Genome-wide characterization of clonal hematopoiesis reveals extensive non-coding putative driver mutations

Authors: Joshua S. Weinstock[1], Karen Conneely[1], Janghee Woo[2], Marios Arvanitis[3], Mitchell J. Machiela[4], Cameron Russell[1]

Affiliations:

1. Department of Human Genetics, Emory University, Atlanta, GA
2. Division of Hematology, Emory University, Atlanta, GA
3. Division of Cardiology, The Ohio State University, Columbus, Ohio
4. Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD

Lead contact: Josh Weinstock (josh.weinstock@emory.edu).

Abstract

As humans age, we acquire somatic mutations in our blood, leading to clonal hematopoiesis (CH). Despite the prevalence of clonal hematopoiesis (CH) in aged individuals, recent searches for selective sweeps in single-cell derived colonies have revealed that most clones have expanded without a known driver mutation. This extensive, unexplained CH motivated our search for novel driver mutations across ~490K blood whole genome sequences from the UK Biobank. We searched across variants with a minor allele count of at least 10 (~147M variants) to discover alleles that are enriched in aged individuals. We identified 45 variants including known CH driver genes (e.g., *DNMT3A*, *ASXL1, SF3B1*) and 35 novel variants. Among the novel age-associated variants, we identified 72 carriers of a somatic mutation in the *TERT* promoter. Although previously reported in pan-cancer analyses, *TERT* promoter mutations have typically been excluded from population searches for CH. We also observed a somatic intronic insertion in *UGT2B7* in 1,165 carriers, a cluster of *IGH* point mutations, and centromeric variation. We conducted a phenome-wide association study (PheWAS) among 30 common disease phenotypes to characterize the phenotypic correlates of these mutations, finding that non-canonical CH mutations contributed ~47% of liability scale heritability. We then performed a genome-wide association study of the *IGH* mutations, finding that common germline variation at *GRAMD1B* is strongly associated with *IGH* mutations. Overall, we characterize CH at both increased breadth and resolution and characterize the entire cascade from upstream germline risk haplotypes to downstream clinical correlates. We release our summary statistics in a publicly accessible portal, somatic.emory.edu.

Main Text

As humans age, hematopoietic stem cells (HSCs) acquire somatic mutations, some of which result in an expansion of a mutated lineage of cells ("clone"). Recent analyses of hematopoiesis across differently aged human donors have revealed that as we age, hematopoiesis transitions from polyclonality to oligoclonality[1]. This phenomenon is commonly referred to as aging-related clonal hematopoiesis (CH), and has recently been invoked as a novel contributor to disease across several distinct organ systems, including hematologic malignancy[2–7], cardiovascular disease[8–11], kidney disease[12], liver disease[13], among others. Initial CH discovery efforts sought the characterization of specific 'driver' mutations, reporting a class of recurrent point mutations across ~80 genes[14], along with genetic lesions altering large stretches of DNA sequence, including mosaic chromosomal alterations[15–19] , loss of the Y chromosome[6,7,20], and loss of the X chromosome[21]. However, sequencing efforts of single HSC derived colonies[1] have revealed that the majority of CH occurs in the absence of known driver mutations, indicating the current characterization of the etiology of CH is largely incomplete. Concordant with these observations, evolutionary modeling of the abundance of clonal synonymous mutations[22] indicates substantial positive selection in HSCs that is unexplained by known genetic lesions. Emerging theories for unexplained clonal expansions include the possibility of epigenetic alterations ('epimutations') sufficient to alter HSC fitness[23], along with the possibility of undiscovered somatic driver mutations.

Initial searches for genetic drivers focused on coding mutations in manually curated lists of genes with links to hematologic malignancy[2,3]. Recent efforts have been defined by the ratio of non-synonymous to synonymous (dN/dS) mutations across all protein coding genes, resulting in the discovery beyond 'canonical' CH genes[24]. However, no analysis has searched across the entire genome for drivers; searches defined based on dN/dS exclude non-coding mutations, and are likely biased towards longer genes[25] which have more opportunities for non-synonymous mutations resulting in differences in statistical power. Although searches for non-coding drivers have occurred in pan-cancer cohorts[26], few analyses have performed in the biobank CH context[27].

In this work, we performed an unbiased search for drivers across ~147M coding and non-coding variants detected in ~490K blood whole genomes sequenced by the UK Biobank[28] (UKB). We defined drivers as those variants that rapidly increased in frequency during aging, which previous evolutionary models have demonstrated is highly unlikely to occur due to stochastic drift[29]. We discovered 35 novel drivers, including an intronic insertion in *UGT2B7*, a mutation in *TERT* promoter, and non-coding mutations in the Immunoglobulin Heavy Chain (IGH). We then comprehensively mapped the common disease correlates of these mutations, performing a phenome-wide association study (PheWAS) among 30 common aging-related disease phenotypes, finding 881 links. We then quantified the relative fitness estimates of these mutations using the passenger approximated clonal expansion rate[30] (PACER). Finally, to discover the inherited determinants of these mutations, we performed GWAS of the *IG* mutations, finding that common genetic variation at *GRAMD1B* was associated. Overall, we perform the first

genome-wide search for age-related CH driver point mutations and their germline and disease correlates across ~490K genomes. An interactive web portal is now available at somatic.emory.edu.

Genome-wide search for novel drivers at single-variant resolution

Driver mutations were defined as somatic variants with alleles strongly enriched in older individuals, indicating selective clonal sweeps during aging. We included all SNP and indel variants present in at least 10 carriers (n = 147M variants) among the ~490K UKB individuals with whole genome sequences and performed a genome-wide association study of the age at recruitment using a linear mixed model, resulting in 68 mutations (**Figure 1)** where the alternate allele was directly or inversely associated with age at genome-wide significance (5×10^{-8}). Among the 68 age-associated variants, 23 variants were likely of germline origin, as they were substantially more frequent and had smaller effect sizes than the 45 other age-associated variants. Among these 23 likely germline variants, some were linked to genes previously discovered in GWAS of longevity phenotypes[31], including variants linked to *APOE* (chr19_44909967_TGG_T, pvalue = 3.0×10^{-8} Table S1) and *APOC1* (chr19_44919689_A_G, pvalue = 3.3×10^{-8}, Table S1). Among the 45 other likely somatic variants, we observed several canonical CH genes, including *DNMT3A, ASLX1, GNB1, IDH2, SF3B1,* along with a gene linked to "lymphoid CH"[32] (*IGLL5*). As a subset canonical CH mutations were excluded from the Illumina Dragen output used to define the variant call set, we then performed a secondary targeted calling of known CH drivers using a pileup-based approach (Methods) and included these along with the 45 age-associated variants in all downstream analyses. Among the 45 variants, 43 were more frequent in older individuals, and two were less frequent in older individuals than young (chr2_235126205_A_AG, chr10_133690123_G_T). Out of the 45 variants, 35 are not included among canonical CH driver gene lists[14], and we refer to them as putative driver mutations.

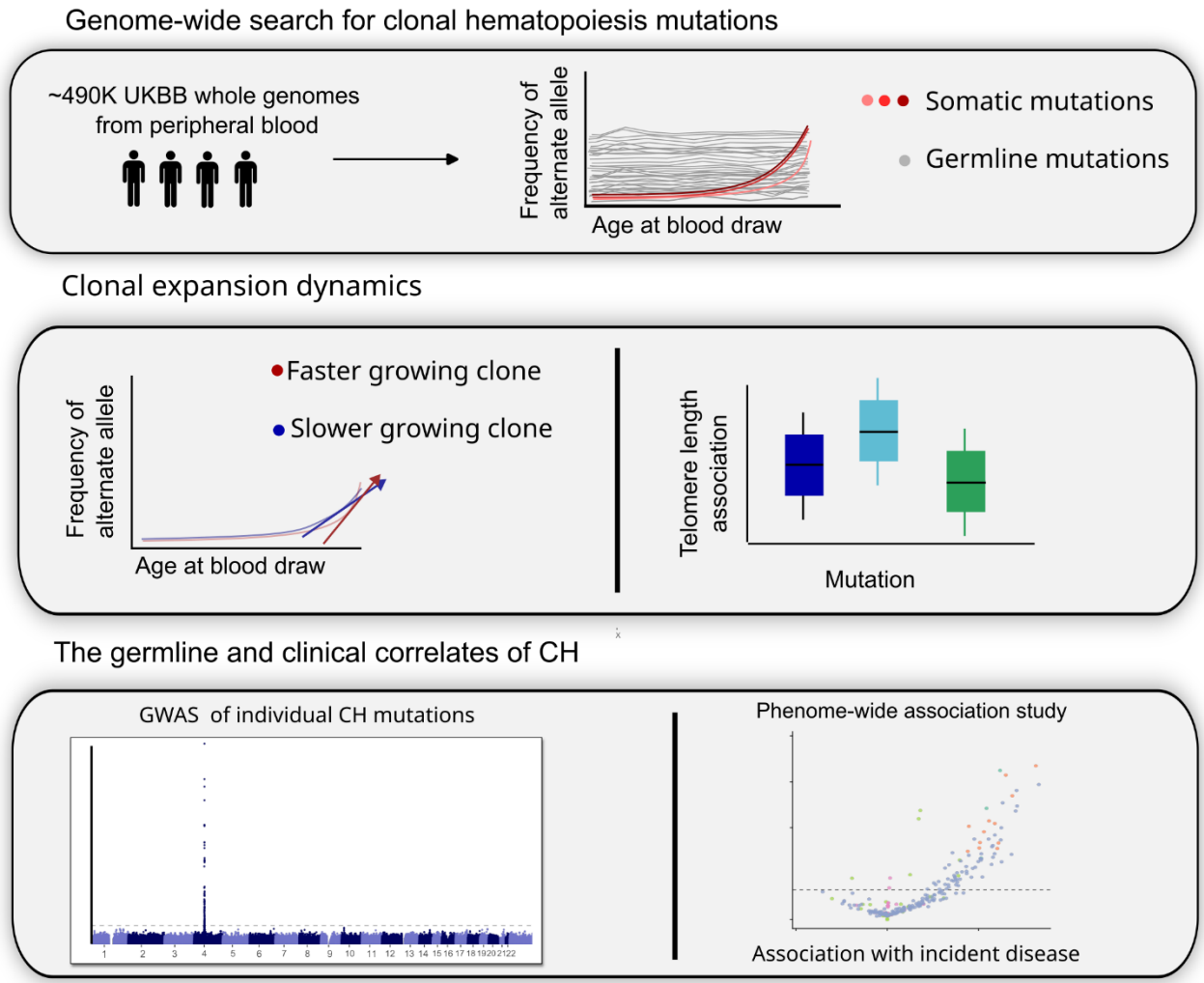


Figure 1: Study overview.

Top*: somatic mutations are called genome-wide by searching for association with age at blood draw.*

Middle*: Rates of clonal expansion are estimated using both telomere lengths and the passenger approximated clonal expansion rate (PACER).*

Bottom*: GWAS and PheWAS analyses were performed to characterize the germline and phenotypic correlates of clonal hematopoiesis.*

Among the 35 novel putative driver mutations, we observed an indel in an intron of *UGT2B7* in 1,165 carriers (chr4_69057335_T_TAC), a mutation in the promoter of *TERT*, five *IGH* mutations (four of which are non-coding: chr14_105860408_G_C, chr14_105860409_C_G, chr14_105863341_C_G, chr14_105864020_G_A), and 17 centromeric variants (**Figure 2, Supplementary Data 1**). Collectively, these mutations were all rare (frequency < 0.5%) and had large effect sizes ($|\beta| > 0.10$ standard deviations in age at recruitment). Of the 35 putative drivers, two are missense (*IGHV2-70D*, *IGLV3-1*), one

may affect splicing (chrX_155066068_C_T, *MTCP1*), and the other 32 are non-coding (**Supplementary Figure 1**).

We then asked whether the novel mutations were merely hitchhiking on clones with driver mosaic chromosomal alterations (mCAs). To assess this hypothesis, we cross referenced the novel putative drivers with mCAs called in the same samples[17]. We observed that the five *IGH* mutations overlapped a relatively frequent mCA (0.25% frequency), suggesting that they may hitchhike on mCA-IGH clones, which is concordant with the presence of a cluster of age-associated point mutations (**Supplementary Figure 2)**. In contrast, the *FRG2B, UGT2B7, DGKB,* and *FGF1* linked mutations did not overlap frequent mCAs (**Supplementary Figure 3**).

In contrast to germline variation, where the proportion of reads containing alternate alleles occurs in expectation 50% of the time for heterozygotes in diploid regions, somatic variation is likely to depart from this expectation, as the mutated reads are likely present in only a fraction of sequenced cells (sub-clonal). We used this signature to assess whether these mutations were truly somatic in origin. After performing beta-binomial hypothesis tests (**Methods**), we observed that five of the 35 putative drivers had variant allele fractions (VAFs) compatible with germline genesis (chr8_92974059_GT_G; *TRIQK*, chr2_235126205_A_AG; *SH3BP4*, chr11_112589659_G_A, chr16_6647985_G_A; *RBFOX1*, chr12_8544303_C_CA; *CLEC4E*), suggesting that they could be germline despite their association with age, or might reach these proportions due to allelic imbalances associated with overlapping mCAs. Notably, all centromeric variants had VAFs that were greater than 0.5.

We then sought to validate that our age associated variants are present in other large-scale sequencing cohorts of blood by cross-refencing variant calls derived from 184,878 blood whole genome sequences generated by the NHLBI TOPMed consortium[33]. Importantly, TOPMed uses distinct sequencing and variant calling procedures that differ from the UKB calls, suggesting that presence in both TOPMed and UKB is unlikely to be explained by shared variant calling artifacts. We found that among the 45 age-associated variants, 43 were also present in TOPMed. We observed that alternative allele frequencies were highly concordant among non-centromeric variants ($\rho = 0.63, p = 0.008$, **Supplementary Figure 4**). However, we observed that centromeric variants were reported at a much greater frequency in TOPMed than in UKB. We then recalled the age-associated centromeric variants in UKB using an orthogonal pileup based variant calling procedure, finding that coverage of reads at the locus was strongly associated with both age at recruitment and genotype (**Supplementary Figure 5)**, suggesting that the centromeric point mutations may tag broader age-associated copy number variation at the chromosome 17 centromere (chr17-CEN).

To assess this hypothesis, we then quantified coverage at the chr17-CEN, which we normalized within sample by estimating the centromere coverage relative to the coverage at *Oct4* (Methods), which was selected due to its intolerance to copy number variation. We similarly estimated normalized coverage estimates for *TET2, DNMT3A, and JAK2*, along with the sex chromosomes. Given the challenges of aligning short-read sequence to

centromeric reasons, we then performed sex stratified analyses to assess whether chr17-CEN may simply tag incorrectly mapped sequence from sex chromosomes. We observed that chr17-CEN depth is negatively associated with mosaic loss of Y (LOY) in males (beta = -0.008, $p = 5.3\ x\ 10^{-4}$), and that it is strongly associated with relative depth at *TET2* (beta = 0.199, $p < 1\ x\ 10^{-300}$, **Supplementary Figure 6**), and chr17-CEN coverage is associated with greater age at blood draw ($p = 3.0\ x\ 10^{-15}$). Taken together, these observations suggest that centromere coverage may tag a broader genomic instability phenotype, although long-read sequencing technologies may be needed to more carefully evaluate this hypothesis.

Given the challenges in assigning causal genes to non-coding variants, we then applied AlphaGenome[34], a recently developed sequence-to-function deep learning model, to classify the most likely causal gene in *cis* of the associated variants. We used AlphaGenome to predict the effects of each mutation on the expression of genes within a 1 Mb window of the mutation in hematopoietic multipotent progenitors (Methods, **Supplementary Data 2).** Given that the *TERT* promoter has previously been reported in pan cancer analyses[26], we selected this region as an exemplar locus, as it contains one of the 35 putative drivers (chr5_1295113_G_A). As anticipated, AlphaGenome predicted that this SNP is likely to upregulate *TERT* in hematopoietic progenitor cells, resulting in increased telomere lengths (**Supplementary Figure 7)**. Surprisingly, for many of the variants, AlphaGenome predicted altered regulation of multiple genes. For example**,** chrX_155066068_C_T is predicted to increase the expression of both *MTCP1* and *CMC4,* and the intronic indel in *UGT2B7* is predicted to both down-regulate *UGT2B7* and up-regulate *UGT2B11*. Notably, *MTCP1* is a member of the *TCL1* family, and is paralogous to *TCL1A*, which we previously discovered as a potent regulator of clonal expansion in CH[30].

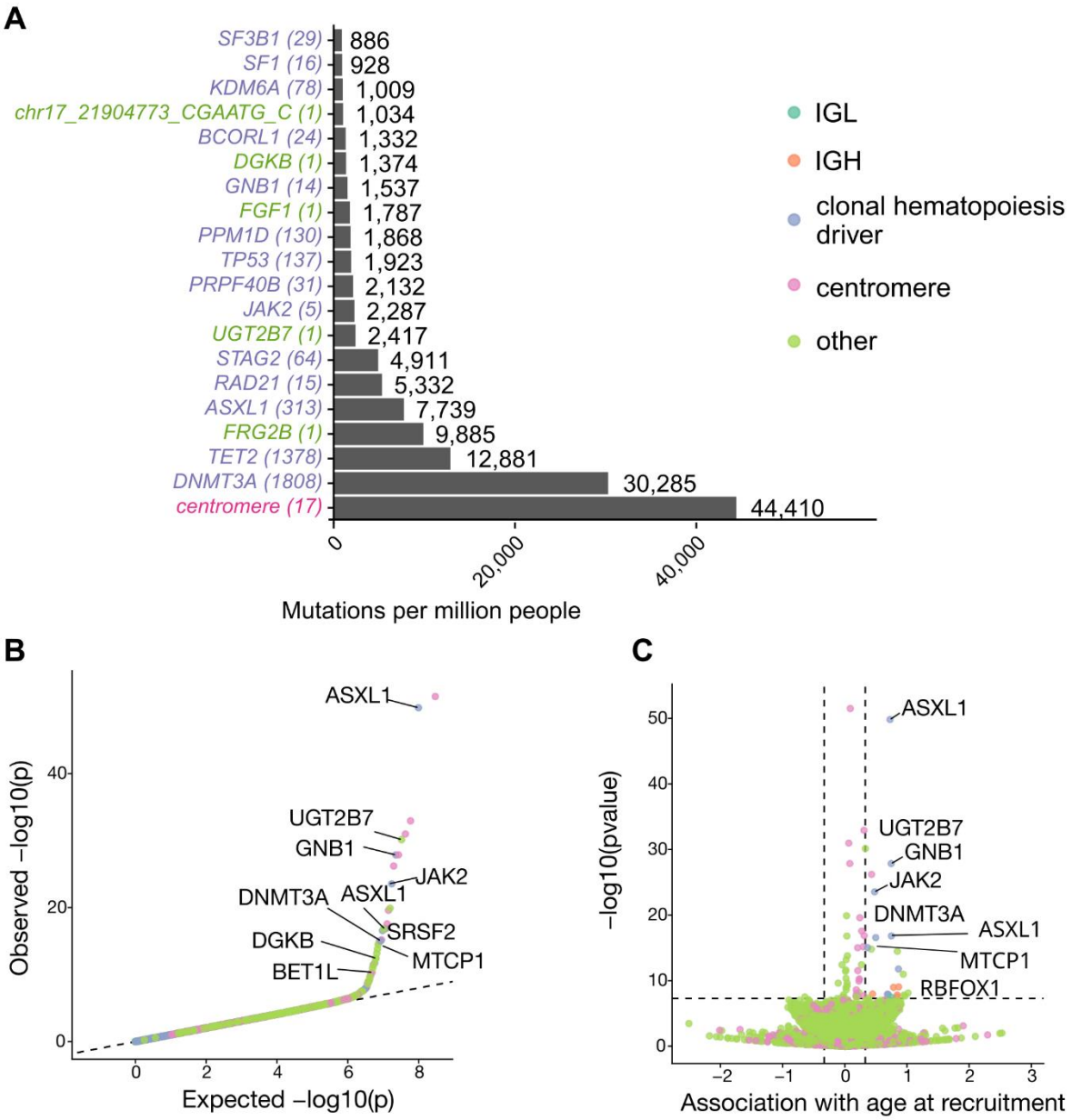


Figure 2: Genome-wide search for clonal hematopoiesis identifies novel putative drivers.

A*. The number of individuals with at least one mutation present in the annotated genes per million people. Variants were classified as CH based on a genome-wide search for alleles that were associated with age at blood draw. Canonical CH driver mutations were called with a pileup approach at candidate loci, and other CH mutations are called by the Illumina Dragen variant caller. Genetic ancestry principal components were included as covariates. All variants (n = ~147M) with at least 10 minor alleles were included, and REGENIE was used to estimate test statistics.*

B*, quantile-quantile plot from the discovery analysis, indicating limited inflation ($\lambda_{gc} = 1.01$).*

C*, A volcano plot displaying the effect sizes in terms of standard deviations of age at blood draw.*

Finally, we asked whether gene-based analyses would result in additional discovery. We performed two gene-based tests; first, we used the aggregated Cauchy association test (ACAT) [35] to meta-analyze the single-variant results within each protein gene, and we limited variants to those that were predicted to have moderate deleteriousness by AlphaMissense[36] (Methods). Second, we used ACAT on all rare variants (MAF $\leq 0.5\%$) within each protein coding gene. After adjustment for exome-wide significance ($2.5\ x\ 10^{-6}$), we observed that gene-based analyses did not increase discovery over the single-variant analyses.

Clonal dynamics of driver mutations

Several recent reports have demonstrated that larger clones are much more likely to associate with disease than smaller clones[8,11,37], underscoring the importance of estimating the rate of clonal expansion. Using a method we previously developed to estimate the relative fitness of driver genes based on quantifying the enrichment of hitchhiking passenger mutations, an approach referred to as the Passenger Approximated Clonal Expansion Rate[30] (PACER), we then estimated the fitness of the novel putative driver genes referent to *DNMT3A* R882H (chr2-25234373-C-T, **Methods**). Consistent with previous reports, mutations in splicing factors (*SF3B1, SRSF2, PRPF8)* and *JAK2* had larger PACER estimates than *DNMT3A* R882H (**Figure 3A**). Among the putative novel drivers, the *TERT* promoter mutation is estimated to have similar fitness to splicing factors and *IDH2,* and greater fitness than *JAK2*. To our surprise, the *IGLL5* and *IGH* mutations had larger fitness estimates than even the splicing factors.

As an orthogonal measure of clonal dynamics, we then asked how CH mutations associated with age-adjusted telomere length quantified with qPCR[38]. We reasoned that rapid clonal expansion would associate with increased attrition of telomere lengths, suggesting that effect of CH mutations should be inversely associated with mutation fitness. Consistent with this hypothesis, we observed that age-adjusted associations with telomere lengths were negatively correlated with PACER estimates ($\rho = -0.37$, $p = 0.016$**,** **Figure 3B**).

Concordant with a recent report describing a telomere length attrition mechanism that is preferential to splicing factor mutations[39], we observed that *SRSF2* and the *TERT* promoter were associated with smaller reduction in telomere lengths than predicted by its fitness estimate. However, we observed that this phenomenon is broader, and likely also applies to the putative drivers in immunoglobulin loci (*IGLL5, IGH)*, underscoring the considerable heterogeneity present in clonal dynamics across the CH spectrum.

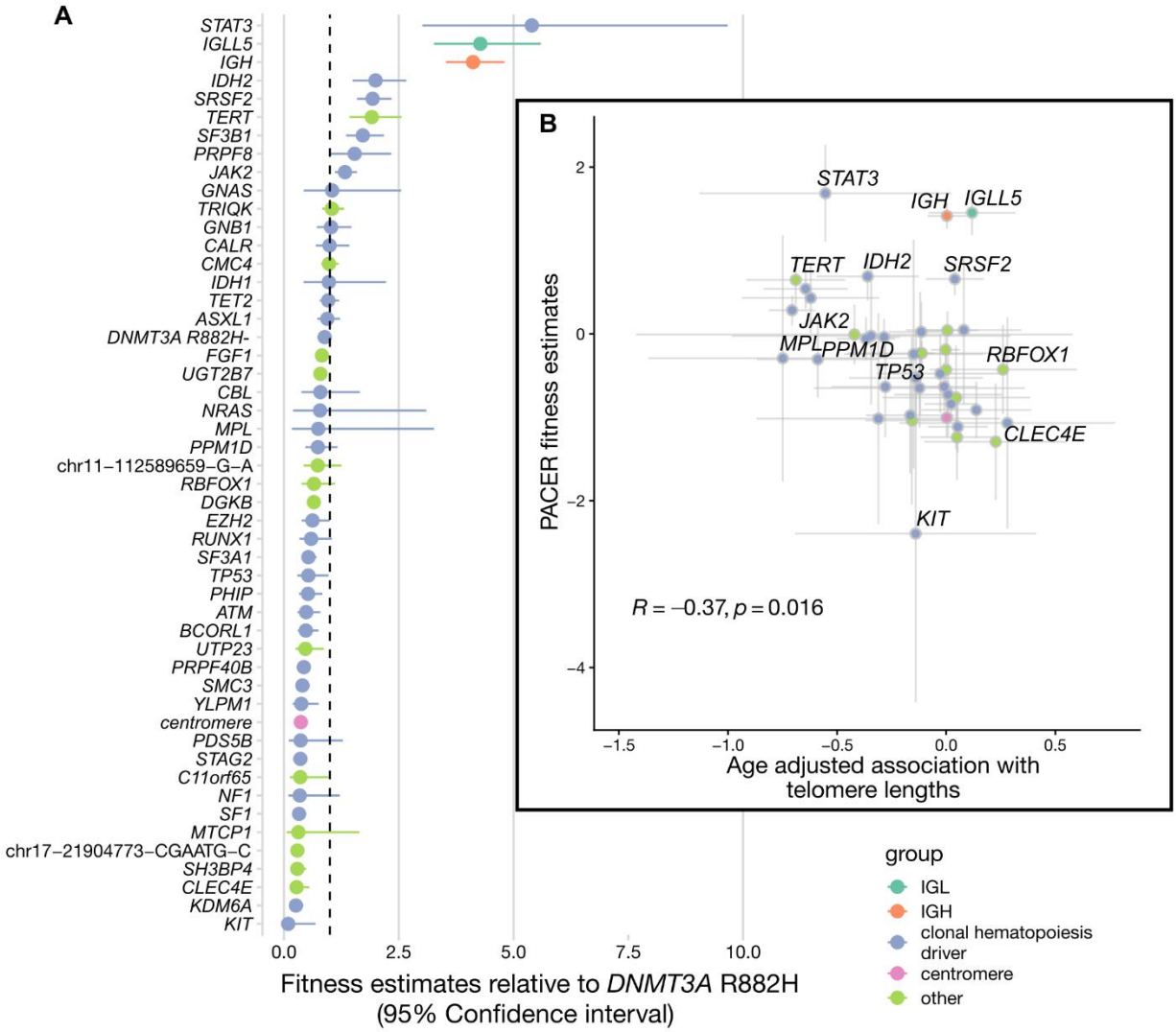


Figure 3: Fitness spectrum of clonal hematopoiesis.

A*. PACER estimates of relative driver gene fitness, with DNMT3A R882H as the reference. Excess passenger mutations are estimated using negative binomial regression after non-parametric adjustment for age and variant allele fraction as covariates.*

B*, inset. Scatter plot with x-axis indicating the association between CH mutations and telomere lengths after adjustment for age, smoking, and genetic ancestry principal components. Association analysis with telomere lengths was performed with REGENIE.*

Common disease correlates of clonal hematopoiesis

Due to an extensive literature reporting associations between CH and common aging-related disease, we then performed single-variant association analyses between CH and 30 common diseases from across the phenome. Phenotypes were defined based on "phewas" processing of the incident ICD10 codes[40], resulting in 421,584 individuals with

incident case-control information. Given the extensive possible sources of confounding in CH epidemiology[41], we included a non-parametric specification of age at blood draw, genetic ancestry principal components, sex, and smoking history as covariates. To mitigate possible effects from cryptic relatedness, we used a generalized linear mixed model to estimate test statistics (**Methods**).

To increase power, we performed a Bayesian multi-phenotype analysis to pool signals across variants and phenotypes using mashr, an approach that has previously been shown to increase power for mapping tissue-shared QTLs[42]. Using this approach, we quantified statistical significance using Bayesian posterior estimates of the local false-sign rate (LFSR), which estimates the probability that the estimated sign of the association coefficient is correct, given the data. Overall, we performed 8,460 tests across 282 variants and 30 phenotypes (**Supplementary Data 3)**. We observed 881, 503, and 396 significant associations at LFSR thresholds of 5.0%, 1.0%, and 0.5%. At a Bonferroni threshold $\left(p < \frac{.05}{8{,}460}\right)$, which is conservative given the dependent nature of the tests performed, we observed 60 significant mutation-phenotype associations, which were only observed among cancer phenotypes.

Although CH has previously been linked to several phenotypes, there are no prior reports that estimate the total variance explained on a liability scale for each phenotype – an analysis analogous to heritability estimation in complex traits. After converting effect scales to liability estimates (**Methods)**, we aggregated the variance explained by all variants for each phenotype (**Figure 4A**), which we refer to as somatic h^2 estimates. Concordant with previous CH phenotype analyses, we observed that the total contribution of CH was largest for hematologic malignancies (somatic $h^2 = 0.003$ for leukemia [phecode = 204], **Figure 4B**).

We then portioned the variance explained between canonical CH and putative novel CH mutations, finding that non-canonical CH mutations contributed on average 47% of the somatic h^2, ranging from 19% (obesity) to 90% (lymphoid leukemia). For example, we observed that the *UGT2B7* somatic indel was associated with increased risk for myeloproliferative neoplasms (odds-ratio of 1.89, LFSR = 3.7×10^{-7}). Surprisingly, we also observed a substantial contribution of *IGH* mutations to atrial fibrillation risk; chr14-105860408-G-T was associated with ~6.5 fold increase in the odds of atrial fibrillation (LFSR = $1.2\ x\ 10^{-3}$). Overall, our analysis underscores the substantial contribution of CH to hematologic and non-hematologic phenotypes, and the value of performing phenome-wide searches to place observations in context.

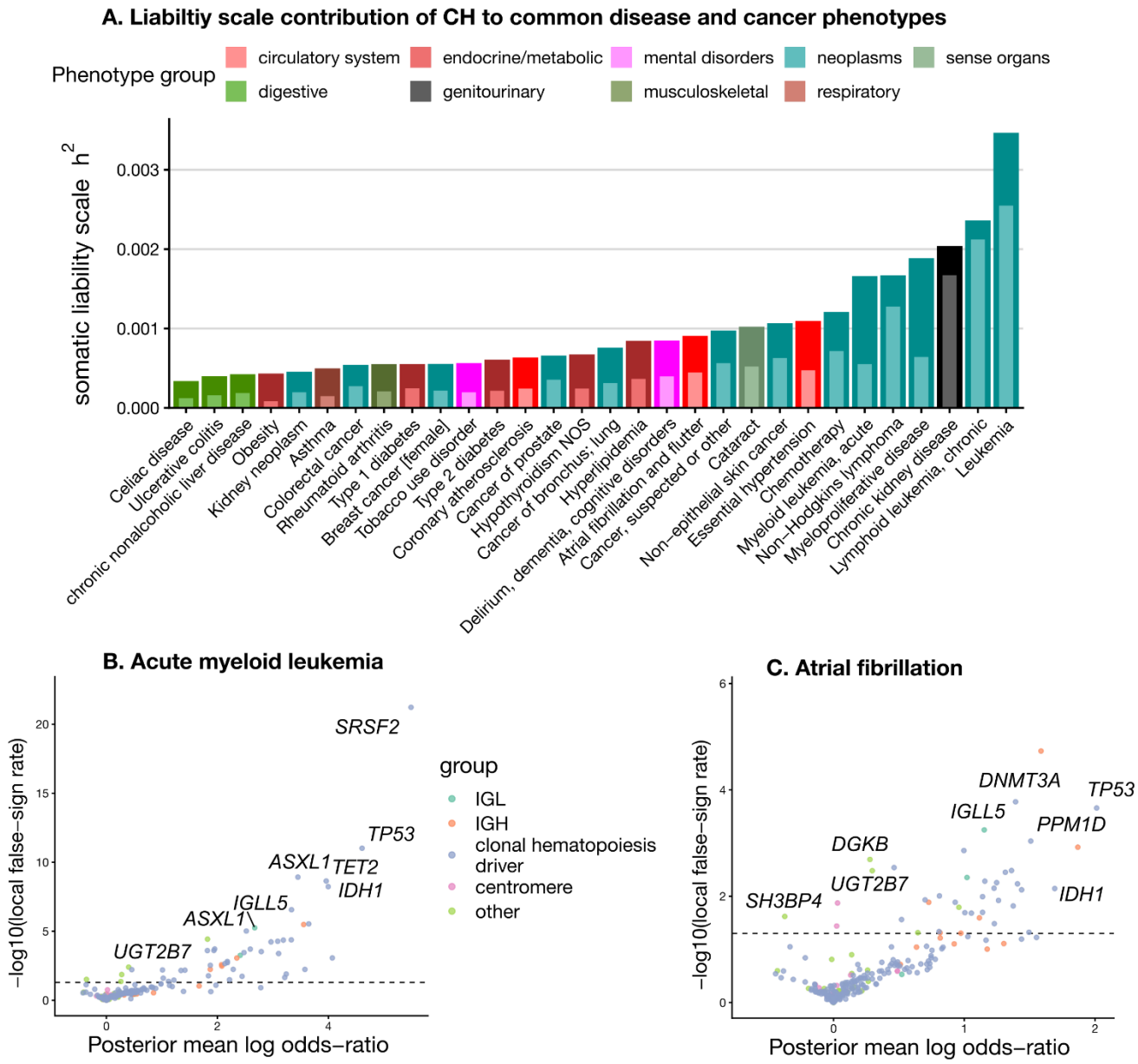


Figure 4: Common-disease correlates of clonal hematopoiesis (CH).

***A**, Liability-scale variance explained (somatic h2) by CH across 30 incident phenotypes in UK Biobank. Bars show the summed liability-scale variance from all CH variants per phenotype, with colors denoting phenotype groups. Lighter colors denote the contribution from novel putative drivers.*

***B–C**, Association landscapes for acute myeloid leukemia (B) and celiac disease (C). Points are CH variants colored by annotation group; the x-axis is the mashr posterior mean log odds ratio and the y-axis is −log10(local false-sign rate, LFSR). Selected loci are labeled. Positive values indicate increased risk; negative values indicate protection. The dashed horizontal line marks LFSR = 0.05. All models adjust for age at blood draw, sex, smoking, and genetic ancestry principal components, and account for relatedness via a generalized linear mixed model before multi-phenotype shrinkage with mashr.*

Germline determinants of clonal hematopoiesis

Analyses of the genetic architecture of clonal hematopoiesis have revealed several common variant loci that collectively implicate HSC self-renewal, DNA repair, cell-cycle regulation, and cytokine signaling as key underlying pathways[7,17,19,21,27,30,41,43,44]. Given that previous genome-wide association studies (GWAS) have largely performed analyses that aggregate all CH mutations together (*DNMT3A / TET2 / ASLX1 / JAK2* analyses excepted), we performed a search for germline determinants of individual putative drivers. Given the substantial fitness effects of mutations in immunoglobulin loci, we first performed a GWAS of case-control phenotype for whether individuals carried at least one *IGH / IGL* mutation (n = 343 cases, 482,326 controls). After inclusion of age at blood draw, smoking history, sex, and genetic ancestry as covariates, we estimated single variant test statistics with REGENIE[45] including both common and low-frequency ($\geq 0.5\%$) variants. To avoid circularity from cross-mapping artifacts, we use stringent measures to exclude possible artifacts (**Methods**).

We observed a single genome-wide significant ($p < 5\ x\ 10^{-8}$) locus including *GRAMD1B* (**Figure 5**). The G allele variant at this variant, rs12797056, is common (10% AF), and was associated with greatly increased odds for an immunoglobulin mutation (odds-ratio = 1.87, $p = 1.6\ x\ 10^{-9}$). We cross referenced this locus with phenome-wide association studies performed in the same cohort[46], finding that *GRAMD1B* locus is also reported in GWAS of chronic lymphoid leukemia (CLL). Overall, these observations suggest that our variant calling workflow recapitulates a CLL linked phenotype *de novo* from sequence alone.

We then performed a GWAS of the *UGT2B7* mutation, finding two distinct genome-wide significant loci flanking the mutation (chr4:69057335, **Supplementary Figure 8**). The first locus is in *cis* of *UGT2B7* (< 100Kb from TSS), and the lead variant (rs4348158) is low-frequency (MAF = 0.6%); the G allele is associated with greatly reduced odds of acquiring the *UGT2B7* mutation (odds-ratio = 0.26, $p = 3.0\ x\ 10^{-44}$). The second locus is 800Kb downstream of the mutation, and the lead variant (rs72851490) is ~2KB from TSS of *SULT1E1*. Large *cis* germline hits have previously been reported for *JAK2*[47,48] and for several mCAs[17]. Both *UGT2B7* and *SULT1E* have previously been linked to sex-hormone phenotypes[45,49]; androgens have previously shown to stimulate myelopoiesis through increased expression of *TERT* in CD34+ cells[50]. Sex-hormone binding globulin has also been nominated as a causal contributor mosaic loss of Y in males[51]. Collectively, these observations suggest a possible sex-hormone mediated pathway through which *UGT2B7* and its germline determinants affect hematopoiesis.

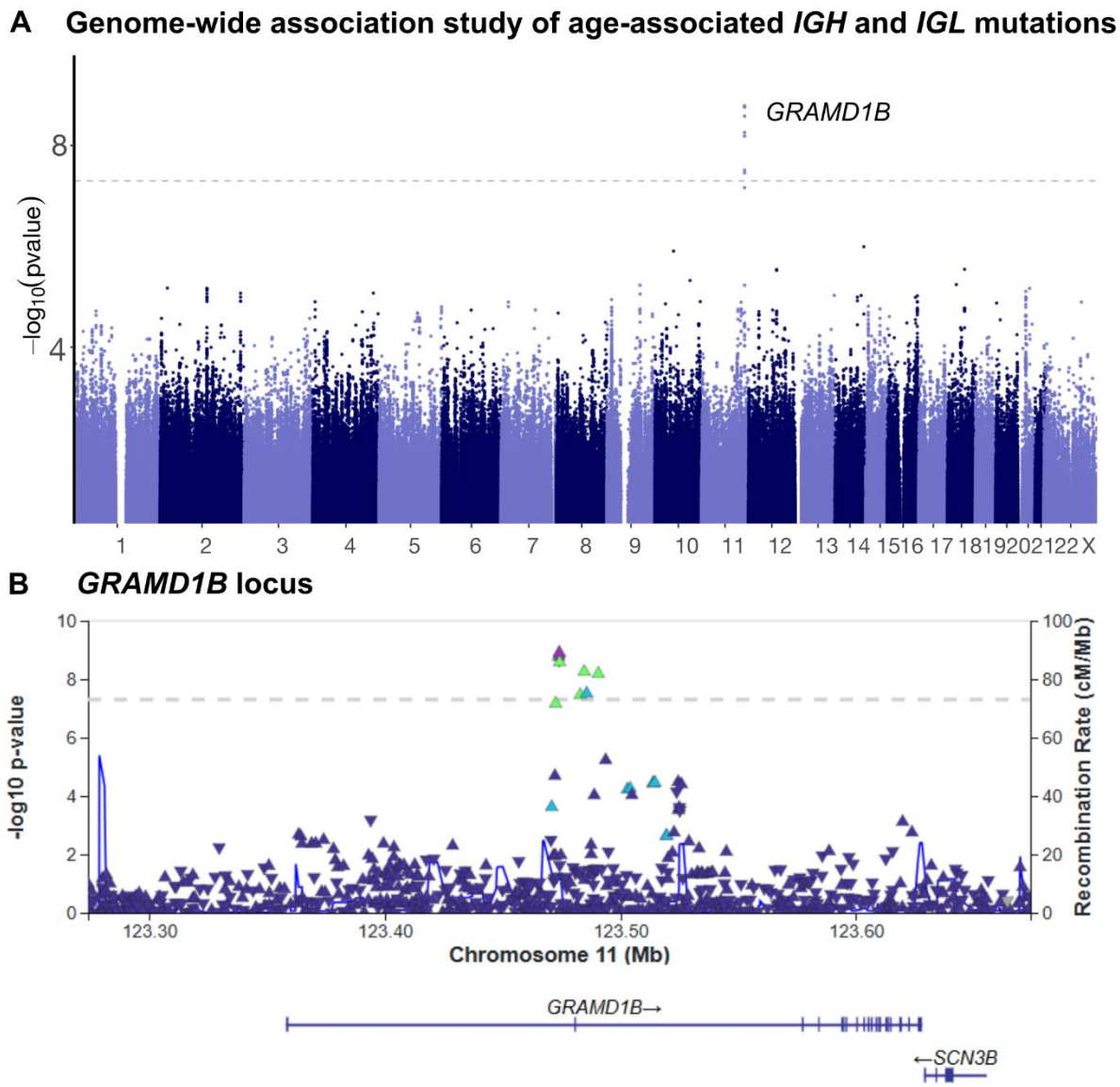


Figure 5: GWAS of IGH and IGL mutations (343 cases, 482,326 controls) recapitulates CLL locus.

A*. GWAS was performed with REGENIE including all variants with MAF > 0.5%. A spline of age, sex, smoking history, and genetic ancestry PCs were included as covariates. Variants falling within contigs that differed between hg19 and hg38 were excluded. Variants falling within UCSC "unusual" regions and hg38 genome reference consortium "exclusions" list were also excluded to mitigate cross-mapping artifacts.*

B*. Locuszoom plot of GRAMD1B locus. LD is calculated using EUR 1000G reference samples.*

Discussion

Enabled by population scale whole-genome sequencing, we conducted a single-variant search for CH across the entire genome. We utilized a simple study design to

329 search for age-associated rare variants, finding that this design was sufficient to
330 recapitulate the majority of canonical CH driver mutations. This search revealed several
331 non-coding point mutations beyond the *TERT* promoter, including mutations that likely
332 affect *UGT2B7, FGF1, MTCP1/TCL1C,* and *DGKB*. We also observed clusters of age-
333 associated mutations in *IGH* and other sites with frequent mCAs, suggesting that a scan for
334 clustered point mutations may be sufficient to nominate mCAs. Using our genome-wide
335 single-variant calls, we then quantified the fitness of these mutations, their germline
336 determinants, and their common disease correlates.

337 To estimate the relative fitness of these mutations, we utilized an approach that we
338 previously developed that uses passenger mutations to estimate an approximate time of
339 mutation acquisition, which is proportional to mutation fitness in a single-mutation HSC
340 population after adjustment for time of blood draw[30]. We observed that relative to
341 canonical myeloid CH driver mutations, *TERT* promoter mutations had similar fitness to
342 mutations in splicing genes and *IDH2,* which have previously been shown to be highly
343 predictive of progression from CH to AML[52]. Stronger still were *IGH* and *IGL* mutations,
344 which had substantially greater excess passenger mutations, suggesting that they were
345 acquired later in life and thus expanded more rapidly to detectable clone sizes.
346 Concordant with recent work describing telomere attrition mechanisms for driver
347 acquisition[39], we observed substantial heterogeneity in the rate of telomere attrition
348 associated with each mutation. We observed that along with *SRSF2* and *STAT3*,
349 immunoglobulin mutations were far less likely to associate with telomere length attrition
350 after adjustment for age at blood draw, suggesting that telomere sensitive clonal expansion
351 mechanisms are likely more frequent than previously described. We found that that the
352 most frequent novel putative drivers were consistent with fitness estimates slightly lower
353 than *DNMT3A-R882H*$^{(-)}$*,* which would explain why they have received relatively little
354 attention in hematologic malignancies; similar to genetic discovery in germline contexts,
355 there is often a long-tail of non-null smaller effects sitting just below the threshold of
356 discovery.

357 Although prior reports have performed PheWAS of CH[5,41,44], none have reported an
358 estimate of the proportion of variance explained by CH mutations on disease risk on a
359 liability scale, which determines the possible utility of CH for population screening
360 programs. Unsurprisingly, we find that CH has a very large contribution to hematologic
361 malignancies. However, we also find a contribution to chronic kidney disease (CKD) that is
362 comparable to some hematologic phenotypes, consistent with recent reports that describe
363 links between CH and acute kidney injury[12]. In accordance with the majority of CH PheWAS
364 reports in the UKB[5,41,44], we find that CH explains little of the variance in coronary
365 atherosclerosis (CAD) at a population level, suggesting that its value as a risk factor is likely
366 substantially less than traditional CAD risk factors.

367 Although *IGH* mutations have historically been only surveilled in lymphoid
368 malignancies, we find that *IGH* point mutations are strongly associated with increased risk
369 for myeloid malignancies as well, suggesting that myeloid CH analyses that exclude *IG*
370 mutations are missing a substantial contributor to disease risk. Although generally
371 considered to be restricted to B cells, our results suggest that *IGH* mutations are likely to

present frequently in myeloid progenitors as well, otherwise they would be extremely difficult to detect with ~30x sequencing of peripheral blood. We observed that *IGH* mutations may affect other organ systems as well, given the large increase in risk of atrial fibrillation associated with *IGH* mutations.

Intriguingly, a GWAS of a case-control *IG* phenotype defined by point mutations revealed a single significant locus at *GRAMD1B*. Given that *GRAMD1B* has previously been reported in CLL GWAS, this suggests that it may be possible to define a CLL phenotype *de novo* from sequence by examining *IG* genes for specific point mutations. This suggests that essentially any large-scale biobank with whole-genome sequencing is likely able to ascertain a CLL precursor phenotype.

Our study is not without limitations. First, as with any epidemiologic analysis of CH, inference of causality is challenging; in contrast to germline variation, the time at which somatic variation is acquired is generally unknown, precluding temporal ordering of disease onset and mutation acquisition. Furthermore, confounders that are known to cause CH are common and difficult to completely adjust for, including smoking effects[41]. Chronic low grade inflammation also likely has a causal effect on CH[53], indicating likely complex causal structure implicating inflammation, CH, and several aging-related diseases. Secondly, although we report novel mutations that are age-associated – and thus indicating positive selection in HSCs -- distinguishing between bona-fide 'drivers' and mutations that 'hitchhike' on drivers remains challenging. Functional characterization of *UGT2B7, FGF1, MTCP1 / TCL1C,* and *DGKB*, which are all expressed much more highly in non-blood tissues than blood, is an important area of future work; their function here remains largely uncharacterized. Finally, although we observed extensive variation at the centromere associated with age, our sensitivity is highly limited at the centromere given the challenges of aligning short reads to highly repetitive regions. Future work using long-read sequencing in large cohorts may be required to fully characterize age associated variation at centromeres.

Overall, we map somatic point mutations across the entire genome, finding novel disease associated mutations in parts of the genome that have typically been excluded from searches for CH driver mutations. Our summary statistics are available in a publicly available portal at somatic.emory.edu.

Methods

Cohorts, ethical approvals, and consent

Whole-genome sequencing (WGS) and clinical phenotypes were obtained from the UK Biobank (UKB) under application 454659. Participants provided informed consent at enrollment; the UKB study protocol received ethics approval from the UK National Research Ethics Service.

Age association single variant analysis

A 'gwas' of age at blood draw was performed using all variants present in "/Bulk/DRAGEN WGS/DRAGEN population level WGS variants, PLINK format [500k release]/" plink files. All variants with at least 10 minor alleles were included, and sex, 'ever smoker', and the first 10 genetic ancestry PCs were included as covariates. Regenie[45] was used to perform association analyses. Age associated variants were declared at a genome-wide significance threshold.

Gene-based association analyses

Two kinds of gene-based analyses were performed. First, single-variant pvalues were meta-analyzed within all protein-coding genes with ACAT[35], including all variants with minor allele frequencies < 0.5% were included. Secondly, an RVAS was performed by using only rare (< 0.5% MAF) variants with AlphaMissense pathogenicity scores greater than 0.3. Pvalues were again meta-analyzed with ACAT.

Loss of Y calling

Mosaic loss of Y was called on the whole-genomes by first quantifying sequencing depth at a 12 pre-specified regions using Mosdepth[54], including three on the X chromosome, three on the Y chromosome, the *DNMT3A* locus, *TET2* locus, *JAK2* locus, *OCT4* locus, *ASXL1* locus, and the chr17 centromere. Depth was then normalized per sample by dividing by the depth at the *OCT4* locus using the following WDL: https://github.com/weinstockj/mca_calling_workflow . Mosaic loss of Y was defined in males as samples with a relative Y chromosome depth of less than 0.369.

Variant calling of canonical CH mutations

Non-reference alleles were called at a set of pre-specified list of 4,737 variants defined based on the union of CH mutations detected in TOPMed[43] and a previous UKB analysis[5]. Variants were then called on the UKB exomes at these sites using a bcftools mpileup approach:

```
bcftools mpileup -Ou \
    -q ~{q} \
    -Q ~{Q} \
    -R ~{region_file} \
    --indels-2.0 \
    -d 500 \
    -a FORMAT/AD,FORMAT/DP,INFO/AD,FORMAT/SP,FORMAT/SCR \
```

```
                -f ~{reference_fasta} ${bam_file} |
                bcftools call -mv -Ou |
                bcftools norm -d none -f ~{reference_fasta} -Ob -o
${prefix}.candidates.bcf
```

Where “q”, the minimum mapping quality of the alignment was set to 10, and “Q”, the minimum base quality was set to 20. After preliminary calling, all variants with mean variant allele fractions (VAFs) between [.46, .54] were excluded, which totaled 577 variants. Canonical mutations were detected in 18,010 individuals.

Variant calling of putative novel drivers

After identifying age-associated variants, we then extracted variant allele fractions and other QC metrics from the Dragen single sample VCFs at the specified 45 variants, using a custom Rust binary (https://github.com/weinstockj/parseSomaticDragen).

Passenger burden quantification

Passengers were quantified by parsing the Illumina Dragen UKB pVCFs “/Bulk/DRAGEN WGS/DRAGEN population level WGS variants, pVCF format [500k release]”. pVCFs were first filtered to bi-allelic singleton SNPs (mac = 1). Then, using a custom Rust binary (https://github.com/weinstockj/count_somatic) , output was filtered to variants with variant allele fractions between [0.05, 0.35], and finally again filtered to variants with the number of alt reads between [4, 6].

PACER gene level estimates

Relative fitness was estimated using PACER, which leverages enrichment of hitchhiking passenger mutations to infer the relative time of driver acquisition and clone growth. Briefly, excess passengers per clone were modeled with negative binomial regression, adjusting for age at blood draw and observed VAF with natural splines.

$$E(Y_i) = f(VAF_i) * Sex_i + f(age\ at\ recruitment) + \boldsymbol{X}_i,$$

$$Y \sim NB(E(Y_i), \phi)$$

Where $f(x)$ denotes a natural spline, and $\boldsymbol{X_i}$ denotes an indicator function for mutation carrier status.

PACER estimates are reported relative to *DNMT3A* R882H (chr2:25234373 C>T).

Telomere length estimates

Regenie was used to compute test statistics for an association analysis between CH mutations and qPCR quantified telomere lengths, which have been previously adjusted for

technical covariates[38] ("z_adjusted_telomere_ratio"). A natural spline of age, sex, "ever smoker", and the first 10 genetic ancestry PCs were included as covariates.

Phenome-wide association study

ICD10 codes were extracted and processed with the "phecode" PheWAS mapping[40]. 30 common diseases were selected for analysis. Single-variant association analyses were performing using CH point mutations with at least 10 minor alleles using Regenie. Smoking history, a spline of age, sex, and the first 10 genetic ancestry PCs were included as covariates. After obtaining single-phenotype analyses, a multivariate analysis was performed with mashr[42], which also estimates posterior estimates of log odds-ratios, qvalues, and local false-sign rates[55].

Liability scale variance explained

To quantify the contribution of clonal hematopoiesis (CH) mutations to disease risk, we first estimated the observed variance of each phenotype on the binary scale. For each phenotype, the prevalence (p) was estimated based on the prevalence in the UKB cohort.

Next, we converted single-variant effect sizes from odds ratios (OR) to liability-scale beta coefficients using a threshold model under a probit approximation. For each variant, the liability-scale effect $\beta_{liability}$ was derived from a log odds ratio conversion to probit scale.

We aggregated the liability-scale variance explained across all CH variants for each phenotype using:

$$h^2 = \sum \beta^2_{liability} AF * (1 - AF)$$

This calculation was performed separately for (i) all CH variants and (ii) novel non-canonical CH variants (excluding known CHIP genes). Covariance between mutations was ignored as an approximation.

GWAS

GWAS of the *IG* and *UGT2B7* mutation phenotypes were performed with Regenie. A spline of age, sex, smoking history, and the first 10 genetic ancestry PCs were included as covariates. All variants were included with minor allele frequencies of at least 0.5%. Variants were then restricted to those were marked "PASS" among variants with at least 10 minor allele counts in TOPMed Freeze 12[33]. Three additional exclusion lists were then applied:

1. Variants falling within contigs that differed between hg19 and hg38
2. Variants in "UCSC unusual" regions
3. Variants in the GRC exclusions list for hg38

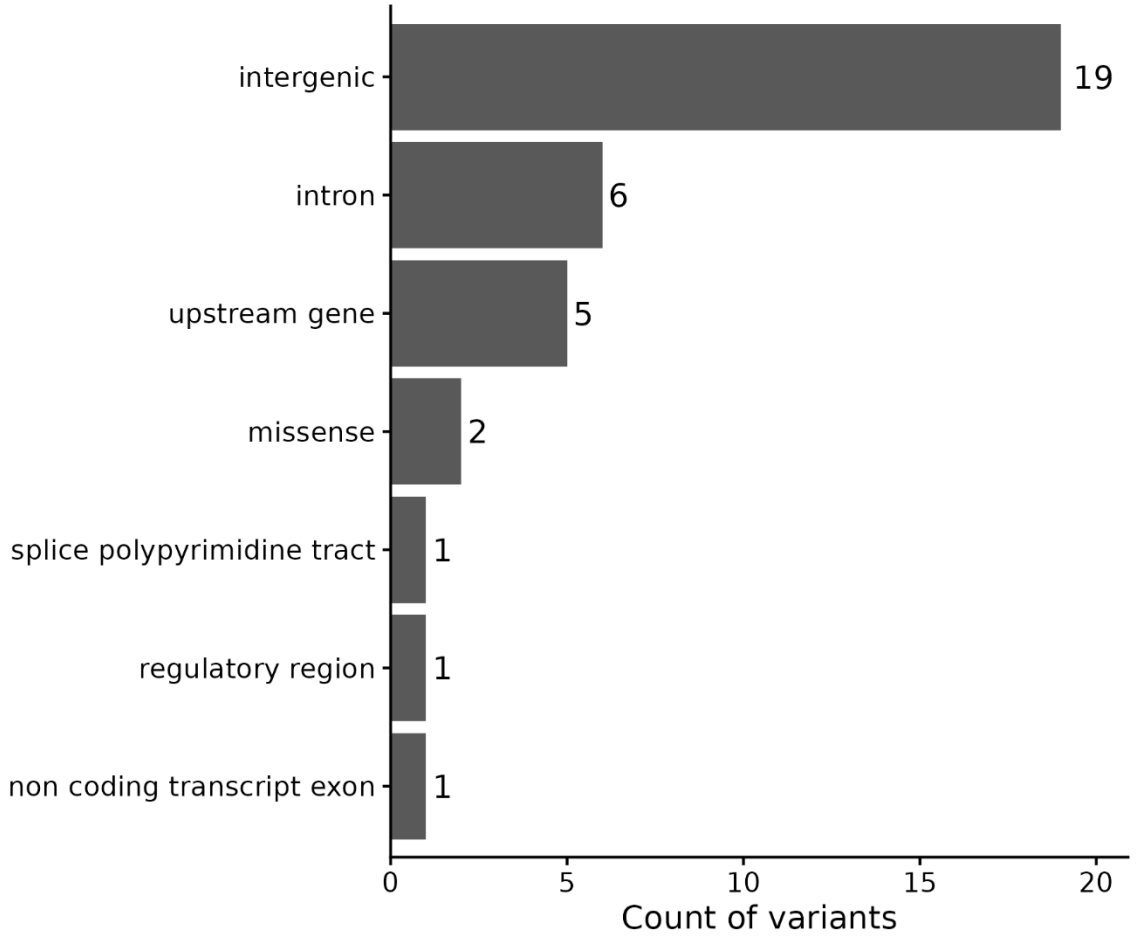
Code availability

1. Regenie WDL workflow: https://github.com/weinstockj/UKB_regenie_workflow
2. Illumina Dragen VCF to parquet: https://github.com/weinstockj/parseSomaticDragen
3. Loss of Y calling: https://github.com/weinstockj/mca_calling_workflow
4. CHIP calling: https://github.com/weinstockj/UKB_CHIP_calling
5. Variant annotation, filtering, plotting: https://github.com/weinstockj/gwasplot
6. Passenger mutation Rust binary: https://github.com/weinstockj/count_somatic
7. Passenger mutation calling WDL: https://github.com/weinstockj/count_somatic_workflow

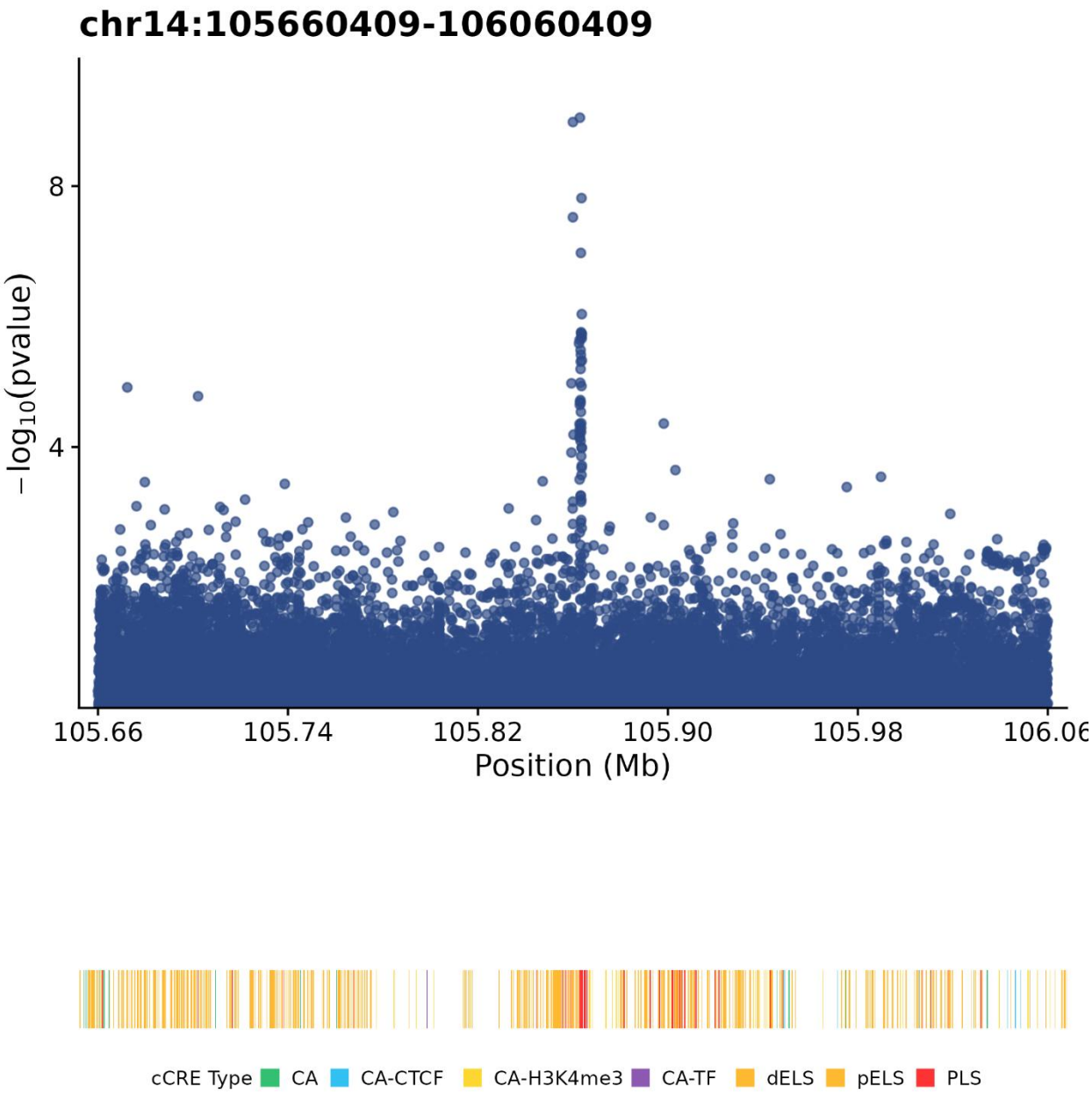
Data Availability

UKB data were accessed under application number 454659. Summary statistics for PheWAS analyses are available at a public web portal, somatic.emory.edu. GWAS summary statistics for the *IG* GWAS are available here: https://my.locuszoom.org/gwas/493597/ .

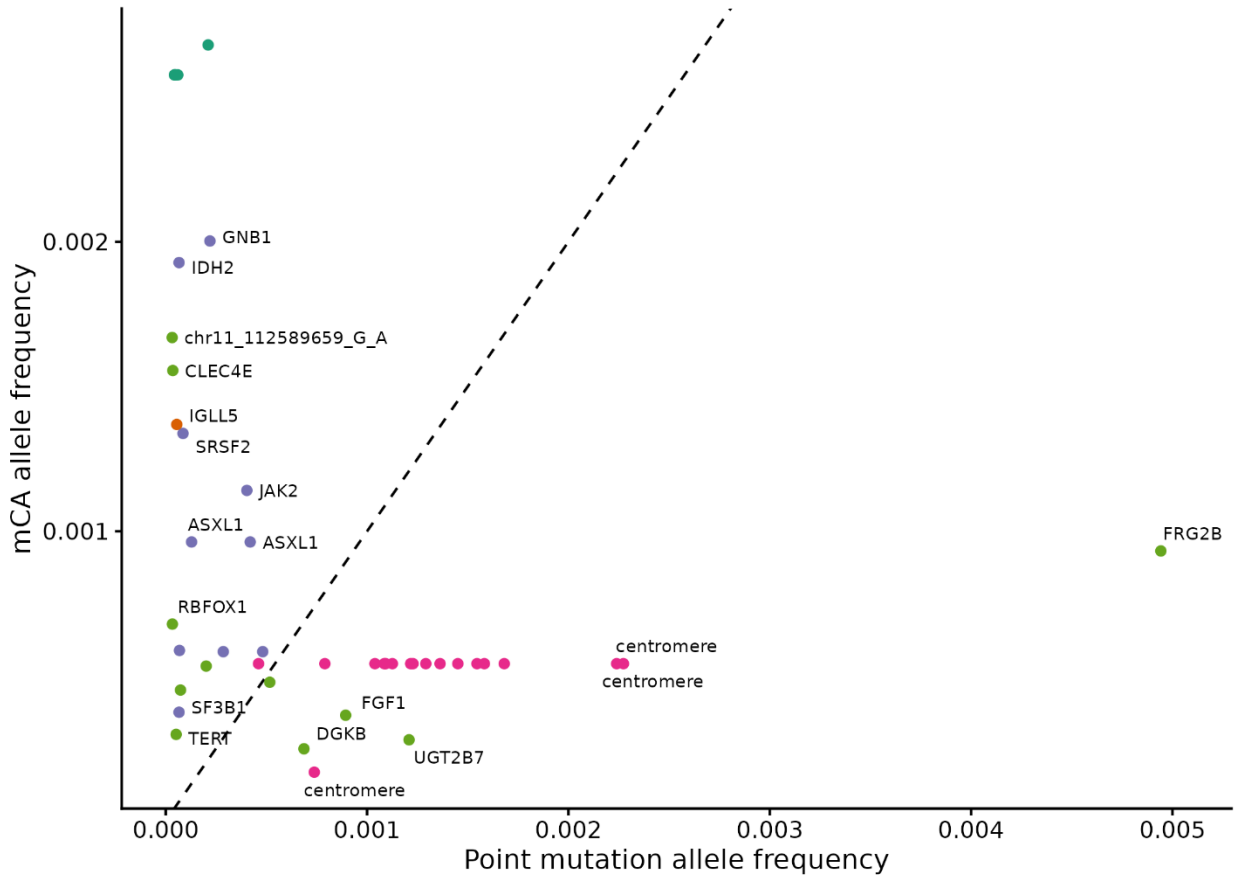
Supplementary Figures



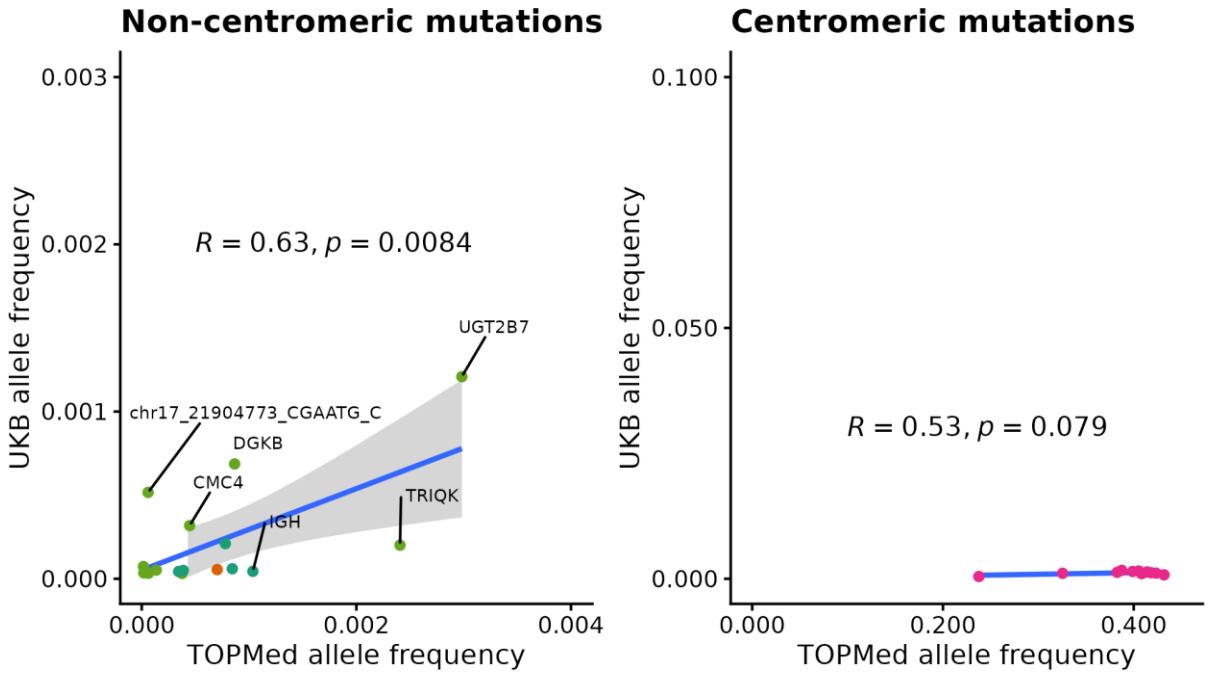
Supplementary Figure 1: Distribution of functional annotations for 35 novel age-associated variants, as annotated by the Ensembl variant effect predictor. Bars indicate counts of missense, splicing, and non-coding variants.



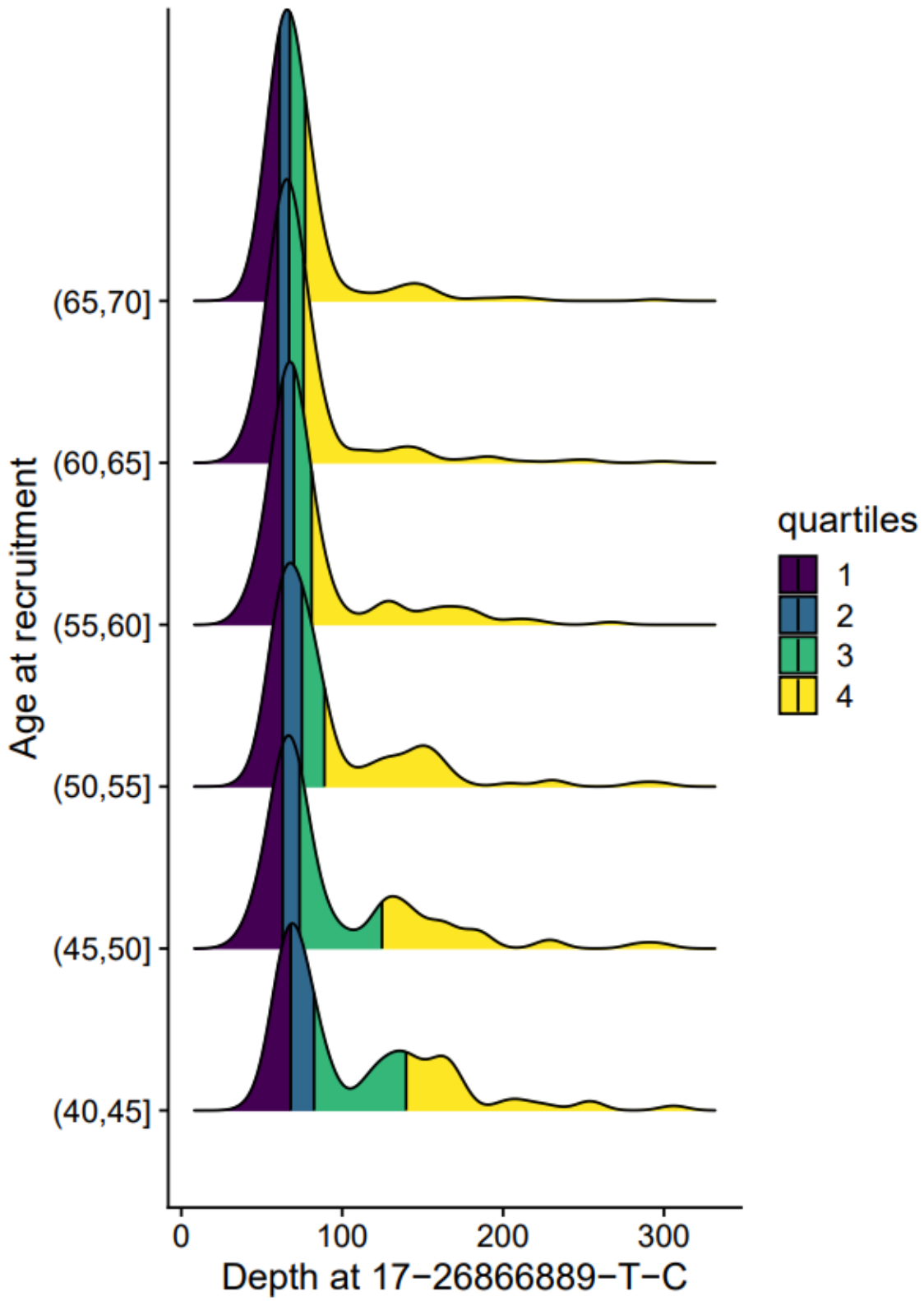
Supplementary Figure 2: View of IGH locus showing five age-associated point mutations. Track on the bottom displays cCRE elements from ENCODE for hematopoietic multipotent progenitor biosample ENCDO116AAA.



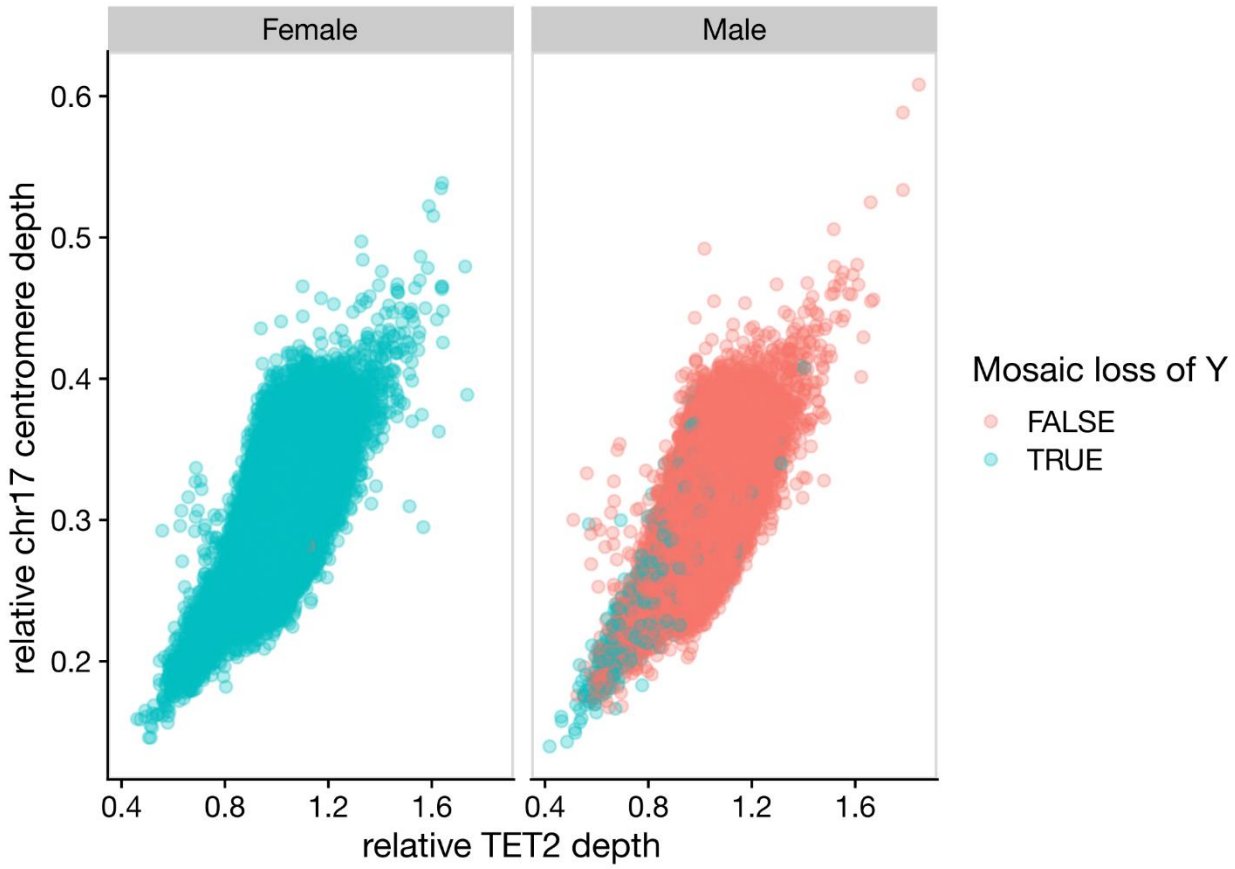
Supplementary Figure 3: X-axis indicates the frequency of point mutations in UKB. Y axis indicates the frequency of overlapping mCAs as reported in Loh et al., 2020.



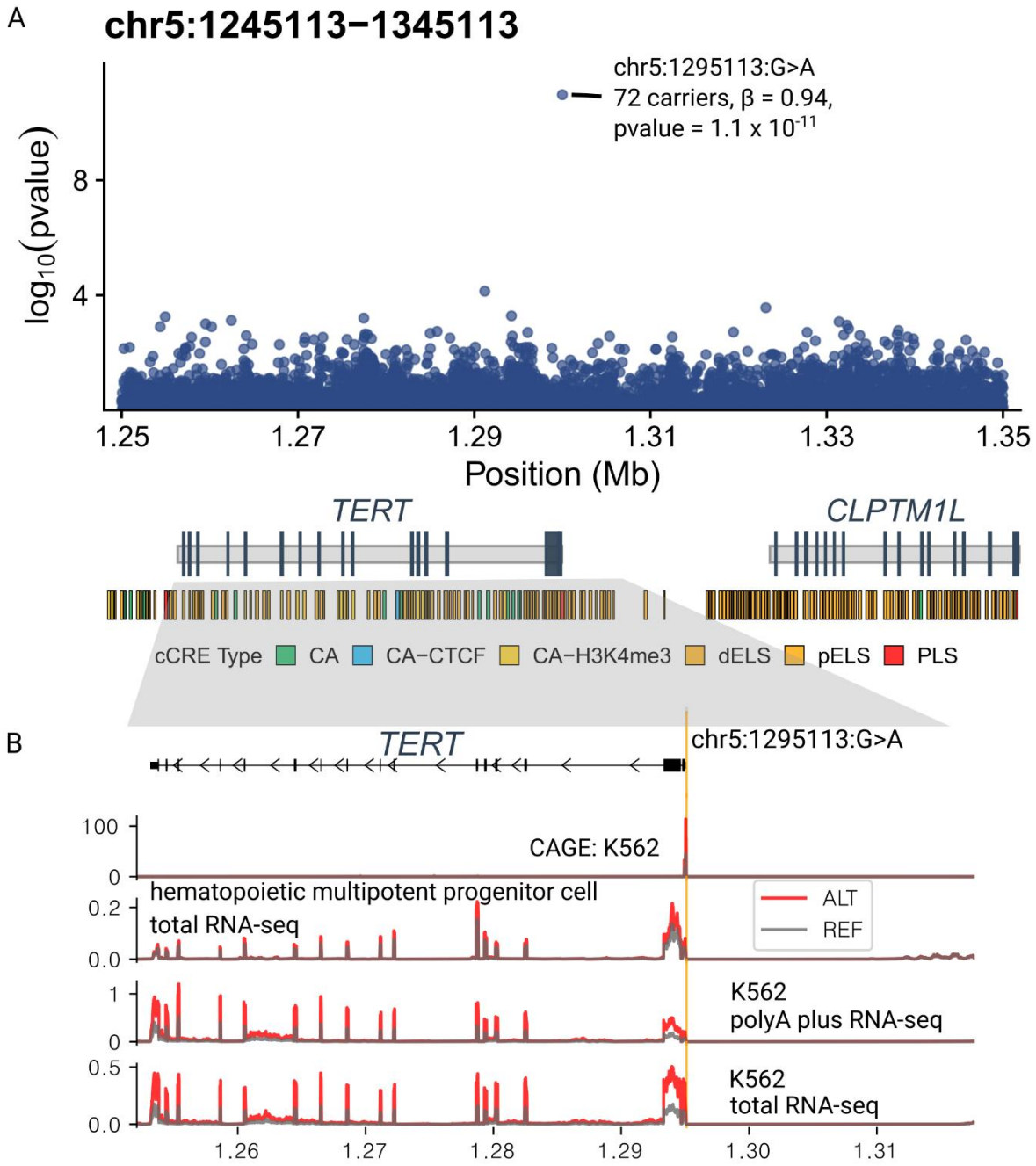
Supplementary Figure 4: Scatter plot of alternative allele frequencies for 45 age-associated variants across UK Biobank and TOPMed cohorts. Non-centromeric variants show strong concordance (ρ = 0.63, p = 0.008), while centromeric variants exhibit higher frequencies in TOPMed.



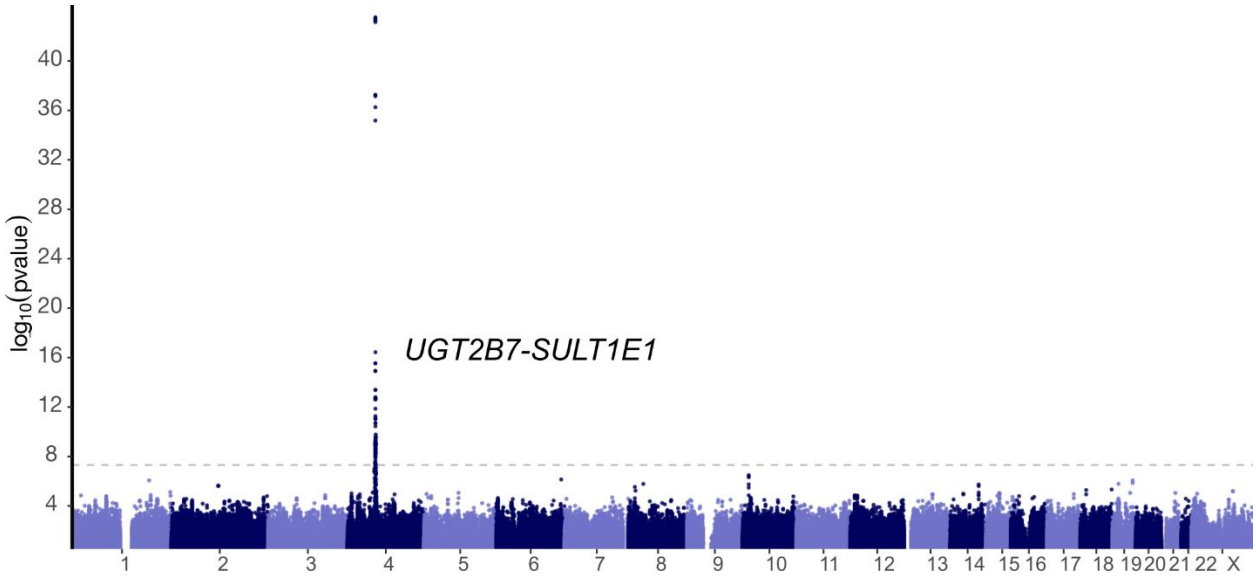
Supplementary Figure 5: Y-axis indicates the density of age at blood draw; colors indicate quartiles of the distribution. X-axis indicates the sample level sequencing depth at 17-26866889-T-C.



Supplementary Figure 6: Depth was quantified using Mosdepth at the chr 17 centromere, OC4, TET2, and sex chromosomes. Depths were then "normalized" by dividing by the depth at OCT4. Mosaic loss of Y was called in males by thresholding the relative Y chromosome depth below 0.369.



Supplementary Figure 7: A. Locuszoom of association between age at blood draw and variants (both germline + somatic) in the TERT locus. The age-associated putative driver, chr5:1295113:G>A is highlighted. B., AlphaGenome predicted RNA-seq tracks in the locus.



Supplementary Figure 8: GWAS was performed with REGENIE including all variants with MAF > 0.5%. A spline of age, sex, smoking history, and genetic ancestry PCs were included as covariates. Variants falling within contigs that differed between hg19 and hg38 were excluded. Variants falling within UCSC "unusual" regions and hg38 genome reference consortium "exclusions" list were also excluded to mitigate cross-mapping artifacts.

Works Cited

1. Mitchell, E. *et al.* Clonal dynamics of haematopoiesis across the human lifespan. *Nature* **606**, 1–8 (2022).
2. Jaiswal, S. *et al.* Age-related clonal hematopoiesis associated with adverse outcomes. *N Engl J Med* **371**, 2488–2498 (2014).
3. Xie, M. *et al.* Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nature Medicine* **20**, 1472–1478 (2014).
4. Genovese, G. *et al.* Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. *New England Journal of Medicine* **371**, 2477–2487 (2014).
5. Kar, S. P. *et al.* Genome-wide analyses of 200,453 individuals yield new insights into the causes and consequences of clonal hematopoiesis. *Nat Genet* **54**, 1155–1166 (2022).
6. Thompson, D. J. *et al.* Genetic predisposition to mosaic Y chromosome loss in blood. *Nature* **575**, 652–657 (2019).
7. Zhou, W. *et al.* Mosaic loss of chromosome Y is associated with common variation near TCL1A. *Nature Genetics* **48**, 563–568 (2016).
8. Jaiswal, S. *et al.* Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *New England Journal of Medicine* https://doi.org/10.1056/NEJMoa1701719 (2017) doi:10.1056/NEJMoa1701719.
9. Sano, S. *et al.* Hematopoietic loss of Y chromosome leads to cardiac fibrosis and heart failure mortality. *Science* **377**, 292–297 (2022).

590 10. Fuster, J. J. *et al.* Clonal hematopoiesis associated with TET2 deficiency accelerates
591 atherosclerosis development in mice. *Science* **355**, 842–847 (2017).
592 11. Bick Alexander G. *et al.* Genetic Interleukin 6 Signaling Deficiency Attenuates
593 Cardiovascular Risk in Clonal Hematopoiesis. *Circulation* **141**, 124–131 (2020).
594 12. Vlasschaert, C. *et al.* Clonal hematopoiesis of indeterminate potential is associated
595 with acute kidney injury. *Nat Med* **30**, 810–817 (2024).
596 13. Wong, W. J. *et al.* Clonal haematopoiesis and risk of chronic liver disease. *Nature* **616**,
597 747–754 (2023).
598 14. Beauchamp, E. M. *et al.* ZBTB33 Is Mutated in Clonal Hematopoiesis and
599 Myelodysplastic Syndromes and Impacts RNA Splicing. *Blood Cancer Discov*
600 https://doi.org/10.1158/2643-3230.BCD-20-0224 (2021) doi:10.1158/2643-3230.BCD-
601 20-0224.
602 15. Loh, P.-R. *et al.* Insights about clonal hematopoiesis from 8,342 mosaic chromosomal
603 alterations. *Nature* **559**, 350–355 (2018).
604 16. Terao, C. *et al.* Chromosomal alterations among age-related haematopoietic clones in
605 Japan. *Nature* **584**, 130–135 (2020).
606 17. Loh, P.-R., Genovese, G. & McCarroll, S. A. Monogenic and polygenic inheritance
607 become instruments for clonal selection. *Nature* **584**, 136–141 (2020).
608 18. Brown, D. W. *et al.* Shared and distinct genetic etiologies for different types of clonal
609 hematopoiesis. *Nat Commun* **14**, 5536 (2023).
610 19. Jakubek, Y. A. *et al.* Mosaic chromosomal alterations in blood across ancestries using
611 whole-genome sequencing. *Nat Genet* 1–8 (2023) doi:10.1038/s41588-023-01553-1.

20. Jakubek, Y. A. *et al.* Genomic and phenotypic correlates of mosaic loss of chromosome Y in blood. 2024.04.16.24305851 Preprint at https://doi.org/10.1101/2024.04.16.24305851 (2024).
21. Liu, A. *et al.* Genetic drivers and cellular selection of female mosaic X chromosome loss. *Nature* 1–8 (2024) doi:10.1038/s41586-024-07533-7.
22. Poon, G. Y. P., Watson, C. J., Fisher, D. S. & Blundell, J. R. Synonymous mutations reveal genome-wide levels of positive selection in healthy tissues. *Nat Genet* **53**, 1597–1605 (2021).
23. Scherer, M. *et al.* Clonal tracing with somatic epimutations reveals dynamics of blood ageing. *Nature* **643**, 478–487 (2025).
24. Bernstein, N. *et al.* Analysis of somatic mutations in whole blood from 200,618 individuals identifies pervasive positive selection and novel drivers of clonal hematopoiesis. *Nat Genet* **56**, 1147–1155 (2024).
25. Zeng, T., Spence, J. P., Mostafavi, H. & Pritchard, J. K. Bayesian estimation of gene constraint from an evolutionary model with gene features. 2023.05.19.541520 Preprint at https://doi.org/10.1101/2023.05.19.541520 (2023).
26. Rheinbay, E. *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, 102–111 (2020).
27. Weinstock, J. S. *et al.* The genetic determinants of recurrent somatic mutations in 43,693 blood genomes. *Science Advances* **9**, eabm4945 (2023).
28. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

29. Watson, C. J. *et al.* The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science* **367**, 1449–1454 (2020).
30. Weinstock, J. S. *et al.* Aberrant activation of TCL1A promotes stem cell expansion in clonal haematopoiesis. *Nature* **616**, 755–763 (2023).
31. Rosoff, D. B. *et al.* Multivariate genome-wide analysis of aging-related traits identifies novel loci and new drug targets for healthy aging. *Nat Aging* **3**, 1020–1035 (2023).
32. Niroula, A. *et al.* Distinction of lymphoid and myeloid clonal hematopoiesis. *Nat Med* **27**, 1921–1927 (2021).
33. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
34. Avsec, Ž. *et al.* AlphaGenome: advancing regulatory variant effect prediction with a unified DNA sequence model. 2025.06.25.661532 Preprint at https://doi.org/10.1101/2025.06.25.661532 (2025).
35. Liu, Y. *et al.* ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *Am J Hum Genet* **104**, 410–421 (2019).
36. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
37. Vlasschaert, C. *et al.* A practical approach to curate clonal hematopoiesis of indeterminate potential in human genetic data sets. *Blood* **141**, 2214–2223 (2023).
38. Codd, V. *et al.* Measurement and initial characterization of leukocyte telomere length in 474,074 participants in UK Biobank. *Nat Aging* **2**, 170–179 (2022).

39. McLoughlin, M. A. *et al.* Telomere attrition becomes an instrument for clonal selection in aging hematopoiesis and leukemogenesis. *Nat Genet* **57**, 2215–2225 (2025).

40. Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**, 1205–1210 (2010).

41. Stacey, S. N. *et al.* Genetics and epidemiology of mutational barcode-defined clonal hematopoiesis. *Nat Genet* 1–11 (2023) doi:10.1038/s41588-023-01555-z.

42. Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat Genet* **51**, 187–195 (2019).

43. Bick, A. G. *et al.* Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature* **586**, 763–768 (2020).

44. Kessler, M. D. *et al.* Common and rare variant associations with clonal haematopoiesis phenotypes. *Nature* 1–9 (2022) doi:10.1038/s41586-022-05448-9.

45. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet* **53**, 1097–1103 (2021).

46. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics* **50**, 1335–1341 (2018).

47. Bao, E. L. *et al.* Inherited myeloproliferative neoplasm risk affects haematopoietic stem cells. *Nature* **586**, 769–775 (2020).

48. Jones, A. V. *et al.* JAK2 haplotype is a major risk factor for the development of myeloproliferative neoplasms. *Nat Genet* **41**, 446–449 (2009).

677 49. Sinnott-Armstrong, N. *et al.* Genetics of 35 blood and urine biomarkers in the UK
678 Biobank. *Nature Genetics* **53**, 185–194 (2021).
679 50. Calado, R. T. *et al.* Sex hormones, acting on the *TERT* gene, increase telomerase activity
680 in human primary hematopoietic cells. *Blood* **114**, 2236–2243 (2009).
681 51. Dawoud, A. A. Z., Tapper, W. J. & Cross, N. C. P. Age-related loss of chromosome Y is
682 associated with levels of sex hormone binding globulin and clonal hematopoiesis
683 defined by TET2, TP53, and CBL mutations. *Sci Adv* **9**, eade9746 (2023).
684 52. Abelson, S. *et al.* Prediction of acute myeloid leukaemia risk in healthy individuals.
685 *Nature* **559**, 400–404 (2018).
686 53. Jakobsen, N. A. *et al.* Selective advantage of mutant stem cells in clonal hematopoiesis
687 occurs by attenuating the deleterious effects of inflammation and aging.
688 2023.09.12.557322 Preprint at https://doi.org/10.1101/2023.09.12.557322 (2023).
689 54. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and
690 exomes. *Bioinformatics* **34**, 867–868 (2018).
691 55. Stephens, M. False discovery rates: a new deal. *Biostatistics* **18**, 275–294 (2017).
692