Spring 2022

Dissecting the Impact of Clonal Hematopoiesis on Age-Related Disease

Seyedeh Maryam Zekavat
Yale University Graduate School of Arts and Sciences, seyedehmaryamzekavat@gmail.com

Abstract

Dissecting the Impact of Clonal Hematopoiesis on Age-Related Disease

Seyedeh Maryam Zekavat

2022

Aging is the strongest risk factor for a number of diseases. Despite current risk prediction, prevention, and therapeutic strategies, age-related diseases including coronary artery disease (CAD), cancer, and now COVID-19, are the leading causes of death in the US and worldwide.

The aging hematopoietic system is characterized by increased prevalence of acquired somatic variants predisposing to clonal expansion. Carriers of somatic mutations predisposing to clonal expansion in hematopoietic stem cells (clonal hematopoiesis of indeterminate potential, CHIP) are at increased risk for not only hematologic cancer but also atherosclerosis. Other classes of somatic variation besides CHIP, including larger somatic structural variants known as mosaic chromosomal abnormalities (mCAs), have also been identified to increase with age and increase risk of cancer.

These data raise several unanswered questions. First, what other age-related diseases are associated with somatic mutations contributing towards clonal hematopoiesis such as CHIP and mCAs? Second, what inherited germline factors influence risk of acquired somatic variants? Third, how does the presence of CHIP influence DNA transcription in human blood cells? My dissertation addresses these questions by integrating genomic data across multiple cohorts with transcriptomic and deep phenotypic data.

Dissecting the Impact of Clonal Hematopoiesis on Age-Related Disease

A Dissertation

Presented to the Faculty of the Graduate School

Of

Yale University

In Candidacy for the Degree of

Doctor of Philosophy

By

Seyedeh Maryam Zekavat

Dissertation Directors: Hongyu Zhao

May 2022

Table of Contents

Acknowledgments

1. Introduction
    1.1. Dissertation Aims
2. Cohort descriptions and methods for CHIP and mCA calling
    2.1. Cohorts and exclusion criteria
    2.2. CHIP calling methods and sensitivity analyses
    2.3. mCA calling methods and sensitivity analyses
3. Phenome-wide association of CHIP and mCAs
    3.1. Association of CHIP and mCAs with age, blood counts, and hematological cancer
    3.2. Comparative phenome-wide association of CHIP and mCAs
    3.3. Association of CHIP with peripheral artery disease and pan-vascular atherosclerosis
    3.4. Association of CHIP with stroke
    3.5. Association of CHIP with heart failure
    3.6. Association of mCAs with diverse infectious diseases, including COVID-19 infection
4. Inherited genetic basis of somatic variation
    4.1. Genome-wide association of CHIP
    4.2. Genome-wide association of mCAs
5. Transcriptome-wide association of CHIP and mCAs
6. Conclusion
7. References

Acknowledgements:
I'm immensely grateful to all who have supported me throughout my life.

- Family: I would not be here if it weren't for the one constant, and yet continuously evolving element in my life, my family -- my parents, Reza Zekavat and Fatemeh Emdad, my sisters Melica and Mona, and my love, Saman Doroodgar, and parents in-law, Sonbol Bayani, Behrooz Doroodgar, and my sister in-law Sahar Doroodgar, and their infinite love and support.
- Pre-Yale: I would not be at Yale if it weren't for all those who helped me see the beauty in academia, medicine, and in the combination of the two via research prior to Yale, in particular:
    - MIT's Biological Engineering department and especially MIT's International Science & Technology Initiatives through which I had my first taste combining ophthalmology with computational modeling at Roche through Dr. Norman Mazer in Switzerland, resulting in my first 1st author paper.
    - Broad Institute of Harvard & MIT: I'm immensely thankful to Dr. Sekar Kathiresan who took the initiative to take me on when I had zero background in computational genomics. Through his lab I was able to start picking up computational biology, worked on the first large-scale whole-genome sequencing datasets, and got introduced to incredible people who set the stage for my future, including Dr. Pradeep Natarajan, then a post-doc in Sek's lab who later became my PhD co-advisor and committee member on the present dissertation.
- During Yale:

- My PhD co-advisors: I'd like to thank Drs. Hongyu Zhao and Pradeep Natarajan and all their lab members for keeping an open, collaborative, and creative mind on research projects, and continuously helping me learn.
- The participants and organizers of all of the biobanks, cohorts, and international collaborative projects I've engaged in, including the Trans-Omics for Precision Medicine program, the COVID-19 Human Genetics Initiative, the UK Biobank, Mass-General-Brigham Biobank, and others.
- Yale's CBB and MSTP programs: I'd like to thank Yale's CBB and MSTP programs and the friends I've made for cultivating a supportive, collegial, and flexible environment that fostered my growth.
- Yale's Ophthalmology department and the Massachusetts Eye & Ear Institute: I'd like to thank Yale and Massachusetts Eye & Ear Institute's Ophthalmology departments for collaboration on the second half of my MD-PhD projects not represented in this thesis: ophthalmology-related projects, which helped shape my decision to apply for ophthalmology residency.

- Post-Yale:
  - I look forward to joining Massachusetts Eye & Ear Institute's Ophthalmology residency over the next four years and thank everyone who made this next chapter a possibility.

Last but not least, a huge heart of love to all the pets that have been newly part of my life during my Yale MD-PhD story as of the COVID-19 pandemic, including our chicken

coop of 10+ chickens, our 4 cats (Mango, Maloose, Luna, and Sunny), and our sweet-hearted Siberian Huskey, Mademoiselle!

Chapter 1: Introduction

Age is the strongest risk factor contributing towards a variety of diseases, including atherosclerosis[3]. Despite current risk prediction, prevention, and therapeutic strategies, age-related diseases including coronary artery disease (CAD) continue to remain the leading cause of death in the US and worldwide[4]. Here, I investigate a novel, independent mechanism contributing towards CAD and other diseases: age-related somatic mutations in bone marrow hematopoietic stem cells predisposing to clonal hematopoiesis. By integrating germline genomic data with somatic variant calls, transcriptomics, and clinical data, this dissertation aims to improve understanding of the mechanistic link between somatic hematopoietic genetic variants and disease.

Our group has discovered a link between the aging hematopoietic system and CAD using whole exome sequencing (WES)[5]. In particular, carriers of somatic mutations predisposing to clonal expansion in hematopoietic stem cells (clonal hematopoiesis of indeterminate potential, CHIP[1]) are at increased risk for not only hematologic cancer but also atherosclerosis[5]. CHIP is defined as the presence of an expanded (i.e.: variant allele fraction, VAF, >2%) small somatic variant (i.e.: SNP, INDEL) in white blood cells among individuals that do not have hematologic cancer. CHIP-related somatic mutations in peripheral blood cells occur across 74 genes known to be implicated in myeloid cancers[5], with the most common mutations being in *DNMT3A*, *TET2*, *JAK2*, and *ASXL1*[6]. The prevalence of such mutations increases with age, with carriers among more than 10% of individuals >70-years. CHIP carriers have 10-fold increased risk for hematologic cancer, particularly myeloid leukemias and myelodysplastic syndrome, and independently, a 4-fold increased risk of early-onset myocardial infarction[5]. CHIP carriers with somatic variants in *TET2* have significantly reduced major adverse

cardiovascular events when treated with canakinumab[7 8], an IL-1B antibody. Thus, knowledge of CHIP status may additionally inform therapeutic strategies.

Animal models also support a connection between CHIP and atherosclerosis. Hematologic knock-out of *Tet2* in mice causes larger atherosclerotic lesions[5]. Transcriptomics of cultured bone-marrow-derived macrophages from these mice show up-regulated expression of genes involved with cytokines, chemokines, and their receptors, and down-regulated expression of genes involved with lysosomal function. This suggests that *Tet2* mutations influence monocyte adhesion, inflammatory signaling, and macrophage phagocytosis[5].

Separate from CHIP, other classes of somatic mutations have also been categorized, including structural somatic mutations known as mosaic chromosomal alterations (mCAs)[2 9 10]. Age-related mosaic chromosomal alterations (mCAs), are large-scale somatic variants (deletions, duplications, and copy-neutral loss of heterozygosity CN-LOH) detected within peripheral leukocytes predisposing to clonal hematopoiesis[2 9 10]. These mCAs have previously been associated with aberrant lymphocyte cell counts, and predispose to chronic lymphocytic leukemia (Hazard ratio, HR~100x) and increased mortality (HR~2)[2 9 10] (**Figure 1.1**).

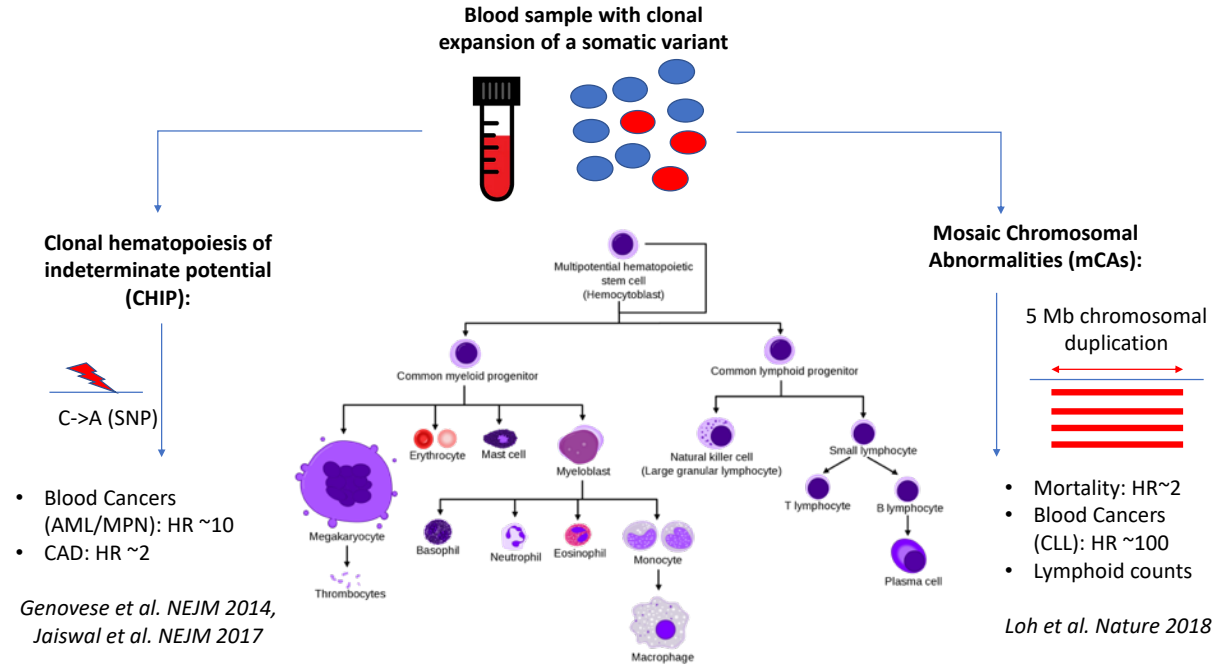


Figure 1.1: Schematic of CHIP and mCAs, showing their respective associations with myeloid (CHIP) and lymphoid (mCAs) leukemias.

Chapter 1.1: Dissertation Aims

These data raise several unanswered questions. First, what other age-related diseases are associated with somatic variants contributing to clonal hematopoiesis (ie: CHIP and mCAs)? Second, what inherited germline factors influence risk of development of somatic variants? Third, how does the presence of CHIP or mCAs influence DNA transcription in human blood cells? This dissertation addresses these questions by integrating whole genome sequence (WGS) data from NHLBI's Trans-Omics for Precision Medicine (TOPMed) program as well as genotype data and whole exome sequencing (WES) data from the UK Biobank as well as other cohorts with somatic, transcriptomic, and deep phenotypic data (**Figure 1.2**).

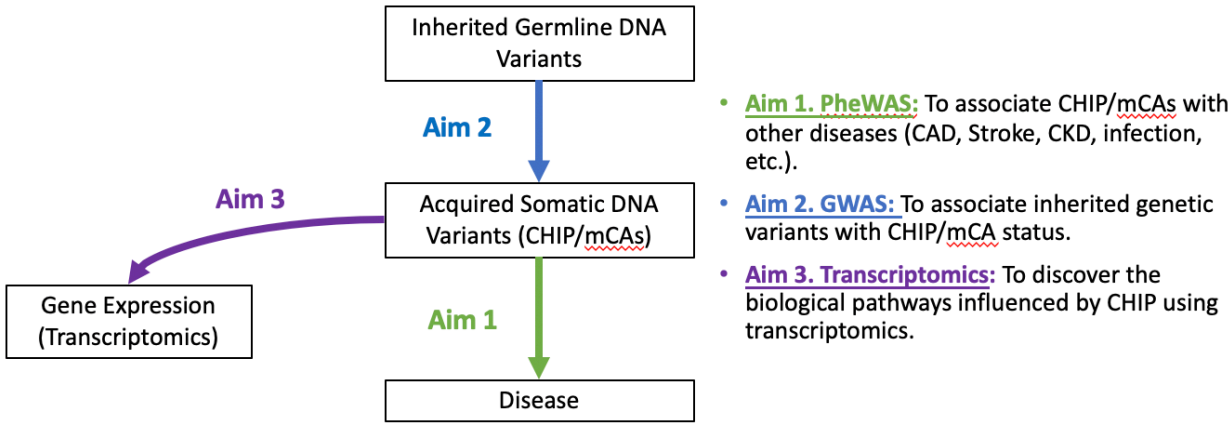


Figure 1.2: Schematic of dissertation aims. *Aim 1: phenome-wide association (PheWAS) of CHIP and mCAs across incident diseases. Aim 2: genome-wide association (GWAS) of CHIP and mCAs to identify inherited basis for acquired somatic mutations. Aim 3: transcriptome-wide association (TWAS) of CHIP and mCAs to identify changes in gene expression and biological pathways influenced by these somatic mutations.*

Chapter 2: Cohort descriptions and methods for CHIP and mCA calling

Chapter 2.1: Cohorts and exclusion criteria

Cohorts used in CHIP analyses:

The UK Biobank is a population-based cohort of approximately 500,000 participants recruited from 2006-2010 with existing genomic and longitudinal phenotypic data and median 10-year follow-up[11]. Baseline assessments were conducted at 22 assessment centres across the UK with sample collections including blood-derived DNA. Of ~49,960 individuals with WES data available, we analyzed 37,657 participants consenting to genetic analyses after our exclusion criteria. Use of the data was approved by the Massachusetts General Hospital Institutional Review Board (protocol 2013P001840) and facilitated through UK Biobank Application 7089.

The Massachusetts General Brigham Biobank (MGBB) contains genotypic and clinical data from >105,000 patients who consented to broad-based research across 7 regional hospitals and median 3-year follow-up[12]. Baseline phenotypes were ascertained from the electronic medical record and surveys. We analyzed 12,465 whole-exome sequenced individuals consenting to genetic analysis after our exclusion criteria. Use of the data was approved by the Massachusetts General Hospital Institutional Review Board (protocol 2020P000904).

Analyses of CHIP acquired from whole genome sequence data in the Trans-Omics for Precision Medicine (TOPMed) program was across 6 major cohort studies (ARIC, CHS, FHS, JHS, MESA, and WHI), cohort descriptions of which are provided in prior publications[13-16].

Across all cohorts, we excluded individuals with prevalent hematologic cancer, individuals without genotypic-phenotypic sex concordance, and one of each pair of 1st or 2nd degree relatives at random. Follow-up time was defined as time from enrollment to disease diagnosis for cases, or to censorship or death for controls.

Cohorts used in mCA analyses:

The UK Biobank, a population-based cohort of approximately 500,000 participants recruited from 2006-2010, had existing genomic and longitudinal phenotypic data[11]. Baseline assessments were conducted at 22 assessment centres across the UK with sample collections including blood-derived DNA. Of 488,377 genotyped individuals, we analyzed 445,101 participants consenting to genetic analyses and who passed sample quality control criteria for mCA calling, had genotypic-phenotypic sex concordance, no 1st or 2nd degree relatives (random exclusion of one from each pair), and no prevalent hematologic cancer at time of blood draw. Genome-wide genotyping of blood-derived DNA was performed by UK Biobank using two genotyping arrays sharing 95% of marker content: Applied Biosystems UK BiLEVE Axiom Array (807,411 markers in 49,950 participants) and Applied Biosystems UK Biobank Axiom Array (825,927 markers in 438,427 participants) both by Affymetrix (Santa Clara, CA)[11]. Secondary use of the data was approved by the Massachusetts General Hospital Institutional Review Board (protocol 2013P001840) and facilitated through UK Biobank Applications 7089 and 21552.

The MGBB contains genotypic and clinical data from >105,000 patients who consented to broad-based research across 7 regional hospitals[12]. Baseline phenotypes were ascertained from the electronic medical record (EMR) and surveys on lifestyle, environment, and family history. Of the approximately 36,000 genotyped individuals, 27,778 samples had available probe raw intensity data (IDAT) files for mCA calling. Blood-derived DNA samples were genotyped using three versions of the Multi-Ethnic Genotyping Array (MEGA) SNP array offered by Illumina. Secondary use of the data was approved by the Massachusetts General Hospital Institutional Review Board (protocol 2020P000904).

The FinnGen project (https://www.finngen.fi/en), launched in 2017, covers the whole of Finland and aims to improve health of people around the world through genetic studies. The latest released version (R6) contains genotypic, demographic, and extensive health (e.g. national inpatient/outpatient registers since 1969/1998, cancer register since 1953, and drug reimbursement register since 1964) information from 269,077 Finnish individuals. Blood-derived DNA samples were genotyped using two versions of FinnGen ThermoFisher Axiom custom array (https://www.finngen.fi/en/researchers/genotyping) provided by the Thermo Fisher genotyping service facility.

Biobank Japan (BBJ) is a hospital-based registry that collected clinical, DNA, and serum samples from approximately 200,000 consented patients with one or more of 47 target diseases at a total of 66 hospitals between 2003-2007[17]. Blood DNA was genotyped in three batches using different arrays or set of arrays, namely: (1) a

combination of Illumina Infinium Omni Express and Human Exome; (2) Infinium Omni Express Exome v.1.0; and (3) Infinium Omni Express Exome v.1.2, which capture very similar SNPs. These analyses were approved by the ethics committees of RIKEN Center for Integrative Medical Sciences and the Institute of Medical Sciences, the University of Tokyo.

Chapter 2.2: CHIP calling methods and sensitivity analyses

GATK Mutect2[18] (https://software.broadinstitute.org/gatk) was used on BAM files for somatic variant calling of SNPs and INDELs using a "panel of normal samples" consisting of 100 randomly selected individuals less than 40 years old. The Mutect2 variant caller uses a Bayesian classifier for detection of low-allele fraction mutations requiring only a few supporting reads, followed by tuned filters that remove artifacts (i.e.: strand bias, poor mapping, triallelic sites, clustered position), and utilizes the panel of normal in addition to the gnomad germline resource as a reference for recurrent sequencing artefacts and germline variation to filter out these sites and thereby calls variants at sites with evidence for somatic variation. To filter out poor-quality somatic variant calls, raw somatic SNVs and indels are filtered to variants that PASS filters upon using FilterMutectCalls with default settings. Mutect2 caller was run separately for each sample with the same settings. Further additional filters were utilized to increase the probability of filtering to true pathogenic somatic CHIP variants, including: filtering to variants with variant allele fraction (VAF) > 2% (i.e.: variants showing evidence of clonal expansion), that were among a pre-specified list of putative pathogenic somatic CHIP variants across 74 genes linked to myeloid leukemias as previously described[5 19]. Variants were annotated with SNPeff. Samples were annotated as CHIP carriers if they carried any CHIP variant, and as Large CHIP carriers (variant allele frequency >10%), since larger CHIP clones have previously been more strongly associated with adverse clinical outcomes[20].

I performed additional sensitivity analyses as part of quality control to assess the change in somatic variant count across successive filters after filtering to FilterMutectCalls PASS variants with alt-allele read depth > 2, an alt-allele called in both forward and reverse strands, +/- a low germline probability via binomial probability of <1% of being inherited with variant allele fraction 50% (i.e.: BinomP(VAF, 0.5, 'less') < 0.01). **Figure 2.2.1** below visualizes the successive drop in somatic variant count per individual by age in the UK Biobank across several filters, showing:

1) all somatic variants that the aforementioned filters, which filtered out 90% of original Mutect2 somatic variant calls,
2) rare (allele frequency, AF<0.01 or NA in each ethnicity of the gnomad exomes and genomes, and overall in gnomad), deleterious variants (annotated as frameshift, transcript ablation, splice acceptor, splice donor, stop gained, start lost, missense deleterious as predicted by MetaSVM)
3) rare deleterious variants across 400 known leukemia genes
4) rare deleterious variants across 74 described CHIP genes

Furthermore, **Figure 2.2.2** shows the last row of Figure 2.2.1 with and without CHIP carriers, showing some preliminary evidence that even with the exclusion of CHIP carriers, there is some residual predicted deleterious somatic genetic variants associated with age across other somatic variant grouping strategies. In particular, further analyses of the overlap between CHIP, clonal hematopoiesis with unknown drivers, or CHUD, herein defined as the rare deleterious somatic variants across 400 leukemia genes with variant allele fraction > 10%, and autosomal mCAs found that 1.2% of carriers carry all three , 14.6% of autosomal mCA carriers also carry a CHUD variant, 9.5% of autosomal

mCA carriers carry CHIP, and 6.7% of CHUD carriers also have an autosomal mCA (**Figure 2.2.3**).

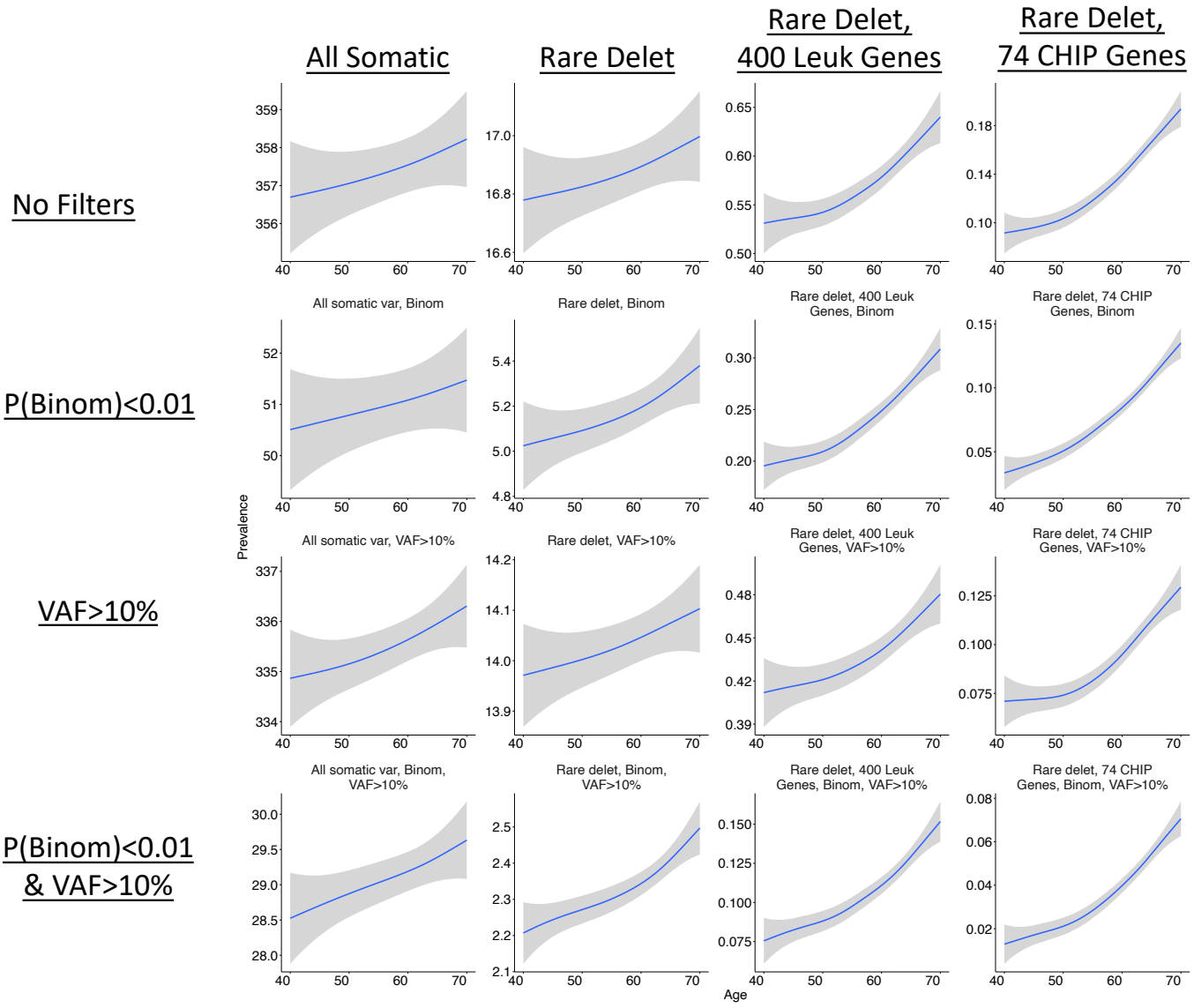


Figure 2.2.1: Number of somatic variants per individual by age across different variant filtration criteria. *Variant counts per individual are reported after filtering to FilterMutectCalls PASS variants with alt-allele read depth > 2, an alt-allele called in both forward and reverse strands, +/- a low germline probability via binomial probability of <1% of being inherited with variant allele fraction 50% (i.e.: BinomP(VAF, 0.5, 'less') < 0.01), and +/- VAF>10% (ie: expanded clones).*

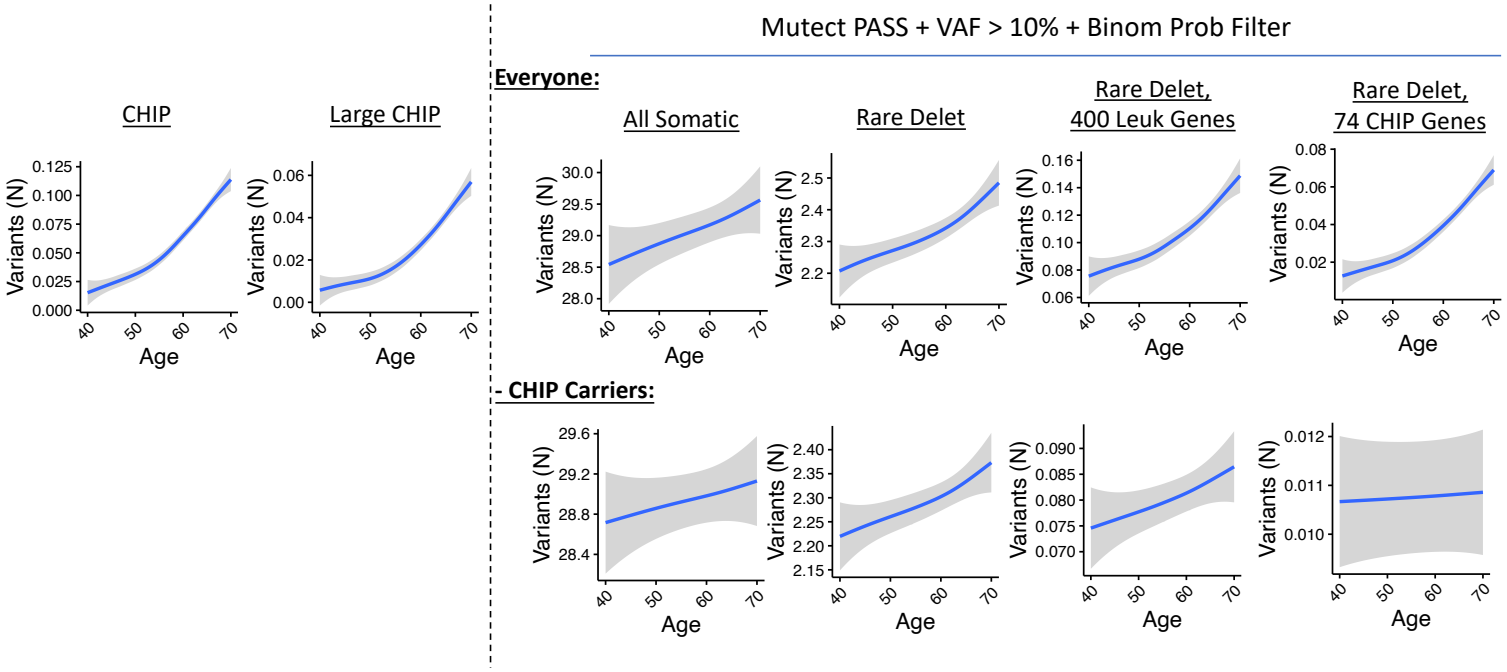


Figure 2.2.2: Number of somatic variants per individual by age across different variant filtration criteria and minus CHIP carriers. *The left panel shows the association of CHIP and Large CHIP calls with age among individuals in the UK Biobank. The right hand panel shows variant counts per individual after filtering to FilterMutectCalls PASS variants with alt-allele read depth > 2, an alt-allele called in both forward and reverse strands, a low germline probability via binomial probability of <1% of being inherited with variant allele fraction 50% (i.e.: BinomP(VAF, 0.5, 'less') < 0.01), and VAF>10% (ie: expanded clones). Associations of somatic counts with age with and without CHIP carriers are provided.*

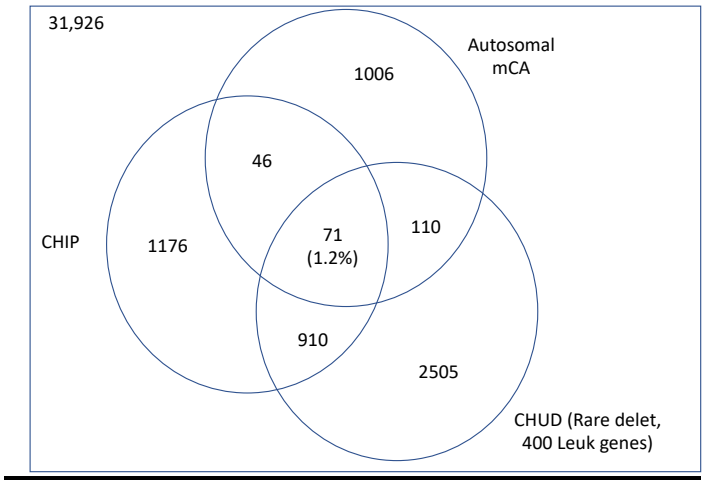


Figure 2.2.3: Overlap of CHIP, CHUD, and mCA carriers among UK Biobank individuals. *CHUD is herein defined as the rare deleterious somatic variants across 400 leukemia genes with variant allele fraction > 10%.*

Further sensitivity analyses was done to further understand how CHIP detection changes at various VAFs across sequencing depths (**Figure 2.2.4**). Sequence data was analyzed from 30 samples with CHIP from a previously published cohort[21] sequenced to >400x depth. The samples were bioinformatically down-sampled to different median depths. Across median depth ~40x (range 30-50x) as seen in the TOPMed WGS, excellent sensitivity was observed for CHIP variants with VAF>10%, while ~50% of CHIP variants with VAF 5-10% were called, and the majority of CHIP variants with VAF 2-5% were not reliably detected. Slightly better sensitivity is observed with the UK Biobank given a median sequencing depth of ~55x (**Figure 2.2.4**).

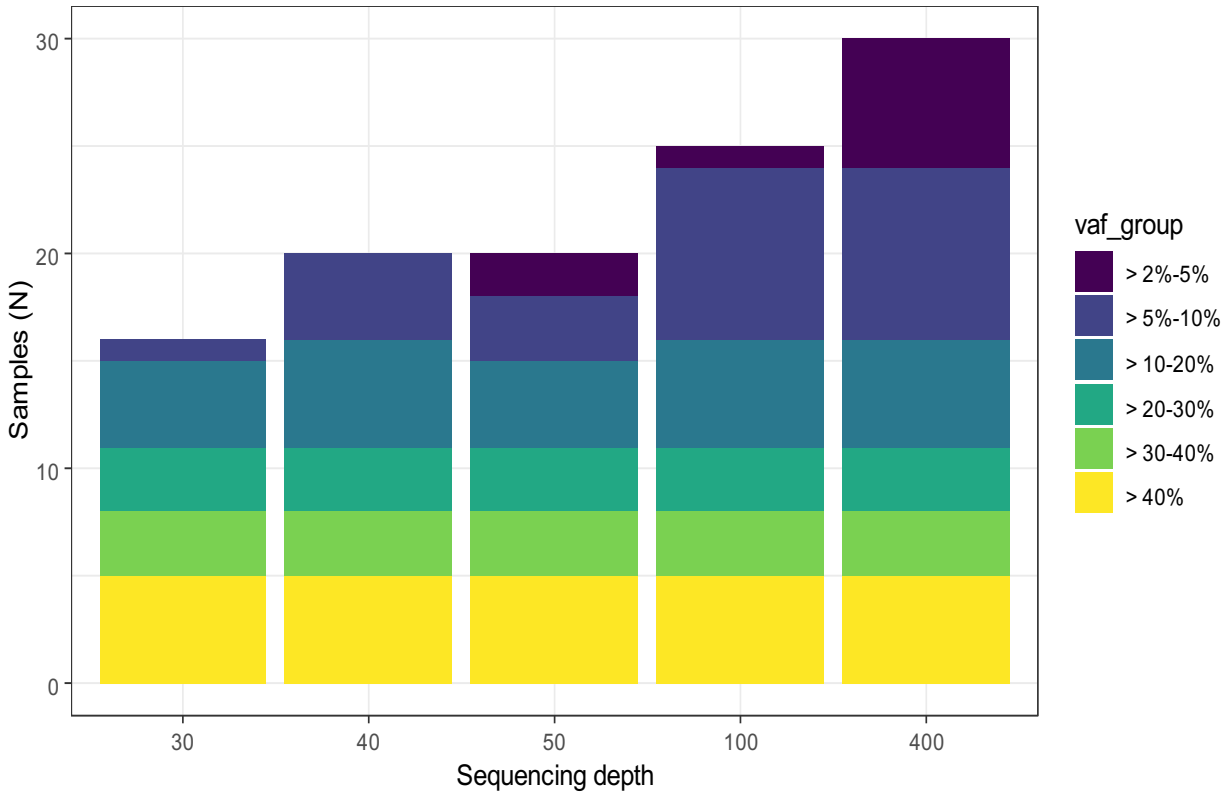


Figure 2.2.4: Sensitivity of CHIP detection at various variant allele fractions (VAFs) across sequencing depths. *A set of 30 samples from a previously published CHIP cohort[21] were bioinformatically down-sampled to different sequencing depths to enable better understanding of somatic variant detection sensitivity across different sequencing depths and VAFs[22].*

Further sensitivity analyses were performed comparing the efficacy of CHIP detection from WGS (~50x depth) versus WES (~100x depth) in the Jackson Heart Study cohort among ~2,000 samples with both WGS and WES performed. 33% of CHIP calls were shared between the two, while 33% of calls were detected by WGS but not WES (due to capture issues, in the JHS exomes, 6/12 TET2 exons were not included), and 33% of calls were detected by WES but not WGS due to lower depth. Out of 18 CHIP calls made by WGS that were included in the exome capture region, all 18 were also identified by WES. Further technical validation of 76 CHIP mutations across 72 samples from the

Women's Health Initiative (WHI) cohort was performed using targeted amplicon deep sequencing (1000x), replicating all 76/76 CHIP mutations from WGS.

Chapter 2.3: mCA calling methods and sensitivity analyses

mCA detection in the MGBB and in FinnGen was newly performed with the Mosaic Chromosomal Alterations (MoChA) software and pipeline (https://github.com/freeseek/mocha). Briefly, genotype intensities were transformed to $\log_2$(R ratio) (LRR) and B-allele frequency (BAF) values to estimate total and relative allelic intensities, respectively, as previously described[23]. Detection of mCAs in the MGB Biobank was performed using raw IDAT intensity files from the Illumina Multi-Ethnic Global Array (MEGA), genotyped using the Illumina GenCall algorithm. The resulting GTC genotype files were converted to VCF files using the bcftools gtc2vcf plugin (https://anaconda.org/bioconda/bcftools-gtc2vcf-plugin). Phasing across the whole cohort was performed using SHAPEIT4[24] in windows of a maximum of 20 centimorgans with 2 centimorgans of overlap between consecutive windows. Genotype phase was ligated across windows using bcftools concat (https://github.com/samtools/bcftools). mCA detection in the MGB Biobank was performed with MoChA[2 10] using a pipeline with default parameters (https://github.com/freeseek/mocha/tree/master/wdl). We excluded 164 samples with phased BAF auto-correlation >0.05, indicative of contamination or other potential sources of poor DNA quality, and 72 samples with phenotype-genotype sex discordance (**Figure 2.3.1**). We removed likely germline copy number polymorphisms (lod_baf_phase <20), constitutional or inborn duplications (mCAs <2Mb with relative coverage >2.25, and mCAs 2-10Mb with relative coverage >2.5) and deletions (filtering out mCAs with relative coverage <0.4) (**Figure 2.3.2**).

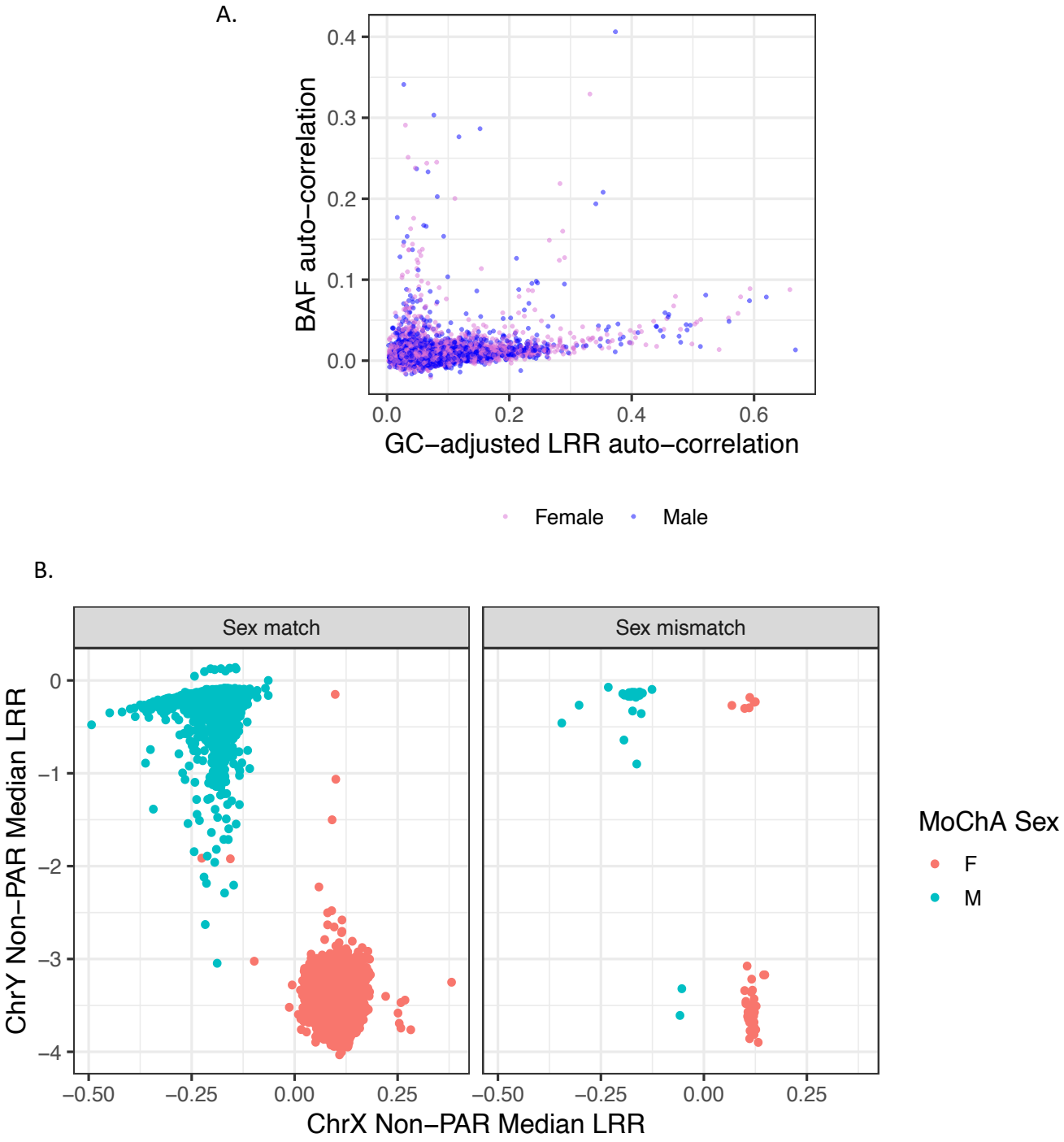


Figure 2.3.1: MGB Biobank mCA sample quality control analyses. *A. plotting sample-level phased B allele frequency (BAF) auto-correlation across consecutive phased heterozygous sites versus Log R Ratio (LRR) of intensities using local GC content. B. Showing sex mismatches between MoChA-derived sex imputed using the chrX nonPAR region versus reported sex.*

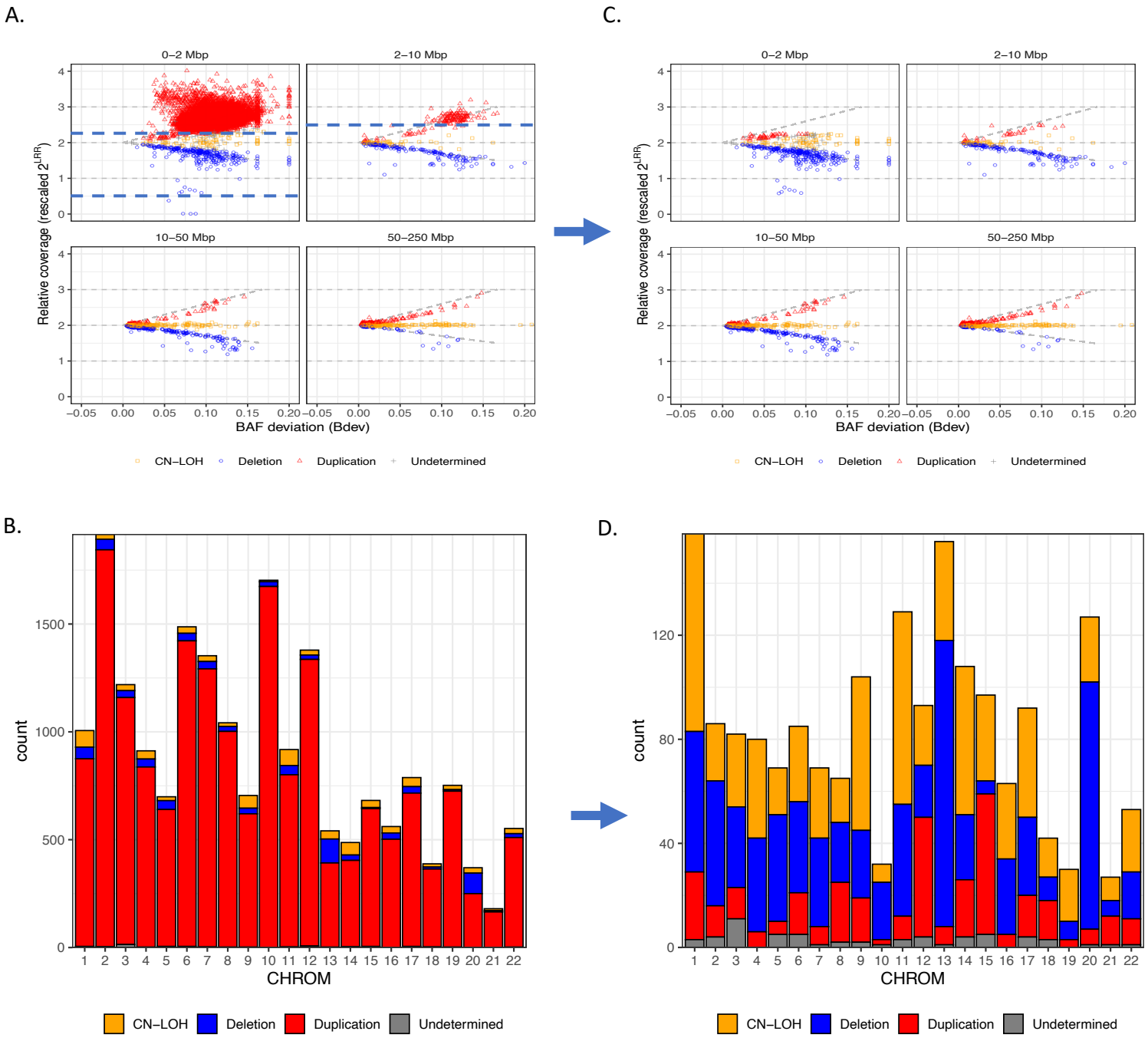


Figure 2.3.2: MGB Biobank mCA variant quality control analyses. *Plots A. and B. represent mCAs carried among the quality-control filtered sample set, and after basic variant quality control filters including removal of likely germline variants (LOD_BAF_PHASE <20 or mCAs annotated as known CNPs). Plots C. and D. reflect additional variant quality control filters to remove constitutional duplications (0-2Mbp mCAs with relative coverage >2.25 and 2-10Mbp mCAs with relative coverage >2.5) and remove constitutional deletions (mCAs with relative coverage <0.5).*

Mosaic chromosomal alteration (mCA) detection in the UK Biobank was as described previously[2 10]. Briefly, genotype intensities were transformed to log2(R ratio) (LRR) and B-allele frequency (BAF values) to estimate total and relative allelic intensities, respectively. Re-phasing was performed using Eagle2[25] and mCA calling was performed by leveraging long-range phase information to search for local imbalances between maternal and paternal allelic fractions. Possible constitutional duplications and low-quality calls were filtered out and cell fraction was estimated as previously described[2]. UK Biobank mCA calls were obtained from dataset Return 2062 generated from UK Biobank application 19808.

The detection of mCAs in the BBJ is as described previously[9]. Briefly, genotyping intensity data was analysed across variants shared between the three primary arrays, and used to compute BAF and LRR. Phasing was performed using the Eagle2 software. Mosaic events were called as previously described[2].

Across all studies, expanded mCA refers to the presence of at least one detectable mCA present in >10% of circulating leukocytes (e.g., cell fraction >10%). A 10% cell fraction threshold was employed since this has been previously linked to greater clonal haematopoiesis-related risk for incident mortality[26] and myocardial infarction[20], additionally this subset of it was observed to most strongly associate with phenotypes in the UK Biobank including aberrant blood cell counts, incident hematologic cancer, and incident infections. Autosomes and sex chromosomes were also separately considered; only autosomal mCAs were available for BBJ.

Chapter 3: Phenome-wide association of CHIP and mCAs

Published across multiple papers[13-16] as:

Bhattacharya R*, **Zekavat SM***, Haessler J, et al. Clonal Hematopoiesis Is Associated With Higher Risk of Stroke. *Stroke* 2021:STROKEAHA. 121.037388.

Yu B, Roberts MB, Raffield LM, **Zekavat SM**, et al. Supplemental Association of Clonal Hematopoiesis With Incident Heart Failure. *J Am Coll Cardiol* 2021;78(1):42-52. doi: 10.1016/j.jacc.2021.04.085 [published Online First: 2021/07/03]

Zekavat SM, Lin SH, Bick AG, et al. Hematopoietic mosaic chromosomal alterations increase the risk for diverse types of infection. *Nat Med* 2021;27(6):1012-24. doi: 10.1038/s41591-021-01371-0 [published Online First: 2021/06/09]

Zekavat SM, Viana-Huete V, Zuriaga MA, et al. TP53-mediated clonal hematopoiesis confers increased risk for incident peripheral artery disease. *medRxiv* 2021:2021.08.22.21262430. doi: 10.1101/2021.08.22.21262430

Please refer to the papers above for additional methodological details, including phenotype definitions, cohort descriptions, and genotyping platforms.

Chapter 3.1: Association of CHIP and mCAs with age, blood counts, and hematological cancer

Association of CHIP with blood counts and hematological cancer:

After excluding individuals with a known history of hematologic malignancy at enrollment, we identified 37,657 unrelated individuals from the UK Biobank (UKB) and 12,465 individuals from Mass General Brigham Biobank (MGBB) with whole exome sequencing data available for downstream analysis. Using a previously validated somatic variant detection algorithm[27], we identified 2,194 (5.8%) and 657 (5.4%) CHIP carriers in the UKB and MGBB, respectively (**Table 3.1.1**). Demographic and clinical characteristics of these individuals, stratified by CHIP status, are depicted in **Table 3.1.2**. CHIP carriers tended to be older, male, previous smokers, and have a history of coronary

artery disease, hypertension, and hyperlipidemia (two-tailed chi-squared and Wilcoxon-rank sum P< 0.05). The association of CHIP with age is provided in **Figure 2.2.2**.

We first replicated known CHIP associations[27] with white blood cell (Beta 0.09 SD; 95% CI 0.05-0.13; P=1.6x10^{-5}), monocyte (Beta 0.05 SD; 95% CI 0.01-0.09; P=0.009), neutrophil (Beta 0.10 SD; 95% CI 0.06-0.14; P=2.1x10^{-6}), and platelet counts (Beta 0.07 SD; 95% CI 0.03-0.11; P=0.0005) in UKB, with larger CHIP clone size as measured by variant allele fraction (VAF) having stronger effects on blood counts (**Figure 3.1.1**). Consistent with the existing literature[6 27], CHIP also associated with incident hematologic malignancy (HR 2.20; 95% CI 1.70-2.85; P=1.8x10^{-9}) - specifically for acute myeloid leukemia (HR 8.08; 95% CI 4.36-14.97; P=3.2x10^{-11}), myeloproliferative neoplasms (HR 5.89; 95% CI 3.69-9.89; P=9.7x10^{-14}), and polycythemia vera (HR 12.37; 95% CI 4.85-31.54; P=1.4x10^{-7}). This risk increased with larger VAF (**Figure 3.1.2**).

Table 3.1.1 - CHIP gene carrier count by cohort. *Splicing Factor Mutations refer to the following CHIP genes: LUC7L2, PRPF8, SF3B1, SRSF2, U2AF1, and ZRSR2. Large CHIP refers to mutations with variant allele frequency > 10%.*

	UKBB (N=37,657)		*MGBB (N=12,465)*	
	All CHIP	Large CHIP	All CHIP	Large CHIP
CHIP (%)	2194 (5.8)	911 (2.4)	657 (5.3)	314 (2.5)
>1 CHIP Mutation (%)	191 (0.5)	70 (0.2)	55 (0.4)	16 (0.1)
DNMT3A (%)	1401 (3.8)	489 (1.4)	311 (2.6)	144 (1.2)
TET2 (%)	347 (1.0)	181 (0.5)	132 (1.1)	61 (0.5)
JAK2 (%)	17 (0.0)	17 (0.0)	5 (0.0)	5 (0.0)
ASXL1 (%)	152 (0.4)	100 (0.3)	47 (0.4)	21 (0.2)
Splicing Factor Mutation (%)	49 (0.1)	28 (0.1)	17 (0.1)	8 (0.1)
TP53 (%)	36 (0.1)	11 (0.0)	20 (0.2)	12 (0.1)
PPM1D (%)	32 (0.1)	12 (0.0)	32 (0.3)	13 (0.1)
TP53 or PPM1D (%)	68 (0.2)	23 (0.1)	52 (0.4)	25 (0.2)

Table 3.1.2 - Demographic and clinical characteristics for CHIP carriers and controls in the UK and Mass General Brigham Biobanks. *P-values reflect chi-square tests comparing CHIP carriers to controls across each phenotypic category.*

	UK Biobank			*MGB Biobank*		
	-CHIP	**+CHIP**	**p**	**-CHIP**	**+CHIP**	**p**
n	35463	2194		11808	657	
age (mean (SD))	56.81 (7.84)	60.59 (6.57)	<0.001	46.13 (14.65)	60.12 (12.05)	<0.001
Sex = Male (%)	16379 (46.2)	1042 (47.5)	0.242	4937 (41.8)	304 (46.3)	0.027
Race (%)			NA			0.002
White	35463 (100.0)	2194 (100.0)		9449 (80.0)	566 (86.1)	
Black				723 (6.1)	27 (4.1)	
Asian				465 (3.9)	19 (2.9)	
Other				474 (4.0)	12 (1.8)	
Unknown				697 (5.9)	33 (5.0)	
Smoking Status (%)			<0.001			<0.001
Current	3027 (8.5)	220 (10.0)		288 (2.4)	17 (2.6)	
Previous	12664 (35.7)	900 (41.0)		3662 (31.0)	261 (39.7)	
Never	19772 (55.8)	1074 (49.0)		7183 (60.8)	351 (53.4)	
Alcohol intake (drinks in last 4wk) (mean (SD))	11.37 (9.89)	11.67 (10.11)	0.156			
Exercise frequency (days in last 4wk) (mean (SD))	8.34 (6.38)	8.45 (6.43)	0.591			
Townsend Deprivation Index (mean (SD))	-1.55 (2.81)	-1.65 (2.75)	0.137			
Significant life stressor in last 2y (%)	16915 (47.8)	1030 (47.1)	0.535			
Handfulls of sweets/day (mean (SD))	1.09 (1.17)	0.93 (1.19)	0.399			
Vegetable servings/day (mean (SD))	1.08 (0.54)	1.02 (0.49)	0.283			
BMI (mean (SD))	27.39 (4.76)	27.48 (4.55)	0.414	28.14 (6.35)	28.55 (6.33)	0.132
Prevalent Type 2 Diabetes Mellitus (%)	956 (2.7)	70 (3.2)	0.189	505 (4.3)	40 (6.1)	0.035
Prevalent Coronary Artery Disease (%)	2040 (5.8)	171 (7.8)	<0.001	378 (3.2)	42 (6.4)	<0.001
Prevalent Hypertension (%)	10650 (30.0)	782 (35.6)	<0.001	1893 (16.0)	193 (29.4)	<0.001
Prevalent Hypercholesterolemia (%)	6159 (17.4)	448 (20.4)	<0.001	1739 (14.7)	172 (26.2)	<0.001

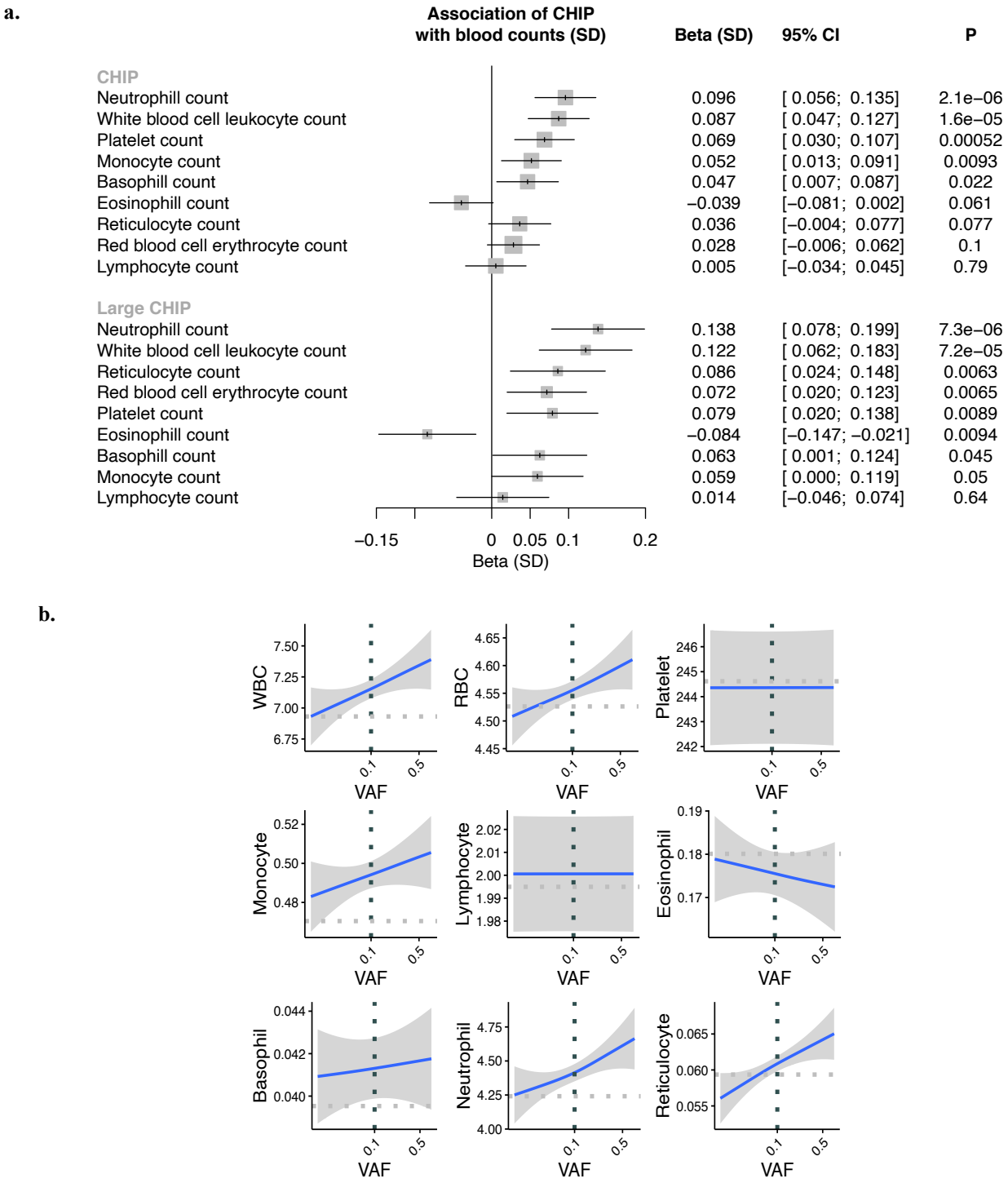


Figure 3.1.1: Association of CHIP with blood counts among individuals without prevalent hematologic malignancy in the UK Biobank. *Blood counts were acquired at time of blood draw for whole exome sequencing. a) Association of CHIP and Large CHIP with normalized blood counts (SD). Associations are adjusted for age, age^2, sex, smoking status, and the first ten principal components of genetic ancestry. b) Association of CHIP variant allele frequency (VAF) with blood counts (in units of 10^9 cells/L). The gray horizontal dotted lines reflect average counts across non-CHIP carriers. The vertical black dotted line reflects the cutoff VAF for Large CHIP (VAF>0.1). CHIP = clonal hematopoiesis of indeterminate potential; VAF = variant allele fraction*

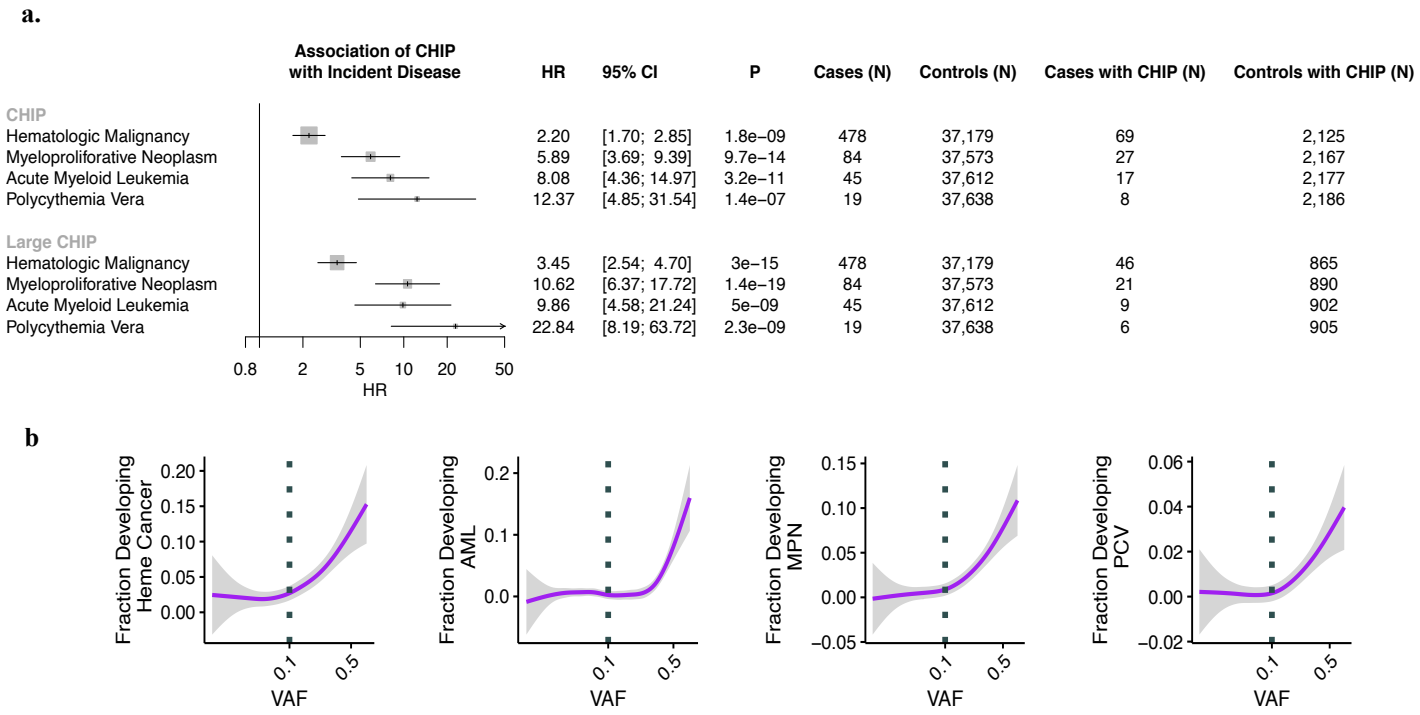


Figure 3.1.2: Association of CHIP (a) and VAF (b) with incident hematologic malignancy among individuals without prevalent hematological malignancy in the UK Biobank. *Associations are adjusted for age, age^2, sex, smoking status, Townsend deprivation index, and the first ten principal components of genetic ancestry. CHIP = clonal hematopoiesis of indeterminate potential; VAF = variant allele fraction*

Association of mCAs with age, blood counts, and hematological cancer:

Population characteristics and mCA prevalence

A total of 767,891 unrelated, multi-ethnic individuals across the UK Biobank (UKB) (N=444,199), Mass General Brigham Biobank (MGBB) (22,461), FinnGen (N=175,690), and BioBank Japan (BBJ) (N=125,541) passing genotype and mCA quality control criteria (**Figure 2.3.1-2**) were analyzed (**Table 3.1.3**). Among the UKB participants, mean age at DNA collection was 57 (standard deviation [SD] 8) years, 204,579 (46.1%) were male, 188,875 (45.0%) were prior or current smokers, and 66,551 (15.0%) had a history of solid cancer. In the MGBB, mean age was 55 (SD 17) years, 10,306 (45.9%)

were male, 9,094 (40.5%) were prior or current smokers, and 6,080 (27.1%) had a history of solid cancer. In FinnGen, mean age was 53 (SD 18) years, 71,000 (40.4%) were male, 42.7% were prior or current smokers (when smoking status was available), and 31,855 (18.1%) had a history of solid cancer. In BBJ, mean age was 65 (SD 12) years, 72,186 (57.5%) were male, and 66,913 (53.3%) were prior or current smokers, and 25,987 (20.7%) had a history of solid cancer.

Table 3.1.3: Baseline summary statistics across the UK Biobank, MGB Biobank, FinnGen, and Biobank Japan among individuals analyzed.

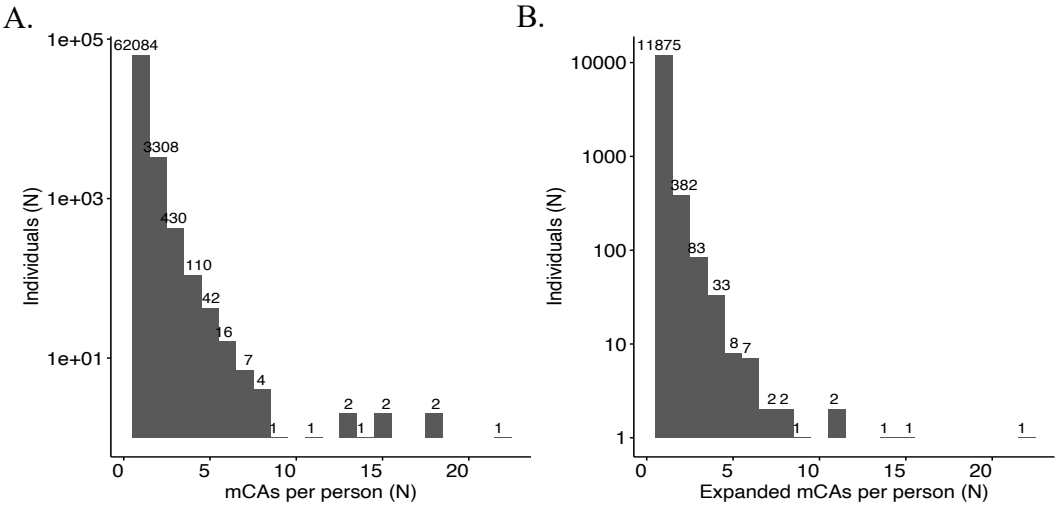
	UK Biobank	***MGB Biobank***	***FinnGen****	***Biobank Japan***
N	444,199	22,461	175,690	125,541
Age of DNA collection (mean (SD))	56.5 (8)	55.0 (16.8)	53.4 (18.4)	64.6 (12.4)
Sex (Male (%))	204,579 (46.1%)	10,306 (45.9%)	71,000 (40.4)	72,186 (57.5%)
Prior or Current Smoker (%)	188,875 (45.0%)	9,094 (40.5%)	30,554 (42.7)	66,913 (53.3%)
Race	White: 417,828 (94.1%) Asian: 10,277 (2.3%) Black: 7,173 (1.6%) Mixed: 2,634 (0.6%) Other: 4,160 (0.9%) Unknown 187 (0.04%)	White: 18,933 (84.3%) Asian: 569 (2.5%) Black: 1,056 (4.7%) Other: 744 (3.3%) Unknown: 1,159 (5.2%)	White: 175,690 (100%)	Asian: 125,541 (100%)
BMI (mean (SD))	27.4 (4.8)	28.5 (6.2)	NA	23.4 (3.7)
Prevalent Solid Cancer	66,551 (15.0%)	6,080 (27.1%)	31,855 (18.1%)	25,987 (20.7%)
Prevalent Type 2 Diabetes	10,835 (2.4%)	1,782 (7.9%)	22,326 (13.2%)	31,636 (25.2%)
Prevalent Coronary Artery Disease	25,287 (5.7%)	3,908 (17.4%)	19,474 (11.1%)	23,099 (18.4%)
Prevalent Hypertension	129,888 (29.2%)	11,010 (49.0%)	NA	37,913 (30.2%)
Prevalent Hypercholesterolemia	66,483 (15.0%)	9,881 (44.0%)	8,583 (5.2%)	35,026 (27.9%)

Table 3.1.4: mCA counts by cohort.

	UK Biobank	***MGB Biobank***	***FinnGen***	***Biobank Japan***
N	444,199	22,461	175,690	125,541
Any mCA(%)	66,011 (14.9)	3,784 (16.8)	22,040 (12.5)	NA
Autosomal mCA (%)	15,350 (3.5)	1,025 (5.2)	3,164 (2.0)	20,440 (16.3)
ChrX (%)	12,265 (5.1)	820 (7.0)	7,058 (6.8)	NA
ChrY (%)	41,284 (20.1)	2,201 (22.0)	12,599 (18.0)	NA
Any expanded mCA (%)	12,398 (3.2)	1,026 (5.2)	9,558 (5.9)	NA
expanded autosomal mCA (%)	2,985 (0.8)	337 (1.8)	1,620 (1.0)	1,676 (1.3%)
expanded ChrX (%)	397 (0.2)	44 (0.2)	479 (0.5)	NA
expanded ChrY (%)	9168 (4.5)	669 (3.4)	7663 (11.8)	NA

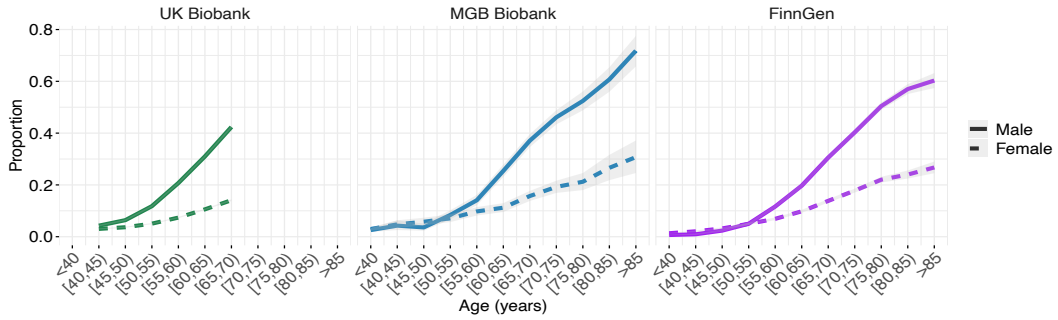
In the UKB, among 444,199 unrelated individuals without a known history of hematologic malignancy, 66,011 (14.9%) carried an mCA (15,350 autosomal) and 12,398 (3.2%) carried an expanded mCA clone, defined as an mCA mutation present in at least 10% of peripheral leukocytes (2,985 autosomal) (**Table 3.1.4**). While most of carriers only carried one mCA, 6% of individuals carried between 2 to 22 non-overlapping mCAs (**Figure 3.1.3**). In the MGBB, across 22,461 unrelated individuals without a history of hematologic cancer, 3,784 (16.8%) carried an mCA (1,025 autosomal) and 1,026 (5.2%) carried an expanded mCA clone (337 autosomal). In FinnGen, across 175,690 individuals without a history of hematologic cancer, 22,040 (12.5%) carried an mCA (3,164 autosomal), and 9,558 (5.9%) carried an expanded mCA clone (1,620 autosomal). In BBJ, across 125,541 individuals without a history of hematologic cancer, only autosomal mCAs were available, with 20,440 carriers (16.3%) and 1,676 (1.3%) that carried an expanded clone (**Table 3.1.4**).

Figure 3.1.3: Total number of mCAs (A) and expanded mCAs (B) per individual in the UK Biobank for mCA carriers.

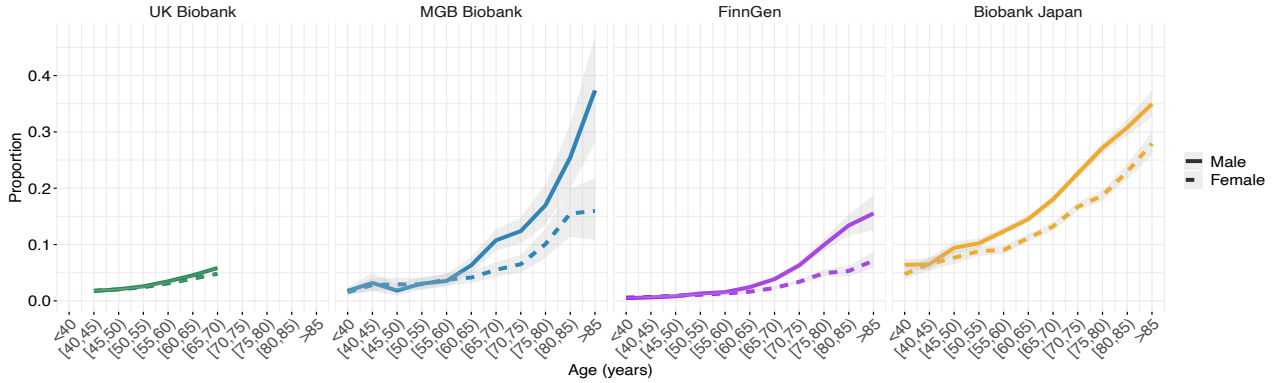


Consistent with previous reports, the frequency of mCAs increased with age and was higher among men (**Figure 3.1.4-5**). The frequency of expanded autosomal mCAs across the UKB, MGBB, FinnGen, and BBJ cohorts combined was 0.27% among individuals <40 years, 0.52% among 40-60 years, 1.5% among 60-80 years, and 4.6% among those greater than 80 years.

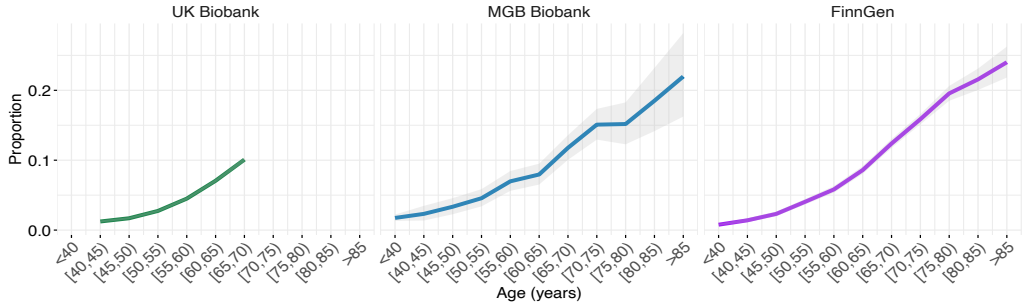
A. Any mCA:



B. Autosomal mCA:



C. ChrX:



D. ChrY:

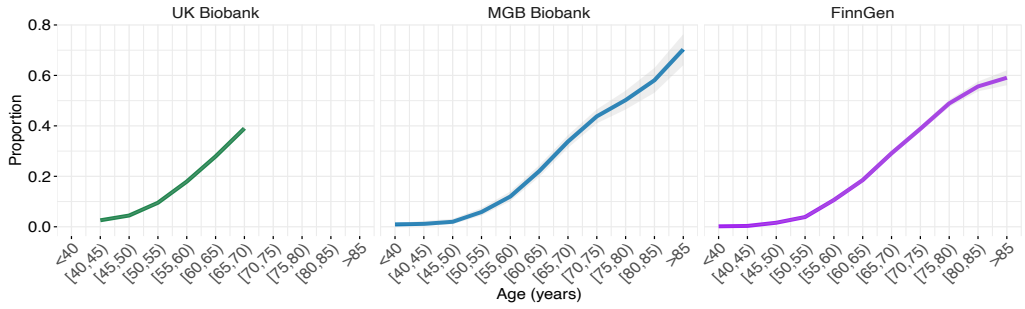


Figure 3.1.4: Prevalence of mCA categories by age bin across cohorts.

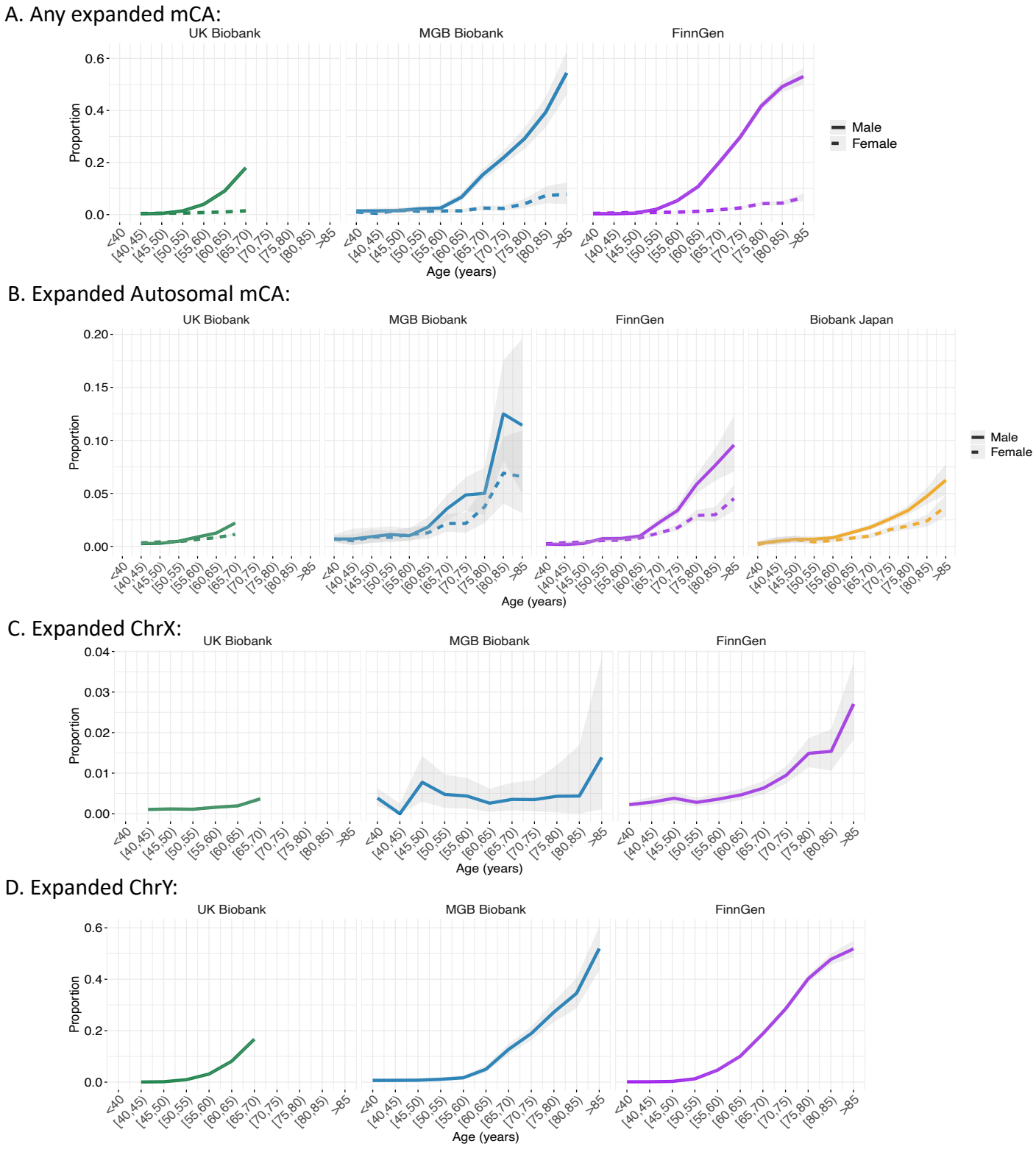


Figure 3.1.5: Prevalence of mCA categories by age bin across cohorts.

Association of mCAs with hematologic traits

We observed a striking association of mCA cell fraction with aberrant cell blood counts acquired at the same visit as blood for genotyping (**Figure 3.1.6**). Increased mCA cell fraction was associated with overall increased white blood cell count with general consistency across the cell differential components, with distinct inflections at around cell fraction of 0.1 (**Figure 3.1.6**). The strongest association across all mCAs groupings (autosomal/chrX/chrY) with blood counts was between expanded autosomal mCAs and increased lymphocyte count at enrollment (Beta 0.40 SD or 0.25 $x10^9$ cells/L; 95% CI 0.36 to 0.44 SD; P=$4.2x10^{-84}$) (**Figure 3.1.7**).

Similarly, incident hematologic cancer risk was also strongly dependent on cell fraction, with a distinct inflection at cell fraction of 10% (**Figure 3.1.8**). We reproduced the associations of mCAs with hematologic cancers with similar effects as previously described in the UKB[2 10]. We found that expanded autosomal mCAs with cell fraction >10% were most strongly associated with incident hematologic cancer (**Figure 3.1.8**), with the strongest association being for incident chronic lymphocytic leukemia (HR 121.9; 95% CI 93.6 to 158.9; P=$4.2x10^{-277}$); although an association with myeloid leukemia was also present (HR 12.3; 95% CI 7.7 to 19.7; P=$2.3x10^{-25}$) (**Figure 3.1.9**). While expanded chrX and chrY mCAs were also associated with chronic lymphocytic leukemia, their effects were considerably lower (chrX: HR 24.1, 95% CI 5.8 to 99.9, P=$1.1x10^{-5}$ and chrY: HR 2.0, 95% CI 1.0 to 4.0, P=0.038) (**Figure 3.1.9**).

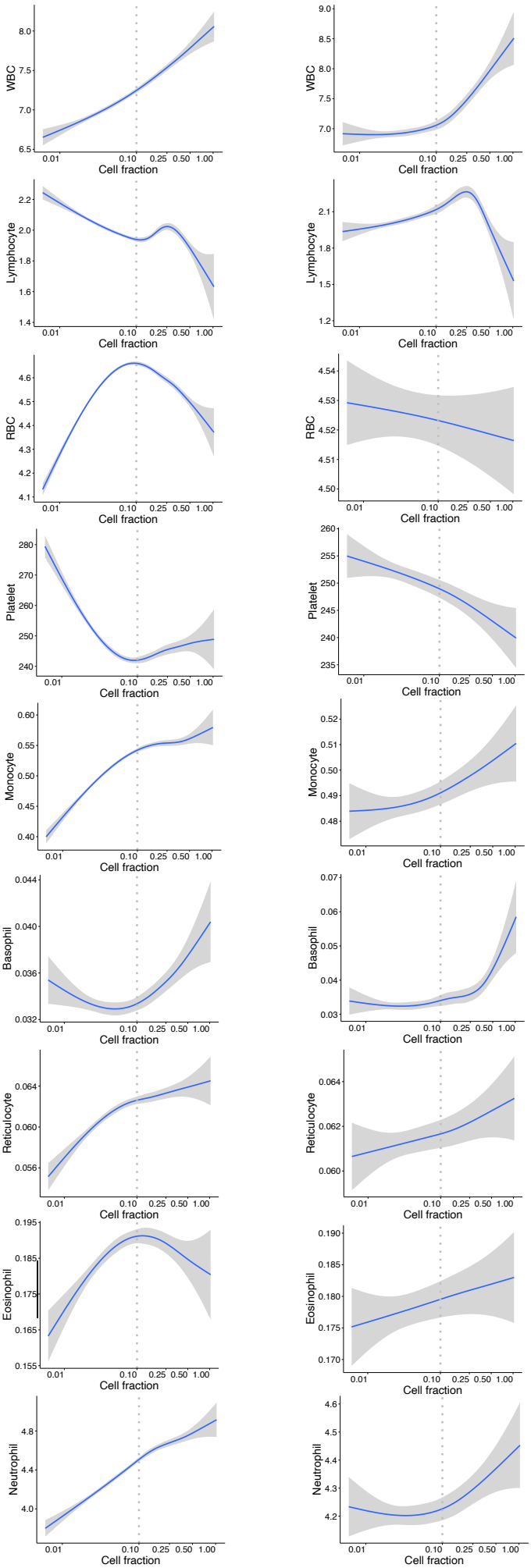


Figure 3.1.6: Associations of mCA cell fraction with blood counts (in units of 10^9 cells/L) among individuals without prevalent hematologic cancer at time of blood draw for genotyping and cell count measurement. *The dotted vertical line at cell fraction of 0.10 represents the cutoff for the expanded mCA definition.*

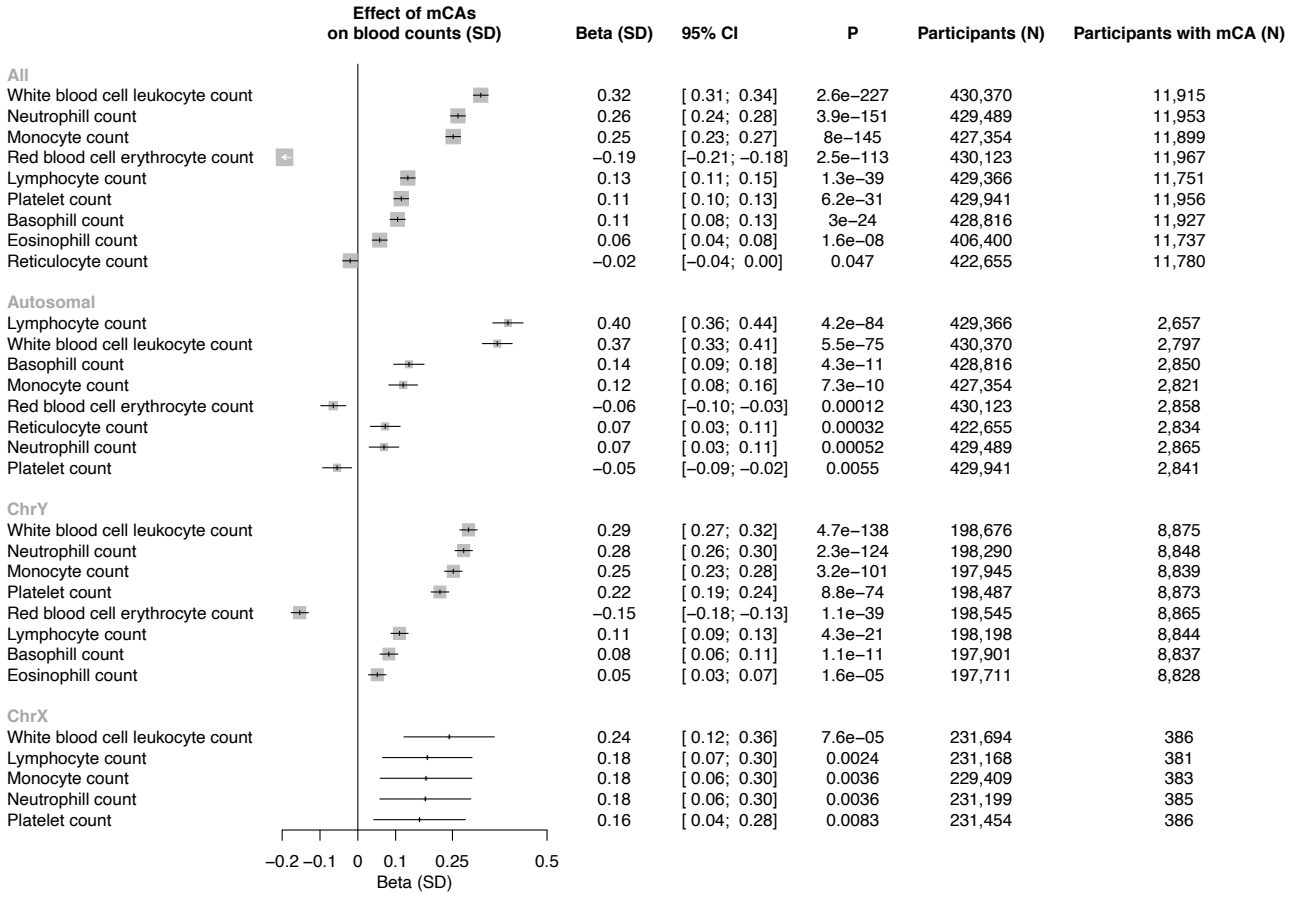


Figure 3.1.7: Association of blood counts with expanded mCAs. *Associations are adjusted for age, age^2, sex, smoking status, and principal components of ancestry.*

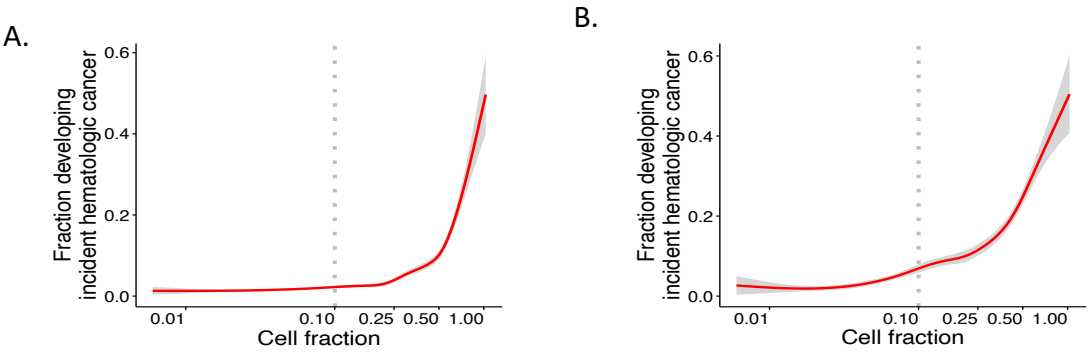


Figure 3.1.8: Association of A) all mCA and B) autosomal mCA cell fraction with incident hematologic cancer. *The dotted vertical line at cell fraction of 0.1 shows the cutoff point for expanded mCAs (defined as mCAs with cell fraction >10%).*

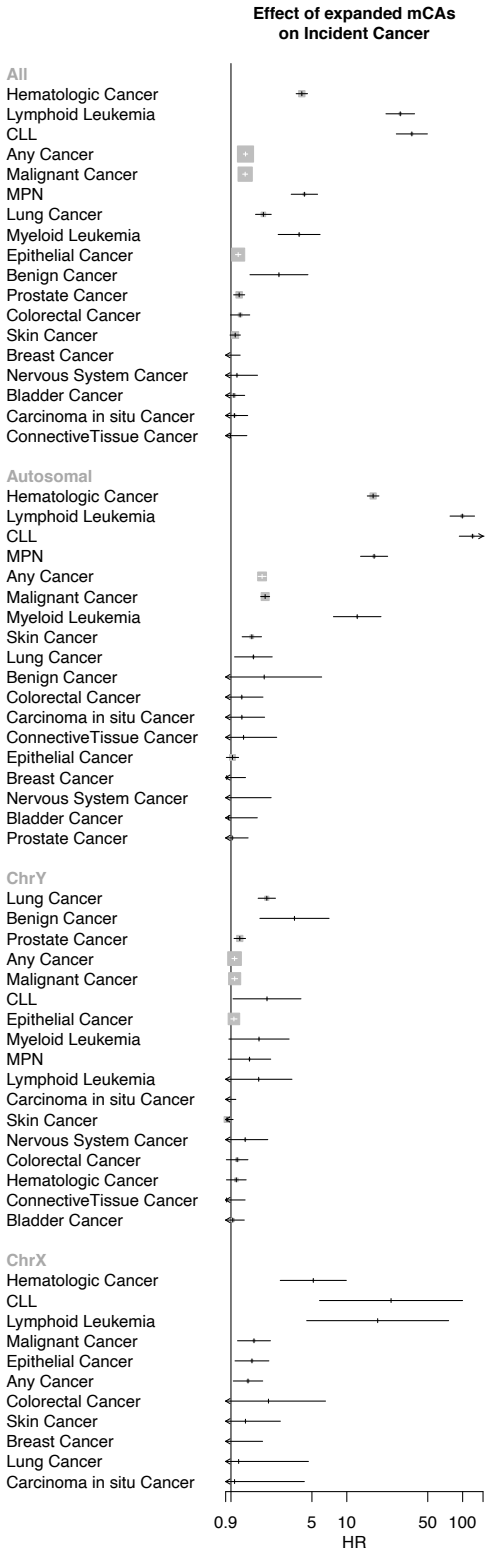


Figure 3.1.9: Association of expanded mCA categories (ie: with cell fraction>10%) with incident cancer in the UK Biobank. *Analyses are adjusted for age, age^2, sex, smoking status, and principal components of ancestry. Individuals with a history of hematologic cancer at enrollment were removed from analysis. CLL = chronic lymphocytic leukemia, MPN = myeloproliferative neoplasm*

Chapter 3.2: Comparative phenome-wide association of CHIP and mCAs

Numerous associations have been identified between clonal hematopoiesis, hematologic malignancy, and non-malignant diseases linked to aging. The present datasets assembled permit a comprehensive and well powered phenome-wide analysis of CHIP and mCAs. Cohorts incorporated in the PheWAS analyses below include the UK Biobank for CHIP (N=37,657) and also the UK Biobank (N=448,100) for mCAs. Here, I performed phenome-wide association of CHIP and mCAs across all of the 1,866 hierarchical phenotypes defined from the Phecode Map 1.2[28] ICD-9 (https://phewascatalog.org/phecodes) and ICD-10 (https://phewascatalog.org/phecodes_icd10) phenotype groupings[29]. Associations with incident phenotypes were performed using Cox proportional hazards models after excluding individuals with the corresponding diagnosis at or prior to enrollment. Models were adjusted for age, age^2, sex, smoking status (25-factor smoking status for the UK Biobank and current/prior/never smoker for other cohorts), and the first ten principal components of genetic ancestry. Analysis was performed across disease phenotypes with at least 9 cases with CHIP or mCA carriers available. Statistical significance was defined using false discovery rate <0.05.

Given the novel suggestive associations observed in the UK Biobank between CHIP and incident cardiovascular phenotypes (i.e.: cardiac arrest and ventricular fibrillation, aortic aneurysms, peripheral vascular disease) (**Figure 3.2.1**), and between autosomal mCAs and incident infectious diseases (i.e.: sepsis, pneumonia) (**Figure 3.2.2**), further analyses were performed meta-analyzing across multiple cohorts to further assess the association of CHIP with 1) pan-vascular atherosclerosis, 2) heart failure, and 3) stroke, as well as the association of mCAs with infectious diseases.

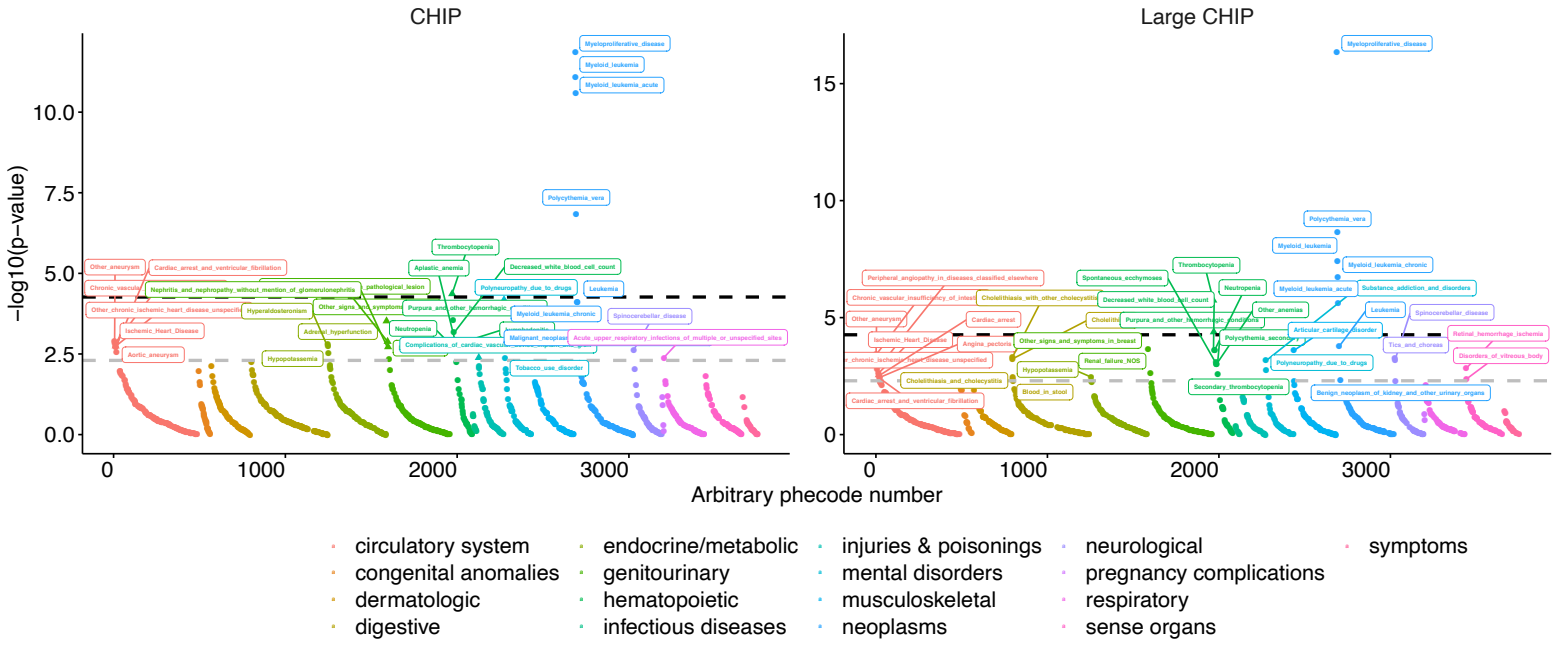


Figure 3.2.1: Association of CHIP and large CHIP with 1,866 incident phenotypes. *Dotted black line reflects the Bonferroni significance cutoff based on the number of incident phenotypes with at least 9 incident case CHIP carriers. Labeled are phenotypes with P<0.05 of association.*

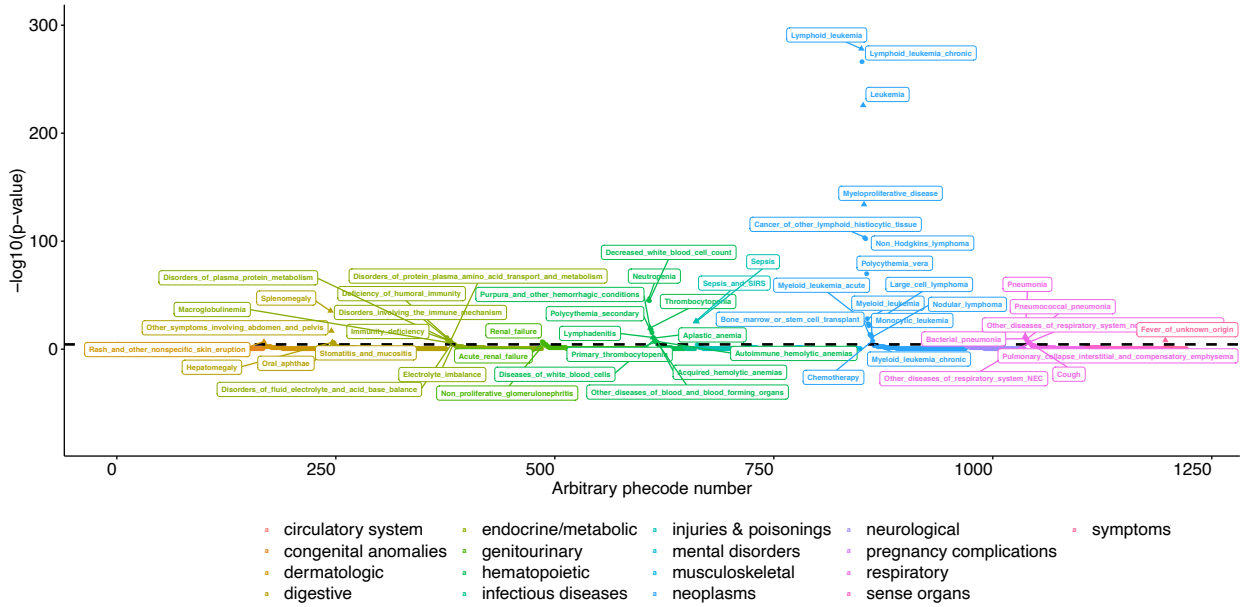


Figure 3.2.2: Association of autosomal mCAs with 1,866 incident phenotypes. *Dotted black line reflects the Bonferroni significance cutoff based on the number of incident phenotypes with at least 9 incident case CHIP carriers. Labeled are phenotypes passing the Bonferroni multiple-testing threshold for significance.*

Chapter 3.3: Association of CHIP with peripheral artery disease (PAD) and pan-vascular atherosclerosis

Peripheral artery disease (PAD) is a leading cause of cardiovascular morbidity and mortality worldwide, and age is among its strongest risk factors. PAD associates with an extremely high cardiovascular mortality and unmitigated can progress to limb loss[30]. CHIP associates with coronary artery disease in multiple studies[20 31]. However, whether CHIP links with increased risk of atherosclerosis in other arterial beds, such as through PAD is unknown. Here, we leveraged 50,122 whole exome sequences from two genetic biobanks (UK Biobank, MassGeneral Brigham Biobank) and tested whether CHIP was associated with increased risk of PAD and atherosclerosis across multiple arterial beds, and additionally whether these associations varied by putative CHIP driver gene. Based on these results, we then performed functional analyses in *Ldlr*-null mice transplanted with 20% *Trp53*-/- bone marrow cells, a murine model of atherosclerosis and clonal hematopoiesis driven by *TP53* mutations.

Using available electronic health record (EHR) data and a previously validated PAD definition[32], we identified 338 and 419 incident PAD cases in UKB and MGBB, respectively. CHIP associated with a 58% increased risk of incident PAD in the UKB (HR_{UKB} = 1.58, 95% CI: 1.11-2.25; P=0.01, **Figure 3.3.1**), results that were replicated in MGBB (Overall HR = 1.66, 95% CI: 1.31-2.11; $P=2.4x10^{-5}$). We then sought to evaluate whether those with larger CHIP clone sizes (i.e., higher VAF) had greater risk for PAD, as larger CHIP clones associate more strongly with adverse clinical outcomes[20]. We observed a graded relationship between CHIP VAF and PAD, as those with a VAF > 10% had even greater risk for an incident PAD event (Overall HR = 1.97, 95% CI: 1.44-

2.71; P=2.3x10^{-5}, **Figure 3.3.1**). Additional sensitivity analyses, including propensity score adjustment and a marginal structural Cox proportional hazards model estimated through stabilized inverse-probability-treatment-weight revealed similar results in the UKB (**Figure 3.3.2**). Subsequent analyses showed no significant interaction between CHIP status and either age, sex, or smoking status on incident PAD risk.

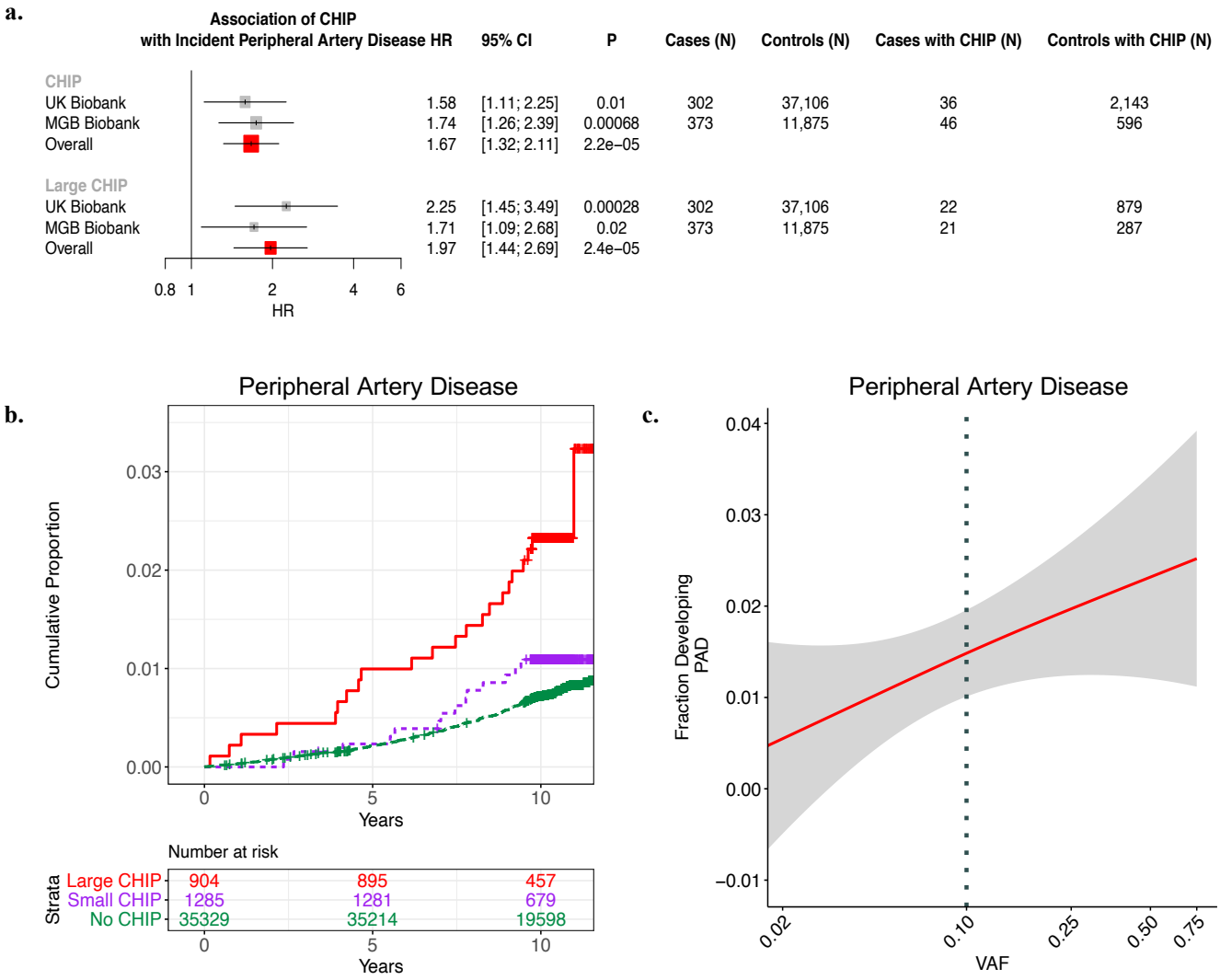


Figure 3.3.1: CHIP and incident PAD risk. *a) Association of CHIP and large CHIP (VAF>10%) carrier state with incident PAD events in the UK Biobank (UKB) and Mass General Brigham Biobank (MGBB). Results were combined using an inverse-variance weighted fixed effects meta-analysis. b) Cumulative proportion of individuals developing PAD stratified by CHIP VAF clone size category in the UK Biobank. c) Fraction of individuals developing incident PAD by CHIP VAF in the UK Biobank.*

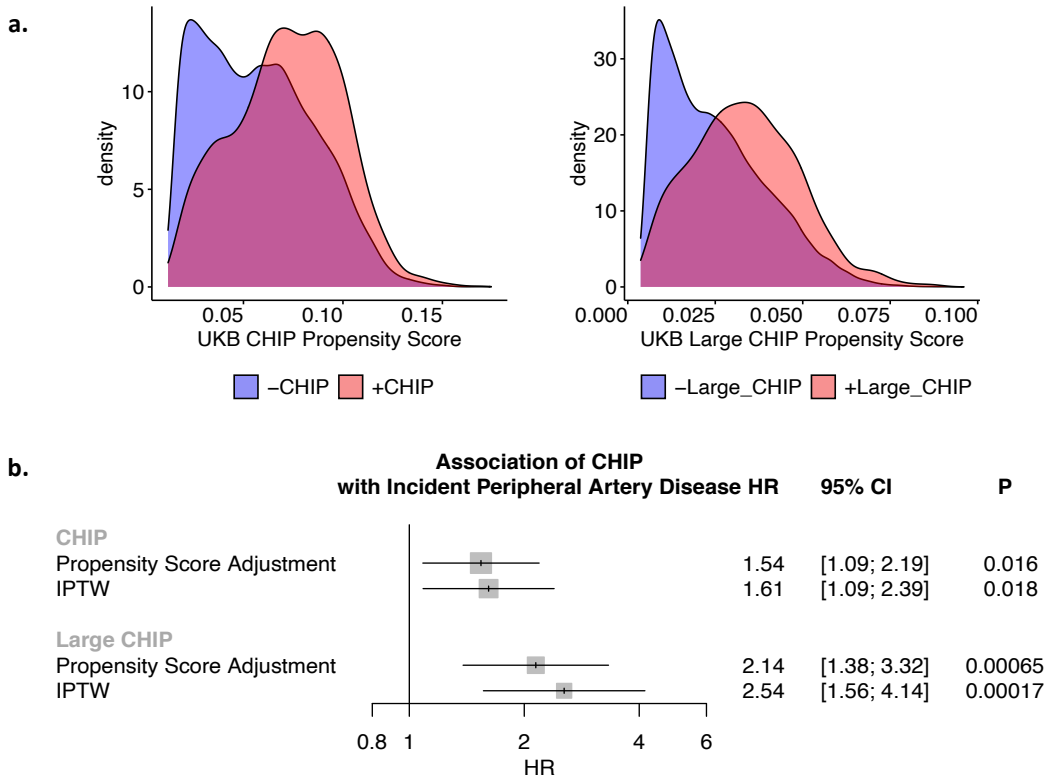


Figure 3.3.2: Epidemiological causal inference analysis for CHIP on incident peripheral artery disease in the UK Biobank. a) *Propensity scores by CHIP and Large CHIP status in the UKB.* ***b)*** *Propensity score adjustment and stabilized inverse probability treatment weighting (IPTW) for the CHIP and Large CHIP association with incident PAD in the UKB. CHIP = clonal hematopoiesis of indeterminate potential; VAF = variant allele fraction; PAD = peripheral artery disease*

CHIP and Incident Atherosclerosis Across Multiple Vascular Beds

We next assessed whether CHIP was associated with 9 other incident atherosclerotic diseases across multiple vascular beds. Using EHR-based disease definitions[33], we tested the association of CHIP with atherosclerotic disease across the mesenteric (acute and chronic), coronary, and cerebral vascular beds, as well as with aneurysmal disease (aortic and any other aneurysm). We observed significant associations for coronary artery disease (HR 1.40, 95% CI: 1.20 to 1.63; $P=1.9x10^{-5}$), any aortic aneurysm (HR 1.74; 95% CI: 1.21 to 2.51; P=0.0028), other aneurysms (HR 1.70;

95% CI: 1.23 to 2.34; P=0.0013), and chronic mesenteric ischemia (HR 9.12; 95% CI: 2.34 to 35.63; P=0.0015) across both cohorts, with directionally consistent effect estimates observed for all the tested phenotypes (**Figure 3.3.3**). These associations were consistently stronger for large CHIP clones (**Figure 3.3.4)**. We then created a composite, incident atherosclerosis outcome combining all nine atherosclerotic phenotypes ("pan-arterial atherosclerosis"). CHIP associated with this combined incident pan-arterial atherosclerosis endpoint (HR 1.31, 95% CI: 1.14 to 1.49, $P=9.7x10^{-5}$), again with stronger effects conferred by large CHIP clones (HR 1.45; 95% CI: 1.20 to 1.75; P=0.00013) (**Figure 3.3.3b,c**).

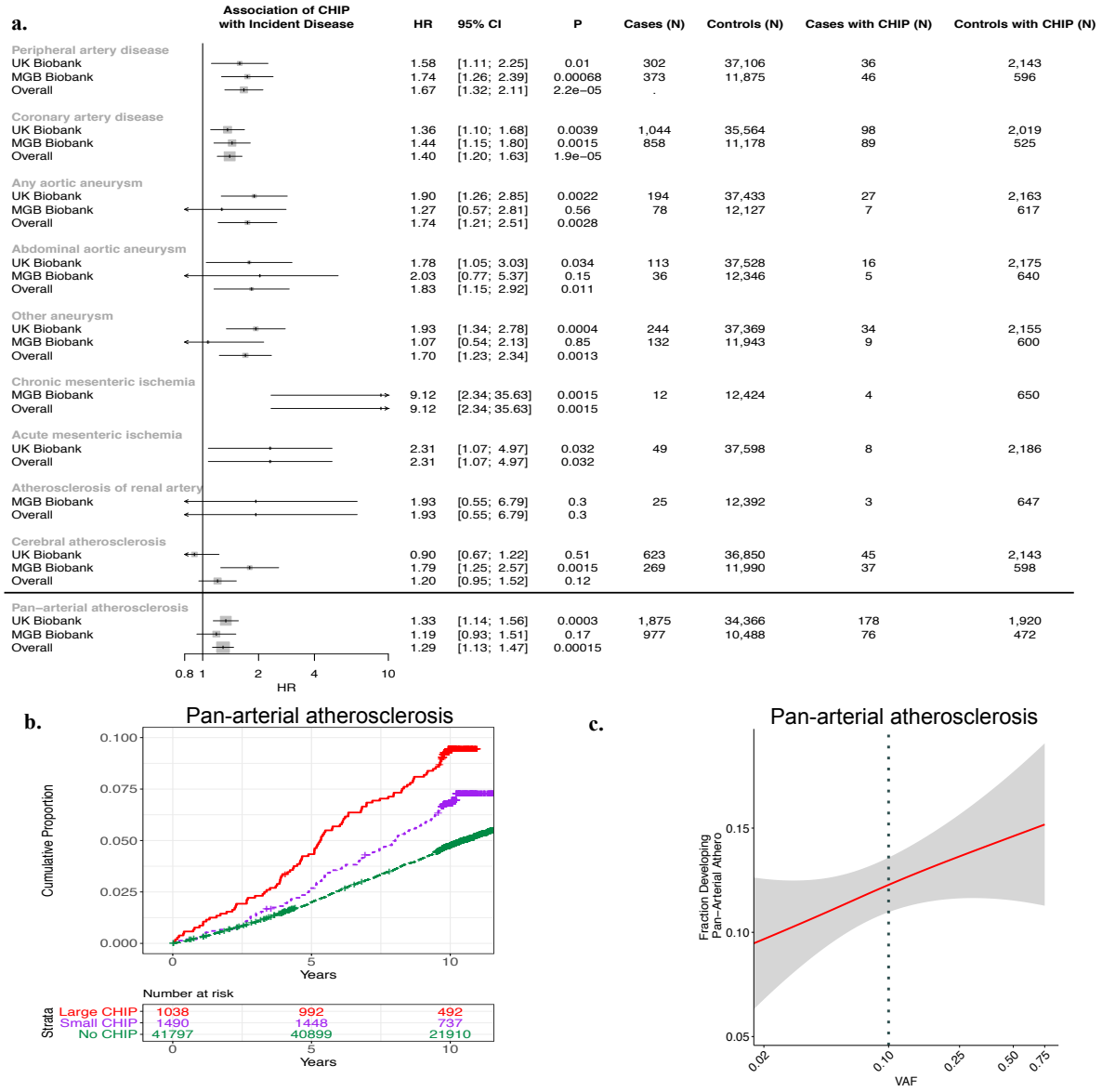


Figure 3.3.3: CHIP and incident pan-arterial atherosclerosis risk. *a) Association of CHIP with 9 incident atherosclerotic diseases separately and combined in a 'Pan-arterial atherosclerosis' phenotype in the UKB, MGBB, and meta-analyzed across both studies ("Overall"). b) Cumulative risk of incident atherosclerosis across the composite 'pan-arterial atherosclerosis' phenotype stratified by no CHIP, small CHIP (VAF<10%), and large CHIP (VAF≥10%) carrier state in the UK Biobank. c) Association of CHIP VAF with fraction of individuals developing pan-arterial atherosclerosis in the UK Biobank. CHIP = clonal hematopoiesis of indeterminate potential; VAF = variant allele fraction; PAD = peripheral artery disease*

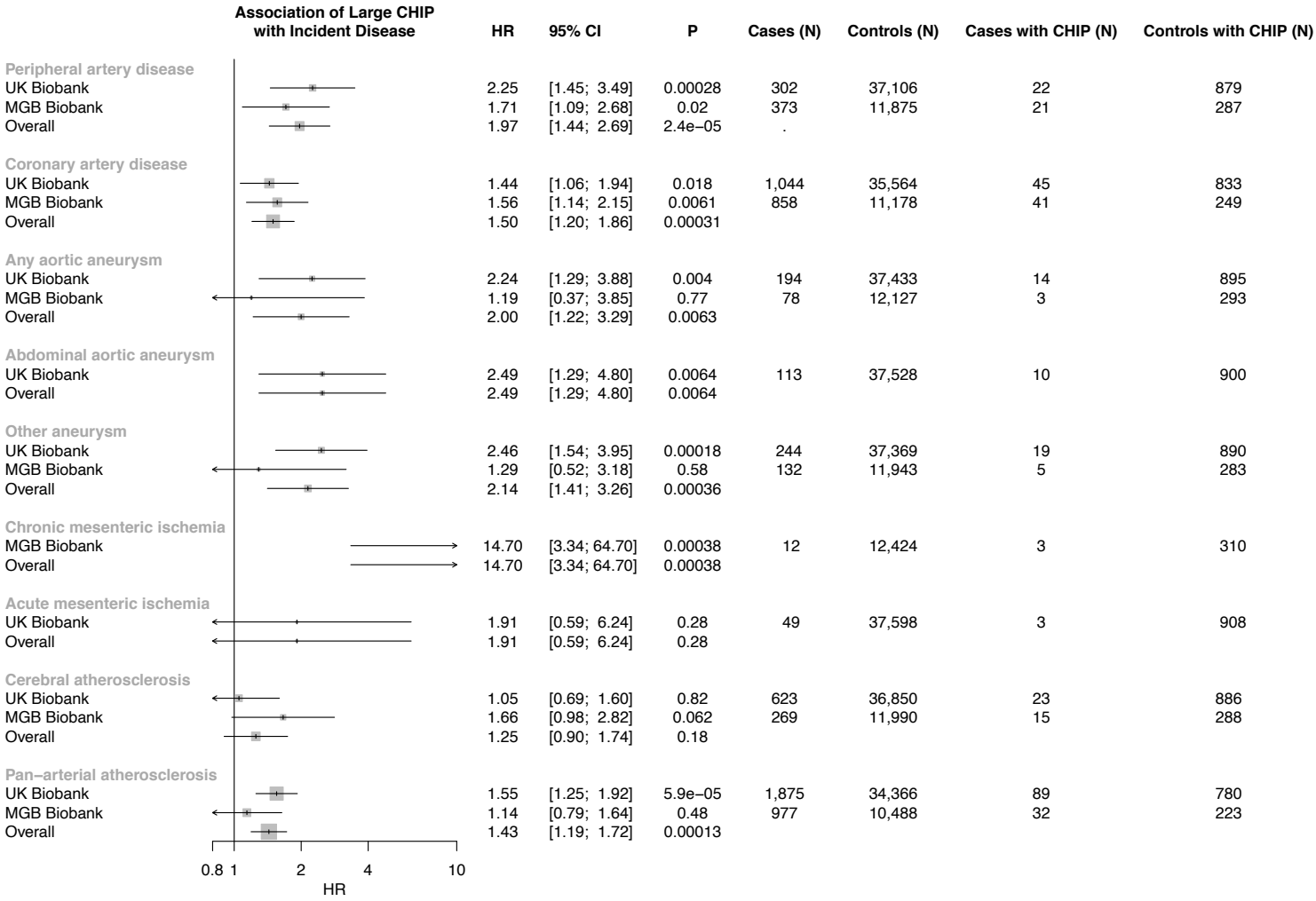


Figure 3.3.4: Association of Large CHIP (VAF>10%) with incident pan-arterial atherosclerosis, combined across peripheral artery disease, coronary artery disease, aneurysms, chronic and acute mesenteric ischemia, cerebral atherosclerosis, and renal artery stenosis. *CHIP = clonal hematopoiesis of indeterminate potential; VAF = variant allele fraction*

Gene-specific analyses of CHIP with incident atherosclerotic diseases

Next, we sought to understand whether the clonal hematopoiesis putative driver gene differentially affected the risk of acquiring atherosclerosis. Previous work has focused primarily on the epigenetic regulators *DNMT3A* and *TET2*[34 35], and whether DDR CHIP confers an increased risk of atherosclerosis is unknown. We stratified the CHIP-PAD and

CHIP pan-arterial atherosclerosis analyses by putative driver genes and specific mutations - focusing on *DNMT3A*, *TET2*, *ASXL1*, *JAK2*, the DDR genes *PPM1D* and *TP53*, and mutations that specifically disrupt splicing factor genes *(LUC7L2, PRPF8, SF3B1, SRSF2, U2AF1,* and *ZRSR2)*[36]. We observed an association of CHIP with PAD across the four common CHIP genes (*DNMT3A*, *TET2*, *ASXL1*, and *JAK2*), with significant heterogeneity of incident PAD effect sizes across the CHIP genes ($P_{heterogeneity}$ = 0.03) (**Figure 3.3.5a**). This heterogeneity persisted in sensitivity analysis after excluding JAK2 carriers ($P_{heterogeneity}$ = 0.046). These data also revealed the novel finding that DDR *TP53* and *PPM1D* CHIP associates with incident PAD (HR 2.72; 95% CI: 1.20 to 1.75; P=0.00013) and incident CAD (HR 2.51; 95% CI: 1.52-4.14; P=0.00032), with a stronger effect on PAD conferred by *TP53* (HR 4.98; 95% CI: 1.23-20.09; P=0.024, **Figure 3.3.5a-c).** Similar findings were observed for the incident pan-arterial atherosclerosis outcome when stratifying by putative driver gene (**Figure 3.3.6**). Further sensitivity analysis for DDR-CHIP and incident PAD when excluding hematologic or solid organ malignancy did not significantly change the associations ($P_{heterogeneity} > 0.05$).

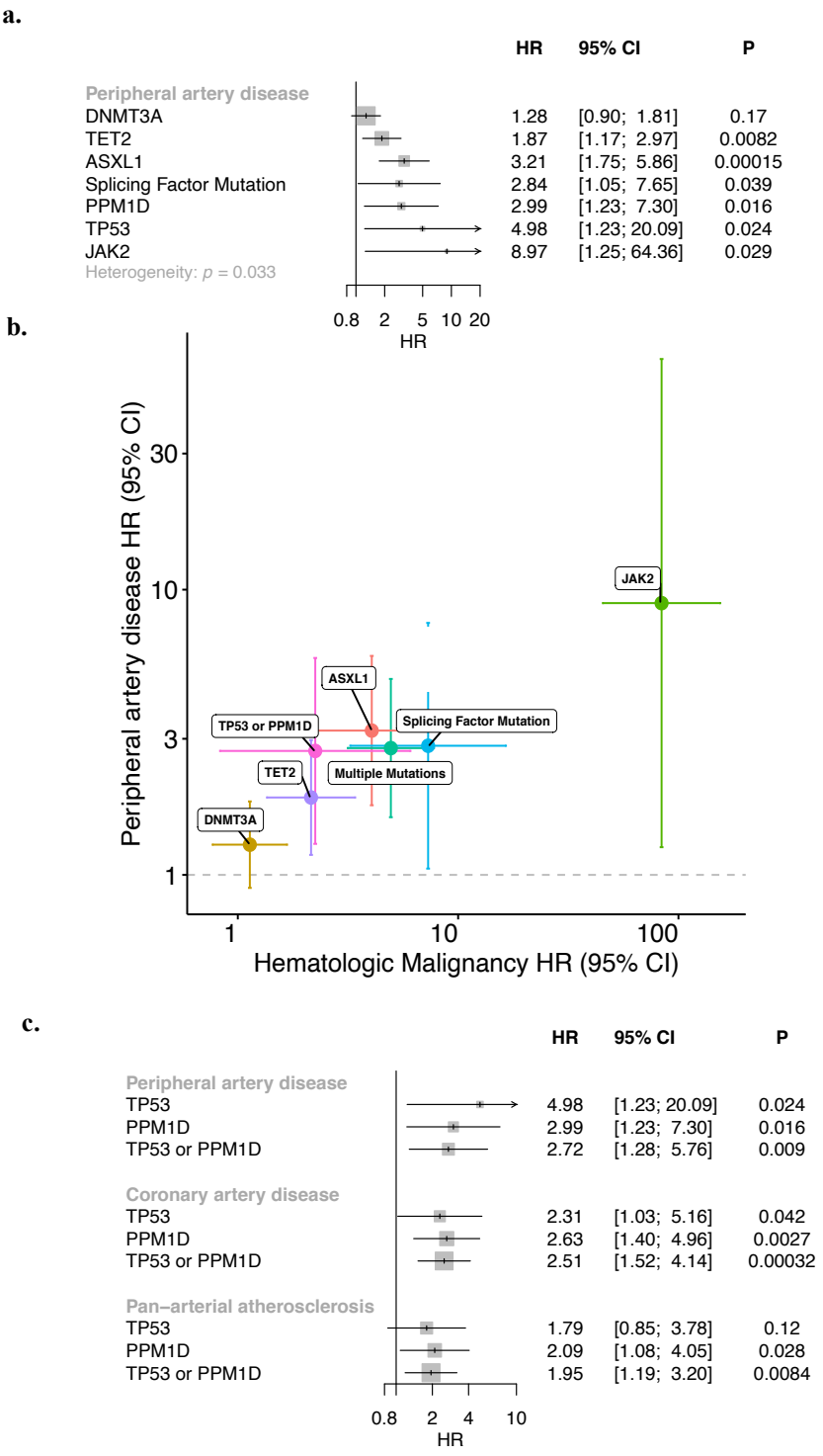


Figure 3.3.5: Gene-specific association of CHIP with incident peripheral artery disease (PAD). *a) CHIP-PAD association analyses stratified by putative CHIP driver gene. Results following meta-analysis across the UKB and MGBB are shown. b) Gene-specific comparison of HR and 95% CI for hematologic malignancy (x-axis) and PAD (y-axis) in the UKB. c) Association of DDR CHIP (PPM1D or TP53) with incident peripheral artery disease, coronary artery disease, and pan-vascular atherosclerosis. Results across UK Biobank and MGB Biobank were combined using an inverse-variance weighted fixed effects meta-analysis. CHIP = clonal hematopoiesis of indeterminate potential; DDR = DNA-damage repair; VAF = variant allele fraction; PAD = peripheral artery disease*

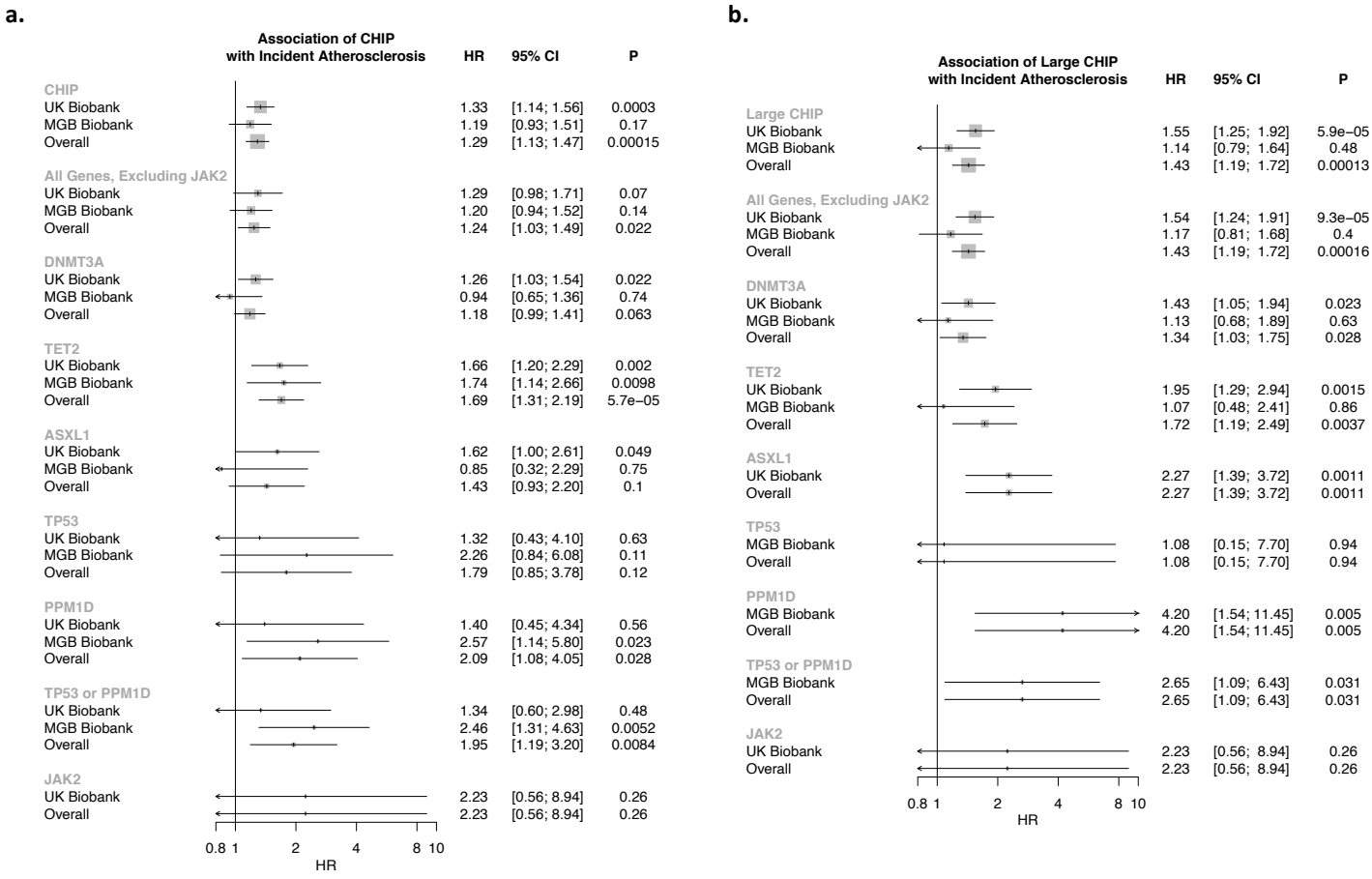


Figure 3.3.6: Association of **a)** CHIP and **b)** Large CHIP genes with incident pan-arterial atherosclerosis, combined across peripheral artery disease, coronary artery disease, aneurysms, chronic and acute mesenteric ischemia, cerebral atherosclerosis, and renal artery stenosis. CHIP = clonal hematopoiesis of indeterminate potential; VAF = variant allele fraction

Atherosclerosis development in p53-/ CHIP mice

Working collaboratively with José J Fuster's laboratory group in the Centro Nacional de Investigaciones Cardiovasculares (Madrid, Spain), based on our gene specific findings, we next further characterized the effects of reduced function of hematopoietic p53 in atherosclerotic mice. To mimic the human scenario of clonal hematopoiesis and test whether the expansion of p53-deficient hematopoietic cells contributes to atherosclerosis,

a competitive bone marrow transplantation (BMT) strategy was used to generate atherosclerosis-prone *Ldlr*-/- chimeric mice carrying 20% *Trp53*-/- hematopoietic cells (20% KO-BMT mice). These mice then consumed a high fat/high cholesterol diet for 9 weeks to induce atherosclerosis development. The presence and expansion of *Trp53*-/- cells led to a significant □40% increase in plaque size in the aortic root of male *Ldlr*-/- mice (**Figure 3.3.7**), without affecting body weight, spleen weight or serum cholesterol levels. Similar results were obtained in female *Ldlr*-/- mice. Increased atherogenesis in mice carrying *Trp53*-/- cells was paralleled by a substantial increase in plaque macrophage content, as assessed by immunohistological staining of Mac2, with no significant changes in other cell components (**Figure 3.3.8**), suggesting a contribution of increased arterial macrophage burden to accelerated atherosclerosis in conditions of p53 CHIP.

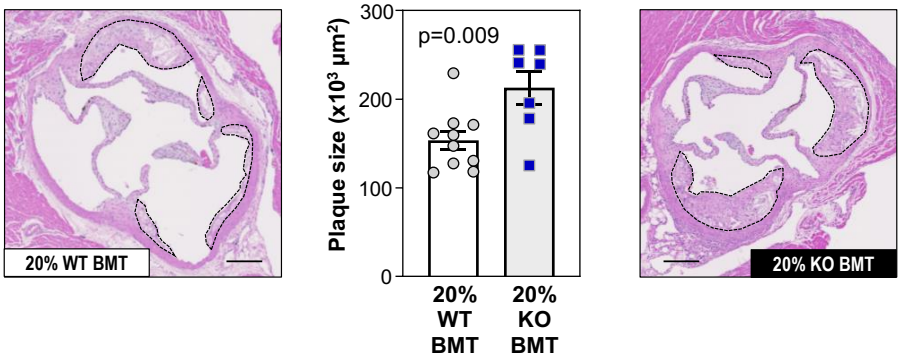


Figure 3.3.7: **Accelerated atherosclerosis in a murine model of *TP53* mutation-driven CHIP.** *20% KO-BMT male mice and 20% WT-BMT controls were fed a high-fat/high-cholesterol (HF/HC) diet for 9 weeks, starting 4 weeks after BMT (n=10 20% WT-BMT, n=7 20% KO-BMT, unless otherwise noted). Representative images of hematoxylin and eosin-stained sections from aortic root are shown; atherosclerotic plaques are delineated by dashed lines. Scale bars, 100 µm. [Figure and analyses performed by Jose J Fuster's group, and included here with permission].*

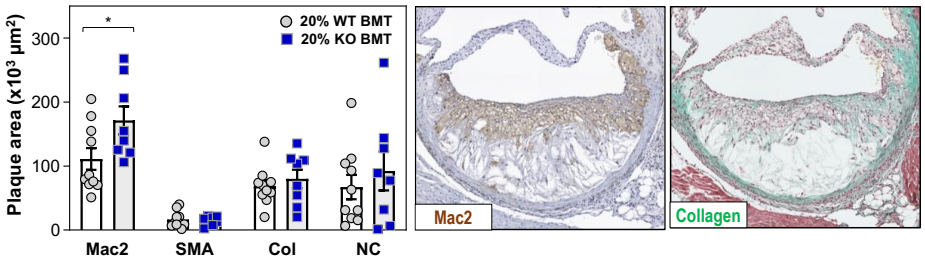


Figure 3.3.8: Increased proliferation and expansion of p53-deficient macrophages. *a) Plaque composition in 20% KO BMT female mice (n=10) and controls (n=8) quantified as absolute intimal content of macrophages (Mac2 antigen immunostaining), vascular smooth muscle cells (smooth muscle α-actin, SMA immunostaining), collagen (Masson's trichrome staining) and necrotic core (collagen-free acellular regions). Representative images of Mac2- and collagen-stained histological sections of 20% KO BMT mice are shown. [Figure and analyses performed by Jose J Fuster's group, and included here with permission].*

Discussion

In this study, we combined exome sequencing data across two biobanks to detect somatic mutations in over 50,000 individuals and observed that CHIP carriers were at significantly increased risk of developing PAD and atherosclerosis across multiple arterial beds. Findings were consistent across CHIP driver genes, including the DDR genes *PPM1D* and *TP53*, with evidence of dose dependent effect of CHIP VAF, with large CHIP clones conferring greater risk of disease, similar to prior observations with coronary artery disease[34]. Lastly, through analysis of p53 CHIP using a BMT murine model, we observed evidence of increased aortic atherosclerotic plaque among CHIP carriers via expansion of plaque macrophages (**Figure 3.3.9**).

These findings permit several conclusions. First, CHIP appears to promote atherosclerosis across the entire arterial system in humans. Previous work demonstrated that CHIP was associated with an increased risk of coronary artery disease and early-onset MI[34]. We further demonstrate that CHIP is also associated with PAD, aortic

aneurysms – commonly driven by atherosclerotic disease[37], and a composite pan-arterial atherosclerosis outcome reflective of an increased burden of atherosclerosis throughout the vascular system. Based on these results, therapies aimed at mitigating the cardiovascular consequences of CHIP are likely to be efficacious throughout the arterial tree, and the link between CHIP and aneurysmal disease warrants further investigation.

Second, CHIP variants specifically in DDR genes (*TP53, PPM1D*) confer an increased risk of atherosclerotic cardiovascular disease. Prior work demonstrated CHIP carriers with *DNMT3A, TET2, ASXL1,* and *JAK2* somatic driver mutations have increased risk of CAD[34]. Somatic variants in DDR genes are often observed following cytotoxic chemotherapy for cancer treatment; however, prior work linking DDR CHIP carriers and cardiovascular disease risk have been limited. In the current study, we demonstrate CHIP related to DDR-genes (*TP53, PPM1D*) confer higher risk of developing atherosclerosis compared to the more common CHIP epigenetic regular genes (DNMT3A, TET2). Furthermore, through experimental mouse studies we show that *TP53* mutations promote atherosclerosis risk via expansion of p53-deficient macrophages in occlusive plaque lesions.

Several limitations exist in the present study. First, our PAD and cardiovascular disease phenotypes are based on EHR data and may result in misclassification of case status. Such misclassification should, however, reduce statistical power for discovery and on average bias results toward the null. Second, selection bias from differential loss-of-follow up, volunteer bias, and missingness in covariates may be present given the nature of the genetic biobanks used in this study. Lastly, the cohorts in these studies are largely of European ancestry; while it seems mechanistically plausible that the same results

would be applicable to individuals of other ancestries, further analyses using ethnically diverse individuals would help assess the generalizability of this finding.

In conclusion, here we newly identified that CHIP, and particularly DDR CHIP, is associated with incident atherosclerosis across multiple vascular beds, with supporting murine evidence of increased plaque among *TP53* CHIP carriers through an expansion of plaque macrophages. This observation enhances our understanding of CHIP mediated atherosclerosis, and may aid risk stratification of DDR gene CHIP patients in a cardio-oncology setting.

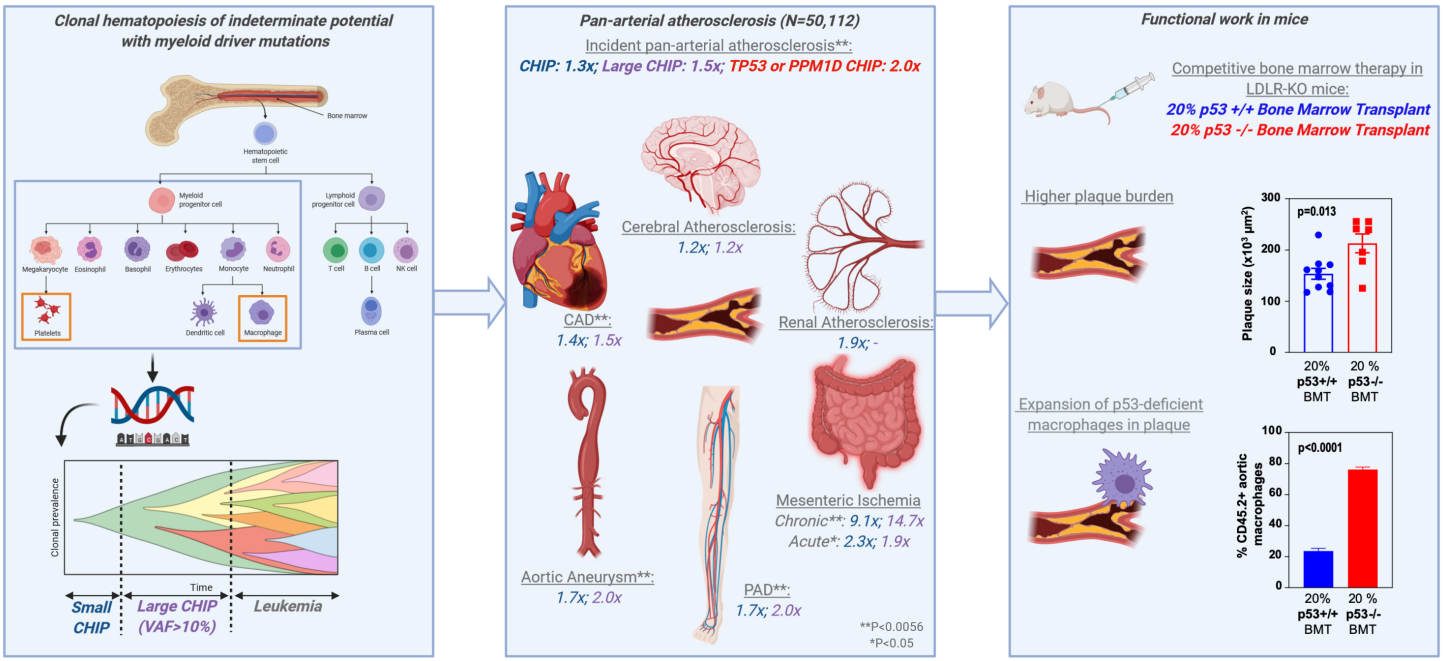


Figure 3.3.9: Study schematic. *In this study, we assessed the association of clonal hematopoiesis of indeterminate potential (CHIP) with myeloid driver mutations with pan-arterial atherosclerosis. CHIP is a category of age-related somatic variants which are associated with incident leukemia and thought to be implicated in atherosclerosis primarily by altering macrophage function and promoting thrombosis. CHIP clones can be characterized by the fraction of blood cells carrying the clone, referred to as the variant allele fraction (VAF); here we categorized large CHIP clones as variants with VAF>10%. Across 50,112 individuals from the UK Biobank and Mass-General Brigham Biobank, we observed that CHIP is associated with increased risk of incident peripheral and pan-arterial atherosclerosis, with stronger effects conferred by large CHIP clones (HR 1.5x). In addition, we observed and a novel associations for TP53 and PPM1D CHIP (HR 2.0x). CHIP was found to be individually associated with a variety of atherosclerotic conditions, with Bonferroni-significant associations (double-starred, **) identified for peripheral artery disease (PAD), coronary artery disease (CAD), aortic aneurysm, and chronic mesenteric ischemia. HR for CHIP are displayed in blue and for large CHIP in purple. Functional analysis was performed to further investigate the observed TP53-PAD association. Ldlr-KO 20% p53 -/- bone-marrow transplanted mice had a significant increase in plaque size, with significant expansion of p53-deficient macrophages in plaque (P<0.001) at 12 weeks.*

Chapter 3.4: Association of CHIP with stroke

The extent to which CHIP associates with stroke risk is not well understood. The association of CHIP with risk of incident ischemic stroke was first reported by Jaiswal et al (2014) in an analysis conducted within two cohorts comprising 2,420 people (hr, 2.2; 95% CI, 1.1 to 4.6; P=0.03) independent of traditional risk factors[19]. The ischemic stroke risk appeared to be somewhat greater among persons who had a variant allele fraction of >10%, or at least 10% of circulating blood DNA with a CHIP mutation. Since brain parenchymal microglial cells and perivascular cells are derived from HSCs, somatic mutations in these cells acquired through CHIP represent an additional potential mechanism by which CHIP might influence the occurrence or severity of cerebral ischemia during infarction or hemorrhage[38-40]. Nevertheless, this initial report was limited by the relatively small number of incident stroke cases and lack of stroke sub-phenotyping. Moreover, whether CHIP is additionally a risk factor for hemorrhagic stroke, another common type of stroke, is unknown. The purpose of this study was to discover whether CHIP is a risk factor for ischemic or hemorrhagic stroke.

Here, CHIP genotypes were obtained from 8 studies [the Atherosclerosis Risk in Communities Study (ARIC), the Cardiovascular Health Study (CHS), the Framingham Heart Study (FHS), the Jackson Heart Study (JHS), the Multi-Ethnic Study of Atherosclerosis (MESA), the Women's Health Initiative (WHI), UK Biobank (UKBB), and Massachusetts General Brigham Biobank (MGBB). Incident stroke was ascertained by physician adjudicators in the cohort studies, and by ICD codes in the biobanks. Cox proportional hazards models were fitted with adjustment for age, sex, diabetes mellitus,

smoking status (never, past, current) and race. Fixed-effects meta-analysis was used to estimate pooled effect sizes.

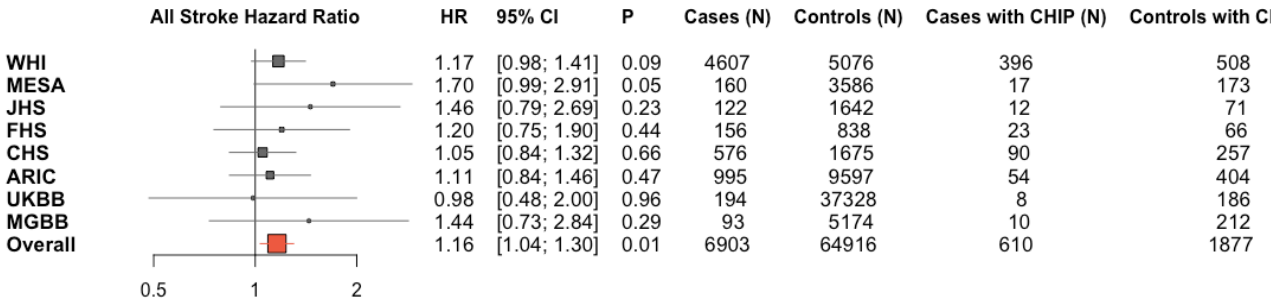
A total of 78,752 participants from 8 studies were included in the final analyses, after excluding individuals with prevalent hematological cancer at enrollment. In the fixed-effect meta-analysis, CHIP mutations were associated with an increased risk of total stroke (HR= 1.17, 95% CI 1.05, 1.28; P=7.1x10^{-90}) (**Figure 3.4.1**). In analysis of stroke subgroups, the risk was greater for hemorrhagic (HR= 1.37, 95% CI 1.15, 1.59; P=4.1x10^{-34}) than ischemic stroke (HR= 1.13, 95% CI 1.00, 1.26; P=2.4x10^{-66}); however no significant heterogeneity was detected between the two stroke subtypes. Further gene-specific analyses in the WHI cohort suggested the *TET2* CHIP gene as having the most strongest effect on future stroke risk (HR 1.85, p=0.004) (**Figure 3.4.2**). *TET2* was associated with increased risk for ischemic stroke (HR 1.93, p=0.006), and the effect sizes for the association of *TET2* (HR=1.50, p=0.15) and *DMNT3A* (HR 1.44, p=0.03) with hemorrhagic stroke were similar.

Table 3.4.1: Baseline Characteristics*. Baseline characteristics of the study population presented by cohort.*

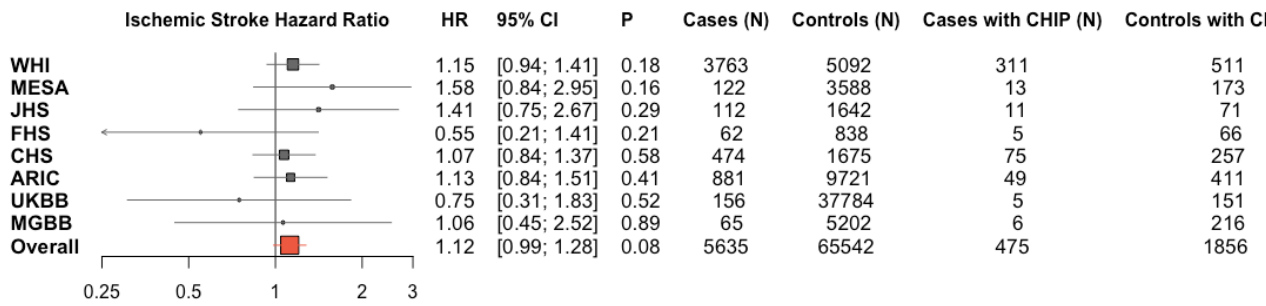
	WHI	MESA	JHS	FHS	CHS	ARIC	MGBB	UKBB
N	9683	3963	1764	994	2315	10355	11962	45186
AGE	68.9 (6.8)	61.1 (9.8)	56.8 (11.4)	66.4 (12.6)	73.9 (5.6)	57.81 (6.0)	46.5 (14.8)	56.5 (8.0)
FEMALE	9683 (100)	2018 (50.9)	1077 (61.1)	539 (54.2)	1297 (56.0)	5890 (56.9)	6968 (58.3)	24656 (54.6)
RACE								
WHITE	7988 (82.5)	1692 (42.7)	0 (0.0)	994 (100)	1889 (81.6)	7552 (72.9)	9595 (80.2)	42110 (93.2)
BLACK	1195 (12.3)	875 (22.1)	1764 (100)	0 (0.0)	397 (17.2)	2783 (26.9)	717 (6.0)	936 (2.1)
OTHER	500 (5.2)	1396 (35.2)	0 (0.0)	0 (0.0)	29 (1.3)	0 (0.0)	1650 (13.8)	2140 (4.7)
HYPERTENSION	4446 (45.9)	1531 (41.9)	1047 (60.6)	217 (21.9)	1523 (65.9)	3765 (36.4)	1905 (15.9)	13442 (29.7)
PRIOR STROKE	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
INCIDENT STROKE	4607 (47.6)	160 (4.0)	122 (6.9)	156 (15.7)	576 (24.9)	995 (9.6)	130 (1.1)	680 (1.5)
CURRENT SMOKER	719 (7.4)	446 (12.2)	231 (13.2)	338 (34.1)	279 (12.1)	2266 (21.9)	290 (2.4)	4050 (9.0)
BMI	28.6 (6.2)	28.1 (5.2)	31.6 (7.1)	25.7 (4.7)	26.5 (4.5)	28.19 (5.6)	28.3 (10.8)	27.4 (4.78)

FOLLOW UP YEARS	10.8 (6.4)	13.5 (2.5)	12.6 (3.6)	7.6 (3.5)	11.3 (7.0)	20.4 (8.0)	3.0 (2.0)	9.9 (2.7)

A.



B.



C.

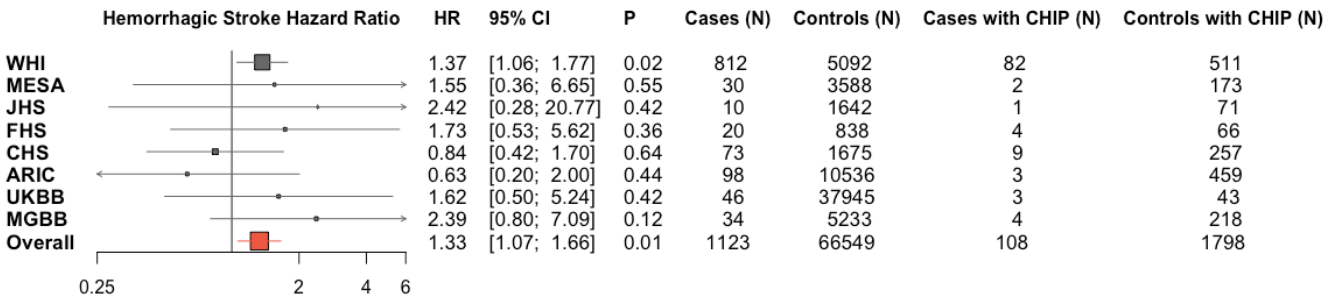


Figure 3.4.1: Association of CHIP with incident stroke. *Cox proportional hazards models were fitted with adjustment for age, sex, diabetes mellitus, smoking status (never, past, current) and the first 10 principal components of genetic ancestry. Fixed-effects meta-analysis was used to estimate pooled effect sizes.*

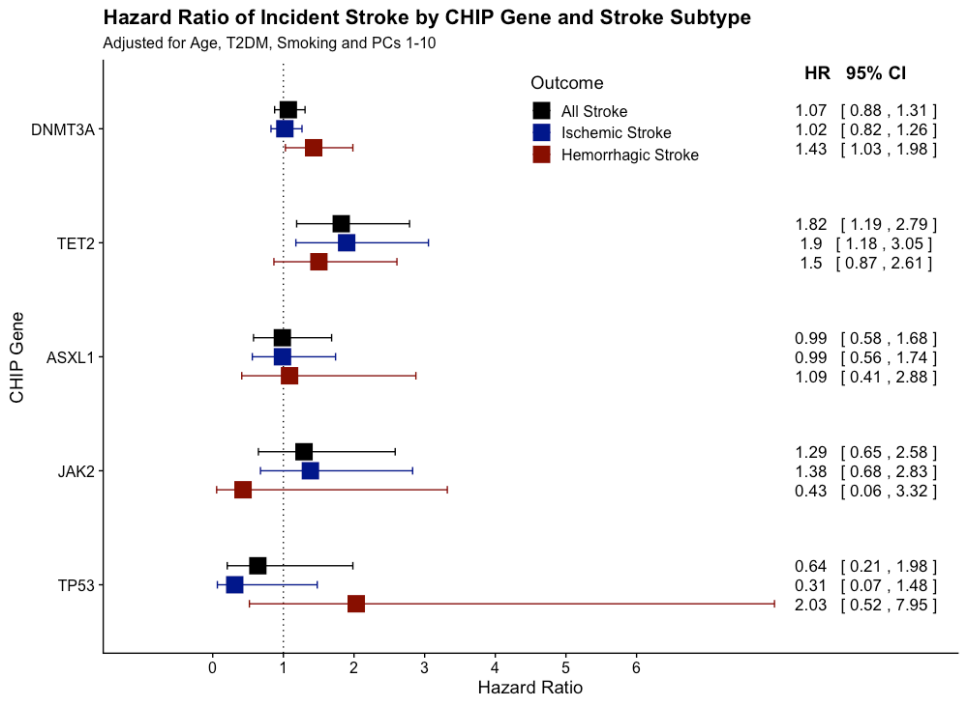


Figure 2: Forest plot of gene-specific hazard ratios for the association between CHIP and Stroke, amongst the WHI cohort. *Cox proportional hazards models were fitted, adjusted for age, type 2 diabetes, smoking history, and the first 10 principal components of genetic ancestry.*

This analyses has several limitations. Firstly, the heterogeneity of study protocols, recruitment and adjudication of patients and clinical events is challenging to harmonize. We attempted through collaboration and rigorous attention to outcome definitions to ensure standard treatment of subjects and events but acknowledge some heterogeneity may persist. However, the inclusion of multiple datasets with diverse individuals improves generalizability of the study findings and simultaneously adds to the strength of the study. Secondly, though these data were prospectively ascertained, they are observational data and thus cannot provide strong causal evidence. Additionally, CHIP was ascertained at a single time point. Having CHIP at multiple time points would allow for stronger evidence linking CHIP and risk of stroke. Lastly, our results were

unexpected in linking CHIP to both hemorrhagic and ischemic stroke (particularly to small-vessel disease). Mechanistic links have not yet been robustly investigated that explain this finding in full.

In summary, our findings identify that CHIP is associated with an increased risk of stroke, with stronger effects for *TET2* CHIP. The finding that CHIP was more strongly associated with hemorrhagic stroke compared to ischemic stroke requires further replication and investigation of the role of CHIP in vascular fragility and the formation of intracranial aneurysms.

Chapter 3.5: Association of CHIP with heart failure

Heart failure (HF) is a leading cause of death in the elderly [41]. Lifetime risk for HF is 1 in 5, and HF is associated with short-term mortality rates exceeding those of many cancers in western countries [42 43]. Coronary heart disease (CHD), along with hypertension, atrial fibrillation, and chronic kidney disease, are all risk factors for incident HF and strongly associated with aging. Age remains the strongest independent predictor for HF, but the age-related factors promoting HF development are incompletely understood.

Recently in a cohort of patients with HF, Dorsheimer et al found during 4.4 years of median follow-up, those with either *TET2* or *DNMT3A* mutations had increased risk of death or HF hospitalization (HR=2.1, 95% CI 1.1-4.0) [44]. Murine models with hematopoietic or myeloid-specific deficiency of *Tet2* or with myeloid-specific transgenic *Jak2*^{V617F} are more prone to cardiac dysfunction after coronary artery ligation-induced myocardial infarction or aortic constriction-induced pressure overload [45-47]. Therefore, we tested the hypothesis that CHIP driver mutations are associated with incident HF in four cohorts from the NHLBI Trans-Omics for Precision Medicine (TOPMed) Program and the United Kingdom Biobank (UKBB) study.

A total of 56,597 study participants were analyzed in the present study to assess the association between CHIP and incident HF. 4,694 of them developed HF with up to 20 years follow-up. The mean age of each study ranged from 54.5 to 74.6 (SD between 5.4 and 13.0) years old, 6% of the participants had CHIP, and 3.3% of the participants had high-VAF CHIP. **Table 3.5.1** shows baseline characteristics for those participants with CHIP compared to those without CHIP. In brief, CHIP carriers were older and more

likely to have comorbidities. Prevalent CHIP did not appear to be related to BMI or lipid profiles. Consistent with prior observations, the most common CHIP genes were *DNMT3A*, *TET2*, *ASXL1* and *JAK2*, as shown in **Table 3.5.2.**

Table 3.5.1: Characteristics by clonal hematopoiesis of indeterminate potential status for individuals included in stroke analyses.

	ARIC		CHS		JHS		UKBB		WHI		All studies	
Category	**CHIP**	**No CHIP**	**CHIP**	**No CHIP**	**CHIP**	**No CHIP**	**CHIP**	**No CHIP**	**CHIP**	**No CHIP**	**CHIP**	**No CHIP**
N	427	9473	337	2063	91	2332	2143	34517	408	4806	3406	53191
Age (years)	60 (5.9)	57.4 (6.1)	74.6 (5.6)	73.4 (5.4)	65.6 (9.0)	54.5 (13.0)	60.6 (6.6)	56.8 (7.8)	67.4 (6.6)	65.2 (6.9)	62.9 (7.9)	58.2 (8.6)
Female	239 (56)	5325 (56)	173 (51)	1161 (56)	55 (60)	1464 (63)	1131 (53)	18593 (54)	408 (100)	4806 (100)	2006 (58.9)	31349 (58.9)
Race												
White	290 (68)	6884 (73)	285 (85)	1673 (81)	0 (0)	0 (0)	2143 (100)	34517 (100)	288 (71)	3147 (66)	3006 (88.3)	46221 (86.9)
Black	137 (32)	2589 (27)	52 (15)	390 (19)	91 (100)	2332 (100)	0 (0)	0 (0)	93 (23)	1296 (27)	373 (11)	6607 (12.4)
Other	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	25 (6)	354 (7)	25 (0.7)	354 (0.7)
DM	67 (16)	1337 (14)	62 (18)	336 (16)	26 (28)	527 (23)	63 (3)	885 (3)	22 (5)	364 (8)	240 (7)	3449 (6.5)
HTN	174 (41)	3296 (35)	159 (47)	991 (48)	65 (71)	1183 (51)	761 (36)	10277 (30)	169 (41)	2082 (43)	1328 (39)	17829 (33.5)
CHD	22 (5)	484 (5)	35 (10)	212 (10)	5 (5)	69 (3)	155 (7)	1846 (5)	13 (3)	186 (4)	230 (6.8)	2797 (5.3)
Stroke	16 (4)	164 (2)	16 (5)	78 (4)	2 (2)	83 (4)	31 (1)	459 (1)	3 (1)	64 (1)	68 (2)	848 (1.6)
Current smoker	117 (27)	2040 (22)	35 (10)	249 (12)	10 (11)	290 (12)	213 (10)	2913 (9)	27 (7)	474 (10)	402 (11.8)	5966 (11.2)
BMI (kg/m^2)	27.5 (5.4)	28.1 (5.4)	26.9 (4.6)	26.8 (4.7)	31.0 (6.6)	31.8 (7.4)	27.5 (4.5)	27.4 (4.7)	29.8 (6.2)	29.7 (6.2)	27.8 (5)	27.9 (5.2)
SBP (mmHg)	125.2 (19.4)	121.9 (18.3)	135.2 (20.4)	136.9 (21.7)	133.9 (16.7)	126.9 (16.2)	145.3 (20.8)	141.4 (20.6)	132 (17.4)	131 (17.8)	139.9 (21.5)	136.2 (21.3)
HF events	125	2046	139	803	11	177	75	695	64	562	414	4283
Follow-up years	17.7 (8.5)	20.0 (7.8)	10.5 (6.4)	11.8 (6.7)	8.4 (3.3)	9.7 (2.5)	10.1 (1.3)	10.2 (1.5)	14.7 (6.2)	15.7 (6.2)	11.6 (5.2)	12.5 (5.7)

Frequencies and percentages are displayed for categorical variables. Mean and SD are displayed for continuous variables. CHIP, clonal hematopoiesis of indeterminate potential; DM, prevalent diabetes mellitus; HTN, prevalent hypertension; CHD, prevalent coronary heart disease; BMI, body mass index; SBP, systolic blood pressure.

Table 3.5.2: Most frequent genes with somatic mutations by each study.

Somatic Mutations	**ARIC N (%)**	**CHS N (%)**	**JHS N (%)**	**UKBB N (%)**	**WHI N (%)**	**All Studies N (%)**
ASXL1	51 (0.5)	33 (1.4)	3 (0.1)	148 (0.4)	24 (0.5)	259 (0.5)
DNMT3A	253 (2.6)	172 (7.2)	55 (2.3)	1370 (3.7)	251 (4.8)	2101 (3.7)
JAK2	8 (0.1)	9 (0.4)	2 (0.1)	21 (0.05)	15 (0.3)	55 (0.1)
TET2	48 (0.5)	81 (3.4)	17 (0.7)	334 (0.9)	89 (1.7)	569 (1.0)
Any mutation	427 (4.3)	337 (14.0)	91 (3.8)	2143 (5.8)	408 (7.8)	3406 (6.0)
Large CHIP	257 (2.6)	287 (12)	82 (3.4)	879 (2.4)	342 (6.6)	1847 (3.3)

Frequencies and percentages are displayed

In the fixed-effect meta-analysis, we observed that the presence of a CHIP mutation was associated with a 25% increased risk of HF (HR= 1.25, 95% CI 1.13, 1.38), with consistent direction of effect in four of the five studies (**Figure 3.5.1**). *TET2* (HR=1.59, 95%CI 1.18, 2.14), *JAK2* (HR=2.50, 95%CI 1.35, 4.64) and *ASXL1* (HR=1.58, 95%CI 1.20, 2.08) somatic mutations were strongly associated with an increased risk of HF, while *DNMT3A* mutations were not associated with HF (**Figure 3.5.2).** In secondary analyses, we observed a slightly stronger association between high-VAF CHIP and the risk of HF (HR=1.29, 95% CI 1.15, 1.44). The associations for CHIP mutations on HF without prior CHD (HR=1.21, 95%CI 1.07, 1.36) and HF with prior CHD (HR=1.26, 95% CI 0.97, 1.64, **Figure 3.5.3**) were homogeneous (p=0.78 for test of homogeneity).

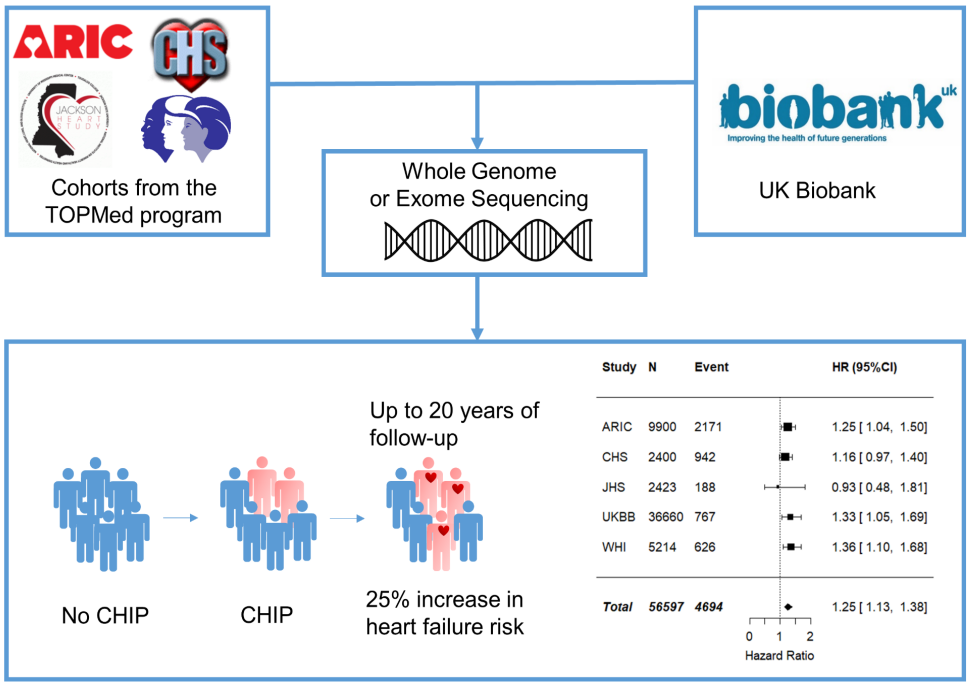


Figure 3.5.1: Clonal hematopoiesis of indeterminate potential mutation and incident heart failure. *Clonal hematopoiesis of indeterminate potential, determined by whole exome or genome sequencing, was significantly associated with an increased risk of heart failure in five prospective studies including 56,597 African, European and Hispanic populations with up to 20 years follow-up. Multivariable adjusted hazard ratios and 95% CIs were calculated separately in each study adjusting for age, sex, education, diabetes mellitus, smoking status, stroke, coronary heart disease, systolic blood pressure, hypertension medication use, body mass index, and race (if more than one) and combined using a fixed-effect meta-analysis.*

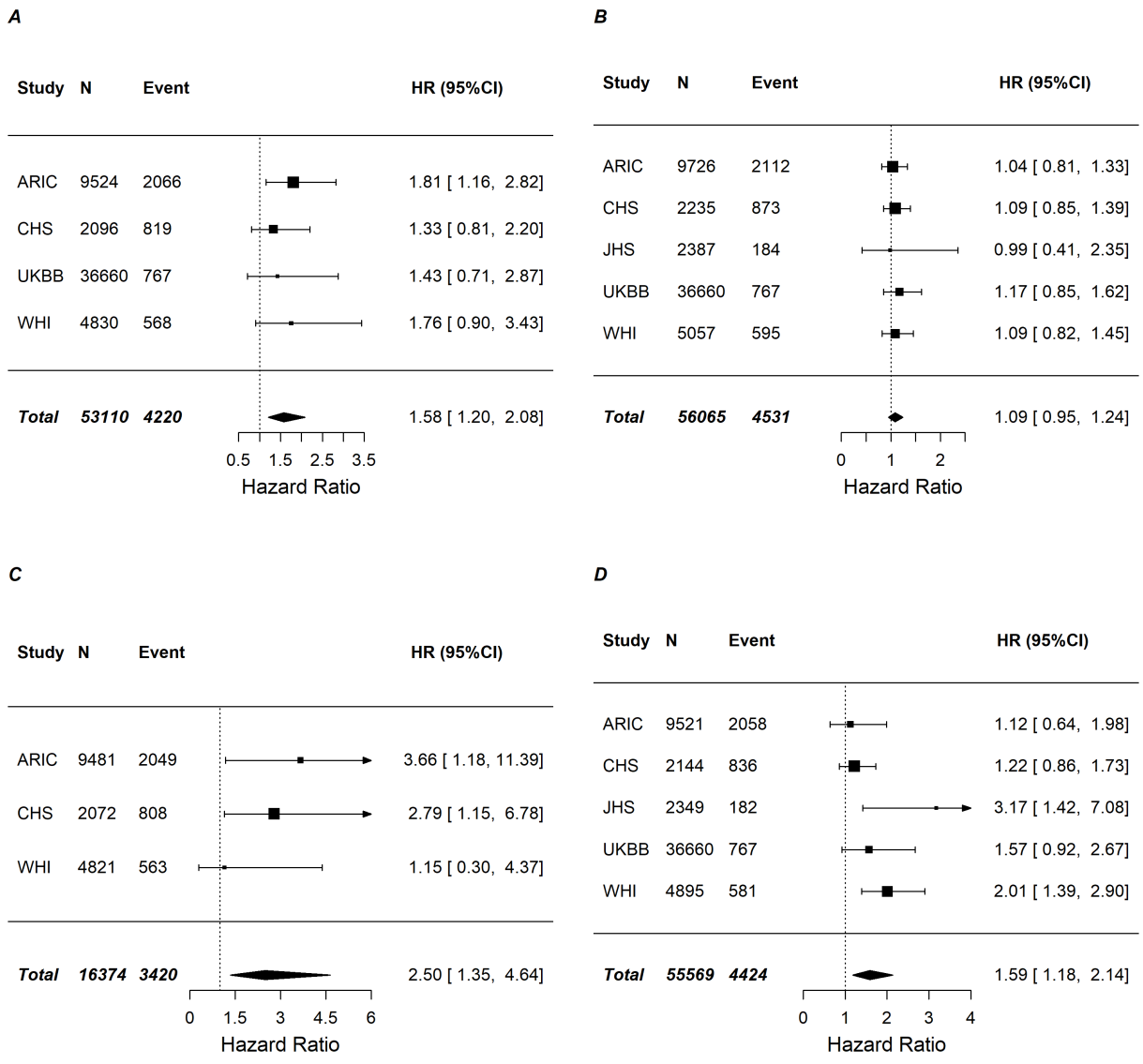


Figure 3.5.2: Clonal hematopoiesis in individual genes and incident heart failure. *Individual genes analyzed include a) ASXL1, b) DNMT3A, c) JAK2 and d) TET2. Event represents the number of incident heart failure cases. For each gene, multivariable adjusted hazard ratios and 95% CIs were calculated separately in each study adjusting for age, sex, education, diabetes mellitus, smoking status, stroke, coronary heart disease, systolic blood pressure, hypertension medication use, body mass index, and race (if more than one) and combined using a fixed-effect meta-analysis.*

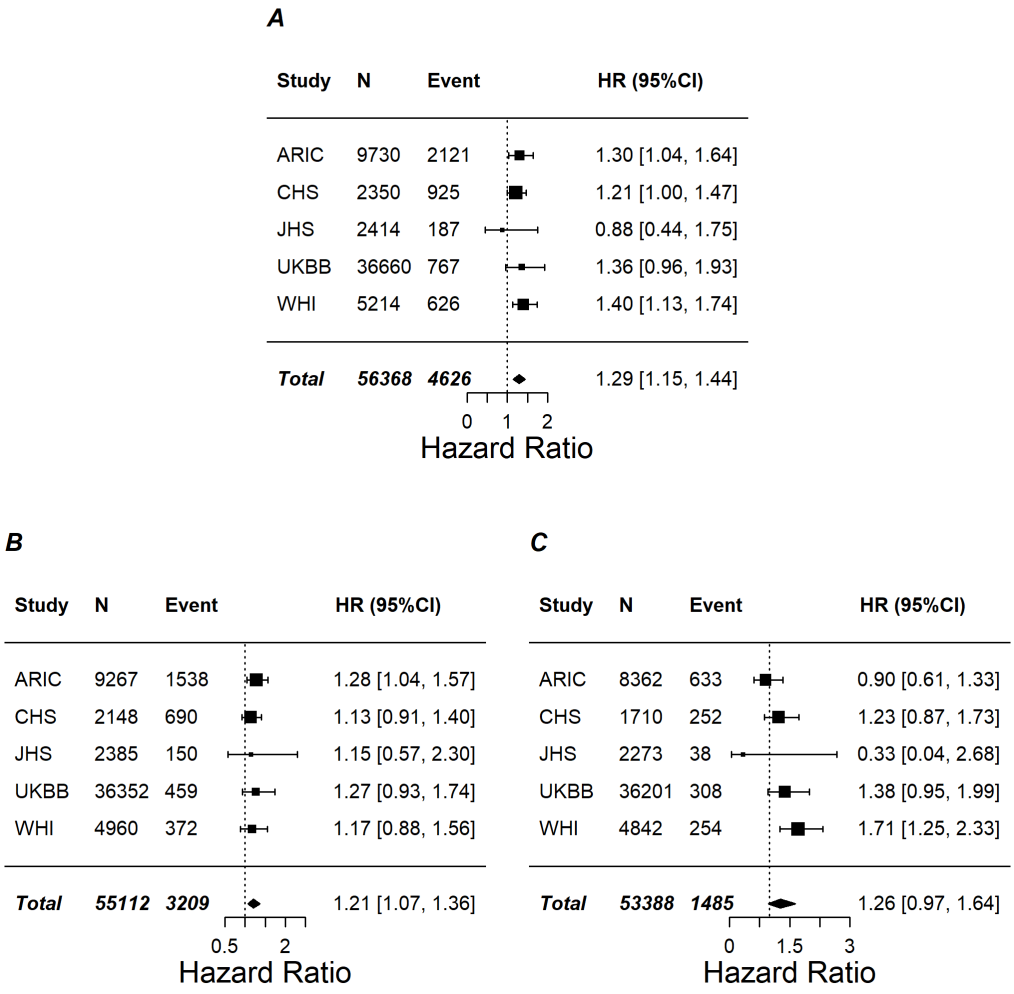


Figure 2. Associations for somatic mutation and heart failure subgroups. *a) clonal hematopoiesis of indeterminate potential with variant allele frequency > 10% and incident heart failure, b) clonal hematopoiesis of indeterminate potential and incident heart failure without prior coronary heart disease, and c) clonal hematopoiesis of indeterminate potential and incident heart failure with prior coronary heart disease. Event represents the number of incident heart failure cases. For each model, multivariable adjusted hazard ratios and 95% CIs were calculated separately in each study adjusting for age, sex, education, diabetes mellitus, smoking status, stroke, coronary heart disease, systolic blood pressure, hypertension medication use, body mass index, and race (if more than one) and combined using a fixed-effect meta-analysis. Coronary heart disease status was not adjusted in the associations of heart failure with or without prior coronary heart disease.*

Follow-up analyses in UKBB were conducted to further investigate the association between CHIP and LVEF. We found that any CHIP was not significantly associated with reduced LVEF (p = 0.07). However, *ASXL1* somatic mutations were

significantly associated with reduced LVEF (beta -4.02%, 95% CI -6.97, -1.06, p=0.008). We did not observe significant associations across *DNMT3A*, *TET2*, *JAK2* specific somatic mutations (**Figure 3.5.4**).

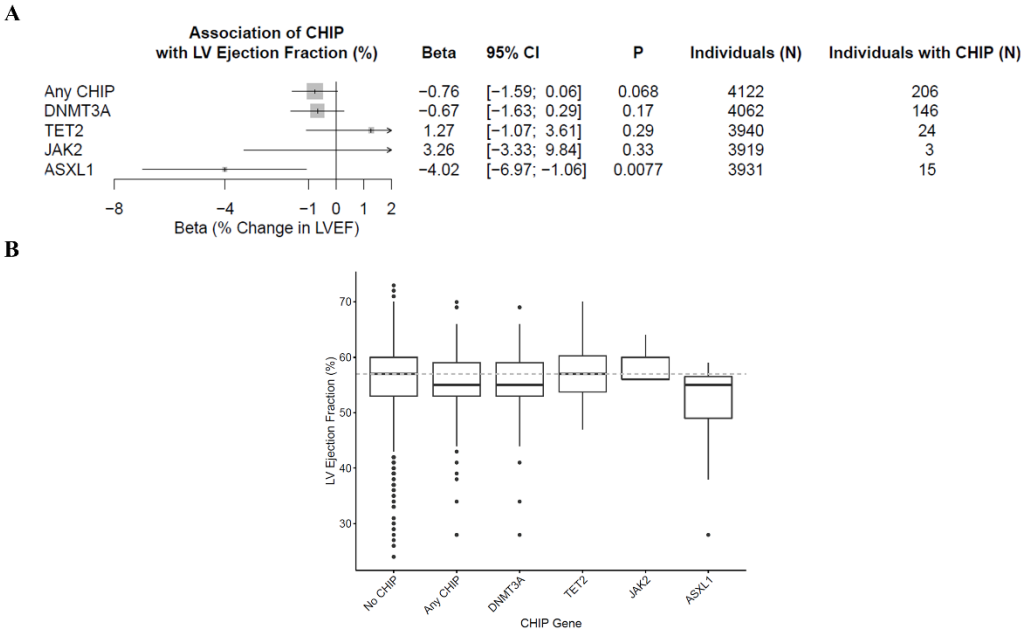


Figure 3.5.4. Clonal hematopoiesis and left ventricular ejection fraction in UK Biobank. *Association of clonal hematopoiesis of indeterminate potential status with left ventricular ejection fraction was performed using a linear regression with the adjustments of age, sex, smoking status, prevalent coronary heart disease, diabetes, systolic blood pressure, and self-reported race in the UK Biobank participants. Unadjusted first quartile, median, and third quartile of left ventricular ejection fraction were presented in the boxplots, and outliers were presented as dots. LVEF = left ventricular ejection fraction.*

While our results suggest a promising link between CHIP and heart failure, some limitations exist. Our results suggesting that gene-specific driver mutations in *TET2, JAK2* and *ASXL1* may be preferentially associated with incident HF risk require confirmation in additional larger studies. One might hypothesize that differences in the

kinetics of clonal expansion of driver mutations (which tend to be greater for *TET2* and *JAK2*) may explain the gene-specific differences in HF risk. Longitudinal studies of CHIP measured at multiple time points in humans may be needed to address this question. In addition, the recent association of multiple CHIP driver mutations with higher HF-related mortality[48] suggest that the presence of multiple CHIP driver mutations may be a surrogate measure for more extensive accumulation of DNA damage or reduced DNA repair or bone marrow-derived endothelial progenitor cell regenerative capacity[49]. Another limitation of the current study was the lack of availability of HF subtype information in a substantial proportion of our overall sample, which limited our ability to explore these associations with adequate power and merits further investigation.

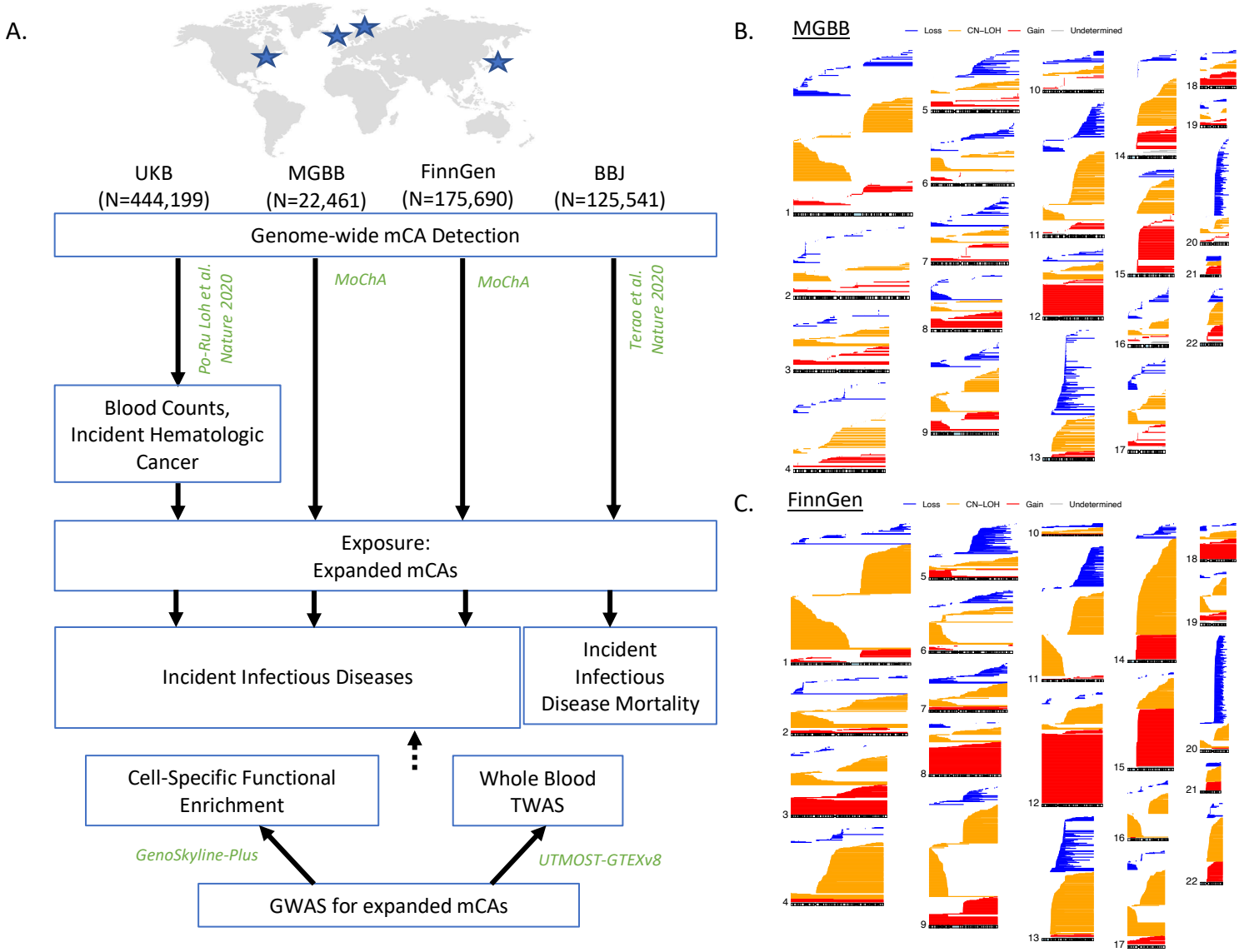
In summary, our findings identify CHIP as a potentially important novel age-related risk factor for HF, consistent with previous findings of the role of CHIP as a risk factor for age-related atherosclerotic CVD more broadly. If confirmed, these findings ultimately may have potential implications for development or targeting of anti-inflammatory therapies such IL-1beta or NLRP3 inflammasome inhibitors in HF patients.

Chapter 3.6: Association of mCAs with diverse infectious diseases, including COVID-19 infection

With advancing age comes increased susceptibility to infectious diseases[50 51]. Immunosenescence is the age-related erosion of immune function, particularly with respect to adaptive immunity[52-55]. Leukocytes, including T-cells and B-cells, are key mediators of adaptive host defenses against infections, with impaired immune responses increasing risk for infections[56-58]. Age-related mosaic chromosomal alterations (mCAs) detected from blood-derived DNA, are clonal structural somatic alterations (deletions, duplications, or copy neutral loss of heterozygosity) present in a fraction of peripheral leukocytes that can indicate clonal hematopoiesis (CH)[2 9 10]. mCAs are associated with aberrant leukocyte cell counts, and increased risks for hematological malignancy and mortality[2 9 10 26 59-63].

While the relationship between mCAs and increased hematologic cancer risk is well established[2 9 10], the impact of mCAs on age-related diminishment in immune function is poorly understood. We hypothesized that mCAs increase risk of infection since mCAs are somatic variants that increase in abundance with age and are associated with alterations in leukocyte count. In this study, we harnessed DNA genotyping array intensity data and long-range chromosomal phase information inferred from 767,891 individuals across four countries to analyze the associations between expanded mCA clones (i.e., mCAs present in at least 10% of peripheral leukocyte DNA indicative of clonal expansion) and diverse infections, including severe coronavirus disease 2019

(COVID-19) from SARS-CoV-2 infection (**Figure 3.6.1**). To elucidate genetic risk factors for the development of expanded mCA clones, we performed a genome-wide



association study (GWAS) in the UK Biobank and subsequent *in silico* cell-specific, transcriptomic, and pathway analyses.

Figure 3.6.1: *Study schematic.* ***a.*** *Genome-wide mCAs were detected across the UKB[10], MGBB (via the MoChA pipeline), FinnGen (via the MoChA pipeline), and BBJ[9]. Association of expanded mCAs (cell fraction >10%) with incident infectious diseases in*

UKB, MGBB, and FinnGen and with incident infectious disease mortality in BBJ was performed. A GWAS for expanded mCAs was then performed in the UKB to discover causal factors for expanded mCAs. Using the GWAS results, cell-specific functional enrichment analyses were performed using GenoSkyline-Plus, which combines epigenetic and transcriptomic annotations with GWAS summary statistics to estimate the relative contribution of cell-specific functional markers to the GWAS results. Additionally, to prioritize putative causal genes and pathways promoting the development of expanded mCAs, whole blood TWAS was performed using UTMOST via GTEx v8. ***b-c.*** *mCA pileup plots across chromosomes 1-22, showing the calls made in MGBB and FinnGen, where each mCA is a separate horizontal line. Blue refers to loss, yellow to CN-LOH, red to gain, and grey to undetermined mCAs.*

mCA presence across the genome was associated with diverse incident infections (as defined in Zekavat et al. Nature Medicine 2021[15]) (HR 1.06; 95% CI 1.04 to 1.09; $P=8.6x10^{-8}$) (**Figure 3.6.2**), independent of age, age^2, sex, smoking status, and first 10 principal components of ancestry in the combined UKB, MGBB, and FinnGen meta-analysis. The dependence of this association with mCA cell fraction is further visualized in **Figure 3.6.3**, which shows an increase in proportion of incident infection cases and incident sepsis cases with cell fraction, with stronger slopes at approximately cell fraction >10%, the cutoff for our expanded mCA definition. Accordingly, the association across diverse infections was stronger for expanded mCA clones, (HR 1.12; 95% CI 1.1 to 1.2; $P=6.3x10^{-7}$) (**Figure 3.6.2-A**). Furthermore, among expanded mCA clones, the strongest association was observed among expanded autosomal mCAs (HR 1.3; 95% CI 1.1 to 1.4; $P=1.8x10^{-7}$) (**Figure 3.6.2-B)**. In particular, expanded autosomal mCAs were associated with sepsis (HR 2.7; 95% CI 2.3 to 3.2; $P=3.1x10^{-28}$), respiratory system infections (HR 1.4; 95% CI 1.2 to 1.5; $P=3.8x10^{-10}$), digestive system infections (HR 1.5; 95% CI 1.3 to 1.7; $P=2.2x10^{-9}$), and genitourinary system infections (HR 1.3; 95% CI 1.1 to 1.4; $P=3.7x10^{-4}$) (**Figure 3.6.2-B)**. The specific mCAs implicated for infection were diverse in nature – across all chromosomes, of different sizes, and mixed across gain, loss, and

copy-number neutral loss of heterozygosity (CNN-LOH) mCAs (**Figure 3.6.4**). Further associations across 20 specific infectious disease subcategories identified significant associations for pneumonia (HR 1.8, 95% CI 1.5 to 2.0, $P=2.3x10^{-15}$), any infection within the ICD-10 A00-B99 category (HR 1.4, 95% CI 1.2 to 1.5, $P=1x10^{-10}$), gastroenteritis (HR 1.4, 95% CI 1.2 to 1.7, $P=9.0x10^{-6}$), other lower respiratory infections (HR 1.3, 95% CI 1.2 to 1.5, $P=2.8x10^{-5}$), and pyelonephritis or urinary tract infection (HR 1.2, 95% CI 1.1 to 1.4, P=0.0018) (**Figure 3.6.5**).

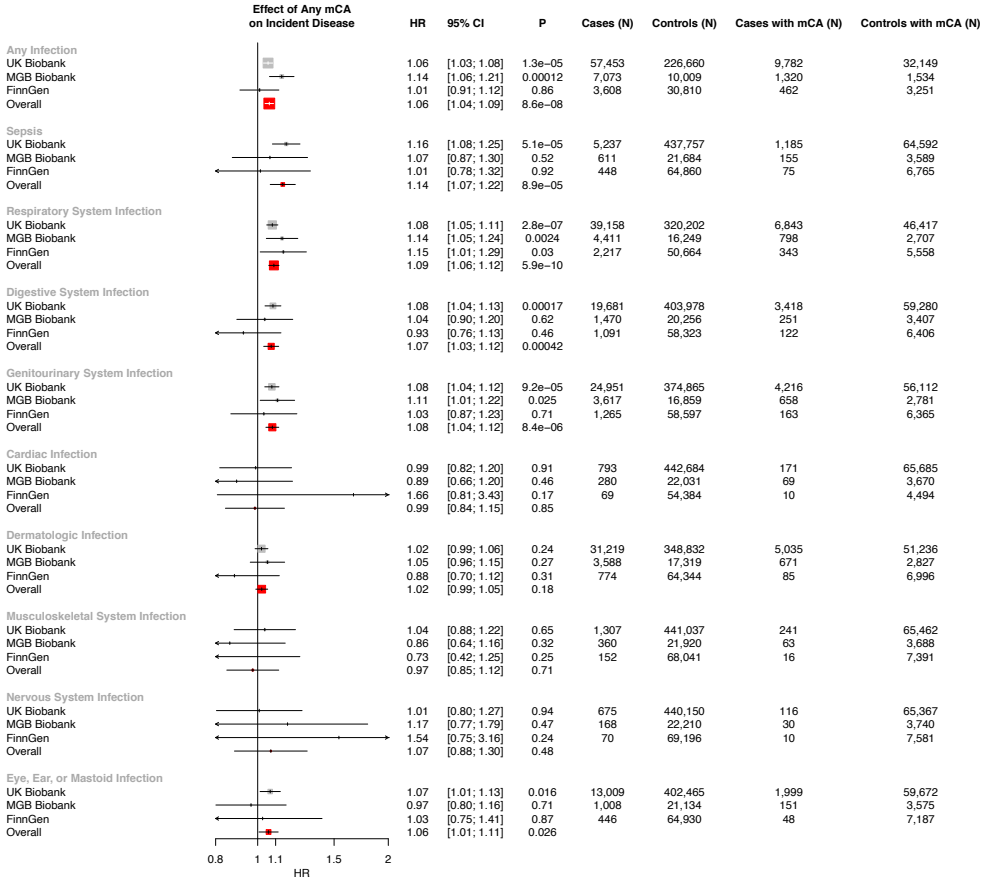
Risks for incident fatal infections were assessed in BBJ since non-fatal incident infectious disease events are currently unavailable in BBJ. Among individuals without any cancer history in BBJ, autosomal mCAs showed nominal associations with fatal incident infections (any infection: HR 1.12, 95% CI 1.0 to 1.2 P=0.04; nervous system infection: HR 2.8, 95% CI 1.1 to 6.9, P=0.02; respiratory system infection: HR 1.15, 95% CI 1.0 to .3, P=0.03), with expanded autosomal mCAs being associated with incident sepsis mortality (HR 2.0; 95% CI 1.0 to 4.2; P=0.05) (**Figure 3.6.6)**, as well as pneumonia history (OR 1.3; 95% CI: 1.1 to 1.5; P=0.0019).

Sensitivity analysis for the association of expanded autosomal mCAs and incident sepsis found that the association was consistently significant across different age groups (**Figure 3.6.7**), and that it was additionally independent of a 25-factor smoking covariate[26], body mass index, type 2 diabetes mellitus, leukocyte count, lymphocyte count, and lymphocyte percentage.

Stratified analyses indicated expanded autosomal mCAs in individuals with cancer prior to infection (either any solid tumors, or hematologic malignancy after time of blood draw for genotyping) conferred stronger effects for sepsis (HR 2.8; 95% CI 2.3 to 3.4; $P=9.7x10^{-26}$) and respiratory system infections (HR 1.6; 95% CI 1.4 to 1.8; $P=6.1x10^{-12}$) compared to individuals without a prior cancer history (sepsis: HR 1.3; 95% CI 0.8 to 2.0; P=0.33, $P_{heterogeneity}$=0.001; respiratory system infections: HR 1.2; 95% CI 1.0 to 1.3; P=0.045, $P_{interaction}$=0.001) (**Figure 3.6.8**). Interestingly, this interaction was driven by prevalent solid cancer, not hematologic cancer after DNA acquisition for mCA genotyping (**Table 3.6.1**). Further multivariable adjustment indicated that incident sepsis and infection were independent of chemotherapy, neutropenia, aplastic anemia, decreased white blood cell count, bone marrow or stem cell transplant, and radiation effects prior to infection (with these phenotypes defined using ICD-10 and ICD-9 phecode groupings[28]) (**Table 3.6.2**).

For sex chromosome mCAs, while none of the incident infections achieved statistical significance (P<0.005) in meta-analysis across the three cohorts, expanded chrX and chrY mCAs were suggestively associated with respiratory system infections (expanded chrX: HR 1.5; 95% CI 1.01to 1.9; P=0.0068; expanded chrY: HR 1.09; 95% CI 1.0 to 1.2; P=0.005), independent of age, age^2, sex, smoking status, and first 10 principal components of ancestry (**Figure 3.6.9**).

A.



B.

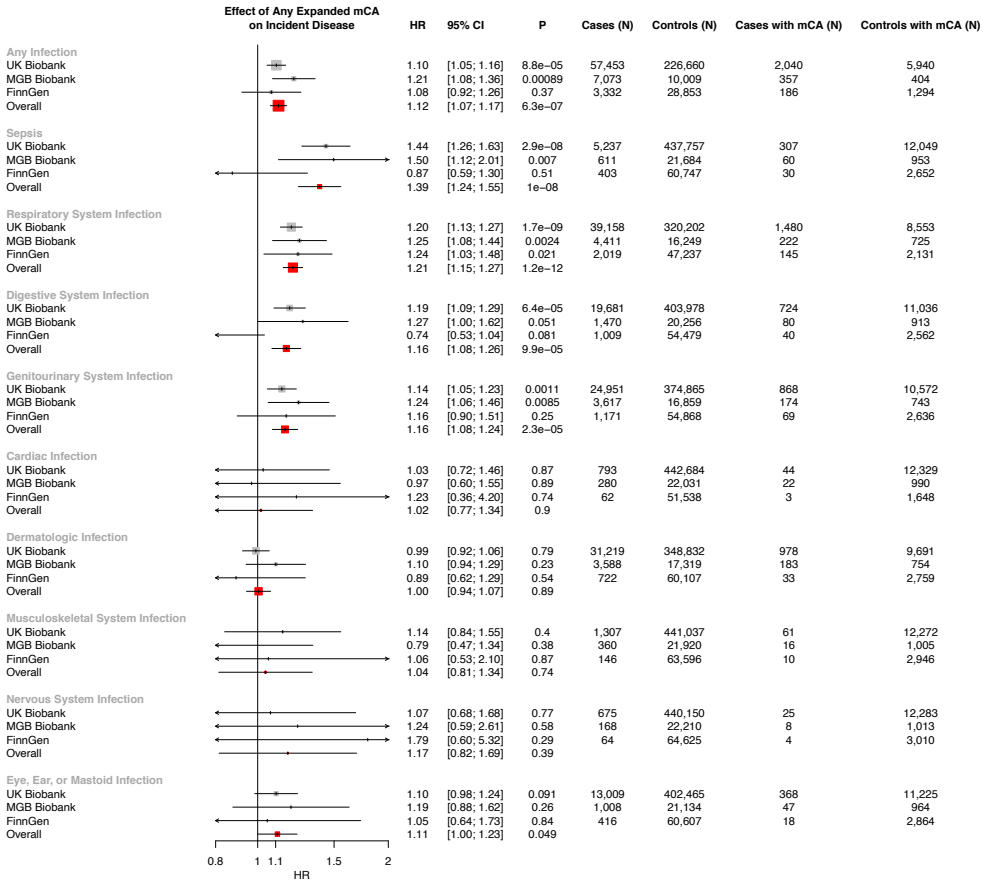


Figure 3.6.2: Associations of A) any mCA and B) any expanded mCA with incident infections. mCA = mosaic chromosomal alterations.

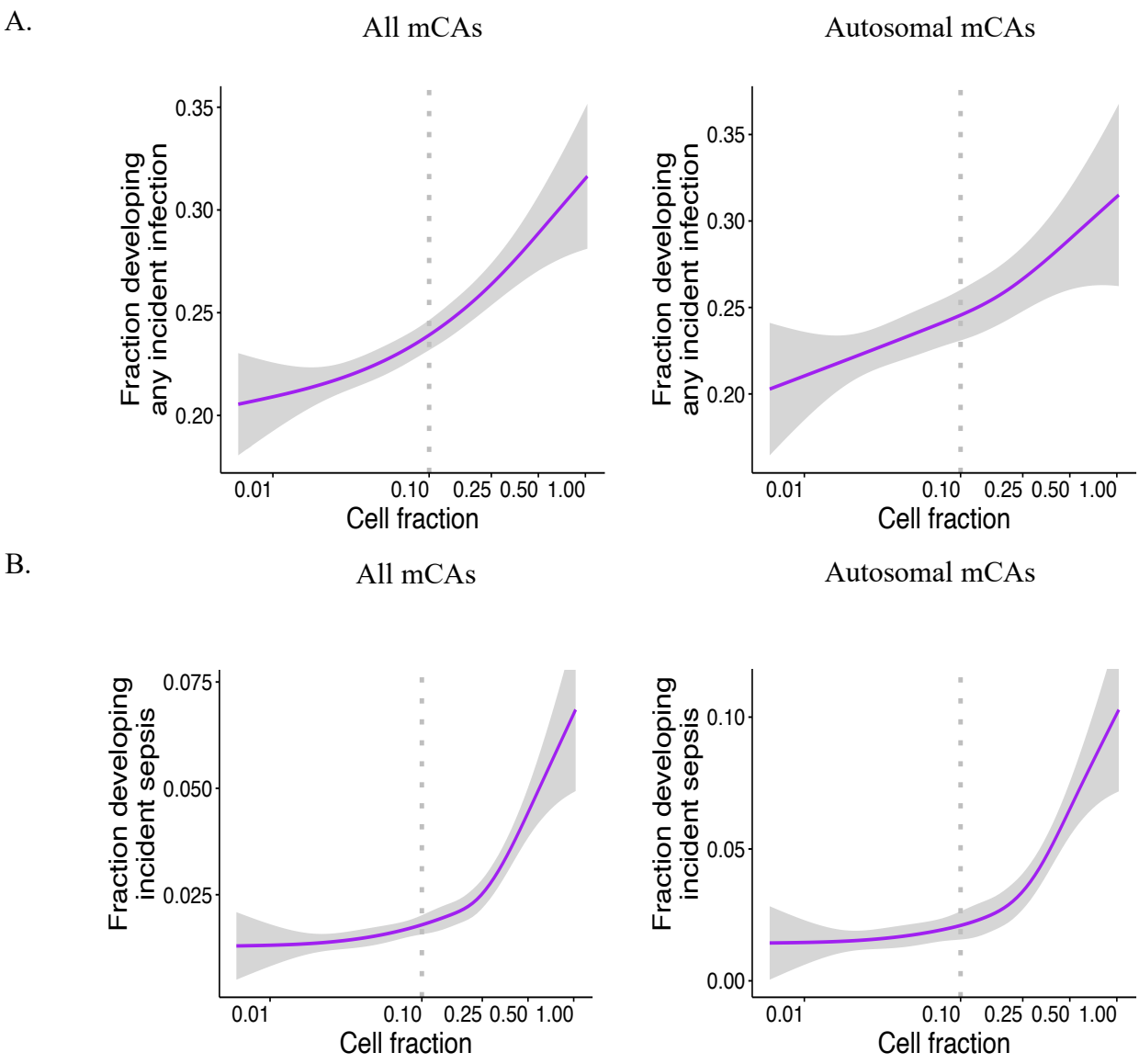


Figure 3.6.3: Associations of mCA cell fraction with A. any incident infection and B. incident sepsis in the UKB among individuals without prevalent hematologic cancer at time of blood draw for genotyping across all mCAs and separately, autosomal mCAs. *The dotted vertical lines at cell fraction of 0.10 represents the cutoff for the expanded mCA definition. Individuals with known hematologic cancer at time of or prior to blood draw for genotyping were excluded.*

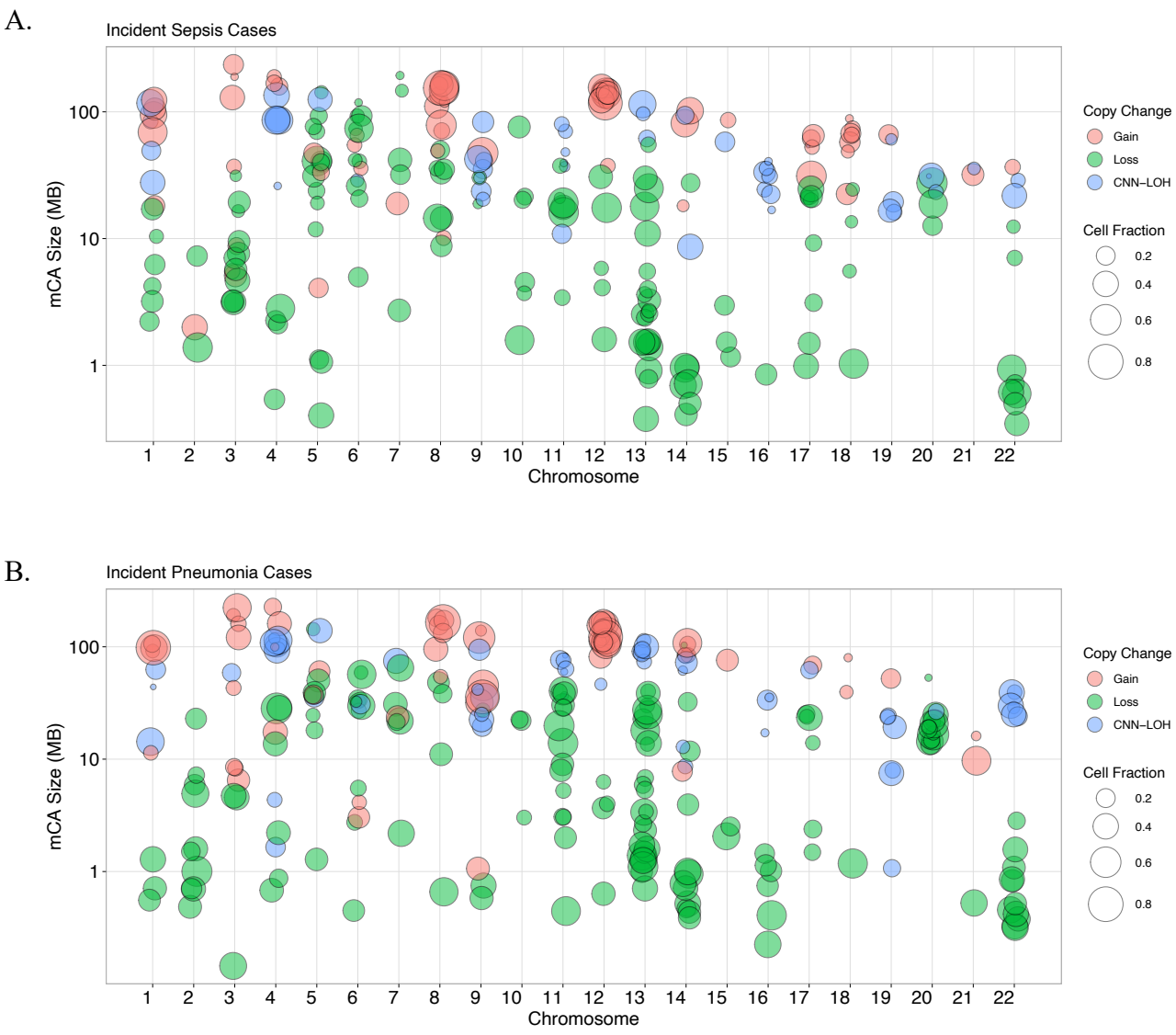


Figure 3.6.4: Visualization of the diverse range of expanded autosomal mCAs detected across the genome among individuals with A. incident sepsis and B. incident pneumonia in the UKB. *Each point represents one mCA carried by a case, with the x-axis as the chromosome, y-axis as the mCA size in mega-bases of DNA (MB), color as the copy change, and size of the point as the cell fraction of that mCA. CNN-LOH=copy number neutral loss of heterozygosity, MB = megabases of DNA, mCA = mosaic chromosomal alterations*

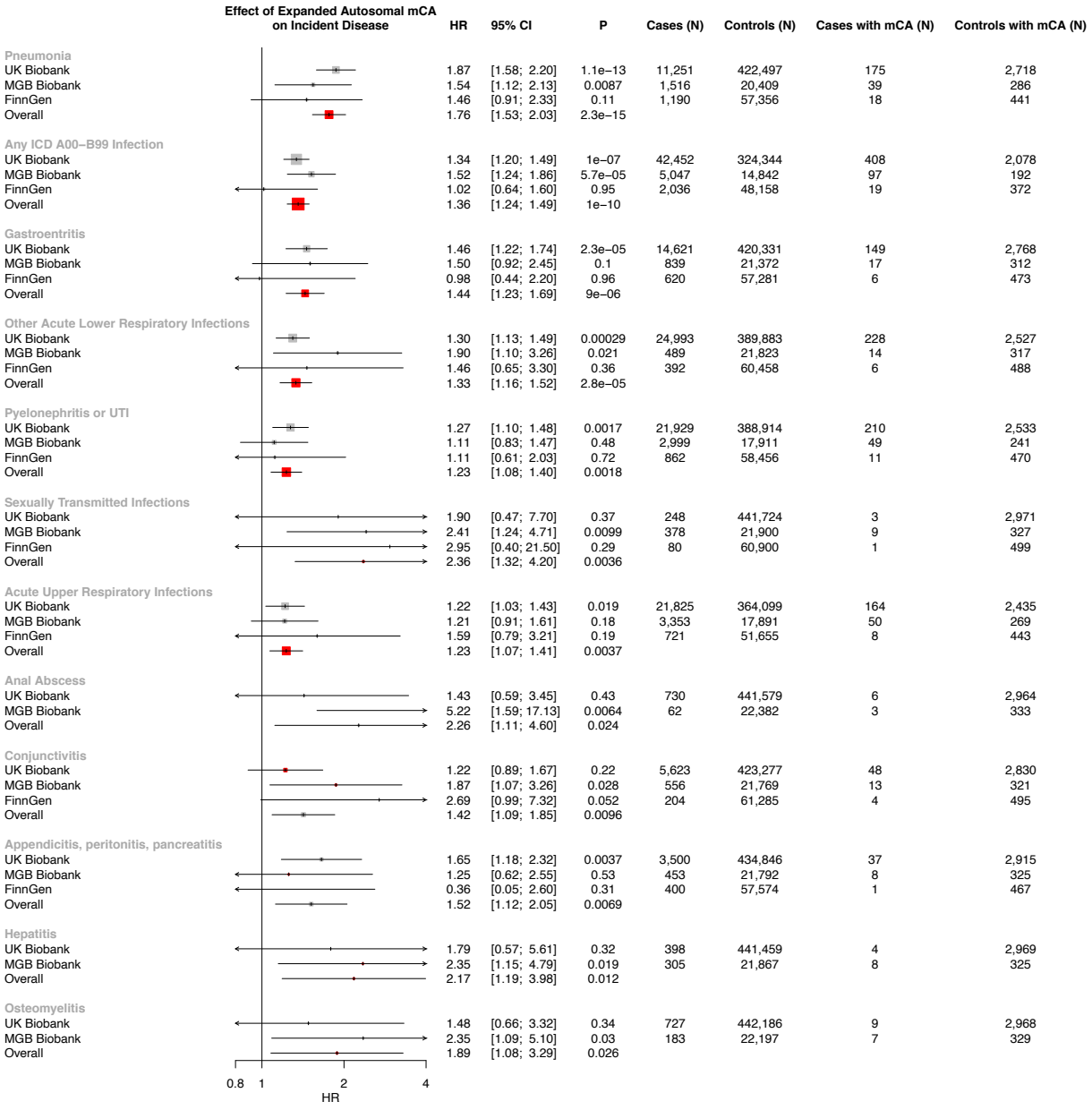
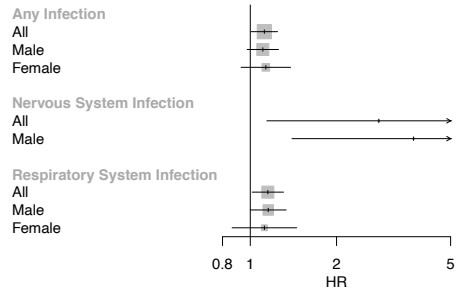

Effect of Expanded Autosomal mCA on Incident Disease
HR
95% CI
P
Cases (N)
Controls (N)
Cases with mCA (N)
Controls with mCA (N)
Pneumonia
UK Biobank 1.87 [1.58; 2.20] 1.1e−13 11,251 422,497 175 2,718
MGB Biobank 1.54 [1.12; 2.13] 0.0087 1,516 20,409 39 286
FinnGen 1.46 [0.91; 2.33] 0.11 1,190 57,356 18 441
Overall 1.76 [1.53; 2.03] 2.3e−15
Any ICD A00−B99 Infection
UK Biobank 1.34 [1.20; 1.49] 1e−07 42,452 324,344 408 2,078
MGB Biobank 1.52 [1.24; 1.86] 5.7e−05 5,047 14,842 97 192
FinnGen 1.02 [0.64; 1.60] 0.95 2,036 48,158 19 372
Overall 1.36 [1.24; 1.49] 1e−10
Gastroentritis
UK Biobank 1.46 [1.22; 1.74] 2.3e−05 14,621 420,331 149 2,768
MGB Biobank 1.50 [0.92; 2.45] 0.1 839 21,372 17 312
FinnGen 0.98 [0.44; 2.20] 0.96 620 57,281 6 473
Overall 1.44 [1.23; 1.69] 9e−06
Other Acute Lower Respiratory Infections
UK Biobank 1.30 [1.13; 1.49] 0.00029 24,993 389,883 228 2,527
MGB Biobank 1.90 [1.10; 3.26] 0.021 489 21,823 14 317
FinnGen 1.46 [0.65; 3.30] 0.36 392 60,458 6 488
Overall 1.33 [1.16; 1.52] 2.8e−05
Pyelonephritis or UTI
UK Biobank 1.27 [1.10; 1.48] 0.0017 21,929 388,914 210 2,533
MGB Biobank 1.11 [0.83; 1.47] 0.48 2,999 17,911 49 241
FinnGen 1.11 [0.61; 2.03] 0.72 862 58,456 11 470
Overall 1.23 [1.08; 1.40] 0.0018
Sexually Transmitted Infections
UK Biobank 1.90 [0.47; 7.70] 0.37 248 441,724 3 2,971
MGB Biobank 2.41 [1.24; 4.71] 0.0099 378 21,900 9 327
FinnGen 2.95 [0.40; 21.50] 0.29 80 60,900 1 499
Overall 2.36 [1.32; 4.20] 0.0036
Acute Upper Respiratory Infections
UK Biobank 1.22 [1.03; 1.43] 0.019 21,825 364,099 164 2,435
MGB Biobank 1.21 [0.91; 1.61] 0.18 3,353 17,891 50 269
FinnGen 1.59 [0.79; 3.21] 0.19 721 51,655 8 443
Overall 1.23 [1.07; 1.41] 0.0037
Anal Abscess
UK Biobank 1.43 [0.59; 3.45] 0.43 730 441,579 6 2,964
MGB Biobank 5.22 [1.59; 17.13] 0.0064 62 22,382 3 333
Overall 2.26 [1.11; 4.60] 0.024
Conjunctivitis
UK Biobank 1.22 [0.89; 1.67] 0.22 5,623 423,277 48 2,830
MGB Biobank 1.87 [1.07; 3.26] 0.028 556 21,769 13 321
FinnGen 2.69 [0.99; 7.32] 0.052 204 61,285 4 495
Overall 1.42 [1.09; 1.85] 0.0096
Appendicitis, peritonitis, pancreatitis
UK Biobank 1.65 [1.18; 2.32] 0.0037 3,500 434,846 37 2,915
MGB Biobank 1.25 [0.62; 2.55] 0.53 453 21,792 8 325
FinnGen 0.36 [0.05; 2.60] 0.31 400 57,574 1 467
Overall 1.52 [1.12; 2.05] 0.0069
Hepatitis
UK Biobank 1.79 [0.57; 5.61] 0.32 398 441,459 4 2,969
MGB Biobank 2.35 [1.15; 4.79] 0.019 305 21,867 8 325
Overall 2.17 [1.19; 3.98] 0.012
Osteomyelitis
UK Biobank 1.48 [0.66; 3.32] 0.34 727 442,186 9 2,968
MGB Biobank 2.35 [1.09; 5.10] 0.03 183 22,197 7 329
Overall 1.89 [1.08; 3.29] 0.026
0.8
1
2
4
HR

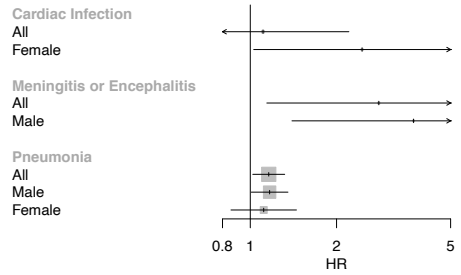
Figure 3.6.5: Suggestive associations (P<0.05) of expanded autosomal mCAs with incident infection categories.

A.



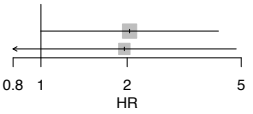
Effect of Autosomal mCAs on Incident Disease Mortality	HR	95% CI	P	Cases (N)	Controls (N)	Cases with mCA (N)	Controls with mCA (N)
Any Infection							
All	1.12	[1.01; 1.2]	0.037	1,998	72,317	476	10,485
Male	1.11	[0.97; 1.2]	0.12	1,329	39,978	346	6,298
Female	1.13	[0.93; 1.4]	0.22	669	32,339	130	4,187
Nervous System Infection							
All	2.81	[1.14; 6.9]	0.025	21	72,317	8	10,485
Male	3.71	[1.40; 9.9]	0.0085	17	39,978	8	6,298
Respiratory System Infection							
All	1.15	[1.02; 1.3]	0.028	1,353	72,317	346	10,485
Male	1.16	[1.00; 1.3]	0.05	971	39,978	269	6,298
Female	1.12	[0.86; 1.4]	0.39	382	32,339	77	4,187

B.



Effect of Autosomal mCAs on Incident Disease Mortality	HR	95% CI	P	Cases (N)	Controls (N)	Cases with mCA (N)	Controls with mCA (N)
Cardiac Infection							
All	1.11	[0.56; 2.2]	0.77	50	72,317	11	10,485
Female	2.46	[1.03; 5.9]	0.043	25	32,339	8	4,187
Meningitis or Encephalitis							
All	2.81	[1.14; 6.9]	0.025	21	72,317	8	10,485
Male	3.71	[1.40; 9.9]	0.0085	17	39,978	8	6,298
Pneumonia							
All	1.16	[1.02; 1.3]	0.023	1,299	72,317	334	10,485
Male	1.17	[1.01; 1.4]	0.039	928	39,978	259	6,298
Female	1.11	[0.86; 1.4]	0.42	371	32,339	75	4,187

C.



Effect of Expanded Autosomal mCAs on Incident Disease Mortality	HR	95% CI	P	Cases (N)	Controls (N)	Cases with mCA (N)	Controls with mCA (N)
Sepsis							
All	2.04	[1.0; 4.2]	0.05	276	72,317	8	753
Male	1.95	[0.8; 4.8]	0.14	164	39,978	5	471

Figure 3.6.6: Suggestive associations (P<0.05) of mCAs with incident infection-related mortality in Biobank Japan. *Associations of autosomal mCAs with A) organ-system level infections and B) specific infection categories. C) Association of expanded autosomal mCAs with Sepsis. Associations are presented among individuals without any cancer history.*

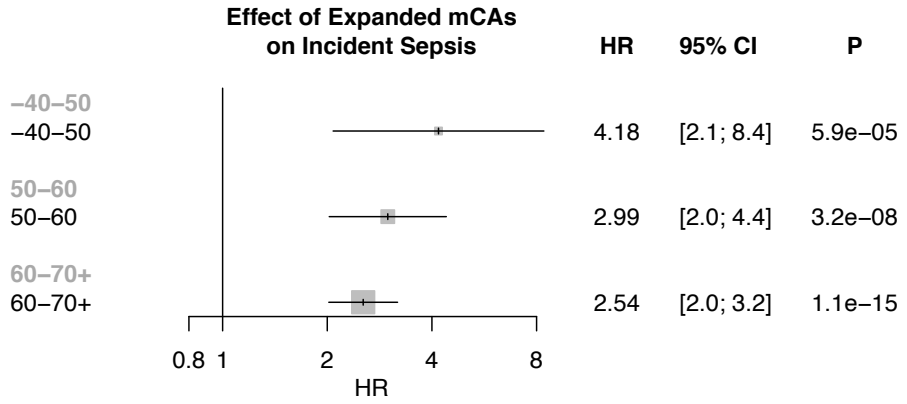


Figure 3.6.7: Associations of expanded autosomal mCAs with incident sepsis and among different age strata in the UK Biobank. Individuals with prevalent hematologic cancer were excluded from analyses. Associations were adjusted for sex, ever smoking status, and principal components 1-10 of ancestry. mCA = mosaic chromosomal alterations.

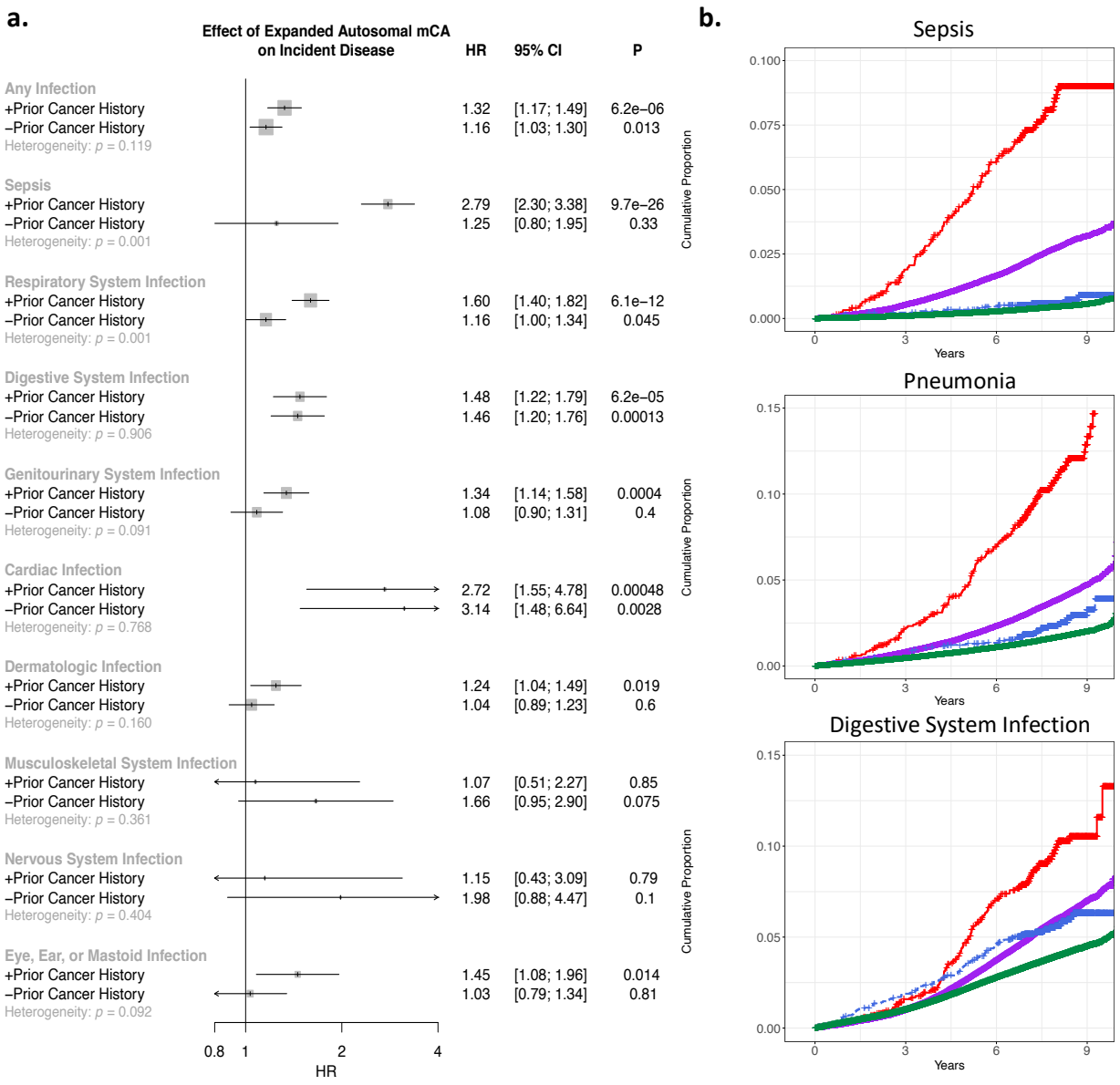


Figure 3.6.8: Association of expanded autosomal mCAs and incident infections, stratified by antecedent cancer history. a. *Association of expanded autosomal mCAs with incident infections across individuals with and without a cancer history before their incident infection, meta-analyzed across UKB, MGBB, and FinnGen combined assuming a fixed effect. Error bars show the 95% confidence interval for estimates. Bonferroni correction was used to determine the level of statistical significance. Individuals with known hematologic cancer at time of or prior to blood draw for genotyping were excluded. Analyses are adjusted for age, age2, sex, smoking status, and principal components of ancestry.* ***b.*** *Cumulative incidence curves for various infections in UKB. Top: sepsis, middle: pneumonia, bottom: digestive system infection.* ***Red:*** *mCA+ Cancer+,* ***Purple:*** *mCA- Cancer+,* ***Blue:*** *mCA+ Cancer-,* ***Green:*** *mCA- cancer-. Individuals with known hematologic cancer at time of or prior to blood draw for genotyping were excluded.*

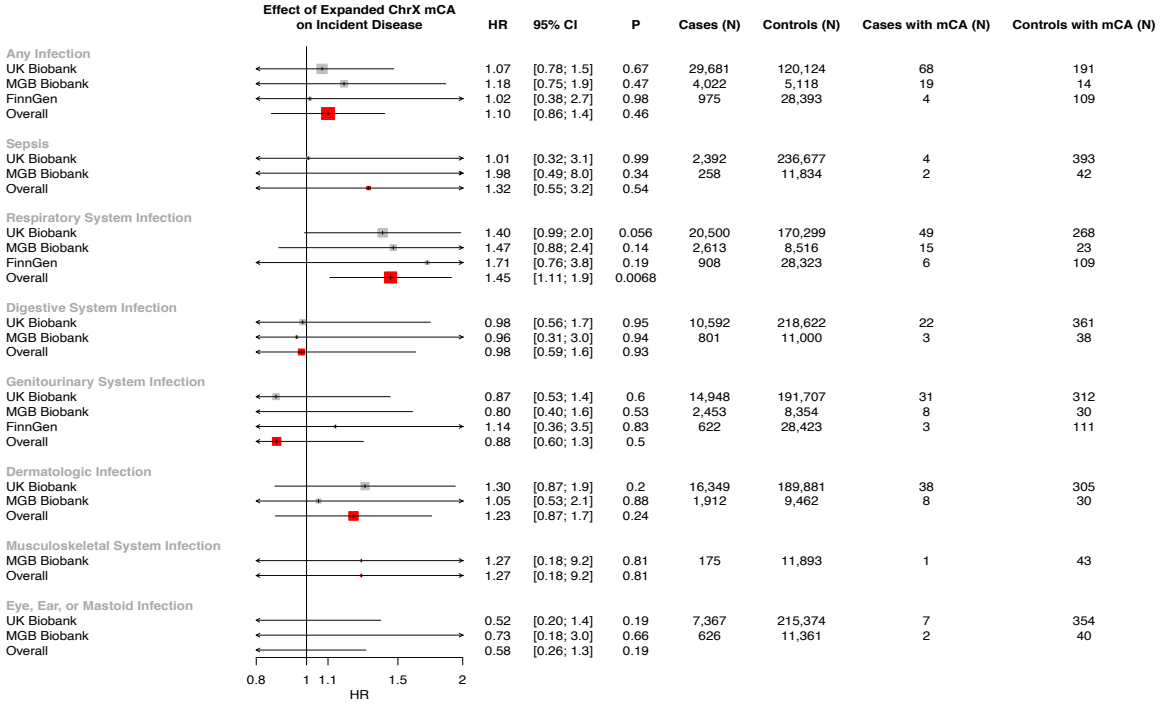
Table 3.6.1: *Sensitivity analysis of incident sepsis and pneumonia association in the UK Biobank among populations of individuals with different types of cancer prior to incident infection, where solid cancer is defined as any non-hematologic cancer. Other covariates in the model included age, age^2, sex, smoking status, and PC1-10 of ancestry.*

Outcome	*Population of people with cancer prior to infection*	*HR*	*P*	*Lower 95% CI*	*Upper 95% CI*	*Cases (N)*	*Controls (N)*	*Cases with mCA (N)*	*Controls with mCA (N)*
Sepsis	Prevalent Solid Cancer	1.86	0.0097	1.16	2.96	1258	64921	20	521
	Incident Solid Cancer Prior to Infection	0.76	0.44	0.38	1.53	1619	51867	10	361
	Incident Hematologic Cancer Prior to Infection	0.98	0.88	0.77	1.25	833	2864	83	312
	Incident Hematologic Cancer and Prevalent Solid Cancer Prior to Infection	0.94	0.85	0.52	1.72	144	546	15	63
	Any Cancer Prior to Infection	**2.82**	**5.28E-22**	**2.28**	**3.48**	**3575**	**119106**	**99**	**1131**
Pneumonia	Prevalent Solid Cancer	1.68	0.0057	1.16	2.43	2382	62325	40	480
	Incident Solid Cancer Prior to Infection	1.33	0.18	0.87	2.03	2369	49466	24	323
	Incident Hematologic Cancer Prior to Infection	1.19	0.18	0.92	1.54	655	2886	80	300
	Incident Hematologic Cancer and Prevalent Solid Cancer Prior to Infection	1.73	0.076	0.94	3.18	119	528	16	55
	Any Cancer Prior to Infection	**2.26**	**5.08E-17**	**1.86**	**2.73**	**5295**	**114149**	**130**	**1048**

Table 3.6.2: *Sensitivity analysis of incident sepsis and pneumonia association in the UK Biobank among those with cancer prior to incident infection, adjusting for chemotherapy, neutropenia, aplastic anemia, decreased white blood cell count, bone marrow or stem cell transplant, and radiation effects prior to infection (as defined using the Vanderbilt ICD-10 and ICD-9 phecode groupings[28]). Other covariates in the model included age, age^2, sex, smoking status, and PC1-10 of ancestry.*

	Adjustment	*HR*	*P*	*Lower 95% CI*	*Upper 95% CI*	*Cases (N)*	*Controls (N)*	*Cases with mCA (N)*	*Control mCA*
Sepsis	Chemotherapy	2.48	3.04E-17	2.01	3.06	3575	119106	99	113
	Neutropenia	1.65	3.98E-06	1.33	2.04	3575	119106	99	113
	Aplastic anemia	2.58	1.84E-18	2.09	3.19	3575	119106	99	113
	Decreased white blood cell count	1.65	3.98E-06	1.33	2.04	3575	119106	99	113
	Bone marrow or stem cell transplant	2.77	3.25E-21	2.24	3.42	3575	119106	99	113
	Effects radiation NOS	2.84	2.85E-22	2.30	3.51	3575	119106	99	113
Pneumonia	Chemotherapy	2.11	1.47E-14	1.74	2.55	5295	114149	130	104
	Neutropenia	1.99	1.38E-12	1.65	2.41	5295	114149	130	104
	Aplastic anemia	2.16	2.17E-15	1.79	2.62	5295	114149	130	104
	Decreased white blood cell count	1.99	1.38E-12	1.65	2.41	5295	114149	130	104
	Bone marrow or stem cell transplant	2.20	5.04E-16	1.82	2.66	5295	114149	130	104
	Effects radiation NOS	2.27	2.59E-17	1.88	2.75	5295	114149	130	104

A.



B.

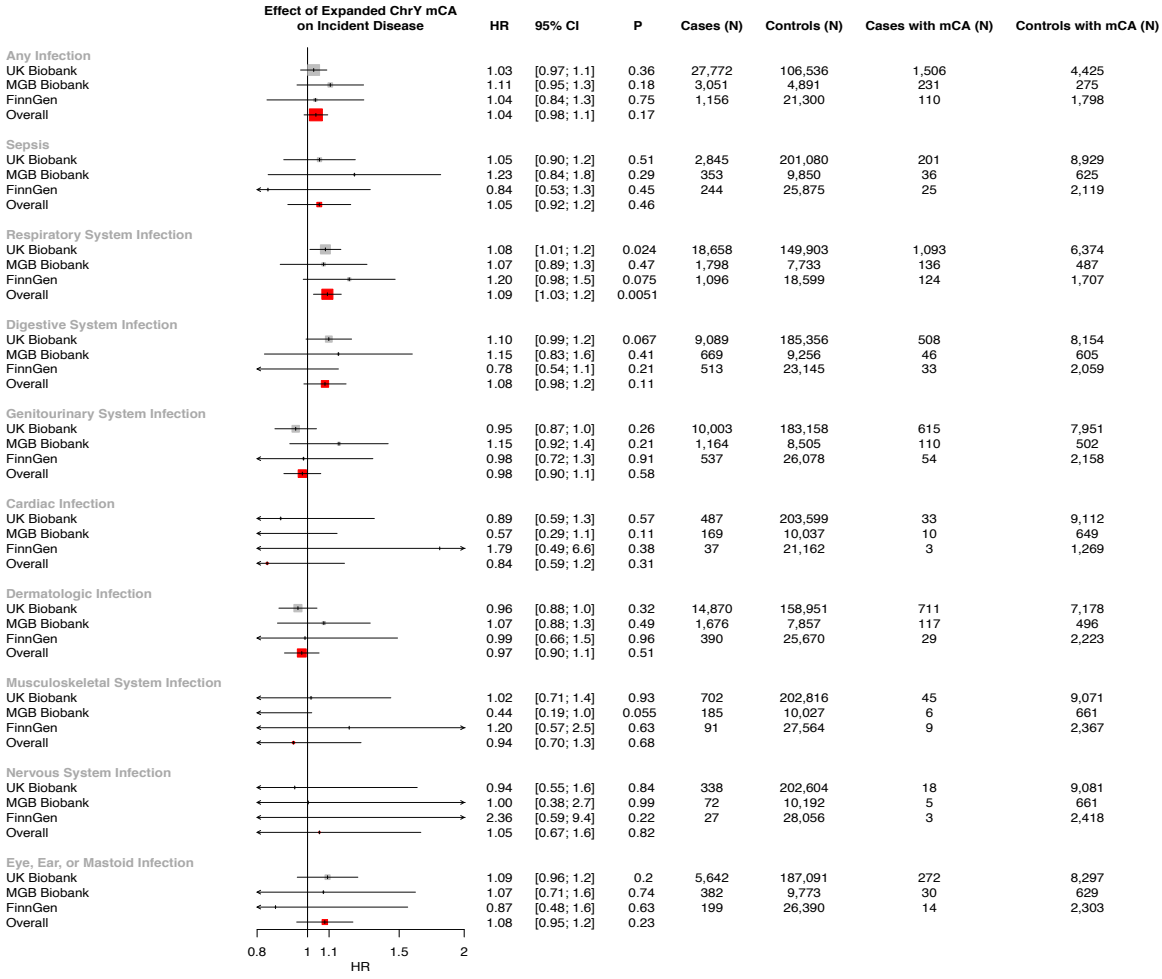


Figure 3.6.9: *Associations of A) expanded ChrY and B) expanded ChrX mCAs with incident infections.*

Association with COVID-19 hospitalization

Across 719 COVID-19 hospitalized cases in the UKB, 44 individuals (6%) carried an expanded mCA clone at time of enrollment (in 2010), versus 3% among 337,877 controls. Adjusting for age, age^2, sex, prior or current smoking status, and principal components of ancestry, expanded mCAs were associated with COVID-19 hospitalizations (OR 1.6; 95% CI 1.1 to 2.2; P=0.0082), with similar effects with expanded autosomal mCAs (OR 2.2; 95% CI 1.2 to 4.1; P=0.02) (**Figure 3.6.10-A**). Analyses in FinnGen showed evidence of replication albeit with a relatively small number of events. The meta-analyzed associations across UKB and FinnGen of expanded autosomal mCAs on COVID-19 hospitalization was OR 2.4, 95% CI 1.3 to 4.5, P=0.004 (**Figure 3.6.10-A**). In the UKB, further sensitivity analysis was performed; the associations persisted with additional adjustment for normalized Townsend deprivation index, normalized body mass index, type 2 diabetes mellitus, hypertension, coronary artery disease, any cancer, asthma, and chronic obstructive pulmonary disease, finding similar associations (**Figure 3.6.11-A**). Additionally, similar associations were observed in the UKB when comparing COVID-19 hospitalization to tested negative controls, and COVID-19 positive versus all from English provinces and, separately, versus tested negative controls (**Figure 3.6.11-B**). Similar effects associations of expanded mCAs with COVID-19 across expanded mCAs were also observed with incident pneumonia in the UKB (**Figure 3.6.12**).

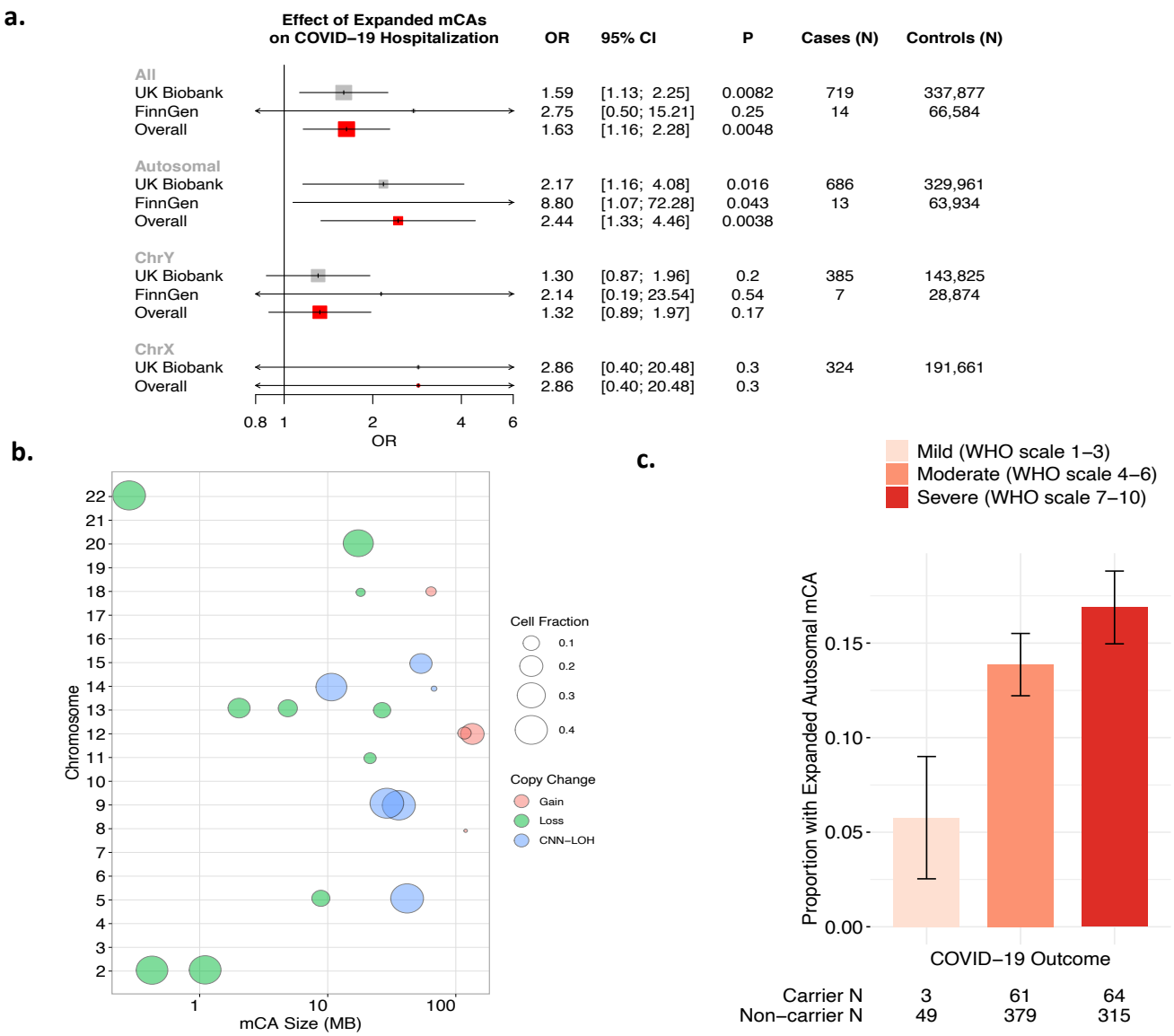
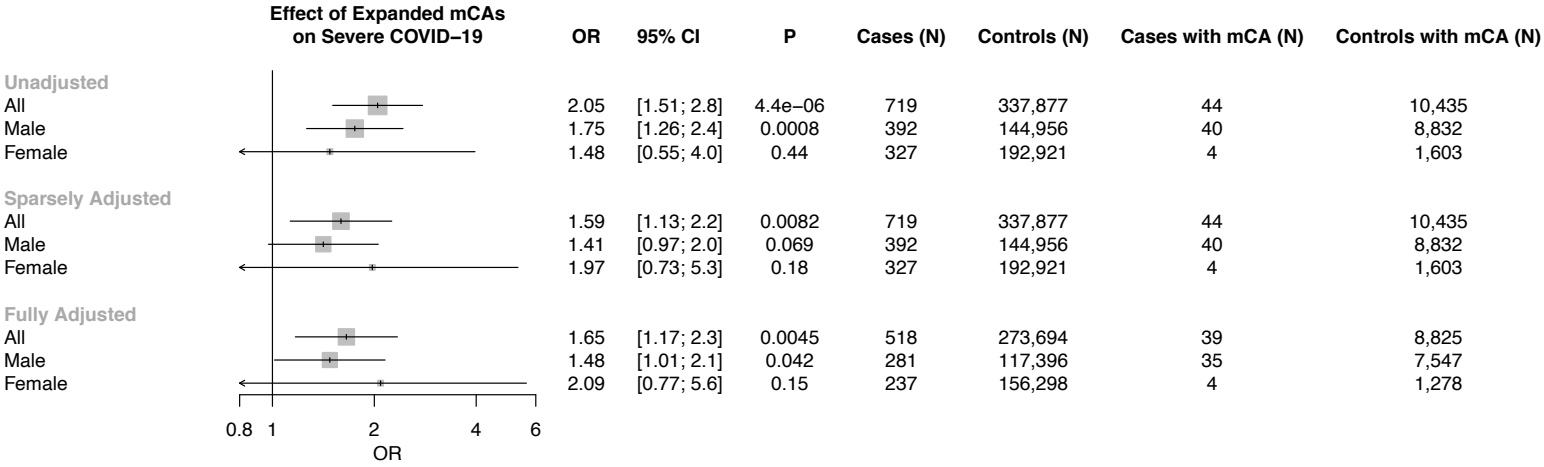


Figure 3.6.10: Association of expanded mCAs with COVID-19 severity. ***a.*** *Association of expanded mCAs with COVID-19 Hospitalization across the UKB and FinnGen determined by logistic regression. Error bars show the 95% confidence interval for estimates. Bonferroni correction was used to determine the level of statistical significance. Individuals with known hematologic cancer at time of or prior to blood draw for genotyping were excluded. Analyses are adjusted for age, age^2, sex, ever smoking status, and principal components of ancestry.* ***b.*** *Visualization of the diverse range of expanded autosomal mCAs detected across the genome among individuals hospitalized with COVID-19 in the UK Biobank. Each point represents one mCA carried by a case, with the x-axis as the chromosome, y-axis as the mCA size in mega-bases of DNA (MB).* ***c.*** *Proportion of expanded autosomal mCAs in each category of COVID-19 outcomes for the CUB COVID-19 cohort, defined using the WHO COVID-19 scale (n=871 participants). 95% binomial proportion confidence intervals are shown. The table below the bar chart shows the counts of expanded autosomal mCA carriers and*

non-carriers in each outcome category. In CUB, the adjusted association between expanded autosomal mCAs and these ordinal COVID-19 outcomes is evaluated by ordinal regression and has OR of 1.52 (CI 95% 1.04 to 2.21, P= 0.031, two-tailed). MGBB = Mass General Brigham Biobank, UKB = UK Biobank, MB=megabase, CNN-LOH = copy number neutral loss of heterozygosity, CUB = Columbia University Biobank, WHO = World Health Organization

A.



B.

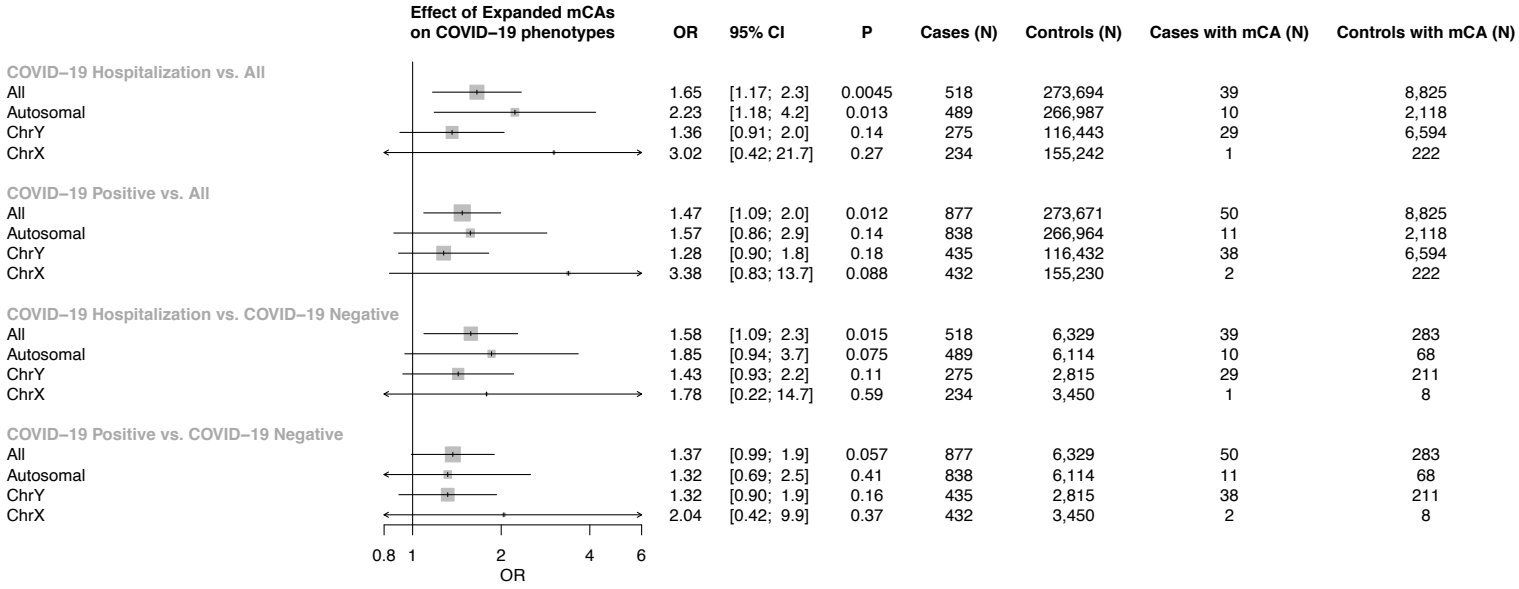


Figure 3.6.11: Associations of expanded mCAs in the UK Biobank with A. COVID-19 hospitalization across different adjustment models, and B. different COVID-19 phenotypes in a fully adjusted model. Adjustment models include 1) an unadjusted model, 2) a sparsely adjusted model which adjusts for age, age2, sex, smoking status, and

principal components of ancestry, and 3) a fully adjusted model which additionally adjusts for Townsend deprivation index, BMI, and the following comorbidities: Asthma, COPD, CAD, T2D, any cancer, and HTN. mCA = mosaic chromosomal alterations, COPD = chronic obstructive pulmonary disease, CAD = coronary artery disease, T2D = type 2 diabetes mellitus.

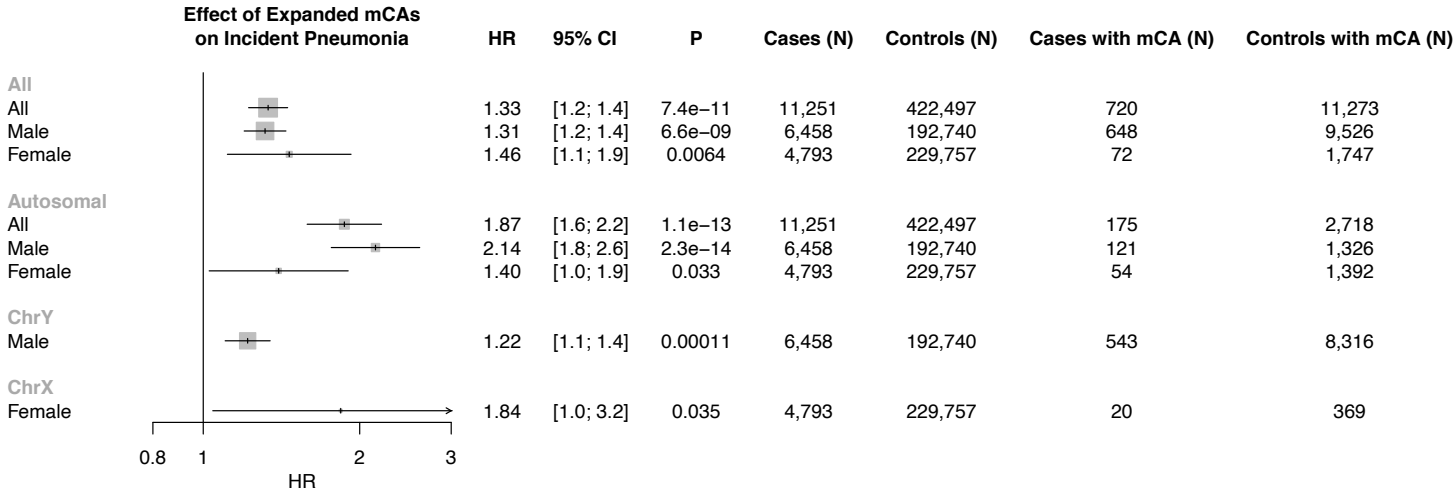


Figure 3.6.12: Association of expanded mCAs with incident pneumonia by sex in the UKB. *Individuals with known hematologic cancer at time of or prior to blood draw for genotyping were excluded. Analyses are adjusted for age, age^2, sex, ever smoking status, and principal components of ancestry.*

Discussion:

Across four geographically distinct biobanks comprising 767,891 individuals without known hematologic malignancy, clonal hematopoiesis (CH) represented by expanded mCAs is increasingly prevalent with age but not readily detectable by conventional blood tests. In addition to strongly predicting future risk of hematologic malignancy, expanded mCAs were also associated with risk for diverse incident infections, particularly sepsis and respiratory infections. These findings were robust across age, sex, tobacco smoking, and were strongest among those who develop cancer. Consistent with these observations, expanded mCAs were also associated with increased odds for COVID-19 hospitalization.

These results support several conclusions. First, mCA-driven CH is a potential risk factor for infection. Recent work showed that CH with myeloid malignancy driver mutations, also referred to as 'clonal hematopoiesis of indeterminate potential' (CHIP), predisposes to myeloid malignancy and coronary artery disease[5 6 19 20 64]. Meanwhile, CH with larger clonal chromosomal rearrangements (i.e., mCAs) predisposes primarily to lymphoid malignancy but not coronary artery disease[2 9 10 61 62]. Our observations suggest CH defined by the presence of mCAs is a risk factor for infection. Since the relationship between mCAs and infection risk was not substantially attenuated when adjusting for leukocyte or lymphocyte counts at baseline visit, the impact of mCAs on infection risk likely acts through mechanisms independent of the impact of CH on cell counts. For example, as mCAs alter gene dosage (e.g., via duplications and deletions) and remove allelic heterogeneity (e.g., copy neutral loss-of-heterozygosity events) in leukocytes, potential impacts on the differentiation, function, and survival of leukocytes are mechanisms that could lead to altered infection risk. In particular, many of the mCA variants are the same lesions found in chronic lymphocytic leukemia, a condition in which lymphocyte differentiation and function is altered promoting infection risk[65-68]. Therefore, molecular changes in leukocytes that promote clonal expansion may occur at the expense of reduced ability to combat infection.

Second, the infectious disease risk associated with mCAs is exacerbated in the setting of cancer. It is well-established that mCAs in blood-derived DNA increase risk for hematologic cancer[2 9 10]. Furthermore, recent evidence suggests an association between

mCAs detected in blood-derived DNA and increased risk of select solid tumor[26 60 69]. Our analysis identified an interaction between mCAs and prior cancer that amplified sepsis and pneumonia risk. Importantly, this interaction was restricted to individuals with solid cancers, not antecedent blood cancer. While this observation could be partially due to synergistic immunosuppressive side effects of cancer therapies[70], the observed associations persisted despite adjustment for these treatments. Alternatively, abnormal regulation of immune inflammatory pathways that release cytokines and inflammatory cells may create chronic states of inflammation in individuals with mCAs[71 72]. Surveillance for expanded mCA clones, particularly among those who develop solid cancer, may help identify individuals at high risk for infection that could benefit from targeted interventions.

Third, our findings could have particular relevance for the ongoing COVID-19 pandemic. We observed that mCAs are associated with elevated risk for COVID-19 hospitalization, with greater than two-fold risk linked to expanded autosomal mCAs. Maladaptive immune responses, particularly in leukocytes, increase risk for severe COVID-19 infections[73-76]. Awareness of COVID-19 risk associated with mCAs may help with the prioritization of emerging prophylactic treatments and initial vaccination programs.

This analysis of mCAs and infection had some limitations. Our study only measures mCAs at one time point for each participant. While our sampled mCA time point is likely correlated with CH at time of infection, CH dynamically changes over time potentially leading to differences in cellular fraction or additional undetected events that were

acquired prior to infection. Additionally, despite the robust adjustment and sensitivity analyses performed in our statistical analysis, including adjustment for chemotherapy, bone marrow transplant, radiation, and other features associated with poor cancer prognosis (neutropenia, aplastic anemia, decreased white blood cell count), we cannot completely rule out the impact of residual confounding in our results from unknown or unmeasured sources.

In conclusion, we report evidence for increased susceptibility to a spectrum of infectious diseases in individuals carrying mCAs in a detectable fraction of leukocytes particularly when cancer is concurrently present. The impacts of mCA on infection risk are systemic, with increased susceptibility to infection observed for a variety of organ systems, including severe COVID-19 presentations.

Chapter 4: Inherited genetic basis of somatic variation

Inherited germline genetic risk factors can predispose to somatic variation. Germline genetic variants have been previously associated with clonal hematopoiesis, either by somatic mosaicism of SNVs and indels through CHIP[77] or by large scale chromosomal rearrangements through mCAs[2], in individuals of European ancestry, and identified variants at a single locus, *TERT*, that associates with clonal hematopoiesis. Here, using the TOPMed WGS, we have not only replicated this finding but also identified several additional genome-wide significant loci across a multi-ethnic cohort, including near the *TET2* and *KPNA4/TRIM59* genes[78]. Furthermore, using the UK Biobank genotype data, we have identified 63 genome-wide significant loci associated with expanded mCA clones. Further understanding the germline genetic risk factors for somatic variants contributing to clonal hematopoiesis (ie: CHIP and mCAs) may suggest therapeutic targets which can modify the progression of somatic variants to disease.

The work in this chapter has been published across multiple papers as[15 22]:

Bick AG, Weinstock JS, Nandakumar SK, Fulko CP, Bao EL, **Zekavat SM**, et al. Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature* 2020;586(7831):763-68.

Zekavat SM, Lin SH, Bick AG, et al. Hematopoietic mosaic chromosomal alterations increase the risk for diverse types of infection. *Nat Med* 2021;27(6):1012-24. doi: 10.1038/s41591-021-01371-0 [published Online First: 2021/06/09]

Please refer to these papers for additional details on methods, cohorts, and other supplementary results.

Chapter 4.1: Genome-wide association of CHIP

Given the distinct association of clonal hematopoiesis with known leukemogenic mutations (i.e., CHIP) with both cancer and atherosclerotic cardiovascular disease, we sought to discover germline genetic variations conferring increased risk for CHIP acquisition.

Methods:

GWAS: Single variant association for each variant in TOPMed WGS Freeze 8 with MAF > 0.1% and MAC > 20 was performed with SAIGE[79], and analysis was performed using the TOPMed Encore analysis server (https://encore.sph.umich.edu). CHIP driver status was dichotomized into a case-control phenotype based on the presence of at least one driver mutation. Prior to running single variant association tests, a logistic mixed model was fit using the lme4 R package[80] to estimate the probability of the CHIP case control status conditional on a spline transformation of the centered age, genotype inferred sex, and cohort. The cohort was included as a random intercept which represents study specific contributions to the log-odds of CHIP at the mean sample age. Age was modeled with a spline to capture the non-linearity of the relationship between age and CHIP. This model was chosen over comparable models based on its AIC. Combining the age, inferred sex, and study into a single quantity aided the convergence of SAIGE compared to the inclusion of these terms separately. The first 10 principal components were also included as covariates.

Given that CHIP is unlikely to manifest in younger individuals, these individuals are effectively censored in our analysis set – that is, a young individual that does not presently have CHIP may still develop CHIP in the future. To avoid the power loss associated with misclassification of controls, we pruned these individuals from our analysis set. The single variant association analysis was run on a pruned set of samples that excluded those which had less than a 1% probability CHIP as estimated by the aforementioned model. This excluded 21,712 samples leading to a final analysis set of 65,405 which was used for downstream association analyses.

Transcriptome-wide association analyses using UTMOST: Multi-tissue gene expression and eQTL data were retrieved from the Genotype-Tissue Expression (GTEx) project (https://www.gtexportal.org). We applied the unified test for molecular signatures (UTMOST)[81] to perform cross-tissue transcriptome-wide association analysis for CHIP. We used cross-tissue gene expression imputation models trained from 44 tissues in GTEx. Gene-level association meta-analysis was performed using the generalized Berk-Jones test implemented in UTMOST (https://github.com/Joker-Jerome/UTMOST). Statistical significance was determined using a Bonferroni corrected P-value cut-off of 2.9×10^{-6}.

MESA RNA-Sequencing and Analysis: RNA-Sequencing was performed on peripheral blood mononuclear cells in MESA[82]. Alignment to the GRCh38 reference genome was done using STAR 2.5.3a[83] and gene quantification and quality control was performed using RNA-SeQC 1.19[33]. Annotation was performed using GENCODE26. For RNA-SeQC, isoforms were collapsed into a single transcript per gene using the

procedure described at https://github.com/broadinstitute/gtex-pipeline/blob/master/gene_model/. Samples that failed the RNA-Seq QC, fingerprinting, or expression-based sex check were filtered out. Further details on the RNASeq pipeline are provided here: https://www.nhlbiwgs.org/sites/default/files/TOPMed_RNAseq_pipeline_COREyr2.pdf Analysis was performed among 247 African Americans from Exam 1 who also had Exam 1 CHIP calls available. Transcript expression was converted to TPM units (transcripts per million) and log2-transformed for analysis. Analysis of rs79901204 with Tet2 expression was performed using a linear mixed model adjusting for age at blood draw, sex, PC1-10 of population stratification from the WGS data, sequencing batch, and kinship relatedness matrix.

Results:

We performed a single variant genome-wide association analysis in a subset of 65,405 individuals (3,831 CHIP driver cases). The trait heritability explained by the analysis with LD score-regression was 3.6%.

We replicated the lead variant of the single locus previously associated at genome wide significance with clonal hematopoiesis (defined based on somatic mosaicism of SNVs and indels)[77], rs34002450 (OR 1.2, $p=2.0 \times 10^{-13}$). rs34002450 is in strong LD ($r^2=0.55$) with our lead variant at this locus rs7705526, a common variant (MAF 0.29) in the 5th intron of *TERT*, which encodes telomere enzyme reverse transcriptase. In TOPMed, carriers of the rs34002450 A (minor) allele have a 1.3-fold risk of developing CHIP ($p=8.4 \times 10^{-24}$). This variant was previously significantly associated with increased

leukocyte telomere length[84]. This variant was also associated with myeloproliferative neoplasms[85] and clonal chromosomal mosaicism[2]. In a phenome-wide association analysis (PheWAS) of rs34002450 in UK Biobank, we identified significant increased risk of MPN ($p=2.6 \times 10^{-13}$), uterine leiomyoma ($p=3.2 \times 10^{-9}$), brain cancer ($p=3.6 \times 10^{-8}$) and a decreased risk of Seborrheic keratosis ($p=1.4 \times 10^{-7}$).

We performed a conditional analysis of the 14 other genome-wide significant SNPs at the TERT locus, conditioning on the lead SNP, to see if there were any additional signals that were independent of rs7705526. We identified a second intronic TERT variant rs13167280 (MAF 0.11, $r^2=0.2$ with rs7705526) that independently associates with CHIP status (OR 1.3, $p=6.1 \times 10^{-10}$; conditional OR: 1.1, $p=4.7 \times 10^{-4}$).

In the TOPMed single-variant association analysis, we additionally identified 2 other novel genome-wide significant genetic loci, including one locus on chromosome 3 in an intergenic region spanning *KPNA4*/*TRIM59* and one locus on chromosome 4 near *TET2* (**Figure 4.1.1**).

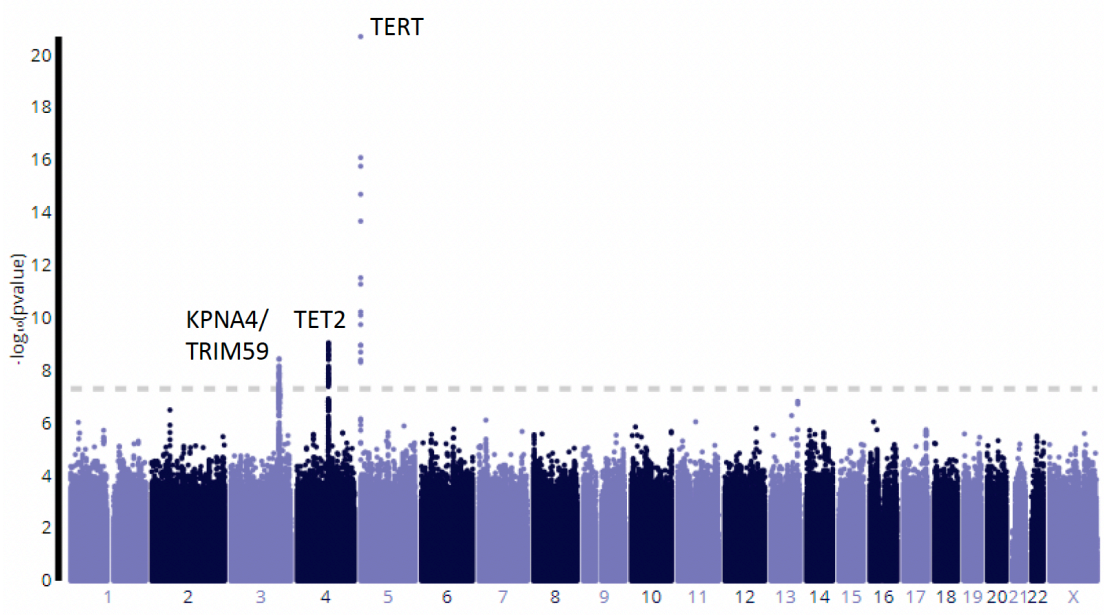


Figure 4.1.1: Genetic determinants of CHIP. Single variant genetic association analyses of CHIP identified 3 genome-wide significant loci.

rs1210060191 is a common variant (MAF 0.54) in a locus with an association signal that spans a 300kb region that includes *KPNA4, TRIM59, IFT80, and SMC4.* The lead variant is a 1 bp intronic deletion in *TRIM59*. Carriers of the del(T) allele have a 1.16-fold increased risk of CHIP ($p=5.3x10^{-10}$) Variants in LD with this variant have been identified as associated with MPN[85]. No other significant phenotype associations were noted in UK Biobank PheWAS analyses.

rs144418061 is an African ancestry specific variant (MAF 0.035 in African Ancestry samples, not present in non-African-ancestry samples) in an intergenic region near *TET2.* Carriers of the A allele have a 2.4-fold increased risk for CHIP ($p=4.0x10^{-9}$). We replicated this association in a distinct set of 570 TOPMed CHIP cases and 8,819 TOPMed controls (OR: 2.1, p=0.026). The association is equally robust for *DNMT3A* CHIP, *TET2* CHIP and *ASXL1* CHIP, suggesting that the germline variant does not specifically predispose to *TET2* CHIP. Although other variants in the vicinity of *TET2*

have been associated with MPN[85], this variant has not been previously identified as associated with any traits in the literature likely due to the under-representation of African ancestry genomes in existing association studies.

We performed a transcriptome-wide association analysis using UTMOST[81] to quantify the relationship between changes in gene expression and genetic predisposition to CHIP. This approach identified the Chr3 *KPNA4/TRIM59* locus and six additional loci including: *AHRR*, *ASL*, *KREMN2*, *LEAP2*, *JSRP1*, *RASEF* (**Figure 4.1.2-3).** *AHRR* directs hematopoietic progenitor cell expansion and differentiation[86].

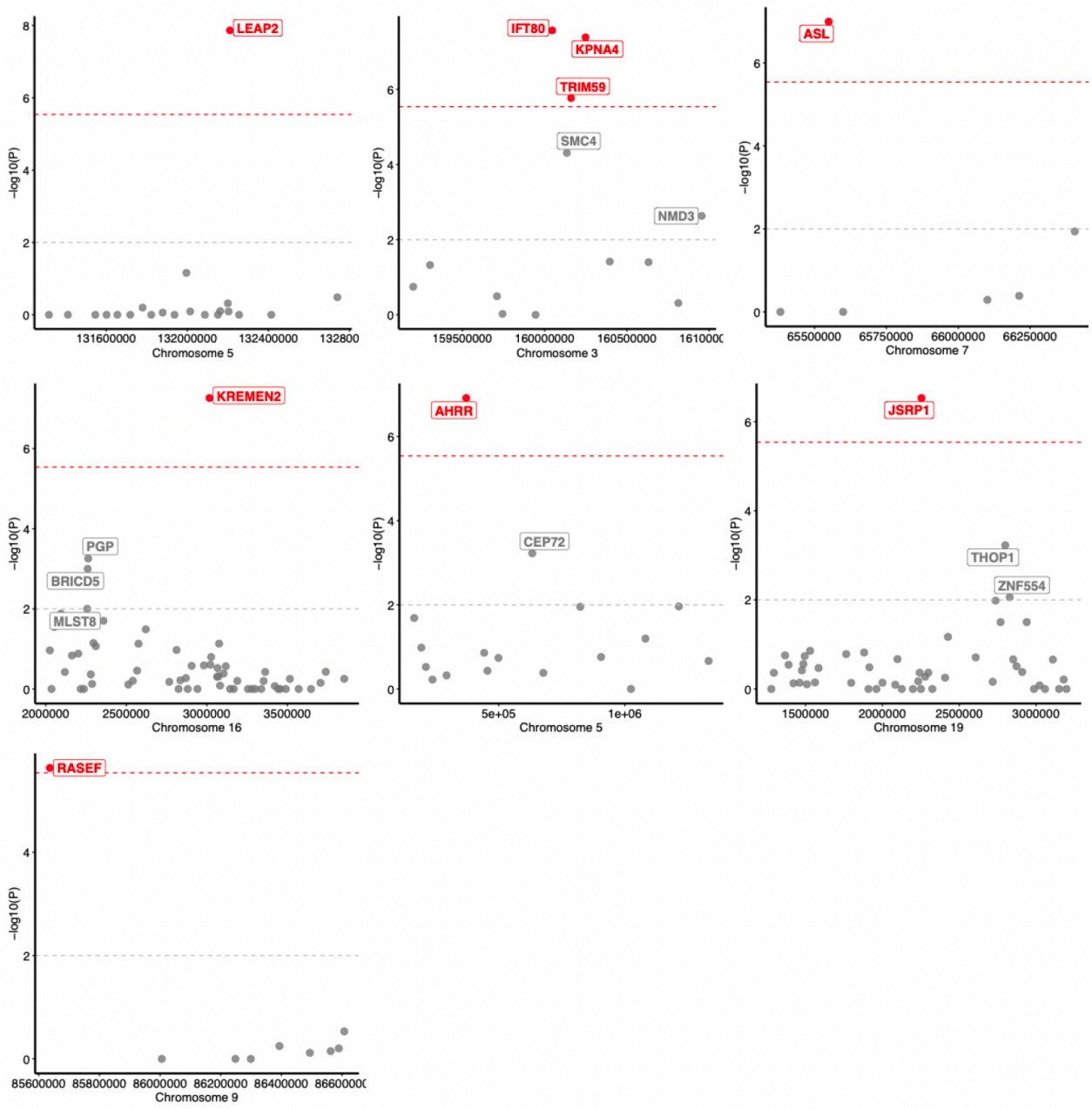


Figure 4.1.2: UTMOST[81] combined CHIP TWAS results across 48 tissues identified 7 significant loci ($P<2.9x10^{-6}$).

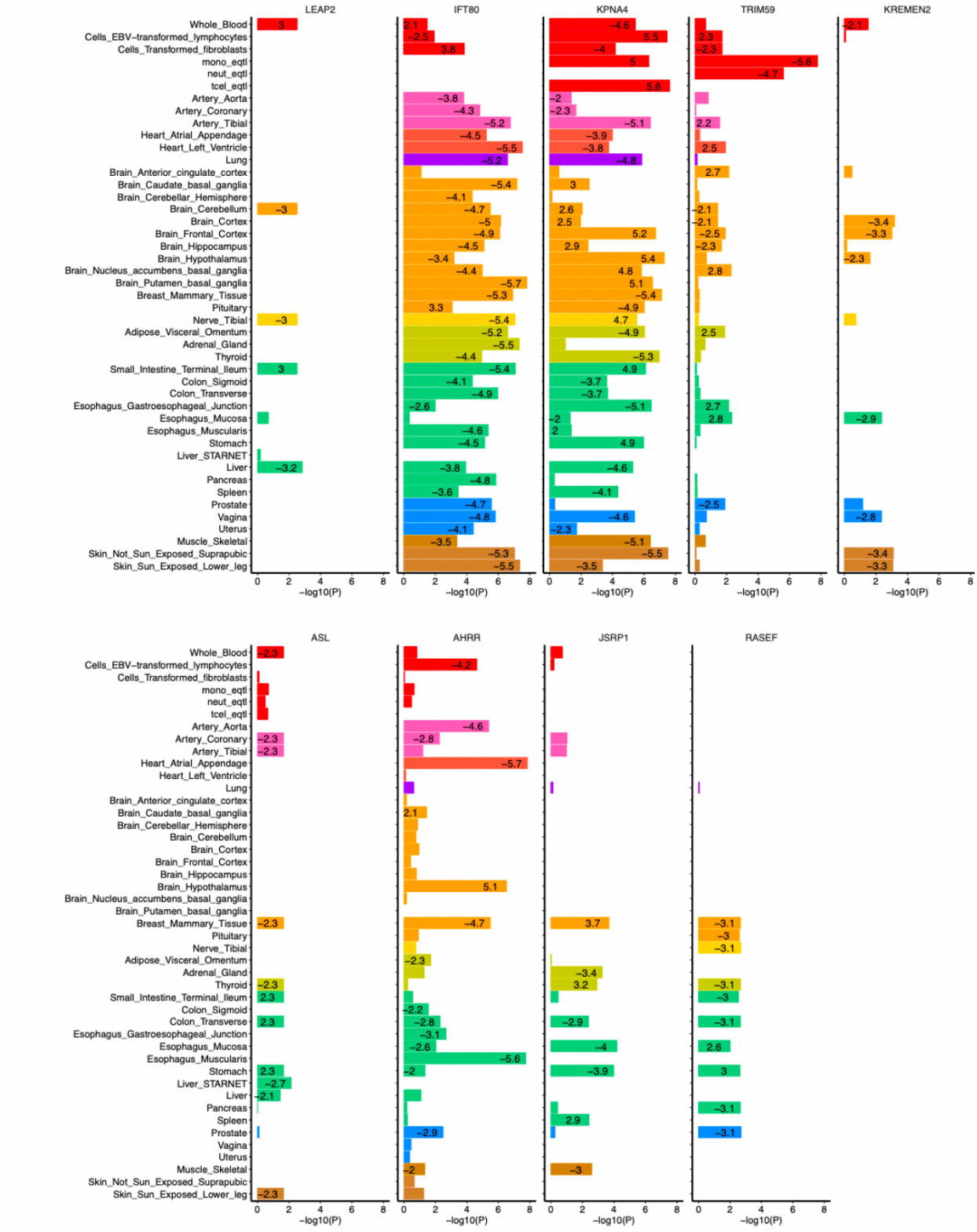


Figure 4.1.3: Tissue-specific results from the top 9 overall UTMOST-significant genes.

eQTL z-scores for associations with P<0.05 are displayed in each bar.

We bioinformatically and experimentally (in collaboration with Joshua Weinstock et al.) characterized the mechanism by which the non-coding African American-specific variant at the *TET2* locus influenced risk for CHIP. First, iterative conditional analysis at the locus suggested that there was most likely only a single causal variant. Fine-mapping prioritized 25 variants in the credible set (>99% posterior probability), none of which overlaps the coding sequence or promoter of a protein-coding gene. We hypothesized that the causal variant affects an enhancer for *TET2* in hematopoietic stem cells, because heterozygous *Tet2* knockout in mice increases the self-renewal of hematopoietic stem cells *in vivo* and recapitulates the clonal expansion observed in humans with somatic mutations in *TET2*[5]. Accordingly, we used the Activity-by-Contact (ABC) model to predict which noncoding elements act as enhancers in CD34+ hematopoietic stem and progenitor cells (HSPC, see Methods in Bick et al. Nature 2020[27]). Only a single variant (rs79901204) in this credible set overlapped an element predicted to regulate any gene, and that element was indeed predicted to regulate *TET2* expression. (**Figure 4.1.4-a**). The T risk allele disrupts a consensus GATA/E-Box motif, likely resulting in reduced binding of the activating transcription factor complex GATA1/GATA2 (**Figure 4.1.4-b,c**). To test whether this variant affects enhancer activity, we tested a 600 base pair region containing the regulatory element using a plasmid-based luciferase enhancer assay in hematopoietic cells. The reference sequence activated luciferase expression by 40-fold (versus control constructs with no enhancer sequence), while the T risk allele activated expression by only 10-fold (**Figure 4.1.4-d**). Lastly, among a subset of 247 African American individuals with whole blood RNAseq, 16 of whom were heterozygotes for rs79901204 and one who was a homozygote, the T risk allele led to a dose-dependent decrease in

whole blood *TET2* expression (Beta: -0.27, SE: 0.11, p=0.012, **Figure 4.1.4-e**). Together, these results suggest that the T risk allele acts to decrease the activity of this enhancer, which in turn reduces expression of *TET2*.

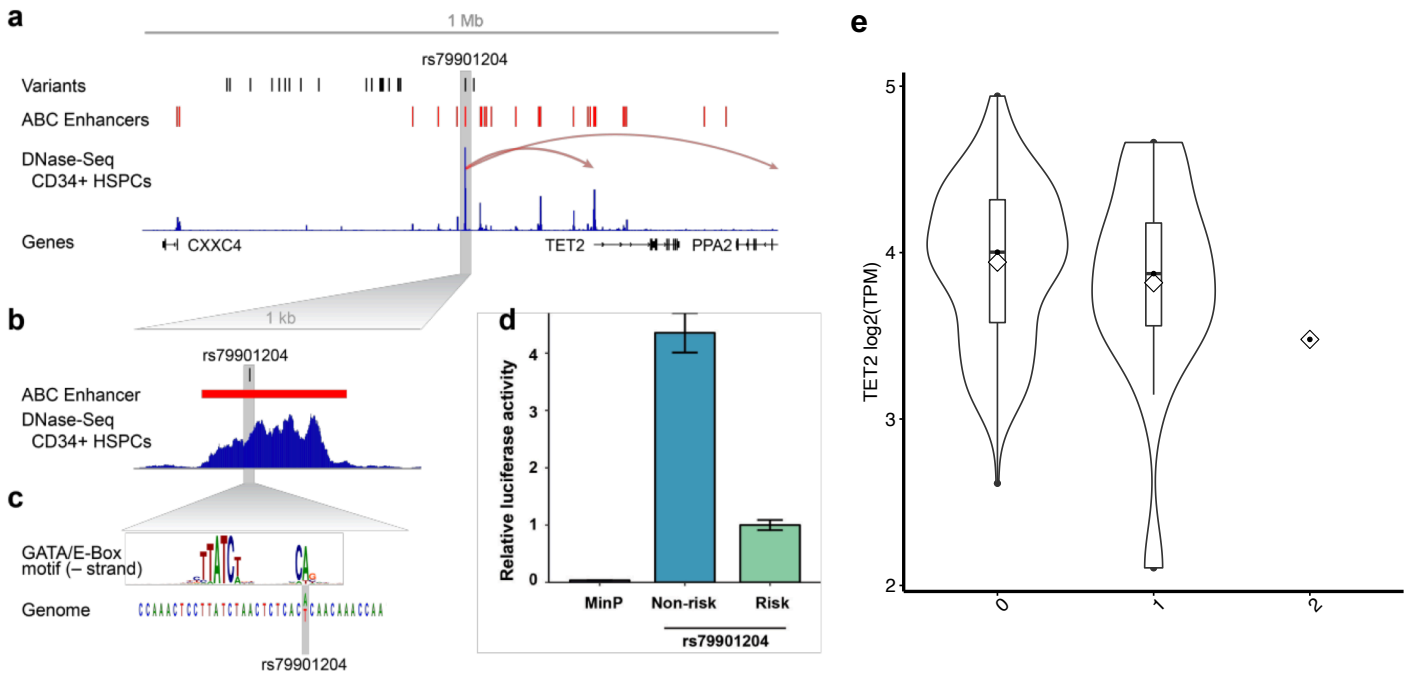


Figure 4.1.4: African ancestry specific TET2 locus risk variant disrupts hematopoietic stem cell TET2 enhancer. *a. the TET2 locus with fine-mapped risk variants, Activity-by-Contact (ABC) hematopoietic stem cell (HSPC) enhancers, Dnase-Seq CD34+ HSPC and RefSeq genes. ABC model predicts that rs79901204 disrupts a TET2 enhancer resulting in decreased TET2 expression. b. expanded view of TET2 enhancer element. c. rs79901204 disrupts a GATA motif/E-Box motif. d. luciferase assay in CD34+ primary cells demonstrates four-fold attenuation of enhancer activity by the rs79901204 risk allele relative to the non-risk allele. e. rs79901204 is associated with decreased TET2 expression in peripheral blood RNA-Seq (p=0.012).*

Discussion

In summary, our work highlights multiple mechanisms through which germline genetic variation can shape somatic variation in hematopoietic stem cells. A set of the germline loci are associated with increased propensity to acquire mutations due to failure of genes that maintain genome integrity (e.g. *TERT*) and which have been implicated in stem cell

maintenance/self-renewal[85]. These loci are associated with acquisition of somatic mutations resulting in neoplasm in multiple tissues. Other germline loci are associated with increased hematopoietic stem cell self-renewal (e.g. *TET2*). While the *TET2* locus is associated with increased risk of acquiring any CHIP driver mutations, it is not associated with cancer outside of the hematopoietic stem cell compartment. Furthermore, our work underscores the benefits of studying genomes from individuals of diverse ancestries. The inclusion of a significant number of African Ancestry samples in TOPMed permitted the discovery of the *TET2* locus which was not present in other ancestries. Further inclusion of diverse individuals in genomic analyses is likely to highlight additional new biological pathways.

Important limitations of our study include reduced sensitivity for detecting CHIP with low allele fractions (VAF 2-5%) even with high-coverage whole genome sequencing. Ultrasensitive targeted sequencing can facilitate detection of such leukemogenic mutations at exceedingly low VAFs but the clinical consequences of this much more pervasive phenomenon as well as determinants of progression to CHIP is not well understood currently.

Overall, comprehensive simultaneous germline and somatic analyses of blood-derived whole genome sequence data demonstrates that germline variation influences the acquisition of somatic mutations in blood cells. Importantly, we anticipate that the TOPMed CHIP dataset defined here will be a valuable tool in establishing associations of CHIP with diverse heart, lung, blood and sleep traits.

Chapter 4.2: Genome-wide association of mCAs

To further understand the inherited germline genetic risk factors for expanded mCA clones, we performed a genome-wide association study (GWAS) of expanded mCA in the UK Biobank.

Methods:

Genome-wide association study (GWAS): GWAS was performed using Hail-0.2 software (https://hail.is/) on the Google cloud. Variants were filtered to high-quality imputed variants (INFO score >0.4), with minor allele frequency >0.005, and with Hardy-Weinberg Equilibrium $P \geq 1x10^{-10}$, as previously performed. A Wald-logistic regression model was used for analysis, adjusting for age, age^2, sex, ever smoking, PC1-10, and genotyping array. Significant, independent loci were identified using $P<5x10^{-8}$ and clumping in Plink-2.0 using an r^2 threshold of 0.1 across 1MB genomic windows using the 1000-Genomes Project European reference panel. An additive mLOY polygenic risk score was developed as such: $\sum_{i=1}^{63} Beta \times SNP_{ij}$, where $Beta$ is the weight for each of the 156 independent genome-wide significant variants previously identified in UKB males[87] and SNP_{ij} is the number of alleles (i.e., 0, 1, or 2) for SNP_i in female j in the UKB.

Cell-type enrichment analyses: We applied partitioned LD score regression using the LDSC software[88] to perform enrichment analysis using the expanded mCA GWAS summary statistics in combination with tissue-specific epigenetic and transcriptomic functionality annotations from GenoSkyline-Plus[89]. In addition to the baseline annotations for diverse genomic features as suggested in the LDSC user manual, we specifically examined the enrichment signals on two tiers of annotations of different

resolutions: GenoSkyline-Plus functionality scores of 7 broad tissue clusters (immune, brain, cardiovascular, muscle, gastrointestinal tract, epithelial, and others); and GenoSkyline-Plus functionality scores of 11 tissue and cell types within the immune cluster (listed in **Figure 4.2.1-D**).

Transcriptome-wide association and pathway enrichment analysis:

Transcriptome-wide association was performed using the expanded mCA GWAS summary statistics in combination with the UTMOST[90] whole blood model updated to GTEXv8 (N=670). Significant genes were identified using a Bonferroni cutoff of $P<0.05/15,625$ or $3.2x10^{-6}$. Pathway enrichment analyses was performed using genes with TWAS $P<0.001$ using the Elsevier Pathways through the EnrichR web server[91].

Results:

We identified 63 independent genome-wide significant loci associated with expanded mCAs ($r^2< 0.1$ across 1MB windows of the genome) (**Figure 4.2.1-A, Table 4.2.1**). Across the 63 germline variants, significant correlation was seen between different mCA categories (**Figure 4.2.2**), suggesting the presence of shared germline genetic variants predisposing to mCAs across the genome. Follow-up analyses using an additive polygenic risk score comprised of 156 independent genome-wide significant variants associated with mosaic loss-of-chromosome Y (mLOY) from males from a prior study in the UKB[87], found significant associations with expanded autosomal mCAs and expanded ChrX mCAs in females, further highlighting the shared germline contributors towards mCAs across the genome (**Figure 4.2.3**).

To further understand what tissues are most implicated in these loci, tissue enrichment analyses using GenoSkyline-Plus was performed. Significant enrichment was

identified in immune-specific epigenetic and transcriptomic functional regions of the genome ($P=7.1x10^{-9}$) (**Figure 4.2.1-B,C**). Further stratification of the immune category identified specific enrichment across CD4+ T-cells, with suggestive evidence of enrichment ($P<0.05$) also present for CD14+ monocytes and the spleen (**Figure 4.2.1-D**).

Additionally, to further understand the transcriptomic effects of the germline inherited risk factors for expanded mCAs, TWAS was performed by combining the GWAS results with GTEXv8[92] whole blood expression quantitative trait loci (eQTLs) using the UTMOST[89] platform. The TWAS identified 62 significant genes whose expression levels in whole blood were significantly influenced by the germline variants in the expanded mCA GWAS (**Figure 4.2.1-E**). While gene enrichment analyses with the Elsevier Pathway Collection did not identify significantly associated pathways after multiple testing correction, top pathways were linked to DNA damage repair and lymphoid processes (**Figure 4.2.1-G**). In particular, the strongest enrichment was identified for immunoglobulin class-switch recombination via classical non-hmologous end-joining involving the MDC1 and ATM genes which are also enriched in the double strand DNA homologous repair pathway. Additionally, we observe an enrichment of genes associated with myeloblast -> neutrophil surface expression markers (involving CD164, FLT3, NCAM1). Moreover, other immune-related pathways included IL23A and IL17A provoked- cancer-association inflammation (involving IL12RB1 and NFKB1), which may provide some connection with our observation of an interaction between cancer and mCAs whereby individuals with cancer and mCAs have a stronger risk of infection compared to individuals without cancer. Additionally, we observe an enrichment of genes associated with myocarditis (NCAM1, PDCD1LG2, F2, IL12RB1,

NFKB1), as well as those associated with the B-cell lineage (FLT3, IKZF1), as well as the T-cell lineage (TAL1, ATM, IKZF1, TCF12). The corresponding GWAS locus-zoom plots for some of these immune-related genes are shown in **Figure 4.2.1-H**.

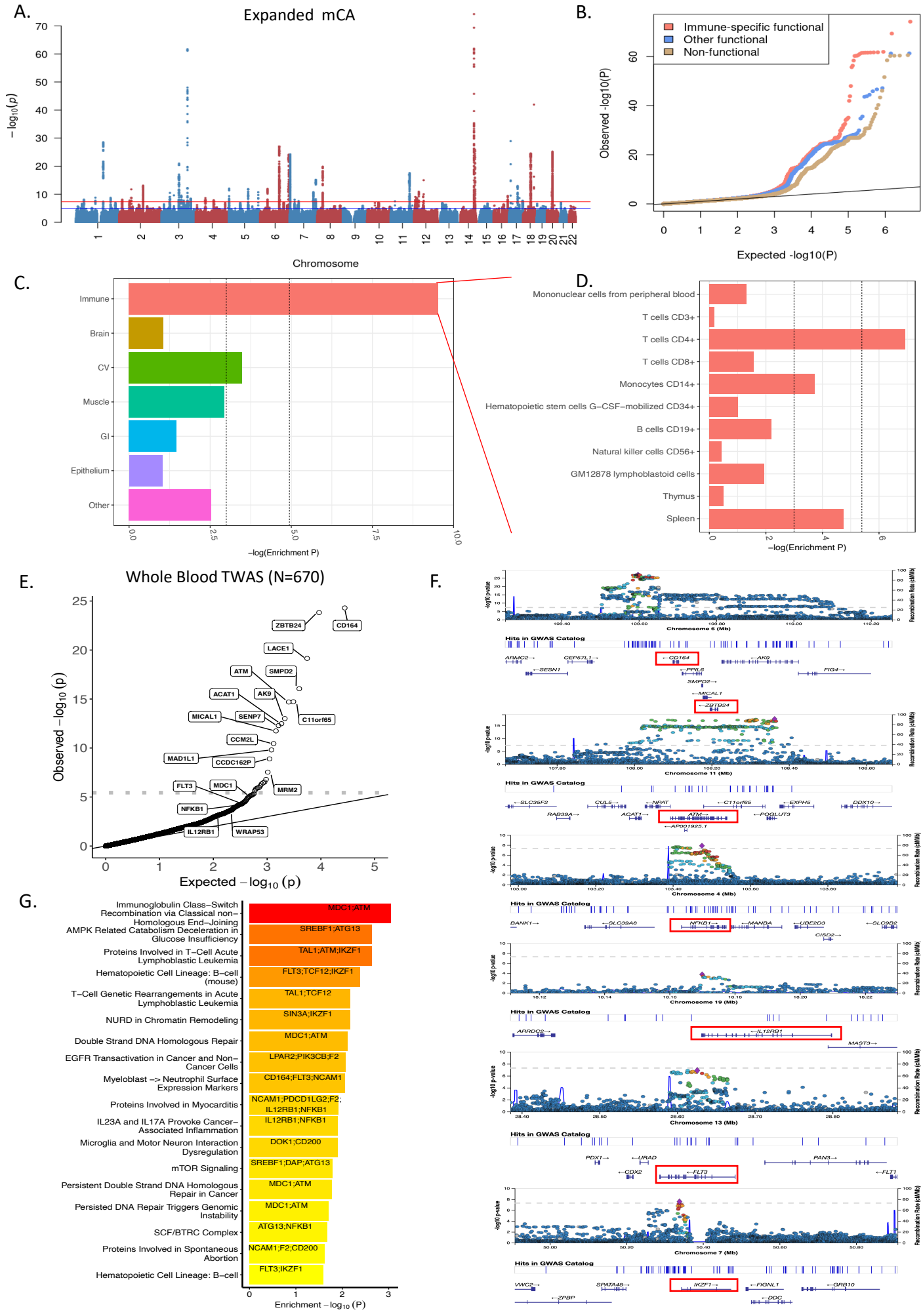

A.
Expanded mCA
Chromosome
B.
Immune-specific functional
Other functional
Non-functional
Observed -log10(P)
Expected -log10(P)
C.
Immune
Brain
CV
Muscle
GI
Epithelium
Other
–log(Enrichment P)
D.
Mononuclear cells from peripheral blood
T cells CD3+
T cells CD4+
T cells CD8+
Monocytes CD14+
Hematopoietic stem cells G–CSF–mobilized CD34+
B cells CD19+
Natural killer cells CD56+
GM12878 lymphoblastoid cells
Thymus
Spleen
–log(Enrichment P)
E.
Whole Blood TWAS (N=670)
Observed −log10 (p)
Expected −log10 (p)
ZBTB24
CD164
LACE1
ATM
SMPD2
ACAT1
AK9
MICAL1
SENP7
C11orf65
CCM2L
MAD1L1
CCDC162P
FLT3
MDC1
MRM2
NFKB1
IL12RB1
WRAP53
F.
Chromosome 6 (Mb)
Chromosome 11 (Mb)
Chromosome 4 (Mb)
Chromosome 19 (Mb)
Chromosome 13 (Mb)
Chromosome 7 (Mb)
Hits in GWAS Catalog
G.
Immunoglobulin Class–Switch Recombination via Classical non–Homologous End–Joining
AMPK Related Catabolism Deceleration in Glucose Insufficiency
Proteins Involved in T–Cell Acute Lymphoblastic Leukemia
Hematopoietic Cell Lineage: B–cell (mouse)
T–Cell Genetic Rearrangements in Acute Lymphoblastic Leukemia
NURD in Chromatin Remodeling
Double Strand DNA Homologous Repair
EGFR Transactivation in Cancer and Non–Cancer Cells
Myeloblast –> Neutrophil Surface Expression Markers
Proteins Involved in Myocarditis
IL23A and IL17A Provoke Cancer–Associated Inflammation
Microglia and Motor Neuron Interaction Dysregulation
mTOR Signaling
Persistent Double Strand DNA Homologous Repair in Cancer
Persisted DNA Repair Triggers Genomic Instability
SCF/BTRC Complex
Proteins Involved in Spontaneous Abortion
Hematopoietic Cell Lineage: B–cell
MDC1;ATM
SREBF1;ATG13
TAL1;ATM;IKZF1
FLT3;TCF12;IKZF1
TAL1;TCF12
SIN3A;IKZF1
MDC1;ATM
LPAR2;PIK3CB;F2
CD164;FLT3;NCAM1
NCAM1;PDCD1LG2;F2;IL12RB1;NFKB1
IL12RB1;NFKB1
DOK1;CD200
SREBF1;DAP;ATG13
MDC1;ATM
MDC1;ATM
ATG13;NFKB1
NCAM1;F2;CD200
FLT3;IKZF1
Enrichment −log10 (P)

Figure 4.2.1: Inherited risk factors for expanded mCAs: GWAS, Cell Type Enrichment, and TWAS. *A. GWAS for expanded mCA identified 63 independent loci. B. Quantile-quantile plot for the GWAS stratified by variants overlying 1) immune-specific functional regions, 2) other functional regions, and 3) non-functional regions as identified by annotations from GenoSkyline-Plus. C. cell-type enrichment results from the Expanded mCA GWAS across immune, brain, cardiovascular (CV), muscle, gastrointestinal (GI), epithelium, and other tissues as annotated using GenoSkyline-Plus. D. Zooming in to show the stratified enrichment by specific categories of immune cells and tissues. Across panels C. and D., the vertical dotted lines indicate (1)P=0.05 for suggestive enrichment, and (2) the Bonferroni-adjusted P-value for significant enrichment. E. Quantile-quantile plot of the whole blood TWAS of the expanded mCA GWAS using 670 samples from GTEXv8 shows enrichment across 62 genes. The horizontal dotted line reflects the Bonferroni-adjusted p-value for significance. Genes with TWAS $P<5x10^{-8}$ or those important in the pathway-enrichment analyses from panel G are labeled. G. Top results from pathway enrichment analysis of the TWAS results using the Elsevier Pathways. H. Highlighting the GWAS locus-zoom plots for some of the TWAS genes implicated in the top pathways from panel G. Red boxes highlight the gene(s) with strongest association in the TWAS analyses.*

Table 4.2.1: 63 independent genome-wide significant loci associated with expanded mCAs

locus	*REF*	*ALT (effect allele)*	*rsid*	*Gene*	*Consequence*	*AF*	*beta*	*P*
14:96180695	G	T	rs2887399	TCL1A	upstream_gene_variant	0.21	-0.36	6.28E-75
3:150014399	T	G	rs6440668	NA	regulatory_region_variant	0.84	-0.31	2.83E-62
18:60920854	C	T	rs17758695	BCL2	intron_variant	0.03	-0.76	1.11E-42
17:7571752	T	G	rs78378222	TP53	3_prime_UTR_variant	0.01	0.61	1.17E-29
1:156200671	T	C	rs2842870	PMF1-BGLAP	intron_variant	0.36	0.17	3.31E-29
6:109597641	T	C	rs6925716	NA	intergenic_variant	0.52	-0.16	1.01E-27
14:101178715	C	G	rs72698720	NA	regulatory_region_variant	0.14	-0.24	1.29E-26
20:30431070	G	T	rs7266148	FOXS1	downstream_gene_variant	0.21	-0.19	6.28E-26
18:42078951	G	A	rs188050966	CTC-782O7.1	intron_variant	0.13	0.21	2.81E-25
6:164472121	G	A	rs2874705	NA	intergenic_variant	0.41	-0.15	8.05E-25
7:1919539	C	T	rs4721146	MAD1L1	intron_variant	0.40	0.15	8.38E-25
18:42161643	G	T	rs1849209	NA	intergenic_variant	0.77	0.17	1.15E-22
3:101267385	T	C	rs13062095	NA	intergenic_variant	0.34	0.14	1.10E-21
8:30285091	G	C	rs2979469	RBPMS	intron_variant	0.74	0.16	2.45E-20
11:108314362	A	C	rs4255510	C11orf65	intron_variant	0.41	0.13	3.43E-18
6:109799923	C	T	rs6911838	ZBTB24	intron_variant	0.34	-0.13	4.35E-16
7:149424769	T	C	rs57003278	KRBA1	intron_variant	0.18	-0.16	7.32E-16
12:54685880	C	T	rs35979828	RP11-968A15.8	intron_variant	0.07	0.21	1.00E-15
11:108149207	T	G	rs141379009	ATM	intron_variant	0.03	0.32	1.34E-15
18:42231958	A	G	rs2852752	NA	regulatory_region_variant	0.67	-0.12	3.33E-15
2:136925439	T	C	rs10193587	NA	intergenic_variant	0.23	0.13	8.04E-14
17:47780716	A	G	rs200689359	SLC35B1	intron_variant	0.04	-0.30	1.12E-13
6:109578530	G	A	rs4946952	C6orf183	intron_variant	0.80	-0.13	4.28E-13
5:1287194	G	A	rs2853677	TERT	intron_variant	0.58	-0.10	9.82E-13
6:42037628	C	G	rs12194781	TAF8	intron_variant	0.13	-0.16	1.22E-12
5:111061881	C	T	rs57201028	STARD4-AS1	intron_variant	0.07	0.18	1.56E-12
2:68962137	G	A	rs10048745	ARHGAP25	5_prime_UTR_variant	0.25	0.11	1.84E-12
6:41986273	A	C	rs4714550	RNU6-761P	upstream_gene_variant	0.75	0.12	3.38E-12
20:29428748	T	G	rs11905279	NA	intergenic_variant	0.17	-0.14	7.08E-12
14:96162418	G	A	rs78986913	TCL1B	downstream_gene_variant	0.04	-0.27	7.55E-12
5:169015479	G	A	rs116483731	SPDL1	missense_variant	0.01	-0.71	1.86E-11
12:26589770	G	A	rs16930705	ITPR2	intron_variant	0.07	0.18	2.60E-11
7:135312572	G	A	rs4073627	NUP205	intron_variant	0.13	-0.15	2.65E-11
3:48638801	C	T	rs62618742	UQCRC1	missense_variant	0.03	-0.30	3.36E-11
3:114574027	G	T	rs12695310	ZBTB20	intron_variant	0.53	-0.09	3.07E-10
16:81049800	T	C	rs12928638	CENPN	intron_variant	0.13	0.13	3.87E-10
18:42252408	C	T	rs138775024	RP11-456K23.1	downstream_gene_variant	0.04	0.22	4.04E-10

12:52306687	C	G	rs35960167	ACVRL1	intron_variant	0.45	0.09	5.82E-10
14:96153765	C	T	rs56111147	TCL1B	intron_variant	0.47	-0.09	7.96E-10
12:6493351	A	G	rs10849448	LTBR	5_prime_UTR_variant	0.75	0.11	9.07E-10
3:47087837	T	A	rs13063578	SETD2	intron_variant	0.40	0.09	1.08E-09
6:109596552	C	T	rs72940976	C6orf183	downstream_gene_variant	0.06	-0.19	1.43E-09
1:111208718	A	T	rs56795609	NA	intergenic_variant	0.17	-0.12	1.62E-09
1:198750522	G	A	rs10800586	NA	intergenic_variant	0.57	-0.09	1.82E-09
12:14595456	G	A	rs7299037	ATF7IP	intron_variant	0.51	0.09	1.85E-09
4:55408875	A	T	rs218264	NA	regulatory_region_variant	0.25	-0.10	2.30E-09
14:101191007	A	T	rs67022228	DLK1	upstream_gene_variant	0.27	-0.10	2.98E-09
3:168860010	G	A	rs2859868	MECOM	intron_variant	0.55	-0.09	3.66E-09
18:42041131	T	G	rs72899729	CTC-782O7.1	intron_variant	0.01	0.33	4.38E-09
1:44937451	G	A	rs11211005	RNF220	intron_variant	0.22	-0.10	7.64E-09
2:16627651	A	T	rs7580707	NA	intergenic_variant	0.74	-0.09	7.86E-09
7:27143757	T	C	rs2522828	HOXA2	upstream_gene_variant	0.91	-0.14	8.22E-09
4:103475444	G	T	rs4648011	NFKB1	intron_variant	0.58	0.08	1.36E-08
3:150238789	A	G	rs4390958	NA	intergenic_variant	0.64	-0.08	1.68E-08
2:58936057	A	T	rs10865307	LINC01122	intron_variant	0.58	0.08	1.76E-08
3:128336298	G	T	rs2492286	RPN1	downstream_gene_variant	0.15	-0.12	1.84E-08
17:76684970	G	A	rs7225707	CYTH1	intron_variant	0.55	0.08	2.03E-08
3:150255127	C	T	rs79022866	SERP1	downstream_gene_variant	0.06	0.16	2.62E-08
7:50338499	G	A	rs1993444	NA	intergenic_variant	0.33	-0.09	2.73E-08
11:24084913	C	T	rs72881160	NA	intergenic_variant	0.13	0.11	3.22E-08
8:30251002	C	A	rs2979484	RBPMS	intron_variant	0.29	0.09	3.26E-08
3:150021924	A	T	rs28582771	NA	regulatory_region_variant	0.30	-0.09	3.59E-08
8:30631471	C	T	rs113406715	PPP2CB	downstream_gene_variant	0.03	0.23	4.92E-08

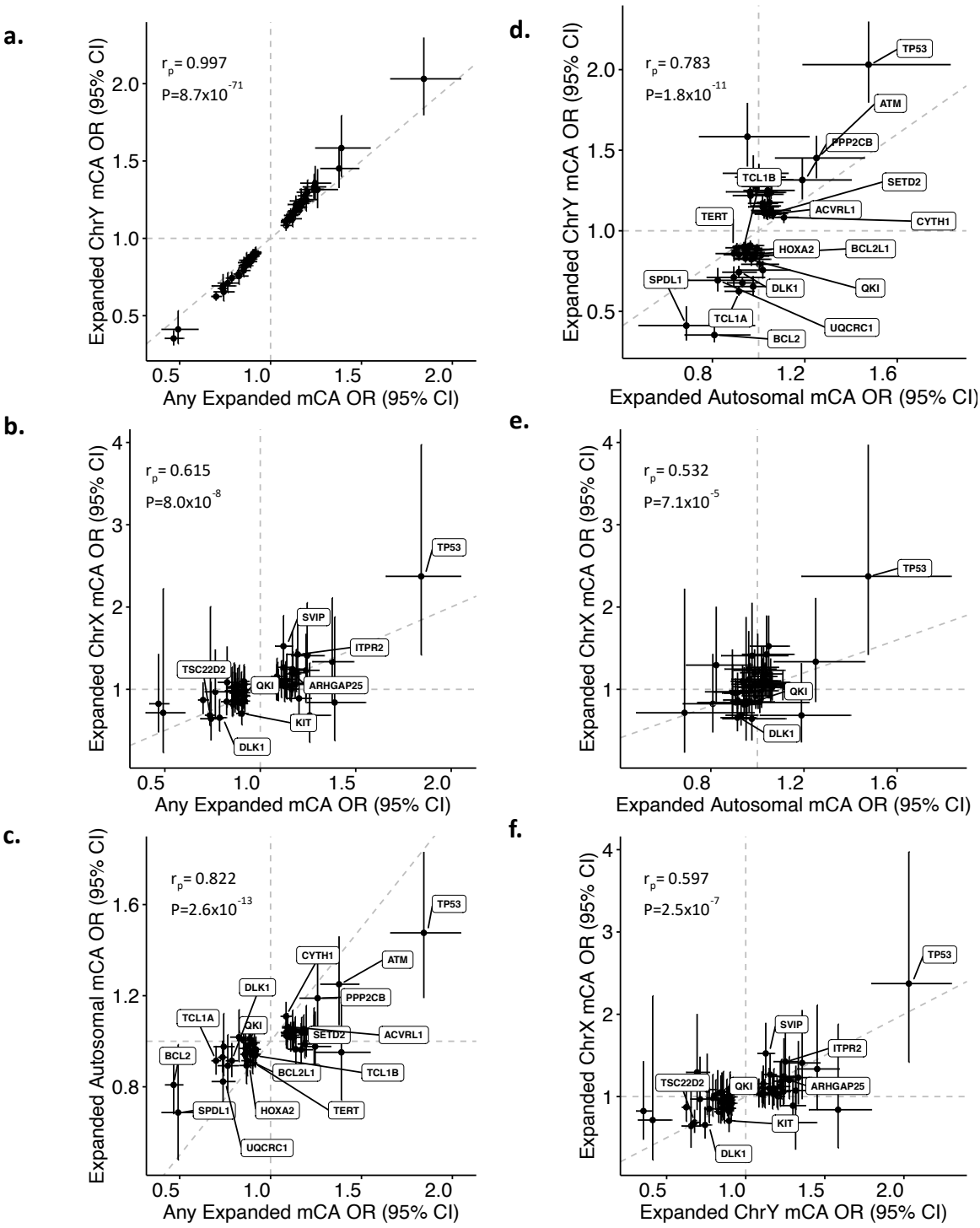


Figure 4.2.2: Correlated associations of 63 independent genome-wide significant variants associated with expanded mCAs (from Table 4.2.1) between different mCA categories (expanded autosomal mCAs, expanded ChrX mCAs, expanded ChrY mCAs) in the UKB. *Across all panels except for panel (a), the labeled genes represent genes attributed to variants that have P<0.05 across the mCA categories in both axes. mCA = mosaic chromosomal alterations, r_p = Pearson correlation*

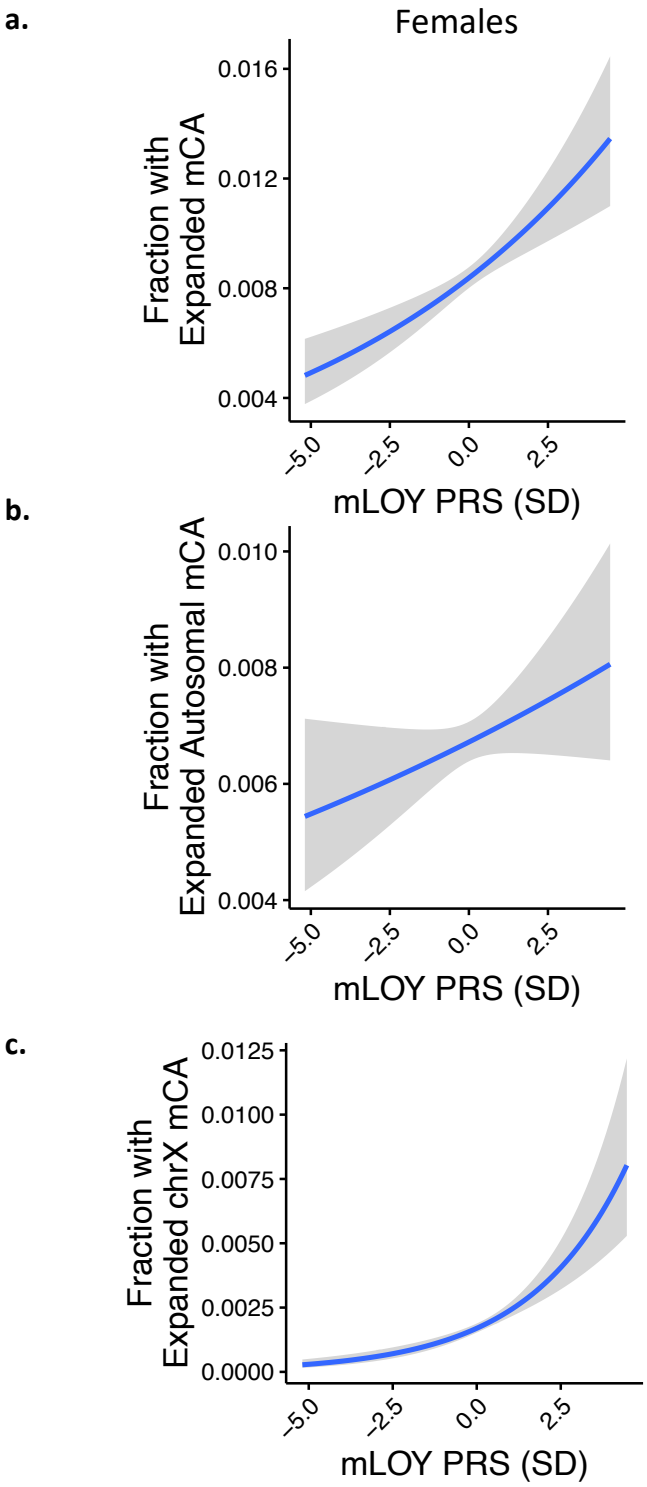


Figure 4.2.3: Association of a mLOY PRS consisting of 156 previously identified[87] independent genome-wide significant variants associated with mLOY, with different expanded mCA categories in UKB Females. *mLOY = mosaic Loss-of-chromosome Y, PRS = polygenic risk score.*

Discussion:

In summary, we explored the heritable basis for expanded mCAs by evaluating common germline genetic variation, and have identified significant enrichment of associating loci among immune cells, and pathways influencing leukemogenic potential, genomic instability, and cellular immunity. We identified 63 independent genome-wide significant loci linked to expanded mCA clones, and our ancillary analyses suggests that these loci are enriched in functional regions of immune cells, particularly CD4+ T-cells. Additionally, our TWAS results point to multiple pathways that promote expanded mCA development and point to genes involved broadly in hematopoiesis, DNA-damage repair pathways and genome instability, and the immune system as important contributors towards promoting expanded mCA development. Therapeutic drugs that modulate the identified germline risk factors for expanded mCAs may also protect against incident infection.

Chapter 5: Transcriptome-wide association of CHIP and mCAs

Transcriptomic analyses of CHIP, in particular of the *Tet2* gene, has previously been done within hematologic *Tet2* knockout mice models, identifying significant changes in expression among genes in inflammatory pathways (ex: among cytokines/chemokines and lysosomal function)[1]. These mice also develop larger atherosclerotic lesions[1]. Further transcriptomics *in humans* may identify changes in gene expression influenced by clonal hematopoiesis, thereby discovering biological pathways influenced by somatic CHIP variants in monocytes.

The earliest stages of atherosclerosis involve monocyte infiltration into vessel walls and differentiation into macrophages. Dr. Hongyu Zhao's lab previously determined that the inferred functional regulatory regions of the genome for monocytes (from the Roadmap Epigenomics Project) show 5-fold enrichment in CAD (5-fold enrichment, $P=1.5x10^{-5}$), AD (11-fold enrichment, $P=2.0x10^{-5}$), and AMD (3-fold enrichment, $P=9.9x10^{-4}$) genome-wide association studies[89], suggesting that monocytes play a significant and causal role in the pathogenesis of these diseases. Prior studies using mouse models have also suggested that the specific hematological cell type influencing atherosclerosis through CHIP mutations are monocytes, and that these mutations influence expression of genes involved with inflammation and phagocytosis central to monocyte-derived macrophages[5].

Previous work using mouse models suggests that loss of function of *Tet2* in myeloid-specific cells increases risk of atherosclerosis. *Ldlr* knockout mice that received bone marrow from mice that lacked *Tet2* in myeloid-lineage specific cells developed

larger atherosclerotic lesions[5]. Transcriptomics of cultured bone-marrow-derived macrophages from the *Tet2*-knockout mice show up-regulated expression of genes involved with cytokines, chemokines, and their receptors, and down-regulated expression of genes involved with lysosomal function, suggesting that these mutations influence monocyte adhesion, inflammatory signaling, and macrophage phagocytosis[5].

Interestingly, among the list of up-regulated genes in *Tet2*-knockout monocytes is *IL1b*[5]. Recent analyses of the CANTOS clinical trial have shown that CHIP carriers with somatic variants in *TET2* have over 4-fold higher improved response to canakinumab, an IL-1B antibody (HR=0.36, P=0.03)[8], compared to all individuals in the trial (HR=0.85, P=0.02)[7] with respect to major adverse cardiovascular events. These findings motivate further transcriptomic analyses of CHIP in *human* CD14+ monocytes, suggesting that resulting findings may implicate therapeutic strategies especially impactful among CHIP carriers to reduce disease risk.

Methods:

Here, I performed analysis of CHIP carrier state with individual gene expression from RNASeq of peripheral blood cells in 899 TOPMed individuals from the Multi-Ethnic Study of Atherosclerosis (MESA) Exam 1, and characterized enriched biological pathways associated with CHIP.

RNA-sequencing and quality control: RNA-sequencing was performed on peripheral blood mononuclear cells in MESA (PBMCs) using microarrays[82]. Alignment to the GRCh38 reference genome was done using STAR 2.5.3a[83] and gene quantification and quality control was performed using RNA-SeQC 1.19. Annotation was performed

using GENCODE2. For RNA-SeQC, isoforms were collapsed into a single transcript per gene using the procedure described at https://github.com/broadinstitute/gtex-pipeline/blob/master/gene_model/. Samples that failed the RNA-Seq QC, fingerprinting, or expression-based sex check were filtered out. Further details on the RNASeq pipeline are provided here: https://www.nhlbiwgs.org/sites/default/files/TOPMed_RNAseq_pipeline_COREyr2.pdf. Transcript expression data was converted from RPKM to TPM, low-expression transcripts with TPM<0.1 were excluded from analysis and the TPM data was then normalized using the TMM method followed by inverse rank normalization (as done in GTEx-v8). 23,017 transcripts (14,599 protein-coding) are expressed in the PBMCs from MESA (with TPM>0.1).

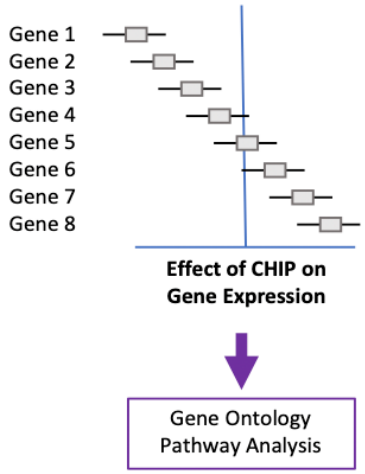


Figure 5.1.1: Schematic of transcriptomics analysis design.

Association of CHIP with gene expression: I first associated CHIP carrier status with gene expression across 14,599 protein-coding genes from 899 participants in the MESA cohort (aged 55-94yr) using a mixed model approach taking into account a

kinship relatedness matrix (used the lmekin package in R-3.5) (**Figure 5.1.1**). In these analyses, I adjusted for age, sex, smoking status, the first 10 genotyping principal components of ancestry, and the first 30 PEER factors (probabilistic estimation of expression residuals)[93] to account for complex non-genetic factors in gene expression levels as previously done in GTEx (https://gtexportal.org/home/documentationPage). Significant transcripts with false discovery rate (FDR)<0.05 were labeled.

Association of CHIP with cell differential: Cibersort[94], a cell-type deconvolution method, was applied to the MESA gene expression data to provide an estimation of the abundance of cell types in the mixed PBMC data using gene expression data. Association of CHIP with percent of each cell type in each sample was performed to understand how CHIP associates with cell abundance.

Association of CHIP with monocyte-specific gene expression: Given the strong biological prior that CHIP influences monocyte function[5], additional analyses was also performed using an interaction term for each gene (CHIP x monocyte %) to understand monocyte-specific associations with CHIP.

Results:

Association of CHIP with RNA expression levels across 14,599 protein-coding transcripts led to one FDR-significant gene (**Figure 5.1**), PSMD1, whose expression in peripheral blood mononuclear cells was lower in CHIP cases (beta = 0.30 SD, $P=3.01 \times 10^{-6}$). PSMD1 encodes the Proteasome 26S Subunit, Non-ATPase 1 protein, a component of the proteasome which plays a key role in removing misfolded or damaged

proteins during cellular processes including cell cycle progression, apoptosis, and DNA damage repair.

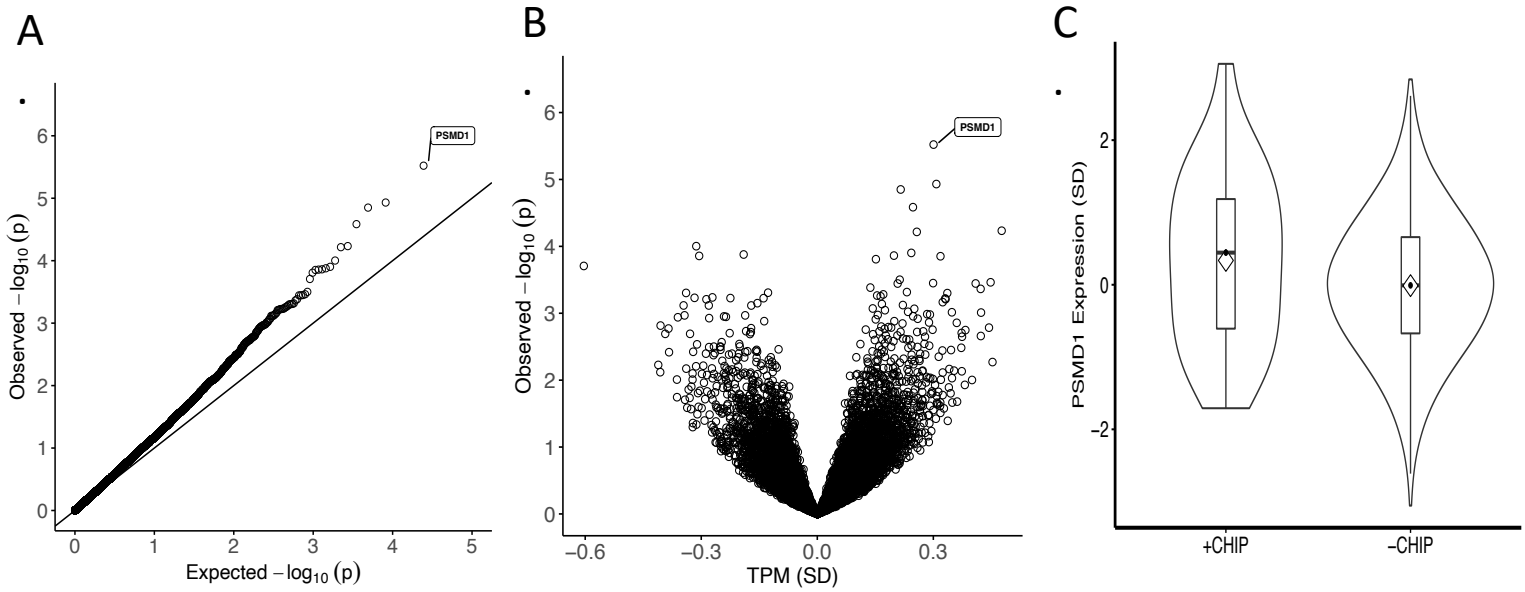


Figure 5.1: Transcriptome-wide association of CHIP in the MESA cohort (30 CHIP carriers, 853 controls)

	Homo sapiens (REF)	upload_1 (Hierarchy NEW!)					
Reactome pathways	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
Removal of the Flap Intermediate from the C-strand	10	3	.13	22.34	+	5.97E-04	2.90E-02
↳Processive synthesis on the C-strand of the telomere	11	3	.15	20.31	+	7.52E-04	2.91E-02
↳Telomere C-strand (Lagging Strand) Synthesis	24	5	.32	15.51	+	3.61E-05	5.88E-03
↳Extension of Telomeres	30	5	.40	12.41	+	9.24E-05	7.28E-03
↳Chromosome Maintenance	90	6	1.21	4.96	+	1.76E-03	4.83E-02
Loss of function of MECP2 in Rett syndrome	11	3	.15	20.31	+	7.52E-04	2.96E-02
↳Pervasive developmental disorders	11	3	.15	20.31	+	7.52E-04	3.01E-02
HSF1 activation	12	3	.16	18.62	+	9.31E-04	3.48E-02
↳Cellular responses to stress	548	19	7.36	2.58	+	2.18E-04	1.42E-02
PCNA-Dependent Long Patch Base Excision Repair	21	5	.28	17.73	+	2.06E-05	4.29E-03
↳Resolution of AP sites via the multiple-nucleotide patch replacement pathway	25	6	.34	17.87	+	2.86E-06	1.63E-03
↳Resolution of Abasic Sites (AP sites)	36	6	.48	12.41	+	1.80E-05	5.15E-03
↳Base Excision Repair	70	6	.94	6.38	+	5.16E-04	2.68E-02
Polymerase switching on the C-strand of the telomere	14	3	.19	15.96	+	1.36E-03	4.21E-02
Mismatch repair (MMR) directed by MSH2:MSH3 (MutSbeta)	14	3	.19	15.96	+	1.36E-03	4.15E-02
↳Mismatch Repair	15	3	.20	14.89	+	1.62E-03	4.51E-02
Removal of the Flap Intermediate	14	3	.19	15.96	+	1.36E-03	4.10E-02
↳Processive synthesis on the lagging strand	15	3	.20	14.89	+	1.62E-03	4.57E-02
↳Lagging Strand Synthesis	20	4	.27	14.89	+	2.60E-04	1.61E-02
↳DNA strand elongation	32	4	.43	9.31	+	1.27E-03	3.98E-02
↳Synthesis of DNA	118	9	1.58	5.68	+	5.02E-05	5.46E-03
↳S Phase	160	11	2.15	5.12	+	1.84E-05	4.68E-03
↳Cell Cycle, Mitotic	495	19	6.65	2.86	+	6.12E-05	6.08E-03
Mismatch repair (MMR) directed by MSH2:MSH6 (MutSalpha)	14	3	.19	15.96	+	1.36E-03	3.99E-02
Polymerase switching	14	3	.19	15.96	+	1.36E-03	3.94E-02
↳Leading Strand Synthesis	14	3	.19	15.96	+	1.36E-03	4.05E-02
Attenuation phase	14	3	.19	15.96	+	1.36E-03	3.89E-02
Gap-filling DNA repair synthesis and ligation in GG-NER	25	5	.34	14.89	+	4.28E-05	5.43E-03
↳Global Genome Nucleotide Excision Repair (GG-NER)	84	8	1.13	7.09	+	3.06E-05	5.37E-03
↳Nucleotide Excision Repair	110	10	1.48	6.77	+	4.58E-06	1.49E-03
Dual Incision in GG-NER	41	7	.55	12.71	+	3.08E-06	1.41E-03
Recognition of DNA damage by PCNA-containing replication complex	30	5	.40	12.41	+	9.24E-05	7.03E-03
↳DNA Damage Bypass	48	5	.64	7.76	+	6.73E-04	3.07E-02
Termination of translesion DNA synthesis	32	5	.43	11.64	+	1.21E-04	8.95E-03
↳Translesion synthesis by Y family DNA polymerases bypasses lesions on DNA template	39	5	.52	9.55	+	2.80E-04	1.69E-02
Regulation of MECP2 expression and activity	30	4	.40	9.93	+	1.02E-03	3.49E-02
↳Generic Transcription Pathway	1196	31	16.06	1.93	+	6.70E-04	3.12E-02
↳RNA Polymerase II Transcription	1318	34	17.70	1.92	+	3.07E-04	1.75E-02
Regulation of TP53 Activity through Acetylation	30	4	.40	9.93	+	1.02E-03	3.44E-02
↳Regulation of TP53 Activity	159	9	2.14	4.22	+	4.18E-04	2.33E-02
↳Transcriptional Regulation by TP53	360	14	4.83	2.90	+	5.01E-04	2.66E-02
Gap-filling DNA repair synthesis and ligation in TC-NER	64	7	.86	8.14	+	4.29E-05	5.16E-03
↳Transcription-Coupled Nucleotide Excision Repair (TC-NER)	78	8	1.05	7.64	+	1.87E-05	4.26E-03
RNA Polymerase I Transcription Initiation	46	5	.62	8.09	+	5.63E-04	2.79E-02
↳RNA Polymerase I Promoter Clearance	78	6	1.05	5.73	+	8.78E-04	3.34E-02
↳RNA Polymerase I Transcription	79	6	1.06	5.66	+	9.35E-04	3.44E-02
Dual incision in TC-NER	65	7	.87	8.02	+	4.70E-05	5.37E-03
Cross-presentation of soluble exogenous antigens (endosomes)	49	5	.66	7.60	+	7.33E-04	3.05E-02
Interleukin-1 signaling	100	10	1.34	7.45	+	2.08E-06	1.58E-03
↳Interleukin-1 family signaling	137	10	1.84	5.44	+	2.74E-05	5.21E-03
↳Signaling by Interleukins	447	17	6.00	2.83	+	1.66E-04	1.15E-02
↳Cytokine Signaling in Immune system	823	25	11.05	2.26	+	1.80E-04	1.21E-02
Regulation of RUNX3 expression and activity	53	5	.71	7.03	+	1.02E-03	3.52E-02
SCF-beta-TrCP mediated degradation of Emi1	54	5	.73	6.90	+	1.10E-03	3.64E-02
Mitochondrial translation elongation	88	8	1.18	6.77	+	4.16E-05	6.34E-03
↳Mitochondrial translation	94	8	1.26	6.34	+	6.43E-05	6.12E-03
↳Translation	293	16	3.93	4.07	+	4.07E-06	1.55E-03
Mitochondrial translation termination	88	8	1.18	6.77	+	4.16E-05	5.94E-03
Mitochondrial translation initiation	88	8	1.18	6.77	+	4.16E-05	5.59E-03
Formation of RNA Pol II elongation complex	61	5	.82	6.10	+	1.82E-03	4.90E-02
↳RNA Polymerase II Transcription Elongation	61	5	.82	6.10	+	1.82E-03	4.95E-02
Cellular response to hypoxia	74	6	.99	6.04	+	6.79E-04	3.04E-02
Downstream signaling events of B Cell Receptor (BCR)	79	6	1.06	5.66	+	9.35E-04	3.39E-02
Signaling by NOTCH4	80	6	1.07	5.59	+	9.94E-04	3.49E-02
CLEC7A (Dectin-1) signaling	96	7	1.29	5.43	+	4.46E-04	2.43E-02
↳C-type lectin receptors (CLRs)	138	8	1.85	4.32	+	7.50E-04	3.06E-02
↳Innate Immune System	1105	31	14.84	2.09	+	1.40E-04	1.00E-02
DNA Replication Pre-Initiation	84	6	1.13	5.32	+	1.26E-03	4.05E-02
RNA Polymerase II Pre-transcription Events	84	6	1.13	5.32	+	1.26E-03	4.00E-02
G1/S Transition	130	8	1.75	4.58	+	5.17E-04	2.62E-02
↳Mitotic G1 phase and G1/S transition	147	8	1.97	4.05	+	1.11E-03	3.61E-02
PTEN Regulation	137	8	1.84	4.35	+	7.17E-04	3.03E-02
↳PIP3 activates AKT signaling	248	11	3.33	3.30	+	7.15E-04	3.08E-02
↳Intracellular signaling by second messengers	287	12	3.85	3.11	+	6.83E-04	3.00E-02
Regulation of expression of SLITs and ROBOs	169	9	2.27	3.97	+	6.36E-04	3.02E-02
HIV Infection	227	12	3.05	3.94	+	8.83E-05	7.47E-03
↳Infectious disease	465	22	6.24	3.52	+	6.41E-07	7.32E-04
Diseases of signal transduction	366	16	4.91	3.26	+	5.53E-05	5.74E-03
Neutrophil degranulation	478	16	6.42	2.49	+	9.73E-04	3.47E-02

Figure 5.2: Gene Ontology pathway enrichment analysis of CHIP transcriptomic results using transcripts with suggestive evidence of association with CHIP (P<0.01).

Further pathway enrichment analysis using the Gene Ontology and STRING resources across transcripts with suggestive evidence of association with CHIP ($P<0.01$) detected multiple FDR-significant pathways (**Figures 5.2-3**) in processes related to DNA damage repair, telomere extension, regulation of TP53 activity, IL-1 signaling, the innate immune system, HIV infection, and infectious diseases, and neutrophil degranulation (**Figure 5.2-3**).

Reactome Pathways			
pathway	*description*	*count in gene set*	*false discovery rate*
HSA-1643685	Disease	36 of 1018	0.00036
HSA-9020702	Interleukin-1 signaling	10 of 98	0.00100
HSA-5696400	Dual Incision in GG-NER	7 of 41	0.00100
HSA-5696398	Nucleotide Excision Repair	10 of 109	0.00100
HSA-5663205	Infectious disease	18 of 363	0.00100
HSA-110373	Resolution of AP sites via the multiple-nucleotide patch repla...	6 of 25	0.00100
HSA-110314	Recognition of DNA damage by PCNA-containing replication ...	6 of 31	0.0012
HSA-73933	Resolution of Abasic Sites (AP sites)	6 of 36	0.0021
HSA-73884	Base Excision Repair	6 of 36	0.0021
HSA-72766	Translation	15 of 288	0.0021
HSA-69242	S Phase	11 of 156	0.0021
HSA-6781827	Transcription-Coupled Nucleotide Excision Repair (TC-NER)	8 of 78	0.0021
HSA-5696399	Global Genome Nucleotide Excision Repair (GG-NER)	8 of 83	0.0021
HSA-5651801	PCNA-Dependent Long Patch Base Excision Repair	5 of 21	0.0021
HSA-446652	Interleukin-1 family signaling	10 of 134	0.0021
HSA-174417	Telomere C-strand (Lagging Strand) Synthesis	5 of 24	0.0022
HSA-69239	Synthesis of DNA	9 of 114	0.0025
HSA-6782210	Gap-filling DNA repair synthesis and ligation in TC-NER	7 of 64	0.0025
HSA-6782135	Dual incision in TC-NER	7 of 65	0.0025
HSA-5696397	Gap-filling DNA repair synthesis and ligation in GG-NER	5 of 25	0.0025
HSA-69278	Cell Cycle, Mitotic	19 of 483	0.0028
HSA-69306	DNA Replication	9 of 122	0.0032
HSA-73893	DNA Damage Bypass	6 of 49	0.0038
HSA-180786	Extension of Telomeres	5 of 30	0.0038
HSA-168249	Innate Immune System	30 of 1012	0.0038
HSA-1640170	Cell Cycle	21 of 586	0.0038
HSA-162906	HIV Infection	12 of 224	0.0038
HSA-5656169	Termination of translesion DNA synthesis	5 of 32	0.0043
HSA-392499	Metabolism of proteins	47 of 1948	0.0046
HSA-74160	Gene expression (Transcription)	36 of 1366	0.0058
HSA-73857	RNA Polymerase II Transcription	33 of 1233	0.0080
HSA-69186	Lagging Strand Synthesis	4 of 20	0.0080
HSA-5419276	Mitochondrial translation termination	7 of 88	0.0083
HSA-5389840	Mitochondrial translation elongation	7 of 88	0.0083
HSA-5368286	Mitochondrial translation initiation	7 of 88	0.0083
HSA-110313	Translesion synthesis by Y family DNA polymerases bypasse...	5 of 39	0.0083

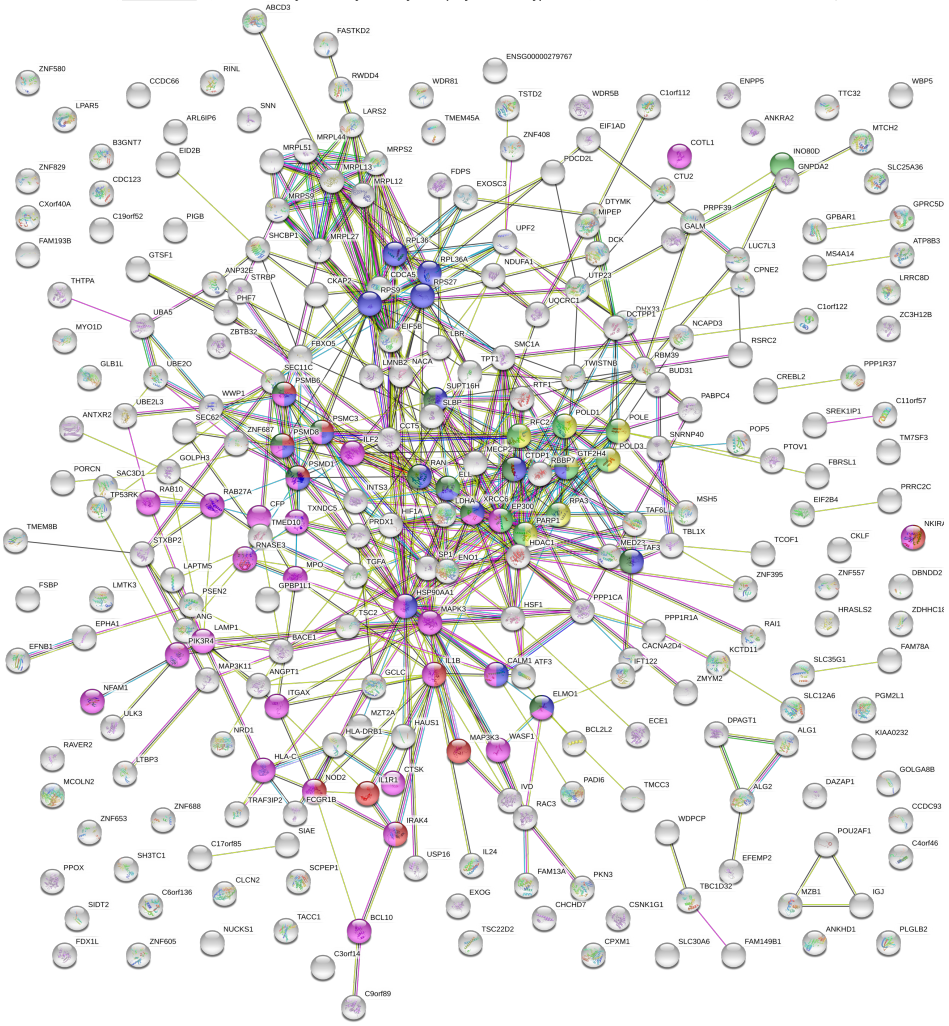


Figure 5.3: STRING-based pathway enrichment analysis and protein-protein interaction visualization across selected reactome pathways in color in the top panel.

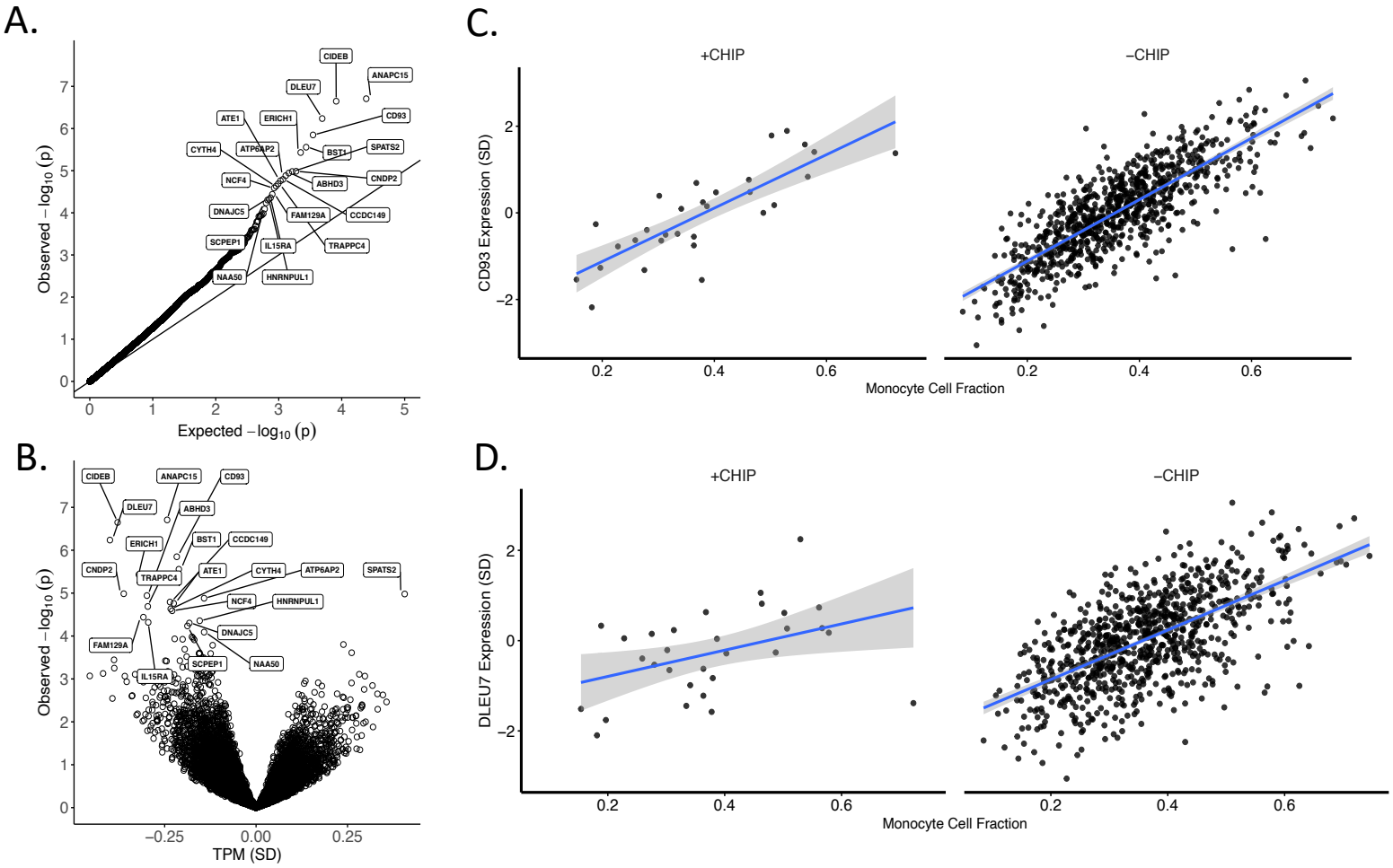


Figure 5.4: Association of CHIP x monocyte percentage interaction with transcript expression across the transcriptome. *A. quantile-quantile plot and B. volcano plot of associations, with transcripts with FDR<0.05 labeled. C-D: examples of interactions of monocyte cell fraction with CHIP on transcript levels across two sample transcripts, CD93 and DLEU7, for whom there was a significant interaction (FDR<0.05)*

Further interaction analysis between CHIP and monocyte percentage from Cibersort monocyte percentage estimation on transcript expression levels identified several transcripts which passed FDR<0.05 correction (**Figure 5.4**), including CD93 which is a myeloid cell-specific marker thought to be involved in intercellular adhesion and in clearance of apoptotic cells, as well as the DLEU7 (Deleted in Lymphocytic

Leukemia 7 protein). Of note, both of these transcripts show decreased slopes of associations between the respective gene expression and monocyte cell fraction in CHIP carriers compared to controls (**Figure 5.4-C,D**).

Similarly, transcriptome-wide association analyses was performed between expanded mCAs and transcripts expressed in blood (TPM>1), finding 3 significantly expressed transcripts associated with any expanded mCA: 1) CSF2RA (colony stimulating factor 2 receptor subunit alpha, also known as GMCSF-receptor), which is decreased in expression among carriers of expanded mCAs, as well as PRKAR1B (protein kinase CAMP-dependent type I regulatory subunit Beta) and RNF5 (ring finger protein 5, which has ubiquitin-protease ligase activity) which are both associated with increased expression among carriers of expanded mCAs. Other FDR<0.05 genes are as labeled in **Figure 5.5**. No significantly enriched pathways were observed through Gene Ontology pathway analysis of genes with P<0.01.

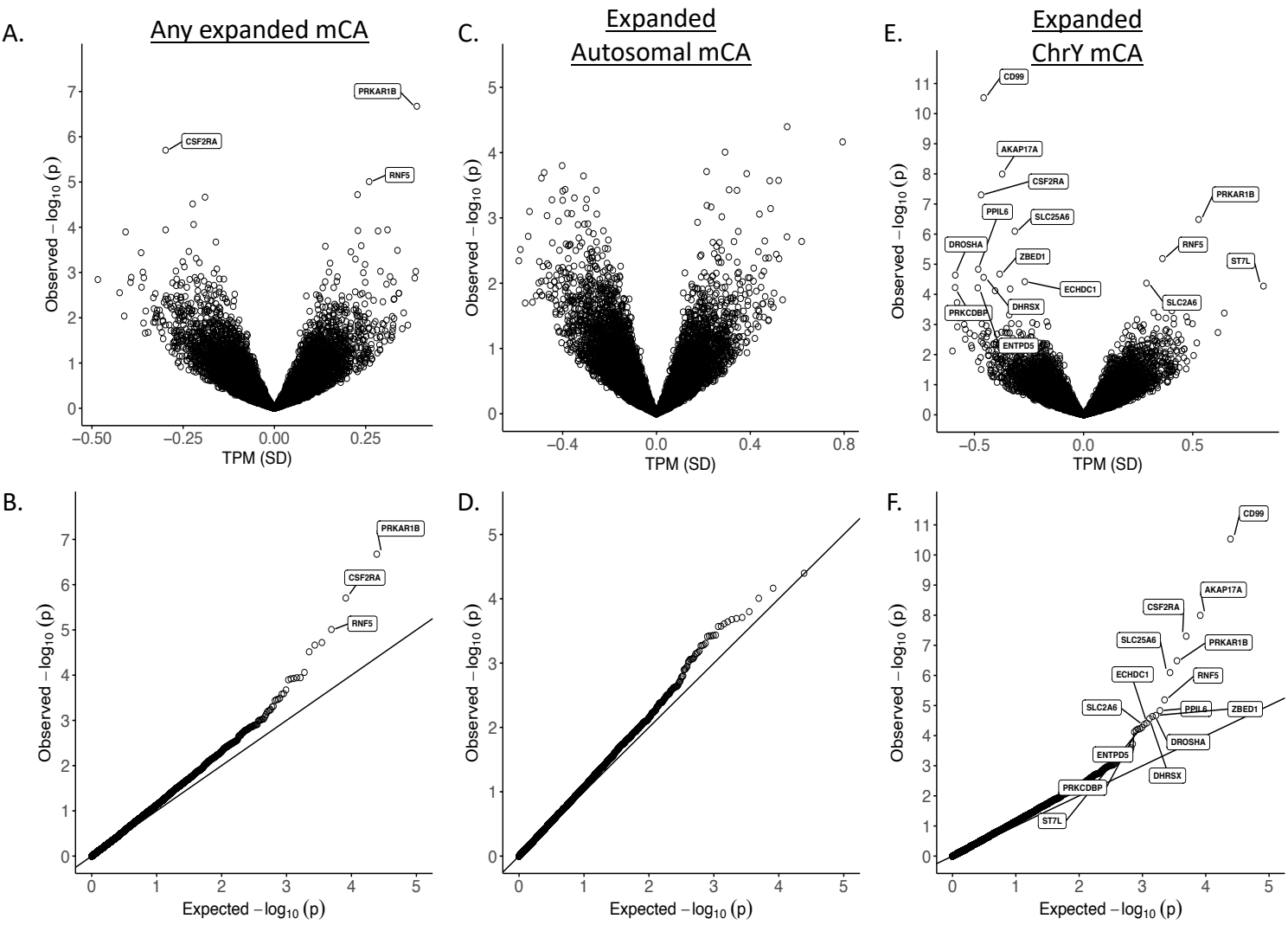


Figure 5.5: Transcriptome-wide association of mCA classes. *A, C, E: Volcano plots for expanded mCA, expanded autosomal mCA, and expanded chrY mCA. B. D. F: quantile-quantile plots for expanded mCA, expanded autosomal mCA, and expanded chrY mCA. Labeled are transcripts with FDR<0.05.*

Discussion:

Overall, I identified an significant enrichment of multiple pathways related to DNA damage repair and immune function linked to CHIP. These findings are concordant with the aforementioned phenotypic link between clonal hematopoiesis from mCAs and incident infections from Chapter 3, and the cell cycle and DNA damage repair genes involved with CHIP and mCAs in the GWAS analyses in Chapter 4. While the mCA transcriptomic analyses resulted in more individual genes identified, the lack of any enriched pathways may be due to multiple factors, including possible confounders, diverse pathways influenced by individual mCA subtypes, or lack of power. The individual associations observed merit further validation.

Several limitations exist in the present analyses. First, given the limited sample size, and the large number of tests performed, power was be limited for transcriptomics. Second, the paucity of datasets with both RNA-sequencing, genotyping for mCA calling, and next-generation sequencing for CHIP calling, all performed at the same exam visit makes replication of these results difficult. Furthermore, due to the observational and cross-sectional nature of these analyses, there is potential for reverse confounding and pleiotropic effects in the association of CHIP somatic variants with gene expression, especially by factors strongly linked to CHIP such as age. Future work with additional datasets may help resolve these issues.

Chapter 6: Conclusion

The accumulation of somatic variants contributing towards clonal hematopoiesis may reflect an aging hematopoietic system whereby senescent blood cells, in particular largely the myeloid lineage for CHIP and lymphoid for mCAs, are affected. The present dissertation permits several conclusions. First, we show through phenome-wide association analyses the link between CHIP and not only myeloid leukemias but also cardiovascular diseases including pan-vascular atherosclerosis, heart failure, and stroke. Additionally, similar analyses link mCAs with lymphoid leukemias as well as diverse infectious diseases. Second, genome-wide analyses link CHIP with several variants also linked to myeloproliferative neoplasms, and mCAs with inherited genetic regions linked to immune cells. Third, transcriptome-wide analyses, while underpowered, suggest an enrichment of pathways linked to DNA damage repair and immune function for CHIP.

Overall, the present analyses were unique in combining multiple levels of 'omics, across multiple ethnicities and cohorts around the world. With the onset of the precision medicine initiative and consortiums such as the NHLBI's Trans-Omics for Precision Medicine (TOPMed) program, great effort is being made to translate scientific findings towards clinical applications. Further efforts to connect acquired somatic mutations across diverse tissues may uncover new pathways towards common diseases. Further overlap of this data with other 'omic datasets will enable improved depth of understanding of inherited and environmental causes of somatic mutations as well as potential means of slowing or preventing the development of somatic clones towards treatment of malignancies and other age-related diseases, thereby meeting significant unmet biological, clinical, and public health needs.

Citations:

1. Steensma DP, Bejar R, Jaiswal S, et al. Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood* 2015;126(1):9-16. doi: 10.1182/blood-2015-03-631747 [published Online First: 2015/05/02]
2. Loh PR, Genovese G, Handsaker RE, et al. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* 2018;559(7714):350-55. doi: 10.1038/s41586-018-0321-x [published Online First: 2018/07/12]
3. Karmali KN, Goff DC, Jr., Ning H, et al. A systematic examination of the 2013 ACC/AHA pooled cohort risk assessment tool for atherosclerotic cardiovascular disease. *J Am Coll Cardiol* 2014;64(10):959-68. doi: 10.1016/j.jacc.2014.06.1186 [published Online First: 2014/09/06]
4. Lozano R, Naghavi M, Foreman K, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* 2012;380(9859):2095-128. doi: 10.1016/S0140-6736(12)61728-0 [published Online First: 2012/12/19]
5. Jaiswal S, Natarajan P, Silver AJ, et al. Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *N Engl J Med* 2017;377(2):111-21. doi: 10.1056/NEJMoa1701719 [published Online First: 2017/06/22]
6. Genovese G, Kahler AK, Handsaker RE, et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N Engl J Med* 2014;371(26):2477-87. doi: 10.1056/NEJMoa1409405 [published Online First: 2014/11/27]
7. Ridker PM, Everett BM, Thuren T, et al. Antiinflammatory Therapy with Canakinumab for Atherosclerotic Disease. *N Engl J Med* 2017;377(12):1119-31. doi: 10.1056/NEJMoa1707914 [published Online First: 2017/08/29]
8. Svensson EC, Madar A, Campbell CD, et al. TET2-Driven Clonal Hematopoiesis Predicts Enhanced Response to Canakinumab in the CANTOS Trial: An Exploratory Analysis. *Circulation* 2018;138(Suppl_1):A15111-A11.
9. Terao C, Suzuki A, Momozawa Y, et al. Chromosomal alterations among age-related haematopoietic clones in Japan. *Nature* 2020 doi: 10.1038/s41586-020-2426-2 [published Online First: 2020/06/26]
10. Loh PR, Genovese G, McCarroll SA. Monogenic and polygenic inheritance become instruments for clonal selection. *Nature* 2020 doi: 10.1038/s41586-020-2430-6 [published Online First: 2020/06/26]
11. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018;562(7726):203-09. doi: 10.1038/s41586-018-0579-z [published Online First: 2018/10/12]
12. Smoller JW, Karlson EW, Green RC, et al. An eMERGE Clinical Center at Partners Personalized Medicine. *J Pers Med* 2016;6(1) doi: 10.3390/jpm6010005 [published Online First: 2016/01/26]
13. Bhattacharya R, Zekavat SM, Haessler J, et al. Clonal Hematopoiesis Is Associated With Higher Risk of Stroke. *Stroke* 2021:STROKEAHA. 121.037388.

14. Yu B, Roberts MB, Raffield LM, et al. Supplemental Association of Clonal Hematopoiesis With Incident Heart Failure. *J Am Coll Cardiol* 2021;78(1):42-52. doi: 10.1016/j.jacc.2021.04.085 [published Online First: 2021/07/03]
15. Zekavat SM, Lin SH, Bick AG, et al. Hematopoietic mosaic chromosomal alterations increase the risk for diverse types of infection. *Nat Med* 2021;27(6):1012-24. doi: 10.1038/s41591-021-01371-0 [published Online First: 2021/06/09]
16. Zekavat SM, Viana-Huete V, Zuriaga MA, et al. TP53-mediated clonal hematopoiesis confers increased risk for incident peripheral artery disease. *medRxiv* 2021:2021.08.22.21262430. doi: 10.1101/2021.08.22.21262430
17. Nagai A, Hirata M, Kamatani Y, et al. Overview of the BioBank Japan Project: Study design and profile. *J Epidemiol* 2017;27(3S):S2-S8. doi: 10.1016/j.je.2016.12.005 [published Online First: 2017/02/13]
18. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 2013;31(3):213-9. doi: 10.1038/nbt.2514 [published Online First: 2013/02/12]
19. Jaiswal S, Fontanillas P, Flannick J, et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N Engl J Med* 2014;371(26):2488-98. doi: 10.1056/NEJMoa1408617 [published Online First: 2014/11/27]
20. Bick AG, Pirruccello JP, Griffin GK, et al. Genetic Interleukin 6 Signaling Deficiency Attenuates Cardiovascular Risk in Clonal Hematopoiesis. *Circulation* 2020;141(2):124-31. doi: 10.1161/CIRCULATIONAHA.119.044362 [published Online First: 2019/11/12]
21. Gibson CJ, Lindsley RC, Tchekmedyian V, et al. Clonal Hematopoiesis Associated With Adverse Outcomes After Autologous Stem-Cell Transplantation for Lymphoma. *J Clin Oncol* 2017;35(14):1598-605. doi: 10.1200/JCO.2016.71.6712 [published Online First: 2017/01/10]
22. Bick AG, Weinstock JS, Nandakumar SK, et al. Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature* 2020;586(7831):763-68.
23. Peiffer DA, Le JM, Steemers FJ, et al. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res* 2006;16(9):1136-48. doi: 10.1101/gr.5402306 [published Online First: 2006/08/11]
24. Delaneau O, Zagury JF, Robinson MR, et al. Accurate, scalable and integrative haplotype estimation. *Nat Commun* 2019;10(1):5436. doi: 10.1038/s41467-019-13225-y [published Online First: 2019/11/30]
25. Loh PR, Danecek P, Palamara PF, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* 2016;48(11):1443-48. doi: 10.1038/ng.3679 [published Online First: 2016/10/28]
26. Loftfield E, Zhou W, Graubard BI, et al. Predictors of mosaic chromosome Y loss and associations with mortality in the UK Biobank. *Sci Rep* 2018;8(1):12316. doi: 10.1038/s41598-018-30759-1 [published Online First: 2018/08/19]
27. Bick AG, Weinstock JS, Nandakumar SK, et al. Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature* 2020;586(7831):763-68. doi: 10.1038/s41586-020-2819-2 [published Online First: 2020/10/16]

28. Wu P, Gifford A, Meng X, et al. Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Med Inform* 2019;7(4):e14325. doi: 10.2196/14325 [published Online First: 2019/09/26]
29. Denny JC, Ritchie MD, Basford MA, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 2010;26(9):1205-10. doi: 10.1093/bioinformatics/btq126 [published Online First: 2010/03/26]
30. Conte MS, Bradbury AW, Kolh P, et al. Global vascular guidelines on the management of chronic limb-threatening ischemia. *J Vasc Surg* 2019;69(6s):3S-125S.e40. doi: 10.1016/j.jvs.2019.02.016 [published Online First: 2019/06/05]
31. Jaiswal S, Fontanillas P, Flannick J, et al. Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *New England Journal of Medicine* 2014;371(26):2488-98. doi: 10.1056/NEJMoa1408617
32. Klarin D, Lynch J, Aragam K, et al. Genome-wide association study of peripheral artery disease in the Million Veteran Program. *Nature medicine* 2019;25(8):1274-79. doi: 10.1038/s41591-019-0492-5 [published Online First: 2019/07/10]
33. Denny JC, Bastarache L, Ritchie MD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* 2013;31(12):1102-10. doi: 10.1038/nbt.2749
34. Jaiswal S, Natarajan P, Silver AJ, et al. Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *New England Journal of Medicine* 2017;377(2):111-21. doi: 10.1056/NEJMoa1701719
35. Fuster JJ, MacLauchlan S, Zuriaga MA, et al. Clonal hematopoiesis associated with TET2 deficiency accelerates atherosclerosis development in mice. *Science (New York, NY)* 2017;355(6327):842-47. doi: 10.1126/science.aag1381 [published Online First: 2017/01/21]
36. Visconte V, M ON, H JR. Mutations in Splicing Factor Genes in Myeloid Malignancies: Significance and Impact on Clinical Features. *Cancers (Basel)* 2019;11(12) doi: 10.3390/cancers11121844 [published Online First: 2019/11/27]
37. Klarin D, Verma SS, Judy R, et al. Genetic Architecture of Abdominal Aortic Aneurysm in the Million Veteran Program. *Circulation* 2020 doi: 10.1161/circulationaha.120.047544 [published Online First: 2020/09/29]
38. Moskowitz MA, Lo EH, Iadecola C. The science of stroke: mechanisms in search of treatments. *Neuron* 2010;67(2):181-98. doi: 10.1016/j.neuron.2010.07.002 [published Online First: 2010/07/31]
39. Shuaib A, Hachinski VC. Mechanisms and management of stroke in the elderly. *Cmaj* 1991;145(5):433-43. [published Online First: 1991/09/01]
40. Yang SJ, Shao GF, Chen JL, et al. The NLRP3 Inflammasome: An Important Driver of Neuroinflammation in Hemorrhagic Stroke. *Cell Mol Neurobiol* 2018;38(3):595-603. doi: 10.1007/s10571-017-0526-9 [published Online First: 2017/07/29]
41. Li H, Hastings MH, Rhee J, et al. Targeting Age-Related Pathways in Heart Failure. *Circ Res* 2020;126(4):533-51. doi: 10.1161/CIRCRESAHA.119.315889

42. Huffman MD, Berry JD, Ning H, et al. Lifetime risk for heart failure among white and black Americans: cardiovascular lifetime risk pooling project. *J Am Coll Cardiol* 2013;61(14):1510-7. doi: 10.1016/j.jacc.2013.01.022
43. Stewart S, Ekman I, Ekman T, et al. Population impact of heart failure and the most common forms of cancer: a study of 1 162 309 hospital cases in Sweden (1988 to 2004). *Circ Cardiovasc Qual Outcomes* 2010;3(6):573-80. doi: 10.1161/CIRCOUTCOMES.110.957571
44. Dorsheimer L, Assmus B, Rasper T, et al. Association of Mutations Contributing to Clonal Hematopoiesis With Prognosis in Chronic Ischemic Heart Failure. *JAMA Cardiol* 2019;4(1):25-33. doi: 10.1001/jamacardio.2018.3965
45. Patel AP, Natarajan P. A New Murine Model of Clonal Hematopoiesis Investigates JAK2 (V617F) in Heart Failure. *JACC Basic Transl Sci* 2019;4(6):698-700. doi: 10.1016/j.jacbts.2019.09.003
46. Sano S, Wang Y, Yura Y, et al. JAK2 (V617F) -Mediated Clonal Hematopoiesis Accelerates Pathological Remodeling in Murine Heart Failure. *JACC Basic Transl Sci* 2019;4(6):684-97. doi: 10.1016/j.jacbts.2019.05.013
47. Sano S, Oshima K, Wang Y, et al. Tet2-Mediated Clonal Hematopoiesis Accelerates Heart Failure Through a Mechanism Involving the IL-1beta/NLRP3 Inflammasome. *J Am Coll Cardiol* 2018;71(8):875-86. doi: 10.1016/j.jacc.2017.12.037
48. Cremer S, Kirschbaum K, Berkowitsch A, et al. Multiple Somatic Mutations for Clonal Hematopoiesis Are Associated With Increased Mortality in Patients With Chronic Heart Failure. *Circ Genom Precis Med* 2020;13(4):e003003. doi: 10.1161/CIRCGEN.120.003003
49. Djohan AH, Sia CH, Lee PS, et al. Endothelial Progenitor Cells in Heart Failure: an Authentic Expectation for Potential Future Use and a Lack of Universal Definition. *J Cardiovasc Transl Res* 2018;11(5):393-402. doi: 10.1007/s12265-018-9810-4
50. Gardner ID. The effect of aging on susceptibility to infection. *Rev Infect Dis* 1980;2(5):801-10. doi: 10.1093/clinids/2.5.801 [published Online First: 1980/09/01]
51. Gavazzi G, Krause KH. Ageing and infection. *Lancet Infect Dis* 2002;2(11):659-66. doi: 10.1016/s1473-3099(02)00437-1 [published Online First: 2002/11/01]
52. Aw D, Silva AB, Palmer DB. Immunosenescence: emerging challenges for an ageing population. *Immunology* 2007;120(4):435-46. doi: 10.1111/j.1365-2567.2007.02555.x [published Online First: 2007/02/23]
53. Franceschi C, Bonafe M, Valensin S. Human immunosenescence: the prevailing of innate immunity, the failing of clonotypic immunity, and the filling of immunological space. *Vaccine* 2000;18(16):1717-20. doi: 10.1016/s0264-410x(99)00513-7 [published Online First: 2000/02/26]
54. Ongradi J, Kovesdi V. Factors that may impact on immunosenescence: an appraisal. *Immun Ageing* 2010;7:7. doi: 10.1186/1742-4933-7-7 [published Online First: 2010/06/16]

55. Panda A, Arjona A, Sapey E, et al. Human innate immunosenescence: causes and consequences for immunity in old age. *Trends Immunol* 2009;30(7):325-33. doi: 10.1016/j.it.2009.05.004 [published Online First: 2009/06/23]
56. Aoshi T, Koyama S, Kobiyama K, et al. Innate and adaptive immune responses to viral infection and vaccination. *Curr Opin Virol* 2011;1(4):226-32. doi: 10.1016/j.coviro.2011.07.002 [published Online First: 2012/03/24]
57. Holly MK, Diaz K, Smith JG. Defensins in Viral Infection and Pathogenesis. *Annu Rev Virol* 2017;4(1):369-91. doi: 10.1146/annurev-virology-101416-041734 [published Online First: 2017/07/19]
58. Pallett LJ, Schmidt N, Schurich A. T cell metabolism in chronic viral infection. *Clin Exp Immunol* 2019;197(2):143-52. doi: 10.1111/cei.13308 [published Online First: 2019/05/01]
59. Lin SH, Loftfield E, Sampson JN, et al. Mosaic chromosome Y loss is associated with alterations in blood cell counts in UK Biobank men. *Sci Rep* 2020;10(1):3655. doi: 10.1038/s41598-020-59963-8 [published Online First: 2020/02/29]
60. Forsberg LA, Rasi C, Malmqvist N, et al. Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. *Nat Genet* 2014;46(6):624-8. doi: 10.1038/ng.2966 [published Online First: 2014/04/30]
61. Jacobs KB, Yeager M, Zhou W, et al. Detectable clonal mosaicism and its relationship to aging and cancer. *Nat Genet* 2012;44(6):651-8. doi: 10.1038/ng.2270 [published Online First: 2012/05/09]
62. Laurie CC, Laurie CA, Rice K, et al. Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat Genet* 2012;44(6):642-50. doi: 10.1038/ng.2271 [published Online First: 2012/05/09]
63. Machiela MJ, Zhou W, Sampson JN, et al. Characterization of large structural genetic mosaicism in human autosomes. *Am J Hum Genet* 2015;96(3):487-97. doi: 10.1016/j.ajhg.2015.01.011 [published Online First: 2015/03/10]
64. Xie M, Lu C, Wang J, et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat Med* 2014;20(12):1472-8. doi: 10.1038/nm.3733 [published Online First: 2014/10/20]
65. Wang L, Fan J, Francis JM, et al. Integrated single-cell genetic and transcriptional analysis suggests novel drivers of chronic lymphocytic leukemia. *Genome Res* 2017;27(8):1300-11. doi: 10.1101/gr.217331.116 [published Online First: 2017/07/07]
66. de Weerdt I, van Hoeven V, Munneke JM, et al. Innate lymphoid cells are expanded and functionally altered in chronic lymphocytic leukemia. *Haematologica* 2016;101(11):e461-e64. doi: 10.3324/haematol.2016.144725 [published Online First: 2016/11/02]
67. Bartik MM, Welker D, Kay NE. Impairments in immune cell function in B cell chronic lymphocytic leukemia. *Semin Oncol* 1998;25(1):27-33. [published Online First: 1998/03/03]
68. Arruga F, Gyau BB, Iannello A, et al. Immune Response Dysfunction in Chronic Lymphocytic Leukemia: Dissecting Molecular Mechanisms and

Microenvironmental Conditions. *Int J Mol Sci* 2020;21(5) doi: 10.3390/ijms21051825 [published Online First: 2020/03/12]
69. Zhou W, Machiela MJ, Freedman ND, et al. Mosaic loss of chromosome Y is associated with common variation near TCL1A. *Nat Genet* 2016;48(5):563-8. doi: 10.1038/ng.3545 [published Online First: 2016/04/12]
70. Galluzzi L, Buque A, Kepp O, et al. Immunological Effects of Conventional Chemotherapy and Targeted Anticancer Agents. *Cancer Cell* 2015;28(6):690-714. doi: 10.1016/j.ccell.2015.10.012 [published Online First: 2015/12/19]
71. Balkwill F, Mantovani A. Inflammation and cancer: back to Virchow? *Lancet* 2001;357(9255):539-45. doi: 10.1016/S0140-6736(00)04046-0 [published Online First: 2001/03/07]
72. de Visser KE, Eichten A, Coussens LM. Paradoxical roles of the immune system during cancer development. *Nat Rev Cancer* 2006;6(1):24-37. doi: 10.1038/nrc1782 [published Online First: 2006/01/07]
73. Lucas C, Wong P, Klein J, et al. Longitudinal analyses reveal immunological misfiring in severe COVID-19. *Nature* 2020 doi: 10.1038/s41586-020-2588-y [published Online First: 2020/07/28]
74. Giamarellos-Bourboulis EJ, Netea MG, Rovina N, et al. Complex Immune Dysregulation in COVID-19 Patients with Severe Respiratory Failure. *Cell Host Microbe* 2020;27(6):992-1000 e3. doi: 10.1016/j.chom.2020.04.009 [published Online First: 2020/04/23]
75. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 2020;395(10223):497-506. doi: 10.1016/S0140-6736(20)30183-5 [published Online First: 2020/01/28]
76. Cunha LL, Perazzio SF, Azzi J, et al. Remodeling of the Immune Response With Aging: Immunosenescence and Its Potential Impact on COVID-19 Immune Response. *Front Immunol* 2020;11:1748. doi: 10.3389/fimmu.2020.01748 [published Online First: 2020/08/28]
77. Zink F, Stacey SN, Norddahl GL, et al. Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* 2017;130(6):742-52. doi: 10.1182/blood-2017-02-769869 [published Online First: 2017/05/10]
78. Bick AG, Weinstock JS, Nandakumar SK, et al. Inherited Causes of Clonal Hematopoiesis of Indeterminate Potential in TOPMed Whole Genomes. *bioRxiv* 2019:782748. doi: 10.1101/782748
79. Zhou W, Nielsen JB, Fritsche LG, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* 2018;50(9):1335-41. doi: 10.1038/s41588-018-0184-y [published Online First: 2018/08/15]
80. Jelsema CM, Peddada SD. CLME: An R Package for Linear Mixed Effects Models under Inequality Constraints. *J Stat Softw* 2016;75 doi: 10.18637/jss.v075.i01 [published Online First: 2016/01/01]
81. Hu Y, Li M, Lu Q, et al. A statistical framework for cross-tissue transcriptome-wide association analysis. *Nature genetics* 2019;51(3):568-76.

82. Reynolds LM, Ding J, Taylor JR, et al. Transcriptomic profiles of aging in purified human immune cells. *BMC Genomics* 2015;16:333. doi: 10.1186/s12864-015-1522-4 [published Online First: 2015/04/23]
83. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;29(1):15-21. doi: 10.1093/bioinformatics/bts635 [published Online First: 2012/10/30]
84. Bojesen SE, Pooley KA, Johnatty SE, et al. Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. *Nat Genet* 2013;45(4):371-84, 84e1-2. doi: 10.1038/ng.2566 [published Online First: 2013/03/29]
85. Bao EL, Nandakumar SK, Liao X, et al. Inherited myeloproliferative neoplasm risk affects haematopoietic stem cells. *Nature* 2020;586(7831):769-75. doi: 10.1038/s41586-020-2786-7 [published Online First: 2020/10/16]
86. Smith BW, Rozelle SS, Leung A, et al. The aryl hydrocarbon receptor directs hematopoietic progenitor cell expansion and differentiation. *Blood* 2013;122(3):376-85. doi: 10.1182/blood-2012-11-466722 [published Online First: 2013/06/01]
87. Thompson DJ, Genovese G, Halvardson J, et al. Genetic predisposition to mosaic Y chromosome loss in blood. *Nature* 2019;575(7784):652-57. doi: 10.1038/s41586-019-1765-3 [published Online First: 2019/11/22]
88. Finucane HK, Bulik-Sullivan B, Gusev A, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* 2015;47(11):1228-35. doi: 10.1038/ng.3404 [published Online First: 2015/09/29]
89. Lu Q, Powles RL, Abdallah S, et al. Systematic tissue-specific functional annotation of the human genome highlights immune-related DNA elements for late-onset Alzheimer's disease. *PLoS Genet* 2017;13(7):e1006933. doi: 10.1371/journal.pgen.1006933 [published Online First: 2017/07/26]
90. Hu Y, Li M, Lu Q, et al. A statistical framework for cross-tissue transcriptome-wide association analysis. *Nat Genet* 2019;51(3):568-76. doi: 10.1038/s41588-019-0345-7 [published Online First: 2019/02/26]
91. Kuleshov MV, Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 2016;44(W1):W90-7. doi: 10.1093/nar/gkw377 [published Online First: 2016/05/05]
92. Consortium GT. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 2020;369(6509):1318-30. doi: 10.1126/science.aaz1776 [published Online First: 2020/09/12]
93. Stegle O, Parts L, Durbin R, et al. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput Biol* 2010;6(5):e1000770. doi: 10.1371/journal.pcbi.1000770 [published Online First: 2010/05/14]
94. Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* 2015;12(5):453-7. doi: 10.1038/nmeth.3337 [published Online First: 2015/03/31]