medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

The Genetic Determinants and Genomic Consequences of Non-Leukemogenic Somatic Point Mutations

Authors: Joshua S. Weinstock[1,#], Sharjeel A. Chaudhry[2,3], Maria Ioannou[4], Maria Viskadourou[2], Paula Reventun[2], Yasminka A. Jakubek[5], L. Alexander Liggett[6], Cecelia Laurie[7], Jai G. Broome[8], Alyna Khan[9], Kent D. Taylor[10], Xiuqing Guo[10], Patricia A. Peyser[11], Eric Boerwinkle[12], Nathalie Chami[13,14], Eimear E. Kenny[15], Ruth J. Loos[13,14], Bruce M. Psaty[16,17,18], Tracy P. Russell[19], Jennifer A. Brody[16], Jeong H. Yun[20], Michael H. Cho[21], Ramachandran S. Vasan[22], Sharon L. Kardia[23], Jennifer A. Smith[23,24], Laura M. Raffield[25], Aurelian Bidulescu[26], Emily O'Brien[27], Mariza de Andrade[28], Jerome I. Rotter[10], Stephen S. Rich[29], Russell P. Tracy[19], Yii Der Ida Chen[10], C. Charles. Gu[30], Chao A. Hsiung[31], Charles Kooperberg[32], Bernhard Haring[33,34], Rami Nassir[35], Rasika Mathias[36], Alex Reiner[32], Vijay Sankaran[6], Charles J. Lowenstein[37], Thomas W. Blackwell[38], Goncalo R. Abecasis[38,39], Albert V. Smith[38], Hyun M. Kang[38], Pradeep Natarajan[40,41,42], Siddhartha Jaiswal[43], Alexander Bick[44], Wendy S. Post[37], Paul Scheet[45], Paul Auer[46], Theodoros Karantanos[4], Alexis Battle[47,48,49,#], Marios Arvanitis[2,47,#]

1 - Department of Human Genetics, School of Medicine, Emory University, Atlanta, GA, USA; 2 - Division of Cardiology, Department of Medicine, Johns Hopkins University, Baltimore, MD; 3 - Department of Surgery, Division of Vascular and Endovascular Surgery, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA; 4 - Division of Hematological Malignancies, Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine; 5 - Department of Internal Medicine, University of Kentucky; 6 - Division of Hematology/Oncology, Boston Childrens Hospital and Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115, USA; 7 - Department of Biostatistics, University of Washington, Seattle, WA 98195, USA.; 8 - Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA 98195, USA; 9 - Department of Biostatistics, University of Washington, Seattle, WA 98195, USA; 10 - The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA USA; 11 - Department of Epidemiology, School of Public Health, Boston University, Boxton, MA USA; 12 - Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA; 13 - The Charles Bronfman Institute of Personalized Medicine; 14 - The Mindich Child Health and Developlement Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 15 - Institute for Genomic Health; 16 - Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA; 17 - Department of Epidemiology, University of Washington, Seattle, WA, USA; 18 - Department of Health Systems and Population Health, University of Washington, Seattle, WA, USA; 19 - Department of Pathology & Laboratory Medicine and Biochemistry, Larner College of Medicine at the University of Vermont, Colchester, VT, USA; 20 - Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA USA; 21 - Channing Division of Network Medicine and Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, MA USA; 22 - National Heart Lung and Blood Institute's, Boston University's Framingham Heart Study, Framingham, MA, USA; 23 - Department of Epidemiology, University of Michigan, Ann Arbor, MI; 24 - Survey Research Center, Institute for Social Research, University of Michgian, Ann Arbor, MI; 25 - Department of Genetics, University of North Carolina, Chapel Hill, NC, 27514; 26 - Department of Epidemiology and Biostatistics, Indiana University School of Public Health Bloomington, Bloomington, IN, USA; 27 - Duke Clinical Research Institute, Durham, NC, USA; 28 - Mayo Clinic, Department of Health Sciences Research, Rochester, MN, USA; 29 - Department of Public Health Sciences, Center for Public Health Genomics,

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

University of Virginia, Charlottesville, VA USA; 30 - Center for Biostatistics and Data Sciences, Washington University, St. Louis, MO USA; 31 - Department of Medicine, Taipei Veterans General Hospital, Taipei Taiwan; 201 Shi-Pai Rd. Sec. 2, Taipei Taiwan; 32 - Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA; 33 - Department of Medicine III, Saarland University Hospital, Homburg, Saarland, Germany; Department of Medicine I, University of Wrzburg, Wrzburg, Bavaria, Germany; 34 - Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, New York, USA. Electronic address; 35 - University of California Davis, Davis, CA, USA; 36 - Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA; 37 - Department of Medicine, Cardiology Division, Johns Hopkins University; 38 - Center for Statistical Genetics, Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI, USA; 39 - Regeneron Pharmaceuticals, Tarrytown, NY, USA; 40 - Center for Genomic Medicine and Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA; 41 - Program in Medical and Population Genetics, Broad Institute of Harvard & MIT, Cambridge, MA; 42 - Department of Medicine, Harvard Medical School, Boston, MA; 43 - Department of Pathology, Stanford University, Stanford, CA, USA; 44 - Division of Genetic Medicine, Department of Medicine, Vanderbilt University, Nashville, TN, USA; 45 - Department of Epidemiology, University of Texas M.D. Anderson Cancer Center, Houston, TX, USA.; 46 - Department of Biostatistics, Medical College of WisconsinDivision of Biostatistics, Institute for Health and Equity, and Cancer Center, Medical College of Wisconsin, Milwaukee, WI, USA; 47 - Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA; 48 - Malone Center for Engineering in Healthcare, Johns Hopkins University, Baltimore, MD; 49 - Department of Computer Science, Johns Hopkins University, Baltimore, MD

Address correspondence to:

Marios Arvanitis (marvani1@jhmi.edu), Josh Weinstock (josh.weinstock@emory.edu), Alexis Battle (ajbattle@jhu.edu)

Abstract

Clonal hematopoiesis (CH) is defined by the expansion of a lineage of genetically identical cells in blood. Genetic lesions that confer a fitness advantage, such as point mutations or mosaic chromosomal alterations (mCAs) in genes associated with hematologic malignancy, are frequent mediators of CH. However, recent analyses of both single cell-derived colonies of hematopoietic cells and population sequencing cohorts have revealed CH frequently occurs in the absence of known driver genetic lesions. To characterize CH without known driver genetic lesions, we used 51,399 deeply sequenced whole genomes from the NHLBI TOPMed sequencing initiative to perform simultaneous germline and somatic mutation analyses among individuals without leukemogenic point mutations (LPM), which we term CH-LPMneg. We quantified CH by estimating the total mutation burden. Because estimating somatic mutation burden without a paired-tissue sample is challenging, we developed a novel statistical method, the Genomic and Epigenomic informed Mutation (GEM) rate, that uses external genomic and epigenomic data sources to distinguish artifactual signals from true somatic mutations. We performed a genome-wide association study of GEM to discover the germline determinants of CH-LPMneg. After fine-mapping and variant-to-gene analyses, we identified seven genes associated with CH-LPMneg (*TCL1A, TERT, SMC4, NRIP1, PRDM16*, *MSRA*, *SCARB1*), and one locus associated with a sex-associated mutation pathway (*SRGAP2C)*. We performed a secondary analysis excluding individuals with mCAs, finding that the genetic architecture was largely unaffected by their inclusion. Functional analyses of *SMC4* and *NRIP1* implicated altered HSC self-renewal and proliferation as the primary mediator of

medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

mutation burden in blood. We then performed comprehensive multi-tissue transcriptomic analyses, finding that the expression levels of 404 genes are associated with GEM. Finally, we performed phenotypic association meta-analyses across four cohorts, finding that GEM is associated with increased white blood cell count and increased risk for incident peripheral artery disease, but is not significantly associated with incident stroke or coronary disease events. Overall, we develop GEM for quantifying mutation burden from WGS without a paired-tissue sample and use GEM to discover the genetic, genomic, and phenotypic correlates of CH-LPMneg.

medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Introduction

As we age, our cells accumulate mutations. The vast majority of these mutations are inconsequential because they do not alter cell fitness. However, a small proportion of these mutations, termed drivers, can cause expansions of cell lineages they reside in. Recently, the age-related acquisition of leukemogenic point mutations (LPM) in whole blood, termed clonal hematopoiesis of indeterminate potential (CHIP), has been described as a prevalent aging-related phenomenon[1–4]. CHIP has previously been associated with increased risk for hematologic malignancy, cardiovascular disease, and increased mortality[4–6]. However, CHIP is a highly specific clonal phenomena defined as the presence of a driver mutation in 74 genes that have previously been associated with hematologic malignancy[7], which is a small proportion of the entire spectrum of somatic variation. Non-CHIP somatic variation in blood, which we term CH-LPMneg, has previously been shown to be a prevalent phenomenon, including mosaic chromosomal alterations (mCAs)[8–10], X-chromosome inactivation skewing[11], and even clonal expansions without known drivers[11,12]. However, the germline determinants and clinical consequences of CH-LPMneg remain uncharacterized.

We previously used the count of high variant-allele fraction (VAF) passenger mutations in 5,071 CHIP carriers in TOPMed to identify the genetic determinants of clonal expansion, an approach termed PACER[13], which uses age at blood-draw and passenger burden to infer the date at which a driver mutation was acquired. However, PACER is only defined for donors with a single driver mutation. Here, we seek to extend our inference of the sample level mutation burden for donors that may have no known driver point mutations. Non-CHIP clonal phenomena, which we refer to as CH-LPMneg, including mCAs, LOY, X-chromosome inactivation, and clonal expansions without known drivers have been previously associated with infection[14], hematologic malginancy[11], and heart failure[15], highlighting the value of quantifying CH-LPMneg.

The accurate detection of somatic mutations in CHIP non-carriers from a single whole-blood draw is likely to be more challenging than the identification in CHIP carriers because the passenger count no longer tracks the history of a single expanded clone. We reasoned that improved estimation of the somatic mutation rate would facilitate more accurate passenger burden inference in this more challenging setting. Previous reports have identified chromatin state as among the primary determinants of mutation rate. Indeed, Shuster-Bockler and Lehner[16] reported that variation in chromatin organization explains 55% of the variation in mutation rate. Thus, external epigenomic annotations are informative for estimating the likelihood that a candidate somatic variant call is a true mutation by altering the prior probability that a given variant call is accurate.

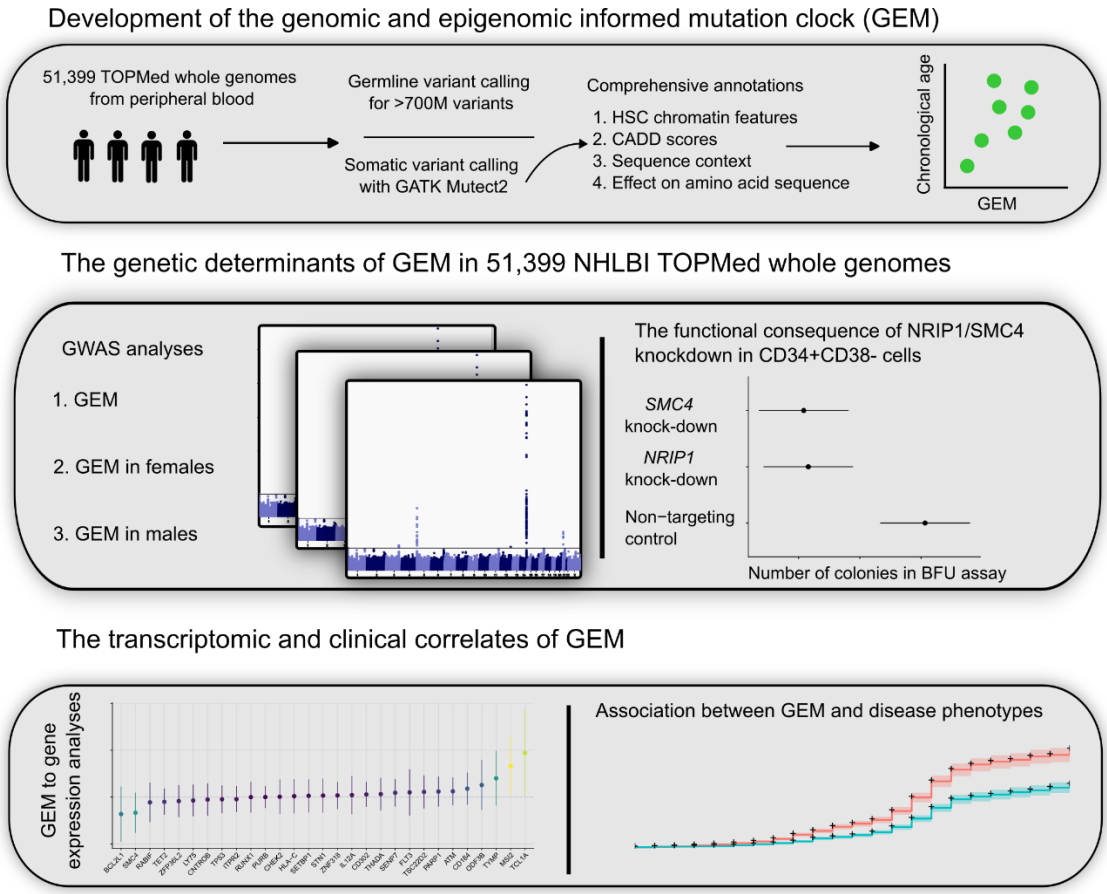
We and others have previously reported that CHIP and other clonal phenomena have germline genetic determinants[9,11,13,17–20]. CHIP has been previously associated with two primary pathways influencing HSC self-renewal and DNA damage pathways. Similarly, GWAS of mCAs have similarly reported associations with HSC self-renewal and DNA repair related loci[21–23]. A recent analysis that defined CH based on the dichotomization of low-VAF mutation burden, termed barcode-CH[24], observed several hits linked to both mCAs and CHIP. These analyses have demonstrated that germline variation is associated with acquired genetic variation and have demonstrated the utility of such analyses for discovering critical regulators of clonal expansion rate, including *TCL1A*.

Here, using 51,399 donors from NHLBI TOPMed consortium[25] without CHIP, we developed a mutation burden estimator, the Genomic and Epigenomic Mutation (GEM) rate (Figure 1). We then used this estimator as a phenotype to discover the genetic determinants of CH-LPMneg. In contrast to barcode-CH, this is a continuous phenotype that excludes individuals with CHIP mutations. This analysis revealed multiple novel loci, including the previously underappreciated role of *NRIP1*, a highly conserved

medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

transcriptional co-activator, in modulating mutation burden. We performed a sensitivity analysis, excluding all individuals with mCAs, finding that the bulk of our genetic discovery was unchanged, suggesting that CH-LPMneg signals are not merely mediated through mCAs. Fine-mapping of the *TCL1A* locus revealed a more complex cis-regulatory architecture than observed in CHIP. Functional characterization of *SMC4* and *NRIP1* with colony forming unit (CFU) assays revealed convergent effects on HSC self-renewal as the primary mechanism of these genes. Sex-stratified analyses of GEM revealed that the *TRIM59-KPNA4-SMC4* locus, which has previously been associated with CHIP and MPNs[17,26,27] is a female-specific signal, and a novel male-specific signal near *MSRA*. Principal component analysis of mutation burden revealed a sex specific mutation pathway. GWAS of this sex-specific mutation pathway identified a novel locus near *SRGA2PC*. Through transcriptomic analyses of blood and non-blood tissues, we identified the genomic consequences of elevated mutation burden in whole blood, which include the systematic down-regulation of the interferon-alpha pathway across hematopoietic lineages. Finally, we show that GEM is useful for predicting risk of incident peripheral artery disease, and associates with altered blood cell indices. Overall, we demonstrate a novel computational approach for quantifying mutation burden, which enabled the discovery of novel genetic determinants of CH-LPMneg.

Figure 1: Study design schematic, describing the development of GEM, the use of GEM to discover the genetic determinants of mutation burden in blood, and the use of GEM to identify the transcriptomic and clinical correlates of mutation burden in blood.



Results

Using 51,399 WGS samples from NHLBI TOPMed (Supplementary Tables 1-2), we first called candidate somatic variants using Mutect2 as previously described[17]. We then performed stringent filtering, including filtering known germline variants and likely sequencing artifacts (Methods). As

medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

distinguishing somatic variants from germline variants in single-tissue variant calling is challenging, we then took careful measures to determine the optimal alt-allele threshold. We observed that excluding variants with higher alt-alleles substantially improved the association of the burden of such mutations with chronological age, suggesting that a stricter alt-allele threshold than we previously applied in the PACER pipeline is useful for excluding germline variation.

Genomic and Epigenomic Annotations Inform Mutation Rate

Next, as chromatin state is among the primary mediators of mutation rate[16], we sought to determine the association between mutation burden and several genomic and epigenomic annotations. We used chromHMM[28] annotations in CD34+ cells from the Roadmap Epigenomic[29] resource as a measure of chromatin state in HSCs, which previous analyses have reported as the causal cell type in clonal phenomena[27]. We calculated the mutation burden stratified by chromatin annotation and examined the association with age, reasoning that the strength of association between mutation burden and chronological age would reflect the proportion of artifactual mutations. We observed that mutations in quiescent chromatin (Figure 2A) are much more strongly associated with age than mutations in transcriptionally active chromatin, recapitulating the role of chromatin in modifying mutation rate.

As mutations that are functional and are not mutated at the stem cell level undergo extensive negative selection[30,31], we then asked whether mutation burden stratified by functional consequence on protein coding sequence modified the association with chronological age. We observed that mutation burden of missense and UTR variants was much more weakly associated with chronological age than mutation burden from intronic and intergenic mutations (Figure 2B), lending credence to our hypothesis that functional consequence is informative for refining mutation burden estimates. We performed a similar analysis with stratified CADD scores, finding that the mutations in the highest quintile of CADD scores were the most weakly associated with age (Figure 2C), again suggesting that deleterious mutations are depleted of association with age and likely enriched for false positives. We performed an analyses stratified by both CADD quantile and chromHMM annotation, finding that the two were not redundant (Extended Data Figure 1). Collectively, these analyses suggest that variant deleteriousness and chromatin annotations are useful for the construction of a mutation derived molecular clock.

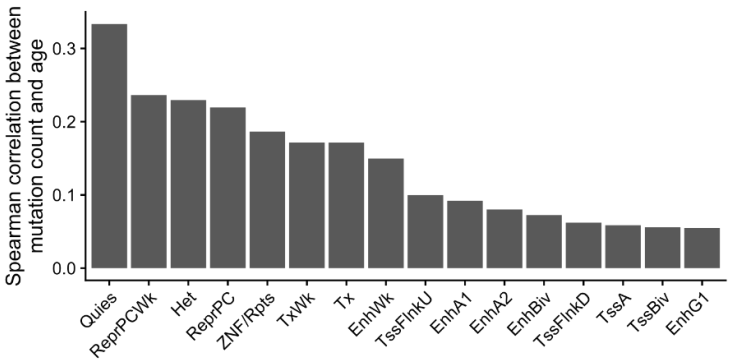
We then developed a weakly-supervised probabilistic graphical modeling approach that incorporates genomic and epigenomic annotations to distinguish somatic mutations from artifacts, a method we term GEM (genomic and epigenomic mutation rate). Weakly-supervised probabilistic graphical modeling approaches have been previously used to identify functional rare-variants[32], demonstrating the utility of such approaches towards annotation of genetic variation. We first comprehensively annotated all candidate somatic variants based on their chromHMM[28] annotations, their functional consequence, CADD[33], their population allele frequency in TOPMed, surrounding sequence context, among others (Methods). GEM uses chronological age as an external annotation to identify which candidate somatic mutations were functional based on their annotations (Figure 2D). GEM enables the classification of candidate somatic mutations as either truly somatic or artifacts, thus increasing power in downstream analyses by facilitating the depletion of likely artifactual variants.

Figure 2: Development of GEM | A, The Spearman correlation between mutation burden and chronological age stratified by chromHMM annotations in CD34+ cells. B, The Spearman correlation between mutation burden and chronological age stratified by functional consequence as annotated by the variant effect predictor (VEP). C, The Spearman correlation between mutation burden and chronological age, stratified by quintiles of CADD scores. D, Plate annotation for the GEM statistical model. θ_0 and

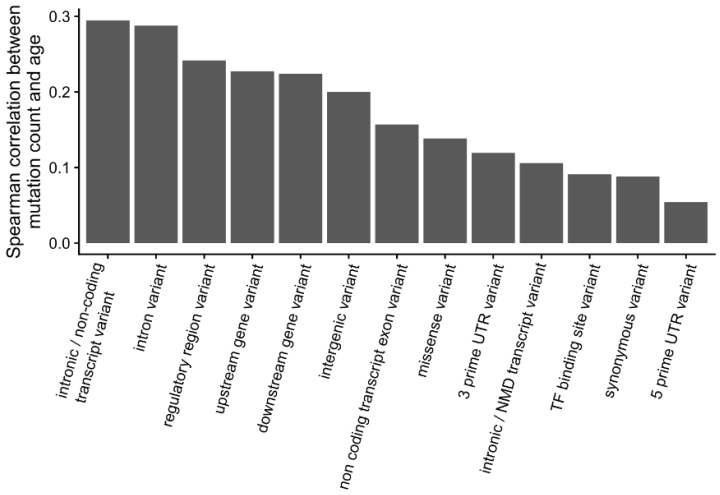
medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

are intercepts; θ_1 reflects the association between log2 transformed value of $\sum z_{ij}$ and chronological age Y_i; z_{ij} *denotes the probability that the* j*th mutation in the* i*th individual is a true somatic mutation. X is a matrix of annotations.*

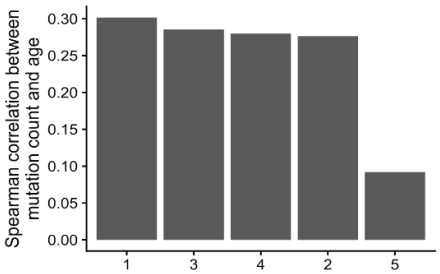
A Mutations in quiescent CD34+ chromatin are enriched for association with age



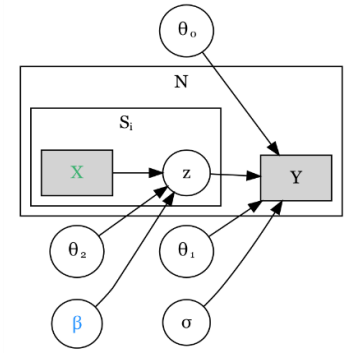
B Intronic and intergenic mutations are enriched for association with age



C Deleterious mutations are depleted of association with age



D The genomic and epigenomic mutation rate (GEM)



The Genetic Determinants of Mutation Burden

Next, as clonal phenomena have been shown to have germline genetic determinants, we performed a GWAS with GEM as the phenotype in 51,399 carriers of diverse ancestry. We computed summary statistics using SAIGE[34]. We detected six genome-wide significant loci, including *TERT, TCL1A,*

medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

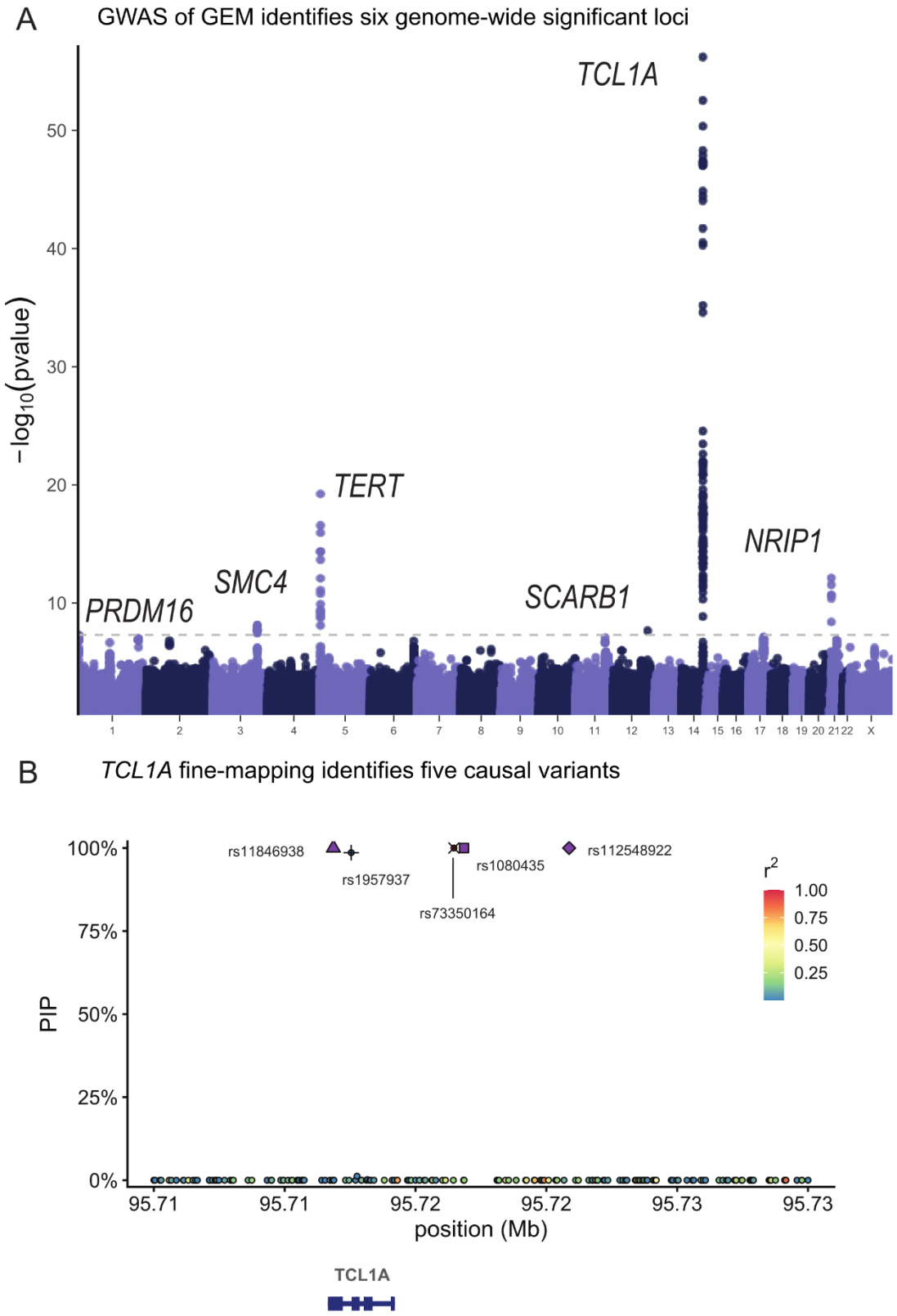
TRIM59-SMC4-KPNA4, NRIP1, SCARB1, and PRMD16 (Figure 3A), and an overall h^2_{SNP} of 9.3%. We then performed a similar analysis based on the burden of mutations in quiescent chromatin and heterochromatin, which resulted in reduced power at *TCL1A* (GEM minimum pvalue of 6×10^{-57} vs mutation burden minimum pvalue of 6×10^{-50}), demonstrating the value of the GEM over simply using mutation burden stratified by chromatin context (Extended Data Figure 2). *TERT*, *TCL1A*, *TRIM59-SMC4-KPNA4* have been previously associated with CHIP[17,18,35]and barcode-CH[24], while *NRIP1* and *PRDM16*, *TERT*, and *TCL1A* have all been associated with mCAs[21–23] and barcode-CH[24]. *SCARB1* has not been previously reported with a related phenotype. To nominate causal SNPs and genes, we fine-mapped each locus using SuSIE[36] and cross-referenced the credible sets with the Open Targets V2G estimates[37] and cell-type specific enhancers-gene pairs from the activity by contact model[38]. These signals collectively highlight the convergence of germline variation influencing CHIP, mCAs, and clonal hematopoiesis without known drivers.

We then asked whether the association between these GWAS loci and GEM was mediated through an association with mCAs. Using a recently developed atlas of mCAs in TOPMed[20], we excluded all mCAs or LOX carriers (n = 13,399) and performed another GWAS as a sensitivity analysis. After filtering to variants that were genome-wide significant in either GWAS, we observed that the effect sizes were remarkably concordant (R^2 = 0.997, Extended Data Figure 3), suggesting that the GWAS of GEM is not merely mediated by the effect of mCAs/LOX.

The lead variant at *TCL1A* was rs2887399, which we previously discovered as among the primary mediators of clonal expansion in CHIP carriers[39], and has been previously reported in GWAS of LOY[21,22]. Fine-mapping[36] revealed five credible sets, each with a single causal variant (rs11846938, rs112548922, rs1080435, rs1957937, and rs73350164, all PIP > .98). V2G[37] identified *TCL1A* as the mostly likely causal gene at each of the five causal SNPs. rs11846938 is 10bp from and in very high LD with rs2887399 (EUR R^2 = 0.88, AFR R^2 = 0.92), and both reside in the core promoter of *TCL1A*. Given high LD between rs2887399 and rs11846938, and the functional evidence we previously observed for the effect of rs2887399, we refer to this signal as "rs2887399/rs11846938." rs112548922 is 9kb upstream of *TCL1A*, suggesting that cis-regulatory elements besides the core promoter are implicated in altered mutation burden. We previously described a CHIP-mutation specific mechanism, whereby mutations in *TET2/ASXL1/SF3B1* are associated with the aberrant chromatin opening of the *TCL1A* promoter in HSCs, but not *DNMT3A* mutations. This led us to hypothesize that rs2887399 is chromatin accessibility-QTL (caQTL) that is a *TET2/ASLX1/SF3B1* mutation (and possibly LOY) specific and thus the risk allele is more likely to lead to the aberrant activation of proto-oncogene *TCL1A* only when a mutation sufficient for chromatin modification at the *TCL1A* promoter has been acquired. Indeed, rs2887399 has been since reported as a caQTL[40] in lymphoblastoid cell lines (LCL). We previously showed that *TCL1A* expression is sufficient for promoting clonal expansion and altering stress response in HSCs[39] and is a key regulator of clonal expansion in CHIP clones. The discovery of additional casual variants at the *TCL1A* locus highlights the utility of applying GEM to samples unascertained for specific genetic lesions, providing increased power for genetic discovery and indicates that the context specific up-regulation of *TCL1A* is a broader phenomenon than previously appreciated, and likely occurs in HSCs without mCAs, LOY, or CHIP, possibly mediated through stochastic epigenetic phenomena that result in increased accessibility at cis-regulatory elements of *TCL1A.*

medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 3: The genetic determinants of GEM. A, The GWAS of GEM. Summary statistics were estimated with SAIGE. B, Fine-mapping of the TCL1A locus. Note rs11846938 is 10bp from rs2887399. Fine-mapping was performed with SuSIE.



medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

The lead variant at *NRIP1* is rs2229742, a common (MAF = 6%) missense variant (p.Arg448Gly) predicted to be deleterious by SIFT[41]. *NRIP1* has previously been discovered in the context of mCAs[14] and barcode-CH[24] but has not been discovered in CH-LPMneg. Fine-mapping of the locus identified one credible set containing three variants (rs2229742, rs2823020, rs2823025). The C allele of rs2229742 is associated with increased GEM burden (beta = 0.08 standard deviations, 72% PIP, pvalue = 7.6×10^{-13}). *NRIP1* is a highly conserved (pLI[42] = 0.99) transcription co-regulator that is highly expressed in HSCs[43] and has been previously reported as a positive regulator of stemness in HSCs[44]. rs2229742 is strongly associated with multiple blood cell index GWAS [45] indicating that altered protein sequence of *NRIP1* results in altered HSC function. *NRIP1* ablation has previously been shown to extend lifespan in murine models[46], indicating a role in aging related phenotypes, although this is reported to be likely mediated through the interaction between *NRIP1* and estrogen signaling rather than modulation of HSC function. As the C allele is associated with increased GEM, this suggests that the C allele may either increase function of the *NRIP1* product or increase the abundance of *NRIP1* through indirect mechanisms, perhaps through increased translation efficiency. To elucidate the consequences of altered amino acid sequence in *NRIP1,* we cross-referenced a recently released catalogue of *trans*-pQTLs from plasma[47]. rs2229742 is a trans-pQTL for both SDC4 (beta = 0.09, pvalue = 3.8×10^{-13}) and PGLYPR2 (beta = 0.07, pvalue = 8.7×10^{-14}). SDC4 is a syndecan, which are cell-surface proteins that can interact with a broad range of ligands. The mouse-genome informatics resource[48] reported that *SDC4* ablation in mice led to several altered hematopoietic phenotypes. *PGLYRP2* is a peptidoglycan recognition protein that has been implicated in interferon regulation and innate immune response[49].

Previous GWAS of clonal phenomena, including CHIP and MPNs have reported the *TRIM59-SMC4-KPNA4* locus, though none have conclusively identified the causal gene. Fine-mapping this locus identified a credible set containing 19 variants. The three variants with the highest PIP were rs11718121 (PIP = 9.2%), rs1451760 (PIP = 8.8%), and rs6790951 (PIP = 8.2%). V2G estimated that *SMC4* was the mostly likely causal gene for each of these three variants, and alt-alleles at the three variants were associated with increased expression of *SMC4* in eQTLGen in whole blood[50] and increased expression of *SMC4* in lipopolysaccharide stimulated monocytes[51]. The interval spanned by the credible set contains a predicted *SMC4* enhancer in CD34+ cells by the ABC model. *SMC4* is a sub-unit of the condensin complex, which is involved in chromosome assembly and segregation during mitosis. Collectively, fine-mapping, V2G estimates, and the ABC model nominate *SMC4* as the mostly likely causal gene in the locus, highlighting the role of *SMC* related proteins in modulated mutation burden.

Fine-mapping of the *SCARB1* locus identified one credible set with a single SNP, rs11057853, a common (MAF = 46%) variant intronic to *SCARB1*. V2G estimated that *SCARB1* was the most likely causal gene for rs11057853, supported by both its presence within the *SCARB1* gene body and its role as an eQTL for *SCARB1* in blood[50]. The C allele was associated with reduced GEM (beta = -0.03 GEM standard deviations) and reduced expression of *SCARB1* (beta = -0.23, pvalue = 5.5×10^{-175}), suggesting that *SCARB1* may be protective against CH-LPMneg. *SCARB1* is a receptor for HDL and rare variant burden tests of *SCARB1* in UK Biobank have identified several associations with lipid traits[52]. Previous reports have described a possible role for *SCARB1* in mediating the metabolic adaptation of long-term HSCs using murine models[53].

We then asked whether rare variants are associated with GEM. We performed a genome-wide RVAS using STAAR[54], including all missense and loss of function (LOF) variants within protein-coding genes. We identified 33 and 18 hits at pvalue thresholds of 5×10^{-6} and 5×10^{-7} (Supplementary table 3). The strongest hit was *CELF2* (pvalue = 3.4×10^{-28}), where coding variants were associated with a higher GEM value among carriers (mean of 0.38, 95% CI: [0.36, 0.40]) than non-carriers (mean of 0.00, 95% CI: [-0.01, 0.01]). *CELF2* is a highly constrained (pLI[42] = 1.00) RNA binding protein and is highly expressed in

medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

neutrophils[37]. A recent report described the role of *CELF2* as a suppressor of the AKT/PI3K signaling pathway[55] in lung carcinoma, which is presumably the signaling pathway mediating the effect of *TCL1A.*

We performed an non-coding RVAS using SCANG[56], a dynamic window approach for identifying sets of SNPs that associate with phenotypes. We applied to SCANG to 100kb regions flanking 1,688 genes that Open Targets[37] has identified as previously associated with cancer (Supplementary table 4). We identified 52 and 6 hits at pvalue thresholds of 5×10^{-6} and 5×10^{-7} (Supplementary table 5). We observed a locus on chr14 flanking the MARK3 gene which was strongly associated with mutation burden (2.5×10^{-8}). To identify the likely causal variants within this window, we then performed a joint analysis of all rare variants that were included. This analysis highlighted three variants as significantly associated including rs190231639, a rare (MAF = 0.04%) variant intronic to *COA8*. To identify the likely causal gene in this gene dense locus, we cross-referenced the V2G results from Open Targets, which nominated *COA8*, *KLC1*, *XRCC3*, and *ZFYVE21* as equivalently likely causal genes. *XRCC3* is involved in the homologous recombination repair pathway of double-stranded DNA, though we are unable to conclusively identify it as the causal gene. Collectively, we identify diverse signals among the rare variants implicating DNA repair and post-translational modifications as key molecular processes contributing to GEM variation.

Mutation Burden Is Indirectly Regulated by the Size of the HSC Pool

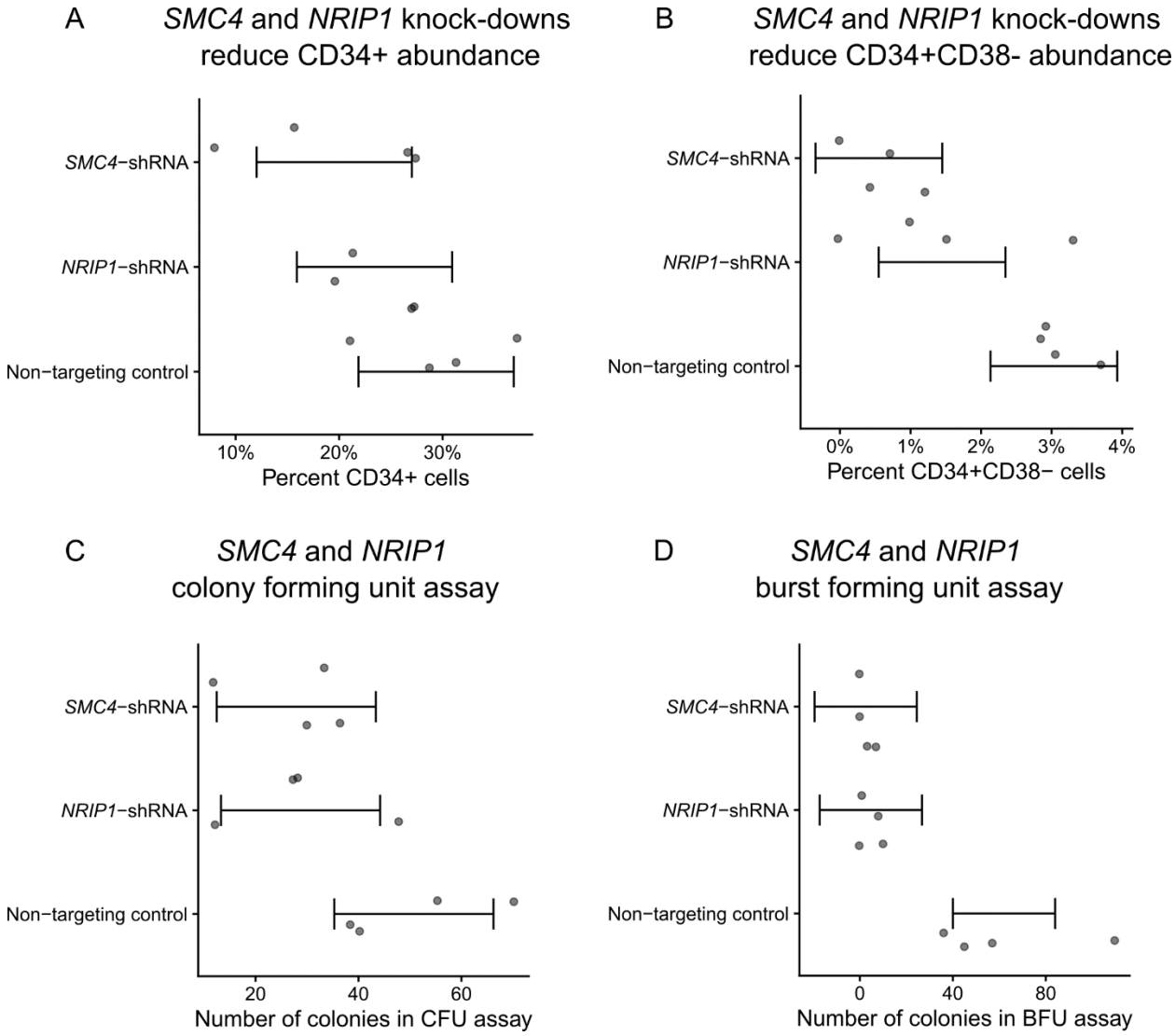
We then asked whether *SMC4* and *NRIP1* contributed to mutation burden by altering HSC self-renewal. To test this hypothesis, we separately knocked down *SMC4* and *NRIP1* in CD34+ bone-marrow derived human HSCs with shRNAs and performed a colony forming unit (CFU) assay. Relative to a non-targeting shRNA control, both *SMC4* and *NRIP1* knockdown cells were more likely to lose stemness in culture (1.6% and 2.5% fewer CD34+CD38- cells, pvalues = 2.0×10^{-3}, 1.7×10^{-3}, Figure 4A-B) and the cells formed fewer burst-forming unity colonies (22.0 and 22.8 fewer colonies, pvalues = 2.4×10^{-3}, 1.9×10^{-3}, Figure 4C-D). These results are consistent with a role of both *SMC4* and *NRIP1* as positive regulators of CD34+ HSC self-renewal, and led us to propose the following mechanisms for their roles as indirect regulators of mutation burden: Either *SMC4/NRIP1* may regulate the fitness of HSC in response to noxious stimuli such as the infection inherent to the shRNA knockdown assay, or *SMC4/NRIP1* directly regulates the size of the HSC pool; a larger active HSC pool increases the likelihood that at least one HSC obtains a fitness advantage through either a genetic lesion or stochastic epigenetic phenomena, leading to a clonal expansion, which will increase the passenger count burden. Similar models have been previously been proposed in the context of myeloproliferative neoplasms[27].

We then sought to explore the causal relationship between HSC pool size and GEM through simulation using a stochastic process that describes realistic HSC population (Methods). We simulated several HSCs which acquire passenger mutations at a constant rate per cell through a Poisson point process. We simulated the size of individual HSC clone populations using a Poisson birth-death process, where a single parameter s governs the relative likelihood of an HSC self-renewing into two identical HSCs as opposed to dividing into two differentiated cell types. At a rate of 1 driver mutation per 10,000 HSCs per year, we simulated modest increases to s in each HSC to model modest increases in cell fitness that some clones may acquire. We stratified these simulations across varying initial sizes of the HSC pool, finding that larger pools were much more likely to contain at least one clone with a substantial increase in fitness (Extended Data Figure 4) and many more high-VAF passengers (Extended Data Figure 5). Importantly, the burden of high VAF passengers increased as the number of increases to s increased. Taken together, this model provides a formal exposition for why GEM may be associated with both the

medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

overall HSC pool and the likelihood of at least one clone obtaining a substantial increase in self-renewal capacity.

Figure 4: The functional consequence of SMC4 and NRIP1 on HSCs. A, SMC4 and NRIP1 were knocked-down with shRNA and the proportion of CD34+ cells was quantified with FACs. Quantities were compared referent to a non-targeting control. B, proportion of CD34+CD38- was quantified with FACS. C, Number of colonies formed in a colony-forming unit (CFU) assay. D, Number of colonies in a burst-forming unit assay.



Regional Mutation Burden and the Genetic Determinants of Sex Specific Mutation Pathways

We then sought to perform more granular analyses of mutation burden, examining heterogeneity by sex and position in the genome. As the importance of mutation burden may vary based on genome position, we estimated the mutation burden in 49 non-overlapping intervals approximately 5 x 10^7 bases in length. We then estimated the association between age and mutation count stratified by interval, indicating heterogeneous associations across the intervals (Extended Data Figure 6). To characterize the underlying structure, we then performed PCA on these mutation counts.

medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

We observed that PC1 was strongly associated with overall mutation burden and explained 60% of the variance, revealing a single general factor associated with mutation burden genome-wide. We observed that the loadings of PC2 were enriched for mutations appearing chromosome-X, suggesting a sex-specific mutation pathway. We observed that PC2 is significantly associated with genotype inferred sex (R2 = 7.3%, pvalue < 2.2×10^{-16}), highlighting a sex-specific contribution to somatic variation in whole blood.

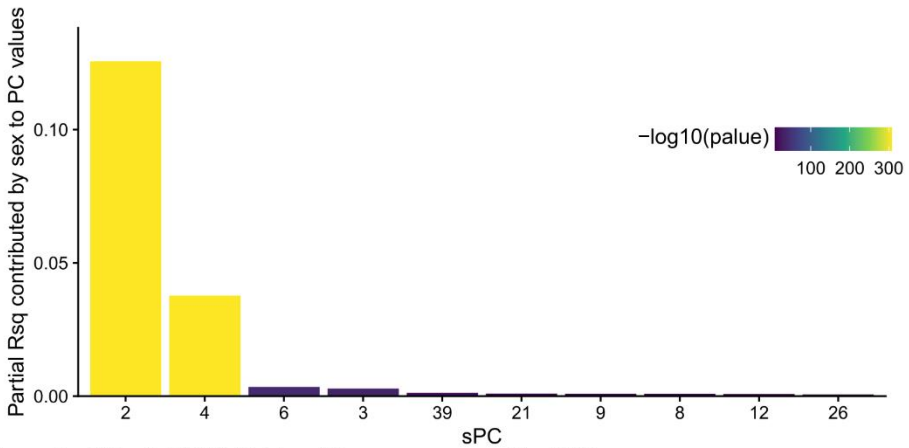
We then asked whether this sex-specific mutation factor had distinct germline determinants from GEM. We observed that a single locus on chromosome 1 near *SRGAP2C* was associated with PC2. The lead variant at this locus is rs61804016, a common variant 62kb away from the transcription start sites of *SRGAP2C* that has been previously reported as an eQTL in whole blood for *SRGAP2C*, *NBPF8*, *NBPF26*, *PFN1P2*, and *SRGAP2C*. However, in monocytes and T-cells, rs61804016 is only an eQTL for *SRGAP2C*[57]. The eQTL associations and proximity to the TSS of *SRGAP2C* suggest that *SRGAP2C* is the most likely causal gene in the locus. *SRGAP2C* is a GTPase activating protein that is expressed in hematopoietic progenitor cells[43]. We then cross-referenced phewas[58] results in UK Biobank[59] and Finngenn [60]. The C allele of rs61804016 is nominally associated with increased risk for breast cancer (odds-ratios of 1.08, 1.07, pvalues of 1.2×10^{-5}, 3.2×10^{-6}), further supporting the sex-specific nature of PC2. *SRGAP2C* is on chromosome 1, suggesting sex-specific regulation of an autosomal gene in the genesis of sex-specific mutation burden. No SNP near *SRGAP2C* was associated at genome-wide significance with the GEM phenotype. These analyses highlight the value of subtyping in mutation burden estimation by revealing sex-specific factors.

Given findings of sex-specific mutation pathways, we then performed sex-stratified GWAS of GEM (Figure 5), which revealed two sex specific signals. At the *SMC4* locus, rs11718121 was associated with increased mutation count in females (beta = 0.043, pvalue = 5.1×10^{-9}) but much more weakly associated in males (beta = 0.019, pvalue= 0.038). At a locus not identified in the standard analysis, near *MSRA*, rs117344298 was associated with decreased mutation count in males (beta = -0.17, pvalue = 2.4×10^{-8}), but unassociated in females (beta = -0.013, pvalue = 0.58). Other signals were largely shared between males and females. Cross-referencing of sex-biased eQTLs reported in GTEx[61] found that *MSRA* had nominally significant sex-biased eQTLs in tibial nerve tissue and Brain cortex, but not in whole blood. *SMC4* did not have sex-biased eQTLs, which may be the result of limited power to detect such effects.

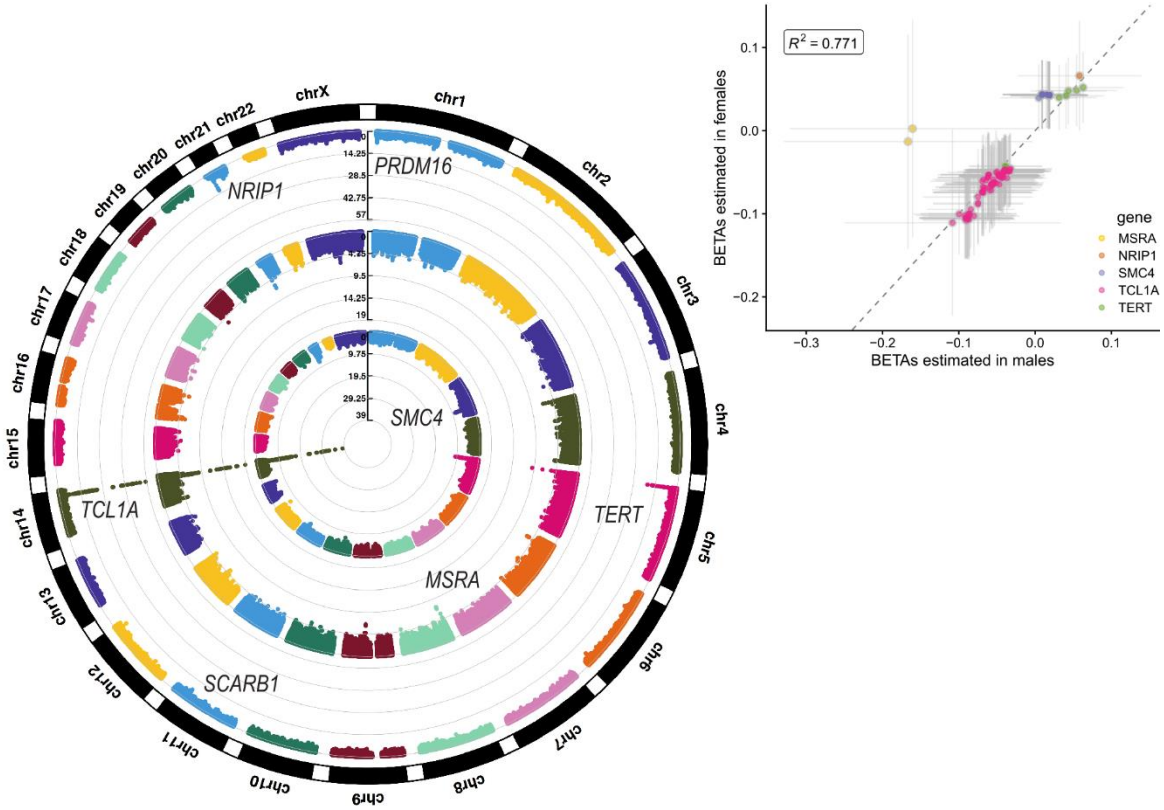
medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 5: The sex specific genetic determinants of mutation burden. A, Regressions were performed for each quantile-transformed somatic principal component (sPC) on study and sex as covariates. The partial variance explained by sex is displayed on the y-axis. B, Circular Manhattan plot. The outer-most ring is the GWAS of GEM on all individuals, the middle ring is the GWAS of GEM on males, and the inner-most ring is the GWAS of GEM in females. Inset, a scatter plot of the two sex-specific GWAS plotting all SNPs with pvalues < 1 x 10^{-8} in either GWAS. Asymptotic confidence intervals are plotted with a width corresponding to genome-wide significance.

A Somatic PC2 is associated with genotype inferred sex



B Sex stratified GWAS identifies sex-specific CH genes



medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

The Transcriptomic Correlates of High Mutation Burden

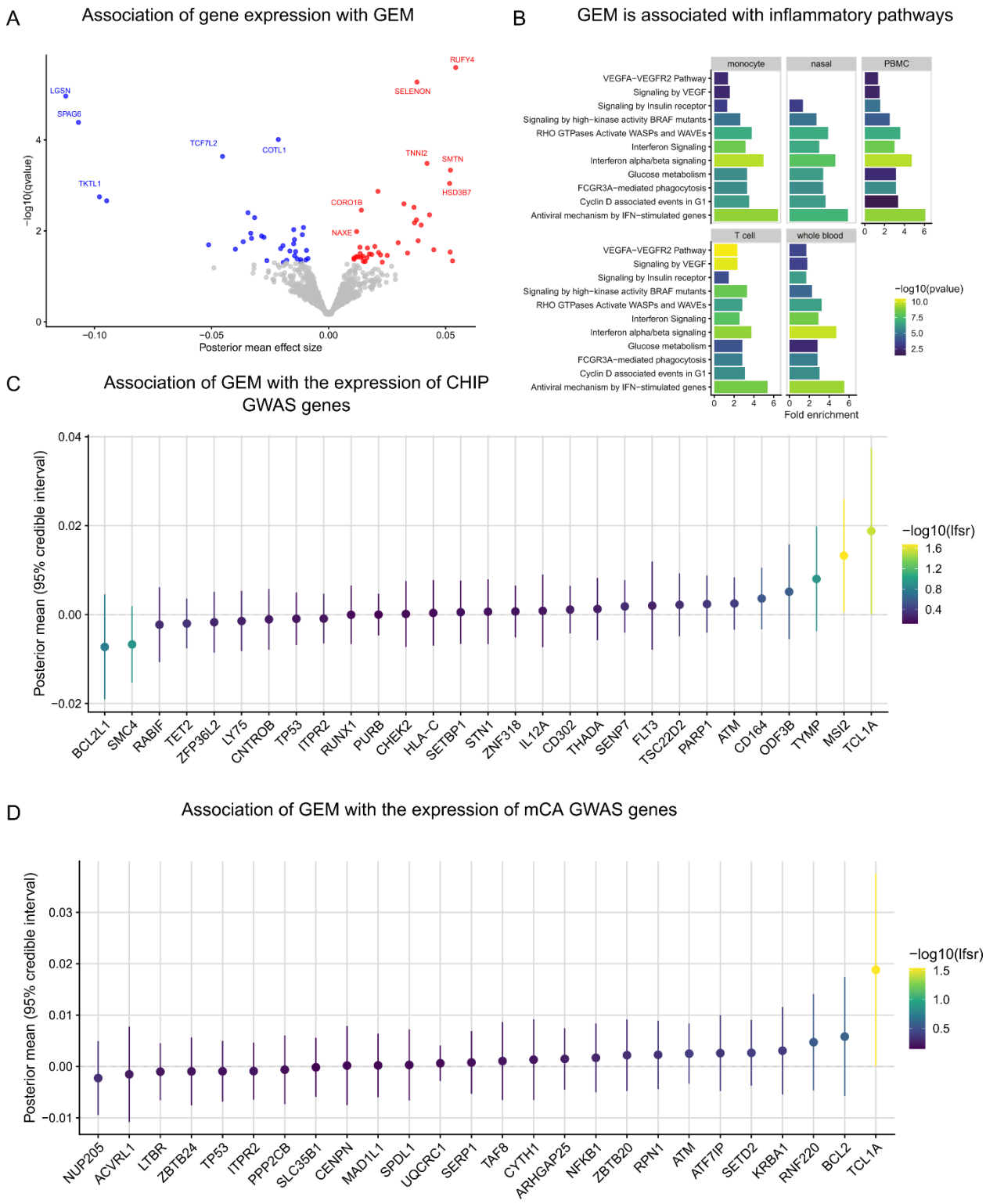
Next, we asked whether GEM associated with altered gene expression across five tissue types available in TOPMed, including whole blood, PBMC, monocytes, T cells, and nasal epithelial cells. We performed a search for GEM-gene associations by regressing the inverse normal transformed expression values of each gene (n = 17,741) on the inverse normalized GEM estimates, including age, genotype inferred sex, 15 genotype PCs, 20 expression PCs, and cohort indicators as covariates. To increase power, we then used mashr [62] to apply shrinkage across the 88,705 GEM-gene associations. We identified 404 GEM-gene associations at a local false sign rate (lfsr)[63] < 0.05 (Supplementary Table 6). Within whole blood, we observed the up-regulation of *RUFY4* with increased GEM. *RUFY4* is highly expressed in dendritic cells[37] and is involved in response to the anti-inflammatory cytokine IL-4[49] (Figure 6A). We also observed the down-regulation of *LGSN*, which although annotated for its role in differentiation cells in the lens, is highly expressed in HSCs and was recently reported as a candidate causal gene in asthma[64].

To characterize the gene programs associated with GEM, we performed pathway enrichment analyses using KEGG[65] as a reference. We observed a striking down-regulation of genes involved in interferon signaling (Fig. 6B). Interferon-alpha is a cytokine with anti-proliferative properties that was previously considered as a candidate therapeutic for AML[66]. Interferon-alpha is thought to reduce clonal expansion through direct and indirect mechanisms, including inducing apoptosis and activating the adaptive immune system. We then performed a tissue specific pathway analysis among the 404 GEM-genes, which similarly identified interferon alpha/beta signaling as greater than 3x fold enriched in each of the five tissue types. We also identified enrichment of the VEGFA-VEGFR2 pathway that is largely specific to T cells (Fig. 6B). Deletion of VEGFA in CD8+ T cells has previously been shown to reduce effector function[67,68]. Collectively, these results highlight the importance of anti-proliferative cytokines to inhibiting clonal expansion and suggest that transcriptomic responses to mutation burden in disparate tissues (nasal epithelial and blood samples) are more similar than anticipated.

Next, we asked whether specific loci that either define CHIP, or have been discovered in GWAS of CHIP or mCAs (Supplementary Tables 7-8), implicate genes whose expression levels are also associated with GEM. Among loci identified in either CHIP or mCA GWAS, we observed that expression of *TCL1A* and *MSI2* are positively associated with GEM (Fig. 6C-D). The association between *TCL1A* and GEM is consistent with *TCL1A* expression as a key mechanism in modulating mutation burden and clonal expansion in whole blood. Among CHIP mutations (Supplementary Table 9), we observed that *IDH2* expression is negatively associated with GEM (Extended Data Figure 7), which corroborates its protective effects against clonal expansion. We then asked whether within blood, there were specific genes with heterogeneous effects. We found that although effect sizes generally were highly concordant, *TCL1A* had a much stronger association with GEM in T cells than in monocytes (Extended Data Figure 8), highlighting the need the tissue and cell-specific transcriptomic analyses when performing searches in blood for the transcriptomic correlates of mutation burden.

medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 6: The transcriptomic correlates of GEM. A, Association analyses were performed between GEM and gene expression in whole blood, including age, sex, genotype PCs 1-5, and expression PCs 1-20 as covariates. B, Enrichment analyses were performed using pathfindR and KEGG pathways as reference. C, Association statistics among CHIP GWAS genes. D, Association statistics among mCA GWAS genes.



medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

The Clinical Correlates of GEM

Because clonal hematopoiesis phenomena have been previously associated with cardiovascular disease[4,6,15], we asked whether GEM associates with vascular and heart disease phenotypes in TOPMed. We first asked whether GEM was associated with coronary artery disease (CAD). We restricted our analyses to those NHLBI TOPMed cohorts with harmonized longitudinal assessment of CAD events; this enabled separate analyses of the association between GEM and incident CAD phenotypes (i.e., CAD events that occurred after the blood draw from which GEM was assessed) and the association between GEM and prevalent CAD (i.e., CAD events prior to the GEM blood draw). Within four NHLBI cohorts (WHI, FHS, CHS, COPDGene), we performed a Cox-proportional hazards regression analyses for incident CAD events after excluding individuals with prevalent CAD disease, including GEM, age at baseline, smoking history, body-mass index (BMI), sex, and germline genotype PCs as covariates. We observed that GEM was not associated with incident CAD events (meta-analysis hazard ratio: 1.00, 95% CI: [0.97, 1.04], pvalue = 0.84, Figure 7A).

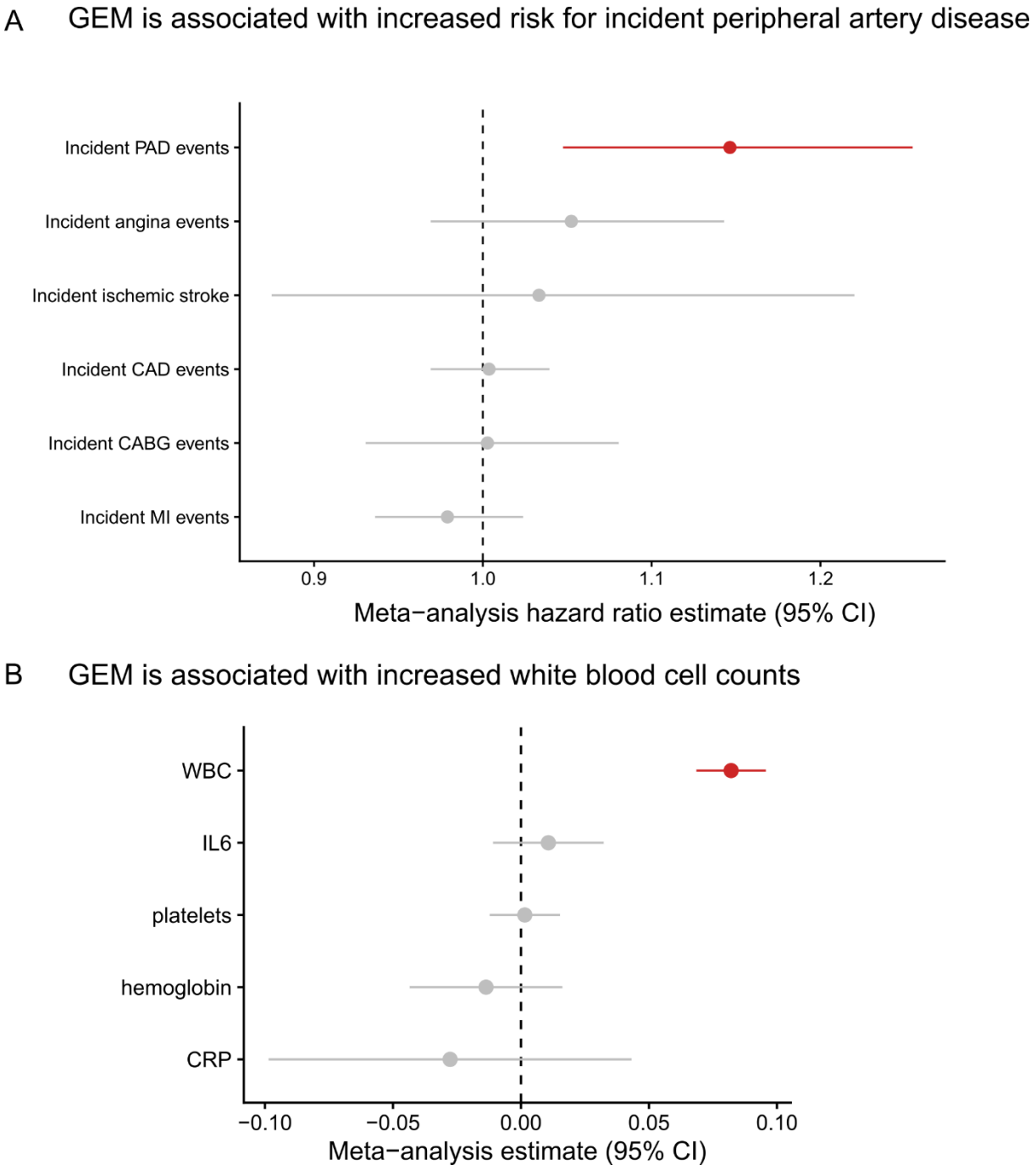
We performed similar analyses with incident ischemic stroke. After meta-analyzing results from three cohorts (WHI, CHS, ARIC), we observed that there was substantial heterogeneity (I^2 = 86%) in results, with a positive association observed in WHI (hazard ratio of 1.19, 95% CI: [1.12, 1.26], pvalue = 2.8×10^{-8}) and null or negative effects observed in ARIC and CHS (Extended Data Figure 9). Meta-analysis resulted in no association between incident stroke and GEM (hazard ratio of 1.03, 95% CI: [0.88, 1.22]), Figure 7A). Given the differences in the distribution of sex across the three cohorts, we then performed a female only analysis in CHS and ARIC, finding no evidence for a sex-specific effect after meta-analyzing with WHI (hazard ratio of 1.05, 95% CI: [0.87, 1.26], Extended Data Figure 10). We then asked whether GEM is associated with incident peripheral artery disease (PAD). After meta-analysis, we observed that GEM was associated with increased risk (hazard ratio 1.15, 95% CI: [1.05, 1.26], pvalue = 3×10^{-3}, Figure 7A) for incident PAD events.

Because evidence within the clonal hematopoiesis literature (Heyde et al.[69]) suggests that prior CAD events are causal contributors to the up-regulation of HSC proliferation, we then asked whether prior CAD was associated with GEM. In a meta-analysis across five NHLBI cohorts (WHI, FHS, CHS, COPD, ARIC), we observed that prior CAD was suggestively associated with a modest increase in GEM (effect of prior CAD on standardized GEM: 0.12, 95% CI: [-0.03, 0.28], pvalue = 0.13, Extended Data Figure 11).

To reconcile these disparate phenotypic associations, we performed analyses between GEM and biomarkers, including complete blood cell counts (CBC) measurements and inflammation measurements. We observed that GEM was positively associated with increased white blood cell count after adjustment for age and smoking status at baseline (Fig. 7B). These results indicate that GEM is associated with altered hematopoiesis. In contrast, after meta-analysis, we observed no association with CRP (Figure 7B). This suggests that the association between GEM and PAD is not mediated through systemic inflammatory markers like CRP, but we note that markers of systemic inflammation like CRP may not be sufficiently sensitive to capture the association between inflammation and HSC activity within the bone-marrow microenvironment.

medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 7: The phenotype correlates of GEM. A, Cox proportional-hazard regressions were performed, regressing incident events on a spline of age, sex, smoking status, and germline PCs. Individuals with prevalent disease were excluded. CAD = coronary artery disease, PAD = peripheral artery disease, CABG = coronary artery bypass graft, MI = myocardial infarction. CAD events were defined as at least one of an MI, CABG, angina, or angioplasty during the follow-up period. A random effects meta-analysis was performed. GEM was inverse normal transformed. Sex was excluded from the WHI regression, and smoking was excluded from the COPD regression. B, A linear regression of the inverse normal transformed biomarker, including a spline of age, sex, smoking status, and germline PCs as covariates. GEM was inverse normal transformed. Sex was excluded from the WHI regression, and smoking was excluding from the COPD regression.



medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Conclusion

Using 51,399 diverse TOPMed whole genomes, we derived a semi-supervised model of mutation rate, GEM, that increases power for discovery of germline determinants of mutation burden by better distinguishing somatic mutations from sequencing artifacts. Using GEM to identify the germline determinants of CH-LPMneg, we observed the convergence of common variant loci influencing multiple types of clonal phenomena, including CHIP and mCAs, demonstrating that genetic predisposition to mutation burden is shared across several different clonal contexts. We observed that altered protein sequence of *NRIP1* was strongly associated with increased mutation rate, which along with its documented role in GWAS of blood cell indices, collectively implicate *NRIP1* as an important regulator of aberrant hematopoiesis. We also observed that *TCL1A*, which we previously identified as a critical moderator of clonal expansion in CHIP carriers, is also associated with mutation burden in samples ascertained for not having CHIP, which may reflect its contribution towards clonal expansion in other kinds of clonal phenomena. Using a sex specific mutation pathway revealed by PCA analysis, we identified a breast-cancer locus, *SRGAP2C*, that associates with GEM.

We anticipate that our approach, which identifies distinct underlying mutation pathways through PCA analysis can be extended to identify other factors that may contribute to specific mutation pathways. Overall, our approach identifies several loci that have been previously discovered in other analyses of clonal hematopoiesis phenomena, suggesting that studying clonal hematopoiesis without known drivers represents an under-appreciated model for discovering the germline determinants of mosaicism in blood. Importantly, analysis of mutation burden without a known CHIP genetic lesion greatly expands the sample size available to perform these analyses; our analysis here is an order of magnitude larger than the number of CHIP carriers discovered in our previous analyses of TOPMed analyses[13,17].

We performed the first multi-tissue analysis of the transcriptomic consequences of mutation burden in whole blood. We observed the striking down-regulation of interferon signaling across five tissues, including four from blood and one from epithelial tissue. Collectively, this analysis highlighted the need for additional characterization of the *in-vivo* transcriptomic consequences of anti-proliferative cytokines on HSC growth. Interferon-alpha, among other cytokines with similar effects, may represent candidates for therapeutic intervention.

We observed that mutation burden in whole blood was associated with altered blood cell indices and increased risk for peripheral artery disease. However, GEM was not associated with incident CAD events. This is consistent with the observation that the association between CH and CAD is highly heterogenous across different forms of clonal phenomena. Within CHIP, the largest phenotype analysis to date[35] reported an association (1.31 hazard ratio) with *TET2* CHIP but not *DNMT3A* CHIP. Several analyses within smaller cohorts have also reported associations between CHIP and CAD phenotypes[4,6,70]. CH mediated through mCAs have no reported association with CAD[9], while both positive and negative reports exists regarding LOY and CAD related phenotypes[15,23]. A recent analysis that examined barcode-CH[24], which includes several different forms of CH, reported no association between barcode-CH and CAD, while finding an association between barcode-CH and PAD, concordant with our results. These observations reflect the need to examine the associations between CH and CAD stratified by the particular genetic lesion(s), size of clone, and potentially the rate at which a clone is expanding. Indeed, recent reports[52,71] on the plasma proteomic associates of CH have found substantial heterogeneity across different CHIP mutations, highlighting the substantial heterogeneity observed at both epidemiologic and molecular levels.

medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

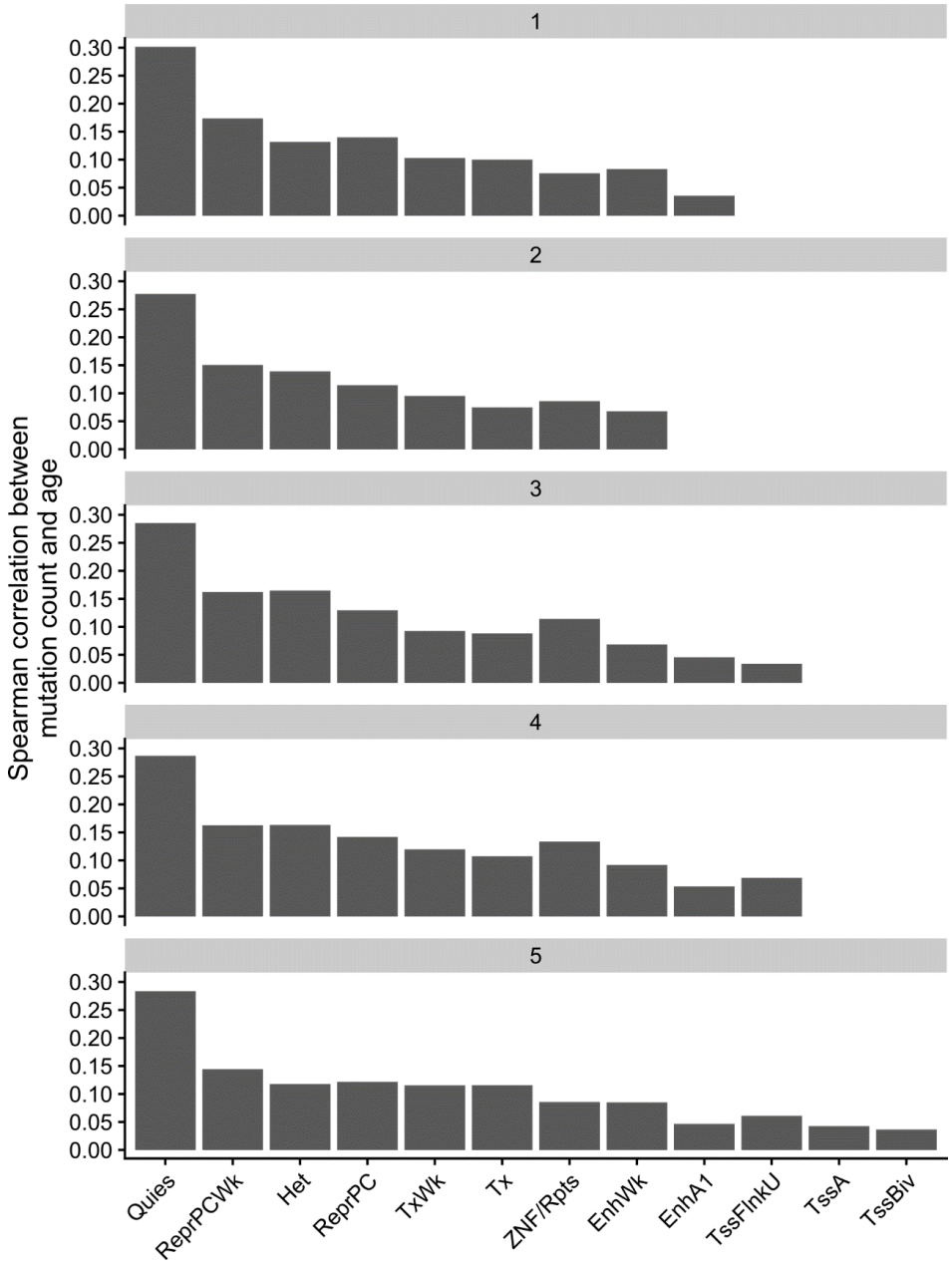
Our approach is not without limitations. The precise estimation of mutations from whole blood remains challenging. Although this approach has been shown to be promising in large cohorts in the context of epidemiologic association analyses, more sensitive sequencing assays are needed for clinical application. Additionally, the genesis of several mutations remains unclear. Although clonal expansion without known drivers clearly occurs[12], elucidating the underlying mechanism remains an open question.

Overall, we develop a novel estimator of mutation burden that is not specific to CHIP carriers. We find that measuring mutation burden, even in individuals without known genetic lesions, is informative for aging related phenotypes. In contrast to surveillance for CHIP, which is relatively rare in individuals less than 80 years old, GEM can be used to monitor mutation burden in a larger proportion of adults. We anticipate that our approach will prove useful in non-blood tissues for the discovery of the germline basis of mutagenesis and will facilitate epidemiologic association analyses, ultimately elucidating the genesis and consequences of mutation burden.

medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

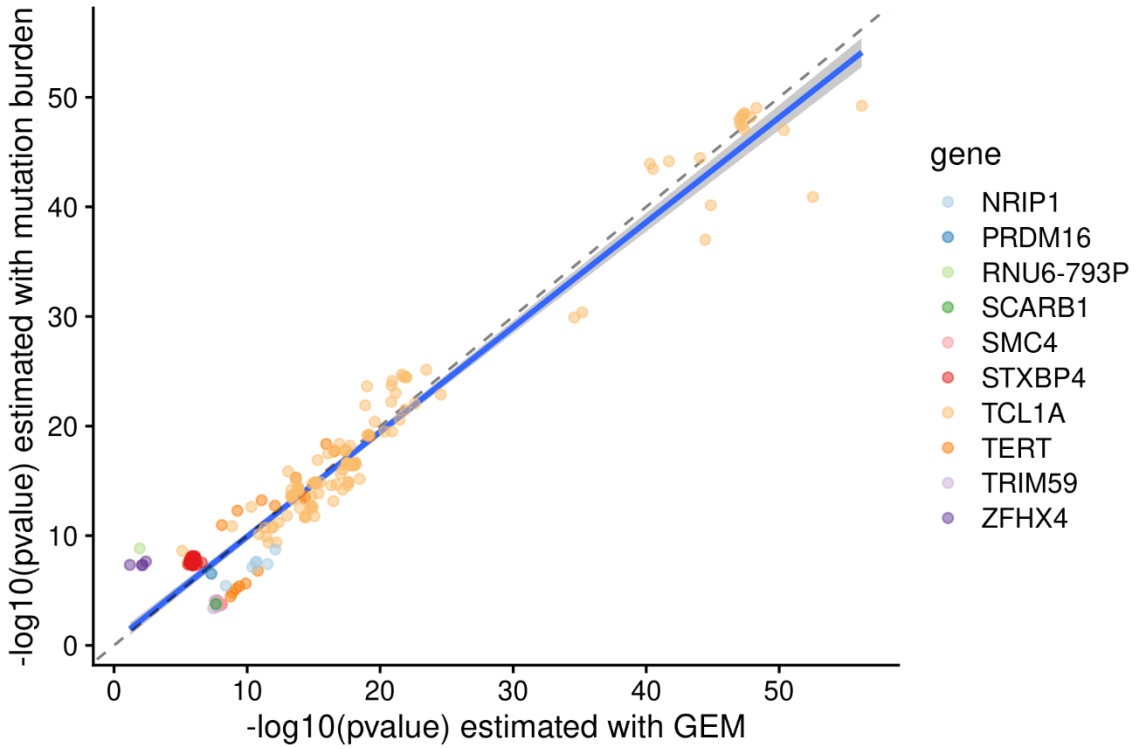
Extended Data Figures

Extended Data Figure 1: Spearman correlation between mutation burden and chronological age was calculated for each of the strata defined by chromHMM 15 state model in CD34+ cells and CADD derived quintiles. A CADD score of 5 indicates a score within the top 20% most deleterious variants.



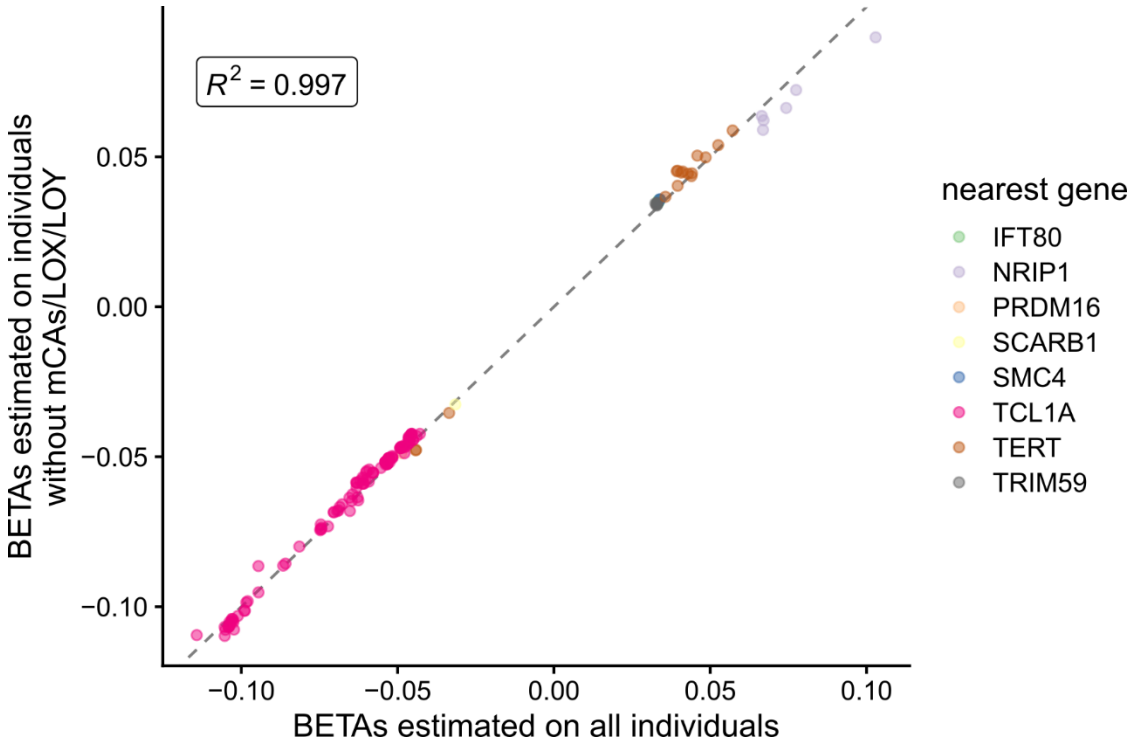
medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Extended Data Figure 2: Scatter plot comparing the -log10 pvalues from GWAS where the phenotype was either GEM (x-axis) or the burden of mutations falling in either heterochromatin or quiescent chromatin in CD34+ cells. Genes are colored by the likely causal gene, which was manually curated. Variants shown have pvalue < 5 x 10^{-8} in at least one of the two GWAS.



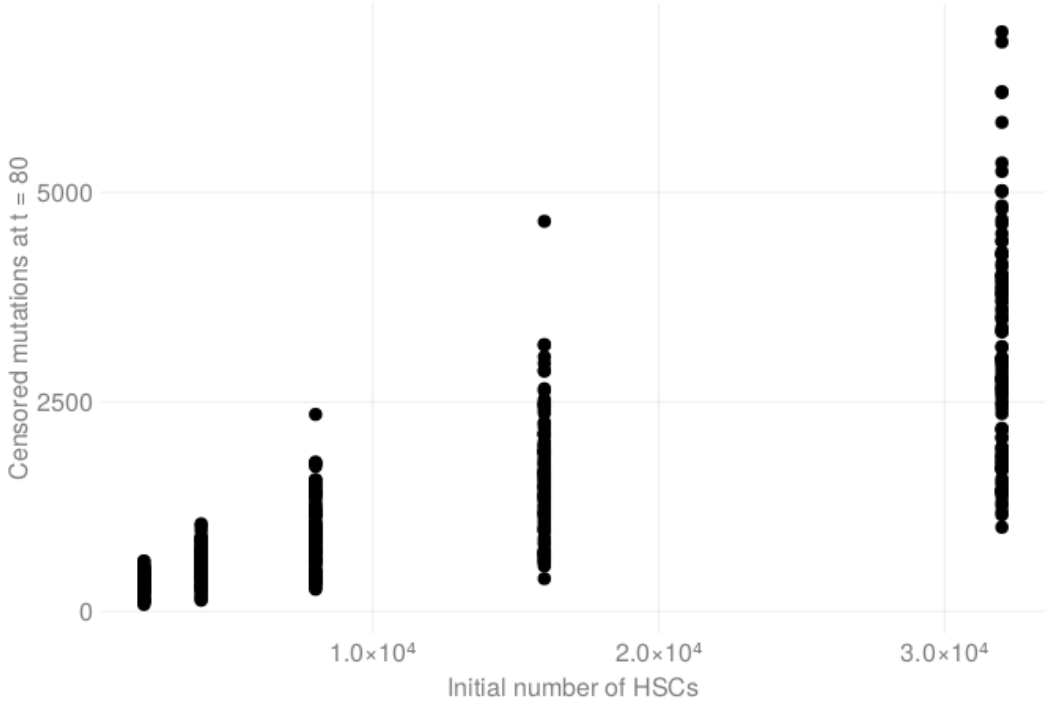
medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Extended Data Figure 3: Scatter plot comparing the beta values from GWAS where the phenotype was either GEM on all individuals (x-axis, n= 51,399) or GEM on individuals that did not have an mCA (n = 38,000). Genes are colored by the likely causal gene, which was manually curated. Variants shown have pvalue < 5 x 10^{-8} in at least one of the two GWAS.

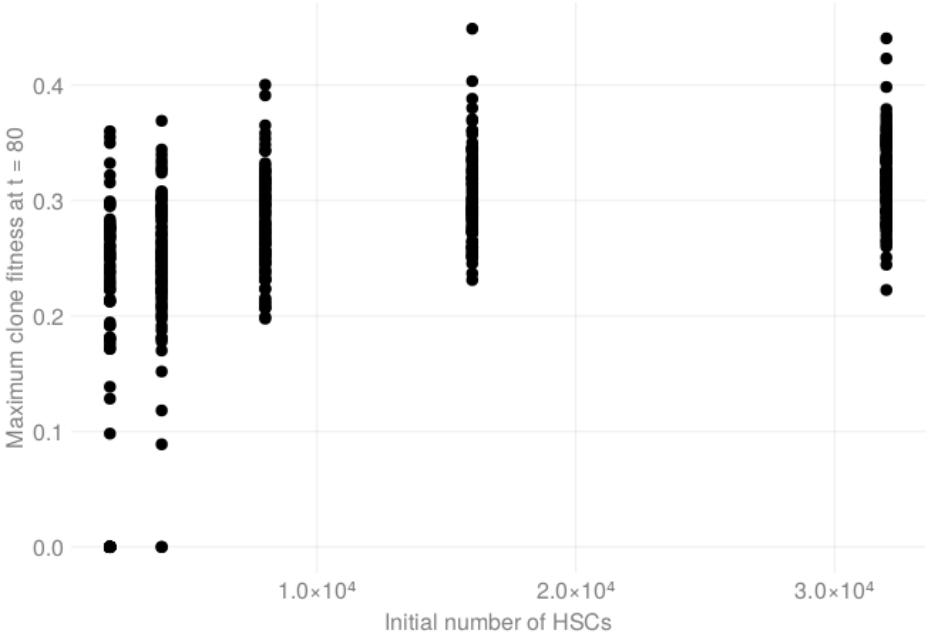


medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Extended Data Figure 4: HSC stochastic process simulation, showing that the number of active HSCs has a large effect on the number of high-VAF mutations at the end of the simulation

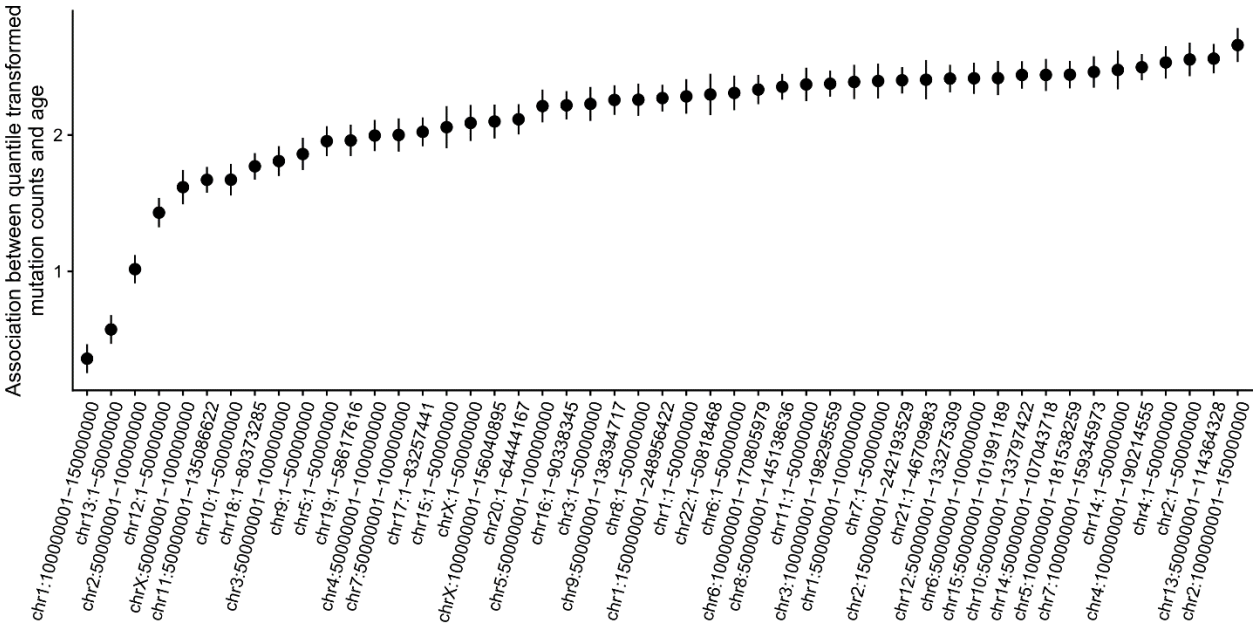


Extended Data Figure 5: HSC stochastic process simulation, showing that the number of active HSCs has a large effect on likelihood of obtaining at clone with high fitness



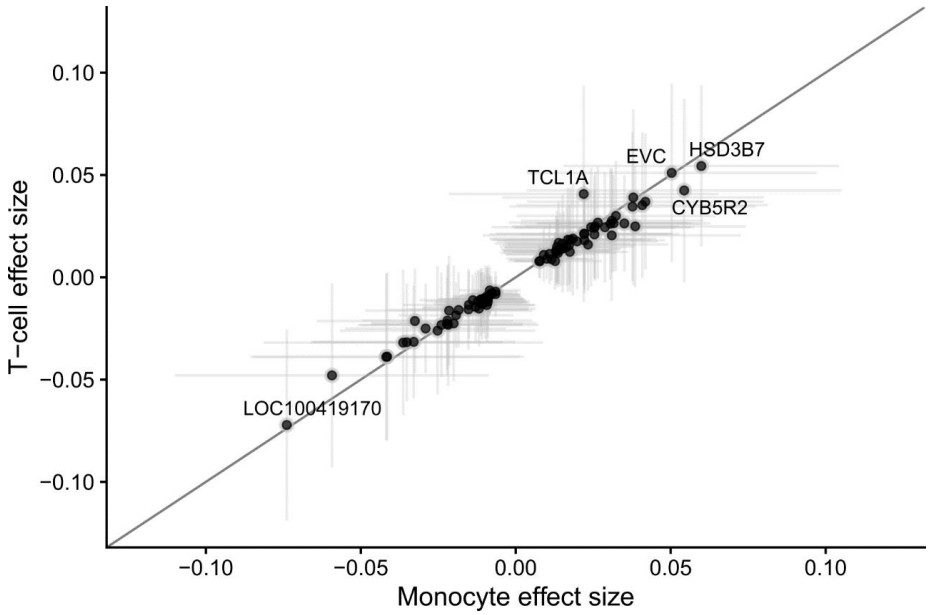
medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Extended Data Figure 6: Linear regressions were performed between the inverse normal transformed mutation burden in each genomic bin with chronological age on the y-axis. Each regression include a study indicator as a covariate.

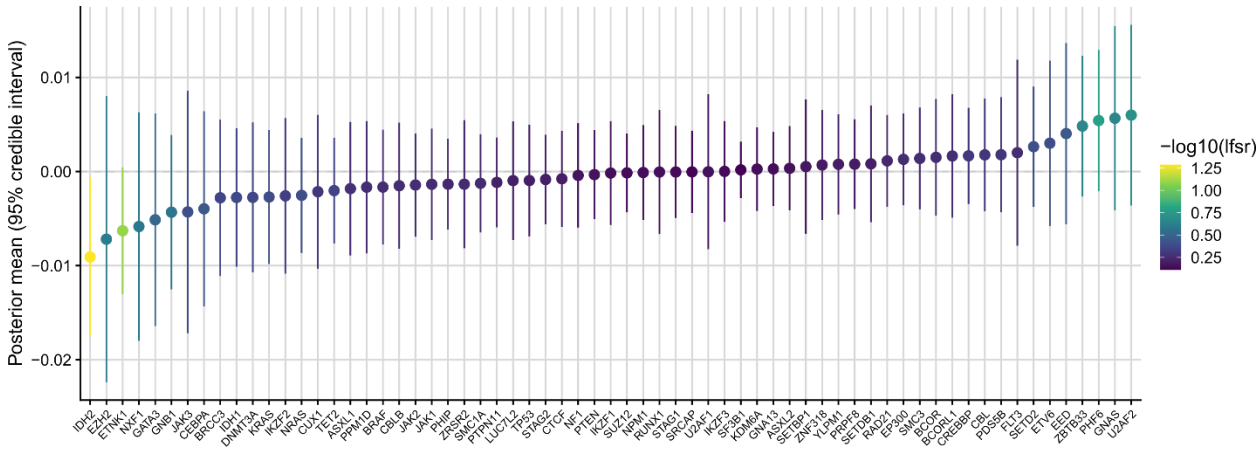


medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Extended Data Figure 8: The association between GEM and gene expression in either monocytes or T cells. Effect sizes are estimated after application of mashr shrinkage, and the intervals denote 95% credible intervals.

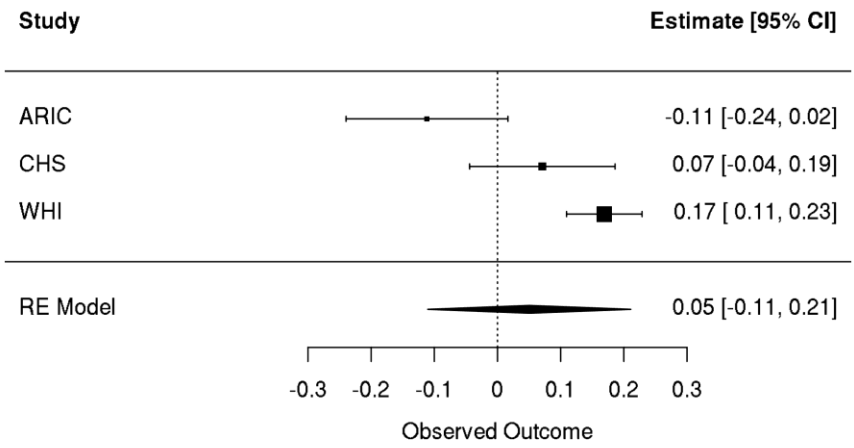


Extended Data Figure 8: The association between GEM and gene expression in whole blood among CHIP genes. Effect sizes are estimated after application of mashr shrinkage, and the intervals denote 95% credible intervals.

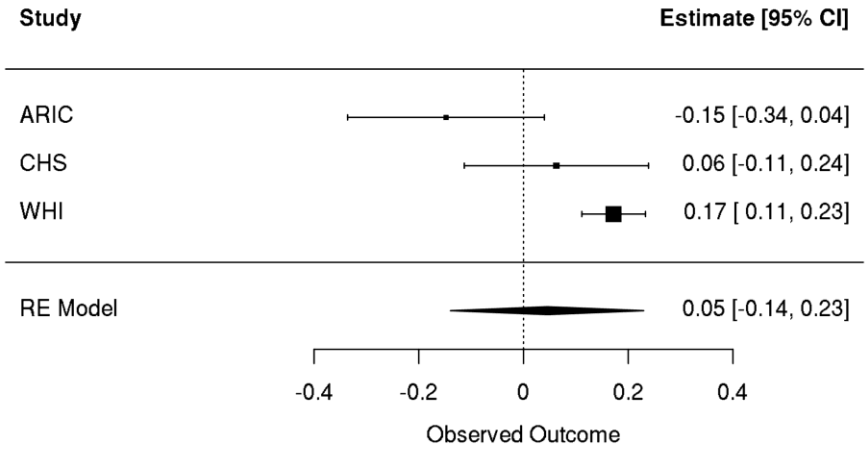


medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Extended Data Figure 9: Meta-analyses of Cox proportional hazards regression with time to ischemic stroke as the outcome. A spline of age, sex, smoking status, and germline PCs were included as covariates. Individuals with prevalent disease were excluded.

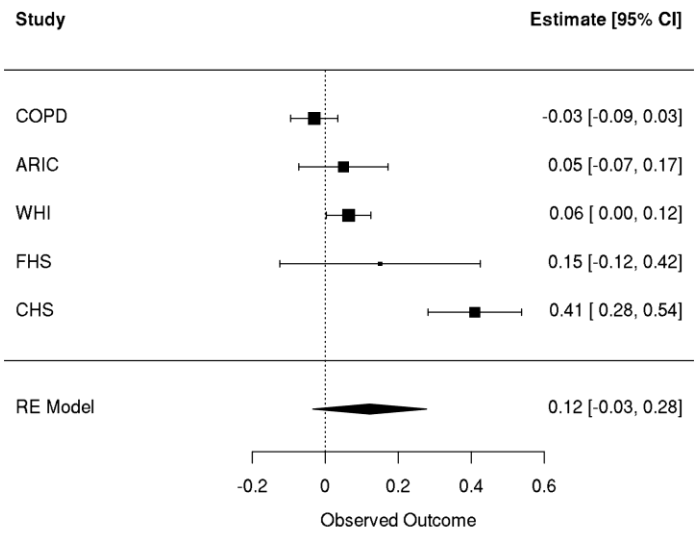


Extended Data Figure 10: Female only meta-analyses of Cox proportional hazards regression with time to ischemic stroke as the outcome. A spline of age, sex, smoking status, and germline PCs were included as covariates. Individuals with prevalent disease were excluded.



medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Extended Data Figure 11: Meta-analyses of linear regressions with inverse normal transformed GEM as the outcome and an indicator for prevalent coronary artery disease events that occurred prior to the blood draw that GEM uses as the covariate of interest. A spline of age, sex, smoking status, and germline PCs were included as covariates.



medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Methods

Germline and somatic variant calling

TOPMed germline variant was performed as previously described[25]. Briefly, TOPMed BAM files were harmonized through the functionally equivalent pipeline[72]. Joint calling of germline SNPs and indels was performed with the Got-Cloud pipeline[73]. Samples were aligned to GRCh38. TOPMed germline SNP and indel freeze 10 was used in this analysis.

Putative somatic variants were first called with GATK Mutect2[74] in "tumor-only" mode with largely default settings. A "panel of normals" was included to exclude sequencing artifacts. Variant calling was performed on Google Cloud using Cromwell[75]. Only bi-allelic variants that passed Mutect2 filters were included in downstream analyses. CHIP calling was performed as previously described[13,17]; briefly, the Mutect2 output was cross-referencing with a list of predominately loss-of-function and missense mutations in a curated set of genes[4,76].

We first identified somatic mutations that occurred once across all individuals, as singleton passenger mutations have a stronger association with chronological age than non-CHIP recurrent somatic mutations[77]. On mutations on the X-chromosome, we halved the variant allele-fraction for all mutations. We then excluded several mutations based on the following filters:

1. All mutations with a depth less than 25x or greater than 100x
2. All mutations falling within low complexity sequence regions
3. All mutations in segmental duplications
4. All mutations falling within genomic regions with germline CNVs with at least 10% minor-allele frequency. Germline CNVs from the TOPMed germline structural variant call-set[78] were used in this filter.
5. All mutations falling within the contigs with sequence that differed between hg19 and hg38, as defined by the "Hg19 diff" track in the UCSC genome table browser.
6. Any germline variant in TOPMed germline SNP and index freeze 10 (derived from 184,878 WGS) with a minor allele count of at least 10 and a variant allele fraction between .26 and .74
7. Any mutation with fewer than 2 alt reads or greater than 6 alt reads. At 38x, this corresponds to a VAF interval of 5%-16%

Annotation of somatic mutations

Singleton mutations after the above filters were first annotated with the variant effect predictor[79] (VEP) including the "—flag-pick", "—check_existing", "—canonical", and "—flag_pick" flags. A CADD[33] plugin was also included. CD34+ chromatin annotations were downloaded for sample BSS00233 from Roadmap epigenomics[29]. Mutations were also annotated with the correspond mutation type (e.g., C->T, G-T, etc.).

The genomic and epigenomic mutation rate (GEM)

GEM is a Bayesian graphical model with the following form. In the outlier layer, standardized chronological age (standardized with (age – 60) / 10) is the outcome variable, denoted as Y_i. We note that GEM is a "weakly-supervised" model in the sense that while individual mutations are unlabeled, the entire training process is "supervised" by chronological age. Y_i conditional on the number of true mutations within an individual is assumed to follow a gaussian distribution. Each individual i has a candidate set of mutations S_i which were identified by the above filtering processes. Instead of using this raw count, we instead replace the count with the expectation of a Bernoulli random variable

medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

$Z_{i,j}$ which denotes whether the jth mutation in the ith individual is a "true mutation" (i.e., takes a value of 1.0) or is an artifact (i.e., takes a value of 0). We include a non-linear transformation g to the sum over the true mutation burden. In practice, $g(x) = log2(x)$ worked well.

$$Y_i \sim N\left(\theta_0 + \theta_1 * g\left(\sum_{j \in S_i} Z_j\right), \sigma\right)$$

The expectation of this random variable is specified through an inverse-logit transformation, i.e.,

$$Z_{j,l} \sim Bernoulli\left(sigmoid\left(\theta_2 + \boldsymbol{X}_j \boldsymbol{\beta}\right)\right)$$

$$E\left(Z_{i.j}\right) = sigmoid\left(\theta_2 + \boldsymbol{X}_j \boldsymbol{\beta}\right)$$

Where $\boldsymbol{X_j}$ represents a length p vector of annotations for the jth mutation and $\boldsymbol{\beta}$ is a length p random vector of weights. θ_2 is included as a bias or intercept term.

The above assumptions express the likelihood of GEM. The prior of GEM is specified as follows:

$$\boldsymbol{\beta} \sim N_p(0, I)$$

$$\sigma \sim LogNormal(0.0, 1.0)$$

$$\theta_1 \sim LogNormal(\epsilon, 1.0)$$

Where ϵ is in practice set to 5 x 10^{-3}. Inference was performed by optimizing the maximum a-posteriori objective using an ADAM optimizer. GEM is implemented in the torch package in R.

Within the matrix of mutation annotations $\boldsymbol{X}$, we include the following annotations:

1. VEP annotated variant impact
2. VEP "somatic" annotation
3. CADD_PHRED score
4. The mutation type
5. The variant allele fraction
6. The chromatin state prediction

GEM was trained on 2,000 randomly sampled individuals with 186,277 total candidate mutations among them.

Genome-wide association studies with GEM

In the context of genome-wide association studies (GWAS), the phenotype was defined as the expected burden of "true" mutations, i.e., $\sum_{j \in S_i} E(Z_j)$ within the ith individual. This phenotype was inverse normal transformed. GWAS summary statistics were estimated with SAIGE on all germline variants where the minor allele count was at least 400 (i.e., MAF $\geq$ 0.4%) among the analyzed samples. Germline principal components 1-10, somatic principal components 3-4, genotype inferred sex, a cohort indicator, chronological age, average sequencing depth per sample, and the residual between the raw and estimated true mutation burden were included as covariates. Somatic principal components 1-2 were excluded as they are strongly associated with total mutation burden and sex respectively. Somatic

medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

mutations that had previously been identified as recurrent[77] were excluded from the summary statistics. All germline variants with a milk-SVM threshold below -0.30 or an individual specific Hardy-Weinberg equilibrium -log10 pvalue above 5.0 were excluded.

Rare-variant association studies with GEM

Rare-variant association studies (RVAS) were performed with the same GEM derived phenotype as the GWAS. We performed a non-coding RVAS by examining rare-variants within 100kb of cancer associated genes as defined by Open Targets[37] using SCANG[56], which performs a scanning procedure for genome regions that contain a set of rare variants that associate with the phenotype. We similarly performed a genome-wide coding variant RVAS using STAAR, including any rare-variant annotated as having a "MODERATE" or "HIGH" impact on amino acid sequence by VEP.

Simulation of mutation burden

We assume an HSC can fall into one of three states:

1. HSCs can divide into two HSCs ("self-renewal")
2. HSCs can divide into two differentiated cells
3. HSCs can divide into one HSC and one differentiated cell

For the purposes of simulating a stochastic process of HSC population size, we treat state 3 as irrelevant because it does not affect the total number of self-renewing HSCs.

We define the HSC clone birth rate as: $\lambda_i(t) \sim Poisson\big(\omega * X_i(t) * \big(1 + s_i(t)\big) * dt\big)$ and the HSC clone death rate as $\psi_i(t) \sim Poisson(\omega * X_i(t) * (1 - s_i(t)) * dt)$. ω is a parameter that controls the rate of births/deaths. dt defines the time interval over which this process is defined.

The total size of the cells within the ith clone at time t as $X_i(t) = \sum_{l \leq t} \lambda_i\,(l) - \psi_i(l)$. A single parameter s_i determines the likelihood of a given HSC falling into state 1. or 2., and thus we refer to this parameter as the clone "fitness."

At any given time t, the VAF of the ith clone is defined as $VAF_i(t) = \frac{X_i(t)}{\sum_j X_j(t)}$. We define the number of passenger mutations at time t in the ith clone as $A_i(t) \sim Poisson\big(X_i(t) * \mu_p * dt\big)$, where μ_p is a per-cell passenger mutation rate. We define $AC_i(t)$ as the count of "censored" passenger mutations at time t for the ith clone, where the censoring occurs due to the limited sensitivity of ~38x sequencing coverage. This censoring is implemented by the following probability $P(Binomial(38, VAF_i) > 2)$.

Association between GEM and gene expression

Separate association analyses were performed for each of the five tissue types available within TOPMed: whole blood, PBMCs, T cells, monocytes, and nasal epithelial tissue. We performed linear regression between the inverse normalized GEM estimate of the true mutation burden and inverse normalized gene expression in the tissue, where chronological age, sex, germline genotype PCs 1-15 and expression PCs 1-20 were included as covariates. In the whole blood analysis, we also included a cohort indicator as a covariate. Summary statistics from each analysis were then included a Bayesian multivariate analysis implemented in mashr [62]. As a measure of "significance", we used the mashr estimate of the local false-sign rate (LFSR) < 0.05. Enrichment analyses were performed with the pathfindR[80] package including all tested genes as the background set and Reactome[81] as the reference database for gene sets.

medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Incident ischemic stroke analysis

Ischemic stroke at most recent visit was chosen for the survival analysis event, and the time to event was defined as the difference in years between baseline and the most recent visit. The WHI, CHS, and ARIC cohorts were included. There was a total of 9,885 individuals included in this analysis from the WHI cohort. In WHI, 9,520 samples were included, there were 1,134 events. In CHS, 2,822 samples were included, and 199 had events. For ARIC, 3,475 samples were included with 231 events. Covariates included Ischemic case status at baseline, BMI measured at baseline, "ever smoker" status at baseline, a spline of age at blood draw, and genetic ancestry PCs 1-4.

Incident coronary artery disease analysis

Incident coronary artery disease at most recent visit was chosen for the survival analysis event, and the time to event was defined as the difference in years between baseline and the most recent visit. A composite coronary artery disease phenotype was defined as an event if at least one of the following occurred during the follow-up period: myocardial infarction, coronary artery bypass graft, angina, angioplasty, or death due to coronary heart disease. Individuals with prevalent disease based on this composite phenotype were excluded. The WHI, CHS, COPDGene, and FHS cohorts were included. In WHI, 9,039 samples were included, there were 1,787 events. In CHS, 2,456 samples were included, and 933 had events. In FHS, 3,786 samples were included and 525 had events. For COPD, 4,987 samples were included with 133 events. Covariates included Ischemic case status at baseline, BMI measured at baseline, "ever smoker" status at baseline, a spline of age at blood draw, and genetic ancestry PCs 1-4.

Lentiviral transduction of healthy CD34+ cells

Lentiviral vectors expressing NRIP1 (V2LHS_172503, V2LHS_172504 and V2LHS_172507) or SMC4-targeting shRNA (V2LHS_21882, V3LHS318029, V3LHS_318030) (Horizon) or non-silencing pGIPZ-puro lentiviral vector was transfected together with pCMV-dR8.9 and vesicular stomatitis virus G–expressing plasmids into HEK 293-FT cells using Lipofectamine 2000 (Thermo Fisher Scientific) for lentiviral supernatant production as previously described[82]. Primary CD34+ cells were obtained as excess material from harvests of normal donors for allogeneic bone marrow transplantation. Specimens were collected by the Johns Hopkins Kimmel Cancer Center Specimen Accessioning Core. Appropriate informed consent was obtained from all donors before specimen collection in accordance with the Declaration of Helsinki and under a research protocol approved by the Johns Hopkins Institutional Review Board. CD34+ cell subsets were isolated using the CD34 MicroBead kit (Miltenyi Biotec) as previously described[83]. CD34+ cells were incubated with the viral supernatant and polybrene (8µg/ml; MilliporeSigma) for transduction in wells pre-coated with retronectin (20ng/ml; MilliporeSigma). After at least 48 hours, cells were treated with puromycin (0.5µg/ml; MilliporeSigma) for 4 days to select resistant cells.

Apoptosis and differentiation assays

Apoptosis was assessed by 7-AAD staining evaluated by flow cytometry (Thermo Fisher Scientific #00-6993-50). Percentages of stem (CD34+CD38-) and progenitor cells (CD34+CD38+) were assessed by CD34 (Thermo Fisher Scientific #11-0349-42) and CD38 (BioLegend #356641) staining evaluated by flow cytometry.

medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Clonogenicity assays

CD34+ cells following puromycin treatment were collected, counted, and plated at a density of 2000 cells/ml in methylcellulose-based media as previously described[82]. After 10 to 14 days of incubation at 37°C in 5% CO2, the recovery of colony-forming units (burst forming unit-erythroid (BFU-E) and colony forming unit-granulocyte/monocyte (CFU-GM)) were determined by colony counting under bright-field microscopy. A cell aggregate composed of >50 cells was defined as a colony.

Code and data availability

Code for the Genomic and Epigenomic Mutation rate pipeline: https://github.com/weinstockj/GEM . Individual whole-genome sequence data for TOPMed whole genomes, individual-level harmonized phenotypes and the CHIP variant call sets used in this analysis are available through restricted access via the dbGaP TOPMed Exchange Area available to TOPMed investigators.

Acknowledgements

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). See Supplementary Information 1 for study omics support information. Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Phenotype harmonization, data management, sample-identity quality control and general study coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We thank the studies and participants who provided biological samples and data for TOPMed. The full study-specific acknowledgments are included in Supplementary Cohort Acknowledgements. The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the US Department of Health and Human Services. The authors wish to acknowledge the contributions of the consortium working on the development of the NHLBI BioData Catalyst ecosystem.

Competing Interests Declaration

L.M.R. is a consultant for the TOPMed Administrative Coordinating Center (through Westat). B.M.P. serves on the Steering Committee of the Yale Open Data Access Project funded by Johnson & Johnson. J.Y. reports grant support from Bayer. M.C. reports grant support from Bayer and GSK, Consulting and speaking fees from Illumina and AstraZeneca. A.G.B. , P.N, and S.J. are cofounders, equity holders, and on the scientific advisory board of TenSixteen Bio. G.R.A. is an employee of Regeneron Pharmaceuticals and receives salary, stock and stock options as compensation.

medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Works Cited

1. Jaiswal, S. *et al.* Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes A BS TR AC T. *NEJM.org. N Engl J Med* **26**, 2488–98 (2014).
2. Genovese, G. *et al.* Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. *New England Journal of Medicine* **371**, 2477–2487 (2014).
3. Xie, M. *et al.* Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nature Medicine* **20**, 1472–1478 (2014).
4. Jaiswal, S. *et al.* Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *New England Journal of Medicine* (2017) doi:10.1056/NEJMoa1701719.
5. Desai, P. *et al.* Somatic mutations precede acute myeloid leukemia years before diagnosis. *Nature Medicine* **24**, 1015–1023 (2018).
6. Bick Alexander G. *et al.* Genetic Interleukin 6 Signaling Deficiency Attenuates Cardiovascular Risk in Clonal Hematopoiesis. *Circulation* **141**, 124–131 (2020).
7. Steensma, D. P. *et al.* Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood* **126**, 9–16 (2015).
8. Loh, P.-R. *et al.* Insights about clonal hematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350–355 (2018).
9. Loh, P.-R., Genovese, G. & McCarroll, S. A. Monogenic and polygenic inheritance become instruments for clonal selection. *Nature* **584**, 136–141 (2020).
10. Terao, C. *et al.* Chromosomal alterations among age-related haematopoietic clones in Japan. *Nature* **584**, 130–135 (2020).
11. Zink, F. *et al.* Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* **130**, 742–752 (2017).

medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

12. Mitchell, E. *et al.* Clonal dynamics of haematopoiesis across the human lifespan. *Nature* **606**, 1–8 (2022).
13. Weinstock, J. S. *et al.* Aberrant activation of TCL1A promotes stem cell expansion in clonal haematopoiesis. *Nature* **616**, 755–763 (2023).
14. Zekavat, S. M. *et al.* Hematopoietic mosaic chromosomal alterations increase the risk for diverse types of infection. *Nat Med* **27**, 1012–1024 (2021).
15. Sano, S. *et al.* Hematopoietic loss of Y chromosome leads to cardiac fibrosis and heart failure mortality. *Science* **377**, 292–297 (2022).
16. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504–507 (2012).
17. Bick, A. G. *et al.* Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature* **586**, 763–768 (2020).
18. Kar, S. P. *et al. Genome-wide analyses of 200,453 individuals yields new insights into the causes and consequences of clonal hematopoiesis*. 2022.01.06.22268846 https://www.medrxiv.org/content/10.1101/2022.01.06.22268846v1 (2022) doi:10.1101/2022.01.06.22268846.
19. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
20. Jakubek, Y. A. *et al.* Mosaic chromosomal alterations in blood across ancestries using whole-genome sequencing. *Nat Genet* 1–8 (2023) doi:10.1038/s41588-023-01553-1.
21. Thompson, D. J. *et al.* Genetic predisposition to mosaic Y chromosome loss in blood. *Nature* **575**, 652–657 (2019).
22. Zhou, W. *et al.* Mosaic loss of chromosome Y is associated with common variation near TCL1A. *Nature Genetics* **48**, 563–568 (2016).

medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

23. Terao, C. *et al.* GWAS of mosaic loss of chromosome Y highlights genetic effects on blood cell differentiation. *Nat Commun* **10**, 4719 (2019).

24. 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.

25. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).

26. 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.

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

28. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods* **9**, 215–216 (2012).

29. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

30. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

31. Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426-3440.e19 (2022).

32. Ferraro, N. M. *et al.* Transcriptomic signatures across human tissues identify functional rare genetic variation. *Science (New York, N.Y.)* **369**, (2020).

33. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research* **47**, D886–D894 (2019).

34. 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).

medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

35. Kessler, M. D. *et al. Exome sequencing of 628,388 individuals identifies common and rare variant associations with clonal hematopoiesis phenotypes*. 2021.12.29.21268342 https://www.medrxiv.org/content/10.1101/2021.12.29.21268342v1 (2022) doi:10.1101/2021.12.29.21268342.

36. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82**, 1273–1300 (2020).

37. Carvalho-Silva, D. *et al.* Open Targets Platform: new developments and updates two years on. *Nucleic Acids Res* **47**, D1056–D1065 (2019).

38. Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nature Genetics* **51**, 1664–1669 (2019).

39. Weinstock, J. S. *et al.* Clonal hematopoiesis is driven by aberrant activation of TCL1A. 2021.12.10.471810 Preprint at https://doi.org/10.1101/2021.12.10.471810 (2021).

40. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* **22**, 1790–1797 (2012).

41. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat Protoc* **11**, 1–9 (2016).

42. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

43. Petryszak, R. *et al.* Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res* **44**, D746–D752 (2016).

44. Toren, A. *et al.* CD133-Positive Hematopoietic Stem Cell “Stemness” Genes Contain Many Genes Mutated or Abnormally Expressed in Leukemia. *Stem Cells* **23**, 1142–1153 (2005).

medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

45. Chen, M.-H. *et al.* Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell* **182**, 1198-1213.e14 (2020).

46. Yuan, R. *et al.* Genetic coregulation of age of female sexual maturation and lifespan through circulating IGF1 among inbred mouse strains. *Proceedings of the National Academy of Sciences* **109**, 8224–8229 (2012).

47. Sun, B. B. *et al.* Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329–338 (2023).

48. Blake, J. A. *et al.* Mouse Genome Database (MGD): Knowledgebase for mouse-human comparative biology. *Nucleic Acids Res* **49**, D981–D987 (2021).

49. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat Genet* **25**, 25–29 (2000).

50. Võsa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet* **53**, 1300–1310 (2021).

51. Fairfax, B. P. *et al.* Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression. *Science* **343**, 1246949 (2014).

52. Dhindsa, R. S. *et al.* Rare variant associations with plasma protein levels in the UK Biobank. *Nature* **622**, 339–347 (2023).

53. Ye, H. *et al.* Leukemic Stem Cells Evade Chemotherapy by Metabolic Adaptation to an Adipose Tissue Niche. *Cell Stem Cell* **19**, 23–37 (2016).

54. Li, X. *et al.* Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat Genet* **52**, 969–983 (2020).

55. Yeung, Y. T. *et al.* CELF2 suppresses non-small cell lung carcinoma growth by inhibiting the PREX2-PTEN interaction. *Carcinogenesis* **41**, 377–389 (2020).

medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

56. Li, Z. *et al.* Dynamic Scan Procedure for Detecting Rare-Variant Association Regions in Whole-Genome Sequencing Studies. *The American Journal of Human Genetics* **104**, 802–814 (2019).

57. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398-1414.e24 (2016).

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

59. Bycroft, C. *et al.* Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv* 166298–166298 (2017) doi:10.1101/166298.

60. FinnGen. FinnGen. *FinnGen Documentation of R3 release* https://finngen.gitbook.io/documentation/.

61. Oliva, M. *et al.* The impact of sex on gene expression across human tissues. *Science* **369**, eaba3066 (2020).

62. 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).

63. Stephens, M. False discovery rates: a new deal. *Biostatistics* **18**, 275–294 (2017).

64. Valette, K. *et al.* Prioritization of candidate causal genes for asthma in susceptibility loci derived from UK Biobank. *Commun Biol* **4**, 700 (2021).

65. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res* **51**, D587–D592 (2023).

66. Anguille, S. *et al.* Interferon-α in acute myeloid leukemia: an old drug revisited. *Leukemia* **25**, 739–748 (2011).

67. de Almeida, P. E. *et al.* Anti-VEGF Treatment Enhances CD8+ T-cell Antitumor Activity by Amplifying Hypoxia. *Cancer Immunol Res* **8**, 806–818 (2020).

medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

68. Palazon, A. *et al.* An HIF-1α/VEGF-A Axis in Cytotoxic T Cells Regulates Tumor Progression. *Cancer Cell* **32**, 669-683.e5 (2017).

69. Heyde, A. *et al.* Increased stem cell proliferation in atherosclerosis accelerates clonal hematopoiesis. *Cell* **184**, 1348-1361.e22 (2021).

70. Yu, B. *et al.* Association of Clonal Hematopoiesis With Incident Heart Failure. *Journal of the American College of Cardiology* **78**, 42–52 (2021).

71. Yu, Z. *et al.* Human Plasma Proteomic Profile of Clonal Hematopoiesis. 2023.07.25.550557 Preprint at https://doi.org/10.1101/2023.07.25.550557 (2023).

72. Regier, A. A. *et al.* Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nature Communications* **9**, 1–8 (2018).

73. Jun, G., Wing, M. K., Abecasis, G. R. & Kang, H. M. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res* **25**, 918–925 (2015).

74. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology* **31**, 213–219 (2013).

75. Voss, K., Gentry, J. & Van der Auwera, G. Full-stack genomics pipelining with GATK4 + WDL + Cromwell. in (F1000 Research, 2017). doi:10.7490/f1000research.1114631.1.

76. Beauchamp, E. M. *et al.* ZBTB33 Is Mutated in Clonal Hematopoiesis and Myelodysplastic Syndromes and Impacts RNA Splicing. *Blood Cancer Discov* (2021) doi:10.1158/2643-3230.BCD-20-0224.

77. Weinstock, J. S. *et al.* The genetic determinants of recurrent somatic mutations in 43,693 blood genomes. *Science Advances* **9**, eabm4945 (2023).

78. Jun, G. *et al.* Structural variation across 138,134 samples in the TOPMed consortium. *bioRxiv* 2023.01.25.525428 (2023) doi:10.1101/2023.01.25.525428.

medRxiv preprint doi: https://doi.org/10.1101/2024.08.22.24312319; this version posted August 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

79. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biology* **17**, 122 (2016).

80. Ulgen, E., Ozisik, O. & Sezerman, O. U. pathfindR: An R Package for Comprehensive Identification of Enriched Pathways in Omics Data Through Active Subnetworks. *Frontiers in Genetics* **10**, (2019).

81. Gillespie, M. *et al.* The reactome pathway knowledgebase 2022. *Nucleic Acids Research* **50**, D687–D692 (2022).

82. Karantanos, T. *et al.* The role of the atypical chemokine receptor CCRL2 in myelodysplastic syndrome and secondary acute myeloid leukemia. *Sci Adv* **8**, eabl8952 (2022).

83. Karantanos, T. *et al.* CCRL2 affects the sensitivity of myelodysplastic syndrome and secondary acute myeloid leukemia cells to azacitidine. *Haematologica* **108**, 1886–1899 (2023).