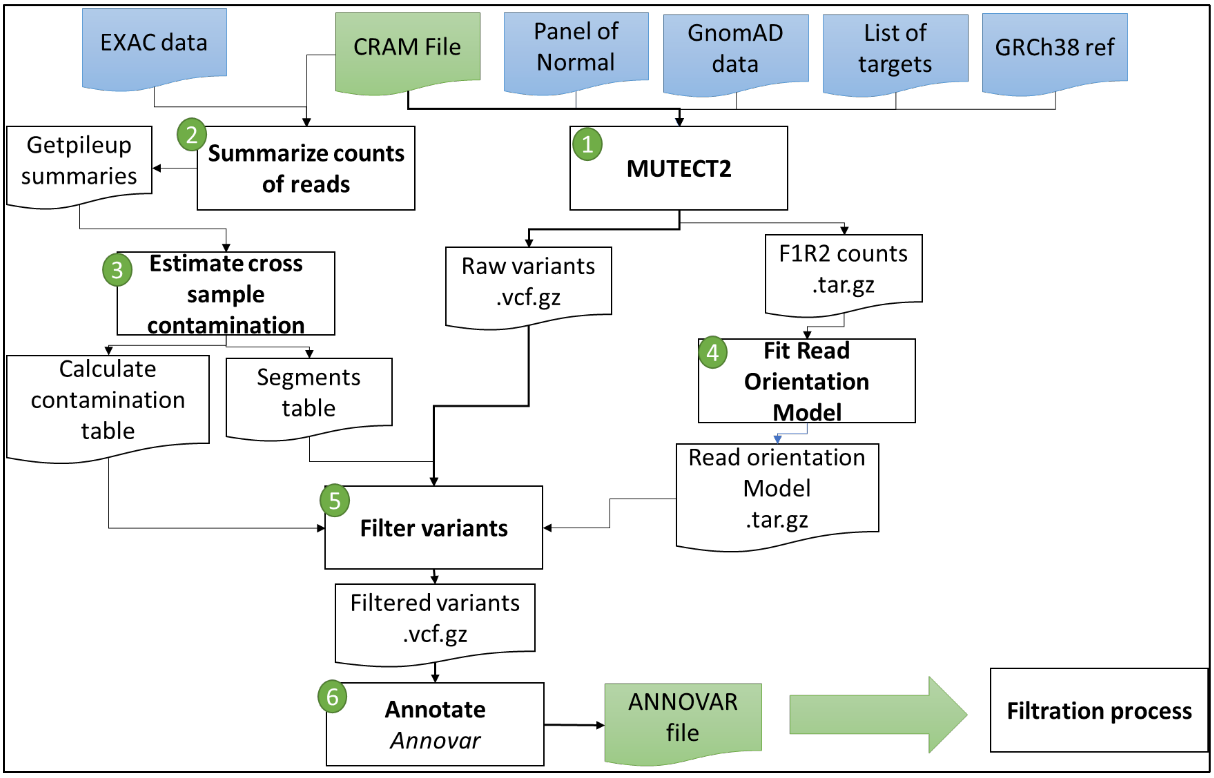


Figure 5-1: Bioinformatic pipeline used to call somatic mutations

The pipeline was run on each sample independently: 1) raw variants were called by Mutect2 by utilising the CRAM file, panel-of-normal, list of targets sequenced by WES, and reference human genome sequence (GRCh38) used by the UK Biobank to generate CRAM files; 2) Getpileup was run on CRAM files and utilised common variants information from EXAC data, and 3) these information were used in step 3 to estimate contamination; 4) F1R2 data were used to fit the orientation model; 5) Data from Mutect2, calculate contamination, and learn read orientations were used to filter variants generated in step 1. Finally, the filtered VCF file was annotated by Annovar adding information from the COSMIC database.

5.3.4 Driver mutations characterization:

To identify putative somatic driver mutations, the analysis was restricted to rare variants with a MAF<0.01 in GnomAD and a minimum number of reads supporting the mutated allele: 3 reads for point mutations and 6 reads for indels. Variants that satisfied either of the following criteria were selected:

First, the variant was defined as a driver mutation in our previous study [203] and did not have any of the following errors in the new Mutect2 calls; strand bias, mapping quality, or clustered events. Second, ultra-rare singleton variants in genes from the cancer gene census list [337] that passed all Mutect2 filters and had a VAF between 0.1 and 0.2 [9]. These variants were further restricted to those with (i) CADD ≥ 20 or (ii) loss of function effect. Driver mutations were classified into myeloid according to the targeted gene using a published list of myeloid genes [89], driver mutations targeted other genes (non-myeloid) were classified as lymphoid.

5.3.5 Predicting the risk of myeloid malignancies

5.3.5.1 Independent variables

(i) Four new features were generated to represent the number of lesions: 1) number of myeloid driver mutations; 2) number of lymphoid driver mutations; 3) number of myeloid mCA; 4) number of lymphoid mCA.

(ii) Driver mutations were classified according to the name of the targeted gene, and their functional impact (nonsynonymous, stopgain, frameshift-deletion, frameshift-insertion, or splicing mutation), and presented by the largest VAF value identified in each group per individual as a representation of the founding clone.

(iii) In previous work, mCA associated with myeloid and lymphoid disease were identified as described in Chapter 3 [202]. I presented mCA by mirrored B-allele frequency (mBAF) of each event.

(iv) Observations of LOY were extracted from previous studies which used a modified version of the MoChA tool and 1,239 germline variants from the Biobank SNP array data to identify regions of allelic imbalance in the PAR1 region [80,253]. The LOY calls were presented by the reported change in BAF.
(v) 29 blood counts and 29 blood biochemistry variables were measured or estimated in the UK Biobank and presented as continuous variables (Table 3-2, and Table 3-3).

(vi) other clinical features used were age, sex, systolic and diastolic blood pressure, smoking status, alcohol consumption status, and BMI.

5.3.5.2 Dependent variable

Incidences of myeloid malignancy were coded as a binary variable (yes/no) and treated as the prediction object. Follow-up times were determined using the “lubridate” [240] package in R to calculate the duration between recruitment and either disease incidence for cases or end of follow-up for controls (The UK Biobank field 191). Data were split randomly into a training and test set with 80% to 20% split, respectively.

5.3.5.3 Developed models

Data imputation, model building, and method evaluation were applied using the available tools in scikit-learn (v.0.23.2), and scikit-survival (v. 0.14.0), Python packages [244,245]. Missing values were replaced by the mean along each independent variable by SimpleImputer class in scikit-learn.

Regularised Cox-proportional hazard's model: A Cox-proportional hazards model was fitted to the training data and regularised using the Elastic Net method [338] to overcome the correlations among the large number of features, and to determine the final model with the most efficient number of features. The Elastic Net method adds a combined ℓ1 (Ridge) and ℓ2 (LASOO) penalty which respectively shrink the coefficients to almost zero and select a subset of features that are more predictive with non-zero coefficient. A default ℓ1:ℓ2 ratio of 0.9 was used along with five-fold cross validation to evaluate a range of 100 hyperparameters (α) with a minimum α value of 0.01 to control shrinkage. The best model was selected according to the α value that achieved the largest *C-index* in the training data.

Random survival forest (RSF) model is the ensemble of trees method that was extended to deal with right censored time to event data. Bootstrapped data are used to build the base tree, whereby each tree is built on different samples, and each node is split on different features from the original data. Finally, prediction is a result of combining individual trees. Recommended settings were used for the number of estimators (n=1000), minimum samples split (n=10), and minimum samples leaf (n=15) [218]. To identify the most important features in the model, permutation was applied using the ELI5 library (https://pypi.org/project/eli5/), which is compatible with scikit-survival.

Gradient Boosted models: A Gradient Boost model is a sequential ensemble of small models. The method was applied to three different base models (i) Cox's partial likelihood with regression tree to maximise the log partial likelihood function, (ii) component-wise least squares base learners which minimises the residual sum of squares, (iii) accelerated failure time (AFT) model with inverse-probability of censoring weighted least squares error.

5.3.5.4 Evaluation methods:

The performance of each model was evaluated on the test data using Harrell's *C-index* [339]. In addition, time-dependent area under the ROC was estimated at different time points across the study time and with intervals of 0.5 years using the *cumulative_dynamic_auc* command in scikit-survival [245], and the results were plotted in a line chart.

5.3.6 Statistical tests for the association of CH with the risk of myeloid malignancies, age, and VAF

The frequency of CH and each of its subcategories (myeloid and lymphoid), were tested for association with the risk of developing myeloid malignancies compared to controls using Fisher's exact tests in statistical functions (scipy.stats) in python. Median age and median VAF distributions were compared between pre-myeloid participants and controls using the Wilcoxon rank-sum test that was applied using scipy.stats. Mann–Whitney U test was used to assess the mean number of lesions between pre-myeloid participants and controls.

5.4 Results

A total of 1,266 out of 200,631 participants with WES data in the UK Biobank had evidence of myeloid malignancy. The study was conducted on the 726 of these individuals who developed myeloid malignancies at least 1 year after recruitment (pre-myeloid; median = 7.1 years, range = 1-13.4). These 726 individuals included MPN (n=321), AML (n=155), MDS (n=141), and others (n=109) as presented in Figure 5-2. As controls, 7,260 individuals were selected that were free from myeloid malignancies at the end of follow-up (median follow-up = 11.7, range = 0.08-14.6) and were matched for age and sex (mean age = 61.2, and males = 53%).

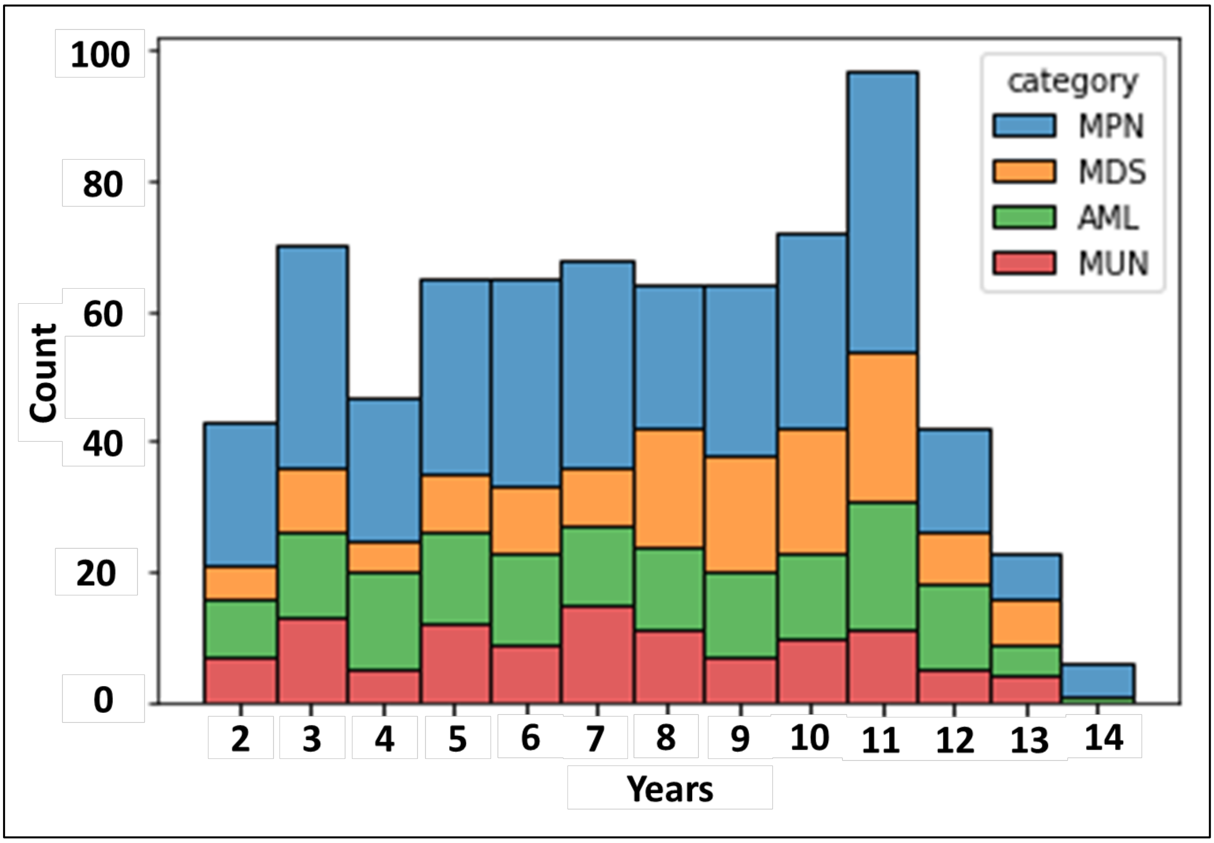


Figure 5-2: The distribution of pre-myeloid cases across the study time

A histogram showing pre-myeloid cases (n=726), who developed myeloid malignancies after 1 year of recruitment; MPN (n=321), AML (n=155), MDS (n=141), and others (n=109). The x-axis shows the time between recruitment and diagnosis. The y-axis shows count of individuals in each category.

5.4.1 The frequency of CH associated with myeloid and lymphoid disease

In my previous work described in Chapter 3, mCA and driver mutations were identified in 5,040 (1%) and 3,863 (2%) participants with SNP array or WES data respectively [202]. In this work, specialised software (Mutect2) and new criteria were used to identify driver mutations in the absence of matched germline samples. This new analysis increased the number of participants with driver mutations from 182 to 264 in pre-myeloid group and from 186 to 896 in the control group (Table 5-1).

Table 5-1: A comparison between GATK calls and Mutect2 calls

	Pre-myeloid$^{\$}$ (n=726)		Control$^{£}$ (n=7260)	
	Old criteria (GATK)	**New criteria (Mutect2)**	**Old criteria (GATK)**	**New criteria (Mutect2)**
All drivers	240 (182)	409 (264)	197 (186)	1110 (896)
DNMT3A	36 (36)	67 (63)	80 (80)	238 (235)
TET2	47 (39)	41 (39)	40 (39)	69 (64)
ASXL1	20 (20)	11 (11)	17 (17)	18 (18)
JAK2	48 (48)	49 (49)	3 (3)	6 (6)
Others	89 (73)	241 (160)	57 (55)	779 (634)

* Number of affected individuals was added between brackets

$^{\$}$ individuals who developed myeloid malignancies after 1 year of recruitment

$^{£}$ individuals who were free from myeloid malignancies

The frequency of CH was significantly higher in pre-myeloid participants versus controls (37.9% versus 13.2%, *P* = 5.87x10^{-56}, Fisher's exact test, Table 5-2). When focusing on myeloid CH, defined by alterations in myeloid-related genes or chromosomal regions (mCA), the difference between pre-myeloid individuals and controls were even more striking (31% versus 6%, P=1.4x10^{-79}, Table 5-2. In contrast, the frequency of lymphoid CH was similar between pre-myeloid participants and controls (6% versus 5%, P=0.12, Fisher's exact test, Table 5-2).

Table 5-2: Summary of CH events and their relationship with pre-myeloid cases

	mCA*		**Driver mutations***		**Clonal Haematopoiesis**		**Fisher's exact test**	
	Pre-myeloid (n=726)	**Control (n=7260)**	**Pre-myeloid (n=726)**	**Control (n=7260)**	**Pre-myeloid (n=726)**	**Control (n=7260)**	**OR**	***P***
Myeloid CH	36 (32)	3 (2)	295 (219)	453 (429)	222	431	6.98	$1.4\text{x}10^{-79}$
Lymphoid CH	3 (3)	14 (9)	114 (68)	658 (537)	46	360	1.32	0.12
All CH	75 (47)	125 (83)	409 (264)	1110 (896)	275	957	4.1	$5.87\text{x}10^{-56}$

* Number of events are indicated, with the number of affected individuals in brackets

Mutations in *DNMT3A* were the most frequent alteration in both groups, occurring in 9% (n=63/726) and 3% (n=235/7260) of pre-myeloid and control participants (Figure 5-3) and accounting for 29% (n= 63/219) and 55% (n=235/429) of CH defined by myeloid genes, in cases and controls respectively. In total, epigenetic regulators (*DNMT3A, TET2* and *ASXL1*) were the most frequently mutated myeloid genes and accounting for 48.9% (n=107/219), and 73% (n=313/429), in cases and controls respectively. When grouped by gene function, the pre-myeloid group was enriched for mutations in genes encoding splicing factors (OR = 39.97; *P* = $3.18\text{x}10^{-50}$, Fisher's exact test), *JAK2* (OR = 110.66; *P* = $1.71\text{x}10^{-49}$), and *IDH1/2* (OR = 101.62; *P* = $5.33\text{x}10^{-16}$). The full data are summarised in Table 5-3.

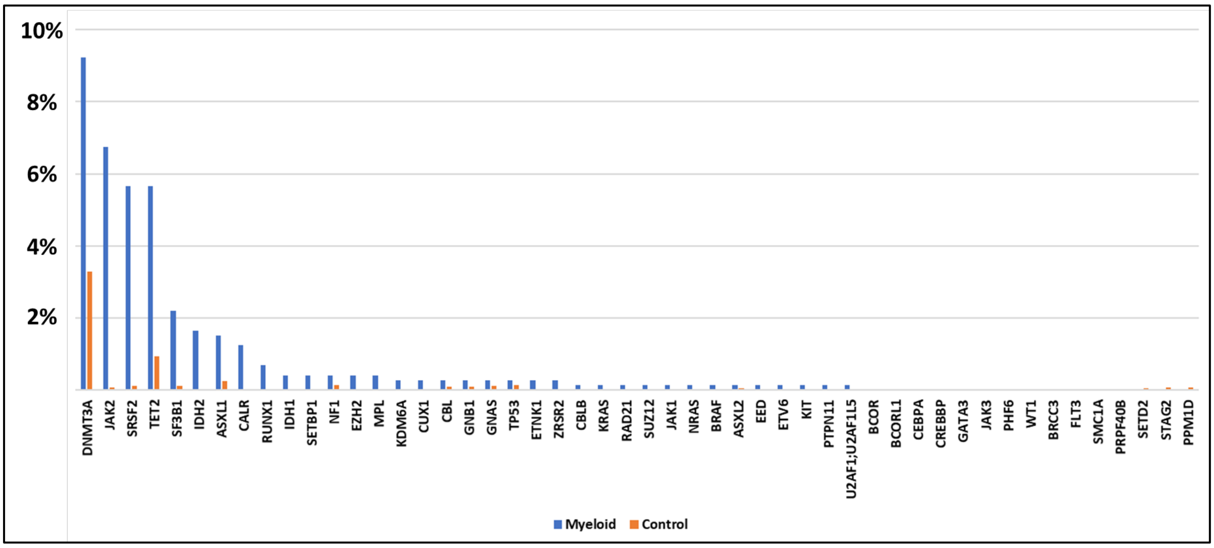


Figure 5-3: The distribution of mutated myeloid genes between pre-myeloid and control groups

Table 5-3: The difference in mutated genes between pre-myeloid and control

CH group	**Pre-myeloid versus controls**		**Age comparison between CH and CH free groups**	
	OR	**Fisher exact (P)**	**Median age$^{\$}$**	**Wilcoxon rank-sum**
All CH	4.12	5.87x10^{-56}	64	1.67x10^{-11}
Myeloid CH (genes + mCA)	6.98	1.40x10^{-79}	64	5.46x10^{-16}
Myeloid CH (myeloid genes)	6.87	8.43x10^{-78}	64	1.28x10^{-15}
Myeloid CH (myeloid mCA)	1.67	1.18x10^{-31}	63	0.17
Lymphoid CH (genes + mCA)	1.32	0.04	63	0.12
Lymphoid CH (Lymphoid genes)	1.29	0.055	63	0.11
Lymphoid CH (Lymphoid mCA)	2.00	0.29	62	0.76
Epigenetic regulator genes	4.64	4.68x10^{-30}	64	4.34x10^{-13}
DNMT3A	3.63	8.45x10^{-15}	63	2.65x10^{-5}
ASXL1	8.28	1.92x10^{-6}	65	0.00054
TET2	8.26	6.00x10^{-19}	66	5.84x10^{-11}

JAK2	110.66	1.71x10^{-49}	63	0.39
Splicing genes	39.97	3.18x10^{-50}	65	8.13x10^{-6}
IDH1/2	101.62	5.33x10^{-16}	63	0.77
Damage genes (*PPM1D/TP53*)	1.69	0.35	63	0.067
GNAS/GNA1	2.46	0.22	66	0.01
Ligase (*CBL/CBLB*)	5.08	0.04	63	0.32

$ median age of CH-free controls is 62

5.4.2 Number of lesions

The mean number of lesions per individual increased with age and was significantly higher in the pre-myeloid group compared with controls (P=3.02x10^{-119}, Mann–Whitney U test). This result is driven by the larger number of myeloid CH lesions in the pre-myeloid group versus controls (myeloid genes, P=7.17x10^{-119}; myeloid mCA, P=3.20x10^{-67}) and not lymphoid CH (lymphoid genes, P=0.04; lymphoid mCA, P=0.18). Number of lesions is shown in Table 5-4, and the mean number of lesions in each age group is shown in Figure 5-4.

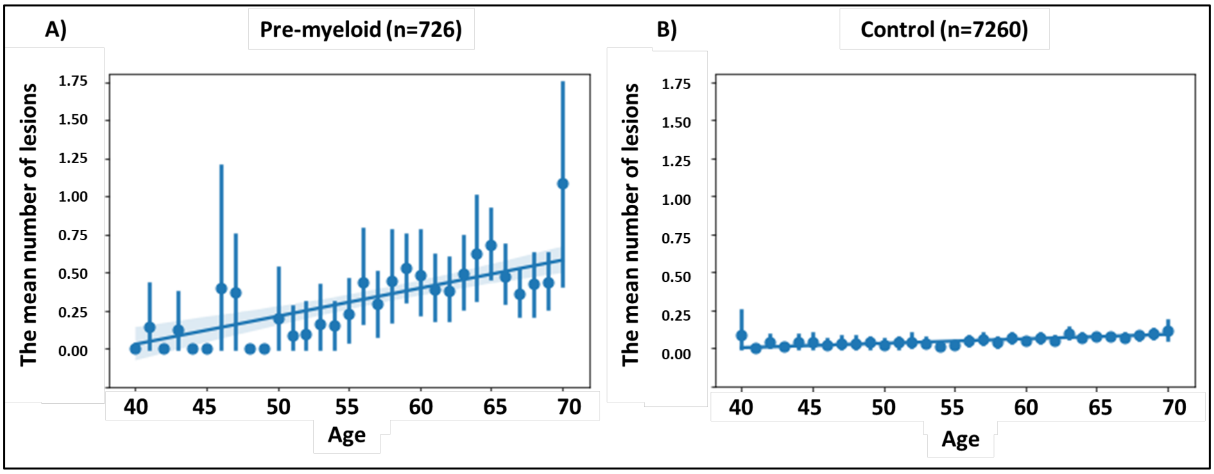


Figure 5-4: The relationship between number of lesions and age

In both groups pre-myeloid (A) and control (B), the mean number of lesions, defined by both myeloid or lymphoid events increase with age. Number of lesions per individual was significantly higher in the pre-myeloid group in comparison to controls (P=6.85x10^{-36}, t-test).

Table 5-4: Number of recurrent alterations in cases and controls

	Pre-myeloid				Control			
N	Myeloid genes	Myeloid mCA	Lymphoid genes	Lymphoid mCA	Myeloid genes	Myeloid mCA	Lymphoid genes	Lymphoid mCA
0	506	694	658	724	6825	7258	6711	7250
1	157	28	53	1	412	2	472	6
2	53	4	11	1	22	0	50	4
>2	10	0	4	0	1	0	27	0

5.4.3 Clone size

The median VAF was significantly different between pre-myeloid (median=0.15, range= 0.03 - 0.9) and control (median = 0.12, range = 0.03 - 0.5) as ($P=1.81\times10^{-24}$, Mann–Whitney U test), this difference is due to mutations in myeloid-related genes ($P=6.53\times10^{-26}$) rather than lymphoid related genes (P= 0.43). The distribution of VAF is illustrated in Figure 5-5.

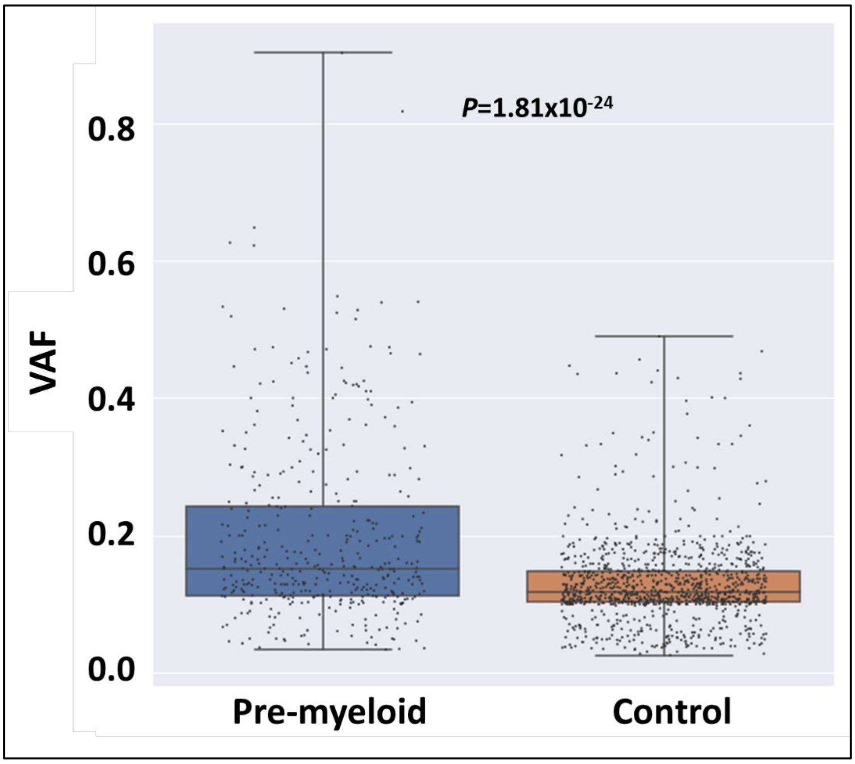


Figure 5-5: The distribution of VAF of driver mutations in pre-myeloid and control

Median VAF was significantly higher in pre-myeloid samples (median = 0.15, range = 0.03 - 0.9) compared with controls (median = 0.12, range = 0.03 - 0.5).

5.4.4 Prediction of myeloid malignancies by a subset of features

In clinical practice, we need to select a small number of predictors to model the risk of developing myeloid malignancies in healthy individuals. The optimal Elastic-Net regulated COX-PH model was used to achieve this using an alpha value of 0.046 and the following 6 features that are listed in descending order of coefficient size; (i) number of lesions in myeloid genes, (ii) VAF for *JAK2* V617F, (iii) RDW, (iv) platelet counts, (v) VAF for *SRSF2* P95, and (vi) number of myeloid mCA (Figure 5-6). When evaluating the optimal Elastic-Net regulated COX-PH model on test data it achieved a C-index of 0.57 which indicates that the model was better at predicting an outcome than random chance.

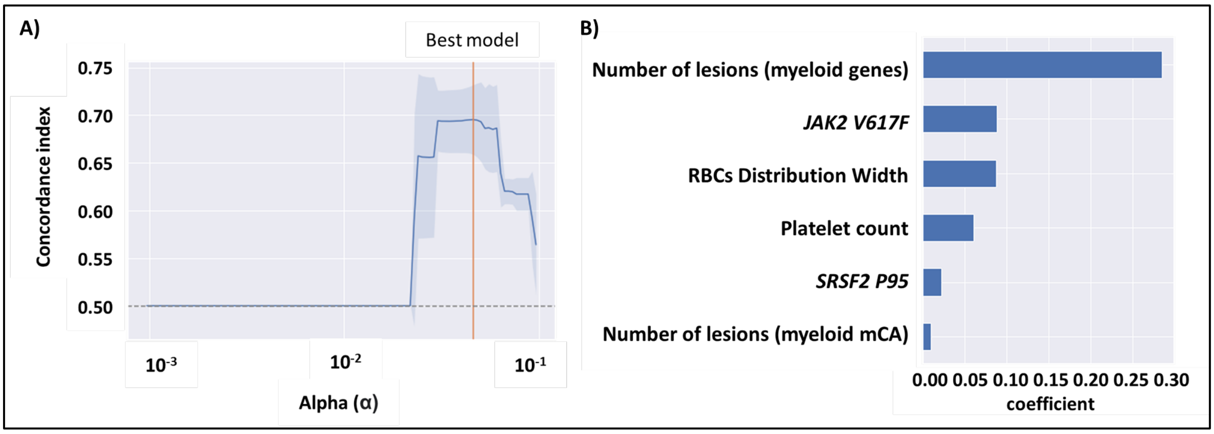


Figure 5-6: The selection of best COX-PH regularised by the Elastic-Net method

A) The relationship between C-index and 100 alpha ranged between 0.01 and 1 and the best model has the highest alpha. B) The selected best model has only six features with non-zero coefficient.

5.4.5 Prediction of myeloid malignancies by machine learning model

Machine learning models have the benefit of utilising high dimensional data. A Random Survival Forest model comprising 1000 trees and default parameters for the minimum number of samples in a leaf node (n=15) and splitting an internal node (n=10) was the most predictive which achieved a *C-index* of 0.78 and attributed the largest weights to platelet indices and number of lesions (myeloid genes) (Figure 5-7 and Figure 5-8).

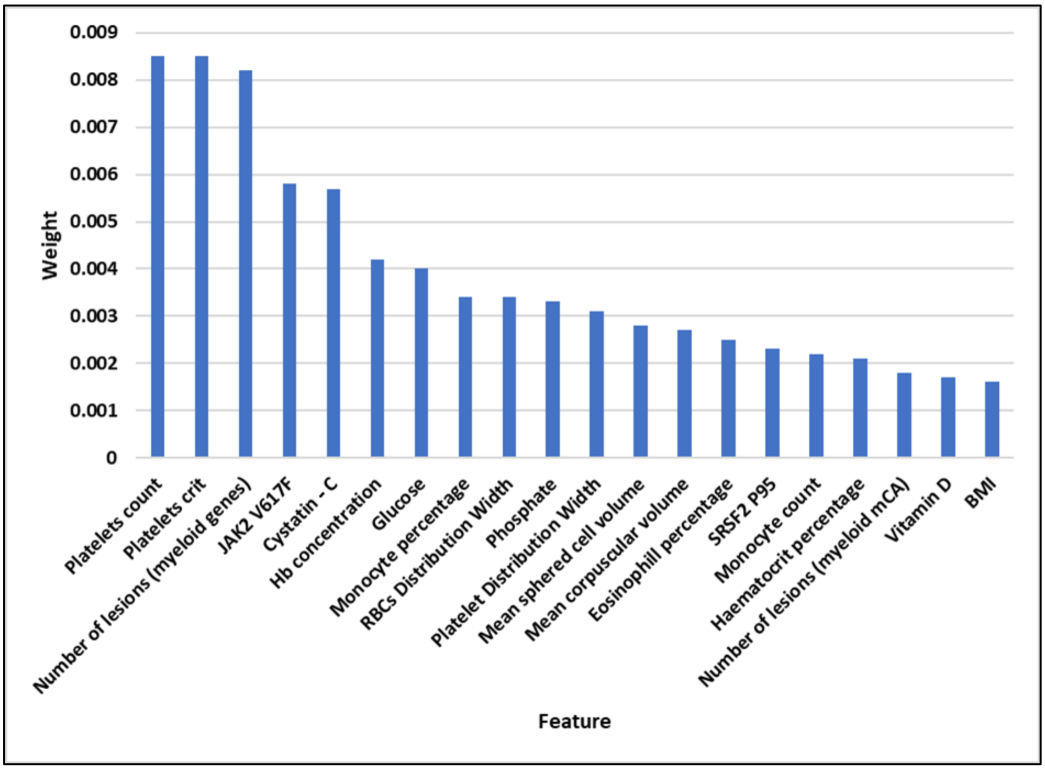


Figure 5-7: Top 20 features according to importance in the random survival forest model

Platelets indices and the number of lesions in myeloid genes were given the largest weights in the random survival model.

Three gradient-boosted models were evaluated under a range of different estimators for each model (i) Cox's partial likelihood with regression trees (ii) component-wise least squares base (iii) Accelerated Failure time model (Figure 5-8). The performance increase was much faster for the Cox's partial likelihood with regression trees which attained a larger C-index with 100 estimators (C-index = 0.76) than the component-wise least squares base (C-index = 0.7), and Accelerated Failure Model (n = 0.69) at the same number of estimators (Table 5-5).

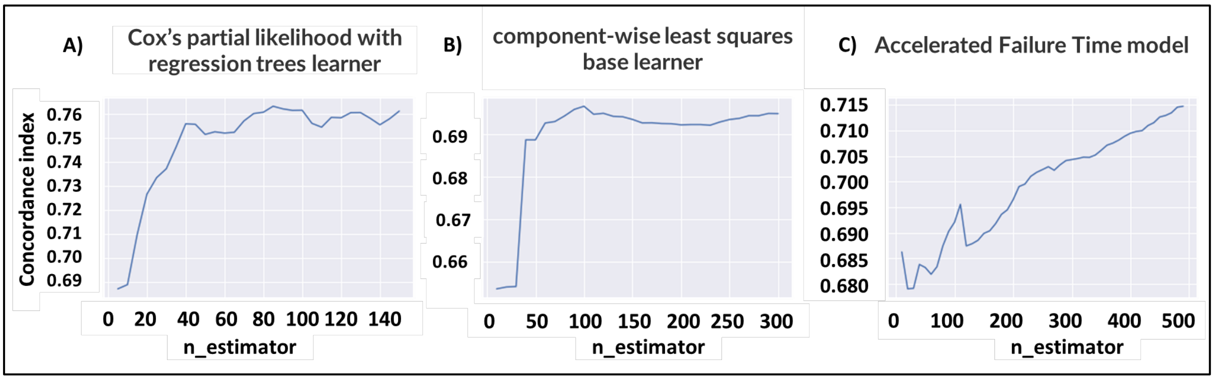


Figure 5-8: Evaluation of the gradient boosting models under three different base learners in comparison to number of estimators

Gradient boosting models were used to enhance 3 different base learners. The performance of each model was evaluated under a range of different estimators with intervals of 10. Cox's partial likelihood with regression trees learner (A) had the best performance and fastest increase in C-index at the lowest number of estimators in comparison to other models (B) and (C)

Table 5-5: The performance of models used in predicting the risk of myeloid malignancies and evaluated on the test data in the UK Biobank

Model	**Parameters**	**C-index**
Elastic-Net regularised Cox-PH	Alpha=0.046	0.57
Random Survival Forest	n=1000 trees	0.78
Gradient boosted models		
Cox's partial likelihood with regression trees learner	n-estimators=100	0.76
Component-wise least squares base	n-estimators=100	0.70
Accelerated Failure time	n-estimators=100	0.69

In survival models, the risk of the disease is not fixed but changes over time. At every time point, there are participants who developed a myeloid neoplasm before this time point, and there are participants who did not develop the disease yet. Consequently, the survival model performance becomes time dependent. I estimated the area under ROC for each model on test data at time intervals of 0.5 year across the study period, as shown in Figure 5-9. The time dependent AUC achieved the best performance with AUC of 0.64 at 6 years after recruitment (Figure 5-9). The random survival forest shows the best performance at all time points. In addition, all the tested models showed an expected decrease in performance with increasing time.

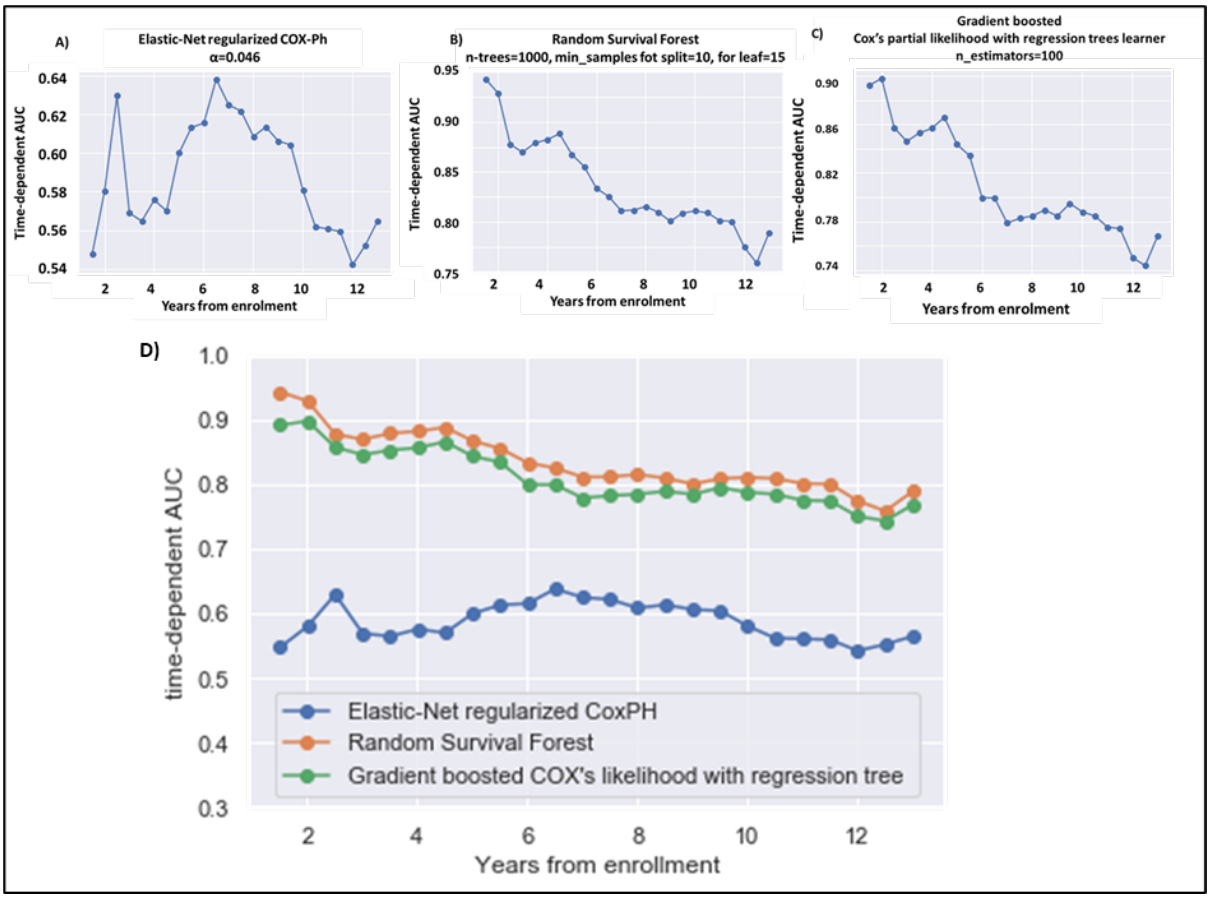


Figure 5-9: Time-dependent area under the ROC of three models evaluated to predict the risk of myeloid malignancies.

Each point presents the relationship of area under the ROC evaluated at time intervals of 0.5 years (years from enrolment) and estimated in the test data. The Random Survival Forest model (B) showed the best performance across all the time points in comparison to Elastic-Net regularised COX-PH model (A), and gradient boosted COX's partial likelihood (C). The performance of the ML models gradually decreased with increase of time from enrolment, as shown in (D).

5.5 Discussion

5.5.1 Models to predict the development of myeloid neoplasms

Several studies have reported the utility of CH in predicting the risk of myeloid malignancies in healthy individuals. These studies collectively proved that the risk of developing myeloid neoplasms was significantly associated with acquiring mutations in myeloid related genes with relatively high VAF. In addition, the number of lesions was a predictor for myeloid malignancies [42,86,87]. The UK biobank provides the opportunity to assess CH defined by both mCA from SNP array data, and driver mutations from WES data. In my analysis, a COX-PH model was regularised with Elastic-Net method to control the high dimensionality of the data by adding a penalty to coefficients that shrink coefficients to almost zero, and to select a subset of features to predict myeloid malignancies. The optimisation of the regularised COX-PH model produced a model that included six features with the number of myeloid targeted genes as the best predictor. Other features include RDW, platelet counts, *JAK2* V617F, *SRSF2* P95, and number of lesions (myeloid mCA). Using time-dependent area under the ROC analysis of the raw values of the six features (without fitting a model, Figure 5-10), the number of lesions (myeloid genes) have similar performance to RDW, a known chronic inflammation marker associated with CH [10], and an independent predictor for MDS [162].

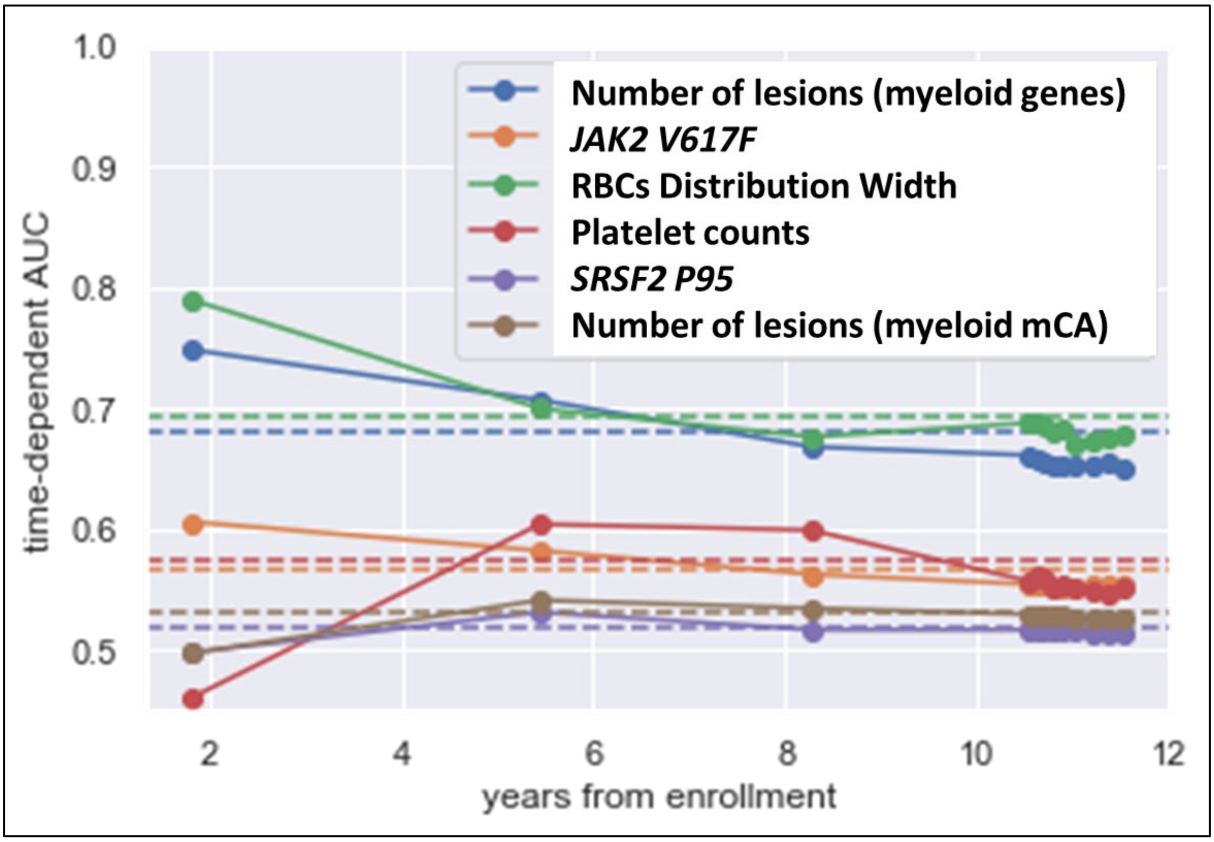


Figure 5-10: Time-dependent area under the ROC of six features selected by best model of Elastic-Net regularised COX-PH

Real-valued features, without fitting a survival model, were used to estimate the performance (area under the time-dependent ROC) of each feature at different time points. Number of lesions (myeloid genes) and RDW were the most discriminative features.

The best model of regularised COX-PH had a low performance when evaluated on test data that achieved *C-index* = 0.57, and time-dependent AUC ranges between = 0.54 and 0.64 during the study time. ML models were developed to handle censored data, which allowed the extension to survival analysis with capability to handle high dimensionality of data. All tested ML models achieved better performance in comparison to the regularised COX-PH model, but the RSF model was the best predictor in test data C-index=0.78, and AUC ranging between 0.9 and 0.74 at 2 years and 12.5 years from recruitment, respectively. In general an AUC between 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent [340]. For prediction, data from individual patients is assigned to each tree in the forest to reach a terminal node. A predictive risk score is the average of risk scores calculated across all trees, as the number of expected events at each terminal node. Platelet indices, and the number of lesions (myeloid genes) have the highest weight in the model. For example, individuals with three mutated myeloid genes have a much higher score in comparison to individuals with no mutated genes (Figure 5-11). In addition, we can notice that the predicted survival probability of individuals with three mutated genes is less than 80%, and 55% after 5 and 10 years, respectively. On the other hand, the selected controls with no mutations have a predicted survival probability of > 90% across all the study time. Clearly the predictive power of the model needs to be validated in an independent cohort, and this could be achieved using the remaining 250,000 UK Biobank participants for whom WES data has been released since my analysis was performed.

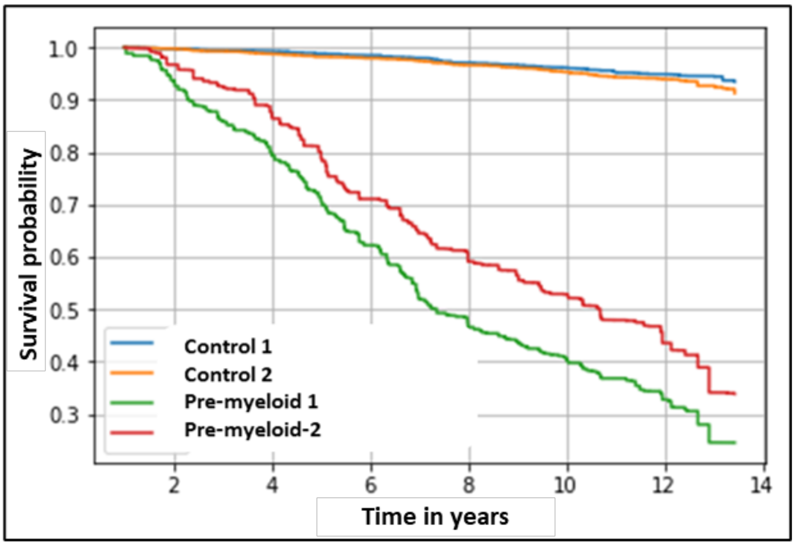


Figure 5-11: Predicted survival plot of 4 selected subjects according to number of lesions (myeloid genes) using RSF model

The survival plot shows 4 individuals. Two pre-myeloid subjects (pre-myeloid 1, and pre-myeloid 2) were chosen randomly from individuals with 3 lesions in myeloid genes. They developed myeloid malignancies after 2.05 and 3.02 years, respectively. Two controls (control 1, and control 2) were chosen randomly from individuals with no clonal lesions. They had follow-up time of 5.6 and 11.35 years, respectively. Survival portability at specific time point, is the mean value computed from all trees across the ensemble at the selected time point.

My analysis was based on using all features, driver mutations, mCA, blood counts, blood biochemistry, and lifestyle to build a model to predict myeloid malignancies, using the capability of ML models to handle high dimensional data. Other frameworks could be implemented but with higher computational and time cost. Ensemble-based integrative framework could be implemented by building separate models for, genetic data, blood biomarkers, and lifestyle factors. Next, the three models could be integrated by using different methods that include bagging, stacking, and blending [341].

In my model, predicted risk scores are largely dependent on platelet indices and number of lesions in myeloid genes. This suggests that platelet counts in particular should be recommended for follow up, e.g. by 6 monthly or annual blood tests. Overall, my study has highlighted that a small number of features that can predict myeloid malignancies. In addition, ML survival models can deal efficiently with large information combining both CH and other healthcare data in a single model with much higher performance in comparison to the traditional COX-PH models. The performance of the COX-PH model (C-index = 0.57) falls behind the previously published models that used CH to predict myeloid malignancies [42], but the best model, RSF model, had C-index 0.78 that was comparable with the previous study [42]. In addition, the performance of my model was likely negatively affected by the heterogeneity of myeloid malignancies that I included, and potentially by the low depth of-coverage of WES compared to targeted sequencing. To the best of my knowledge, no previous model has used biochemistry markers to predict myeloid malignancies in healthy individuals. My findings highlight the significant value of evaluating kidney function to stratify the risk of myeloid malignancies.

5.5.2 Strengths and limitations

Our definition of myeloid malignancies is based on aggregating data from different resources in the UK Biobank. However, the completeness of these resources varies in terms of the number of subjects with data and the date of last reported information. For example, primary care data was only available for 45% of the cohort, and cancer registry data was last updated in 2016. Although, these resources complement each other, we opted to exclude subjects that were diagnosed within one year of recruitment to minimise the number of undiagnosed cases. In addition, the UK Biobank cohort has a “healthy volunteer” bias and achieved a low recruitment response rate of only 5.5% [287,342,343]. Another kind of bias could occur due to high proportion of MPN (n=321) among our pre-myeloid subjects (n=726); other components were AML (n=155), MDS (n=141), and others (n=109). *JAK2* V617F is the most common somatic mutation in MPN [344], which may explain the high weight for platelet indices (1st and 2nd), and *JAK2* V617F (4th) in the RSF model. The prevalence of MPN in the UK Biobank might be explained by the fact that participants were recruited by invitation. It is notable that other cohorts have markedly different structures, for example a study of CH in the Biobank Japan included 215 subjects with myeloid malignancies: AML (n=90), MDS (n=100), MPN (n=5), and others (n=20) [86]. The median follow-up of our study was 11.7 years, and the median time to diagnosis was 7.1 years after recruitment. This is similar to the study to predict the risk of AML (median diagnosis = 7.6

years) [42]. Survival analysis deals with unfixed status of the disease across time such that the AUC varies according to time. In my analysis, the performance of the tested models decreased with follow-up time, so I expect that there will be no significant improvement in the models by increasing the follow-up time. Improvements in the predictive power of the model might have been improved by sequential molecular analysis to monitor changes in clone size over time and the appearance of new mutations, but the UK Biobank was limited to analysis of a single baseline sample. Finally, I considered myeloid malignancies as a single group and it will be important to break down predictive factors for AML, MDS and MPN as specific entities.

My study demonstrates the importance of blood measures in addition to CH to predict the risk of myeloid malignancies. For example, cystatin-C is a kidney function biomarker, and glomerular filtration rate (GFR), estimated from cystatin-C, was negatively associated with CH as showed in Chapter 4 [203]. In the random survival forest model, cystatin-C has the highest weight among 29 blood biochemistry measures and 5th among all features to predict the risk of myeloid malignancies.

Regarding the applied models, my analysis provided enough evidence for the superior performance of ML models, specifically RSF, over traditional COX-PH, however my analysis was restricted to the available ML models in the scikit-survival package. Other ML models could be tested such as Extreme Gradient Boosting (XGB), that uses decision trees in a gradient boosting model to support the decision, and is widely used in health settings [345]. Our models were evaluated by two methods: Harrell's C-index, and time-dependent AUC, and a permutation based-method implemented in Eli5 package was used to detect features that were important in the random survival forest model, but other tools could be applied to explain the model such as SHapley Additive exPlanations (SHAP) values , a method that calculates a value for each feature to explain the model [346]. This method was not compatible with RSF model developed by scikit-survival. As mentioned above, the true value of the models needs to be evaluated using an independent validation cohort of patients.

In summary, my study has shown that the use of ML models allows the integration of available data and generation of a model with good performance to predict the risk of myeloid malignancies, although with decreasing predictive value with increasing of follow-up time. An extension of my analysis could include the development of specialised models for predicting the risk of myeloid malignancies subclasses, AML, MPN, and MDS.

Chapter 6 Sex hormone binding globulin promotes the risk of age-related loss of the Y chromosome

6.1 Summary

Mosaic loss of the Y-chromosome (LOY) is the most common somatic alteration in men and a marker of clonal mosaicism. I aimed to assess the relationship between LOY and serum biomarkers in the UK Biobank (n=222,835 men; 44,558 with LOY) and explore the interaction with genetic factors. LOY was strongly associated with levels of sex hormone binding globulin (SHBG, β=0.11, P_{FDR}=2.34x10^{-86}), a key regulator of testosterone bioavailability associated with diverse disorders including cancer and autoimmune diseases. Furthermore, LOY was associated with total testosterone (TT, β=0.09, P_{FDR}=6.8x10^{-56}), but not bioavailable testosterone (P_{FDR}=0.11) or free testosterone (P_{FDR}=0.06). Mendelian randomisation indicated a causal effect of SHBG on LOY in the BioBank Japan using 8 SNPs (P=6.58x10^{-4}). There was no evidence for a causal effect of LOY, defined by 40 SNPs, on SHBG (P=0.46). Assessment of cis-eQTLs for 13 genes associated with LOY identified two that were also associated with levels of SHBG, however only rs7141210 (imprinted *DLK1-MEG3* locus) modified the relationship between SHBG and LOY (rs7141210-T/T; $P_{interaction}$=0.04) with low levels of SHBG seen specifically in men without LOY (β=-0.02, P=0.001), but not those with LOY (P=0.41). CH defined by somatic driver mutations was not associated with sex hormone levels but was associated with LOY defined as >30% of cells (OR=1.52, *P*=2.92x10^{-4}) and was even stronger for CH without discernible driver mutations (OR=2.46, P=5.09x10^{-34}). *TET2, TP53,* and *CBL* mutations were enriched in LOY cases defined as >30% of cells, but not *DNMT3A* and *ASXL1* mutations. My findings thus characterise the relationship between LOY, sex hormones and CH, and highlight an independent role for SHBG mediated by *DLK1-MEG3*.

6.2 Introduction

Age-related somatic loss of the Y-chromosome (LOY) in peripheral blood leukocytes is the most prevalent chromosomal alteration in men.[76] LOY has been identified in as many as 20% of the UK Biobank male participants (median age = 58), but only 10% of these (2% of all men) had LOY involving >20% of leukocytes [80,347,348]. A series of genome wide association studies (GWAS) have characterised inherited genetic variation that predisposes to LOY,[76,80,347] with 156 independent loci explaining up to 34% of the heritability [80,349]. LOY has also been causally linked to smoking behaviour [76,77,347], indicating that the environment as well as genetics is an important factor. LOY is associated with all-cause mortality, cancer mortality [76,350,351], and a wide range of non-malignant conditions [76,234,352-354]. LOY is also associated with variation in blood cell counts [348,355] and has long been recognised as a recurrent clonal cytogenetic finding in haematological malignancies where, in the absence of other changes, it is associated with a good prognosis [121]. LOY with large clone size has been linked to the presence of somatic mutations associated with haematological malignancies and an elevated risk of developing myeloid neoplasia in two recent, small studies of selected cases [356,357]. It has been suggested that LOY might be a broad marker of genomic instability across different tissues and may exert its effects by altering immune cell function [80,358].

CH is a widespread phenomenon characterised by the presence of expanded mutated clones of blood cells [293], predominantly in individuals over the age of 60. Large autosomal chromosomal alterations including gains, losses and copy number neutral loss of heterozygosity (CNN-LOH) inferred from SNP array data have been used to identify clonality [6,7], but CH is more commonly recognised by sequence analysis and the finding of pathogenic driver mutations associated with haematological malignancies, most commonly in the epigenetic regulators *DNMT3A*, *TET2*, and *ASXL1* but also a wide range of other genes [8-10,37,89,169,359]. Broad screens by whole exome sequencing (WES) or whole genome sequencing (WGS) have revealed that clonality in the absence of known driver mutations (unknown driver CH) is even more prevalent than CH with driver mutations [9]. Like LOY, CH has been linked to a wide range of malignant and non-malignant diseases [10,89,203,294,295,325,360]. Most prominently, CH defined by autosomal chromosomal alterations or driver mutations confers a 10-fold higher risk for the development of haematological malignancies, and recent studies showed a lineage-specific risk for mutations in genes associated with myeloid or lymphoid neoplasms [86,361]. The clinical consequences of unknown driver CH have not been defined, and the reason for clonality in these cases remains unclear. Furthermore, the extent to which mosaic LOY can be considered as CH has not been defined.

Although age is the major risk factor for both CH and LOY, it has become clear that there is significant overlap between genetic factors that predispose to both of these abnormalities. However this shared risk is complex, for example the T allele of rs2887399, located in the promoter of *TCL1A* at 14q32, is associated with a lower risk of LOY [78] as well as a lower risk of common forms of CH defined by 14q CN-LOH and *TET2* mutations [253,362]. This allele, however, is associated with an elevated risk of CH defined by *DNMT3A* mutations [363,364]. Broadly, inherited variants in cancer susceptibility genes and genes that are mutated in cancer feature prominently as risk factors for both CH and LOY [80,169,263,347,349,365]. As for external factors, prior chemotherapy is associated with CH characterised by *PPM1D* and *TP53* mutations [281], whereas smoking is associated with *ASXL1* mutations as well as LOY [202].

It is widely accepted that biochemical profiles change with age in a manner independent of specific disease states, for example ageing is associated with depletion of sex hormones [366]. Sex hormone binding globulin (SHBG) is a glycoprotein that binds to steroids with high affinity, with both 5α-dihydrotestosterone and testosterone binding much more strongly than oestradiol [367]. In men, circulating testosterone levels are regulated by SHBG, with on average 58% bound to SHBG, 40% bound to albumin, and 2% as free testosterone (FT) [368,369]. Binding to albumin is weak and so all non-SHBG-bound testosterone is considered as bioavailable testosterone (BAT) [370]. Ageing is associated with a decline in TT, FT and BAT and an increase in SHBG [366]. In this study I aimed to determine if any common biochemical measures, including sex hormone levels, are associated with LOY and to understand the interaction between genetic and biochemical factors.

6.3 Methods

6.3.1 Study cohort

The UK Biobank is a large prospective cohort described in detail elsewhere [178] involving approximately 500,000 individuals aged between 40 and 69 years at recruitment. Genome wide SNP data derived from peripheral blood leucocytes was available for most participants, and WES data for 200,631 participants at the time of analysis.

6.3.2 Mosaic loss of the Y chromosome

A previous study used the UK Biobank SNP data to identify males with mosaic LOY (n=44,588; 20% of evaluable males) using a method which compared allelic intensities for statistically phased haplotypes of the pseudo-autosomal region 1 (PAR1) [80]. This method for detecting LOY was considered to be less error prone than those based on the median genotyping intensity over the non-pseudoautosomal region of the Y chromosome and was able to detect mosaicism with a clonal fraction down to 1% [80,253]. The spectrum of LOY was categorised according to clonal fraction by considering the median change of B-allele frequency (BAF), specifically BAFs of 0.026, 0.056, and 0.088 corresponding to clonal fractions of 10%, 20%, 30%, respectively as described [78].

6.3.3 Biochemistry markers and sex hormones

Measurements of 29 biochemistry markers were available from serum samples collected on recruitment to the UK Biobank [178]. Mass action equations were used to calculate FT and BAT from measurements of SHBG, TT and albumin as described [371]. Further details regarding the biochemical assay methods and external quality assurance are available at https://biobank.ndph.ox.ac.uk/showcase/showcase/docs/serum_biochemistry.pdf.

6.3.4 The relationship between biochemistry markers and LOY

I focused on the subset of male UK Biobank participants who were evaluable for LOY assessment (n=222,835) [80]. The relationship between LOY and each of 31 biomarkers (29 measured and 2 calculated) was tested using multivariable linear regression in R. Continuous measures for each sex hormone were transformed into a normal distribution using inverse normal rank transformation and used as the dependant variable. The independent variables were LOY as a binary predictor, age, smoking status (never, previous, current), and the first 10 genetic principal components (10 PCA). Effect sizes were reported as beta coefficients (β) with 95% confidence intervals (95% CI). P-values were adjusted for 31 tests using the False Discovery Rate (FDR) method [258]. The average measure of sex hormones in participants with LOY were compared to those in LOY free controls using Mann–Whitney U tests.

6.3.5 Allelic SHBG score

Thirteen genetic variants were associated with circulating SHBG levels by a previous two-stage GWAS, including 10 variants that achieved genome wide significance plus 3 independent cis variants that were identified by conditional analysis of the *SHBG* gene [372]. After excluding one SNP with heterogeneity towards females (*P* = 0.02, rs440837) and a second SNP on the X-chromosome (rs1573036), 11 SNPs were considered in relation to LOY. To consider both precision and power, two allelic scores were constructed. The first consisted of all 11 SNPs (rs12150660, rs1641537, rs1625895, rs6258, rs17496332, rs2411984, rs293428, rs780093, rs7910927, rs8023580 and rs4149056) while the second was based only on the 4 SNPs located in the vicinity of *SHBG* (rs12150660, rs1641537, rs1625895 and rs6258). Each score was calculated as the sum of number of risk factor-increasing alleles per SNP weighted by their corresponding genetic effect size. The allelic scores were calculated using Plink V1.9,[255] and transformed into a normal distribution using inverse normal rank transformation. A multivariable logistic regression model adjusted for age, sex, smoking status and 10 PCA was used to assess the relationship between LOY status (binary, dependent) and *SHBG* allelic scores (independent, continuous).

6.3.6 Mendelian Randomisation

Mendelian randomisation (MR) was used to assess the possibility of a causal relationship between SHBG and LOY using germline SNPs associated with circulating SHBG as instrumental variables, following the STROBE guidelines [320]. Eight SNPs were used as instrumental variables without considering estimated genetic effect sizes for LOY in 95,380 men recruited to the BioBank Japan (BBJ) [81]. Three of the eleven SNPs that were used to generate the allelic score (rs12150660, rs6258 and rs2411984) were excluded as genotypes were unavailable and/or non-informative in BBJ. I applied an inverse variance weighted model (IVW) with fixed effect using the TwoSamplesMR package [242]. To explore the effect of SNP selection, I repeated the MR analysis using a genome wide significance threshold ($P < 5x10^{-8}$) to select four SNPs (rs1641537, rs1625895, rs293428 and rs7910927) associated with SHBG [372]. I also performed leave-one-out analysis to evaluate the effect of each SNP on the analysis. To assess the effect of LOY on SHBG levels I examined SNPs (n=50) previously associated with LOY [81] in BBJ. 10 SNPs were excluded as genotypes were unavailable and/or non-informative in the UK Biobank. The remaining 40 SNPs were used as instrumental variables without considering estimated genetic effect sizes for SHBG in the UK Biobank men.

6.3.7 Identification of CH

A propensity score matching method in R [335] was used to select one control per case and to match for age in comparison to participants with LOY. Somatic mutations were called in LOY samples and matched controls using GATK (Version 4.1.9) and Mutect2 [336] to process individual CRAM files in the tumour-only mode. Following best practice guidelines (https://gatk.broadinstitute.org/hc/en-us/articles/360035531132), a Panel Of Normal (PON; version 23rd August 2017) from the Broad Institute that were generated using Mutect2 on samples from the 1000 genomes project to identify recurrent artifacts, and germline variants from GnomAD were used to remove artefacts and germline variants. To identify putative somatic driver mutations, the analysis was restricted to rare variants with a minor allele frequency (MAF) <0.01 in GnomAD and a minimum number of reads supporting the mutated allele: 3 reads for point mutations and 6 reads for indels. Variants that satisfied either of the following two criteria were selected: first; recurrent driver mutations as defined in Chapter 4, second singleton variants that passed all Mutect2 filters with variant allele frequencies (VAF) between 0.1 and 0.2 [9]. Driver mutations were classified into myeloid or lymphoid according to a published list of genes associated with myeloid neoplasms (n=76) [89] and genes associated with acute lymphoblastic leukaemia or chronic lymphocytic leukaemia in the Cancer Gene Census (CGC),[337] respectively (Table 6-1). Participants with variants in myeloid genes were considered as having myeloid CH (n= 2,890), and those with variants in lymphoid genes as having lymphoid CH (n=532). Cases with mutations in genes involved in both myeloid and lymphoid neoplasms were considered as myeloid CH. To identify participants with evidence of clonality in the absence of pathogenic mutations in known driver genes, I also identified variants in all coding genes that were not defined as myeloid or lymphoid. Participants with singleton variants in any gene with VAF range between 0.1 and 0.2 not defined as myeloid or lymphoid were considered as having unknown driver CH (n=15,874).

Table 6-1: Genes used to define myeloid CH and lymphoid CH

Myeloid genes	*ASXL1,ASXL2,BCOR,BCORL1,BRAF,BRCC3,CBL,CBLB,CEBPA,CREBBP,CSF1R,CSF3R,CTCF CUX1,DNMT3A,EED,EP300,ETNK1,ETV6,EZH2,FLT3,GATA1,GATA2,GATA3,GNA13,GNAS, GNB1,IDH1,IDH2,IKZF1,IKZF2,IKZF3,JAK1,JAK2,JAK3,KDM6A,KIT,KRAS,LUC7L2,MLL,MLL2, MPL,NF1,NPM1,NRAS,PDS5B,PDSS2,PHF6,PHIP,PPM1D,PRPF40B,PRPF8,PTEN,PTPN11, RAD21,RUNX1,SETBP1,SETD2,SETDB1,SF1,SF3A1,SF3B1,SRSF2,SMC1A,SMC3,STAG1, STAG2,SUZ12,TET2,TP53,U2AF1,U2AF2,WT1,ZRSR2,CALR*
Lymphoid genes	*ABL1,AFF3,AFF4,BCL11B,BCL9,BCR,CCNC,CCND1,CDK6,CDKN1B,CNOT3,CRLF2,DNM2, ECT2L,ELN,EPS15,EWSR1,FBXW7,FCGR2B,FOXP1,HLF,IGH,IL7R,IRS4,KMT2A,LCK,LEF1, LMO1,LMO2,LYL1,LYN,MLLT11,MLLT3,MYCL,NCKIPSD,NOTCH1,NT5C2,NUP214,OLIG2, P2RY8,PAX5,PBX1,PICALM,PML,PTPRC,RAP1GDS1,RB1,RPL10,RPL5,SET,SH2B3,SMAD4, STIL,TAF15,TAL2,TBL1XR1,TCF3,TFPT,TLX1,TLX3,TRA,TRB,ZNF384,ZNF521,BCL11A,BCL2, BCL3,BCL6,BIRC3,BTG1,BTK,CCND2,CHD2,CHST11,DDX3X,FAT1,FSTL3,LRP1B,MAPK1, MYC,NFKBIE,POT1,TCL1A,XPO1,ABCA12,ACSL3,ATM,CCND3,CCR7,DYRK1A,FGFR2,FOXO3, GPR158,HIST1H1C,ITPR2,KMT2C,KMT2D,LRP5,MST1,MYD88,NFE2,PTPRA,RBBP4,RPS3A, SIN3A,STAT3,STAT5B,TMEM30A,VMA21,VSTM4,ABCD2,AKAP11,BTBD10,CACNA1C, CARD11,CCDC18,CD70,CD79B,CHL1,CROCC,CRTC1,CTSS,CWH43,CXXC1,FBXO7,FNDC3B, FRYL,GABRG3,GRID2,HERC2,HNRNPCL1,IGLL5,IRF4,ITGA1,ITIH5,KIF3A,KIF5B,KLHL6, KLHL7,MED12,MTFR2,NLK,NOTCH2,PABPC1,PIM1,PLSCR1,POSTN,POU2F2, PSMD10,RPL37,SLC16A7,SLC17A6,TBC1D32,TPR,TUBB2A,VAV1,ZYG11B*

6.3.8 The relationship between LOY and CH

The association between LOY and all variants, driver variants, myeloid CH, lymphoid CH, and unknown driver CH were tested using logistic regression in R. I further assessed the relationship with LOY clone size categories (≤10%, 10%-20%, 20%-30%, and >30%) using Fisher’s exact tests. The strength of the association was reported as odds ratio (OR) with 95% CI. P-values were adjusted for 20 tests using the FDR method. To test the relationship at the individual driver gene level, the analysis was focused on genes mutated in ≥3 males with LOY in >30% of cells.

6.3.9 Assessment of the co-existence of LOY and CH

The co-existence of CH and LOY was tested by assessing the relationship between the BAF for LOY and VAF of driver mutations. For cases with two or more mutations the highest VAF was used. To avoid excess CH with VAF between 0.1-0.2, the analysis for myeloid and lymphoid CH was restricted to VAFs detected for recurrent driver mutations as defined in chapter 4. I assessed the relationship with LOY in ≤10% and >10% of cells. The dependent variables were VAF, age and smoking status (never, previous, current). The strength of the association was reported as β coefficient with 95% CI. P-values were adjusted for 6 tests using the FDR method.

6.3.10 The relationship between CH and sex hormones

The association between CH and sex-hormone levels were tested using linear regression in R, with normally transformed sex hormone as the independent variable. The dependent variables were driver mutation state as a binary predictor, age, smoking status (never, previous, current), and 10 PCA. The association was reported as β coefficient with 95% CI. P-values were adjusted for 4 tests using the FDR method.

6.3.11 Expression quantitative trait analyses

The eQTLGen database incorporates expression quantitative trait locus (eQTL) data from blood samples from a total of 31,684 individuals [373]. To select a genetic proxy for gene expression, I filtered cis-eQTLs within a distance <1 Mb, and FDR < 0.05 and selected the SNP with the smallest FDR value, with no other genes showing a stronger association with the selected SNP to minimise horizontal pleiotropy. I restricted the analysis to directly genotyped SNPs with MAF >0.05 in the UK Biobank. 19 SNPs associated with LOY were associated with 27 genes by position, biological function, expression, or nonsynonymous variants in the gene [347]. My analysis was restricted to 13 of these genes for which an expression proxy was identified.

6.3.12 The assessment of the interaction between eQTL, LOY and SHBG levels

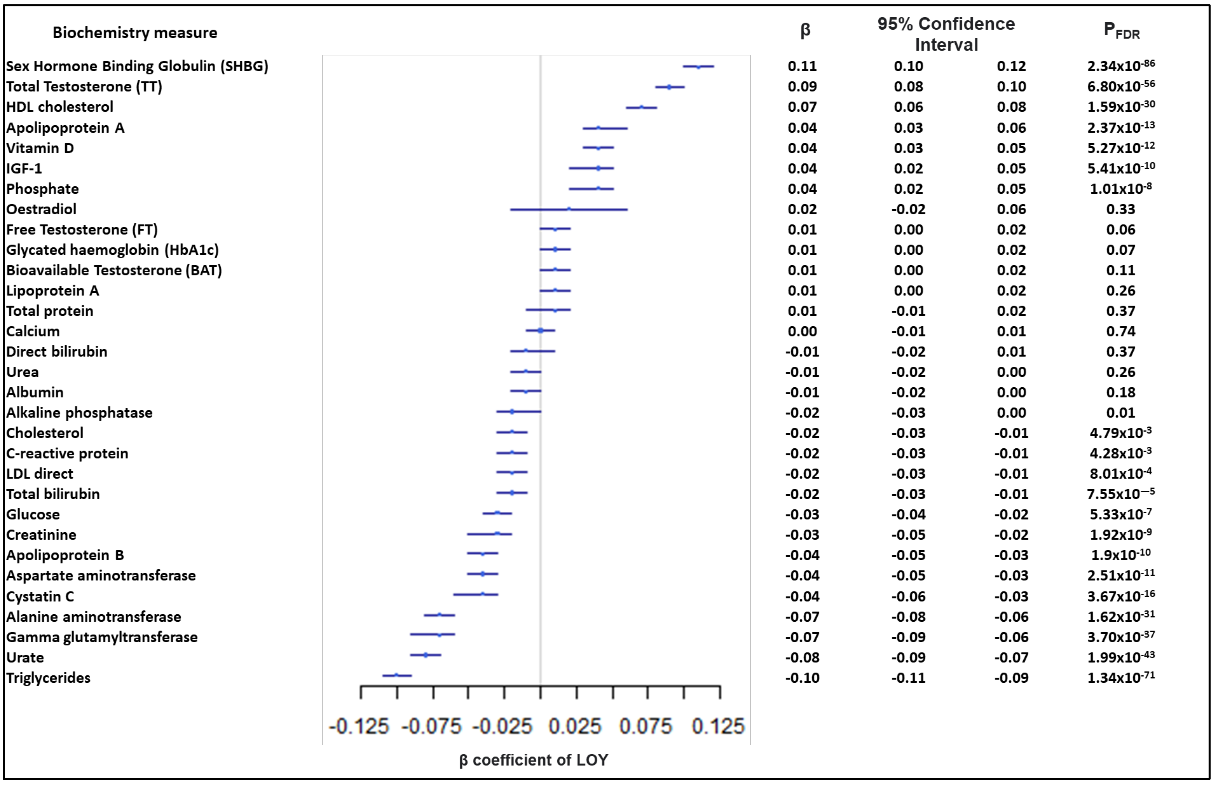
The most significant eQTL SNP for each gene was encoded according to an additive model for the risk allele (0/1/2) in the UK Biobank and was factorised in each model with 0 as the reference (0/1 for heterozygous, and 0/2 if homozygous for the risk allele). The following statistical tests were applied, and adjusted for age, smoking, 10 PCA, and multiple tests by the FDR method. First, the relationship between each eQTL and LOY was assessed using logistic regression in R where LOY (binary) was the dependent variable and eQTL was the independent variable. Second, the relationship between eQTLs and SHBG levels was assessed using linear regression in R where SHBG (continuous) was taken as the dependent variable and transformed to a normal distribution using rank transformation, and eQTLs were the independent variable. Finally, if an eQTL was significantly associated with both LOY and SHBG, the interaction effect of the eQTL and LOY on SHBG regression was assessed by linear regression in R. Inverse normal rank transformed SHBG was considered as a continuous dependent variable, and each of eQTL, LOY, and eQTL x LOY (interaction effect) as the independent variable. To visualise the interaction effect in my models, I used '*interactions'* in R, a tool that was developed to interpret statistical interactions in regression models [374].

6.4 Results

6.4.1 The relationship between LOY and biochemistry markers

To investigate the relationship between LOY and serum biomarkers, I used previously published calls of LOY that were generated by utilizing long-range phasing information to analyse allele-specific genotyping intensities of 1,239 variants in the pseudo-autosomal region 1 [80]. I restricted my analysis to the 222,835 males who passed QC, of whom 44,558 (20%) had LOY. Of these, the majority (n=31,952; 72%), had an estimated LOY clonal fraction of <10%. I compared the presence or absence of LOY with 29 biochemistry parameters that were directly measured by the UK Biobank, as well as estimated levels of FT (median = 0.21 nmol/L; range = 0.003 - 1.93) and BAT (median = 5.1 nmol/L; range = 0.09 - 45.68) that were derived from measurements of SHBG, TT and albumin [375].

On multivariate analysis adjusted for age, sex, 10 PCA, smoking and multiple tests, I found that LOY as a binary predictor was most strongly associated with elevated levels of SHBG (β = 0.11, 95% CI: 0.10 - 0.12, P = 2.34×10^{-86}). SHBG binds steroids [367] and it is notable that the second strongest positive association was with TT (β = 0.09, 95% CI: 0.08 - 0.10, P = 6.80×10^{-56}). There was no association, however, between LOY and FT (P = 0.06) or BAT (P = 0.11) (Figure 6-1 and Table 6-2). Participants with LOY had higher levels of SHBG and TT (SHBG: median nmol/L = 41.54 vs 35.86, P < 0.001; TT: median nmol/L = 11.74 vs 11.58, P < 0.001; Mann-Whitney U tests) but lower levels of FT (median nmol/L = 0.19 vs 0.20, P < 0.001), and BAT (median nmol/L = 4.78 vs 5.18, P < 0.001). My observational results points to a direct relationship between levels of SHBG and LOY, that cannot be explained by age, smoking, population stratification, or free/bioavailable testosterone (Figure 6-2).



Biochemistry measure	β	95% Confidence Interval		P_{FDR}
Sex Hormone Binding Globulin (SHBG)	0.11	0.10	0.12	$2.34x10^{-86}$
Total Testosterone (TT)	0.09	0.08	0.10	$6.80x10^{-56}$
HDL cholesterol	0.07	0.06	0.08	$1.59x10^{-30}$
Apolipoprotein A	0.04	0.03	0.06	$2.37x10^{-13}$
Vitamin D	0.04	0.03	0.05	$5.27x10^{-12}$
IGF-1	0.04	0.02	0.05	$5.41x10^{-10}$
Phosphate	0.04	0.02	0.05	$1.01x10^{-8}$
Oestradiol	0.02	-0.02	0.06	0.33
Free Testosterone (FT)	0.01	0.00	0.02	0.06
Glycated haemoglobin (HbA1c)	0.01	0.00	0.02	0.07
Bioavailable Testosterone (BAT)	0.01	0.00	0.02	0.11
Lipoprotein A	0.01	0.00	0.02	0.26
Total protein	0.01	-0.01	0.02	0.37
Calcium	0.00	-0.01	0.01	0.74
Direct bilirubin	-0.01	-0.02	0.01	0.37
Urea	-0.01	-0.02	0.00	0.26
Albumin	-0.01	-0.02	0.00	0.18
Alkaline phosphatase	-0.02	-0.03	0.00	0.01
Cholesterol	-0.02	-0.03	-0.01	$4.79x10^{-3}$
C-reactive protein	-0.02	-0.03	-0.01	$4.28x10^{-3}$
LDL direct	-0.02	-0.03	-0.01	$8.01x10^{-4}$
Total bilirubin	-0.02	-0.03	-0.01	$7.55x10^{-5}$
Glucose	-0.03	-0.04	-0.02	$5.33x10^{-7}$
Creatinine	-0.03	-0.05	-0.02	$1.92x10^{-9}$
Apolipoprotein B	-0.04	-0.05	-0.03	$1.9x10^{-10}$
Aspartate aminotransferase	-0.04	-0.05	-0.03	$2.51x10^{-11}$
Cystatin C	-0.04	-0.06	-0.03	$3.67x10^{-16}$
Alanine aminotransferase	-0.07	-0.08	-0.06	$1.62x10^{-31}$
Gamma glutamyltransferase	-0.07	-0.09	-0.06	$3.70x10^{-37}$
Urate	-0.08	-0.09	-0.07	$1.99x10^{-43}$
Triglycerides	-0.10	-0.11	-0.09	$1.34x10^{-71}$

Figure 6-1: The relationship between LOY and biochemistry markers

The relationship between LOY and each of 31 biomarkers (29 measured and 2 calculated) was tested using multivariable linear regression in R in 222,835 UK Biobank males.

Table 6-2: Linear regression results of LOY against 31 biochemistry markers.

	Coefficient	95% confidence intervals		P(FDR)
SHBG	0.11	0.10	0.12	2.34×10^{-86}
Testosterone	0.09	0.08	0.10	6.80×10^{-56}
HDL cholesterol	0.07	0.06	0.08	1.59×10^{-30}
Apolipoprotein A	0.04	0.03	0.06	2.37×10^{-13}
Vitamin D	0.04	0.03	0.05	5.27×10^{-12}
IGF-1	0.04	0.02	0.05	5.41×10^{-10}
Phosphate	0.04	0.02	0.05	1.01×10^{-8}
Oestradiol	0.02	-0.02	0.06	0.33
Free Testosterone	0.01	0.00	0.02	0.06

Glycated haemoglobin (HbA1c)	0.01	0.00	0.02	0.07
Bioavailable Testosterone	0.01	0.00	0.02	0.11
Lipoprotein A	0.01	0.00	0.02	0.26
Total protein	0.01	-0.01	0.02	0.37
Calcium	0.00	-0.01	0.01	0.74
Direct bilirubin	-0.01	-0.02	0.01	0.37
Urea	-0.01	-0.02	0.00	0.26
Albumin	-0.01	-0.02	0.00	0.18
Alkaline phosphatase	-0.02	-0.03	0.00	0.01
Cholesterol	-0.02	-0.03	-0.01	4.79×10^{-3}
C-reactive protein	-0.02	-0.03	-0.01	4.28×10^{-3}
LDL direct	-0.02	-0.03	-0.01	8.01×10^{-4}
Total bilirubin	-0.02	-0.03	-0.01	7.55×10^{-5}
Glucose	-0.03	-0.04	-0.02	5.33×10^{-7}
Creatinine	-0.03	-0.05	-0.02	1.92×10^{-9}
Apolipoprotein B	-0.04	-0.05	-0.03	1.93×10^{-10}
Aspartate aminotransferase	-0.04	-0.05	-0.03	2.51×10^{-11}
Cystatin C	-0.04	-0.06	-0.03	3.67×10^{-16}
Alanine aminotransferase	-0.07	-0.08	-0.06	1.62×10^{-31}
Gamma glutamyltransferase	-0.07	-0.09	-0.06	3.70×10^{-37}
Urate	-0.08	-0.09	-0.07	1.99×10^{-43}
Triglycerides	-0.10	-0.11	-0.09	1.34×10^{-71}

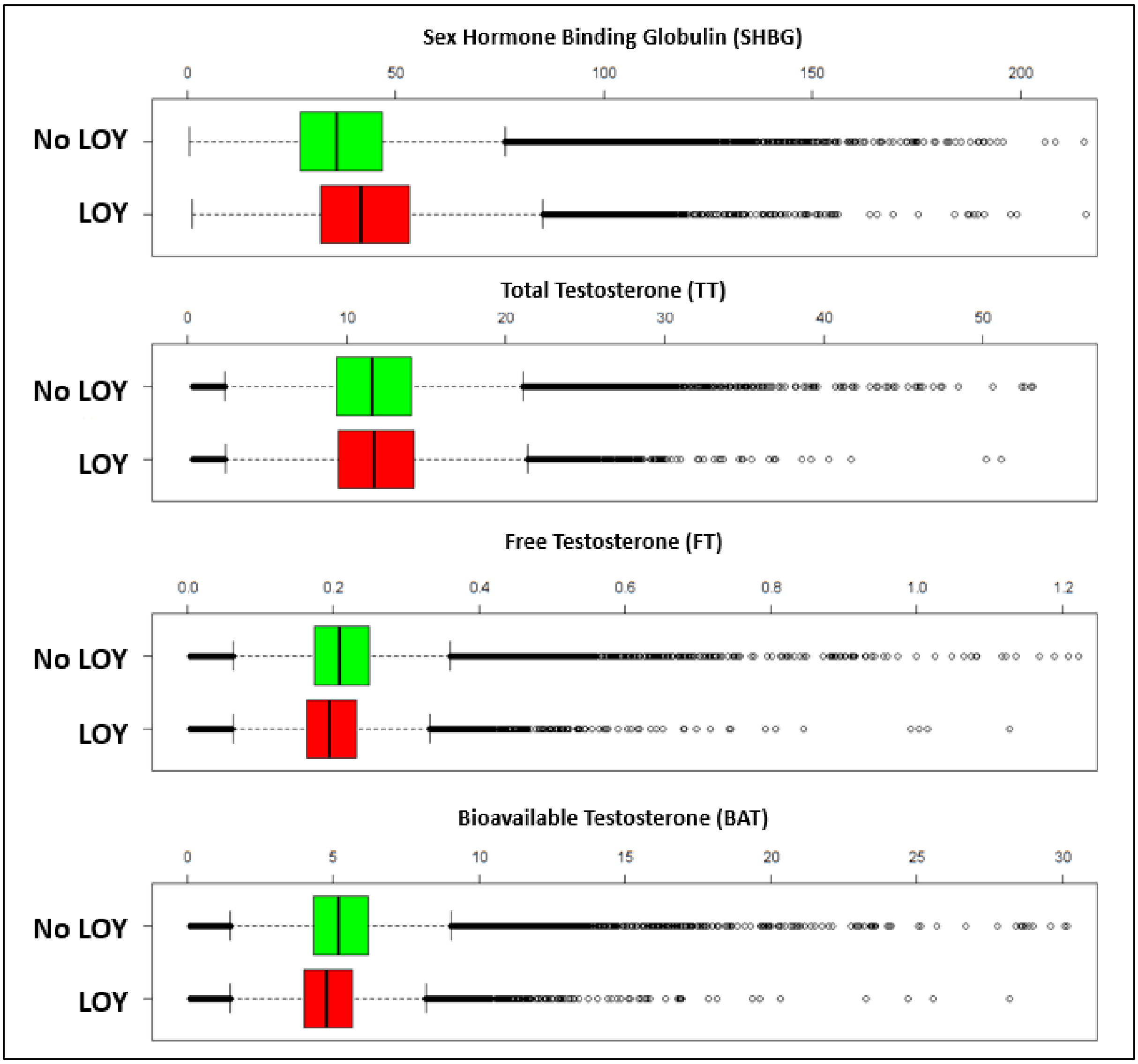


Figure 6-2: The relationship between LOY and levels of sex hormones

The box plots summarise serum sex hormone measurements in participants without LOY (n=178,277) and with LOY (n=4,458). SHBG: median nmol/L = 41.54 vs 35.86, P < 0.001; TT: median nmol/L = 11.74 vs 11.58, P < 0.001; FT: median nmol/L = 0.19 vs 0.20, P < 0.001; BAT: median nmol/L = 4.78 vs 5.18, P < 0.001.

6.4.2 The relationship between genetically defined SHBG and LOY

Published GWAS have identified multiple genetic determinants of SHBG levels in serum [372]. To understand the relationship between LOY and SHBG I generated allelic scores to summarise the genetic variation associated with SHBG and evaluated the scores as predictors of LOY. Since the allelic scores were derived from independent cohorts they represent unbiased instruments to assess the relationship with LOY in the UK Biobank [224]. I found that genetically predicted SHBG was significantly associated with the finding of LOY in the UK Biobank using the score estimated from 11 independent SNPs (OR = 1.02, 95% CI: 1.01 – 1.04, P = 5.59x10^{-5}).

To assess the possibility of a causal relationship between SHBG and LOY I performed MR analysis (Figure 6-3). In a liberal analysis, 8 independent SNPs were used to estimate the effect of SHBG on LOY (Table 6-3). Using an IVW model with fixed effect I identified a positive causal relationship (β = 0.15, 95% CI: 0.06 - 0.23, P = 6.58x10^{-4}). In a more conservative analysis restricted to 4 SNPs associated with SHBG with P < 5x10^{-8}, the effect of SHBG on LOY was confirmed (β = 0.17, 95% CI: 0.07 - 0.26, P = 7.28x10^{-4}). Leave-one-out analysis (

Table 6-4) found that significance was lost when rs7910927 at 10q21.3 within *JMJD1C* was excluded (liberal analysis: β = 0.08, 95% CI: -0.02 - 0.17, P = 0.13; conservative analysis: β = 0.07, 95% CI: -0.04 - 0.19, P = 0.22). To assess the possibility of bidirectional effect I utilised 40 SNPs with P < 5x10^{-8} associated with LOY in Japanese men [81] to measure their effect on SHBG levels in the UK Biobank men [376] but no significant relationship was found (β = 0.02, 95% CI: -0.88 - 0.13, P = 0.46).

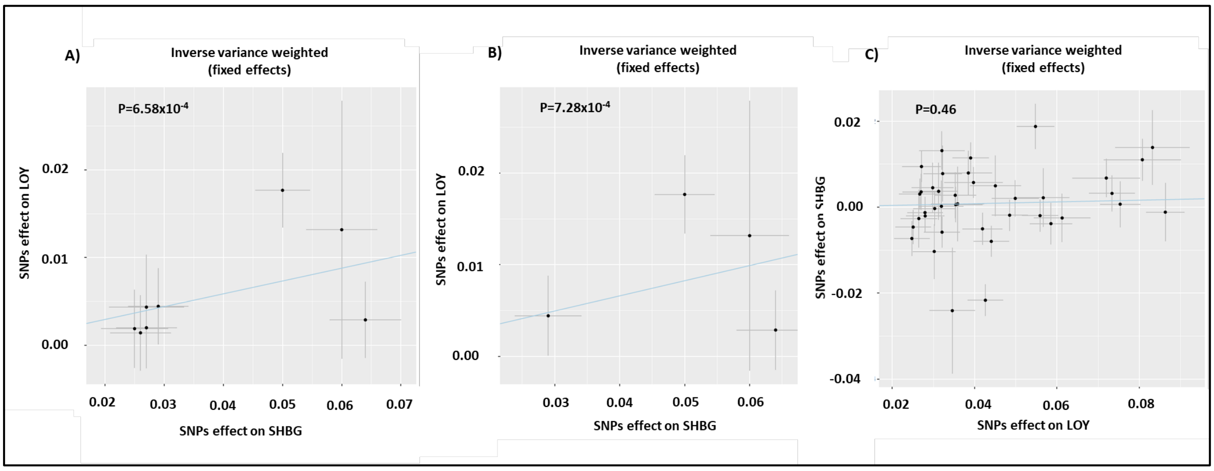


Figure 6-3: Mendelian randomisation using an inverse variance weighted model to estimate the causal relationship between SHBG and LOY

(A) Liberal analysis using 8 independent SNPs associated with SHBG. The IVW test estimated a significant positive effect of SHBG on LOY (P = 6.58x10^{-4}). (B) Conservative analysis using 4 SNPs (P = 7.28x10^{-4}). (C) analysis using 40 independent SNPs associated with LOY. The IVW test estimated no effect of LOY on SHBG (P = 0.46). The line of regression is indicated in blue, and the axes show β coefficients for SNP effects on SHBG and LOY.

Table 6-3: Mendelian randomization using 8 SNPs associated with SHBG as instrumental variables to assess the causal effect of SHBG on LOY

SNP	Position	Effect allele	Other allele	Gene	β exposure	β outcome	Se outcome	P outcome	P exposure	Se exposure	Main analysis	strict analysis
rs4149056	12:21331549	T	C	*SLCO1B1*	0.03	4.34 x 10^{-3}	0.01	0.5	1.50 x 10^{-5}	6.30 x 10^{-3}	Y	N
rs8023580	15:96708291	T	C	*NR2F2*	-0.03	-1.88 x 10^{-3}	4.45 x 10^{-3}	0.65	5.00 x 10^{-6}	5.61 x 10^{-3}	Y	N
rs780093	2:27742603	T	C	*GCKR*	-0.03	-1.40 x 10^{-3}	4.29 x 10^{-3}	0.72	7.00 x 10^{-8}	5.10 x 10^{-3}	Y	N
rs17496332	1:107546375	A	G	*PRMT6*	-0.03	-1.98 x 10^{-3}	4.61 x 10^{-3}	0.69	2.00 x 10^{-7}	5.10 x 10^{-3}	Y	N
rs293428	4:69591782	A	G	*UGT2B15*	-0.03	-4.43 x 10^{-3}	4.34 x 10^{-3}	0.33	3.00 x 10^{-8}	5.10 x 10^{-3}	Y	Y
rs7910927	10:65138910	T	G	*JMJD1C*	-0.05	-0.02	4.26 x 10^{-3}	2.60E-5	1.00 x 10^{-25}	4.59 x 10^{-3}	Y	Y
rs1625895	17:7578115	T	C	*SHBG*	-0.06	-0.01	0.01	0.35	1.75 x 10^{-21}	6.00 x 10^{-3}	Y	Y
rs1641537	17:7545721	T	C	*SHBG*	-0.06	-2.88 x 10^{-3}	4.32 x 10^{-3}	0.52	1.20 x 10^{-24}	6.00 x 10^{-3}	Y	Y

exposure represents the information provided by the GWAS of SHBG

outcome represents the information provided by the GWAS of LOY

beta coefficient (β), P-value)P) , and standard error (se)

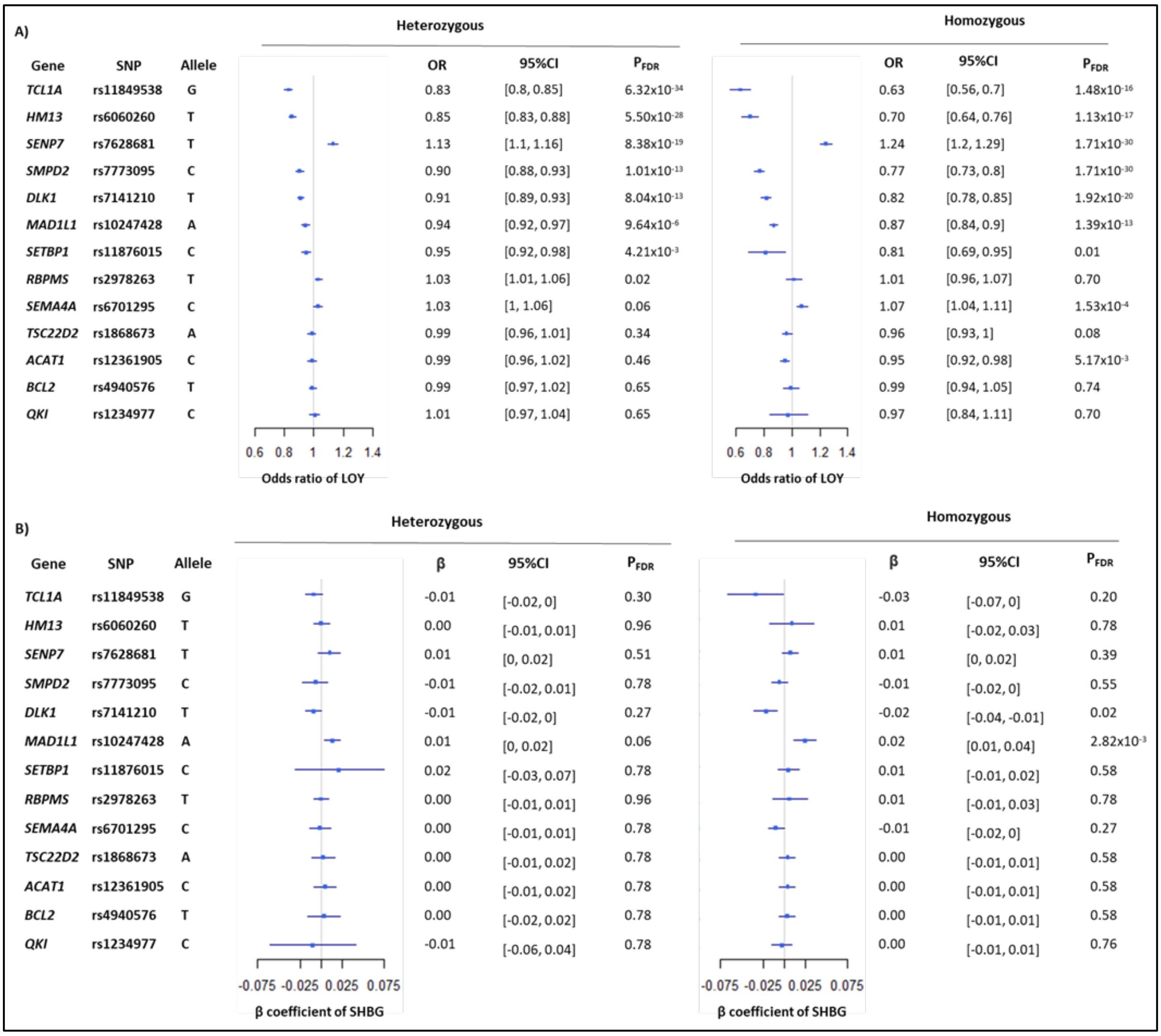
Table 6-4: Leave-one-out analysis

	Analysis with all 8 SNPs				**Analysis with 4 SNPs**			
Excluded SNP	**Coefficient**	**95% confidence interval**		**P**	**Coefficient**	**95% confidence interval**		***P***
rs1641537	0.22	0.11	0.32	1.11×10^{-4}	0.30	0.16	0.44	2.70×10^{-5}
rs1625895	0.14	0.06	0.23	9.65×10^{-4}	0.16	0.07	0.26	1.09×10^{-3}
rs293428	0.15	0.06	0.23	1.15×10^{-3}	0.17	0.07	0.27	1.27×10^{-3}
rs7910927	0.08	-0.02	0.17	0.13	0.07	-0.04	0.19	0.22
rs17496332	0.15	0.06	0.24	6.55×10^{-4}				
rs4149056	0.15	0.06	0.23	8.72×10^{-4}				
rs8023580	0.15	0.06	0.24	6.61×10^{-4}				
rs780093	0.15	0.07	0.24	5.82×10^{-4}				
All	0.15	0.06	0.23	6.58×10^{-4}	0.17	0.07	0.26	7.28×10^{-4}

6.4.3 The effect of gene expression on the relationship between SHBG and LOY

A previous GWAS identified 19 genomic regions associated with LOY, all of which were confirmed in a subsequent follow up study [80,347] and thus can be considered as robust associations. To try and understand the mechanism underlying the relationship between SHGB and LOY I aimed to investigate the impact of gene expression in these regions on the relationship between SHBG and LOY. Using the eQTLGene [373] database I identified SNPs that serve as valid proxies for the expression of 13 genes (*ACAT1, BCL2, DLK1, HM13, MAD1L1, QKI, RBPMS, SEMA4A, SENP7, SENP8, SETBP1, SMPD2, TCL1A,* and *TSC22D2*) from 13 of the 19 regions.

I found that the heterozygous state of 8/13 eQTLs and the homozygous state of 9/13 eQTLs were associated with LOY in the UK Biobank. The eQTL for *TCL1A* at 14q32.13 had the strongest association with LOY status (rs11849538_G/C, OR = 0.83, 95% CI = 0.80 - 0.85, P = 6.32x10^{-34}; rs11849538_G/G, OR = 0.63, 95% CI: 0.56 - 0.69, P = 1.48x10^{-16}), as shown in Figure 6-4 and Table 6-5. Only 2 of the 13 genes, however, were significantly associated with levels of SHBG as shown in Table 6-6. *MAD1L1* at 7p22.3 was positively associated with SHBG (rs10247428_A/A, β = 0.02, 95% CI = 0.01 - 0.03, P = 0.003), but this was in the opposite direction to its relationship with LOY. The *DLK1-MEG3* eQTL at 14q32.2 was negatively associated with SHBG (rs7141210-T/T, β = -0.02, 95% CI = -0.03 - 0.007, P = 0.02) and also negatively associated with LOY (Table 6-5). Of these two signals, only rs7141210-T/T modified the relationship between LOY and SHBG ($P_{interaction}$ = 0.04) (Figure 6-5, Table 6-7). There was no influence of rs7141210-T on TT, FT or BAT (Table 6-8) indicating that the interaction was specific for SHBG.



A)

Gene	SNP	Allele	Heterozygous OR	Heterozygous 95%CI	Heterozygous P_{FDR}	Homozygous OR	Homozygous 95%CI	Homozygous P_{FDR}
TCL1A	rs11849538	G	0.83	[0.8, 0.85]	$6.32x10^{-34}$	0.63	[0.56, 0.7]	$1.48x10^{-16}$
HM13	rs6060260	T	0.85	[0.83, 0.88]	$5.50x10^{-28}$	0.70	[0.64, 0.76]	$1.13x10^{-17}$
SENP7	rs7628681	T	1.13	[1.1, 1.16]	$8.38x10^{-19}$	1.24	[1.2, 1.29]	$1.71x10^{-30}$
SMPD2	rs7773095	C	0.90	[0.88, 0.93]	$1.01x10^{-13}$	0.77	[0.73, 0.8]	$1.71x10^{-30}$
DLK1	rs7141210	T	0.91	[0.89, 0.93]	$8.04x10^{-13}$	0.82	[0.78, 0.85]	$1.92x10^{-20}$
MAD1L1	rs10247428	A	0.94	[0.92, 0.97]	$9.64x10^{-6}$	0.87	[0.84, 0.9]	$1.39x10^{-13}$
SETBP1	rs11876015	C	0.95	[0.92, 0.98]	$4.21x10^{-3}$	0.81	[0.69, 0.95]	0.01
RBPMS	rs2978263	T	1.03	[1.01, 1.06]	0.02	1.01	[0.96, 1.07]	0.70
SEMA4A	rs6701295	C	1.03	[1, 1.06]	0.06	1.07	[1.04, 1.11]	$1.53x10^{-4}$
TSC22D2	rs1868673	A	0.99	[0.96, 1.01]	0.34	0.96	[0.93, 1]	0.08
ACAT1	rs12361905	C	0.99	[0.96, 1.02]	0.46	0.95	[0.92, 0.98]	$5.17x10^{-3}$
BCL2	rs4940576	T	0.99	[0.97, 1.02]	0.65	0.99	[0.94, 1.05]	0.74
QKI	rs1234977	C	1.01	[0.97, 1.04]	0.65	0.97	[0.84, 1.11]	0.70

B)

Gene	SNP	Allele	Heterozygous β	Heterozygous 95%CI	Heterozygous P_{FDR}	Homozygous β	Homozygous 95%CI	Homozygous P_{FDR}
TCL1A	rs11849538	G	-0.01	[-0.02, 0]	0.30	-0.03	[-0.07, 0]	0.20
HM13	rs6060260	T	0.00	[-0.01, 0.01]	0.96	0.01	[-0.02, 0.03]	0.78
SENP7	rs7628681	T	0.01	[0, 0.02]	0.51	0.01	[0, 0.02]	0.39
SMPD2	rs7773095	C	-0.01	[-0.02, 0.01]	0.78	-0.01	[-0.02, 0]	0.55
DLK1	rs7141210	T	-0.01	[-0.02, 0]	0.27	-0.02	[-0.04, -0.01]	0.02
MAD1L1	rs10247428	A	0.01	[0, 0.02]	0.06	0.02	[0.01, 0.04]	$2.82x10^{-3}$
SETBP1	rs11876015	C	0.02	[-0.03, 0.07]	0.78	0.01	[-0.01, 0.02]	0.58
RBPMS	rs2978263	T	0.00	[-0.01, 0.01]	0.96	0.01	[-0.01, 0.03]	0.78
SEMA4A	rs6701295	C	0.00	[-0.01, 0.01]	0.78	-0.01	[-0.02, 0]	0.27
TSC22D2	rs1868673	A	0.00	[-0.01, 0.02]	0.78	0.00	[-0.01, 0.01]	0.58
ACAT1	rs12361905	C	0.00	[-0.01, 0.02]	0.78	0.00	[-0.01, 0.01]	0.58
BCL2	rs4940576	T	0.00	[-0.02, 0.02]	0.78	0.00	[-0.01, 0.01]	0.58
QKI	rs1234977	C	-0.01	[-0.06, 0.04]	0.78	0.00	[-0.01, 0.01]	0.76

Figure 6-4: The relationship between the predicted expression of 13 genes and each of LOY and SHBG.

eQTL SNPs were used as proxies for gene expression and as assessed as predictors for LOY (panel A) and SHBG levels (panel B) incorporating age, smoking status, and 10 PCA as covariates.

Table 6-5: The relationship between 13 eQTLs and LOY

			Heterozygous				Homozygous			
Gene	**SNP**	**Allele assessed**	**OR**	**92.5% CI**	**97.5% CI**	***P***	**OR**	**92.5% CI**	**97.5% CI**	***P***
MAD1L1	**rs10247428**	**A**	0.94	0.92	0.97	9.64×10^{-6}	0.87	0.84	0.90	1.39×10^{-13}
DLK1	**rs7141210**	**T**	0.91	0.89	0.93	8.04×10^{-13}	0.82	0.78	0.85	1.92×10^{-20}
TCL1A	**rs11849538**	**G**	0.83	0.80	0.85	6.32×10^{-34}	0.63	0.56	0.70	1.48×10^{-16}
SEMA4A	**rs6701295**	**C**	1.03	1.00	1.06	0.06	1.07	1.04	1.11	1.53×10^{-4}
SENP7	**rs7628681**	**T**	1.13	1.10	1.16	8.38×10^{-19}	1.24	1.20	1.29	1.71×10^{-30}
SMPD2	**rs7773095**	**C**	0.90	0.88	0.93	1.01×10^{-13}	0.77	0.73	0.80	1.71×10^{-30}
SETBP1	**rs11876015**	**C**	0.95	0.92	0.98	0.00	0.81	0.69	0.95	0.01
TSC22D2	**rs1868673**	**A**	0.99	0.96	1.01	0.34	0.96	0.93	1.00	0.08
ACAT1	**rs12361905**	**C**	0.99	0.96	1.02	0.46	0.95	0.92	0.98	5.17×10^{-3}
BCL2	**rs4940576**	**T**	0.99	0.97	1.02	0.65	0.99	0.94	1.05	0.74
QKI	**rs1234977**	**C**	1.01	0.97	1.04	0.65	0.97	0.84	1.11	0.70
HM13	**rs6060260**	**T**	0.85	0.83	0.88	5.50×10^{-28}	0.70	0.64	0.76	1.13×10^{-17}
RBPMS	**rs2978263**	**T**	1.03	1.01	1.06	0.02	1.01	0.96	1.07	0.70

Table 6-6: The relationship between 13 eQTLs and SHBG

			Heterozygous				Homozygous			
Gene	**SNP**	**Allele assessed**	**β**	**92.5% CI**	**97.5% CI**	***P***	**β**	**92.5% CI**	**97.5% CI**	***P***
MAD1L1	**rs10247428**	**A**	0.01	0.00	0.02	0.06	0.02	0.01	0.04	2.823 x 10^{-3}
DLK1	**rs7141210**	**T**	-0.01	-0.02	0.00	0.27	-0.02	-0.04	-0.01	0.02
TCL1A	**rs11849538**	**G**	-0.01	-0.02	0.00	0.30	-0.03	-0.07	0.00	0.20
SEMA4A	**rs6701295**	**C**	0.00	-0.01	0.01	0.78	-0.01	-0.02	0.00	0.27
SENP7	**rs7628681**	**T**	0.01	0.00	0.02	0.51	0.01	0.00	0.02	0.39
SMPD2	**rs7773095**	**C**	-0.01	-0.02	0.01	0.78	-0.01	-0.02	0.00	0.55
SETBP1	**rs11876015**	**C**	0.02	-0.03	0.07	0.78	0.01	-0.01	0.02	0.58
TSC22D2	**rs1868673**	**A**	0.00	-0.01	0.02	0.78	0.00	-0.01	0.01	0.58
ACAT1	**rs12361905**	**C**	0.00	-0.01	0.02	0.78	0.00	-0.01	0.01	0.58
BCL2	**rs4940576**	**T**	0.00	-0.02	0.02	0.78	0.00	-0.01	0.01	0.58
QKI	**rs1234977**	**C**	-0.01	-0.06	0.04	0.78	0.00	-0.01	0.01	0.76
HM13	**rs6060260**	**T**	0.00	-0.01	0.01	0.96	0.01	-0.02	0.03	0.78
RBPMS	**rs2978263**	**T**	0.00	-0.01	0.01	0.96	0.01	-0.01	0.03	0.78

Table 6-7: The effect of genes expression on the relationship between LOY and SHBG

			Heterozygous				Homozygous			
Gene	**SNP**	**Allele assessed**	**β**	**92.5% CI**	**97.5% CI**	***P***	**β**	**92.5% CI**	**97.5% CI**	***P***
MAD1L1	**rs10247428**	**A**	0.018	-0.005	0.041	0.257	-0.008	-0.040	0.024	0.624
DLK1	**rs7141210**	**T**	0.008	-0.015	0.030	0.507	0.044	0.007	0.080	0.040

Table 6-8: The effect of rs7141210-T on the relationship between LOY and other sex hormones

	Heterozygous				Homozygous			
Gene	**β**	**92.5% CI**	**97.5% CI**	***P***	**β**	**92.5% CI**	**97.5% CI**	***P***
Total testosterone	2.73×10^{-3}	-0.02	0.03	0.82	0.02	-0.02	0.06	0.30
Free testosterone	-2.36×10^{-3}	-0.02	0.02	0.84	-0.02	-0.05	0.02	0.41
Bioavailable testosterone	-3.10×10^{-4}	-0.02	0.02	0.98	-0.01	-0.05	0.02	0.54

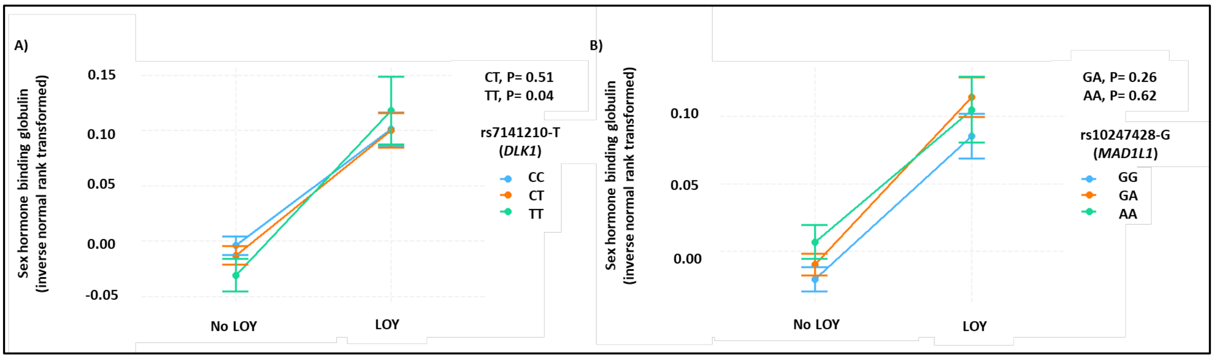


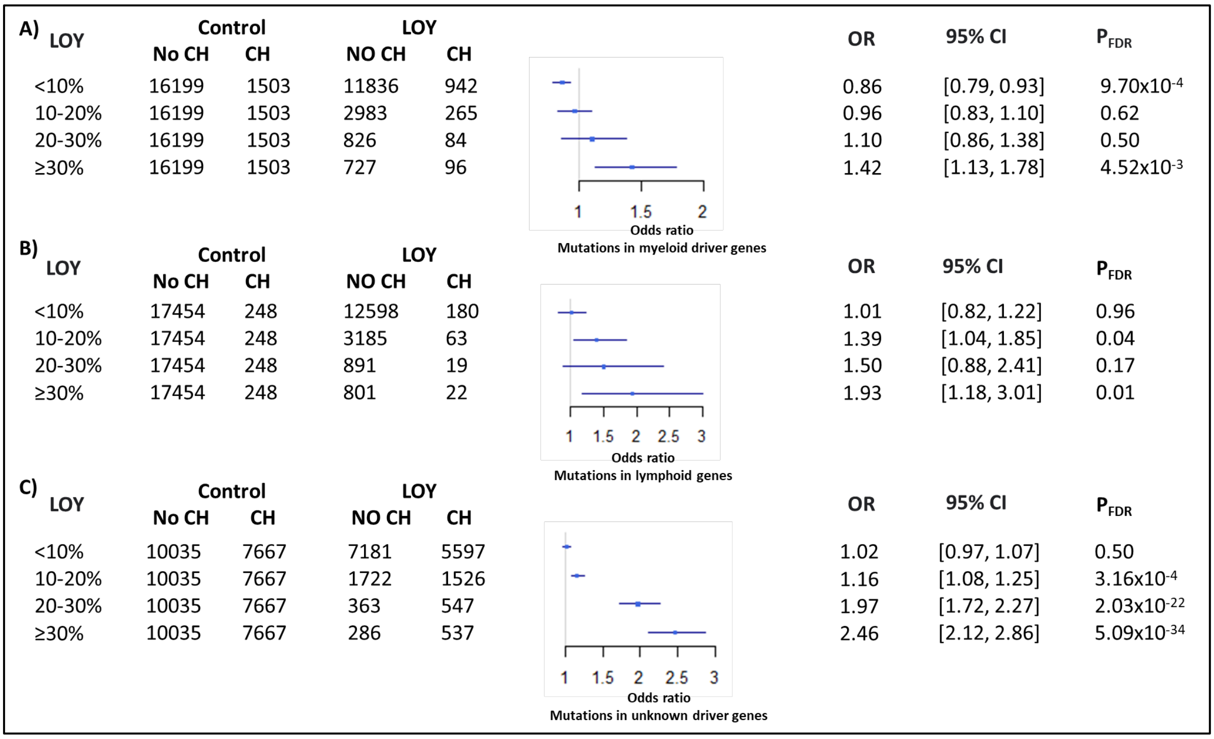
Figure 6-5: The interaction between genetically predicted expression of DLK1-MEG3 and MAD1L1, according to LOY status and SHBG levels

Summary of SHBG values for men with and without LOY for (A) rs7141210 as a proxy for *DLK1-MEG3* expression and (B) rs10247428 as a proxy for *MAD1L1* expression. The interaction effect of each eQTL and LOY on SHBG was assessed by linear regression and was significant for rs7141210-T/T (green; left panel) in comparison to (C/C, blue). No significant difference was seen between rs10247428 genotypes.

6.4.4 The relationship between LOY and CH

To assess the impact of somatic driver mutations and other markers of clonality on the relationship between SHBG, testosterone and LOY, I first assessed the relationship between somatic variants (driver and non-driver) and LOY. WES data was available for 17,759 participants with LOY, of whom 28% (n=4,981) were estimated to have an LOY clone size ≥ 10% of leucocytes. For comparison, I randomly selected the UK Biobank age-matched controls (n=17,702) that were negative for LOY. I identified recurrent somatic mutations in driver genes associated with myeloid CH and lymphoid CH, plus likely somatic variants in other genes that indicated clonality in the absence of known driver mutations, which I refer as unknown driver CH. Overall, the frequency of each CH subtype (myeloid, lymphoid, unknown) was similar between cases with LOY and controls. Striking differences emerged, however, when LOY was stratified by clone size. All CH (myeloid plus lymphoid plus unknown driver) was significantly associated with LOY in ≥ 10% of cells with a clear increase in the strength of the association with increasing LOY clone size (10-20% LOY, OR = 1.17, P = 1.81x10^{-4}; 20-30% LOY, OR = 2.20, P = 4.25x10^{-27}; ≥30% LOY, OR = 3.43, P = 2.42x10^{-52}). Both myeloid CH (OR = 1.42, P = 4.52x10^{-3}), and lymphoid CH (OR = 1.93, P = 0.01) were significantly associated with LOY in ≥30% of cells but not LOY of smaller clone size. The relationship was most prominent, however, for unknown driver CH (OR

= 2.46, P = $5.09x10^{-34}$). Full results are presented in Figure 6-6 and Table 6-9. None of the three CH subtypes were associated with SHBG or the three measures of testosterone, indicating no effect of driver mutations on the relationship between LOY and sex hormones (Table 6-10).



A)

LOY	Control		LOY		OR	95% CI	P_{FDR}
	No CH	CH	NO CH	CH			
<10%	16199	1503	11836	942	0.86	[0.79, 0.93]	$9.70x10^{-4}$
10-20%	16199	1503	2983	265	0.96	[0.83, 1.10]	0.62
20-30%	16199	1503	826	84	1.10	[0.86, 1.38]	0.50
≥30%	16199	1503	727	96	1.42	[1.13, 1.78]	$4.52x10^{-3}$

B)

LOY	Control		LOY		OR	95% CI	P_{FDR}
	No CH	CH	NO CH	CH			
<10%	17454	248	12598	180	1.01	[0.82, 1.22]	0.96
10-20%	17454	248	3185	63	1.39	[1.04, 1.85]	0.04
20-30%	17454	248	891	19	1.50	[0.88, 2.41]	0.17
≥30%	17454	248	801	22	1.93	[1.18, 3.01]	0.01

C)

LOY	Control		LOY		OR	95% CI	P_{FDR}
	No CH	CH	NO CH	CH			
<10%	10035	7667	7181	5597	1.02	[0.97, 1.07]	0.50
10-20%	10035	7667	1722	1526	1.16	[1.08, 1.25]	$3.16x10^{-4}$
20-30%	10035	7667	363	547	1.97	[1.72, 2.27]	$2.03x10^{-22}$
≥30%	10035	7667	286	537	2.46	[2.12, 2.86]	$5.09x10^{-34}$

Figure 6-6: The relationship between LOY and CH

LOY was stratified according to the clonal size (<10%, 10%-20%, 20%-30%, and ≥30%) and the proportion of participants with CH within each group was compared with controls. (A) myeloid CH, (B) lymphoid CH, (C) unknown driver CH.

Table 6-9: The relationship between LOY and driver mutations

		Control		LOY		OR	95% confidence interval		
LOY clonal size	**Driver mutations (CH)**	**No CH**	**CH**	**No CH**	**CH**				P_{FDR}
All mutations (myeloid + lymphoid + unknown)	**<10%**	8284	9418	6059	6719	0.98	0.93	1.02	0.38
All mutations (myeloid + lymphoid + unknown)	**10-20%**	8284	9418	1394	1854	1.17	1.08	1.26	1.81 x 10^{-4}
All mutations (myeloid + lymphoid + unknown)	**20-30%**	8284	9418	260	650	2.20	1.90	2.56	4.25 x 10^{-27}
All mutations (myeloid + lymphoid + unknown)	**≥30%**	8284	9418	168	655	3.43	2.88	4.10	2.42 x 10^{-52}
Known driver genes (myeloid + lymphoid)	**<10%**	15951	1751	11656	1122	0.88	0.81	0.95	2.32 x 10^{-3}
Known driver genes (myeloid + lymphoid)	**10-20%**	15951	1751	2920	328	1.02	0.90	1.16	0.76
Known driver genes (myeloid + lymphoid)	**20-30%**	15951	1751	807	103	1.16	0.93	1.44	0.25
Known driver genes (myeloid + lymphoid)	**≥30%**	15951	1751	705	118	1.52	1.24	1.87	2.92 x 10^{-4}
Known driver genes (myeloid)	**<10%**	16199	1503	11836	942	0.86	0.79	0.93	9.70 x 10^{-4}
Known driver genes (myeloid)	**10-20%**	16199	1503	2983	265	0.96	0.83	1.10	0.62
Known driver genes (myeloid)	**20-30%**	16199	1503	826	84	1.10	0.86	1.38	0.50
Known driver genes (myeloid)	**≥30%**	16199	1503	727	96	1.42	1.13	1.78	4.52 x 10^{-3}
Known driver genes (lymphoid)	**<10%**	17454	248	12598	180	1.01	0.82	1.22	0.96
Known driver genes (lymphoid)	**10-20%**	17454	248	3185	63	1.39	1.04	1.85	0.04

Known driver genes (lymphoid)	**20-30%**	17454	248	891	19	1.50	0.88	2.41	0.17
Known driver genes (lymphoid)	**≥30%**	17454	248	801	22	1.93	1.18	3.01	0.01
unknown driver genes	**<10%**	10035	7667	7181	5597	1.02	0.97	1.07	0.50
unknown driver genes	**10-20%**	10035	7667	1722	1526	1.16	1.08	1.25	3.16×10^{-4}
unknown driver genes	**20-30%**	10035	7667	363	547	1.97	1.72	2.27	2.03×10^{-22}
unknown driver genes	**≥30%**	10035	7667	286	537	2.46	2.12	2.86	5.09×10^{-34}

Table 6-10: The relationship between CH and sex hormones

		Coefficient	95% confidence interval		*P*
All mutations (myeloid + lymphoid + unknown)	Sex hormone binding protein	-0.01	-0.03	0.01	0.48
All mutations (myeloid + lymphoid + unknown)	Total testosterone	-0.01	-0.03	0.02	0.63
All mutations (myeloid + lymphoid + unknown)	Free testosterone	0.00	-0.02	0.02	0.87
All mutations (myeloid + lymphoid + unknown)	Bioavailable testosterone	0.01	-0.02	0.03	0.60
Known driver genes (myeloid + lymphoid)	Sex hormone binding protein	0.01	-0.02	0.05	0.54
Known driver genes (myeloid + lymphoid)	Total testosterone	0.01	-0.03	0.05	0.70
Known driver genes (myeloid + lymphoid)	Free testosterone	0.00	-0.03	0.04	0.94
Known driver genes (myeloid + lymphoid)	Bioavailable testosterone	0.01	-0.03	0.04	0.75
Known driver genes (myeloid)	Sex hormone binding protein	0.01	-0.03	0.05	0.58
Known driver genes (myeloid)	Total testosterone	0.00	-0.04	0.04	0.85
Known driver genes (myeloid)	Free testosterone	-0.01	-0.05	0.03	0.53
Known driver genes (myeloid)	Bioavailable testosterone	-0.01	-0.05	0.03	0.67
Known driver genes (lymphoid)	Sex hormone binding protein	0.01	-0.08	0.10	0.80
Known driver genes (lymphoid)	Total testosterone	0.07	-0.02	0.16	0.13
Known driver genes (lymphoid)	Free testosterone	0.08	-0.01	0.16	0.08
Known driver genes (lymphoid)	Bioavailable testosterone	0.08	0.00	0.17	0.05
unknown driver genes	Sex hormone binding protein	-0.01	-0.03	0.01	0.35
unknown driver genes	Total testosterone	-0.01	-0.03	0.02	0.55
unknown driver genes	Free testosterone	0.00	-0.02	0.02	0.85
unknown driver genes	Bioavailable testosterone	0.01	-0.02	0.03	0.63

To understand the relationship between CH and LOY in more detail, I assessed the association between somatic mutations in specific driver genes in participants with LOY in >30% cells (n=823) compared to LOY free controls (n=17,702). *TET2* was the most significantly enriched mutated gene in LOY cases (4% versus 1.5% in controls, OR = 2.64, P = 9.58x10^{-5}) with *TP53* (OR = 6.96, P = 7.62x10^{-3}) and *CBL* (OR = 7.43, P = 0.04) mutations also showing a significant enrichment (Table 6-11). Other genes, including *DNMT3A* and *ASXL1*, showed no enrichment in high level LOY cases.

Table 6-11: The relationship between LOY with clone size >30% and driver genes

		Control		LOY		OR	95% confidence interval		
LOY clonal size	**Driver mutations (CH)**	**No CH**	**CH**	**No CH**	**CH**				**P_{FDR}**
> 30%	*TET2*	16199	253	727	30	2.64	1.73	3.90	9.58 x 10^{-5}
> 30%	*TP53*	16199	16	727	5	6.96	1.99	19.95	7.62 x 10^{-3}
> 30%	*CBL*	16199	9	727	3	7.43	1.29	29.84	0.04
> 30%	*NF1*	16199	15	727	3	4.46	0.83	15.80	0.09
> 30%	*DNMT3A*	16199	637	727	18	0.63	0.37	1.01	0.11
> 30%	*SF3B1*	16199	36	727	4	2.48	0.64	6.93	0.14
> 30%	*STAG2*	16199	113	727	8	1.58	0.66	3.23	0.33
> 30%	*SRSF2*	16199	40	727	3	1.67	0.33	5.27	0.49
> 30%	*ASXL1*	16199	167	727	7	0.93	0.37	1.98	1.00

The possibility that LOY and CH as defined by somatic mutations might co-exist in the same clone was assessed by analysing the relationship between LOY BAF and the VAFs of driver mutations. Figure 6-7 shows a summary of the results at different ranges of LOY. Myeloid CH VAFs predicted BAF levels in samples with LOY >10% (β = 0.10, 95% CI = 0.05 - 0.15, P = 1.12x10^{-4}). Similar results were seen for lymphoid CH (β = 0.20, 95% CI = 0.06 - 0.35, P = 0.02) and also unknown driver CH (β = 0.19, 95% CI = 0.14 - 0.24, P = 3.72x10^{-12}), which by definition was restricted to VAFs of 0.1 - 0.2 (Table 6-12).

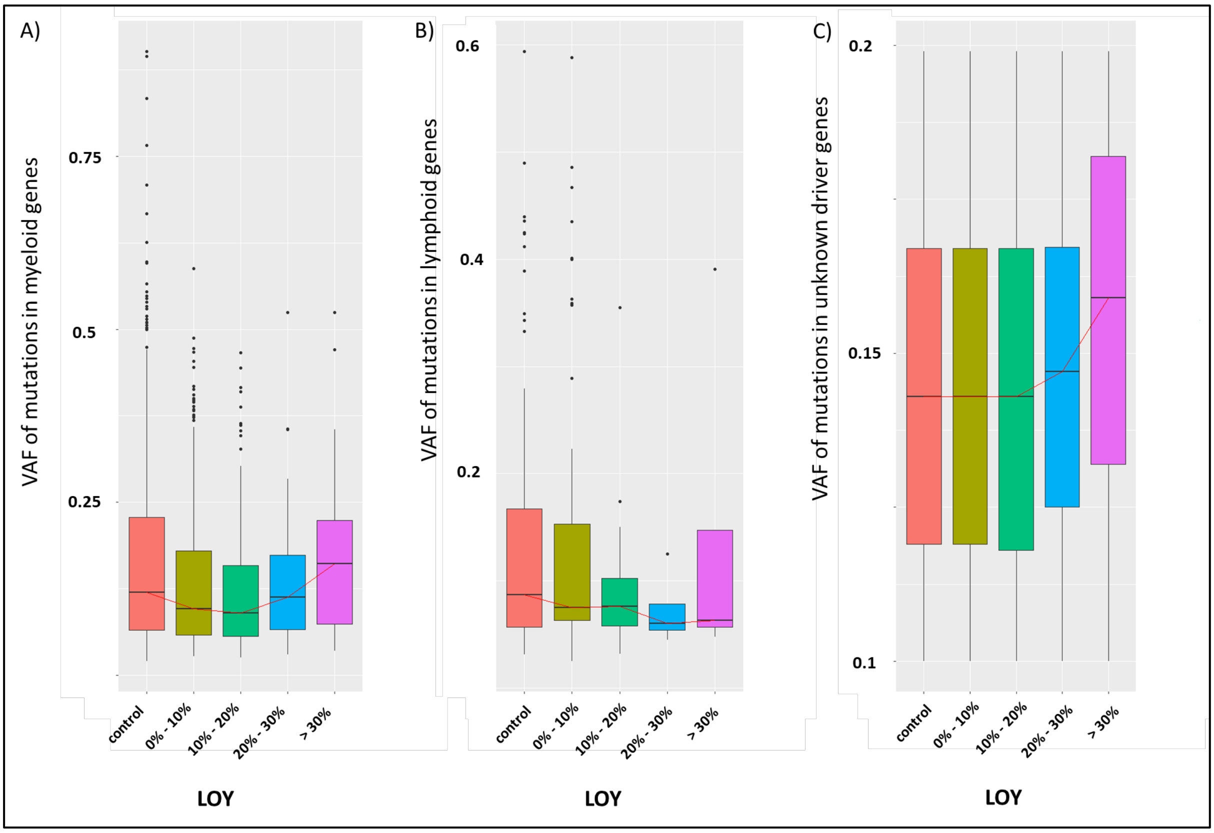


Figure 6-7: The relationship between LOY clonal size and CH VAFs

Boxplots summarizing the distribution of VAFs of somatic mutations in controls and cases with LOY broken down by clone size. (A) myeloid CH, (B) lymphoid CH, (C) unknown driver CH. The red lines connect median values.

Table 6-12: The relationship between driver mutations and LOY

	LOY clonal size	Beta coefficient	95% confidence interval		*P*
Myeloid genes	**< 10%**	0.00	0.00	0.00	0.89
Myeloid genes	**> 10%**	0.10	0.05	0.15	1.12 x10^{-4}
Lymphoid genes	**< 10%**	0.00	-0.01	0.01	0.89
Lymphoid genes	**> 10%**	0.20	0.06	0.35	0.02
Unknown driver gene	**< 10%**	0.00	-0.01	0.00	0.06
Unknown driver gene	**> 10%**	0.19	0.14	0.24	3.72 x10^{-12}

6.5 Discussion

Age-related mosaic LOY in peripheral blood leukocytes is known to be influenced by both genetic and environmental factors. I have found that LOY is also strongly associated with levels of SHBG in serum, and that this association is independent of known confounders (age, smoking history and the first 10 principal genetic components). Furthermore, using Mendelian randomisation I found that SHBG levels are likely causally linked to LOY, but LOY has no effect on SHBG. SHBG regulates the level of circulating testosterone and, although I found that both FT and BAT were lower in men with LOY compared to those without LOY, there was no significant relationship between LOY and either FT or BAT on multivariate analysis. This finding is inconsistent with the free hormone hypothesis, which proposes that only the unbound fraction of testosterone is biologically active in target tissues [377], and instead suggests that involvement of other pathways such as binding and internalization of SHBG-bound testosterone by specific cell types, including B-cells [378,379]. The mechanism by which SHBG promotes LOY is unclear, but from a genetic perspective the effect is not explained by variation at *SHBG* alone. Other loci are involved, particularly *JMJD1C* which encodes a histone demethylase previously linked to SHBG levels [372].

To understand the influence of genetic factors on the relationship between SHBG and LOY in more detail, I focused on genetically predicted expression of genes linked to the development of LOY. I found that rs7141210-T, a marker for expression of genes in the *DLK1–MEG3* region at 14q32 [373], was associated with elevated LOY and SHBG, and that homozygosity for the T-allele of rs7141210 modified the relationship between LOY and SHBG. *DLK1-MEG3* is a large and complex imprinted cluster of genes and non-coding RNAs. The methylated paternally derived chromosome expresses the protein-coding genes *DLK1*, *RTL1* and *DIO3*, while the non-methylated maternally derived chromosome expresses the non-coding genes *MEG3*, *MEG8*, *asRTL1*, multiple miRNAs and lncRNAs [380]. Constitutional uniparental disomy (UPD) at 14q32 is associated with the developmental disorders Temple syndrome (maternal UPD) and Kagami–Ogata syndrome (paternal UPD) whereas somatically acquired paternal UPD is associated with CH and myeloid malignancies.[267] Genome wide significant signals have been identified near *DLK1* in association with CH defined by acquired 14q UPD [362] and somatic driver mutations [289] as well as LOY [80,81,347]. Furthermore, a parent of origin specific effect of rs1555405-A linked to differential methylation has been defined in relation to platelet counts [381]. This SNP is in linkage disequilibrium with rs7141210 (D'=1, R^2=0.7), with rs7141210-T allele being correlated with rs1555405-A allele. Collectively these findings suggest a potential parent of origin impact of rs7141210-T on the relationship between SHBG and LOY.

I have defined the relationship between CH and LOY in the UK Biobank. LOY in ≥30% of cells was associated with both myeloid and lymphoid CH, with 14% of affected individuals having one or more somatic driver mutation compared to 10% of controls ($P = 2.92 \times 10^{-4}$; Table 6-9). At the level of individual genes, the most striking finding was that mutated *TET2* was associated with LOY, but not *DNMT3A* or *ASXL1*. Collectively, mutations in one of these genes accounts for 90% of cases of CH defined by sequence analysis, and my findings are consistent with the notion that CH with *TET2* mutations is different from CH with *DNMT3A* or *ASXL1* mutations. With larger studies, specific disease associations are emerging, for example CH with *TET2* mutations has been linked to chronic obstructive pulmonary diseases [94], but not CH with *DNMT3A* mutations.

Most strikingly, however, unknown driver CH was seen in 65% of cases with high level (≥30% of cells) LOY compared to just 43% of controls ($P = 5.09 \times 10^{-34}$). Overall, 80% of cases with high level LOY had mutational evidence of clonality, with LOY in ≥10% of cells clearly associated with unknown driver CH. For the first time, therefore, my findings provide broad molecular confirmation that LOY ≥10% is indeed clonal, and I predict that comprehensive sequencing by WGS will confirm clonality in most cases. Importantly for my study, neither overall CH nor any CH subtype was associated with SHBG or measures of testosterone (Table 6-10). The driver of clonality in cases with LOY and unknown driver CH remains unclear but I found that the degree of LOY was strongly predicted by the VAF of the somatic variants used to define CH ($P = 3.72 \times 10^{-12}$). This suggests that LOY might itself be a driver of clonality, as has been postulated recently from whole genome sequence data of single cell-derived haematopoietic cells colonies [382], and that LOY therefore accounts for an appreciable proportion of unknown driver CH.

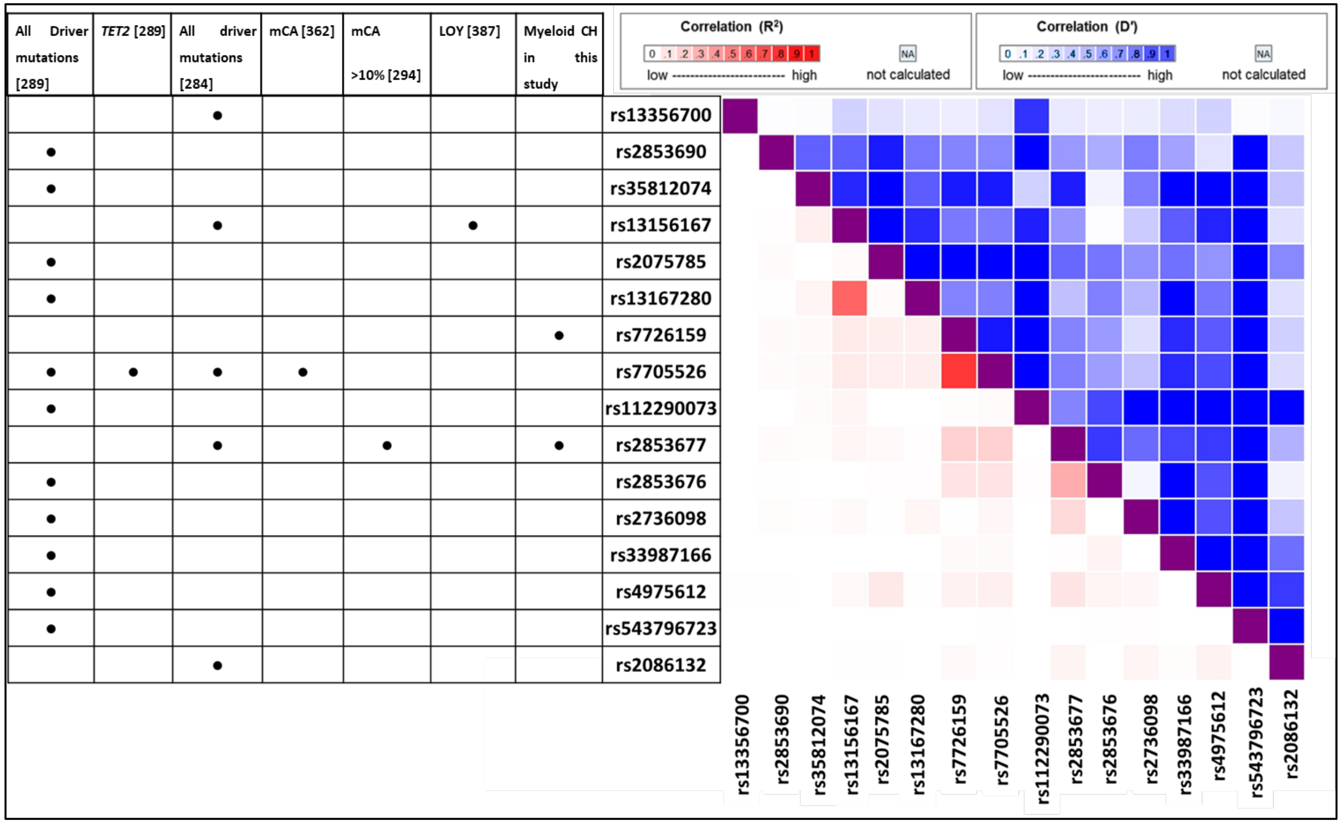
Chapter 7 Conclusion and future plans

This thesis describes the prevalence of CH in healthy volunteers from the UK Biobank, aged between 40 and 69 years at recruitment, and dissects the relationship between CH and risk factors including smoking status as an environmental exposure, genetic predisposition, blood counts and serum biomarkers (Chapter 3). In addition, my study extends the scientific knowledge about the impact of CH on chronic inflammatory diseases focusing on adverse outcomes associated with CKD (Chapter 4), and the utility of machine learning survival models to predict the risk of myeloid neoplasms from highly dimensional data (Chapter 5). Furthermore, my study elucidates the impact of endogenous sex hormones, somatic driver mutations, and gene expression on the risk of developing LOY in men (Chapter 6).

Age is a well-established risk factor for CH having been reported across many studies and using different genetic lesions to define CH, specifically SNVs or indels in genes associated with haematological malignancies, mCAs and LOY [6,7,10,347]. I detected mCAs, either CNG, CNL, or UPD, of relatively large size (≥ 2Mb) and clone size (> 10%) in 1% of the UK Biobank cohort and showed that their incidence increased significantly with age from 0.85% at age 40 - 45 to 1.29% at age 65 - 70 years (Chapter 3). The relationship between age and the risk of acquiring an mCA was strongest for specific lesions associated with myeloid disorders with an estimated annual risk of 1.1 fold, which is similar to that observed for driver mutations in myeloid genes. Recent models of stem cell dynamics that incorporated age, mutation rate, and population size of HSCs, detected an independent fitness advantage for each mutation [83]. These results were mirrored in a recent longitudinal analysis of driver mutations and their VAFs which showed that the majority of clones expand at a constant exponential rate over time but the growth rate varies according to the mutated gene [383]. The greatest annual risk of acquiring an mCA associated with myeloid malignancies could be explained by a faster growth rate as suggested by a recent study which found a positive correlation between dN/dS (the rate of substitutions at non-silent site/ the rate of substitutions at silent site) calculated coefficient for individual position and its correlation strength to myeloid malignancy [383]. However, these findings cannot explain the poor relationship between *GNB1 K57E* and myeloid malignancies. *GNB1* K57E is a highly fit mutation identified in healthy individuals [83]. In my analysis of 200,631 exomes from the UK Biobank, *GNB1* K57E (n=90) was the third most frequent mutation after *DNMT3A* R882 and *JAK2* V617F but only two individuals with *GNB1* K57E developed a myeloid neoplasm after recruitment. Furthermore, *GNB1* K57E was found in only 10 samples with haematological and lymphoid origin in the COSMIC database (version 91) in comparison to 41,923 samples for *JAK2* V617F, and 886 samples for *DNMT3A* R882H. These data suggest that growth rate alone is not enough to determine the malignant potential of a mutation.

Environmental exposures are a mixture of chemical and physical substances in air, water, food, or soil that may have a harmful effect on a person’s health. Smoking has been connected to higher risk of somatic driver mutations, and LOY [8,77]. I confirmed the observed relationship between smoking and CH and extended these previous findings to show that the relationship is most prominent for *ASXL1* mutations due to current rather than previous smoking. This potentially indicates a specific mechanism for the broad health benefit of quitting smoking (Chapter 3). The association between *ASXL1* and smoking was confirmed in an independent validation cohort from the UK Biobank, and in a published study of post-therapy cancer patients [85]. Although the observational analysis indicated an association between CH defined by driver mutations and smoking history, the results from two MR studies were heterogenous. The genetically defined ‘lifetime smoking index’ in the UK Biobank was not associated with CH defined by driver mutations detected in the TopMed cohort [384]. However, a meta-analysis of genetically defined smoking involving 1.2 million individuals [385], detected a significant association with CH defined by driver mutations in the UK Biobank with concordant results for *DNMT3A* and *TET2* [284]. In addition, exposure of *TET2* -/- transplanted mouse model to cigarette smoke or e-cigarette aerosol promoted a clonal expansion over time [386]. The investigation of CH defined by driver mutations in 628,388 individuals from the UK Biobank and the MyCode Community Health Initiative cohort indicated a significant association between CH and lung cancer that was independent of smoking [289]. A recent WGS study of single colonies from bronchial epithelial cells indicated a high impact of smoking on mutational burden and the number of driver mutations, but the study also found significant inter-individual and intra-individual heterogeneity that reached as much as 10 fold between individual cells [384]. Single cell analysis is required to assess if similar heterogeneity exists in HSCs among smokers.

CH has a relatively small heritability of 3.6%[169] in comparison to heritability of LOY which has been estimated at 34%[387], but common variants in *TERT* were linked to all types of CH as well as LOY (Figure 7-1). The *TERT* gene encodes the catalytic subunit of telomerase, an essential enzyme for the *de novo* synthesis of telomeres. *TERT* expression is usually low in normal somatic cells but often elevated in cancer. Different mechanisms have been associated with its activation including non-coding mutations in the promoter of *TERT*. On analysis of the first UK Biobank release of WES data, I identified two independent signals within *TERT* that achieved genome-wide significance for association with myeloid CH: rs7726159 and rs2853677 (Chapter 3). In addition, linkage disequilibrium data and GWAS data of myeloid disorders in the UK Biobank indicated that variation in intron 2 (rs7726159) is associated with CH but does not predict development of MPN whereas variation in intron 3 (rs2853677) does predict development of MPN (http://big.stats.ox.ac.uk). My analysis confirmed that inherited variation in *TERT* is associated with CH but with different levels of phenotypic risk. A recent study of CH defined by driver mutations in a larger cohort of the UK Biobank participants (n=200,453) identified three independent signals near *TERT* (rs2853677, rs13156167 and rs2086132,) in the main analysis, and two additional signals on conditional analysis (rs7705526, rs13356700) [284]. rs7726159 is in LD with rs7705526 (r^2=0.79) and thus findings of [284] confirm and extend my observations (Figure 7-1). Significantly associated variants near *TERT* were concordant for both *DNMT3A* and *TET2* mutated CH, but variations near *TCL1A* gene were not concordant, as rs2887399-T increase the risk of *DNMT3A*-mutated CH, but decrease the risk of *TET2*-mutated CH [169,363].



All Driver mutations [289]	*TET2* [289]	All driver mutations [284]	mCA [362]	mCA >10% [294]	LOY [387]	Myeloid CH in this study	
		●					rs13356700
●							rs2853690
●							rs35812074
		●			●		rs13156167
●							rs2075785
●							rs13167280
						●	rs7726159
●	●	●	●				rs7705526
●							rs112290073
		●		●		●	rs2853677
●							rs2853676
●							rs2736098
●							rs33987166
●							rs4975612
●							rs543796723
		●					rs2086132

Figure 7-1: Heatmap matrix of LD statistics in *TERT* common variants associated with detectable CH

The squared correlation coefficient between pairs of loci (r^2) is shown in shades of red. The coefficient of linkage disequilibrium (D') is shown in shades of blue. The heat map of D' and r^2 includes genome wide significant signals in the region of *TERT* from four different studies, shown in the left table, in addition to my findings.. In this thesis, I identified two independent signals within *TERT* for association with myeloid CH, which have both been confirmed by other studies: (i) rs7726159 is in LD with rs7705526 (r^2=0.79), which has previously been associated with all driver mutations, *TET2*, and mCA. (ii) rs2853677 was confirmed for association with all driver mutations and extended mCA.

The relationship between CH and the elevated risk of all-cause mortality, haematological malignancies, and CVD provides a link between aging and low-grade inflammation. I confirmed the relationship between myeloid CH and both RDW and all-cause mortality after adjusting for age, sex, and smoking status. Next, I reported an association between CH and eGFR estimated from cystatin-C, and moderate CKD defined by eGFR between 15-59 mL/min/1.73 m^2, and that the relationship was due to myeloid rather than lymphoid CH. My findings support the previous argument that eGFR.cys is more informative than eGFR.creat to define CKD [319]. Myeloid CH increases the risk of adverse outcomes in CKD and this increase is only partly explained by incident myeloid neoplasms. My findings were confirmed in a new study, which showed that the 2 and 5 years probability of ESKD calculated by kidney failure risk equations [388] was much higher in individuals with CKD and CH defined by somatic driver mutations in comparison to CKD without CH.[389] Murine models have characterised the relationship between CH and atherosclerosis development mediated by the elevation of the inflammatory markers, such as IL-6, and IL-1β. [89,390] but future studies are essential to investigate the impact of CH on the histopathological features of CKD such as glomerular sclerosis and interstitial fibrosis.

The low-grade inflammation associated with CH was marked by higher RDW, IL-6 levels, and CRP levels. On the gene level, the relationship between myeloid CH and CKD was significant for CH defined by putative somatic mutations in *CBL, TET2, JAK2, PPM1D* and *GNB1* but not *DNMT3A* or *ASXL1*. In the CANTOS clinical trial, individuals with *TET2* mutations had a better response to Canakinumab, a human monoclonal antibody that targets IL-1β, and reduces the risk of major events (non-fatal myocardial infarction, non-fatal stroke, or cardiovascular death), in patients with myocardial infarction and increased inflammation as indicated by an CRP > 2 mg/L [391]. My results suggest that the possibility of using myeloid CH to stratify patients with high risk of CVD and high CRP may extend to CKD patients. The impact of CH on the response to therapeutic monoclonal antibodies, however, is unknown. Ziltivekimab, an IL-6 inhibitor, decreases biomarkers of inflammation and thrombosis such as CRP, and fibrinogen in CKD patients and reduces the risk of adverse outcomes [392]. The incorporation of CH assessment in future clinical trials should be considered to assess the impact of CH on the response to anti-inflammatory medications in CKD patients.

LOY is the most common genetic lesion in men. I observed that LOY is strongly associated with levels of SHBG in serum, and MR analysis suggested a causal effect of SHBG on LOY, but LOY has no effect on SHBG. Furthermore, the calculated levels of BAT and FT were not associated with LOY on multivariate analysis, thus arguing against the long-standing hypothesis of the free-hormone pathway.

SHBG has cell-specific mechanisms for binding and internalisation that varies between T-cells and B-cells [379]. My findings implicate both *JMJD1C* and *DLK1–MEG3* as being involved in the link between SHBG and LOY but further investigations are required to understand the role of these genes.

The link with *DLK1-MEG3* is particularly interesting. Utilising eQTL data from blood samples, I found that rs7141210-T, a marker for expression of genes in the *DLK1–MEG3* imprinted region at 14q32.2 [373], was associated with higher risk of LOY and elevated SHBG serum levels, and that homozygosity for the T-allele of rs7141210 modified the relationship between LOY and SHBG. Although the mechanism is not obvious, this finding suggests that *DLK1-MEG3* might mediate or at least influence the effect of SHBG on LOY. Previous parent of origin studies have identified a specific relationship between paternal rs7141210-T allele and age at menarche in females [393], and the maternal rs7141210-T allele and platelet counts [381]. The *DLK1-MEG3* locus falls in the minimal affected region of acquired 14q UPD which is associated with both MPN and CH, and has a methylation pattern indicative of loss of maternal 14q and gain of paternal chromosome 14q [267]. In addition, genome-wide significant signals near the *DLK1-MEG3* locus have been associated with different forms of CH defined by 14q UPD (rs7141110), driver mutations (rs72698720), and LOY (rs72698720) [80,289,362]. It will be interesting in future studies to assess the role of the parent of origin of rs7141210 in relation to the link with SHBG, and potentially this might be achieved by utilising inferred parent of origin information of individual alleles generated by modelling identity by decent sharing with second and third relatives in the UK Biobank [394].

I assessed the relationship between LOY and CH and identified a significant association between CH and LOY ≥ 10% of cells. At the driver gene level, the relationship was significant between LOY in ≥ 30% cells and mutations targeting myeloid or lymphoid genes, with the effect being most prominent for mutations in *TET2*, *TP53* and *CBL* but not with *DNMT3A* and *ASXL1*. The relationship was even more prominent between LOY and unknown driver CH, an observation that confirms LOY is clonal. In addition, analysis of VAFs indicated a co-occurrence between LOY and CH both in the absence and presence of known driver mutations, an observation that is consistent with the hypothesis that LOY may be a direct driver of clonality. This result is concordant with single cell analysis which showed that many expanded clades within have LOY in the absence of other driver mutations of clonality [382]. A possible mechanism might be loss of *UTY* (also known as *KDM6C*) located at Yq11.221 since this gene and its paralogue *UTX* on the X-chromosome both demonstrated tumour suppressive properties in a mouse model of AML [395].

Sex biases have been identified among markers of CH. As a whole, mCA are more frequent in males [7] but specific mCA are known to have unusual sex biases [253,362] for example chromosome 15 CNG is more frequent in men, while 16p11.2 CNL and 10q CNL are more prevalent in females. Loss of the X chromosome has been reported in 5% of females but the widespread finding of LOY in comparison to other sex biases is consistent with the hypothesis that LOY may be subject to positive selection. Overall, therefore, these findings indicate that LOY, at least of clone size >10%, should be considered as a form of CH and that detailed analysis of outcomes and clinical phenotypes should evaluate all forms of CH, including LOY, and any overlap between them.

Although CH has been shown to confer a highly elevated HR for development of a myeloid neoplasm, the actual probability that a myeloid neoplasm will develop is relatively low and translates to a rate of roughly 1-2% of CH cases per year [293]. It would be highly desirable therefore to identify other factors that might help to identify individuals at elevated risk of progression. CH is a measurable event based on the targeted genes, VAF and individual age. All of these factors have been used to predict the risk of myeloid malignancies in healthy individuals. Previous studies have used univariate methods such as a Kaplan-Meier estimate and log-rank test [43], multivariable models such as regularised COX-PH [42], and Fine–Gray regression [86] to deal with time to event data. However, these methods are not suited for modelling high dimensional data. The expansion of ML models to handle censored data has enabled the prediction of health outcomes with a capability to model highly dimensional data. I used ML survival models to predict the risk of myeloid neoplasms in healthy individuals by utilising CH data, blood counts, and serum biomarkers as predictive features. The RSF model was the best predictor in test data and attributed the largest weights to platelet counts, platelet crit, and number of lesions in myeloid genes. However, the RSF model showed the importance of combining data from CH calls, blood counts, and serum biomarkers, and that no single feature was weighted more than 1% in the model. Interestingly, three serum biomarkers (cystatin-C, glucose and phosphate) were among the top 10 features in the generated RSF model. These markers may indicate the importance of evaluating kidney function and glucose levels as early signs to predict the risk of myeloid malignancies. This model is concordant with my findings described in Chapter 4 that myeloid CH was significantly associated with moderate CKD defined by eGFR.cys between 15-59. In routine clinical practice, it may not be easy to generate all the CH metrics as well as blood counts and biochemistry measures to forecast the risk of developing myeloid neoplasms in healthy individuals. Instead, it may be more productive to focus on individuals at high risk such as moderate CKD patients, and utilise CH to stratify the risk of myeloid malignancies.

The work presented in this thesis was developed using data from the UK Biobank, which provided a wide variety of detailed genotypic and phenotypic information for my project across the last 4 years. Many different research groups have been interested in CH and used data from the UK Biobank as well as other large genetic cohorts such as TOPMed and BBJ in their investigations. These studies collectively provide a valuable resource to validate my results and to elucidate the limitations of my study, for example (i) the UK Biobank cohort has a "healthy volunteer" bias that may be a consequence of the low recruitment response rate [287]. In comparison, the TOPMed cohort was enriched in lung, heart, and blood related diseases (>60% of participants); (ii) the UK Biobank phenotypic data lacks harmonisation, e.g the definition of myeloid malignancies, CKD, and CVD may vary between research groups; (iii) the UK Biobank lacks direct measures for inflammatory cytokines such as IL-6, an IL-1β; (iv) the UK Biobank lacks longitudinal information, which means that CH cannot be tracked over time. This last point is particularly important given the fact that clonal dynamics over time vary substantially between individuals. [382,383,396] Recently, the UK Biobank released WGS data that will allow the study of more genetic features related to CH such as non-coding driver mutations, and mutational signatures. In addition, the release of NMR-metabolomics will improve our understanding of inflammation associated with CH e.g. glycoprotein acetyls is an inflammatory biomarker associated with cardiovascular risk that should provide more sensitive estimates of the inflammatory state in comparison to CRP [397].

In summary, this thesis provides further evidence for the wide reaching significance of CH. The future implementation of CH as a marker to predict the risk of developing haematological malignancies, non-malignant diseases, or the response to anti-inflammatory medications will need the following actions: (i) more extensive genetic and phenotypic data is required to develop more detailed clinical insights at the driver gene level; (ii) large numbers of samples with non-European ancestry will need to be included and considered separately, e.g. rs144418061 near *TET2* was specifically associated with CH in African individuals; (iii) design of a fast and cheap sequencing assay to detect CH will be needed to enable large longitudinal studies of CH, e.g. a recent study has developed a cost-effective panel of 11 genes associated with CH using a single-molecule molecular inversion probe sequencing approach [398]; (iv) development of a standard method for processing NGS data, and calling driver mutations in blood samples, and (v) introduction of CH measurements into clinical trials of anti-inflammatory medications to enable the clinical utility of CH in predicting therapeutic response to be determined, as well as the effect on CH clone size to be evaluated. I anticipate that the very rapid progress in understanding the causes and consequences of CH is likely to lead to real health benefits in the coming years.

Appendices

The supplementary data were deposited under DOI: https://doi.org/10.5258/SOTON/D2351

Appendix A: Supplementary Data File

Description: the accompanying Excel spreadsheet presents supplementary tables (n=6) of Chapter 3

Filename: chapter 3 supplementary tables submit.xls

Appendix B: Supplementary Data File

Description: the accompanying Excel spreadsheet presents supplementary tables(n=3) of Chapter 4

Filename: chapter 4 supplementary tables submit.xls

References

1. Orkin SH, Zon LI. Hematopoiesis and stem cells: plasticity versus developmental heterogeneity. *Nature Immunology* 2002;3(4):323-28.

2. Sun J, Ramos A, Chapman B, et al. Clonal dynamics of native haematopoiesis. *Nature* 2014;514(7522):322-27.

3. Greaves M, Maley CC. Clonal evolution in cancer. *Nature* 2012;481(7381):306-13.

4. Busque L, Patel JP, Figueroa ME, et al. Recurrent somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis. *Nature Genetics* 2012;44(11):1179-81.

5. Xu X, Zhang Q, Luo J, et al. JAK2V617F: prevalence in a large Chinese hospital population. *Blood* 2007;109(1):339-42.

6. Laurie CC, Laurie CA, Rice K, et al. Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nature Genetics* 2012;44(6):642-50.

7. Jacobs KB, Yeager M, Zhou W, et al. Detectable clonal mosaicism and its relationship to aging and cancer. *Nature Genetics* 2012;44(6):651–58.

8. Genovese G, Kähler AK, Handsaker RE, et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *New England Journal of Medicine* 2014;371(26):2477-87.

9. Zink F, Stacey SN, Norddahl GL, et al. Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* 2017;130(6):742-52.

10. Jaiswal S, Fontanillas P, Flannick J, et al. Age-related clonal hematopoiesis associated with adverse outcomes. *New England Journal of Medicine* 2014;371(26):2488-98.

11. Shlush LI, Zandi S, Mitchell A, et al. Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature* 2014;506(7488):328–33.

12. Landgren O, Kyle RA, Pfeiffer RM, et al. Monoclonal gammopathy of undetermined significance (MGUS) consistently precedes multiple myeloma: a prospective study. *Blood* 2009;113(22):5412-17.

13. Kyle RA, Therneau TM, Rajkumar SV, et al. A long-term study of prognosis in monoclonal gammopathy of undetermined significance. *N Engl J Med* 2002;346(8):564-9.

14. Young BD, Debernardi S, Lillington DM, et al. A role for mitotic recombination in leukemogenesis. *Advances in Enzyme Regulation* 2006;46:90-97.

15. Stephens K, Weaver M, Leppig KA, et al. Interstitial uniparental isodisomy at clustered breakpoint intervals is a frequent mechanism of NF1 inactivation in myeloid malignancies. *Blood* 2006;108(5):1684-89.

16. Makishima H, Maciejewski JP. Pathogenesis and consequences of uniparental disomy in cancer. *Clin Cancer Res* 2011;17(12):3913-23.

17. Raghavan M, Lillington DM, Skoulakis S, et al. Genome-wide single nucleotide polymorphism analysis reveals frequent partial uniparental disomy due to somatic recombination in acute myeloid leukemias. *Cancer Research* 2005;65(2):375-78.

18. Maciejewski JP, Mufti GJ. Whole genome scanning as a cytogenetic tool in hematologic malignancies. *Blood* 2008;112(4):965-74.

19. Metzker ML. Sequencing technologies — the next generation. *Nature Reviews Genetics* 2010;11(1):31-46.

20. Beroukhim R, Mermel CH, Porter D, et al. The landscape of somatic copy-number alteration across human cancers. *nature* 2010;463(7283):899.

21. Zack TI, Schumacher SE, Carter SL, et al. Pan-cancer patterns of somatic copy number alteration. *Nature Genetics* 2013;45(10):1134-40.

22. Martincorena I, Raine KM, Gerstung M, et al. Universal patterns of selection in cancer and somatic tissues. *Cell* 2017;171(5):1029-41. e21.

23. Horn S, Figl A, Rachakonda PS, et al. TERT promoter mutations in familial and sporadic melanoma. *Science* 2013;339(6122):959-61.

24. Huang FW, Hodis E, Xu MJ, et al. Highly recurrent TERT promoter mutations in human melanoma. *Science* 2013;339(6122):957-9.

25. Rheinbay E, Nielsen MM, Abascal F, et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* 2020;578(7793):102-11.

26. Erady C, Boxall A, Puntambekar S, et al. Pan-cancer analysis of transcripts encoding novel open-reading frames (nORFs) and their potential biological functions. *npj Genomic Medicine* 2021;6(1):4.

27. Huang JZ, Chen M, Chen D, et al. A Peptide Encoded by a Putative lncRNA HOXB-AS3 Suppresses Colon Cancer Growth. *Mol Cell* 2017;68(1):171-84.e6.

28. Kralovics R, Passamonti F, Buser AS, et al. A gain-of-function mutation of JAK2 in myeloproliferative disorders. *New England Journal of Medicine* 2005;352(17):1779-90.

29. Szpurka H, Gondek L, Mohan S, et al. UPD1p indicates the presence of MPL W515L mutation in RARS-T, a mechanism analogous to UPD9p and JAK2 V617F mutation. *Leukemia* 2009;23(3):610.

30. Klampfl T, Gisslinger H, Harutyunyan AS, et al. Somatic mutations of calreticulin in myeloproliferative neoplasms. *New England Journal of Medicine* 2013;369(25):2379-90.

31. Makishima H, Jankowska A, Tiu R, et al. Novel homo-and hemizygous mutations in EZH2 in myeloid malignancies. *Leukemia* 2010;24(10):1799-804.

32. Ernst T, Chase AJ, Score J, et al. Inactivating mutations of the histone methyltransferase gene EZH2 in myeloid disorders. *Nature Genetics* 2010;42(8):722.

33. Grand FH, Hidalgo-Curtis CE, Ernst T, et al. Frequent CBL mutations associated with 11q acquired uniparental disomy in myeloproliferative neoplasms. *Blood* 2009;113(24):6182-92.

34. Dunbar AJ, Gondek LP, O'Keefe CL, et al. 250K single nucleotide polymorphism array karyotyping identifies acquired uniparental disomy and homozygous mutations, including novel missense substitutions of c-Cbl, in myeloid malignancies. *Cancer research* 2008;68(24):10349-57.

35. Vannucchi AM, Antonioli E, Guglielmelli P, et al. Clinical profile of homozygous JAK2 617V> F mutation in patients with polycythemia vera or essential thrombocythemia. *Blood, The Journal of the American Society of Hematology* 2007;110(3):840-46.

36. Busque L, Patel JP, Figueroa ME, et al. Recurrent somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis. *Nature Genetics* 2012;44(11):1179.

37. Xie M, Lu C, Wang J, et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nature Medicine* 2014;20(12):1472.

38. Young AL, Challen GA, Birmann BM, et al. Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. *Nature Communications* 2016;7(1):1-7.

39. Schmitt MW, Kennedy SR, Salk JJ, et al. Detection of ultra-rare mutations by next-generation sequencing. *Proceedings of the National Academy of Sciences* 2012;109(36):14508-13.

40. Rawstron AC, Bennett FL, O'Connor SJ, et al. Monoclonal B-cell lymphocytosis and chronic lymphocytic leukemia. *N Engl J Med* 2008;359(6):575-83.

41. McKerrell T, Park N, Moreno T, et al. Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hemopoiesis. *Cell reports* 2015;10(8):1239-45.

42. Abelson S, Collord G, Ng SWK, et al. Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* 2018;559(7714):400-04.

43. Desai P, Mencia-Trinchant N, Savenkov O, et al. Somatic mutations precede acute myeloid leukemia years before diagnosis. *Nature Medicine* 2018;24(7):1015-23.

44. Okano M, Xie S, Li E. Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nature Genetics* 1998;19(3):219-20.

45. Ley TJ, Ding L, Walter MJ, et al. DNMT3A mutations in acute myeloid leukemia. *N Engl J Med* 2010;363(25):2424-33.

46. Thol F, Damm F, Lüdeking A, et al. Incidence and prognostic influence of DNMT3A mutations in acute myeloid leukemia. *J Clin Oncol* 2011;29(21):2889-96.

47. Kim SJ, Zhao H, Hardikar S, et al. A DNMT3A mutation common in AML exhibits dominant-negative effects in murine ES cells. *Blood* 2013;122(25):4086-9.

48. Patel JP, Gönen M, Figueroa ME, et al. Prognostic relevance of integrated genetic profiling in acute myeloid leukemia. *N Engl J Med* 2012;366(12):1079-89.

49. Delhommeau F, Dupont S, James C, et al. TET2 Is a Novel Tumor Suppressor Gene Inactivated in Myeloproliferative Neoplasms: Identification of a Pre-JAK2 V617F Event. *Blood* 2008;112(11):lba-3.

50. Itzykson R, Kosmider O, Cluzeau T, et al. Impact of TET2 mutations on response rate to azacitidine in myelodysplastic syndromes and low blast count acute myeloid leukemias. *Leukemia* 2011;25(7):1147-52.

51. Figueroa ME, Abdel-Wahab O, Lu C, et al. Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. *Cancer Cell* 2010;18(6):553-67.

52. Wu D, Hu D, Chen H, et al. Glucose-regulated phosphorylation of TET2 by AMPK reveals a pathway linking diabetes to cancer. *Nature* 2018;559(7715):637.

53. Yin R, Mao S-Q, Zhao B, et al. Ascorbic Acid Enhances Tet-Mediated 5-Methylcytosine Oxidation and Promotes DNA Demethylation in Mammals. *Journal of the American Chemical Society* 2013;135(28):10396-403.

54. Douet-Guilbert N, Laï JL, Basinko A, et al. Fluorescence in situ hybridization characterization of ider(20q) in myelodysplastic syndrome. *Br J Haematol* 2008;143(5):716-20.

55. Carbuccia N, Murati A, Trouplin V, et al. Mutations of ASXL1 gene in myeloproliferative neoplasms. *Leukemia* 2009;23(11):2183-86.

56. Tamburri S, Lavarone E, Fernández-Pérez D, et al. Histone H2AK119 Mono-Ubiquitination Is Essential for Polycomb-Mediated Transcriptional Repression. *Molecular Cell* 2020;77(4):840-56.e5.

57. Abdel-Wahab O, Adli M, LaFave Lindsay M, et al. ASXL1 Mutations Promote Myeloid Transformation through Loss of PRC2-Mediated Gene Repression. *Cancer Cell* 2012;22(2):180-93.

58. Scheuermann JC, de Ayala Alonso AG, Oktaba K, et al. Histone H2A deubiquitinase activity of the Polycomb repressive complex PR-DUB. *Nature* 2010;465(7295):243-47.

59. Vannucchi AM, Lasho TL, Guglielmelli P, et al. Mutations and prognosis in primary myelofibrosis. *Leukemia* 2013;27(9):1861-9.

60. Gelsi-Boyer V, Brecqueville M, Devillier R, et al. Mutations in ASXL1 are associated with poor prognosis across the spectrum of malignant myeloid diseases. *Journal of Hematology & Oncology* 2012;5(1):12.

61. Papaemmanuil E, Cazzola M, Boultwood J, et al. Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *New England Journal of Medicine* 2011;365(15):1384-95.

62. Yoshida K, Sanada M, Shiraishi Y, et al. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* 2011;478(7367):64-69.

63. Abdel-Wahab O, Levine R. The spliceosome as an indicted conspirator in myeloid malignancies. *Cancer Cell* 2011;20(4):420-3.

64. Fiscella M, Zhang H, Fan S, et al. Wip1, a novel human protein phosphatase that is induced in response to ionizing radiation in a p53-dependent manner. *Proc Natl Acad Sci U S A* 1997;94(12):6048-53.

65. Hsu JI, Dayaram T, Tovy A, et al. PPM1D Mutations Drive Clonal Hematopoiesis in Response to Cytotoxic Chemotherapy. *Cell Stem Cell* 2018;23(5):700-13.e6.

66. Kahn JD, Miller PG, Silver AJ, et al. PPM1D-truncating mutations confer resistance to chemotherapy and sensitivity to PPM1D inhibition in hematopoietic cells. *Blood* 2018;132(11):1095-105.

67. Marcellino B, Tripodi J, Bar-Natan M, et al. Significance of Abnormalities of PPM1D in Myeloproliferative Neoplasms. *Blood* 2019;134:4207.

68. Levine RL, Pardanani A, Tefferi A, et al. Role of JAK2 in the pathogenesis and therapy of myeloproliferative disorders. *Nature Reviews Cancer* 2007;7(9):673-83.

69. James C, Ugo V, Le Couédic J-P, et al. A unique clonal JAK2 mutation leading to constitutive signalling causes polycythaemia vera. *Nature* 2005;434(7037):1144-448.

70. Levine RL, Wadleigh M, Cools J, et al. Activating mutation in the tyrosine kinase JAK2 in polycythemia vera, essential thrombocythemia, and myeloid metaplasia with myelofibrosis. *Cancer Cell* 2005;7(4):387-97.

71. Baxter EJ, Scott LM, Campbell PJ, et al. Acquired mutation of the tyrosine kinase JAK2 in human myeloproliferative disorders. *The Lancet* 2005;365(9464):1054-61.

72. Lindauer K, Loerting T, Liedl KR, et al. Prediction of the structure of human Janus kinase 2 (JAK2) comprising the two carboxy-terminal domains reveals a mechanism for autoregulation. *Protein Eng* 2001;14(1):27-37.

73. Ayaz P, Hammarén HM, Raivola J, et al. Structural models of full-length JAK2 kinase. *bioRxiv* 2019:727727.

74. Bousoik E, Montazeri Aliabadi H. "Do We Know Jack" About JAK? A Closer Look at JAK/STAT Signaling Pathway. *Front Oncol* 2018;8:287.

75. Pierre RV, Hoagland HC. Age-associated aneuploidy: loss of Y chromosome from human bone marrow cells with aging. *Cancer* 1972;30(4):889-94.

76. Forsberg LA, Rasi C, Malmqvist N, et al. Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. *Nature Genetics* 2014;46(6):624-28.

77. Dumanski JP, Rasi C, Lönn M, et al. Mutagenesis. Smoking is associated with mosaic loss of chromosome Y. *Science* 2015;347(6217):81-3.

78. Zhou W, Machiela MJ, Freedman ND, et al. Mosaic loss of chromosome Y is associated with common variation near TCL1A. *Nature Genetics* 2016;48(5):563-68.

79. Laine J, Künstle G, Obata T, et al. The protooncogene TCL1 is an Akt kinase coactivator. *Mol Cell* 2000;6(2):395-407.

80. Thompson DJ, Genovese G, Halvardson J, et al. Genetic predisposition to mosaic Y chromosome loss in blood. *Nature* 2019;575(7784):652-57.

81. Terao C, Momozawa Y, Ishigaki K, et al. GWAS of mosaic loss of chromosome Y highlights genetic effects on blood cell differentiation. *Nature Communications* 2019;10(1):4719.

82. Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. *Nat Rev Genet* 2007;8(8):610-8.

83. Watson CJ, Papula A, Poon GY, et al. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science* 2020;367(6485):1449-54.

84. Robertson NA, Latorre-Crespo E, Terradas-Terradas M, et al. Longitudinal dynamics of clonal hematopoiesis identifies gene-specific fitness effects. *BioRxiv* 2021:2021.05.27.446006.

85. Bolton KL, Ptashkin RN, Gao T, et al. Cancer therapy shapes the fitness landscape of clonal hematopoiesis. *Nature Genetics* 2020;52(11):1219-26.

86. Saiki R, Momozawa Y, Nannya Y, et al. Combined landscape of single-nucleotide variants and copy number alterations in clonal hematopoiesis. *Nature Medicine* 2021;27(7):1239-49.

87. Young AL, Tong RS, Birmann BM, et al. Clonal hematopoiesis and risk of acute myeloid leukemia. *Haematologica* 2019;104(12):2410-17.

88. Jaiswal S, Ebert BL. Clonal hematopoiesis in human aging and disease. *Science* 2019;366(6465).

89. Jaiswal S, Natarajan P, Silver AJ, et al. Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease. *New England Journal of Medicine* 2017;377(2):111-21.

90. Rauch PJ, Silver AJ, Gopakumar J, et al. Loss-of-function mutations in dnmt3a and tet2 lead to accelerated atherosclerosis and convergent macrophage phenotypes in mice. *Blood* 2018;132(Supplement 1):745-45.

91. Wang W, Liu W, Fidler T, et al. Macrophage inflammation, erythrophagocytosis, and accelerated atherosclerosis in Jak2 V617F mice. *Circulation Research* 2018;123(11):e35-e47.

92. Bonnefond A, Skrobek B, Lobbens S, et al. Association between large detectable clonal mosaicism and type 2 diabetes with vascular complications. *Nature Genetics* 2013;45(9):1040.

93. Wain LV, Shrine N, Miller S, et al. Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *The Lancet Respiratory Medicine* 2015;3(10):769-81.

94. Buscarlet M, Provost S, Zada YF, et al. DNMT3A and TET2 dominate clonal hematopoiesis and demonstrate benign phenotypes and different genetic predispositions. *Blood, The Journal of the American Society of Hematology* 2017;130(6):753-62.

95. West RR, Hsu AP, Holland SM, et al. Acquired ASXL1 mutations are common in patients with inherited GATA2 mutations and correlate with myeloid transformation. *Haematologica* 2014;99(2):276-81.

96. Bödör C, Renneville A, Smith M, et al. Germ-line GATA2 p. THR354MET mutation in familial myelodysplastic syndrome with acquired monosomy 7 and ASXL1 mutation demonstrating rapid onset and poor survival. *Haematologica* 2012;97(6):890-94.

97. Vogelstein B, Papadopoulos N, Velculescu VE, et al. Cancer genome landscapes. *Science* 2013;339(6127):1546-58.

98. Tomasetti C, Vogelstein B, Parmigiani G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proceedings of the National Academy of Sciences* 2013;110(6):1999-2004.

99. Welch JS, Ley TJ, Link DC, et al. The origin and evolution of mutations in acute myeloid leukemia. *Cell* 2012;150(2):264-78.

100. Steensma DP, Bejar R, Jaiswal S, et al. Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood* 2015;126(1):9-16.

101. Arber DA, Orazi A, Hasserjian R, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* 2016;127(20):2391-405.

102. Grimwade D, Hills RK, Moorman AV, et al. Refinement of cytogenetic classification in acute myeloid leukemia: determination of prognostic significance of rare recurring chromosomal abnormalities among 5876 younger adult patients treated in the United Kingdom Medical Research Council trials. *Blood* 2010;116(3):354-65.

103. Mrózek K, Heerema NA, Bloomfield CD. Cytogenetics in acute leukemia. *Blood reviews* 2004;18(2):115-36.

104. Döhner H, Estey E, Grimwade D, et al. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood* 2017;129(4):424-47.

105. Thiede C, Koch S, Creutzig E, et al. Prevalence and prognostic impact of NPM1 mutations in 1485 adult patients with acute myeloid leukemia (AML). *Blood* 2006;107(10):4011-20.

106. Kottaridis PD, Gale RE, Frew ME, et al. The presence of a FLT3 internal tandem duplication in patients with acute myeloid leukemia (AML) adds important prognostic information to cytogenetic risk group and response to the first cycle of chemotherapy: analysis of 854 patients from the United Kingdom Medical Research Council AML 10 and 12 trials. *Blood* 2001;98(6):1752-59.

107. Papaemmanuil E, Gerstung M, Bullinger L, et al. Genomic classification and prognosis in acute myeloid leukemia. *New England Journal of Medicine* 2016;374(23):2209-21.

108. TCGA. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *New England Journal of Medicine* 2013;368(22):2059-74.

109. Faderl S, Talpaz M, Estrov Z, et al. The biology of chronic myeloid leukemia. *New England Journal of Medicine* 1999;341(3):164-72.

110. Tefferi A, Lasho T, Finke C, et al. CALR vs JAK2 vs MPL-mutated or triple-negative myelofibrosis: clinical, cytogenetic and molecular comparisons. *Leukemia* 2014;28(7):1472.

111. Scott LM, Tong W, Levine RL, et al. JAK2 exon 12 mutations in polycythemia vera and idiopathic erythrocytosis. *New England Journal of Medicine* 2007;356(5):459-68.

112. Pikman Y, Lee BH, Mercher T, et al. MPLW515L is a novel somatic activating mutation in myelofibrosis with myeloid metaplasia. *PLoS medicine* 2006;3(7):e270.

113. Nangalia J, Massie CE, Baxter EJ, et al. Somatic CALR mutations in myeloproliferative neoplasms with nonmutated JAK2. *New England Journal of Medicine* 2013;369(25):2391-405.

114. Kralovics R, Guan Y, Prchal JT. Acquired uniparental disomy of chromosome 9p is a frequent stem cell defect in polycythemia vera. *Experimental Hematology* 2002;30(3):229-36.

115. Cazzola M, Kralovics R. From Janus kinase 2 to calreticulin: the clinically relevant genomic landscape of myeloproliferative neoplasms. *Blood* 2014;123(24):3714-19.

116. Zhang S-J, Rampal R, Manshouri T, et al. Genetic analysis of patients with leukemic transformation of myeloproliferative neoplasms shows recurrent SRSF2 mutations that are associated with adverse outcome. *Blood* 2012;119(19):4480-85.

117. Tefferi A, Guglielmelli P, Lasho TL, et al. Mutation-enhanced international prognostic systems for essential thrombocythaemia and polycythaemia vera. *British journal of haematology* 2020;189(2):291-302.

118. Tefferi A, Guglielmelli P, Lasho TL, et al. MIPSS70+ Version 2.0: Mutation and Karyotype-Enhanced International Prognostic Scoring System for Primary Myelofibrosis. *J Clin Oncol* 2018;36(17):1769-70.

119. Malcovati L, Hellström-Lindberg E, Bowen D, et al. Diagnosis and treatment of primary myelodysplastic syndromes in adults: recommendations from the European LeukemiaNet. *Blood* 2013;122(17):2943-64.

120. Schanz J, Tüchler H, Solé F, et al. New comprehensive cytogenetic scoring system for primary myelodysplastic syndromes (MDS) and oligoblastic acute myeloid leukemia after MDS derived from an international database merge. *Journal of Clinical Oncology* 2012;30(8):820-29.

121. Greenberg PL, Tuechler H, Schanz J, et al. Revised International Prognostic Scoring System for Myelodysplastic Syndromes. *Blood* 2012;120(12):2454-65.

122. Haferlach T, Nagata Y, Grossmann V, et al. Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. *Leukemia* 2014;28(2):241-47.

123. Papaemmanuil E, Gerstung M, Malcovati L, et al. Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood, The Journal of the American Society of Hematology* 2013;122(22):3616-27.

124. Orazi A, Germing U. The myelodysplastic/myeloproliferative neoplasms: myeloproliferative diseases with dysplastic features. *Leukemia* 2008;22(7):1308-19.

125. Patnaik MM, Tefferi A. Cytogenetic and molecular abnormalities in chronic myelomonocytic leukemia. *Blood cancer journal* 2016;6(2):e393.

126. Elena C, Gallì A, Such E, et al. Integrating clinical features and genetic lesions in the risk assessment of patients with chronic myelomonocytic leukemia. *Blood* 2016;128(10):1408-17.

127. Gondek LP, Tiu R, O'Keefe CL, et al. Chromosomal lesions and uniparental disomy detected by SNP arrays in MDS, MDS/MPD, and MDS-derived AML. *Blood* 2008;111(3):1534-42.

128. Tartaglia M, Niemeyer CM, Fragale A, et al. Somatic mutations in PTPN11 in juvenile myelomonocytic leukemia, myelodysplastic syndromes and acute myeloid leukemia. *Nature Genetics* 2003;34(2):148-50.

129. Loh ML, Sakai DS, Flotho C, et al. Mutations in CBL occur frequently in juvenile myelomonocytic leukemia. *Blood* 2009;114(9):1859-63.

130. Sakaguchi H, Okuno Y, Muramatsu H, et al. Exome sequencing identifies secondary mutations of SETBP1 and JAK3 in juvenile myelomonocytic leukemia. *Nature Genetics* 2013;45(8):937-41.

131. Piazza R, Valletta S, Winkelmann N, et al. Recurrent SETBP1 mutations in atypical chronic myeloid leukemia. *Nature Genetics* 2013;45(1):18-24.

132. Maxson JE, Gotlib J, Pollyea DA, et al. Oncogenic CSF3R mutations in chronic neutrophilic leukemia and atypical CML. *New England Journal of Medicine* 2013;368(19):1781-90.

133. Ostergaard P, Simpson MA, Connell FC, et al. Mutations in GATA2 cause primary lymphedema associated with a predisposition to acute myeloid leukemia (Emberger syndrome). *Nature Genetics* 2011;43(10):929-31.

134. Pabst T, Mueller BU, Zhang P, et al. Dominant-negative mutations of CEBPA, encoding CCAAT/enhancer binding protein-α (C/EBPα), in acute myeloid leukemia. *Nature Genetics* 2001;27(3):263-70.

135. Zhang MY, Churpek JE, Keel SB, et al. Germline ETV6 mutations in familial thrombocytopenia and hematologic malignancy. *Nature Genetics* 2015;47(2):180-85.

136. Béri-Dexheimer M, Latger-Cannard V, Philippe C, et al. Clinical phenotype of germline RUNX1 haploinsufficiency: from point mutations to large genomic deletions. *European Journal of Human Genetics* 2008;16(8):1014-18.

137. Polprasert C, Schulze I, Sekeres MA, et al. Inherited and somatic defects in DDX41 in myeloid neoplasms. *Cancer Cell* 2015;27(5):658-70.

138. Jones AV, Chase A, Silver RT, et al. JAK2 haplotype is a major risk factor for the development of myeloproliferative neoplasms. *Nature Genetics* 2009;41(4):446-49.

139. Olcaydu D, Harutyunyan A, Jäger R, et al. A common JAK2 haplotype confers susceptibility to myeloproliferative neoplasms. *Nature Genetics* 2009;41(4):450.

140. Kilpivaara O, Mukherjee S, Schram AM, et al. A germline JAK2 SNP is associated with predisposition to the development of JAK2V617F-positive myeloproliferative neoplasms. *Nature Genetics* 2009;41(4):455-59.

141. Campbell PJ. Somatic and germline genetics at the JAK2 locus. *Nature Genetics* 2009;41(4):385-86.

142. Vilaine M, Olcaydu D, Harutyunyan A, et al. Homologous recombination of wild-type JAK2, a novel early step in the development of myeloproliferative neoplasm. *Blood, The Journal of the American Society of Hematology* 2011;118(24):6468-70.

143. Tapper W, Jones AV, Kralovics R, et al. Genetic variation at MECOM, TERT, JAK2 and HBS1L-MYB predisposes to myeloproliferative neoplasms. *Nature Communications* 2015;6(1):6691.

144. Bao EL, Nandakumar SK, Liao X, et al. Inherited myeloproliferative neoplasm risk affects haematopoietic stem cells. *Nature* 2020;586(7831):769-75.

145. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell* 2011;144(5):646-74.

146. Anderson L, Pfeiffer R, Landgren O, et al. Risks of myeloid malignancies in patients with autoimmune conditions. *British journal of cancer* 2009;100(5):822-28.

147. Kristinsson SY, Landgren O, Samuelsson J, et al. Autoimmunity and the risk of myeloproliferative neoplasms. *Haematologica* 2010;95(7):1216-20.

148. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin* 2020;70(1):7-30.

149. Sánchez-Aguilera A, Arranz L, Martín-Pérez D, et al. Estrogen Signaling Selectively Induces Apoptosis of Hematopoietic Progenitors and Myeloid Neoplasms without Harming Steady-State Hematopoiesis. *Cell Stem Cell* 2014;15(6):791-804.

150. Burney BO, Hayes TG, Smiechowska J, et al. Low testosterone levels and increased inflammatory markers in patients with cancer and relationship with cachexia. *J Clin Endocrinol Metab* 2012;97(5):E700-9.

151. Shahidi NT. Androgens and erythropoiesis. *N Engl J Med* 1973;289(2):72-80.

152. Chan G, DiVenuti G, Miller K. Danazol for the treatment of thrombocytopenia in patients with myelodysplastic syndrome. *Am J Hematol* 2002;71(3):166-71.

153. Tefferi A, Vaidya R, Caramazza D, et al. Circulating interleukin (IL)-8, IL-2R, IL-12, and IL-15 levels are independently prognostic in primary myelofibrosis: a comprehensive cytokine profiling study. *J Clin Oncol* 2011;29(10):1356-63.

154. Garbers C, Monhasery N, Aparicio-Siegmund S, et al. The interleukin-6 receptor Asp358Ala single nucleotide polymorphism rs2228145 confers increased proteolytic conversion rates by ADAM proteases. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* 2014;1842(9):1485-94.

155. Pedersen KM, Çolak Y, Ellervik C, et al. Loss-of-function polymorphism in IL6R reduces risk of JAK2V617F somatic mutation and myeloproliferative neoplasm: A Mendelian randomization study. *EClinicalMedicine* 2020;21:100280.

156. Salvagno GL, Sanchis-Gomar F, Picanza A, et al. Red blood cell distribution width: A simple parameter with multiple clinical applications. *Critical Reviews in Clinical Laboratory Sciences* 2015;52(2):86-105.

157. Lippi G, Targher G, Montagnana M, et al. Relation between red blood cell distribution width and inflammatory biomarkers in a large cohort of unselected outpatients. *Archives of Pathology & Laboratory Medicine* 2009;133(4):628-32.

158. Ludwiczek S, Aigner E, Theurl I, et al. Cytokine-mediated regulation of iron transport in human monocytic cells. *Blood, The Journal of the American Society of Hematology* 2003;101(10):4148-54.

159. Tsuboi S, Miyauchi K, Kasai T, et al. Impact of red blood cell distribution width on long-term mortality in diabetic patients after percutaneous coronary intervention. *Circulation Journal* 2013;77(2):456-61.

160. Arbel Y, Shacham Y, Finkelstein A, et al. Red blood cell distribution width (RDW) and long-term survival in patients with ST elevation myocardial infarction. *Thrombosis Research* 2014;134(5):976-79.

161. Sousa R, Gonçalves C, Guerra IC, et al. Increased red cell distribution width in Fanconi anemia: a novel marker of stress erythropoiesis. *Orphanet journal of rare diseases* 2016;11(1):102.

162. Buckstein R, Jang K, Friedlich J, et al. Estimating the prevalence of myelodysplastic syndromes in patients with unexplained cytopenias: a retrospective study of 322 bone marrows. *Leuk Res* 2009;33(10):1313-8.

163. Tillett WS, Francis T. SEROLOGICAL REACTIONS IN PNEUMONIA WITH A NON-PROTEIN SOMATIC FRACTION OF PNEUMOCOCCUS. *J Exp Med* 1930;52(4):561-71.

164. McCarty M. THE OCCURRENCE DURING ACUTE INFECTIONS OF A PROTEIN NOT NORMALLY PRESENT IN THE BLOOD : IV. CRYSTALLIZATION OF THE C-REACTIVE PROTEIN. *J Exp Med* 1947;85(5):491-8.

165. Vigushin DM, Pepys MB, Hawkins PN. Metabolic and scintigraphic studies of radioiodinated human C-reactive protein in health and disease. *J Clin Invest* 1993;91(4):1351-7.

166. Barbui T, Carobbio A, Finazzi G, et al. Elevated C-reactive protein is associated with shortened leukemia-free survival in patients with myelofibrosis. *Leukemia* 2013;27(10):2084-86.

167. Lussana F, Carobbio A, Salmoiraghi S, et al. Driver mutations (JAK2V617F, MPLW515L/K or CALR), pentraxin-3 and C-reactive protein in essential thrombocythemia and polycythemia vera. *Journal of Hematology & Oncology* 2017;10(1):54.

168. Busque L, Sun M, Buscarlet M, et al. High-sensitivity C-reactive protein is associated with clonal hematopoiesis of indeterminate potential. *Blood Advances* 2020;4(11):2430-38.

169. Bick AG, Weinstock JS, Nandakumar SK, et al. Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature* 2020;586(7831):763-68.

170. Stämpfli MR, Anderson GP. How cigarette smoke skews immune responses to promote infection, lung disease and cancer. *Nature Reviews Immunology* 2009;9(5):377-84.

171. Doll R, Peto R, Wheatley K, et al. Mortality in relation to smoking: 40 years' observations on male British doctors. *BMJ* 1994;309(6959):901-11.

172. Moorman A, Roman E, Cartwright R, et al. Smoking and the risk of acute myeloid leukaemia in cytogenetic subgroups. *British journal of cancer* 2002;86(1):60.

173. Caspersson T, Zech L, Johansson C. Differential binding of alkylating fluorochromes in human chromosomes. *Experimental Cell Research* 1970;60(3):315-19.

174. Levsky JM, Singer RH. Fluorescence in situ hybridization: past, present and future. *Journal of Cell Science* 2003;116(14):2833-38.

175. Pinkel D, Albertson DG. Array comparative genomic hybridization and its applications in cancer. *Nature Genetics* 2005;37(6):S11-S17.

176. du Manoir S, Speicher MR, Joos S, et al. Detection of complete and partial chromosome gains and losses by comparative genomic in situ hybridization. *Human Genetics* 1993;90(6):590-610.

177. Sachidanandam R, Weissman D, Schmidt SC, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 2001;409(6822):928-34.

178. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018;562(7726):203-09.

179. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences* 1977;74(12):5463-67.

180. Luckey JA, Drossman H, Kostichka AJ, et al. High speed DNA sequencing by capillary electrophoresis. *Nucleic Acids Research* 1990;18(15):4417-21.

181. Prober JM, Trainor GL, Dam RJ, et al. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* 1987;238(4825):336-41.

182. Fuller CW, Middendorf LR, Benner SA, et al. The challenges of sequencing by synthesis. *Nature Biotechnology* 2009;27(11):1013-23.

183. Check Hayden E. Technology: The $1,000 genome. *Nature* 2014;507(7492):294-95.

184. Meyer M, Kircher M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols* 2010(6):pdb. prot5448.

185. Mamanova L, Coffey AJ, Scott CE, et al. Target-enrichment strategies for next-generation sequencing. *Nature Methods* 2010;7(2):111-18.

186. Hardenbol P, Banér J, Jain M, et al. Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nature Biotechnology* 2003;21(6):673-78.

187. Tewhey R, Warner JB, Nakano M, et al. Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nature Biotechnology* 2009;27(11):1025-31.

188. Gnirke A, Melnikov A, Maguire J, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology* 2009;27(2):182.

189. Mitra RD, Church GM. In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Res* 1999;27(24):e34.

190. Wang DG, Fan J-B, Siao C-J, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 1998;280(5366):1077-82.

191. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* 2009;10(1):57-63.

192. Hosomichi K, Mitsunaga S, Nagasaki H, et al. A Bead-based Normalization for Uniform Sequencing depth (BeNUS) protocol for multi-samples sequencing exemplified by HLA-B. *BMC Genomics* 2014;15(1):645.

193. Mills RE, Walter K, Stewart C, et al. Mapping copy number variation by population-scale genome sequencing. *Nature* 2011;470(7332):59-65.

194. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv Preprint; arXiv:1303.3997* 2013.

195. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* 2015;12(4):357-60.

196. Broadinstitute. *Picard Toolkit*. http://broadinstitute.github.io/picard/.

197. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 2010;20(9):1297-303.

198. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* 2011;43(5):491-98.

199. Poplin R, Ruano-Rubio V, DePristo MA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv* 2017:201178.

200. Benjamin D, Sato T, Cibulskis K, et al. Calling somatic snvs and indels with mutect2. *BioRxiv* 2019:861054.

201. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* 2010;38(16):e164-e64.

202. Dawoud AAZ, Tapper WJ, Cross NCP. Clonal myelopoiesis in the UK Biobank cohort: ASXL1 mutations are strongly associated with smoking. *Leukemia* 2020;34(10):2660-72.

203. Dawoud AAZ, Gilbert RD, Tapper WJ, et al. Clonal myelopoiesis promotes adverse outcomes in chronic kidney disease. *Leukemia* 2022;36(2):507-15.

204. Brämer GR. International statistical classification of diseases and related health problems. Tenth revision. *World health statistics quarterly. Rapport trimestriel de statistiques sanitaires mondiales* 1988;41(1):32-36.

205. Beroukhim R, Lin M, Park Y, et al. Inferring loss-of-heterozygosity from unpaired tumors using high-density oligonucleotide SNP arrays. *PLoS Computational Biology* 2006;2(5):e41.

206. Lindgren D, Höglund M, Vallon-Christersson J. Genotyping Techniques to Address Diversity in Tumors. In: Gisselsson D (ed.) *Advances in Cancer Research*: Academic Press; 2011 p151-82.

207. Staaf J, Lindgren D, Vallon-Christersson J, et al. Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. *Genome Biology* 2008;9(9):R136.

208. Van Hout CV, Tachmazidou I, Backman JD, et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* 2020;586(7831):749-56.

209. Genomicsplc. *WeCall*. https://github.com/Genomicsplc/wecall.

210. Regier AA, Farjoun Y, Larson DE, et al. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nature Communications* 2018;9(1):4038.

211. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001;29(1):308-11.

212. Yun T, Li H, Chang P-C, et al. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics* 2021;36(24):5582-89.

213. Forbes S, Bhamra G, Bamford S, et al. The catalogue of somatic mutations in cancer (COSMIC). *Current Protocols in Human Genetics* 2008;57(1):10.11. 1-10.11. 26.

214. Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* 1958;53(282):457-81.

215. Cox DR. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)* 1972;34(2):187-202.

216. Spooner A, Chen E, Sowmya A, et al. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific Reports* 2020;10(1):20410.

217. Breiman L. Random Forests. *Machine Learning* 2001;45(1):5-32.

218. Ishwaran H, Kogalur UB, Blackstone EH, et al. Random survival forests. *The Annals of Applied Statistics* 2008;2(3):841-60, 20.

219. Zhang Z, Zhao Y, Canes A, et al. Predictive analytics with gradient boosting in clinical medicine. *Ann Transl Med* 2019;7(7):152.

220. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?*. *International Journal of Epidemiology* 2003;32(1):1-22.

221. Benjamin DJ, Berger JO, Johannesson M, et al. Redefine statistical significance. *Nature Human Behaviour* 2018;2(1):6-10.

222. Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol* 2000;29(4):722-9.

223. Interleukin-6 Receptor Mendelian Randomisation Analysis C, Swerdlow DI, Holmes MV, et al. The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis. *Lancet (London, England)* 2012;379(9822):1214-24.

224. Burgess S, Thompson SG. Use of allele scores as instrumental variables for Mendelian randomization. *International Journal of Epidemiology* 2013;42(4):1134-44.

225. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology* 2013;37(7):658-65.

226. Bowden J, Davey Smith G, Haycock PC, et al. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology* 2016;40(4):304-14.

227. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology* 2015;44(2):512-25.

228. Zhao Q, Wang J, Hemani G, et al. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *Annals of Statistics* 2020;48(3):1742-69.

229. Taliun D, Harris DN, Kessler MD, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* 2021;590(7845):290-99.

230. Collins FS, Varmus H. A New Initiative on Precision Medicine. *New England Journal of Medicine* 2015;372(9):793-95.

231. Köttgen A, Pattaro C. The CKDGen Consortium: ten years of insights into the genetic basis of kidney function. *Kidney Int* 2020;97(2):236-42.

232. Wuttke M, Li Y, Li M, et al. A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nature Genetics* 2019;51(6):957-72.

233. Nagai A, Hirata M, Kamatani Y, et al. Overview of the BioBank Japan Project: Study design and profile. *J Epidemiol* 2017;27(3s):S2-s8.

234. Dumanski JP, Lambert JC, Rasi C, et al. Mosaic Loss of Chromosome Y in Blood Is Associated with Alzheimer Disease. *Am J Hum Genet* 2016;98(6):1208-19.

235. Aho AV, Kernighan BW, Weinberger PJ. Awk — a pattern scanning and processing language. *Software: Practice and Experience* 1979;9(4):267-79.

236. Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics* 1996;5(3):299-314.

237. Gel B, Serra E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* 2017;33(19):3088-90.

238. Lawrence M, Huber W, Pages H, et al. Software for computing and annotating genomic ranges. *PLoS Computational Biology* 2013;9(8).

239. Therneau TM, Grambsch PM. The Cox Model. In: Therneau TM, Grambsch PM (eds.) *Modeling Survival Data: Extending the Cox Model*. New York, NY: Springer; 2000 p39-77.

240. Grolemund G, Wickham H. Dates and times made easy with lubridate. *Journal of Statistical Software* 2011;40(3):1-25.

241. Qiu W, Chavarro J, Lazarus R, et al. powerSurvEpi: power and sample size calculation for survival analysis of epidemiological studies. *R package version 0.0* 2015;9.

242. Hemani G, Zheng J, Elsworth B, et al. The MR-Base platform supports systematic causal inference across the human phenome. *Elife* 2018;7:e34408.

243. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* 2020;17(3):261-72.

244. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* 2011;12:2825-30.

245. Pölsterl S. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. *J. Mach. Learn. Res.* 2020;21(212):1-6.

246. Hunter JD. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* 2007;9(3):90-95.

247. Waskom ML. Seaborn: statistical data visualization. *Journal of Open Source Software* 2021;6(60):3021.

248. TeamHG-Memex. *Eli5*. https://github.com/TeamHGMemex/eli5.

249. Millard LA, Davies NM, Gaunt TR, et al. Software Application Profile: PHESANT: a tool for performing automated phenome scans in UK Biobank. *International Journal of Epidemiology* 2017:47(1): 29–35.

250. Roman E, Smith A, Appleton S, et al. Myeloid malignancies in the real-world: Occurrence, progression and survival in the UK's population-based Haematological Malignancy Research Network 2004–15. *Cancer epidemiology* 2016;42:186-98.

251. Jankowska AM, Szpurka H, Tiu RV, et al. Loss of heterozygosity 4q24 and TET2 mutations associated with myelodysplastic/myeloproliferative neoplasms. *Blood* 2009;113(25):6403-10.

252. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26(6):841-42.

253. Loh P-R, Genovese G, Handsaker RE, et al. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* 2018;559(7714):350.

254. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;25(16):2078-79.

255. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* 2007;81(3):559-75.

256. D Turner S. qqman: an R package for visualizing GWAS results using QQ and manhattan plots. *The Journal of Open Source Software* 2018;3(25):731.

257. Machiela MJ, Chanock SJ. LDassoc: an online tool for interactively exploring genome-wide association study results and prioritizing variants for functional investigation. *Bioinformatics* 2018;34(5):887-89.

258. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 1995;57(1):289-300.

259. Hollander M, Wolfe DA, Chicken E. *Nonparametric statistical methods*: New York, NY : John Wiley & Sons; 2013.

260. Machiela MJ, Zhou W, Sampson JN, et al. Characterization of large structural genetic mosaicism in human autosomes. *The American Journal of Human Genetics* 2015;96(3):487-97.

261. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. *Nature* 2013;500(7463):415-21.

262. Acuna-Hidalgo R, Sengul H, Steehouwer M, et al. Ultra-sensitive sequencing identifies high prevalence of clonal hematopoiesis-associated mutations throughout adult life. *The American Journal of Human Genetics* 2017;101(1):50-64.

263. Jones AV, Chase A, Silver RT, et al. JAK2 haplotype is a major risk factor for the development of myeloproliferative neoplasms. *Nature Genetics* 2009;41(4):446-49.

264. Szustakowski JD, Balasubramanian S, Kvikstad E, et al. Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. *Nature Genetics* 2021;53(7):942-48.

265. Poplin R, Chang P-C, Alexander D, et al. A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology* 2018;36(10):983-87.

266. McDevitt M, Dunbar AJ, O'Keefe C, et al. SNP-a Karyotyping Provides a Clonal Molecular Marker and Is Associated with a High Incidence of Segmental Uniparental Disomy in Patients with CMML. *Blood* 2008;112(11):3657-57.

267. Chase A, Leung W, Tapper W, et al. Profound parental bias associated with chromosome 14 acquired uniparental disomy indicates targeting of an imprinted locus. *Leukemia* 2015;29(10):2069-74.

268. Kaneko H, Misawa S, Horiike S, et al. TP53 mutations emerge at early phase of myelodysplastic syndrome and are associated with complex chromosomal abnormalities. *Blood* 1995;85(8):2189-93.

269. Chase A, Pellagatti A, Singh S, et al. PRR14L mutations are associated with chromosome 22 acquired uniparental disomy, age-related clonal hematopoiesis and myeloid neoplasia. *Leukemia* 2019;33(5):1184-94.

270. Aziz A, Baxter EJ, Edwards C, et al. Cooperativity of imprinted genes inactivated by acquired chromosome 20q deletions. *J Clin Invest* 2013;123(5):2169-82.

271. Jia D, Jurkowska RZ, Zhang X, et al. Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. *Nature* 2007;449(7159):248-51.

272. Xu J, Wang Y-Y, Dai Y-J, et al. DNMT3A Arg882 mutation drives chronic myelomonocytic leukemia through disturbing gene expression/DNA methylation in hematopoietic cells. *Proceedings of the National Academy of Sciences* 2014;111(7):2620-25.

273. Russler-Germain David A, Spencer David H, Young Margaret A, et al. The R882H DNMT3A Mutation Associated with AML Dominantly Inhibits Wild-Type DNMT3A by Blocking Its Ability to Form Active Tetramers. *Cancer Cell* 2014;25(4):442-54.

274. Moran-Crusio K, Reavie L, Shih A, et al. Tet2 loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. *Cancer Cell* 2011;20(1):11-24.

275. Kim E, Ilagan JO, Liang Y, et al. SRSF2 mutations contribute to myelodysplasia by mutant-specific effects on exon recognition. *Cancer Cell* 2015;27(5):617-30.

276. Itzykson R, Kosmider O, Renneville A, et al. Prognostic score including gene mutations in chronic myelomonocytic leukemia. *J Clin Oncol* 2013;31(19):2428-36.

277. Ho YH, Méndez-Ferrer S. Microenvironmental contributions to hematopoietic stem cell aging. *Haematologica* 2020;105(1):38-46.

278. Oddsson A, Kristinsson S, Helgason H, et al. The germline sequence variant rs2736100_C in TERT associates with myeloproliferative neoplasms. *Leukemia* 2014;28(6):1371-74.

279. Hinds DA, Barnholt KE, Mesa RA, et al. Germ line variants predispose to both JAK2 V617F clonal hematopoiesis and myeloproliferative neoplasms. *Blood, The Journal of the American Society of Hematology* 2016;128(8):1121-28.

280. Gillis NK, Ball M, Zhang Q, et al. Clonal haemopoiesis and therapy-related myeloid malignancies in elderly patients: a proof-of-concept, case-control study. *The Lancet Oncology* 2017;18(1):112-21.

281. Coombs CC, Zehir A, Devlin SM, et al. Therapy-related clonal hematopoiesis in patients with non-hematologic cancers is common and associated with adverse clinical outcomes. *Cell Stem Cell* 2017;21(3):374-82. e4.

282. Huang Z, Sun S, Lee M, et al. Single-cell analysis of somatic mutations in human bronchial epithelial cells in relation to aging and smoking. *Nature Genetics* 2022;54(4):492-98.

283. Vivek Kumar PR, Zareena Hamza V, Mohankumar MN, et al. Studies on the HPRT mutant frequency in T lymphocytes from healthy Indian male population as a function of age and smoking. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 2004;556(1):107-16.

284. Kar SP, Quiros PM, Gu M, et al. Genome-wide analyses of 200,453 individuals yield new insights into the causes and consequences of clonal hematopoiesis. *Nature Genetics* 2022;54(8):1155-66.

285. Zhang Q, Zhao K, Shen Q, et al. Tet2 is required to resolve inflammation by recruiting Hdac2 to specifically repress IL-6. *Nature* 2015;525(7569):389-93.

286. Manolio TA, Weis BK, Cowie CC, et al. New models for large prospective studies: is there a better way? *Am J Epidemiol* 2012;175(9):859-66.

287. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *American Journal of Epidemiology* 2017;186(9):1026-34.

288. Bick AG, Pirruccello JP, Griffin GK, et al. Genetic Interleukin 6 Signaling Deficiency Attenuates Cardiovascular Risk in Clonal Hematopoiesis. *Circulation* 2020;141(2):124-31.

289. Kessler MD, Damask A, O'Keeffe S, et al. Exome sequencing of 628,388 individuals identifies common and rare variant associations with clonal hematopoiesis phenotypes. *MedRxiv* 2022:2021.12.29.21268342.

290. Liu N, Guo XH, Liu JP, et al. Role of telomerase in the tumour microenvironment. *Clinical and Experimental Pharmacology and Physiology* 2019:47(3):357-64.

291. Lee J, Taneja V, Vassallo R. Cigarette smoking and inflammation: cellular and molecular mechanisms. *J Dent Res* 2012;91(2):142-9.

292. Patel KV, Ferrucci L, Ershler WB, et al. Red Blood Cell Distribution Width and the Risk of Death in Middle-aged and Older Adults. *Archives of Internal Medicine* 2009;169(5):515-23.

293. Steensma DP, Bejar R, Jaiswal S, et al. Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood, The Journal of the American Society of Hematology* 2015;126(1):9-16.

294. Zekavat SM, Lin S-H, Bick AG, et al. Hematopoietic mosaic chromosomal alterations increase the risk for diverse types of infection. *Nature Medicine* 2021;27(6):1012-24.

295. Bick AG, Popadin K, Thorball CW, et al. Increased prevalence of clonal hematopoiesis of indeterminate potential amongst people living with HIV. *Scientific Reports* 2022;12(1):577.

296. Honigberg MC, Zekavat SM, Niroula A, et al. Premature Menopause, Clonal Hematopoiesis, and Coronary Artery Disease in Postmenopausal Women. *Circulation* 2021;143(5):410-23.

297. Jha V, Garcia-Garcia G, Iseki K, et al. Chronic kidney disease: global dimension and perspectives. *The Lancet* 2013;382(9888):260-72.

298. Zimmermann J, Herrlinger S, Pruy A, et al. Inflammation enhances cardiovascular risk and mortality in hemodialysis patients. *Kidney international* 1999;55(2):648-58.

299. Hamm LL, McCullough PA, Kasiske BL, et al. Kidney disease as a risk factor for development of cardiovascular disease. *Circulation* 2003;108:2154-69.

300. Group KDIGOCW. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney Int Suppl* 2013;3(1):1-150.

301. Israni A, Snyder J, Skeans M, et al. Predicting coronary heart disease after kidney transplantation: Patient Outcomes in Renal Transplantation (PORT) Study. *American Journal of Transplantation* 2010;10(2):338-53.

302. Gansevoort RT, Correa-Rotter R, Hemmelgarn BR, et al. Chronic kidney disease and cardiovascular risk: epidemiology, mechanisms, and prevention. *The Lancet* 2013;382(9889):339-52.

303. Kasiske BL, Guijarro C, Massy ZA, et al. Cardiovascular disease after renal transplantation. *Journal of the American Society of Nephrology* 1996;7(1):158-65.

304. Levey AS, Coresh J. Chronic kidney disease. *The Lancet* 2012;379(9811):165-80.

305. Oberg BP, McMenamin E, Lucas F, et al. Increased prevalence of oxidant stress and inflammation in patients with moderate to severe chronic kidney disease. *Kidney International* 2004;65(3):1009-16.

306. Busque L, Sun M, Buscarlet M, et al. High-sensitivity C-reactive protein is associated with clonal hematopoiesis of indeterminate potential. *Blood Advances* 2020;4(11):2430.

307. Hojs R, Ekart R, Bevc S, et al. Markers of inflammation and oxidative stress in the development and progression of renal disease in diabetic patients. *Nephron* 2016;133(3):159-62.

308. Mihai S, Codrici E, Popescu ID, et al. Inflammation-Related Mechanisms in Chronic Kidney Disease Prediction, Progression, and Outcome. *J Immunol Res* 2018;2018:2180373.

309. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 2011;27(21):2987-93.

310. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;581(7809):434-43.

311. Kircher M, Witten DM, Jain P, et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics* 2014;46(3):310-15.

312. Tate JG, Bamford S, Jubb HC, et al. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Research* 2019;47(D1):D941-D47.

313. Bouaoun L, Sonkin D, Ardin M, et al. TP53 variations in human cancers: new lessons from the IARC TP53 database and genomics data. *Human Mutation* 2016;37(9):865-76.

314. Hu L, Li Z, Cheng J, et al. Crystal structure of TET2-DNA complex: insight into TET-mediated 5mC oxidation. *Cell* 2013;155(7):1545-55.

315. Zhao Z, Chen S, Zhu X, et al. The catalytic activity of TET2 is essential for its myeloid malignancy-suppressive function in hematopoietic stem/progenitor cells. *Leukemia* 2016;30(8):1784-88.

316. Pattaro C, Riegler P, Stifter G, et al. Estimating the glomerular filtration rate in the general population using different equations: effects on classification and association. *Nephron Clinical Practice* 2013;123(1-2):102-11.

317. Levey AS, Stevens LA, Schmid CH, et al. A new equation to estimate glomerular filtration rate. *Annals of Internal Medicine* 2009;150(9):604-12.

318. Grubb A. Shrunken pore syndrome - a common kidney disorder with high mortality. Diagnosis, prevalence, pathophysiology and treatment options. *Clin Biochem* 2020;83:12-20.

319. Lees JS, Welsh CE, Celis-Morales CA, et al. Glomerular filtration rate by differing measures, albuminuria and prediction of cardiovascular disease, mortality and end-stage kidney disease. *Nature medicine* 2019;25(11):1753-60.

320. Skrivankova VW, Richmond RC, Woolf BAR, et al. Strengthening the Reporting of Observational Studies in Epidemiology Using Mendelian Randomization: The STROBE-MR Statement. *JAMA* 2021;326(16):1614-21.

321. Burgess S, Smith GD, Davies NM, et al. Guidelines for performing Mendelian randomization investigations. *Wellcome Open Research* 2019;4.

322. Croxford R. Restricted cubic spline regression: a brief introduction. *Toronto: Institute for Clinical Evaluative Sciences* 2016:1-5.

323. Harrell FE. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*: Springer New York, NY; 2001.

324. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 1997;30(7):1145-59.

325. Buscarlet M, Provost S, Zada YF, et al. DNMT3A and TET2 dominate clonal hematopoiesis and demonstrate benign phenotypes and different genetic predispositions. *Blood* 2017;130(6):753-62.

326. Grubb A. Cystatin C is Indispensable for Evaluation of Kidney Disease. *Ejifcc* 2017;28(4):268-76.

327. Nowak C, Ärnlöv J. Kidney Disease Biomarkers Improve Heart Failure Risk Prediction in the General Population. *Circulation: Heart Failure* 2020;13(8):e006904.

328. Zi M, Xu Y. Involvement of cystatin C in immunity and apoptosis. *Immunol Lett* 2018;196:80-90.

329. Weiner DE, Tighiouart H, Elsayed EF, et al. The relationship between nontraditional risk factors and outcomes in individuals with stage 3 to 4 CKD. *American Journal of Kidney Diseases* 2008;51(2):212-23.

330. Christensen AS, Møller JB, Hasselbalch HC. Chronic kidney disease in patients with the Philadelphia-negative chronic myeloproliferative neoplasms. *Leukemia research* 2014;38(4):490-95.

331. Koschmieder S, Chatain N. Role of inflammation in the biology of myeloproliferative neoplasms. *Blood Rev* 2020;42:100711.

332. Sinnott-Armstrong N, Tanigawa Y, Amar D, et al. Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nature Genetics* 2021:1-10.

333. Kourou K, Exarchos TP, Exarchos KP, et al. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J* 2015;13:8-17.

334. Wang P, Li Y, Reddy CK. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)* 2019;51(6):1-36.

335. Zhang Z. Propensity score method: a non-parametric technique to reduce model dependence. *Ann Transl Med* 2017;5(1):7.

336. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology* 2013;31(3):213-19.

337. Sondka Z, Bamford S, Cole CG, et al. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer* 2018;18(11):696-705.

338. Simon N, Friedman J, Hastie T, et al. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J Stat Softw* 2011;39(5):1-13.

339. Harrell FE, Jr, Califf RM, Pryor DB, et al. Evaluating the Yield of Medical Tests. *JAMA* 1982;247(18):2543-46.

340. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol* 2010;5(9):1315-6.

341. Sharma D, Gotlieb N, Farkouh ME, et al. Machine Learning Approach to Classify Cardiovascular Disease in Patients With Nonalcoholic Fatty Liver Disease in the UK Biobank Cohort. *Journal of the American Heart Association* 2022;11(1):e022576.

342. Swanson JM. The UK Biobank and selection bias. *Lancet (London, England)* 2012;380(9837):110.

343. Collins R. What makes UK Biobank special? *The Lancet* 2012;379(9822):1173-74.

344. Kralovics R, Passamonti F, Buser AS, et al. A gain-of-function mutation of JAK2 in myeloproliferative disorders. *N Engl J Med* 2005;352(17):1779-90.

345. Chen T, Guestrin C. XGBoost: A scalable tree boosting system In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,(pp. 785–794). *New York, NY, USA: ACM* 2016;10(2939672.2939785).

346. Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems* 2014;41(3):647-65.

347. Wright DJ, Day FR, Kerrison ND, et al. Genetic variants associated with mosaic Y chromosome loss highlight cell cycle genes and overlap with cancer susceptibility. *Nature Genetics* 2017;49(5):674-79.

348. Lin SH, Loftfield E, Sampson JN, et al. Mosaic chromosome Y loss is associated with alterations in blood cell counts in UK Biobank men. *Sci Rep* 2020;10(1):3655.

349. Silver AJ, Bick AG, Savona MR. Germline risk of clonal haematopoiesis. *Nature Reviews Genetics* 2021;22(9):603-17.

350. Noveski P, Madjunkova S, Sukarova Stefanovska E, et al. Loss of Y Chromosome in Peripheral Blood of Colorectal and Prostate Cancer Patients. *PLoS One* 2016;11(1):e0146264.

351. Machiela MJ, Dagnall CL, Pathak A, et al. Mosaic chromosome Y loss and testicular germ cell tumor risk. *Journal of Human Genetics* 2017;62(6):637-40.

352. Baliakas P, Forsberg LA. Chromosome Y loss and drivers of clonal hematopoiesis in myelodysplastic syndrome. *Haematologica* 2021;106(2):329-31.

353. Pérez-Jurado LA, Cáceres A, Esko T, et al. Clonal chromosomal mosaicism and loss of chromosome Y in men are risk factors for SARS-CoV-2 vulnerability in the elderly. *MedRxiv* 2022:2020.04.19.20071357.

354. Haitjema S, Kofink D, van Setten J, et al. Loss of Y Chromosome in Blood Is Associated With Major Cardiovascular Events During Follow-Up in Men After Carotid Endarterectomy. *Circ Cardiovasc Genet* 2017;10(4):e001544.

355. Terao C, Suzuki A, Momozawa Y, et al. Chromosomal alterations among age-related haematopoietic clones in Japan. *Nature* 2020;584(7819):130-35.

356. Ouseph MM, Hasserjian RP, Dal Cin P, et al. Genomic alterations in patients with somatic loss of the Y chromosome as the sole cytogenetic finding in bone marrow cells. *Haematologica* 2021;106(2):555-64.

357. Ljungström V, Mattisson J, Halvardson J, et al. Loss of Y and clonal hematopoiesis in blood—two sides of the same coin? *Leukemia* 2022;36(3):889-91.

358. Dumanski JP, Halvardson J, Davies H, et al. Immune cells lacking Y chromosome show dysregulation of autosomal gene expression. *Cell Mol Life Sci* 2021;78(8):4019-33.

359. Miller PG, Qiao D, Rojas-Quintero J, et al. Association of clonal hematopoiesis with chronic obstructive pulmonary disease. *Blood* 2022;139(3):357-68.

360. Jaiswal S. Clonal hematopoiesis and nonhematologic disorders. *Blood* 2020;136(14):1606-14.

361. Niroula A, Sekar A, Murakami MA, et al. Distinction of lymphoid and myeloid clonal hematopoiesis. *Nature Medicine* 2021;27(11):1921-27.

362. Loh P-R, Genovese G, McCarroll SA. Monogenic and polygenic inheritance become instruments for clonal selection. *Nature* 2020;584(7819):136-41.

363. Kessler MD, Damask A, O'Keeffe S, et al. Exome sequencing of 628,388 individuals identifies common and rare variant associations with clonal hematopoiesis phenotypes. *MedRxiv* 2022:2021.12.29.21268342.

364. Gopakumar J, Weinstock J, Burugula BB, et al. Clonal Hematopoiesis Is Driven By Aberrant Activation of TCL1A. *Blood* 2021;138(Supplement 1):597-97.

365. Zhao Y, Stankovic S, Koprulu M, et al. GIGYF1 loss of function is associated with clonal mosaicism and adverse metabolic health. *Nature Communications* 2021;12(1):4178.

366. Gray A, Feldman HA, McKinlay JB, et al. Age, Disease, and Changing Sex Hormone Levels in Middle-Aged Men: Results of the Massachusetts Male Aging Study*. *The Journal of Clinical Endocrinology & Metabolism* 1991;73(5):1016-25.

367. Dunn JF, Nisula BC, Rodbard D. Transport of Steroid Hormones: Binding of 21 Endogenous Steroids to Both Testosterone-Binding Globulin and Corticosteroid-Binding Globulin in Human Plasma. *The Journal of Clinical Endocrinology & Metabolism* 1981;53(1):58-68.

368. Wheeler MJ. The Determination of Bio-Available Testosterone. *Annals of Clinical Biochemistry* 1995;32(4):345-57.

369. Vermeulen A, StoÏCa T, Verdonck L. The Apparent Free Testosterone Concentration, An Index of Androgenicity. *The Journal of Clinical Endocrinology & Metabolism* 1971;33(5):759-67.

370. Manni A, Pardridge WM, Cefalu W, et al. Bioavailability of albumin-bound testosterone. *J Clin Endocrinol Metab* 1985;61(4):705-10.

371. Vermeulen A, Verdonck L, Kaufman JM. A Critical Evaluation of Simple Methods for the Estimation of Free Testosterone in Serum. *The Journal of Clinical Endocrinology & Metabolism* 1999;84(10):3666-72.

372. Coviello AD, Haring R, Wellons M, et al. A genome-wide association meta-analysis of circulating sex hormone-binding globulin reveals multiple Loci implicated in sex steroid hormone regulation. *PLoS Genet* 2012;8(7):e1002805.

373. Võsa U, Claringbould A, Westra H-J, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature Genetics* 2021;53(9):1300-10.

374. Long JA. Interactions: Comprehensive, user-friendly toolkit for probing interactions. R package version 1.1. 0. 2019. https://cran.r-project.org/package=interactions.

375. Emadi-Konjin P, Bain J, Bromberg IL. Evaluation of an algorithm for calculation of serum “Bioavailable” testosterone (BAT). *Clinical Biochemistry* 2003;36(8):591-96.

376. Neale R. UK Biobank GWAS round 2 [http://www.nealelab.is/uk-biobank/]. 01/08/2018 ed. Boston, MA: Neale Lab, 2018.

377. Mendel CM. The Free Hormone Hypothesis: A Physiologically Based Mathematical Model*. *Endocrine Reviews* 1989;10(3):232-74.

378. Hammes A, Andreassen TK, Spoelgen R, et al. Role of Endocytosis in Cellular Uptake of Sex Steroids. *Cell* 2005;122(5):751-62.

379. Balogh A, Karpati E, Schneider AE, et al. Sex hormone-binding globulin provides a novel entry pathway for estradiol and influences subsequent signaling in lymphocytes via membrane receptor. *Scientific Reports* 2019;9(1):4.

380. Rocha STd, Edwards CA, Ito M, et al. Genomic imprinting at the mammalian Dlk1-Dio3 domain. *Trends in Genetics* 2008;24(6):306-16.

381. Zink F, Magnusdottir DN, Magnusson OT, et al. Insights into imprinting from parent-of-origin phased methylomes and transcriptomes. *Nature Genetics* 2018;50(11):1542-52.

382. Mitchell E, Spencer Chapman M, Williams N, et al. Clonal dynamics of haematopoiesis across the human lifespan. *Nature* 2022;606(7913):343-50.

383. Fabre MA, de Almeida JG, Fiorillo E, et al. The longitudinal dynamics and natural history of clonal haematopoiesis. *Nature* 2022;606(7913):335-42.

384. Levin MG, Nakao T, Zekavat SM, et al. Genetics of smoking and risk of clonal hematopoiesis. *Scientific Reports* 2022;12(1):7248.

385. Liu M, Jiang Y, Wedow R, et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nature Genetics* 2019;51(2):237-44.

386. Ramanathan G, Johnson R, Chen JH, et al. Cigarette Smoke and E-Cigarette Aerosols Lead to Clonal Expansion of Tet2 -/- and Dnmt3a R878H Cells In Vivo. *Blood* 2021;138:2167.

387. Thompson DJ, Genovese G, Halvardson J, et al. Genetic predisposition to mosaic Y chromosome loss in blood. *Nature* 2019;575(7784):652-57.

388. Tangri N, Stevens LA, Griffith J, et al. A Predictive Model for Progression of Chronic Kidney Disease to Kidney Failure. *JAMA* 2011;305(15):1553-59.

389. Vlasschaert C, McNaughton AJM, Chong M, et al. Association of Clonal Hematopoiesis of Indeterminate Potential with Worse Kidney Function and Anemia in Two Cohorts of Patients with Advanced Chronic Kidney Disease. *Journal of the American Society of Nephrology* 2022;33(5):985-95.

390. Fuster JJ, MacLauchlan S, Zuriaga MA, et al. Clonal hematopoiesis associated with TET2 deficiency accelerates atherosclerosis development in mice. *Science* 2017;355(6327):842-47.

391. Svensson EC, Madar A, Campbell CD, et al. TET2-driven clonal hematopoiesis predicts enhanced response to canakinumab in the CANTOS trial: an exploratory analysis. *Circulation* 2018;138(Suppl_1):A15111-A11.

392. Ridker PM, Devalaraja M, Baeres FMM, et al. IL-6 inhibition with ziltivekimab in patients at high atherosclerotic risk (RESCUE): a double-blind, randomised, placebo-controlled, phase 2 trial. *Lancet (London, England)* 2021;397(10289):2060-69.

393. Perry JRB, Day F, Elks CE, et al. Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* 2014;514(7520):92-97.

394. Hofmeister RJ, Rubinacci S, Ribeiro DM, et al. Parent-of-origin effects in the UK Biobank. *BioRxiv* 2021:2021.11.03.467079.

395. Gozdecka M, Meduri E, Mazan M, et al. UTX-mediated enhancer and chromatin remodeling suppresses myeloid leukemogenesis through noncatalytic inverse regulation of ETS and GATA programs. *Nature Genetics* 2018;50(6):883-94.

396. Williams N, Lee J, Mitchell E, et al. Life histories of myeloproliferative neoplasms inferred from phylogenies. *Nature* 2022;602(7895):162-68.

397. Chiesa ST, Charakida M, Georgiopoulos G, et al. Glycoprotein Acetyls: A Novel Inflammatory Biomarker of Early Cardiovascular Risk in the Young. *Journal of the American Heart Association* 2022;11(4):e024380.

398. Uddin MM, Zhou Y, Bick AG, et al. Cost effective sequencing enables longitudinal profiling of clonal hematopoiesis. *MedRxiv* 2022:2022.01.31.22270028.