Clonal myelopoiesis in the UK Biobank cohort: *ASXL1* mutations are strongly associated with smoking

Ahmed A.Z. Dawoud[1], William J Tapper[1] & Nicholas C.P. Cross[1,2]

[1] Faculty of Medicine, University of Southampton, Southampton, UK
[2] Wessex Regional Genetics Laboratory, Salisbury NHS Foundation Trust, Salisbury, UK

Correspondence to:

Professor N.C.P. Cross
Wessex Regional Genetics Laboratory
Salisbury NHS Foundation Trust
Salisbury SP2 8BJ, UK

Tel: +(44) 1722 429080
Fax: +(44) 1722 331531
email: ncpc@soton.ac.uk

Abstract

We sought to determine the significance of myeloid clonal hematopoiesis (CH) in the UK Biobank cohort (n=502,524, median age=58 years). Utilizing SNP array (n=486,941) and whole exome sequencing data (n=49,956), we identified 1166 participants with myeloid CH, defined by myeloid-associated mosaic chromosome abnormalities (mCA) and/or likely somatic driver mutations in *DNMT3A*, *TET2*, *ASXL1*, *JAK2*, *SRSF2* or *PPM1D*. Myeloid CH increased by 1.1-fold per annum (myeloid mCA, $P=1.57\times10^{-38}$; driver mutations, $P=5.89\times10^{-47}$). Genome-wide association analysis identified two distinct signals within *TERT* that predisposed to myeloid CH, plus a weaker signal corresponding to the *JAK2* 46/1 haplotype. Specific subtypes of myeloid CH were associated with several blood features and clinical phenotypes, including *TET2* mutations and chronic obstructive pulmonary disease. Smoking history was significantly associated with myeloid CH: 53% of myeloid CH cases were smokers compared to 44% of controls ($P=3.38\times10^{-6}$), a difference principally due to current (OR=1.10; $P=6.14\times10^{-6}$) rather than past smoking (P=0.08). Breakdown of CH by specific mutation type revealed that *ASXL1* loss of function mutations were most strongly associated with current smoking status (OR=1.07; $P=1.92\times10^{-5}$), and the only abnormality associated with past smoking (OR=1.04; P=0.0026). We suggest that the inflammatory environment induced by smoking may promote the outgrowth of *ASXL1*-mutant clones.

Running title: *ASXL1* mutations are associated with smoking history

Keywords: *ASXL1*, smoking, UK Biobank, clonal hematopoiesis

Introduction

Age-related clonal hematopoiesis (CH) is a widespread phenomenon that predisposes to the development of hematological malignancies as well as some non-hematological conditions. First suggested more than 20 years ago by skewed X-chromosome inactivation patterns in myeloid cells of elderly females,[1] clonality in the absence of a hematological neoplasm was proven by the finding of acquired *TET2* driver mutations in a subset of skewed cases.[2] CH as a more general phenomenon was defined by several lines of evidence, including (i) the prevalence of *JAK2* V617F was much greater than myeloproliferative neoplasms (MPN) in randomly selected subjects undergoing hospital-based clinical investigations;[3] (ii) mosaic chromosome alterations (mCA), indicative of myeloid or lymphoid clonality, were seen in large cohorts ascertained for non-hematological conditions[4, 5] and (iii) myeloid malignancy-associated driver mutations, most commonly involving *DNMT3A*, *TET2* or *ASXL1*, were found at a much higher frequency than expected in similar large, non-hematological cohorts.[6-8]

The studies above determined that CH is strongly age dependent, with >10% of individuals aged >65-70 years of age being affected, but only 1% of those <50 years old. Subsequent analysis using more sensitive mutation detection methods indicated that driver mutation-associated clonal expansion is more common[9] than previously thought and indeed nearly ubiquitous when the limit of detection is pushed to as low as 0.03% variant allele frequency (VAF).[10] The picture is further complicated by the fact that whole exome sequencing (WES) and whole genome sequencing (WGS) studies have found that CH often predominates in the absence of discernible driver mutations, potentially as a consequence of stochastic drift acting on a small population of active hematopoietic stem cells.[6-8]

Although individuals with CH have a 10-fold excess risk of developing a hematological neoplasm, the actual rate of progression is only 1-2% per annum, and thus the term CH of indeterminate potential (CHIP) is widely used.[11] However CH is also associated with all-cause mortality, and specifically an increased risk of atherosclerotic cardiovascular disease after

adjusting for known risk factors.[12] CH has also been associated with other disorders, for example chronic obstructive pulmonary disorders.[13]

UK Biobank is an ongoing prospective cohort study that aims to provide new insights into the causes of common and uncommon diseases. Recruitment of approximately 500,000 community-dwelling, generally healthy subjects aged 40-69 years took place between 2006 and 2010 across the UK. Participants provided comprehensive demographic, psychosocial and

medical information during an initial visit to a UK Biobank assessment centre along with baseline blood and urine samples for genomic, biochemical and other laboratory assessments. Long term follow-up is provided via linked medical records.[14] In this study we aimed to assess the prevalence, impact and causes of myeloid CH, defined by the presence of myeloid mCA and/or myeloid malignancy-associated mutations, in the UK Biobank cohort.

Methods

Cohort structure. UK Biobank participants were split into four phenotypic groups: myeloid disorders, lymphoid disorders, other cancers, and cancer free based on the International Classification of Disease codes (ICD version 10) recorded by the national cancer registry (Data-Field 40006) and reason for admission to hospital (Data-Fields 41202, 41204, and 41270; Supplementary Table 1). In subsequent analyses, these four groups are collectively referred to as the phenotypic groups. The ICD-10 coding system was introduced into the national cancer registry in 1995 and to the hospital admissions in 1996. Consequently, any relevant events that occurred before 1995/6 were not captured. The cancer registry data was accessed as of July 2018 (most recent record December 2016); 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 (median age at recruitment = 58 years). The four phenotypic groups were defined by events that occurred at any time during the study period (1995 to 2018). All participants provided informed consent according to the Declaration of Helsinki, and UK Biobank received ethical approval from the North West multi-centre Research Ethics Committee (REC reference 11/NW/0382). The current study was conducted under approved UK Biobank application number 35273.

SNP genotype data. Genotypic data for 488,377 samples were obtained from UK Biobank. These samples were genotyped at Affymetrix laboratories (Santa Clara, USA) in batches of approximately 4,700 samples using two similar microarrays, the UK BiLEVE array (807,411 markers) (n=49,950 samples) and the remainder by UK Biobank axiom array (825,927 markers). The two microarrays share 95% of SNPs and are thus highly comparable. Routine quality control (QC) was performed by Affymetrix, which excluded poorly genotyped SNPs, and multi-allelic SNPs (n= 35,014 across the two platform). Further QC was performed by UK Biobank (Supplementary Methods) that restricted the released data to 805,426 markers, and 486,941 samples with no quality flags.[15] Of the 488,377 participants with available SNP data, we excluded a further 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).

Calling mosaic chromosomal alterations from SNP array data. Raw input files were generated for each sample which contained the 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. Regions of allelic imbalance (AI) were detected using BAF segmentation to analyse the raw input files using the default parameters for Affymetrix array data.[16] To select likely somatic events and exclude potential false positives, a custom script (see Supplementary Methods) was used to process the resulting AI regions as follows. First, bedtools was used to merge AI regions with a minimum density of 1 SNP per 20Kb that were separated by less than 2Mb. AI regions were then scored based on the product of SNP density, heterozygosity rate and coverage (see Supplementary Table 2). Finally, AI regions greater than 2Mb in total length with confidence scores above an empirically defined threshold (≥9) were defined as mCA. These were further broken down into copy number loss (CNL), copy number gain (CNG) or acquired uniparental disomy (aUPD) based on the regions median LRR value. The parameters used have been estimated to identify clonal fractions larger than 0.1, 0.2 and 0.27 for aUPD, CNL and CNG, respectively.[16] Full details are given in the Supplementary Methods. Since mCA may be derived from myeloid or lymphoid cells, we correlated mCA with clinical phenotype to specifically define myeloid mCA (see Results).

Whole exome sequencing data. Whole exome sequencing (WES) was performed by Regeneron Genetics. In brief, read sequences were aligned to the reference human genome (version GRCH38) using BWA-MEM, duplicate reads were marked using Picard tools, and patient level genomic variant call files (gVCF) were generated using the WeCall variant caller with ≥2 alternative reads to call a variant as described.[17] Of the 49,996 subjects that were

157 successfully sequenced, 40 samples flagged by Regeneron Genetics for QC issues were
158 excluded from the analysis.[17]
159
160 *Identification of candidate somatic driver variants from WES data.* The gVCF files from UK
161 Biobank were converted to VCF format and filtered to remove variants with low read depth
162 (DP; <7 for SNVs, <10 for indels). SAMtools/Bcftools[18] was used to merge the separate files
163 into one multi-sample VCF, split multi-allelic positions into separate variants and to
164 normalise the location of indels using their left most position.[18] The combined VCF file was
165 annotated using Annovar and the RefSeq gene database[19]. Putative somatic mutations
166 (regardless of VAF) were identified in six genes known to be associated with myeloid
167 neoplasia that were exonic, had an alternate allele frequency ≤1% in public databases of
168 common variation (1000 Genome, ExAC, ESP6500, gnomAD) and were either loss of function
169 (LOF) mutations (*TET2*, *DNMT3A*, *ASXL1*, *PPM1D*) or known oncogenic hotspots (*DNMT3A*
170 R882, *JAK2* V617F, *SRSF2* P95; see Supplementary Methods).[7]
171
172 *Statistical analyses:*
173 *Association of mCA categories and somatic driver mutations with phenotypic group*. The
174 frequency of mCA events, each of its subcategories (aUPD, CNG or CNL) and somatic driver
175 mutations were tested for association with either myeloid, lymphoid, or other cancers
176 compared to cancer free controls using Fisher's exact tests in SPSS (Version 25). The average
177 number of mCA events per sample in either myeloid, lymphoid, or other cancers were
178 compared against cancer free controls using Mann-Whitney U tests[20].
179
180 *Association of specific mCAs with hematological phenotype*. Autosomal mCAs were stratified
181 by type (aUPD, CNL, CNG), chromosome arm (*p* or *q*), and position (telomeric or interstitial).
182 Each type with at least one observation was tested for association with a hematological
183 phenotype (either myeloid, or lymphoid) in comparison to cancer free controls using
184 Fisher's exact tests in SPSS (version 25). A total of 416 specific mCAs were tested after
185 selecting mCA types with at least one observation. Interstitial events that were associated
186 with a myeloid phenotype and not previously recognised as a recurrent abnormality were

manually reviewed (see Supplementary Methods). 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).

The association of common variants with clonality. Samples with SNP array data were split into cases and controls. Cases were defined by the presence of one or more feature associated with myeloid CH; either a myeloid mCA or at least one putative somatic mutation in six driver genes (*JAK2 V617F*, *SRSF2* P95, *DNMT3A* R882 or any frameshift/stopgain mutation 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 hematological malignancy during the study period. A total of 265,112 SNPs with minor allele frequencies (MAF) greater than 10% and without deviation from Hardy-Weinberg equilibrium (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.[21] These results were visualised using the qqman, qqnorm and qqplot plot procedures in R to generate a Manhattan plot and quantile-quantile plot. In regions with multiple SNPs reaching genome-wide significance, stepwise logistic regression was used to determine the number of independent signals. All SNPs with $P<10^{-8}$ and within 500kb of the index SNP were added to the regression model in order of significance.

The association of clonal hematopoiesis with smoking (past, current, any), clinical phenotype blood traits and biochemistry. We used the PHEnome Scan ANalysis Tool (PHESANT)[22] to test the effect of myeloid CH, which is defined by myeloid mCAs and/or somatic mutations

associated with myeloid diseases, on selected phenotypes from the UK Biobank. Phenotypes were tested using either ordinal (current and previous smoking), multinomial (combined smoking status), logistic (clinical phenotypes, n=395), or linear regression (blood features, n=29 or biochemical markers, n=30). All regressions included covariates for age and sex with the addition of smoking for the analysis of clinical phenotypes and blood features. Where appropriate, inverse normal transformation was applied to counteract departures from normality. Clinical phenotypes were determined using the primary/main diagnoses from hospital inpatient records (Data-field 41202) and limited to phenotypes identified in ≥0.1% of the UK Biobank participants.

Survival analyses: To test the association between myeloid CH and either all-cause mortality, myocardial infarction (MI) or stroke, we used the survival package[23] in R to perform Cox regression analyses with correction for age, sex and smoking status. Follow-up times were calculated using the "lubridate"[24] 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.

Where relevant, P values were corrected for multiple testing using the false discovery rate (FDR) and denoted P_{FDR}. Detailed methods plus associated figures and tables are provided in the Supplementary Methods.

Results

The study cohort. The phenotypic breakdown of the UK Biobank cohort is summarised in Table 1, 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).[25]

Identification of mosaic chromosomal alterations. We analysed genome wide SNP array data to identify autosomal regions of AI in all participants for whom the array data passed QC (n=486,941), and X imbalances for female participants (n=264,083). In the absence of matched constitutional DNA to definitively identify somatic events, blood cell clonality was inferred by using an upper mBAF threshold to exclude events that were likely to be constitutional. In total, our method identified 8,203 mCA >2Mb in size in 5,040 participants (1% of 486,941 analysed samples; Supplementary Table 2) which broke down into aUPD (n=4224), CNG (n=659) or CNL (n=3320). The frequency of these events (mCA, aUPD, CNG and CNL) were all significantly higher in the myeloid and lymphoid disease groups versus cancer free controls while only CNL were more frequent in other types of cancer versus controls (Table 2). The incidence of mCAs was highest in the myeloid disorders group with 11% of samples affected (210/1913). Of these, more than 75% involved aUPD (158/210) which corresponds to a 16-fold enrichment versus cancer free controls and the most significantly associated mCA subcategory with disease (OR=16.39; P_{FDR}=8.78x10^{-124}; Fisher's exact test). 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 2).

The frequency of mCA increased with age and ranged between 0.85% at 40-45 years to 1.29% at 66-70 years over all UK Biobank participants with array data passing QC (n=486,941) (Figure 1a). Using logistic regression, this corresponded to an annual increased risk of acquiring a mCA by 1.02 fold after adjusting for gender (OR=1.02, P=1.80x10^{-19}). The

risk of acquiring each subcategory of mCA also increased with age; aUPD (OR = 1.02, P=3.14x10^{-14}), CNG (OR = 1.04, P=1.09x10^{-6}) and CNL (OR = 1.01, P=3.66x10^{-4}).

Association of specific mCA with hematological phenotype. As expected, distinct mCAs were associated with myeloid and lymphoid disorders. After excluding interstitial events that failed manual review, a total of 17 mCAs involving 15 chromosomal arms were found to be associated with myeloid disorders using Fisher's exact tests (Table 3). Hereafter, these 17 abnormalities are collectively referred to as myeloid mCAs. Of these, 9p aUPD was the most significant individual abnormality associated with myeloid disorders (OR=2858, P_{FDR}=6.28x10^{-191}, Table 3). The risk of acquiring at least one of the myeloid mCAs was more strongly associated with age (OR=1.10, P=1.57x10^{-38}) than mCAs involving all chromosomes (Figure 1b). Interstitial CNL of chromosome 13 was the most significant abnormality associated with lymphoid disorders (OR=23.16; P_{FDR}=1.24x10^{-63}, Supplementary Table 3).

Clonality defined by somatic mutations in individual genes. We restricted our analysis of the WES data to a set of genes/mutations that collectively account for approximately 95% of myeloid CH events in 6 genes (*DNMT3A, TET2, ASXL1, JAK2, SRSF2,* and *PPM1D*)[6-8, 26]. As summarised in Table 4, we identified 721 candidate driver mutations in 678 subjects, with *DNMT3A* being the most commonly affected gene. Only 37 cases had more than one variant, with frequency of myeloid disorders being higher (11.8%) in participants with more than one variant compared to those 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. A detailed list of variants in the 6 genes are given in Supplemental Table 4.

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 less difference in the prevalence of myeloid mutations in participants who had non-hematological cancers versus cancer-free controls (1.4% vs 1.2%; P_{FDR} =0.04; OR=1.2).

The frequency of CH defined by putative somatic mutations also increased with age and ranged between 0.4% at age <45 years to 2.8% at age >65 years (Figure 1e). The risk of acquiring any one of these mutations increased by 1.1 fold per year, which is the same as the effect determined for myeloid mCA (OR=1.1, P=5.89x10^{-47}). The six genes were also significant when tested separately (Supplementary Table 5). The risk of acquiring a somatic mutation in *SRSF2* had the largest annual increase with age at 1.2 fold (P=3.66x10^{-4}) although these mutations were the rarest (n=20) and absent in participants younger than 50 years old (Figure 1e).

The association of CH with constitutional genetic variation. A previous study has associated germline variation at the *TERT* locus (rs34002450) with CH in the Icelandic population.[7] To examine the influence of genetic variation on myeloid CH in the UK Biobank cohort, we performed a genome wide association study (GWAS) to assess the influence of common polymorphisms in the 1,166 participants with myeloid CH defined by the presence of (i) any myeloid malignancy-associated mCA (Table 3) and/or (ii) somatic mutations in the 6 genes of interest. A total of 265,112 SNPs were tested for association with myeloid CH using allelic chi-square tests to compare these cases with 30,892 controls that were free of any clonal marker, or any hematological malignancy (Figure 2a,b). Three SNPs with genome-wide significance were identified in the *TERT* gene. 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 2c; Supplementary Table 6). 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. A second signal

was seen just below the level of genome wide significance (Figure 2b) and included rs3780381, rs17425819 and rs10974944. These SNPs are within *JAK2* and are in linkage disequilibrium (LD) with the 46/1 haplotype, previously shown to be strongly associated with acquisition of *JAK2* V617F.[27] This association signal disappeared when cases with *JAK2 V617F* (n=40) and mCA including *JAK2* (n=115) were removed from the analysis.

The association between clonal hematopoiesis and smoking. To examine the relationship between smoking and CH, we used PHESANT to perform regression analyses of past, current and combined smoking status in 32,058 participants consisting of 1,166 with myeloid CH and 30,892 without myeloid CH and without any hematological malignancy. The odds of having ever smoked (combined status) were significantly higher in participants with myeloid CH (53% smokers) than those without myeloid CH (44% smokers; P_{FDR}=3.38x10^{-6}, Table 5, Supplementary Table 7). This effect was associated with current smoking status (OR=1.10, P_{FDR}=6.14x10^{-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* are strongly associated with current smoking status (OR=1.07, P_{FDR}=1.92x10^{-5}) and the only abnormality associated with past smoking status (OR=1.04, P_{FDR}=0.0026). Indeed, 69% of participants with *ASXL1* mutations were past or current smokers. Participants with myeloid CH without *ASXL1* mutations (n=1066) remained significantly associated with current smoking but the effect was weaker (OR=1.07, P_{FDR}=0.0008). 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 5; Supplementary Table 7; all comparisons were corrected for gender and age).

The association between myeloid CH and clinical phenotype, blood traits and biochemical measures

We sought to determine if myeloid CH in hematological malignancy-free participants (n=911) was associated with 29 blood features, 30 biochemical markers and 395 non-malignant clinical phenotypes. Myeloid CH was associated with 9 blood features including red cell distribution width (RDW), estimated to be 1.02 fold higher in myeloid CH cases (P_{FDR}=9.7x10^{-4}; linear regression; Table 6), in line with previous findings.[8] In addition, myeloid CH was associated with a significant increase in all platelet indices, and particularly platelet distribution width (PDW; OR=1.03, P_{FDR}=6x10^{-6}) and a decrease in basophil indices, most notably basophil counts (OR=0.95, P_{FDR}=5.9x10^{-4}, Table 6). Myeloid CH cases also showed an anemia-like blood profile with low hemoglobin and a decrease in cholesterol, high density lipoprotein and creatinine (Table 7). Clinically, a significant association was seen with urinary tract disorders defined by ICD-10 codes N35.9 (urethral stricture) and N32.0 (bladder constriction) (Table 8).

Looking at individual drivers of myeloid CH, several associations emerge. The strongest multifactorial findings were for *JAK2 V617F*, which was associated with increased red cell and platelet parameters (Table 6). *ASXL1* mutations were associated with increased RDW and PDW, an anemia-like profile and reduced IGF-1. *TET2* mutations were associated with reduced eosinophils, increased monocytes and, clinically, ulcer of the lower limb, agranulocytosis and chronic obstructive pulmonary disease. Participants with *SRSF2* P95 showed an increase in the percentage of reticulocytes and a decrease in HDL cholesterol, and mCA was associated with multiple parameters, particularly low basophils, increased PDW, several decreased biochemical measures as well as urinary tract disorders. Finally, *DNMT3A* and *PPM1D* mutations did not show any associations except for a relatively minor increase in platelet and monocyte counts, respectively. The complete results are shown in Supplementary Tables 8, 9 and 10.

We noted that the incidence of all-cause mortality since study entry was higher in participants with myeloid CH without hematological malignancy (42/911; 4.6%) compared with controls (674/30892; 2.2%). 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. There was no association between CH and subsequent MI or stroke in the absence of a hematological malignancy (MI: myeloid CH, n=873 vs controls, n=30271; HR=1.16; *P*=0.53; stroke: myeloid CH, n=890 vs controls, n=30472; HR=1.18; *P*=0.58).

DISCUSSION

The prevalence and significance of CH has been reported in several cohorts, but our study has several distinctive features. UK Biobank is a very large population-based cohort that includes an extensive repertoire of baseline phenotypic data as well as >9 year prospective collection of clinical follow up information. Genome wide SNP data is available for the great majority of participants (n=486,941), and WES data for a subset (n=49,956), all derived from a single baseline peripheral sample taken at study entry. Thus, we were able to assess CH associated with both mCA and somatic mutations, albeit with a modest limit of detection compared to some published studies. We 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.

In total we identified 1,166 subjects with myeloid CH. Of these, 678 had mutations in one or more genes (1.4% of cases who underwent WES), and 506 had myeloid mCA (0.1% of subjects who had a SNP array; 18 subjects had both mCA and somatic mutations). Given the median age of the UK Biobank cohort (58 years), the limited sensitivity afforded by WES and the fact we focused only on myeloid mCA, the prevalence of CH is broadly comparable to published reports.[6, 8]

Like other studies, we found that *DNMT3A* was the most commonly mutated gene and that there was a strong relationship between both mCA and mutations with age. Clinically, we confirmed associations between CH and all-cause mortality and a specific association between *TET2* mutations and chronic obstructive pulmonary disorders with acute lower

respiratory infection.[13] In common with some other studies,[7, 13] we did not find an association between CH and MI or stroke. The reason that the association between CH and cardiovascular disease is very prominent in some studies[12] but not others is presumably explained by differences in cohort structure, follow up time, and definitions of CH. UK Biobank had an upper recruitment age of 69 years and the follow up is only 9.1 years. Our analysis is estimated to have 86% power (Supplementary Figure 5) to detect an association between CH and MI based on an HR of 1.9, as previously reported,[12] and an overall event rate of 1.4% (439/31144). Our 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 UK Biobank cases are sequenced and the median follow-up is extended, more associations are likely to emerge.

We also confirmed an association with CH and smoking[7] and showed for the first time that this effect is predominantly, but not exclusively, associated with *ASXL1* mutations. Whilst this association, and other novel findings in our study, needs to be confirmed in an independent population-based cohort, we note that *ASXL1* mutations have recently been associated with smoking in a large cohort of post-therapy cancer patients.[28] CH has previously been associated chemotherapy and radiotherapy[29, 30] which, along with the association with ageing, suggest that a link between stress hematopoiesis and the development of clonality. Although it is conceivable that smoking preferentially induces *ASXL1* mutations, it seems more likely that the chronic inflammation induced by smoking creates an environment within which *ASXL1* mutant clones have a selective advantage. A previous study noted an increase in the incidence of C>A transversions in smokers [30] 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.[31] Of note, we found that CH was associated with inflammatory related disorders (e.g. chronic obstructive pulmonary disorders, agranulocytosis, and ulcer of lower limb) but it is unclear whether these might be caused by CH or whether the association reflects a common inflammatory background.

Whilst extrinsic factors play a role in the development of CH, constitutional genetics is also important. Genome wide association analysis identified rs34002450, an intronic variant in *TERT* gene, to be associated with CH in the Icelandic population.[7] This SNP was not included on the array platform used by UK Biobank, but we independently identified two distinct signals within *TERT* that achieved genome wide significance: rs7726159 and rs2853677. One of these SNPs, rs7726159, is in LD with rs34002450 (r^2= 0.70) whereas rs2853677 has been associated with MPN and *JAK2* V617F associated CH.[32] Two additional SNPs in *TERT* have been reported to predispose to the development of MPN and *JAK2* V617F associated CH: rs2736100 [33, 34] and rs7705526.[32] LD analysis for the two primary signals (rs7726159 and rs2853677; Supplementary Table 12) revealed (i) rs7726159 is in LD with rs7705526 (r^2=0.79) but does not reach genome-wide significance for association with self-reported polycythemia vera (PV) in 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 CH but does not predict development of MPN but variation in intron 3 (rs2736100, rs2853677 and rs7705526) does predict development of MPN (Figure 2; Supplementary Table 12). SNP rs2853677 is not in LD with any of the other variants and is thus a unique independent signal for both CH and MPN.

We conclude that 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[35] and impacts on the tumor microenvironment via diverse pathways, including inflammation. It is possible therefore that chronic inflammation provides a link between genetic and environmental predisposition to CH.

ACKNOWLEDGEMENTS

AAZD was supported by Lady Tata International Award; NCPC was supported by Blood Cancer UK

DISCLOSURE OF CONFLICTS OF INTEREST

The authors have no conflicts of interest to declare

REFERENCES

1. Busque L, Mio R, Mattioli J, Brais E, Blais N, Lalonde Y*, et al.* Nonrandom X-inactivation patterns in normal females: lyonization ratios vary with age. *Blood* 1996; **88**(1)**:** 59-65.

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

3. Xu X, Zhang Q, Luo J, Xing S, Li Q, Krantz SB*, et al.* JAK2V617F: prevalence in a large Chinese hospital population. *Blood* 2007; **109**(1)**:** 339-342.

4. Laurie CC, Laurie CA, Rice K, Doheny KF, Zelnick LR, McHugh CP*, et al.* Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nature Genetics* 2012; **44**(6)**:** 642.

5. Jacobs KB, Yeager M, Zhou W, Wacholder S, Wang Z, Rodriguez-Santiago B*, et al.* Detectable clonal mosaicism and its relationship to aging and cancer. *Nature Genetics* 2012; **44**(6)**:** 651.

6. Genovese G, Kähler AK, Handsaker RE, Lindberg J, Rose SA, Bakhoum SF*, et al.* Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *New England Journal of Medicine* 2014; **371**(26)**:** 2477-2487.

7. Zink F, Stacey SN, Norddahl GL, Frigge ML, Magnusson OT, Jonsdottir I*, et al.* Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* 2017; **130**(6)**:** 742-752.

8. Jaiswal S, Fontanillas P, Flannick J, Manning A, Grauman PV, Mar BG*, et al.* Age-related clonal hematopoiesis associated with adverse outcomes. *New England Journal of Medicine* 2014; **371**(26)**:** 2488-2498.

9. Acuna-Hidalgo R, Sengul H, Steehouwer M, van de Vorst M, Vermeulen SH, Kiemeney LA*, 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.

10. Young AL, Challen GA, Birmann BM, Druley TE. Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. *Nature Communications* 2016; **7:** 12484.

11. Steensma DP, Bejar R, Jaiswal S, Lindsley RC, Sekeres MA, Hasserjian RP*, et al.* Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood* 2015; **126**(1)**:** 9-16.

12. Jaiswal S, Natarajan P, Silver AJ, Gibson CJ, Bick AG, Shvartz E*, et al.* Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease. *New England Journal of Medicine* 2017; **377**(2)**:** 111-121.

13. Buscarlet M, Provost S, Zada YF, Barhdadi A, Bourgoin V, Lépine G*, et al.* DNMT3A and TET2 dominate clonal hematopoiesis and demonstrate benign phenotypes and different genetic predispositions. *Blood* 2017; **130**(6)**:** 753-762.

14. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K*, et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018; **562**(7726)**:** 203.

15. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K*, et al.* Genome-wide genetic data on~ 500,000 UK Biobank participants. *BioRxiv* 2017**:** 166298.

16. Staaf J, Lindgren D, Vallon-Christersson J, Isaksson A, Göransson H, Juliusson G*, 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.

17. Van Hout CV, Tachmazidou I, Backman JD, Hoffman JX, Yi B, Pandey A*, et al.* Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank. *bioRxiv* 2019**:** 572347.

18. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N*, et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* 2009; **25**(16)**:** 2078-2079.

19. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* 2010; **38**(16)**:** e164-e164.

20. Hollander M, Wolfe DA, Chicken E. *Nonparametric statistical methods*, vol. 751. John Wiley & Sons, 2013.

21. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D*, et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* 2007; **81**(3)**:** 559-575.

22. Millard LA, Davies NM, Gaunt TR, Davey Smith G, Tilling K. Software Application Profile: PHESANT: a tool for performing automated phenome scans in UK Biobank. *International Journal of Epidemiology* 2017**:** 47(41): 29–35.

23. Therneau TM, Grambsch PM. The Cox model. *Modeling survival data: extending the Cox model*. Springer 2000, pp 39-77.

24. Grolemund G, Wickham H. Dates and times made easy with lubridate. *Journal of Statistical Software* 2011; **40**(3)**:** 1-25.

25. Roman E, Smith A, Appleton S, Crouch S, Kelly R, Kinsey 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:** 186-198.

26. McKerrell T, Park N, Moreno T, Grove CS, Ponstingl H, Stephens J*, et al.* Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hemopoiesis. *Cell Reports* 2015; **10**(8)**:** 1239-1245.

27. Jones AV, Chase A, Silver RT, Oscier D, Zoi K, Wang YL*, et al.* JAK2 haplotype is a major risk factor for the development of myeloproliferative neoplasms. *Nature Genetics* 2009; **41**(4)**:** 446-449.

28. Bolton KL, Ptashkin RN, Gao T, Braunstein L, Devlin SM, Kelly D*, et al.* Oncologic therapy shapes the fitness landscape of clonal hematopoiesis. *bioRxiv* 2019**:** 848739.

29. Gillis NK, Ball M, Zhang Q, Ma Z, Zhao Y, Yoder SJ*, 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)**:** 112-121.

30. Coombs CC, Zehir A, Devlin SM, Kishtagari A, Syed A, Jonsson P*, 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)**:** 374-382. e374.

31. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV*, et al.* Signatures of mutational processes in human cancer. *Nature* 2013; **500**(7463)**:** 415-421.

32. Hinds DA, Barnholt KE, Mesa RA, Kiefer AK, Do CB, Eriksson N*, 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)**:** 1121-1128.

33. Oddsson A, Kristinsson S, Helgason H, Gudbjartsson D, Masson G, Sigurdsson A*, et al.* The germline sequence variant rs2736100_C in TERT associates with myeloproliferative neoplasms. *Leukemia* 2014; **28**(6)**:** 1371-1374.

34. Tapper W, Jones AV, Kralovics R, Harutyunyan AS, Zoi K, Leung W*, et al.* Genetic variation at MECOM, TERT, JAK2 and HBS1L-MYB predisposes to myeloproliferative neoplasms. *Nature Communications* 2015; **6**(1)**:** 1-11.

35. Liu N, Guo XH, Liu JP, Cong YS. Role of telomerase in the tumour microenvironment. *Clinical and Experimental Pharmacology and Physiology* 2019**:** 47(43):357-364.

FIGURE LEGENDS

Figure 1: The relationship between clonal hematopoiesis and age. (A) Total mCA frequency across different age intervals showing increased risk of acquiring a mCA by 1.02 fold ($P=1.80x10^{-19}$). (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; $P=1.80x10^{-19}$). (C) Myeloid mCA frequency across different age intervals showing increased risk of acquiring a myeloid mCA by 1.1 fold ($P=1.57x10^{-38}$). (D) 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; $P=1.57x10^{-38}$). (E) Frequency of individual mutations showing an age-related increase for all genes and increased risk of acquiring mutations by 1.1 fold ($P=5.89x10^{-47}$). (F) 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; $P=5.89x10^{-47}$).

Figure 2: Genome wide association analysis to identify genetic predisposition to myeloid clonal hematopoiesis. (A) 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) (B) 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. (C) Locus zoom plot focusing on SNPs in the region of *TERT* at chromosome band 5p15. (D) Positions and genetic relationships between *TERT* SNPs.

Table 1: Summary of 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 (see Methods).
5) Data available from 49,996 cases of which 40 were excluded following QC (see Methods).

Table 2: 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)	$4.10x10^{-6}$	210 (11)	13.21	$3.74x10^{-145}$	158	16.39	$8.78x10^{-124}$	37	40.88	$3.10x10^{-44}$	34	4.61	$2.18x10^{-12}$
lymphoid disorders	12546	768 (2.1)	0.54	363 (2.9)	3.19	$1.41x10^{-105}$	146	2.15	$1.77x10^{-19}$	81	14.00	$3.09x10^{-55}$	194	4.01	$1.22x10^{-83}$
Other cancers	116921	1854 (1.6)	0.89	1185 (1)	1.10	$4.00x10^{-3}$	657	1.03	0.26	67	1.24	0.09	527	1.16	$3.6x10^{-3}$
Cancer free	355561	5269 (1.6)		3282 (0.9)			1938			165			1386		

The number of mosaic chromosomal abnormalities (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; acquired uniparental disomy (aUPD), copy number gain (CNG) and copy number loss (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}).

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

Event	mCA type	Cancer free, No. positive[†]	Myeloid malignancies*		
			No. positive	OR	P_{FDR}
chr9p	aUPD	7	86	2859	6.30×10^{-191}
chr9p	CNG	1	15	3335	7.93×10^{-33}
chr14q	aUPD	23	12	116	2.67×10^{-18}
chr9q	CNG	1	8	1771	6.93×10^{-17}
chr1p	aUPD	31	10	72	1.25×10^{-13}
chr20i	CNL	36	10	63	3.31×10^{-13}
chr4q	aUPD	13	8	136	1.04×10^{-12}
chr1q	CNG	1	4	833	4.26×10^{-8}
chr8p	CNG	5	4	177	8.45×10^{-7}
chr9p	CNL	1	3	662	5.82×10^{-6}
chr8q	CNG	4	3	166	4.03×10^{-5}
chr7q	aUPD	6	3	110	1.01×10^{-4}
chr7q	CNL	0	2	-	2.31×10^{-4}
chr17p	aUPD	4	2	110	2.85×10^{-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

Table 4: Summary of somatic mutations.

	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

Table 5: The relationship between smoking and clonal hematopoiesis

Marker[1]	No. myeloid	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}
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	2.60×10^{-3}	1.07	1.92×10^{-5}	1.94; 4.68	1.02×10^{-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	6.40×10^{-3}	0.88; 1.82	0.03
All myeloid CH	1166	488	134	1.02	0.09	1.10	6.14×10^{-6}	1.17; 1.76	3.38×10^{-6}
Myeloid CH without *ASXL1*	1066	439	114	1.01	0.36	1.07	8.8×10^{-4}	1.12; 1.59	5.81×10^{-4}

1) Loss of function (LOF); clonal hematopoiesis (CH); mosaic chromosomal alterations (mCA)
2) Number of smokers encoded in the combined smoking status in UK Biobank "Data-Field 20116"
3) Results of ordinal logistic regression, total tests = 16, corrected for age, sex and FDR.
4) 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)

Table 6: 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)	$9.1x10^{-13}$
	Platelet distribution width	%	30282	16.69	16.46	1.04 (1.03-1.05)	$4.3x10^{-9}$
	Hematocrit percentage	%	30282	41.05	41.61	0.98 (0.97-0.99)	$1.0x10^{-4}$
	Basophil percentage	%	30272	0.53	0.61	0.97 (0.96-0.99)	$2.9x10^{-4}$
	Mean corpuscular hemoglobin	g/dL	30282	34.43	34.27	1.02 (1.01-1.03)	$4.2x10^{-3}$
	Red blood cell count	10^9/L	30282	4.47	4.54	0.98 (0.97-0.99)	$4.4x10^{-3}$
	Red blood cell distribution width	%	30282	13.81	13.5	1.02 (1.01-1.03)	$4.9x10^{-3}$
	Hemoglobin concentration	g/dL	30282	14.13	14.26	0.98 (0.98-0.99)	$6.8x10^{-3}$
	Mean sphered cell volume	fL	28916	83.99	84.5	0.98 (0.97-0.99)	$3.0x10^{-2}$
ASXL1	Platelet distribution width	%	30082	16.72	16.46	1.03 (1.01-1.04)	$5.9x10^{-4}$
	Red blood cell distribution width	%	30082	13.93	13.5	1.03 (1.01-1.04)	$7.0x10^{-4}$
	Mean corpuscular volume	fL	30082	90.68	91.8	0.98 (0.97-0.99)	$9.1x10^{-4}$
	Mean corpuscular haemoglobin	pg	30082	31.03	31.47	0.98 (0.97-0.99)	$2.8x10^{-3}$
	Mean sphered cell volume	fL	28712	83.36	84.5	0.98 (0.97-0.99)	$1.0x10^{-2}$
DNMT3A	Platelet count	10^9/L	30289	251.27	242.76	1.02 (1.01-1.03)	$1.8x10^{-2}$
JAK2	Platelet crit	%	30019	0.30	0.22	1.04 (1.03-1.06)	$7.5x10^{-12}$
	Platelet count	10^9/L	30019	341.46	242.76	1.04 (1.03-1.06)	$1.3x10^{-11}$
	Red blood cell distribution width	%	30019	15.31	13.5	1.04 (1.03-1.05)	$3.6x10^{-9}$
	Platelet distribution width	%	30019	17.14	16.46	1.03 (1.02-1.05)	$1.0x10^{-6}$
	High light scatter reticulocyte	10^{12}/L	28656	0.02	0.02	1.02 (1.01-1.03)	$4.0x10^{-2}$
PPM1D	Monocyte count	10^9/L	30007	0.59	0.48	1.02 (1.01-1.03)	$3.5x10^{-2}$
SRSF2	Reticulocyte percentage	%	28643	1.88	1.32	1.02 (1.01-1.03)	$1.3x10^{-2}$
	High light scatter reticulocyte	%	28643	0.62	0.4	1.02 (1.01-1.03)	$2.4x10^{-2}$
TET2	Eosinophil count	10^9/L	30169	0.15	0.17	0.98 (0.97-0.99)	$1.1x10^{-3}$
	Eosinophil percentage	%	30169	2.18	2.53	0.98 (0.97-0.99)	$4.4x10^{-3}$
	Monocyte percentage	%	30169	7.83	7.06	1.02 (1.01-1.03)	$3.0x10^{-2}$
All Myeloid CH	Platelet distribution width	%	30883	16.57	16.46	1.03 (1.02-1.04)	$6.0x10^{-6}$
	Basophil count	10^9/L	30873	0.04	0.04	0.95 (0.93-0.97)	$5.9x10^{-4}$
	Red blood cell distribution width	%	30883	13.71	13.5	1.02 (1.01-1.04)	$9.7x10^{-4}$
	Platelet crit	%	30883	0.23	0.22	1.02 (1.01-1.03)	$1.1x10^{-3}$
	Hematocrit percentage	%	30883	41.46	41.61	0.98 (0.97-0.99)	$1.7x10^{-3}$
	Platelet count	10^9/L	30883	248.32	242.76	1.02 (1.01-1.03)	$3.3x10^{-3}$
	Hemoglobin concentration	g/dL	30883	14.22	14.26	0.99 (0.98-0.99)	$1.7x10^{-2}$
	Basophil percentage	%	30883	0.58	0.61	0.98 (0.97-0.99)	$4.4x10^{-2}$

* Total number of linear regression tests =232, corrected for age, sex, and smoking status.

Table 7: 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 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

* Total number of linear regression tests =150, corrected for age, sex, and smoking status.

Table 8: Clinical phenotypes associated with myeloid markers

Marker	**Phenotype**	**Cases**	**Controls n=30,892**	**OR (CI 97.5%)**	**P_{FDR}**
mCA n=301	N35.9 Urethral stricture, unspecified	10 (3.3%)	189 (0.6%)	1.17 (1.09-1.24)	0.008
	N32.0 Bladder-neck obstruction	7 (2.3%)	98 (0.03%)	1.19 (1.09-1.28)	0.009
TET2 n=189	L97 Ulcer of lower limb, not elsewhere classified	4 (2.1%)	19 (0.06%)	1.28 (1.14-1.39)	0.004
	D70 Agranulocytosis	4 (2.1%)	51 (0.16%)	1.23 (1.10-1.34)	0.009
	J44.0 COPD with acute lower respiratory infection	7 (3.7%)	137 (0.44%)	1.16 (1.07-1.23)	0.009

* Total number of logistic regression tests =3160, corrected for age, sex, and smoking status.

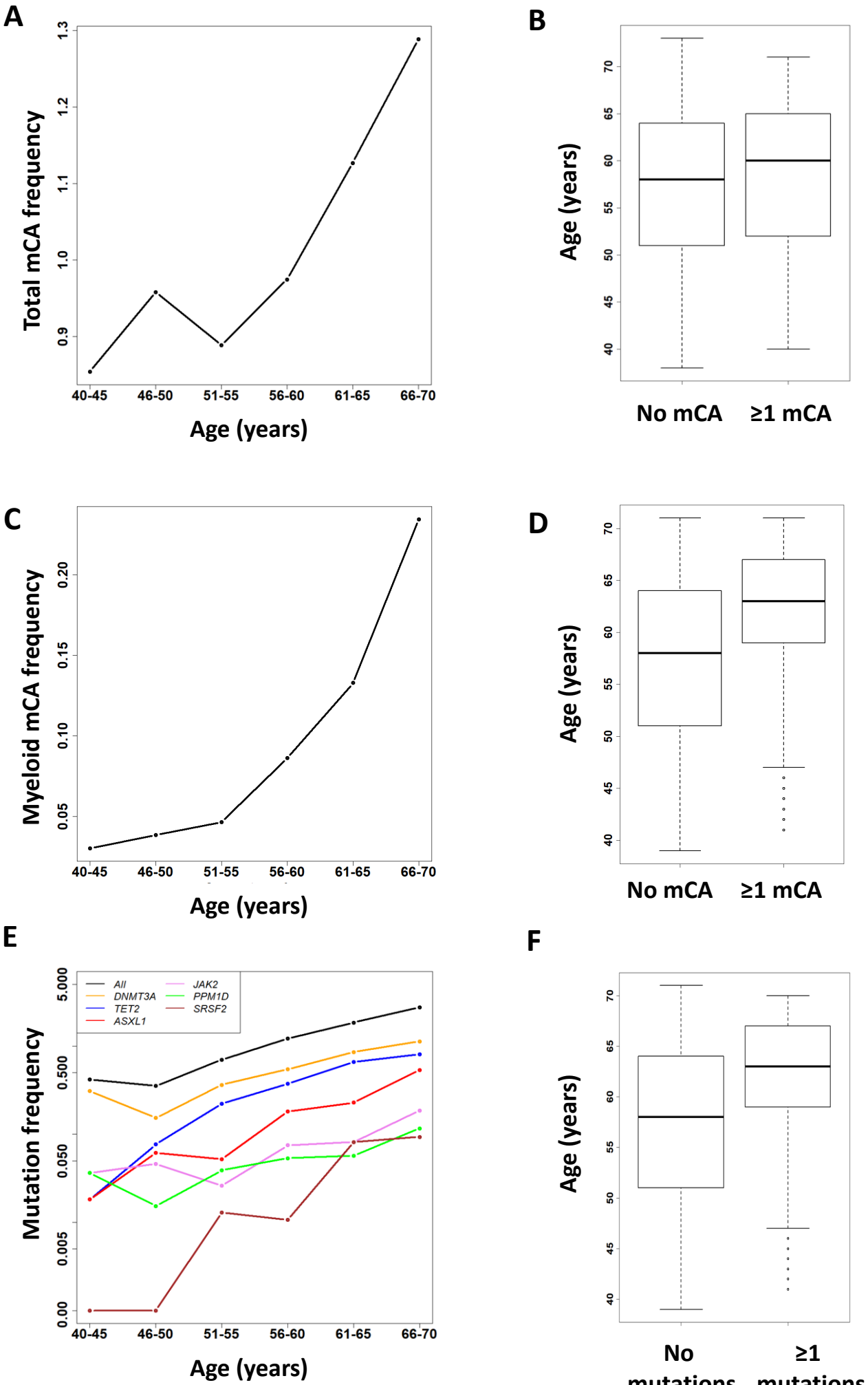
A
Total mCA frequency
0.9
1.0
1.1
1.2
1.3
40-45
46-50
51-55
56-60
61-65
66-70
Age (years)
B
Age (years)
40
45
50
55
60
65
70
No mCA
≥1 mCA
C
Myeloid mCA frequency
0.05
0.10
0.15
0.20
40-45
46-50
51-55
56-60
61-65
66-70
Age (years)
D
Age (years)
40
45
50
55
60
65
70
No mCA
≥1 mCA
E
All
DNMT3A
TET2
ASXL1
JAK2
PPM1D
SRSF2
Mutation frequency
0.00
0.005
0.050
0.500
5.000
40-45
46-50
51-55
56-60
61-65
66-70
Age (years)
F
Age (years)
40
45
50
55
60
65
70
No mutations
≥1 mutations

Figure 1

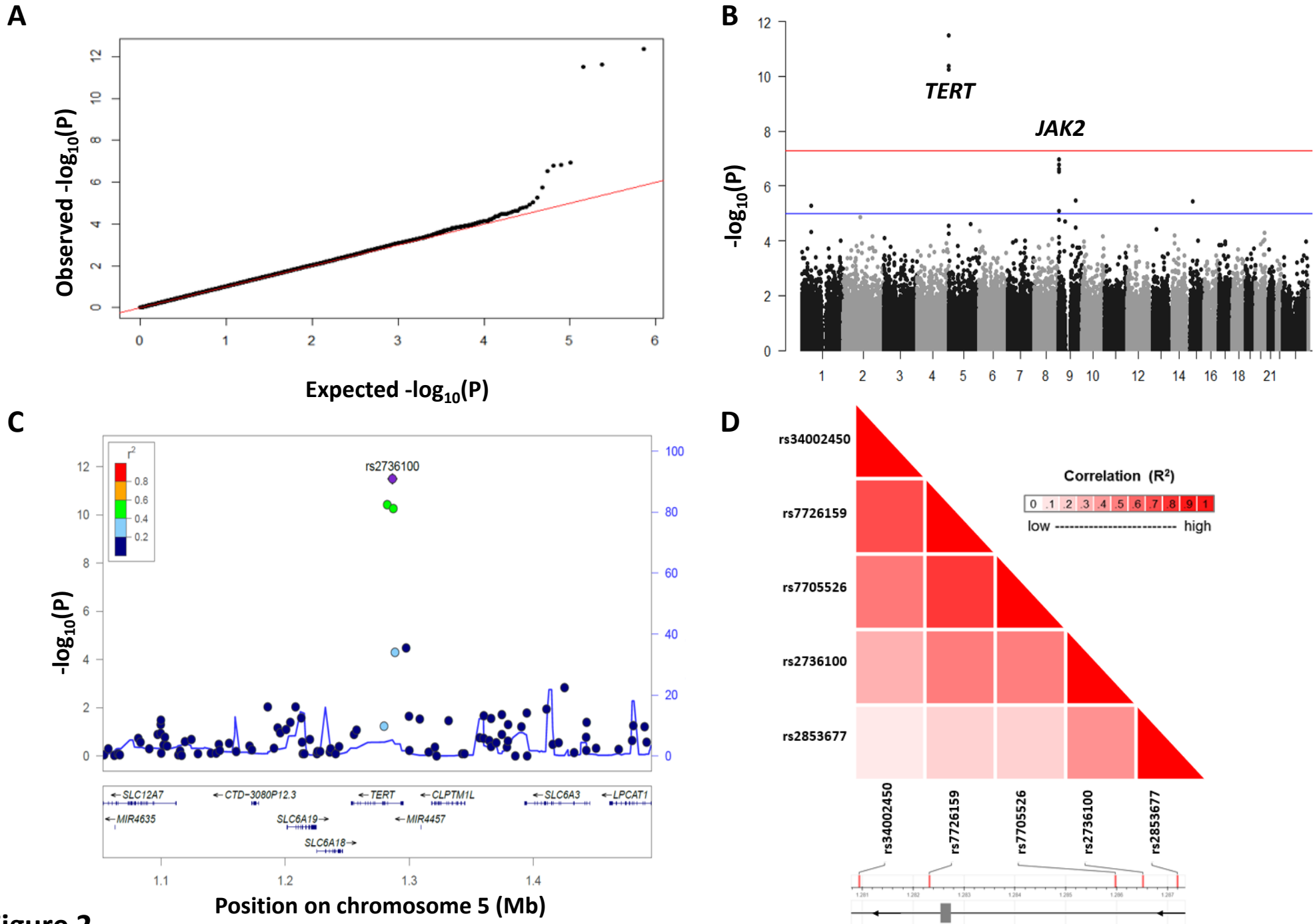
A
Observed -log10(P)
Expected -log10(P)
B
-log10(P)
TERT
JAK2
C
-log10(P)
rs2736100
r2
0.8
0.6
0.4
0.2
SLC12A7
CTD-3080P12.3
TERT
CLPTM1L
SLC6A3
LPCAT1
MIR4635
SLC6A19
MIR4457
SLC6A18
Position on chromosome 5 (Mb)
D
rs34002450
rs7726159
rs7705526
rs2736100
rs2853677
Correlation (R²)
low ---------- high
TERT intron 3
TERT intron 2

Figure 2