## Title
Inherited causes of clonal haematopoiesis in 97,691 whole genomes



## Journal
Nature, 586(7831)

## ISSN
0028-0836

## Authors
Bick, Alexander G
Weinstock, Joshua S
Nandakumar, Satish K
et al.

## Publication Date
2020-10-29

## DOI
10.1038/s41586-020-2819-2

Peer reviewed

HHS Public Access

Author manuscript

*Nature*. Author manuscript; available in PMC 2021 April 14.

Published in final edited form as:

*Nature*. 2020 October ; 586(7831): 763–768. doi:10.1038/s41586-020-2819-2.

# Inherited Causes of Clonal Hematopoiesis in 97,691 TOPMed Whole Genomes

*A full list of authors and affiliations appears at the end of the article.*

## Abstract

Age is the dominant risk factor for most chronic human diseases; yet the mechanisms by which aging confers this risk are largely unknown.[1] Recently, the age-related acquisition of somatic mutations in regenerating hematopoietic stem cell populations leading to clonal expansion was associated with both hematologic cancer [2–4] and coronary heart disease[5], a phenomenon termed 'Clonal Hematopoiesis of Indeterminate Potential' (CHIP).[6] Simultaneous germline and somatic whole genome sequence analysis now provides the opportunity to identify root causes of CHIP. Here, we analyze high-coverage whole genome sequences from 97,691 participants of diverse ancestries in the NHLBI TOPMed program and identify 4,229 individuals with CHIP. We identify associations with blood cell, lipid, and inflammatory traits specific to different CHIP genes. Association of a genome-wide set of germline genetic variants identified three genetic loci associated with CHIP status, including one locus at *TET2* that was African ancestry specific. *In silico*-informed *in vitro* evaluation of the *TET2* germline locus identified a causal variant that disrupts a *TET2* distal enhancer resulting in increased hematopoietic stem cell self-renewal. Overall, we observe that germline genetic variation shapes hematopoietic stem cell function



Please address correspondence to: Pradeep Natarajan, MD MMSc, 185 Cambridge St, CPZN 3.184, Boston, MA 02114 USA, pradeep@broadinstitute.org, Twitter: @pnatarajanmd, Sekar Kathiresan, MD, 75 Ames St, Cambridge, MA 02139 USA, sekar@broadinstitute.org, Twitter: @skathire.

‡Present Address: Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN.

AUTHOR CONTRIBUTIONS

A.G.B., P.N. and S.K. conceived the study. A.G.B. and J.S.W. performed the germline and somatic whole genome sequence analyses. C.P.F., E.L.B., S.M.Z, M.D.S., M.J.L., J.N., K.C., C.J.G., A.E.L., B.B.B., P.S., J.K., J.M.E., A.P.R, B.L.E., and S.J. performed additional bioinformatic analyses. S.K.N., X.L. and V.G.S. experimentally characterized the TET2 locus. M.A.T., F.A., K.A., B.D.M, K.C.B., A.M., M.F., S.R., B.M.P., E.K.S., S.T.W., N.D.P., R.S.V., E.G.B., S.L.R.K., J.H., R.C.K., N.L.S., D.K.A., D.A.S., A.C., M.d.A., X.G.,B.A.K., B.C., J.M.P., H.G., D.A.M., S.T.M., I.Y., M.B.S., P.A.P., J.G.B., S.M.G., F.F.W., Q.W., M.E.M., M.D., E.E.K., K.E.N., L.J.L., B.E.C., J.C.B., M.H.C., J.L.S., D.W.B., L.A.C., A.C.M., L.C.B., J.A.S., T.N.K, S.A., S.R.H., H.K.T.,I.V.Y., J.A.H., S.L., J.M.J., J.E.C., S.E.W., D.E.W., D.C.R., D.D., J.Y.M., R.P.T., E.J.B., N.R., R.J.F.L., P.D., Y.L., L.H., J.L., P.K., B.I.F., D.L., L.F.B., J.E.H., J.S.F., E.A.W., P.T.E., M.R.I., T.E.F., L.M.R., S.M.A., M.M.W., E.C.S., J.B., L.K.W., B.D.L., W.H.S., D.M.R., E.B., J.E.M., R.A.M., P.D., K.D.T., A.D.J., P.L.A., C.K., C.C.L., T.W.B., A.V.S., H.Z.,, E.L., L.L., S.S.R., J.I.R., J.G.W., P.S., J.O.K., E.S.L., J.M.E., and G.A. contributed to sample acquisition, DNA sequencing and phenotypic curation for the NHLBI TOPMed constituent cohorts analyzed here. A.G.B., J.S.W., S.K. and P.N. wrote the manuscript with input from all authors.

*These individuals contributed equally to this work

#These individuals jointly supervised this work

†A list of authors and their affiliations appears at the end of the paper

DATA AVAILABILITY

Individual whole-genome sequence data for TOPMed whole genomes, individual-level harmonized phenotypes, harmonized germline variant call sets, the CHIP somatic variant call sets, RNA-Seq and peripheral blood methylation data used in this analysis are available through restricted access via the dbGaP. Accession numbers for these datasets are provided in Supplementary Table 1. Summary-level genotype data are available through the BRAVO browser (https://bravo.sph.umich.edu/). Full GWAS summary statistics are available for general research use through controlled access at dbGaP accession phs001974: NHLBI TOPMed: Genomic Summary Results for the Trans-Omics for Precision Medicine Program. A subset of the TOPMed cohorts analyzed here are based on sensitive populations, precluding public sharing of full genomic summary results.

leading to CHIP through mechanisms that are both specific to clonal hematopoiesis and shared mechanisms leading to somatic mutations across tissues.

The U.S. National Heart, Lung, and Blood Institute (NHLBI) Trans-omics for Precision Medicine (TOPMed) project seeks to use high-coverage (>35x) whole genome sequencing (WGS) and molecular profiling to improve fundamental understanding of heart, lung, blood, and sleep disorders.[7] Within the TOPMed program, we designed a study to detect CHIP from blood DNA-derived WGS in 97,691 individuals across 52 largely observational epidemiologic studies to discover the inherited genetic causes and phenotypic consequences of CHIP (Supplementary Table 1).

To confidently identify somatic mutations in blood-derived DNA, we applied a somatic variant caller[8] to TOPMed WGS data. We identified CHIP carriers on the basis of a pre-specified list of leukemogenic driver mutations (see Methods, Supplementary Table 2).[5]

In total, we identified 4,938 CHIP mutations in 4,229 individuals (Supplementary Table 3). The median variant allele fraction (VAF) of the CHIP mutations observed was 16%. Consistent with prior reports, >75% of these CHIP mutations were in one of three genes, *DNMT3A*, *TET2*, and *ASXL1*. Approximately 15% of these CHIP mutations were in the five next most frequent genes (*PPM1D*, *JAK2*, *SF3B1*, *SRSF2* and *TP53*, Figure 1). Amongst these 8 genes, there was marked heterogeneity in clonal fraction. For example, *DNMT3A* and *TET2* CHIP clonal fraction of the peripheral blood was ~25% smaller ($p= 1.3 \times 10^{-15}$) and ~14% smaller ($p=2.1 \times 10^{-4}$), respectively, than *ASXL1* clonal fraction, implicating the presence of driver mutation gene-specific differences in clonal selection (Extended Data Figure 1a). 90% of individuals with CHIP driver mutations had only one identified mutation (Extended Data Figure 1b).

## CHIP phenotypic associations

CHIP prevalence was strongly correlated with age at blood draw ($p < 10^{-300}$, Figure 1 inset). CHIP prevalence was highly consistent across studies and comparable to previous reports[2–4] using whole exome sequencing (Extended Data Figure 1c,d). Consistent with prior studies, history of smoking was associated with increased CHIP odds (OR = 1.18, $p=5 \times 10^{-5}$) whereas Hispanic ancestry and East Asian ancestry were each associated with reduced CHIP odds (OR = 0.50, p=0.008 and OR = 0.56, p=0.001 respectively) after adjusting for age (Supplementary Table 4).

Carriers of frameshift CHIP mutations were on average older individuals than carriers of single nucleotide CHIP mutations (Wilcox rank sum test: p=0.01). In the subset of individuals with *ASXL1* CHIP mutations, which are exclusively loss-of-function single nucleotide stop-gain mutations or frameshift mutations, *ASXL1* frameshift mutation carriers were similarly older (Wilcox rank sum test: p=0.009, Extended Data Figure 3a).

*JAK2* CHIP carriers were the youngest among CHIP carriers. Relative to *JAK2*, *ASXL1* and *TET2* carriers were 3.3 (p=0.01) and 3.9 ($p=9.1 \times 10^{-4}$) years older, respectively, while

*PPM1D*, *SF3B1* and *SRSF2* carriers were 5.0, 6.9 and 7.7 years older ($p=5.7 \times 10^{-4}$, $1.8 \times 10^{-6}$, $1.3 \times 10^{-4}$), respectively (Extended Data Figure 3b).

To evaluate the overlap between CHIP and large-scale mosaic chromosomal rearrangements[9], we evaluated a subset of 855 samples with both WGS and array genotyping data. The two somatic events did not co-occur more than expected by chance (hypergeometric p=0.25, Extended Data Figure 3c).

CHIP is distinguished from other clonal hematologic disorders based on the absence of cytopenia, dysplasia, and neoplasia.[6] We observed a modest increase in total white blood cell count ($p=1.1 \times 10^{-5}$) and a modest decrease in hemoglobin (p=0.04), among those with CHIP compared to those without (Extended Data Figure 3a, Supplementary Table 5). In aggregate, CHIP driver mutations were associated with increased red blood cell distribution width (RDW, $p=3.0 \times 10^{-5}$) consistent with prior observations.[10] Notably, RDW is a hematologic parameter that increases with age and predicts overall mortality and poor clinical outcomes in the setting of CVD and in older adults.[11]

Given the prior association of CHIP with atherosclerotic cardiovascular disease[5,12], we asked whether CHIP carriers had altered lipid profiles. Consistent with prior reports[5], we observed negative correlations of *JAK2* CHIP carrier status with total cholesterol ($p=5.1 \times 10^{-4}$) and LDL cholesterol (p=0.0014) but no other significant associations (Extended Data Figure 3b, Supplementary Table 6).

We characterized the human inflammatory profile of CHIP carriers (Extended Data Figure 3c, Supplementary Table 7). In aggregate, CHIP was associated with increased IL-6 (p=0.0035). There was no association of CHIP with quantitative C-reactive protein (CRP) and elevated CRP did not reliably identify carriers of CHIP (AUC: 0.55; for cutoff of CRP>2 mg/L: PPV=6.3%, sensitivity=60%). Driver gene-specific analyses highlighted the association of *TET2* CHIP with increased IL-1b ($p=2.4 \times 10^{-4}$), while *JAK2* and *SF3B1* were associated with increased circulating IL-18 ($p=1.3 \times 10^{-4}$ and $1.27 \times 10^{-20}$ respectively).

To identify underlying determinants of the somatic mutational spectrum, we performed COSMIC mutational signature analysis[13] on passenger somatic mutations identified in CHIP carriers and non-carriers (see Methods). Among CHIP carriers, we observe enrichment of signature 4, which has been associated with smoking, as well as signature 6, which has been associated with defective DNA mismatch repair. (Extended Data Figure 5).

## Germline genetic determinants of CHIP

Germline genetic variants have been previously associated with clonal hematopoiesis, defined either by somatic mosaicism of SNVs and indels[14] or by large scale chromosomal rearrangements[9], in individuals of European ancestry, and identified variants at a single locus, *TERT*, that associates with clonal hematopoiesis. Given the distinct association of clonal hematopoiesis with known leukemogenic mutations (i.e., CHIP) with both cancer[2,15,16] and atherosclerotic cardiovascular disease[5,12], we sought to discover germline genetic variations conferring increased risk for CHIP acquisition. We performed a single variant genome-wide association analysis in a subset of 65,405 individuals (3,831 CHIP

cases) where the likelihood of having a CHIP mutation was >1% (see Methods). The trait heritability explained by the analysis with LD score-regression was 3.6%.

Our WGS-based association analysis of CHIP 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),[14] 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 ($p=8.4\times10^{-24}$). This variant was previously significantly associated with increased leukocyte telomere length[17], myeloproliferative neoplasms (MPN, Bao, co-submitted manuscript) and clonal chromosomal mosaicism[9]. In a phenome-wide association analysis (PheWAS) of rs34002450-A in UK Biobank, we identified significant increased risk of MPN ($p=2.6 \times 10^{-13}$), uterine leiomyoma ($p=3.2 \times 10^{-9}$) and brain cancer ($p=3.6 \times 10^{-8}$).

We performed a conditional analysis at the TERT locus, and 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, $p=6.1\times10^{-10}$; 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 3, Extended Data Figure 6, Supplementary Table 8).

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 ($p=5.3\times10^{-10}$). Variants in LD with this variant have been identified as associated with MPN (Bao et al, co-submitted manuscript). No other significant phenotypic 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 ($p=4.0\times10^{-9}$). We replicated this association in an additional 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 (Bao et al, co-submitted manuscript), 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 published association studies.

We considered whether there might be germline variants that predispose to specific CHIP driver mutations by separately performing a GWAS on *DNMT3A* CHIP and *TET2* CHIP. We identified a single novel locus for *DNMT3A* chip at rs2887399 in an intron of T-cell leukemia/lymphoma 1A (*TCL1A*). Carriers of the T allele (MAF 0.26) are at 1.23-fold risk

of acquiring a *DNMT3A* CHIP mutation ($p=3.9 \times 10^{-9}$). Intriguingly carriers of the T allele are at decreased risk of acquiring a *TET2* CHIP mutation (OR: 0.82, p=0.0012), and consequently it was not identified in the primary CHIP GWAS analysis. This variant is also associated with mosaic loss of chromosome Y.[18]

We evaluated whether our associations between germline loci and CHIP clones were robust across CHIP clone size spectrum, using the association between the *JAK2* 46/1 haplotype (tagged by rs1327494) and *JAK2* CHIP.[19] We find that rs1327494 associates with *JAK2* CHIP presence across VAF thresholds. We evaluated whether this observation generalized beyond JAK2 CHIP to encompass all CHIP. Intriguingly, we find that the TERT locus (tagged by rs7705526) is associated with CHIP presence across all VAF thresholds (Supplementary Table 9). These observations imply that our genetic associations are not dependent on clone size detectable by deep-coverage whole genome sequencing.

As single-variant analyses have limited power to detect rare-variant associations, we next performed several types of variant aggregation association tests. First, we performed a transcriptome-wide association analysis to quantify the relationship between changes in gene expression and genetic predisposition to CHIP[20] (see Methods). This approach identified the Chr3 *KPNA4/TRIM59* locus and six additional loci including: *AHRR*, *ASL*, *KREMN2*, *LEAP2*, *JSRP1*, *RASEF*. (Extended Data Fig. 7–8) *AHRR* directs hematopoietic progenitor cell expansion and differentiation.[21]

We also performed gene-based association tests for aggregations of rare (MAF<0.1%) putative loss-of-function (pLOF) germline variants within genes for CHIP presence. Although no genes reached exome-wide significance, the top associated gene was DNA damage repair gene *CHEK2* (OR 1.7, $p=1.3\times10^{-5}$, Supplementary Table 10). Rare germline variants in *CHEK2* are implicated in a diverse set of hematologic and solid tumor malignancies.[22,23] Common variants in *CHEK2* are associated with MPN[19] and a low frequency frameshift *CHEK2* is associated with somatic chromosomal mosaicism[9]. In recent experimental work, suppression of CHEK2 in human cord blood $Lin^{-}CD34^{+}$ cells increased cellular proliferation in long term culture. (Bao et al, co-submitted manuscript) These results suggest that while *CHEK2* while may ordinarily limit hematopoietic stem cell expansion, loss of *CHEK2* function may promote self-renewal increasing risk of CHIP.

We next sought to determine whether rare variants in non-coding regions associate with CHIP acquisition (see Methods). One set of variants in *HAPLN1* enhancers exceeded a p-value threshold of p<0.05 after Bonferroni-correction (OR: 6.8, $p=1.96 \times 10^{-5}$, Supplementary Table 11). *HAPLN1* is an extracellular matrix protein, produced in bone marrow stromal cells that has previously been implicated in NF-κB signaling.[24]

We asked whether germline genetic variants might be associated with CHIP clonal expansion. No single variants or aggregated rare variants exceeded Bonferroni significance (Supplementary Table 12–13).

## *TET2* CHIP risk locus characterization

Lastly, we bioinformatically and experimentally characterized the mechanism by which the non-coding African American-specific variant at the *TET2* locus influenced risk for CHIP. First, iterative conditional analyses 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*[25] and recapitulates the clonal expansion observed in humans with somatic mutations in *TET2*.[5,10] 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). 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 3a, Supplementary Table 14) The T risk allele disrupts a consensus GATA/E-Box motif, likely resulting in reduced binding of the activating transcription factors GATA1 and GATA2 (Figure 3b,c).

We then evaluated whether rs79901204 affected TET2 expression *in vivo* in human peripheral blood samples. We utilized whole blood RNAseq from 247 African American individuals, 16 of whom were heterozygotes for rs79901204 and one who was a homozygote. In these samples, the T risk allele led to a dose-dependent decrease in whole blood *TET2* expression (Beta: −0.27, SE: 0.11, two-sided linear mixed model p=0.012, Figure 3d). Therefore, we sought to test our hypothesis that that the rs79901204 risk allele acts to decrease the activity of this TET2 enhancer and that decreased enhancer activity reduces expression of *TET2* in vitro.

To test whether rs79901204 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 118-fold (versus control constructs with no enhancer sequence), while the T risk allele activated expression by only 27-fold (two-sided t-test p=0.007, Figure 3e).

To test whether deletion of this enhancer would alter *TET2* gene expression, we performed deletion of the enhancer element in CD34+ HSPCs using a pair of CRISPR/Cas9 guides introduced as ribonucleoproteins, which resulted in decreased *TET2* expression after 48 hours (Figure 3f).

We then sought to establish the effect of decreased *TET2* expression on HSPC expansion using a colony forming unit cellular assay. Human HSPCs were electroporated with Cas9 targeting a coding region of *TET2* and *AAVS1* (a control locus) and plated for primary and secondary colony-forming assays (Figure 3g). To establish a dose response relationship, two *TET2* guides were used with differential editing efficiency (Figure 3h, Extended Data Figure 9). *TET2* coding disruption resulted in expanded secondary colony formation compared to *AAVS1* controls, with greater expansion identified in the *TET2* guide with greater editing

efficiency (Figure 3i). Thus, we demonstrate that reduction of *TET2* activity promotes self-renewal and proliferation of HSPCs, illustrating how both germline noncoding and somatic coding variation at this locus converge to affect *TET2* and influence the development of CHIP.

Given the established role of *TET2* in DNA de-methylation and our finding that rs79901204 is associated with decreased *TET2* expression (Figure 3d), we hypothesized that carriers of rs79901204 T allele may have altered peripheral blood methylation profiles. We performed a methylation-QTL analysis of 1747 African Americans and identified 597 genes across the genome with differentially methylated CpG loci associated with rs79901204 T carrier status. The most strongly differentially methylated sites were at the *TET2* locus itself. (Extended Data Fig. 10, Supplementary Table 15)

Our observations permit several conclusions. First, our sample size which is nearly an order of magnitude larger than prior CHIP analyses[2,3,14] enables refinement of CHIP phenotype associations at the level of CHIP driver genes. We find that considerable heterogeneity exists across CHIP phenotypes by driver gene. For example, IL-1b and IL-18 both activate through the inflammasome and increase IL-6. However, while *TET2* CHIP is associated with increased levels of IL-1b, *JAK2* and *SF3B1* CHIP are associated with IL-18.

Second, 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 *CHEK2*) and which have been implicated in stem cell maintenance/self-renewal (*Bao et al,* co-submitted manuscript). 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. A third set of germline loci are associated with the acquisition of CHIP mutations in specific driver genes. This previously was described in the *JAK2* 46/1 haplotype leading to *JAK2* p.V617F via a *cis* haplotype effect.[26–28] We now identify a novel *DNMT3A* CHIP specific locus at the *TCL1A* promoter specifically associated with increased risk of *DNMT3A* CHIP, but not other CHIP subsets.

We identify a convergence of common and rare germline genetic predisposition to leukocyte telomere length, MPN, large scale somatic chromosomal mosaicism and CHIP, suggesting shared causal mechanisms. Importantly, to date, only CHIP with leukemogenic driver mutations (as opposed to somatic chromosomal mosaicism[9] or CHIP with unknown driver mutations[14]) has been robustly associated with non-oncologic diseases independently of age. The partially overlapping genetic predisposition we observe across these three clonal phenomena suggest that although there may be similar genetic architecture that predispose individuals to acquiring a somatic mutation, the specific change may be particularly relevant to atherosclerotic disease as opposed to the general phenomenon of clonal hematopoiesis itself.

Third, 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.[29] Furthermore, the cross-sectional analyses of CHIP with non-genetic risk factors and biomarkers limit conclusions regarding temporal relationships between CHIP and these features; however, these observations still permit risk prediction for CHIP presence. Notably, inflammatory biomarker analyses are concordant with prior model experiments indicating elevations of observed inflammatory biomarkers as a consequence of CHIP.[5,10] Lastly, given the age-dependence of CHIP, it is likely that many individuals not observed to have CHIP in this study will develop CHIP in the future.

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.

## Methods

### Study Samples

Whole genome sequencing (WGS) was performed on 97,691 samples sequenced as part of 52 studies contributing to the NHLBI TOPMed research program Freeze 6 release as previously described for discovery analyses.[7] An additional distinct set of 9,389 WGS sequenced samples from the NHLBI TOPMed Freeze 8 release were used for replicating the *TET2* genetic association. Study designs include prospective cohorts, families, population isolates, and case-only collections. A subset of the studies focus on heart (~40%) or lung (~30%) phenotypes, with the remainder representing prospective population cohorts or electronic health record linked cohorts which have been assessed for many diverse phenotypes. None of the studies which comprise TOPMed selected individuals for sequencing on the basis of hematologic malignancy. Approximately 82% of participants are U.S. residents with diverse ancestry and ethnicity (40% European, 32% African, 16% Hispanic/Latino, 10% Asian). Each of the constituent studies provided informed consent on the participating samples. Details on participating cohorts and samples is provided in Supplementary Table 1. The age of participants at time of blood draw was obtained for a subset of 82,807 of the samples. The median age was 55, the mean age 52.5, and the maximum age 98. The age distribution varied across the constituent cohorts.

Written informed consent was obtained from all human participants by each of the studies that contributed to TOPMed with approval of study protocols by ethics committees at participating institutions as summarized in Supplementary Table 1. Each study received institutional certification prior to deposition in dbGaP which certified that all relevant institutional ethics committees approved the individual studies and that the genomic and phenotypic data submission was compliant with all relevant ethical regulations. This certification was deposited in dbGaP along with the data. Secondary analysis of the TOPMed dbGaP data as described in this manuscript was approved by the Partners Healthcare Institutional Review Board. All relevant ethics committees approved this study and this work is compliant with all relevant ethical regulations.

## WGS Processing, Variant Calling and CHIP annotation

BAM files were remapped and harmonized through a previously described unified protocol.[30] SNPs and short indels were jointly discovered and genotyped across the TOPMed samples using the GotCloud pipeline.[31] An SVM filter was trained to discriminate between true variants and low-quality sites. Sample quality was assessed through pedigree errors, contamination estimates, and concordance between self-reported sex and genotype inferred sex. Variants were annotated using snpEff 4.3.

Putative somatic SNPs and short indels were called with GATK Mutect2[8] (https://software.broadinstitute.org/gatk). Briefly, Mutect2 searches for sites where there is evidence for variation, and then performs local reassembly. It uses an external reference of recurrent sequencing artifacts termed a “panel of normal samples” to filter out these sites, and calls variants at sites where there is evidence for somatic variation. The panel of normal samples used for our study included 100 randomly selected individuals under the age of 40 years. Absence of a hotspot CHIP mutation was verified prior to inclusion in the panel of normal set. An external reference of germline variants[32] was provided to filter out likely germline calls. We deployed this variant calling process on Google Cloud using Cromwell (https://github.com/broadinstitute/cromwell). The caller was run individually for each sample with the same settings. The Cromwell WDL configuration file is available from the authors upon request.

Samples were annotated as having CHIP if the Mutect2 output contained one or more of a pre-specified list of putative CHIP variants as previously described[2,5] (Supplementary Table 2) at a VAF >2%.

## WGS sensitivity to detect CHIP

To empirically demonstrate the sensitivity of CHIP detection and VAF, we re-analyzed sequence data from 30 samples with CHIP from a previously published cohort.[33] These samples were sequenced to >400x depth. We bioinformatically down-sampled the reads to the range of sequencing depths compatible with whole exome and whole genome sequencing. The TOPMed WGS samples were sequenced to a median depth of ~40x, although sequencing of any particular region was typically 30x-50x. Across this range of sequencing depths we observe robust ability to call CHIP with VAF >10%, which is the most clinically actionable subset of CHIP. We also capture approximately half of the CHIP

calls in the VAF 5–10% range. To reliably capture CHIP in the 5–10% range requires ~100x sequencing depth commonly done in whole exome sequencing, but even at this sequencing depth the majority of the VAF 2–5% CHIP calls are not reliably detected. (Extended Data Figure 11)

## Amplicon sequencing validation

To evaluate the fidelity of our TOPMed WGS CHIP dataset, we performed technical validation of 76 CHIP mutations in 72 samples using targeted deep sequencing. All 76 of 76 CHIP mutations identified with WGS were also identified with targeted deep sequencing. CHIP mutations were validated by single-molecule molecular inversion probe sequencing (smMIPS).[34] Capture probes were designed to tile all coding exons (+/− 5 bp) for 12 of the mostly highly prevalent CHIP genes plus four recurrent mutation hotspots, totaling 44.5 kb. Probes were synthesized as a pool by CustomArray, Inc., amplified using Q5 DNA polymerase (NEB) using outer flanking primers, and digested with BbsI-HF (NEB) to remove adaptors. For each sample, captures were performed with 500 ng gDNA and converted to dual-barcoded Illumina sequencing libraries as described.[35] Sequence capture libraries were pooled for paired-end 150 bp sequencing on a Hiseq 4000 lane. Resulting reads were aligned with bwa mem and processed using the mimips pipeline (source code at https://github.com/kitzmanlab/mimips) to trim capture probe sequences, and to remove reads with duplicated unique molecular identifiers. Somatic variants were called by MuTect2 as described above and confirmed by manual inspection with IGV.

## Somatic Chromosomal Mosaic Detection

In order to assess the relationship between CHIP and clonal mosaicism reflecting chromosomal mutation, we sought to characterize large (megabase-scale) acquired chromosomal alterations leading to allelic imbalance using existing SNP array data on a subset of the samples in this analysis. To do so, we compared statistically reconstructed haplotypes (using MaCH[36]) with the patterns of “B allele” frequencies (BAFs), measured via SNP array. Regions of nonrandom similarities between the estimated haplotypes and BAFs were detected with hapLOH[37], and indicate acquired chromosomal alterations. We identified genomic allelic imbalance events using a threshold of a posterior probability for allelic imbalance > 0.8 and event size > 1Mb. We excluded allelic imbalance events with fewer than ten markers and removed potential germline duplications if a detected event exhibited the following: 1) 50% reciprocal overlap with database of genomic variants (DGV) and 2) was not determined to be a deletion or LRR deviations > 0.08, size < 5Mb and BAF deviations > 0.1. Phasing and event detection was performed in SyQADA.[38]

## Blood traits

Conventionally measured blood cell counts and indices were selected for analysis including: hemoglobin, hematocrit, red blood cell count, white blood cell count, basophil count, eosinophil count, neutrophil count, lymphocyte count, monocyte count, platelet count, mean corpuscular hemoglobin, mean corpuscular hemoglobin concentration, mean corpuscular volume, mean platelet volume and red cell distribution width. Phenotypes were collected by each cohort, centrally harmonized by the TOPMed Data Coordinating Center (DCC). Additional documentation about harmonization algorithms for each specific trait is available

from the TOPMed DCC and accompanies the data on the dbGaP TOPMed Exchange area. Up to 37,653 individuals from 10 cohorts where utilized for this analysis that had one or more blood traits measured concurrently or following the blood draw used for CHIP ascertainment. Traits were first log2 normalized and then analyzed using a general linear regression model with CHIP status, age, sex, study and the first 10 ancestry principal components as covariates.

### Lipid phenotypes

Conventionally measured plasma lipids, including total cholesterol, LDL-C, HDL-C, and triglycerides, were included for analysis. LDL-C was either calculated by the Friedewald equation when triglycerides were <400 mg/dl or directly measured. Given the average effect of statins, when statins were present, total cholesterol was adjusted by dividing by 0.8 and LDL-C by dividing by 0.7. Triglycerides were natural log transformed for analysis. Phenotypes were harmonized by each cohort and deposited into dbGaP TOPMed Exchange area as previously described.[39] Up to 28,310 individuals from 19 cohorts where utilized for this analysis that had one or more lipid trait measured concurrently or following the blood draw used for CHIP ascertainment. Lipid traits were first normalized for age, sex and ancestry principal components and then analyzed using a general linear regression model with CHIP status, age, sex, study and the first 10 ancestry principal components as covariates.

### Inflammatory Markers

A set of makers previously implicated in mediating cardiometabolic disease were analyzed including: CD-40, CRP, E-Selectin, ICAM-1, IL-1b, IL-6, IL-10, IL-18, 8-epi PGF2a, Lp-PLA2 mass and activity, MCP1, MMP9, MPO, OPG, P-selectin, TNF-Alpha, TNF-Alpha Receptor 1, TNF-receptor 2. Phenotypes were collected by each cohort, centrally harmonized by the TOPMed DCC and then deposited into dbGaP TOPMed Exchange area. Additional documentation about harmonization algorithms for each specific trait is available from the TOPMed DCC and accompanies the data on dbGaP. Up to 22,092 individuals from 10 cohorts were utilized for this analysis that had one or more inflammatory marker measured concurrently or following the blood draw used for CHIP ascertainment. Inflammatory markers were first normalized using a log2(x+1) transformation and then analyzed using a general linear regression model with CHIP status, age, sex, study and the first 10 ancestry principal components as covariates.

### Mutational Signatures

We identified all putatively somatic singleton mutations in a subset of the TOPMed samples that included 3,764 cases with a single CHIP driver mutation and a randomly sampled set of 5,000 controls. Variants were filtered to ensure a depth >=25 reads, a VAF < 35% and no overlap with the germline variant site list from TOPMed Freeze 5 (available: https://bravo.sph.umich.edu/freeze5/hg38/). Multiallelic variants and indels were excluded. We used https://cancer.sanger.ac.uk/cosmic/signatures_v2 as a reference for mutation signatures and the MutationalPatterns R package to estimate the contributions of the signatures.[13,40,41] We defined a signature as being “differentially observed” if at least 99% of its observations

are in CHIP cases, or if at most 1% of its observations are in cases (i.e., one of cases or controls contains at least 99% of the signature observations).

## Single Variant Association

Single variant association for each variant in Freeze 8 with MAF > 0.1% and MAC > 20 was performed with SAIGE[42], 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[43] 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.

## Fine mapping

We applied FINEMAP 1.3[44] to the summary statistics from SAIGE, using the z-score and LD matrices as input. We fine-mapped the TET2 locus using the summary statistics from the African ancestry single variant summary statistics and estimated LD on the same set of samples using Plink 1.9. We set the maximum number of causal SNPs in the region to 10 and used a shotgun stochastic search.

## Transcriptome-wide association analysis

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)[20] 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 cutoff $2.9 \times 10^{-6}$.

### Rare Variant Analyses

Collapsing burden tests were applied to specific variant grouping schemes using EPACTS (https://genome.sph.umich.edu/wiki/EPACTS). The same covariates as the single variant tests were used on the same set of samples. We used burden tests due to their limited compute requirements, which were considerable for the number of variants and samples tested. Two grouping schemes were specified: the first groups coding variation, and the second groups putative regulatory elements in a relevant cell line. The first used all putative LOF variants as identified by snpEff. Given that some variants were present in both the Mutect2 calls and the germline variant calls, we pruned the LOF variants to exclude variants that were present in both call sets. The second grouping scheme used all variants in regions that were predicted enhancers for CD34 cells that had CADD scores of at least 10. Predicted enhancers were identified by the activity-by-contact model.[45]

### Predicting enhancer-gene regulation for TET2.

We used the Activity-by-Contact (ABC) model[46] to predict which enhancers regulate which genes in CD34+ hematopoietic progenitor cells, with minor modifications as follows.

Briefly, this model predicts the effect of each putative regulatory element (defined as a DNase peak within 5Mb of a given promoter) by multiplying the Activity of each element (estimated from DNase-seq and H3K27ac ChIP-seq) by its Contact with a target promoter (estimated from Hi-C data). The ABC score of a single element on a gene's expression is the predicted effect of that element divided by the sum of the predicted effects of all elements for a given gene.

We identified putative regulatory elements by using MACS2 to call peaks in DNase-seq data from mobilized CD34+ hematopoietic progenitor cells from the Roadmap Epigenome Project (downloaded from http://egg2.wustl.edu/roadmap/data/byFileType/alignments/consolidated/E050-DNase.tagAlign.gz) Initially we considered all peaks with p-value < 0.1. To further refine this list, we kept the 100,000 peaks with the highest number of DNase-seq reads. We then resized these peaks to be 500 bp in length centered on the peak summit, merging any overlapping peaks, and removed any peaks overlapping ENCODE "blacklisted regions"[47] (regions of the genome previously observed to accumulate anomalous numbers of reads in epigenetic sequencing experiments; downloaded from https://sites.google.com/site/anshulkundaje/projects/blacklists). To this peak list, we added 500 bp regions centered on the transcription start site of all genes. Any overlapping regions resulting from these additions or extensions were merged.

Within each putative regulatory element, we estimated enhancer Activity as the geometric mean of read counts from DNase-seq and H3K27ac ChIP-seq data from the Roadmap Epigenome Project (http://egg2.wustl.edu/roadmap/data/byFileType/alignments/consolidated/E050-DNase.tagAlign.gz, and E050-H3K27ac.tagAlign.gz).

We estimated enhancer-promoter Contact from the KR-normalized Hi-C contact maps in primary CD34+ cells. we then calculated effect of each putative enhancer-gene connection by multiplying the Activity and Contact for that element and gene. Dividing the effect of each element by the sum of effects for all elements for a given gene yields the ABC score:

$$\text{ABC score}_{E-G} = \frac{A_E \times C_{E-G}}{\sum_{e \text{ within 5 Mb}} A_e \times C_{e-G}}$$

To call predicted enhancer-gene connections, we used a threshold on the ABC score of 0.015. The rs79901204 variant overlapped an enhancer with ABC score of 0.0308 for TET2, which, based on comparison of ABC scores to large-scale enhancer perturbation datasets, corresponds to a positive predictive value of approximately 61%.

## Functional Evaluation of TET2 locus

The genomic region containing risk and non-risk allele of the variant rs79901204 (600bp) was synthesized as gblocks (IDT Technologies) and cloned into the Firefly luciferase reporter constructs (pGL4.24) using NheI and EcoRV sites. The Firefly constructs (500ng) were co-transfected with pRL-SV40 Renilla luciferase constructs (50ng) into 100,000 K562 cells (ATCC) using Lipofectamine LTX (Invitrogen) according to manufacturer's protocols. Cells were harvested after 48 hours and the luciferase activity measured by Dual-Glo Luciferase Assay system (Promega). K562 cell identity was validated using STR analysis. Mycoplasma testing was routinely performed on all cells used in the study and confirmed to test negative.

## CRISPR/Cas9 editing of CD34$^+$ human HSPCs

Editing of *TET2* enhancer and *TET2* coding regions was performed by electroporation of Cas9 Ribonucleoprotein complex (RNP) into CD34$^+$ human HSPCs. CD34+ HSPCs from adult donors obtained from the Fred Hutchinson Cancer Research Center, Seattle, USA were thawed 24 hours prior to electroporation and cultured in HSC expansion conditions throughout the experiment (Stemspan II media with CC100 cytokine cocktail from Stem Cell Technologies and TPO (50ng/ul) and small molecule UM171 (35nM)). The RNP complex was made by mixing Cas9 (50 pmol) and modified sgRNAs from Synthego (100 pmol in total). HSPCs (3.75 × 10 5) resuspended in 20 μl P3 solution were mixed with RNP and transferred to Nucleocuvette strips for electroporation with program DZ-100 (Lonza 4D Nucleofector). *TET2* gene expression was measured at 6 days post-electroporation.

For enhancer deletion experiments two guides targeting 5' and 3' ends of the enhancer element was used simultaneously (ENH_sgRNA_1: GGATTCTGTATTCGTCTGTG & ENH_sgRNA_2: TCTACTCACAGGGCCCAATG). For *TET2* coding disruption experiments single guides were used (TET2_CDS1: TGGAGAAAGACGTAACTTCG & TET2_CDS2: TCTGCCCTGAGGTATGCGAT). For negative control, a guide targeting *AAVS1* site was used (GGGGCCACTAGGGACAGGAT). Editing efficiency of *TET2* CDS and AAVS1 guides were measured by Sanger sequencing followed by TIDE analysis. Editing efficiency of *TET2* enhancer deletion was measured by PCR and agarose gel electrophoresis.

## Colony-forming unit cell assays

3 days post RNP-electroporation, 500 CD34+ HSPCs were plated in 1ml methylcellulose media (# H4034, Stem Cell Technologies). Primary CFU-C colonies were counted after 14

days. For the colony replating experiments, 2 weeks after the primary plating, the colonies from three pates were pooled, washed with PBS, and the cells were plated in new methylcellulose media at 25,000 cells/ml for an additional 2 weeks.

### RNA-Sequencing and eQTL Analysis:

RNA-Sequencing was performed on peripheral blood mononuclear cells from a subset of the MESA cohort. Alignment to the GRCh38 reference genome was done using STAR 2.5.3a.[48] Gene Quantification and quality control was performed using RNA-SeQC 1.19.[49] 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 available here: https://www.nhlbiwgs.org/sites/default/files/TOPMed_RNAseq_pipeline_COREyr2.pdf

Analysis was performed using samples from 247 African Americans from MESA cohort Exam 1. Transcript expression was converted to TPM units (transcripts per million) and log2-transformed for analysis consistent with the GTEx consortium[50] best practices. 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.

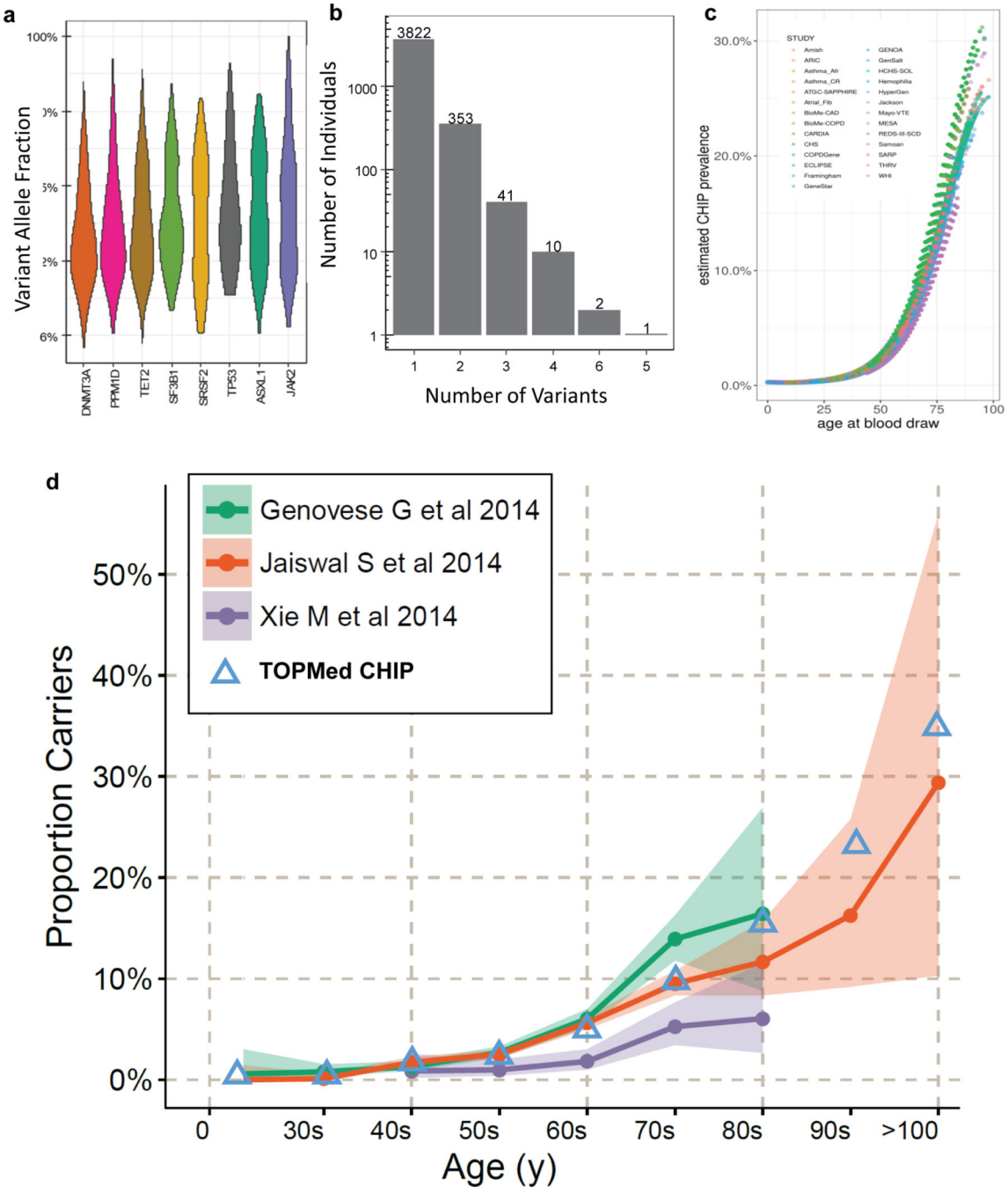
### Genome-wide Methylation-QTL analysis of TET2 risk locus

Illumina Methylation EPIC 850K array data interrogating over 850,000 CpG DNA methylation sites was generated at the University of Washington's Northwest Genomic Center from blood samples collected from African Americans at the Jackson Heart Study baseline exam. Fluorescent signal intensities were preprocessed with the R package *minfi*[51] using the normal-exponential out-of-band (noob) background correction method with dye-bias normalization. N = 1747 total samples (1097 women and 650 men) remained after severe outliers were identified and removed. 71 individuals were positive for CHIP and 100 were carriers of the rs79901204 variant.

Methylation levels at each CpG site were then quantified as β values, defined as the ratio of intensities between methylated (M) and unmethylated (U) signals where $\beta = M/(M+U+100)$. Values therefore ranged from $\beta = 0$ (completely unmethylated) to $\beta = 1$ (completely methylated). Batch correction for assay plate position was performed on the β values via *ComBat*.[52] Relative leukocyte cell counts (CD8+ T-lymphocytes, CD4+ T-lymphocytes, Natural Killer cells, B cells, Monocytes, and Granulocytes) were estimated as previously described by Houseman[53] and Horvath[52].

To investigate methylation in the *TET2* locus, a linear mixed effects model was fitted using CpGassoc[53] in R 3.6.0 with rs79901204 as the predictor and the batch-corrected methylation β levels as the dependent variable, adjusting for age, sex, estimated cell counts, the top 10 principal components of genetic ancestry, and CHIP status. A Bonferroni corrected threshold of $p = 5.8\times10^{-8}$ was used to establish statistical significance.

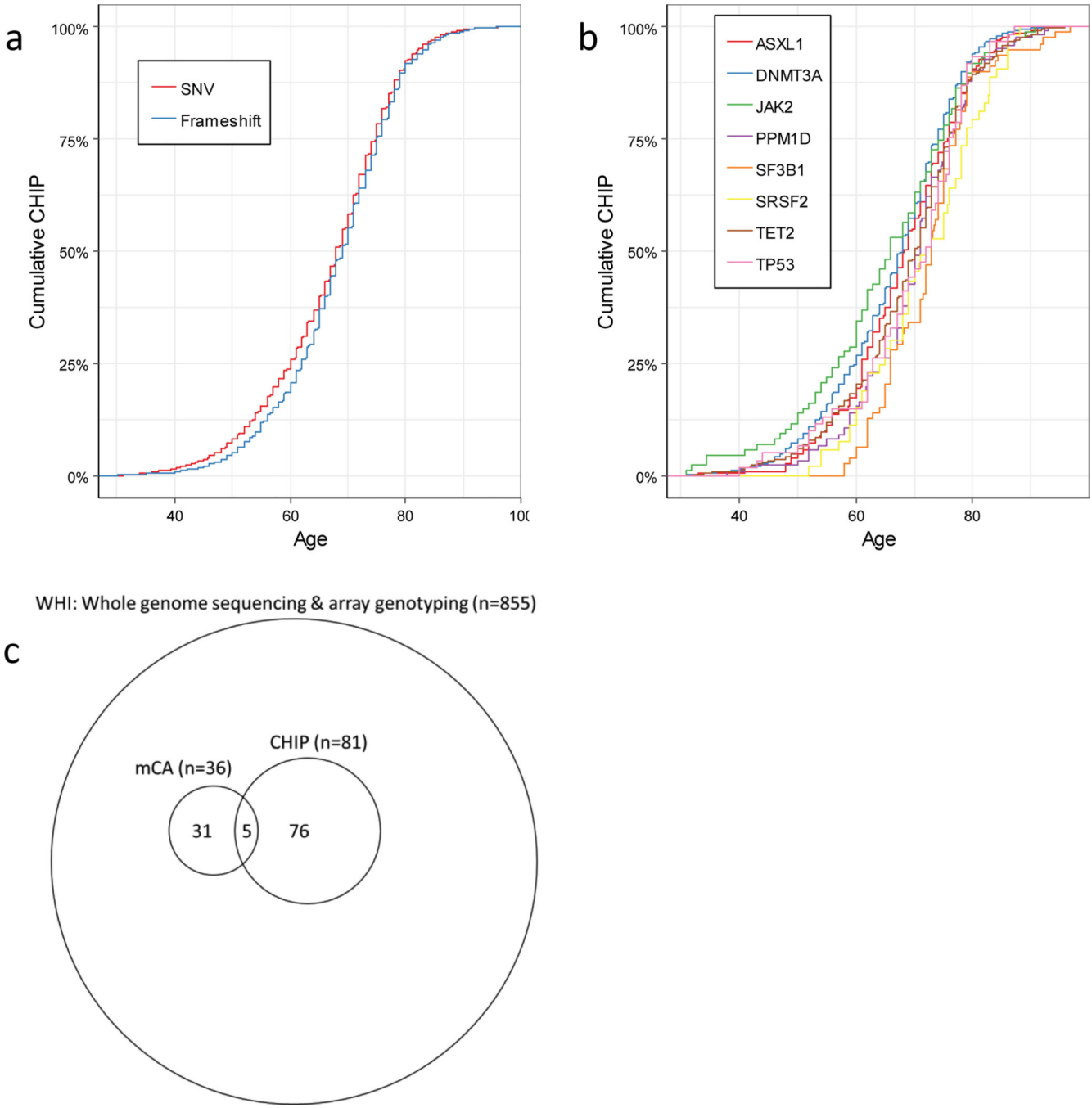
## Extended Data




**Extended Data Fig. 1|. Characterizing TOPMed CHIP.**

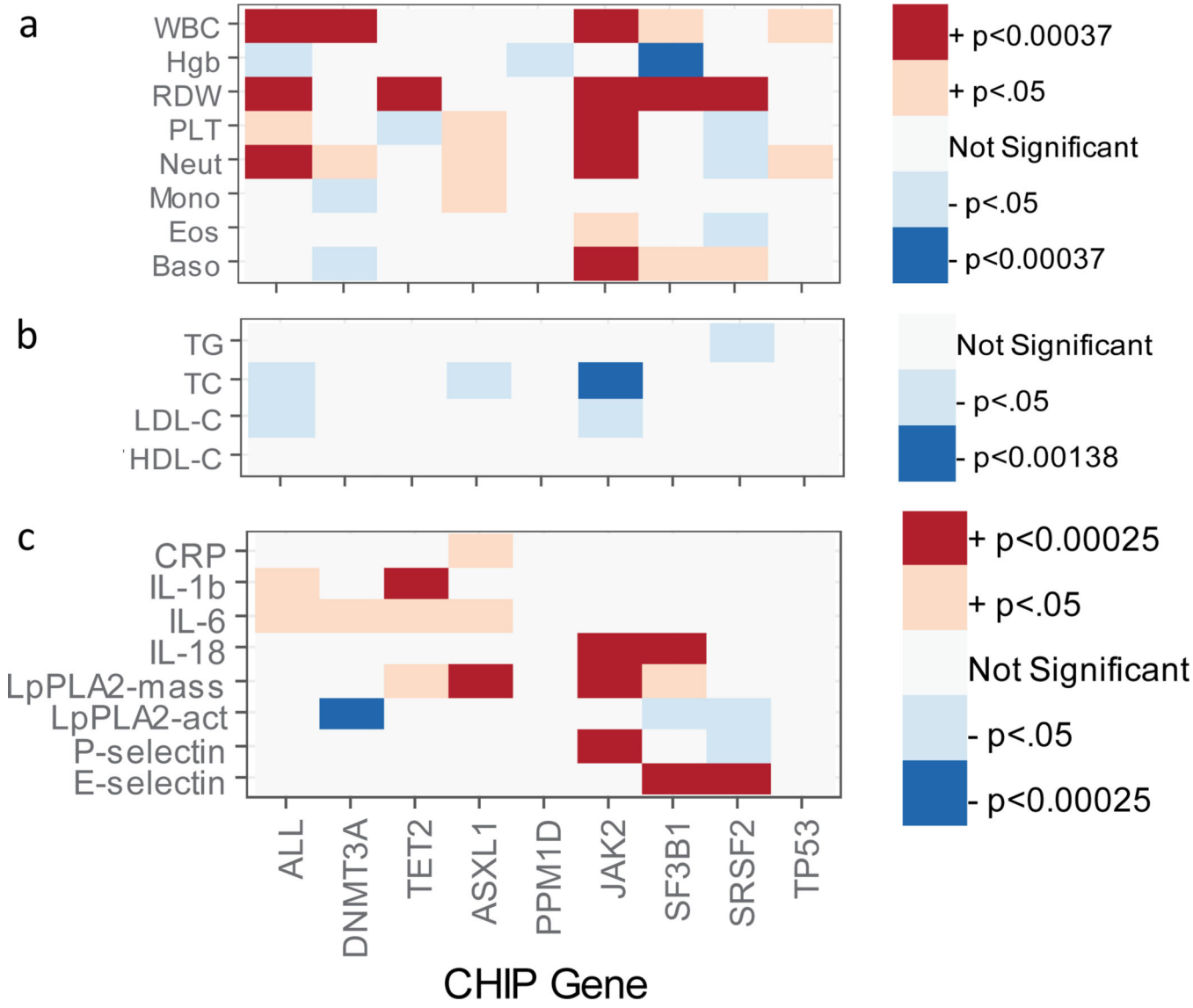
**a**, There was marked heterogeneity of CHIP clone size as measured by variant allele fraction by CHIP driver gene. Violin plot spanning minimum and maximum values calculated on full dataset (Supplementary Table 3). Sample size for each element in violin plot displayed in Fig. 1, **b,** 90% of individuals with CHIP had only one CHIP driver mutation identified **c,** CHIP prevalence with age was highly concordant across sequenced cohorts. CHIP

prevalence was estimated from a logistic mixed model with spline-transformed age, sex, and cohort included as predictors. The cohort was included as a random intercept. Sample size for each cohort listed in Supplementary Table 1. **d,** CHIP prevalence with age in this study (blue triangles, N=82,807) was highly consistent with previously observed CHIP prevalence (dots represent mean point prevalence with shaded area represents 95% confidence interval; $N_{Genovese}$=12,380; $N_{Jaiswal}$ = 17,182; $N_{Xie}$ = 2,728).



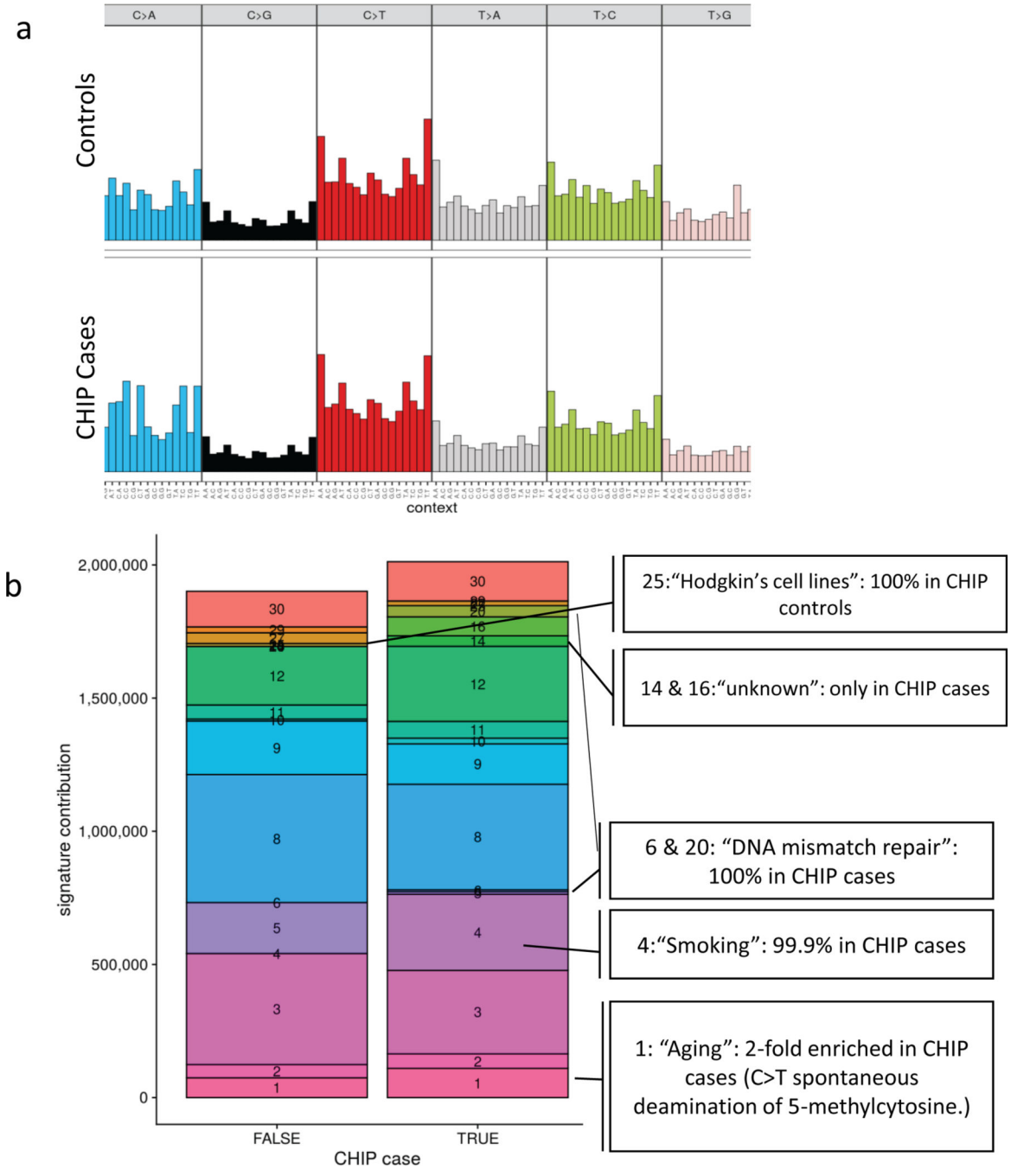

**Extended Data Fig. 2|. CHIP age association by mutational mechanism, gene and overlap with somatic chromosomal mosaicism.**

**a,** cumulative density plot of CHIP incidence with age stratified by single nucleotide variant (SNV) vs frameshift mutations. SNVs were observed in younger individuals than Frameshift mutations (N=4,939; two-sided wilcox rank sum test p=0.01). **b,** cumulative density plot of CHIP incidence with age stratified by driver gene. **c,** 855 elderly WHI individuals (mean age: 70) with both whole genome and the array genotyping data available were interrogated for large-scale mosaic chromosomal rearrangements. The two somatic events did not co-occur more than would be expected by chance (hypergeometric p=0.25).



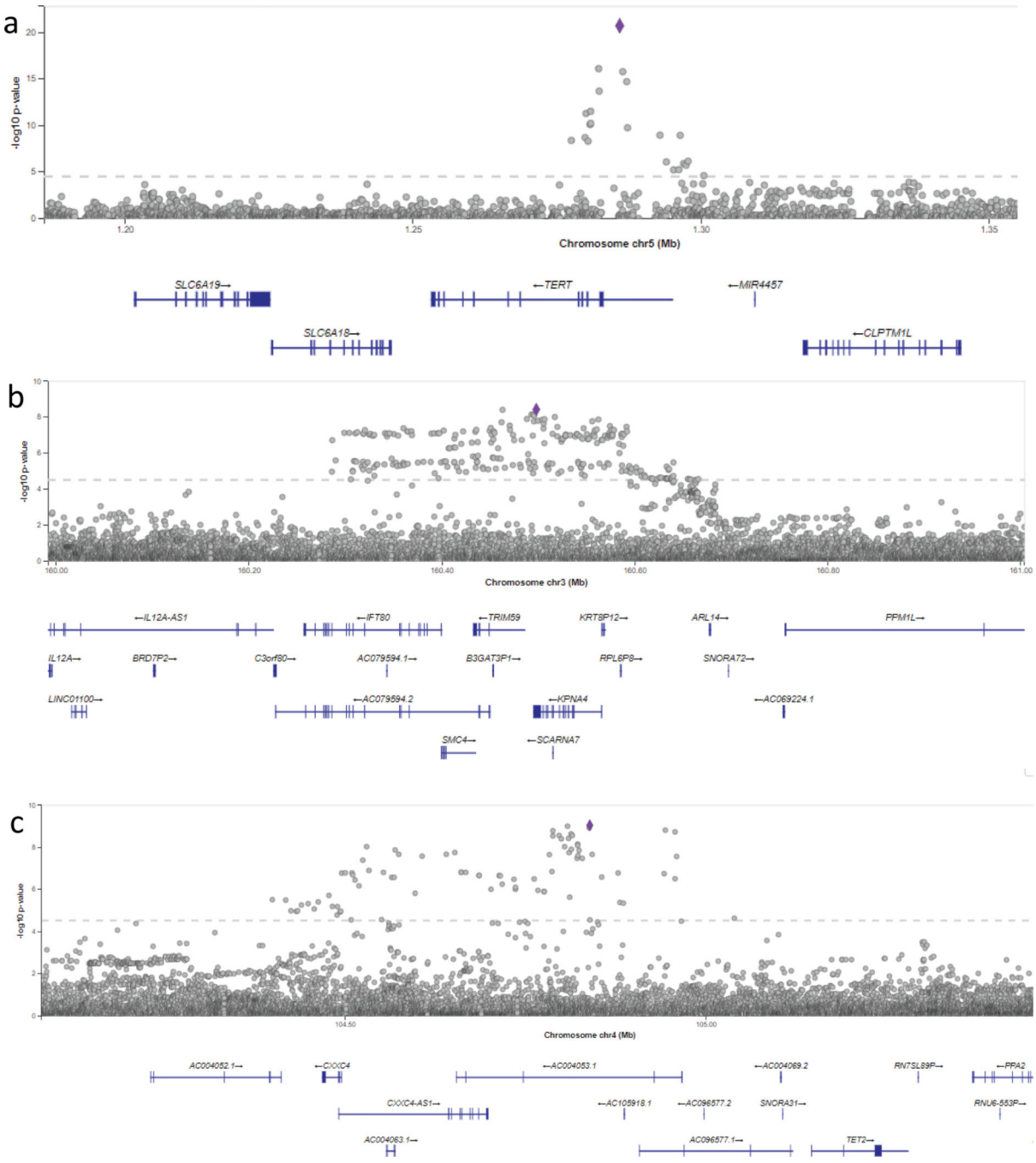

**Extended Data Fig. 3|. CHIP associates with Blood, Lipid, and Inflammatory traits.**
**a,** CHIP consistently associated with increased Red Cell Distribution Width (RDW). *JAK2*, *SF3B1* and *SRSF2* showed driver gene specific effects on blood traits (see Supplementary Table S5) **b,** CHIP status was not consistently associated with lipid traits, other than *JAK2* CHIP which was associated with decreased total cholesterol and a trend towards decreased LDL (see Supplementary Table S6) **c,** CHIP status is associated with inflammatory markers,

however notable heterogeneity existed across CHIP mutations (see Supplementary Table S7). Associations utilized a two-sided t-test from a multivariate general linear model including age, smoking, race and gender and study center and were not adjusted for multiple comparisons. Sample sizes and exact p-values for each phenotype are listed in Supplementary Tables 5–7.

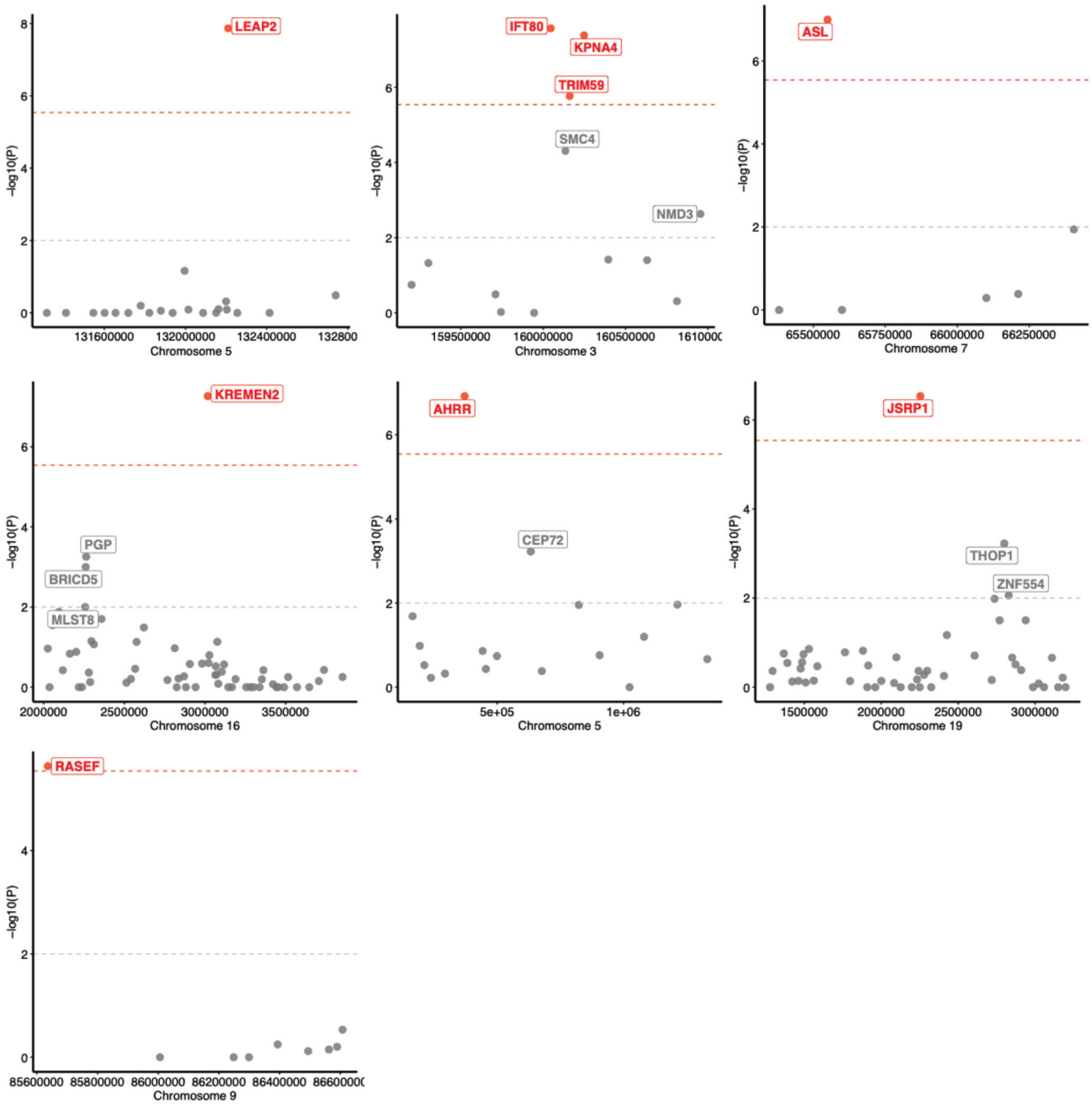



**Extended Data Fig. 4|. CHIP passenger somatic mutation spectrum.**

**a,** Singleton mutation counts by nucleotide context in CHIP Cases and Controls. **b,** Signature contribution in CHIP cases and controls identified differential enrichment

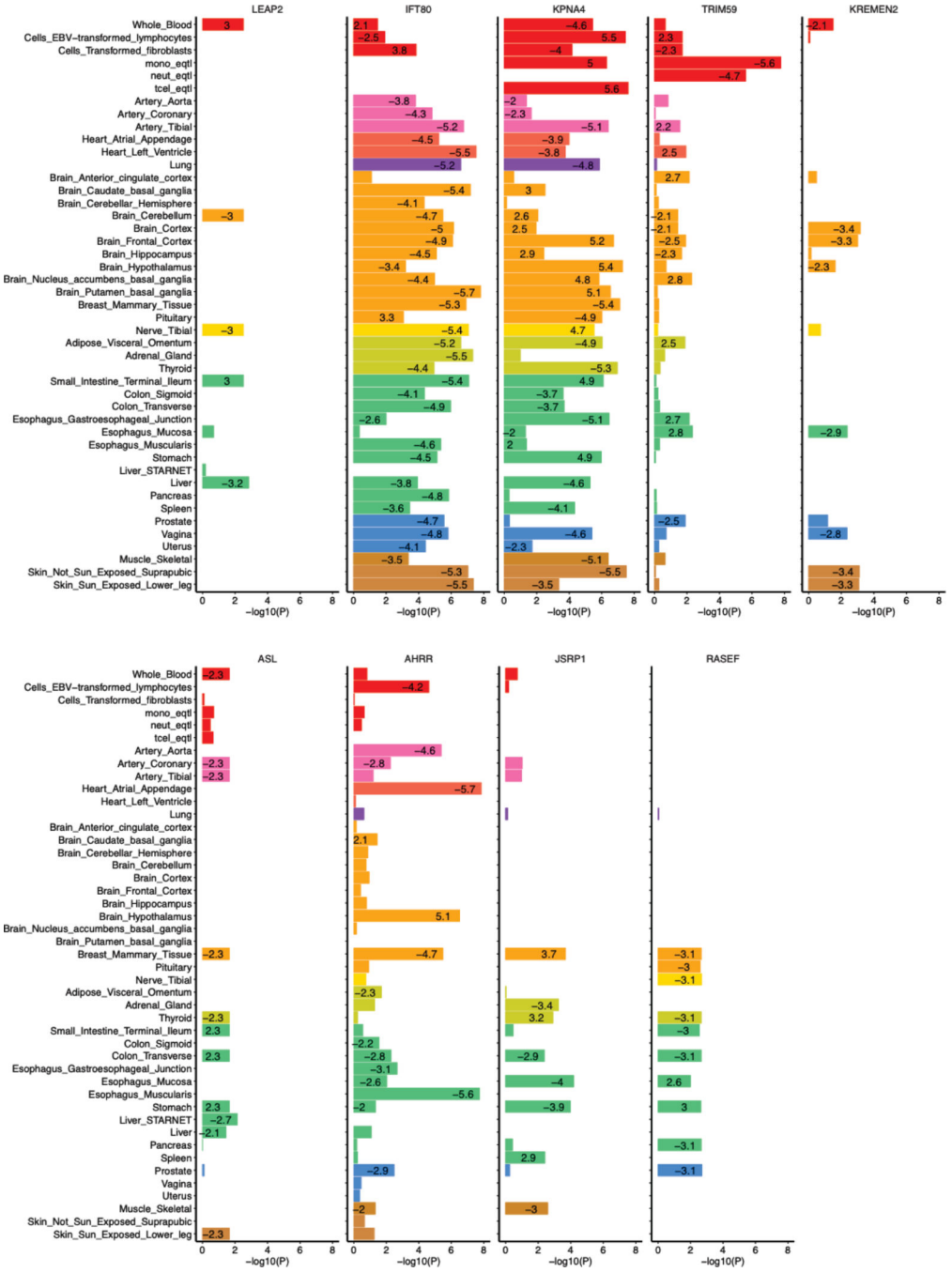



**Extended Data Fig. 5|. CHIP Single variant association regional association plots.**
**a,** TERT locus **b,** TRIM59/KPNA4 locus **c,** TET2 locus. Two-sided association testing performed using SAIGE (N=65,405 individuals, see methods)

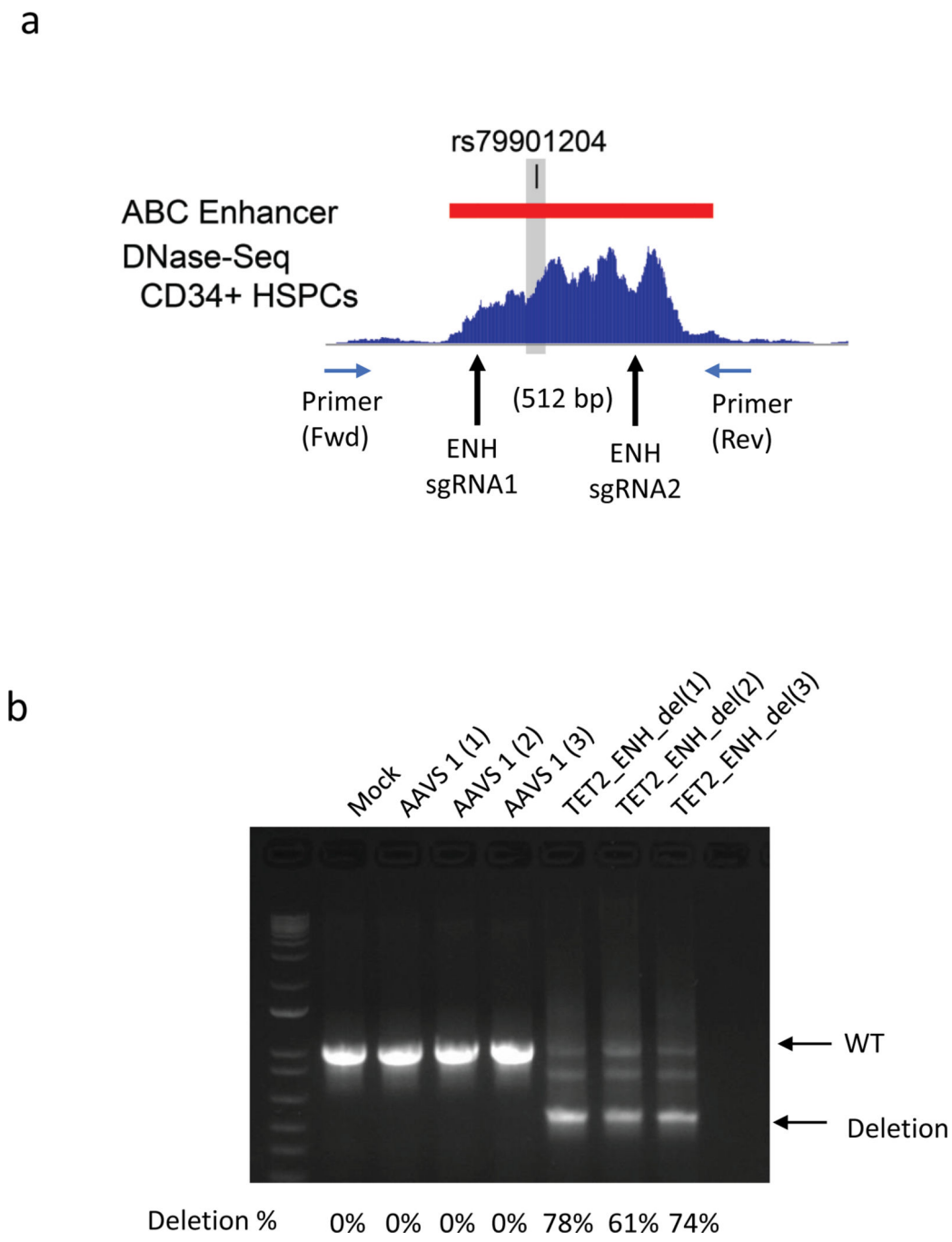



**Extended Data Fig. 6|. CHIP transcriptome-wide association study (TWAS) results across 48 tissues identified 7 significant loci.**

UTMOST algorithm applied to CHIP genome wide association study results from n=65,405 individuals (see methods). Genomic coordinates listed on x-axis. P-value from generalized Berk-Jones test on Y axis. Multiple hypothesis corrected threshold, $p<2.9 \times 10^{-6}$ displayed as dotted red line.

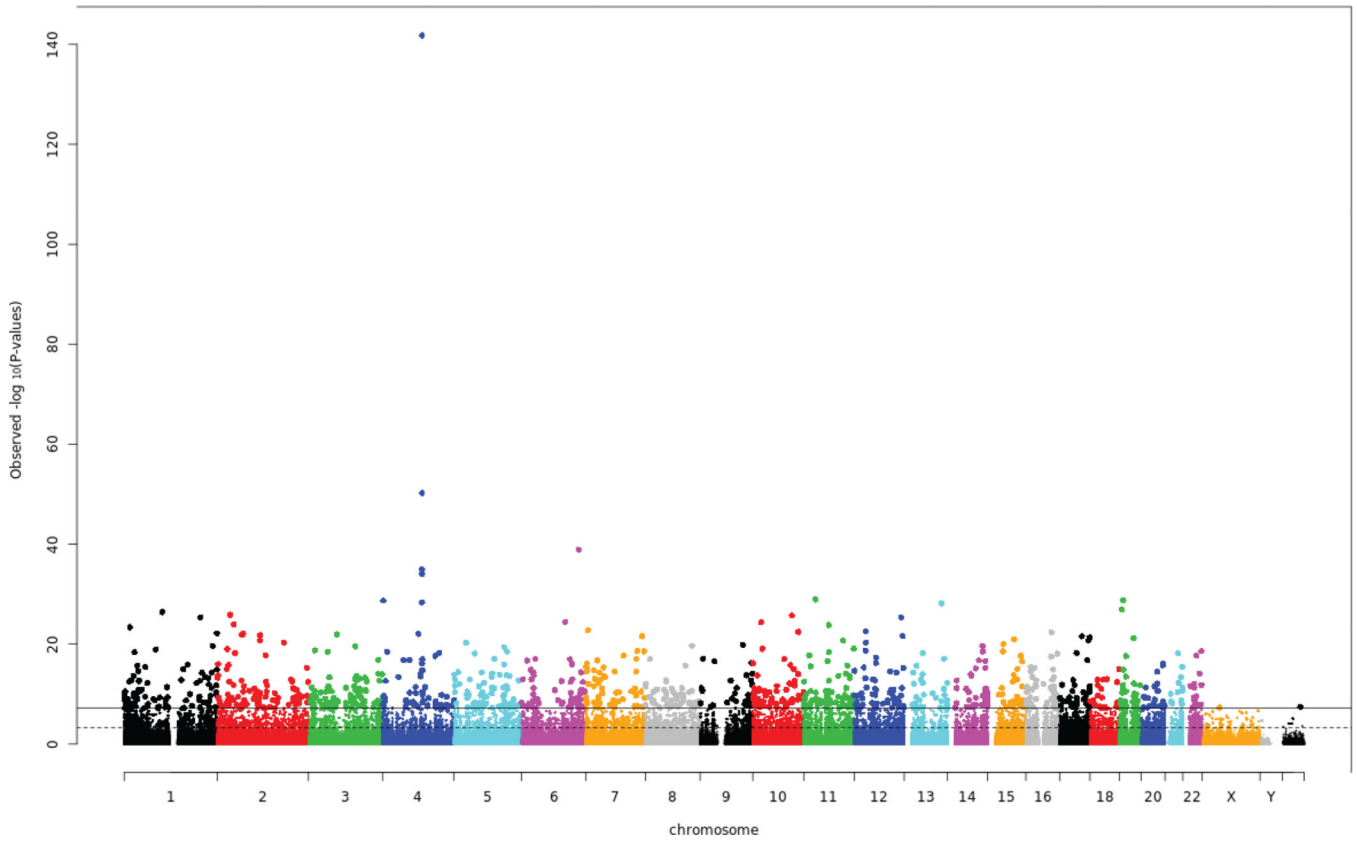



**Extended Data Fig. 7|. Tissue-specific results from the top 9 overall UTMOST-significant genes.** UTMOST algorithm applied to CHIP genome wide association study results from n=65,405 individuals. P-value from generalized Berk-Jones test. eQTL z-scores for associations with P<0.05 are displayed in each bar. GTEX eQTL tissue listed on Y-axis.

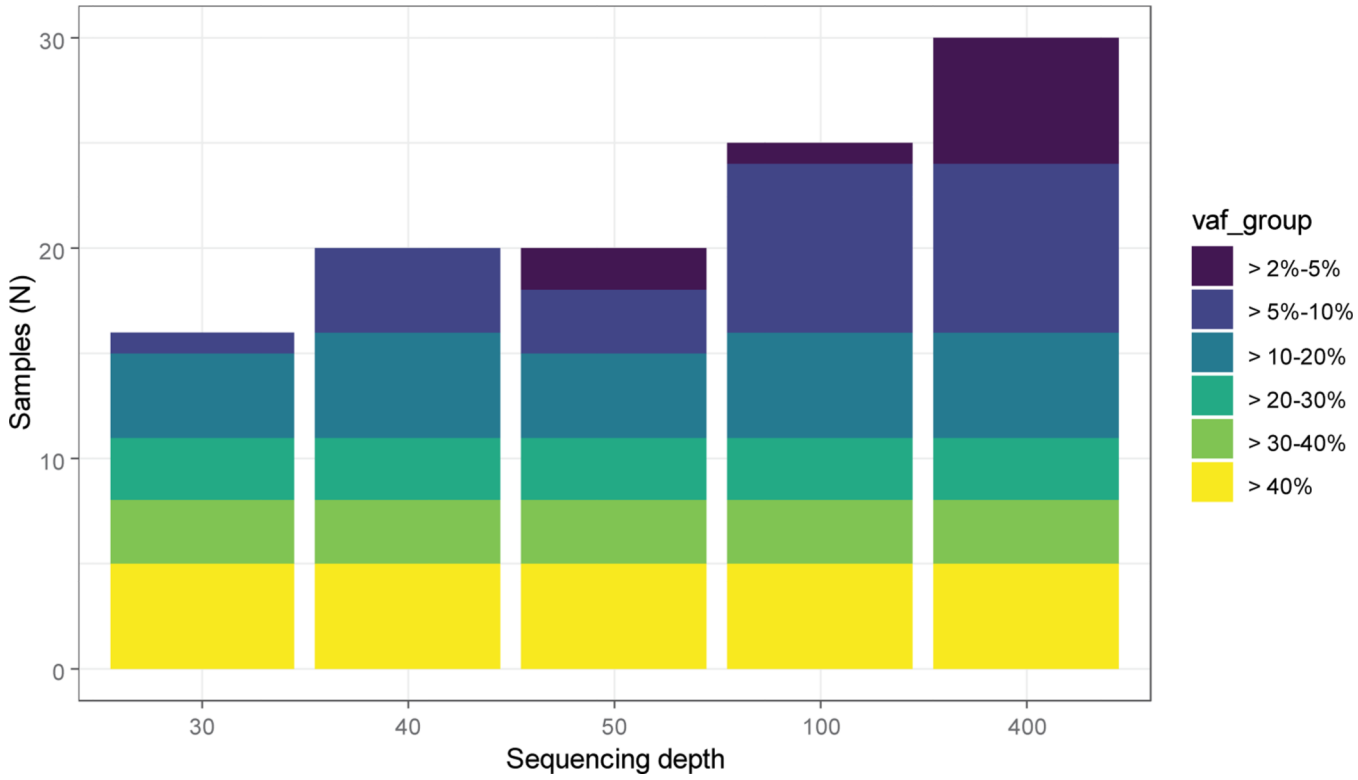



**Extended Data Fig. 8 |. CRISPR/Cas9 editing efficiency of TET2 Enhancer deletion in primary CD34+ HSPCs.**

a, Schematic showing the position of the two sgRNAs used to delete the TET2 enhancer (512bp) containing rs79901204. B, Gel electrophoresis image of PCR products from genomic DNA of edited HSPCs indicating unedited (WT) and deletion bands at sgRNA target site. Percentages of deletion alleles determined by band intensity and is shown below each lane. The experiment contains 3 biological replicates and was performed once.




**Extended Data Fig. 9 |. rs79901204 associated with genome wide differential methylation signal,** Methylation Quantitative Trait association results of rs79901204 variant with cpg methylation probes identify an altered peripheral leukocyte methylation profile genome wide in N = 1747 individuals. The strongest signal is at the chr4 TET2 locus. P-values on Y-axis derived from two-sided linear mixed effects model (see methods). To account for multiple hypothesis testing, a Bonferroni threshold of $p < 5.8 \times 10^{-8}$ was used to establish statistical significance.




**Extended Data Fig. 10 |. Sensitivity of CHIP detection at various VAFs across sequencing depths.** A set of 30 samples from a previously published CHIP cohort (Gibbons et al, 2017) were computationally down sampled to 30x, 40x, 50x, 100x and 400x sequencing depth. TOPMed WGS data was typically in the 40x depth range across CHIP genes. WGS data has excellent sensitivity to detect CHIP clones with VAF >10%, and ~50% sensitivity to detect CHIP VAF 5–10%, with minimal ability to detect CHIP clones <5%.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Authors

Alexander G. Bick[1,2,3,*,‡], Joshua S. Weinstock[4,*], Satish K. Nandakumar[2,5], Charles P. Fulco[2,6], Erik L. Bao[2,5,7], Seyedeh M. Zekavat[2,8], Mindy D. Szeto[9,10], Xiaotian Liao[2,5], Matthew J. Leventhal[2], Joseph Nasser[2], Kyle Chang[11], Cecelia Laurie[12], Bala Bharathi Burugula[13], Christopher J. Gibson[14], Amy E. Lin[15], Margaret A. Taub[16], Francois Aguet[2], Kristin Ardlie[2], Braxton D. Mitchell[17,18], Kathleen C. Barnes[9,19], Arden Moscati[20], Myriam Fornage[21,22], Susan Redline[3,23,24], Bruce M. Psaty[25,26,27,28], Edwin K. Silverman[3,29], Scott T. Weiss[3,29], Nicholette D. Palmer[30], Ramachandran S. Vasan[31], Esteban G. Burchard[31,32], Sharon L. R. Kardia[33], Jiang He[34,35], Robert C. Kaplan[36,37], Nicholas L. Smith[26,28,38], Donna K. Arnett[39], David A. Schwartz[40], Adolfo Correa[41], Mariza de Andrade[42], Xiuqing Guo[43], Barbara A. Konkle[44,45], Brian Custer[46,47], Juan M.

Peralta[48], Hongsheng Gui[49], Deborah A. Meyers[50], Stephen T. McGarvey[51], Ida Yii-Der. Chen[52], M. Benjamin Shoemaker[53], Patricia A. Peyser[33], Jai G. Broome[12], Stephanie M. Gogarten[12], Fei Fei Wang[12], Quenna Wong[12], May E. Montasser[17], Michelle Daya[9], Eimear E. Kenny[54], Kari E. North[55], Lenore J. Launer[56], Brian E. Cade[23,57], Joshua C. Bis[25], Michael H. Cho[29], Jessica Lasky-Su[3,29], Donald W. Bowden[30], L. Adrienne. Cupples[58], Angel C. Mak[32], Lewis C. Becker[59], Jennifer A. Smith[33,60], Tanika N. Kelly[34,35], Stella Aslibekyan[61], Susan R. Heckbert[26,28], Hemant K. Tiwari[62], Ivana V. Yang[40], John A. Heit[63], Steven Lubitz[2,3,64], Jill M. Johnsen[44,45], Joanne E. Curran[48], Sally E. Wenzel[65], Daniel E. Weeks[66], Dabeeru C. Rao[67], Dawood Darbar[68], Jee-Young Moon[36], Russell P. Tracy[69], Erin J. Buth[12], Nicholas Rafaels[19], Ruth J.F. Loos[20,70], Peter Durda[69], Yongmei Liu[71], Lifang Hou[72], Jiwon Lee[23], Priyadarshini Kachroo[3,29], Barry I. Freedman[73], Daniel Levy[74,75], Lawrence F. Bielak[33], James E. Hixson[76], James S. Floyd[77], Eric A. Whitsel[78,79], Patrick T. Ellinor[2,3,64], Marguerite R. Irvin[61], Tasha E. Fingerlin[80], Laura M. Raffield[81], Sebastian M. Armasu[82], Marsha M. Wheeler[83], Ester C. Sabino[84], John Blangero[48], L. Keoki Williams[49], Bruce D. Levy[3,85], Wayne Huey-Herng Sheu[86], Dan M. Roden[87], Eric Boerwinkle[88], JoAnn E. Manson[3,89,90], Rasika A. Mathias[58], Pinkal Desai[91], Kent D. Taylor[92], Andrew D. Johnson[74,75], NHLBI Trans-Omics for Precision Medicine Consortium[†], Paul L. Auer[93], Charles Kooperberg[94], Cathy C. Laurie[12], Thomas W. Blackwell[4], Albert V. Smith[4], Hongyu Zhao[95,96], Ethan Lange[9], Leslie Lange[9], Stephen S. Rich[97], Jerome I. Rotter[92], James G. Wilson[98,99], Paul Scheet[11], Jacob O. Kitzman[13,100], Eric S. Lander[2,101,102], Jesse M. Engreitz[2,103], Benjamin L. Ebert[2,3,14,104], Alexander P. Reiner[26,94], Siddhartha Jaiswal[105], Gonçalo Abecasis[4,106], Vijay G. Sankaran[2,3,5], Sekar Kathiresan[2,3,107,108,#], Pradeep Natarajan[2,3,64,#]

## Affiliations

[1]Department of Medicine, Massachusetts General Hospital, Boston, MA.

[2]Broad Institute of MIT and Harvard, Cambridge, MA.

[3]Harvard Medical School, Boston MA.

[4]Center for Statistical Genetics, Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI.

[5]Division of Hematology/Oncology, Boston Children's Hospital and Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston MA.

[6]Department of Systems Biology, Harvard Medical School, Boston, MA.

[7]Health Sciences and Technology Program, Harvard Medical School, Boston, MA.

[8]Yale School of Medicine, New Haven, CT.

[9]Division of Biomedical Informatics and Personalized Medicine, Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO.

[10]Medical Scientist Training Program, University of Colorado Anschutz Medical Campus, Aurora, CO.

[11]Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, TX.

[12]Department of Biostatistics, University of Washington, Seattle, WA.

[13]Department of Human Genetics, University of Michigan, Ann Arbor, MI.

[14]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA.

[15]Division of Cardiovascular Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA.

[16]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD.

[17]Department of Medicine, University of Maryland School of Medicine, Baltimore, MD.

[18]Geriatrics Research and Education Clinical Center, Baltimore Veterans Administration Medical Center, Baltimore, MD.

[19]Colorado Center for Personalized Medicine, School of Medicine, University of Colorado, Aurora, CO.

[20]Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY.

[21]Brown Foundation Institute of Molecular Medicine, McGovern Medical School, University of Texas Health Science Center at Houston, Houston, TX.

[22]Human Genetics Center, School of Public Health, University of Texas health Science Center at Houston, Houston, TX.

[23]Division of Sleep and Circadian Disorders, Department of Medicine, Brigham and Women's Hospital, Boston, MA.

[24]Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA.

[25]Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA.

[26]Department of Epidemiology, University of Washington, Seattle, WA.

[27]Department of Health Services, University of Washington, Seattle, WA.

[28]Kaiser Permanente Washington Health Research Institute, Seattle, WA.

[29]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA.

[30]Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC.

[31]Departments of Medicine and Epidemiology, Boston University School of Medicine, Boston, MA.

[32]Department of Medicine, University of California, San Francisco, CA.

[33]Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI.

[34]Department of Epidemiology, Tulane University School of Public Health and Tropical Medicine, New Orleans, LA.

[35]Tulane University Translational Science Institute, New Orleans, LA.

[36]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY.

[37]Fred Hutchinson Cancer Research Center, Division of Public Health Sciences, Seattle WA.

[38]Seattle Epidemiologic Information and Research Center, Department of Veterans Affairs, Office of Research & Development, Seattle WA.

[39]College of Public Health, University of Kentucky, Lexington KY.

[40]Department of Medicine, University of Colorado, Aurora, CO.

[41]Departments of Medicine and Population Health Science, University of Mississippi Medical Center, Jackson, MS.

[42]Department of Health Sciences Research, Mayo Clinic, Rochester, MN.

[43]Institute for Translational Genomics and Population Sciences, Department of Pediatrics, Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center, Torrance, CA.

[44]Bloodworks Northwest, Seattle, WA.

[45]Department of Medicine, University of Washington, Seattle, WA.

[46]Vitalant Research Institute, San Francisco, CA.

[47]University of California at San Francisco, San Francisco, CA.

[48]Department of Human Genetics and South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley School of Medicine, Brownsville, TX.

[49]Center for Individualized and Genomic Medicine Research, Department of Internal Medicine, Henry Ford Health System, Detroit, MI.

[50]Division of Genetics, Genomics and Precision Medicine, University of Arizona, Tucson, AZ.

[51]Department of Epidemiology and International Health Institute, Brown University School of Public Health, Providence.

[52]Medical Genetics, Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center, Los Angeles, CA.

[53]Division of Cardiology, Vanderbilt University Medical Center, Nashville, TN.

[54]Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY.

[55]Department of Epidemiology, University of North Carolina, Chapel Hill, NC.

[56]Laboratory of Epidemiology, Demography, and Biometry, Intramural Research Program, National Institute on Aging, Bethesda, MD.

[57]Division of Sleep Medicine, Harvard Medical School, Boston, MA.

[58]Departments of Biostatistics and Epidemiology, Boston University School of Medicine, Boston, MA.

[59]GeneSTAR Research Program, Department of Medicine, Johns Hopkins School of Medicine, Baltimore MD.

[60]Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI.

[61]Department of Epidemiology, University of Alabama at Birmingham, Birmingham, AL.

[62]Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL.

[63]Division of Cardiovascular Diseases, Mayo Clinic, Rochester, MN.

[64]Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA.

[65]Department of Environmental and Occupational Health, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA.

[66]Departments of Human Genetics and Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA.

[67]Division of Biostatistics, Washington University School of Medicine, St. Louis, MO.

[68]Division of Cardiology, University of Illinois at Chicago, Chicago, IL.

[69]Department of Pathology and Laboratory Medicine, Larner College of Medicine, University of Vermont, Burlington, VT.

[70]Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY.

[71]Division of Cardiology, Department of Medicine, Duke University Medical Center, Durham, NC.

[72]Department of Preventive Medicine, Northwestern Feinberg School of Medicine, Northwestern University, Chicago, IL.

[73]Department of Internal Medicine, Section on Nephrology, Wake Forest School of Medicine, Winston-Salem, NC.

[74]Framingham Heart Study, Framingham, MA.

[75]Population Sciences Branch, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD.

[76]Department of Epidemiology, Human Genetics and Environmental Sciences, University of Texas Health Science Center at Houston School of Public Health, Houston, TX.

[77]Departments of Medicine and Epidemiology, University of Washington, Seattle, WA.

[78]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC.

[79]Department of Medicine, School of Medicine, University of North Carolina, Chapel Hill, NC.

[80]Center for Genes Environment and Health, National Jewish Health, Denver, CO.

[81]Department of Genetics, University of North Carolina, Chapel Hill, NC.

[82]Department of Health Sciences Research, Mayo Clinic, Rochester, MN.

[83]Department of Genome Science, University of Washington, Seattle, WA.

[84]Instituto de Medicina Tropical, Faculdade de Medicina da Universidade de Sao Paulo, Sao Paulo, Brazil.

[85]Division of Pulmonary and Critical Care Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA.

[86]Division of Endocrinology and Metabolism, Department of Internal Medicine, Taichung Veterans General Hospital, Taichung, Taiwan.

[87]Departments of Medicine, Pharmacology, and Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN.

[88]Human Genetics Center, University of Texas Health Science Center at Houston and Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX.

[89]Department of Medicine, Brigham and Women's Hospital, Boston, MA.

[90]Department of Epidemiology, Harvard School of Public Health, Boston, MA.

[91]Department of Medicine, Weill Cornell Medical School, New York, NY.

[92]Institute for Translational Genomics and Population Sciences, Departments of Pediatrics and Medicine, Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center, Torrance, CA.

[93]Joseph J. Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI.

[94]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle WA.

[95]Department of Biostatistics, School of Public Health, Yale University, New Haven, CT.

[96]Computational Biology and Bioinformatics Program, Yale University, New Haven, CT.

[97]Department of Public Health Sciences, Center for Public Health Genomics, University of Virginia, Charlottesville, VA.

[98]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS.

[99]Department of Cardiology, Beth Israel Deaconess Medical Center, Boston, MA.

[100]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI.

[101]Department of Biology, MIT, Cambridge, MA, USA.

[102]Department of Systems Biology, Harvard Medical School, Boston, MA, USA.

[103]Harvard Society of Fellows, Harvard University, Cambridge, MA.

[104]Howard Hughes Medical Institute, Boston, MA.

[105]Department of Pathology, Stanford University School of Medicine, Stanford, CA.

[106]Regeneron Pharmaceuticals, Tarrytown, NY.

[107]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA.

[108]Verve Therapeutics, Cambridge, MA.

## ACKNOWLEDGEMENTS

Investigators who conducted this research report individual research support from R35 HL135818 (S. Redline), P01 HL132825 (S. Weiss), R01HL091357 and R01HL055673 (D. Arnett), W81XWH-17–1-0597 (D. Schwartz), K01 HL135405 (B. Cade), P01 HL132825 (J Lasky-Su), K01HL136700 (S. Aslibekyan), R01HL113323 (J Curran), R01HL1333040 (D. Weeks), 1R01HL138737 (D. Darbar), P01 HL132825 (P. Kachroo), T32 HL129982 (L. Raffield), R01HL113323 (J. Blangero), HHS-N2682018000002I (T. Blackwell and A Smith), U54GM115428 (J. Wilson), R01HL148565 and R01HL148050 (P. Natarajan), F30-HL149180 (S. M. Zekavat), R01HL139731 and AHA-18SFRN34250007 (S. Lubitz). Claudia Adams Barr Program for Innovative Cancer Research (V. Sankaran), R01142711, MGH Hassenfeld Scholar Award (P. Natarajan), Fondation Leducq TNE-18CVD04 (A. Bick, B. Ebert, S. Jaiswal, P. Natarajan, S. Kathiresan), Burroughs Wellcome Foundation and Ludwig Cancer Center (S. Jaiswal), UM1-HG008895 (S. Kathiresan).

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626–02S1; contract HHSN268201800002I). Phenotype harmonization, data management, sample-identity QC, and general study coordination were provided by the TOPMed Data Coordinating Center (3R01HL-120393–02S1; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

COMPETING INTERESTS

B. Psaty serves on the Steering Committee of the Yale Open Data Access Project funded by Johnson & Johnson. E. Silverman and M. Cho received grant support from GlaxoSmithKlein and Bayer. S. Weiss receive royalties from UpToDate. S. Aslibekyan reports Employment/equity in 23andMe, Inc. B. Freedman is a consultant for RenalytixAI and AstraZeneca Pharmaceuticals. MEM reports funding from Regeneron Pharmaceuticals Inc. unrelated to this project. M. Cho has received grant support from GlaxoSmithKlein and Bayer and consulting or speaking fees from AstraZeneca and Illumina. J. Floyd has consulted for Shionogi Inc. B. Levy is a co-founder of Nocion Therapeutics; receives grant support from Pieris Pharmaceuticals, Sanofi and Samsung Research America; and has served as a consultant for Bayer, Entrinsic Health, Gossamer Bio, NControl, Novartis, Teva and Thetis Pharmaceuticals. E. Lander serves on the board of directors for Codiak BioSciences and serves on the scientific

advisory board of F-Prime Capital Partners and Third Rock Ventures. B. Ebert reports grant support from Celgene and Deerfield. P. Ellinor has received grant support from Bayer AG and has served on advisory boards or consulted for Bayer AG, Quest Diagnostics, MyoKardia and Novartis. G. Abecasis is an employee of Regeneron Pharmaceuticals and owns stock and stock options for Regeneron Pharmaceuticals. S. Jaiswal is a scientific advisor to Grail. S. Lubitz receives sponsored research support from Bristol Myers Squibb, Pfizer, Bayer AG, Boehringer Ingelheim, and Fitbit, and has consulted for Bristol Myers Squibb, Pfizer and Bayer AG, and participates in a research collaboration with IBM. P. Natarajan reports grants support from Amgen, Apple, and Boston Scientific, and is a scientific advisor to Apple. S. Kathiresan is an employee of Verve Therapeutics, and holds equity in Verve Therapeutics, Maze Therapeutics, Catabasis, and San Therapeutics. He is a member of the scientific advisory boards for Regeneron Genetics Center and Corvidia Therapeutics; he has served as a consultant for Acceleron, Eli Lilly, Novartis, Merck, Novo Nordisk, Novo Ventures, Ionis, Alnylam, Aegerion, Haug Partners, Noble Insights, Leerink Partners, Bayer Healthcare, Illumina, Color Genomics, MedGenome, Quest, and Medscape The remainder of the authors report that they have nothing to disclose. The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the U.S. Department of Health and Human Services.

## Appensix: NHLBI Trans-Omics for Precision Medicine Consortium

Namiko Abe[109], Christine Albert[89], Laura Almasy[110], Alvaro Alonso[111], Seth Ament[112], Peter Anderson[113], Pramod Anugu[114], Deborah Applebaum-Bowden[115], Dan Arking[116], Allison Ashley-Koch[117], Stella Aslibekyan[118], Tim Assimes[119], Dimitrios Avramopoulos[116], John Barnard[120], R. Graham Barr[121], Emily Barron-Casella[116], Lucas Barwick[122], Terri Beaty[116], Gerald Beck[120], Diane Becker[59], Rebecca Beer[115], Amber Beitelshees[112], Emelia Benjamin[31], Panagiotis Benos[123], Marcos Bezerra[124], Larry Bielak[4], Russell Bowler[80], Jennifer Brody[113], Ulrich Broeckel[125], Karen Bunting[109], Carlos Bustamante[119], Jonathan Cardwell[126],Vincent Carey[89], Cara Carty[127], Richard Casaburi[128], James Casella[116], Peter Castaldi[88], Mark Chaffin[2], Christy Chang[112], Yi-Cheng Chang[129], Daniel Chasman[89], Sameer Chavan[126], Bo-Juen Chen[109], Wei-Min Chen[130], Seung Hoan Choi[2], Lee-Ming Chuang[129], Mina Chung[120], Ren-Hua Chung[131], Clary Clish[2], Suzy Comhair[12], Elaine Cornell[69], Carolyn Crandall[128], James Crapo[132], Jeffrey Curtis[4], Coleen Damcott[112], Sayantan Das[4], Sean David[132], Colleen Davis[113], Michael DeBaun[133], Ranjan Deka[134], Dawn DeMeo[137], Scott Devine[112], Qing Duan[135], Ravi Duggirala[136], Susan Dutcher[137], Charles Eaton[138], Lynette Ekunwe[114], Adel El Boueiz[29], Leslie Emery[113], Serpil Erzurum[120], Charles Farber[45], Matthew Flickinger[4], Nora Franceschini[135], Chris Frazar[26], Mao Fu[112], Stephanie M. Fullerton[113], Lucinda Fulton[137], Stacey Gabriel[4], Weiniu Gan[115], Shanshan Gao[126], Yan Gao[114], Margery Gass[37], Bruce Gelb[139], Xiaoqi (Priscilla) Geng[4], Mark Geraci[140], Soren Germer[109], Robert Gerszten[3], Auyon Ghosh[88], Richard Gibbs[141], Chris Gignoux[19], Mark Gladwin[123], David Glahn[8], Da-Wei Gong[112], Harald Goring[136], Sharon Graw[126], Daniel Grine[126], C. Charles Gu[137], Yue Guan[112], Namrata Gupta[3], Jeff Haessler[126], Michael Hall[114], Daniel Harris[112], Nicola L. Hawley[142], Ben Heavner[12], Ryan Hernandez[143], David Herrington[144], Craig Hersh[29], Bertha Hidalgo[61], Brian Hobbs[88], John Hokanson[126], Elliott Hong[112], Karin Hoth[145],Chao (Agnes) Hsiung[131], Yi-Jen Hung[146], Haley Huston[44], Chii Min Hwu[147], Rebecca Jackson[148], Deepti Jain[12], Cashell Jaquish[115], Min A Jhun[4], Craig Johnson[113], Rich Johnston[111], Kimberly Jones[116], Hyun Min Kang[12], Shannon Kelly[59], Michael Kessler[112], Alyna Khan[113], Wonji Kim[29], Greg Kinney[126], Holly Kramer[149], Christoph Lange[150], Meryl LeBoff[47], Seunggeun Shawn Lee[4], Wen-Jane Lee[147], Jonathon LeFaive[4], David Levine[113], Joshua Lewis[112], Xiaohui Li[151], Yun Li[135], Henry Lin[151], Honghuang Lin[152], Keng Han Lin[4], Xihong Lin[150], Simin Liu[153], Yu Liu[154], Kathryn Lunetta[152], James Luo[115], Michael Mahaney[136], Barry Make[116], Ani Manichaikul[97], Lauren Margolin[2], Lisa Martin[110], Susan Mathai[126], Susanne May[12],

Patrick McArdle[112], Merry-Lynn McDonald[118], Sean McFarland[2,3,5], Daniel McGoldrick[156], Caitlin McHugh[12], Hao Mei[114], Luisa Mestroni[126], Julie Mikulla[115], Nancy Min[114], Mollie Minear[115], Ryan L Minster[123], Matt Moll[85], Courtney Montgomery[157], Solomon Musani[41], Stanford Mwasongwe[114], Josyf C Mychaleckyj[130],Girish Nadkarni[158], Rakhi Naik[116], Take Naseri[159], Sergei Nekhai[160], Sarah C. Nelson[12], Bonnie Neltner[126], Deborah Nickerson[156], Jeff O'Connell[112], Tim O'Connor[112], Heather Ochs-Balcom[161], David Paik[154], James Pankow[162], George Papanicolaou[115], Afshin Parsa[112], Marco Perez[119], James Perry[112], Ulrike Peters[94], Patricia Peyser[4], Lawrence S Phillips[111], Toni Pollin[112], Wendy Post[116], Julia Powers Becker[126], Meher Preethi Boorgula[126], Michael Preuss[20], Pankaj Qasba[115], Dandi Qiao[29], Zhaohui Qin[111], Laura Rasmussen-Torvik[72], Aakrosh Ratan[130], Robert Reed[112], Elizabeth Regan[132], Muagututi'a Sefuiva Reupena[134], Ken Rice[12], Carolina Roselli[2], Ingo Ruczinski[116], Pamela Russell[126], Sarah Ruuska[44], Kathleen Ryan[112], Danish Saleheen[121], Shabnam Salimi[112], Steven Salzberg[116], Kevin Sandow[151], Christopher Scheller[4], Ellen Schmidt[4], Karen Schwander[137], Frank Sciurba[123], Christine Seidman[2,3,15,104], Jonathan Seidman[2,3], Vivien Sheehan[164], Stephanie L. Sherman[165], Amol Shetty[112], Aniket Shetty[126], Brian Silver[166], Josh Smith[113], Tanja Smith[109], Sylvia Smoller[36], Beverly Snively[144], Michael Snyder[119], Tamar Sofer[57], Nona Sotoodehnia[113], Adrienne M. Stilp[113], Garrett Storm[126], Elizabeth Streeten[112], Jessica Lasky Su[29], Yun Ju Sung[137], Jody Sylvia[47], Adam Szpiro[113], Carole Sztalryd[112], Daniel Taliun[4], Hua Tang[119], Matthew Taylor[126], Simeon Taylor[112], Marilyn Telen[117], Timothy A. Thornton[113], Machiko Threlkeld[156], Lesley Tinker[126], David Tirschwell[113], Sarah Tishkoff[167], Hemant Tiwari[62], Catherine Tong[12], Michael Tsai[162], Dhananjay Vaidya[116], David Van Den Berg[168], Peter VandeHaar[4], Scott Vrieze[169], Tarik Walker[126], Robert Wallace[145], Avram Walts[126], Heming Wang[23], Karol Watson[128], Bruce Weir[113], Lu-Chen Weng[2,64], Jennifer Wessel[170], Cristen Willer[171], Kayleen Williams[113], Carla Wilson[132], Joseph Wu[154], Huichun Xu[112], Lisa Yanek[116], Rongze Yang[112], Norann Zaghloul[112], Yingze Zhang[172], Snow Xueyan Zhao[32], Wei Zhao[2], Degui Zhi[173], Xiang Zhou[4], Xiaofeng Zhu[174], Michael Zody[109], Sebastian Zoellner[4]

[109]New York Genome Center, New York, NY. [110]Children's Hospital of Philadelphia, University of Pennsylvania, Philadelphia, PA. [111]Emory University, Atlanta, GA. [112]University of Maryland, Baltimore, MD. [113]University of Washington, Seattle, WA. [114]University of Mississippi, Jackson, MS. [115]National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD. [116]Johns Hopkins University, Baltimore, MD. [117]Duke University, Durham, NC. [118]University of Alabama, Birmingham, AL. [119]Stanford University, Stanford, CA. [120]Cleveland Clinic, Cleveland, OH. [121]Columbia University, New York, NY. [122] The Emmes Corporation, LTRC, Rockville, MD. [123]University of Pittsburgh, Pittsburgh, PA. [124]Fundação de Hematologia e Hemoterapia de Pernambuco - HEMOPE, Recife, Brazil. [125]Medical College of Wisconsin, Milwaukee, WI. [126]University of Colorado Anschutz Medical Campus, Aurora, CO. [127]Washington State University, Seattle, WA. [128]University of California, Los Angeles, Los Angeles, CA. [129] National Taiwan University Hospital, Taipei, Taiwan. [130]University of Virginia, Charlottesville, VA. [131]National Health Research Institute, Zhunan, Taiwan. [132]National Jewish Health, Denver, CO. [132]University of Chicago, Chicago, IL. [133]Vanderbilt University, Nashville, TN. [134]University of Cincinnati, Cincinnati, OH. [135]University of North Carolina, Chapel Hill, NC. [136]University

of Texas Rio Grande Valley School of Medicine, Edinburg, TX. [137]Washington University in St Louis, St Louis, MO. [138]Brown University, Providence, RI. [139]Icahn School of Medicine at Mount Sinai, New York, NY. [140]Indiana University, Medicine, Indianapolis, IN. [141]Baylor College of Medicine Human Genome Sequencing Center, Houston, TX. [142]Department of Chronic Disease Epidemiology, Yale University, New Haven, CT. [143]McGill University, Montreal, QC, Canada.[144]Wake Forest Baptist Health, Winston-Salem, NC. [145]University of Iowa, Iowa City, IA. [146]Tri-Service General Hospital National Defense Medical Center, Taipei, Taiwan. [147]Taichung Veterans General Hospital Taiwan, Taichung City, Taiwan. [148] Division of Endocrinology, Diabetes and Metabolism, Department of Medicine, Ohio State University Wexner Medical Center, Columbus, OH. [149]Loyola University, Public Health Sciences, Maywood, IL. [150]Department of Biostatistics, Harvard School of Public Health, Boston, MA. [151]Lundquist Institute, Torrance, CA. [152]Boston University, Boston, MA. [153]Department of Epidemiology, Brown University, Providence, RI. [154]Cardiovascular Institute, Stanford University, Palo Alto, CA. [155]George Washington University, Washington, DC. [156]Department of Genome Sciences, University of Washington, Seattle, WA. [157]Oklahoma Medical Research Foundation, Genes and Human Disease, Oklahoma City, OK. [158]Division of Nephrology, Icahn School of Medicine at Mount Sinai, New York, NY. [159]Ministry of Health, Government of Samoa, Apia. [160]Howard University, Washington, DC. [161]University at Buffalo, Buffalo, NY. [162]University of Minnesota, Minneapolis, MN. [163]Lutia i Puava ae Mapu i Fagalele, Apia, Samoa. [164]Department of Pediatrics, Baylor College of Medicine, Houston, TX. [165]Department of Human Genetics, Emory University, Atlanta, GA. [166]UMass Memorial Medical Center, Worcester, MA. [167]Department of Genetics, University of Pennsylvania, Philadelphia, PA. [168]USC Methylation Characterization Center, University of Southern California, Los Angeles, CA. [169]Department of Psychology, University of Minnesota, Minneapolis, MN. [170]Department of Epidemiology, Indiana University, Indianapolis, IN. [171]Department of Medicine, University of Michigan, Ann Arbor, MI. [172]Department of Medicine, University of Pittsburgh, Pittsburgh, PA. [173]Center for Precision Health, School of Biomedical Informatics, University of Texas Health at Houston, Houston, TX. [174]Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH.

## REFERENCES

1. Kennedy BK et al. Geroscience: linking aging to chronic disease. Cell 159, 709–13 (2014). [PubMed: 25417146]
2. Jaiswal S. et al. Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. New England Journal of Medicine 371, 2488–2498 (2014).
3. Genovese G. et al. Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. New England Journal of Medicine 371, 2477–2487 (2014).
4. Xie M. et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. Nature Medicine 20, 1472–1478 (2014).
5. Jaiswal S. et al. Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. New England Journal of Medicine 377, 111–121 (2017).
6. Steensma DP et al. Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. Blood 126, 9–16 (2015). [PubMed: 25931582]

7. Taliun D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. bioRxiv 563866 (2019). doi:10.1101/563866
8. Cibulskis K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature Biotechnology (2013). doi:10.1038/nbt.2514
9. Loh PR et al. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. Nature (2018). doi:10.1038/s41586-018-0321-x
10. Jaiswal S. et al. Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. New England Journal of Medicine 371, 2488–2498 (2014).
11. Patel K. v. et al. Red Cell distribution width and mortality in older adults: A meta-analysis. Journals of Gerontology - Series A Biological Sciences and Medical Sciences 65 A, 258–265 (2010).
12. Bick AG et al. Genetic Interleukin 6 Signaling Deficiency Attenuates Cardiovascular Risk in Clonal Hematopoiesis. Circulation 141, 124–131 (2020). [PubMed: 31707836]
13. Alexandrov LB et al. Clock-like mutational processes in human somatic cells. Nature Genetics 47, 1402–1407 (2015). [PubMed: 26551669]
14. Zink F. et al. Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. Blood 130, 742–752 (2017). [PubMed: 28483762]
15. Bowman RL, Busque L & Levine RL Cell Stem Cell Review Clonal Hematopoiesis and Evolution to Hematopoietic Malignancies. (2018). doi:10.1016/j.stem.2018.01.011
16. Desai P. et al. Somatic mutations precede acute myeloid leukemia years before diagnosis. Nature Medicine 24, 1015–1023 (2018).
17. Bojesen SE et al. Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. Nature Genetics (2013). doi:10.1038/ng.2566
18. Zhou W. et al. Mosaic loss of chromosome Y is associated with common variation near TCL1A. Nature genetics 48, 563–8 (2016). [PubMed: 27064253]
19. Hinds DA et al. Germ line variants predispose to both JAK2 V617F clonal hematopoiesis and myeloproliferative neoplasms. Blood (2016). doi:10.1182/blood-2015-06-652941
20. Hu Y. et al. A statistical framework for cross-tissue transcriptome-wide association analysis. Nature Genetics (2019). doi:10.1038/s41588-019-0345-7
21. Smith BW et al. The aryl hydrocarbon receptor directs hematopoietic progenitor cell expansion and differentiation. Blood (2013). doi:10.1182/blood-2012-11-466722
22. Cybulski C. et al. CHEK2 Is a Multiorgan Cancer Susceptibility Gene. The American Journal of Human Genetics 75, 1131–1135 (2004). [PubMed: 15492928]
23. Rudd MF, Sellick GS, Webb EL, Catovsky D & Houlston RS Variants in the ATM-BRCA2-CHEK2 axis predispose to chronic lymphocytic leukemia. Blood (2006). doi:10.1182/blood-2005-12-5022
24. Huynh M. et al. Hyaluronan and proteoglycan link protein 1 (HAPLN1) activates bortezomib-resistant NF-κB activity and increases drug resistance in multiple myeloma. The Journal of biological chemistry 293, 2452–2465 (2018). [PubMed: 29279332]
25. Moran-Crusio K. et al. Tet2 Loss Leads to Increased Hematopoietic Stem Cell Self-Renewal and Myeloid Transformation. Cancer Cell (2011). doi:10.1016/j.ccr.2011.06.001
26. Kilpivaara O. et al. A germline JAK2 SNP is associated with predisposition to the development of JAK2V617F-positive myeloproliferative neoplasms. Nature Genetics 41, 455–459 (2009). [PubMed: 19287384]
27. Jones A. v. et al. JAK2 haplotype is a major risk factor for the development of myeloproliferative neoplasms. Nature Genetics 41, 446–449 (2009). [PubMed: 19287382]
28. Olcaydu D. et al. A common JAK2 haplotype confers susceptibility to myeloproliferative neoplasms. Nature Genetics 41, 450–454 (2009). [PubMed: 19287385]
29. Young AL, Challen GA, Birmann BM & Druley TE Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. Nature Communications 7, 1–7 (2016).
30. Regier AA et al. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. Nature Communications (2018). doi:10.1038/s41467-018-06159-4

31. Jun G, Wing MK, Abecasis GR & Kang HM An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. Genome Research 25, 918–925 (2015). [PubMed: 25883319]
32. Karczewski KJ et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv 531210 (2019). doi:10.1101/531210
33. Gibson CJ et al. Clonal Hematopoiesis Associated With Adverse Outcomes After Autologous Stem-Cell Transplantation for Lymphoma. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 35, 1598–1605 (2017). [PubMed: 28068180]
34. Hiatt JB, Pritchard CC, Salipante SJ, O'Roak BJ & Shendure J Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. Genome Research 23, 843–854 (2013). [PubMed: 23382536]
35. Pérez Millán MI et al. Next generation sequencing panel based on single molecule molecular inversion probes for detecting genetic variants in children with hypopituitarism. Molecular Genetics and Genomic Medicine 6, 514–525 (2018).
36. Li Y, Willer CJ, Ding J, Scheet P & Abecasis GR MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetic Epidemiology 34, 816–834 (2010). [PubMed: 21058334]
37. Vattathil S & Scheet P Haplotype-based profiling of subtle allelic imbalance with SNP arrays. Genome Research 23, 152–158 (2013). [PubMed: 23028187]
38. Fowler J, San Lucas FA & Scheet P System for Quality-Assured Data Analysis: Flexible, reproducible scientific workflows. Genetic Epidemiology 43, 227–237 (2019). [PubMed: 30565316]
39. Natarajan P. et al. Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. Nature Communications 9, 3391 (2018).
40. Nik-Zainal S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 534, 47–54 (2016). [PubMed: 27135926]
41. Blokzijl F, Janssen R, van Boxtel R & Cuppen E MutationalPatterns: Comprehensive genome-wide analysis of mutational processes. Genome Medicine 10, 33 (2018). [PubMed: 29695279]
42. Zhou W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nature genetics 50, 1335–1341 (2018). [PubMed: 30104761]
43. Bates D. et al. Package "lme4": Linear Mixed-Effects Models using "Eigen" and S4. Journal of Statistical Software (2015). doi:10.18637/jss.v067.i01
44. Benner C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics (Oxford, England) 32, 1493–1501 (2016).
45. Amemiya HM, Kundaje A & Boyle AP The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Scientific reports 9, 9354 (2019). [PubMed: 31249361]
46. Fulco CP et al. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. Nature Genetics doi:10.1038/s41588-019-0538-0
47. Johnson WE, Li C & Rabinovic A Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics (Oxford, England) 8, 118–27 (2007).
48. Dobin A. et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013). [PubMed: 23104886]
49. DeLuca DS et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics (Oxford, England) 28, 1530–2 (2012).
50. Stranger BE et al. Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease. Nature Genetics 49, 1664–1670 (2017). [PubMed: 29019975]
51. Houseman EA et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics (2012). doi:10.1186/1471-2105-13-86
52. Horvath S & Levine AJ HIV-1 infection accelerates age according to the epigenetic clock. Journal of Infectious Diseases (2015). doi:10.1093/infdis/jiv277
53. Barfield RT, Kilaru V, Smith AK & Conneely KN CpGassoc: An R function for analysis of DNA methylation microarray data. Bioinformatics (2012). doi:10.1093/bioinformatics/bts124

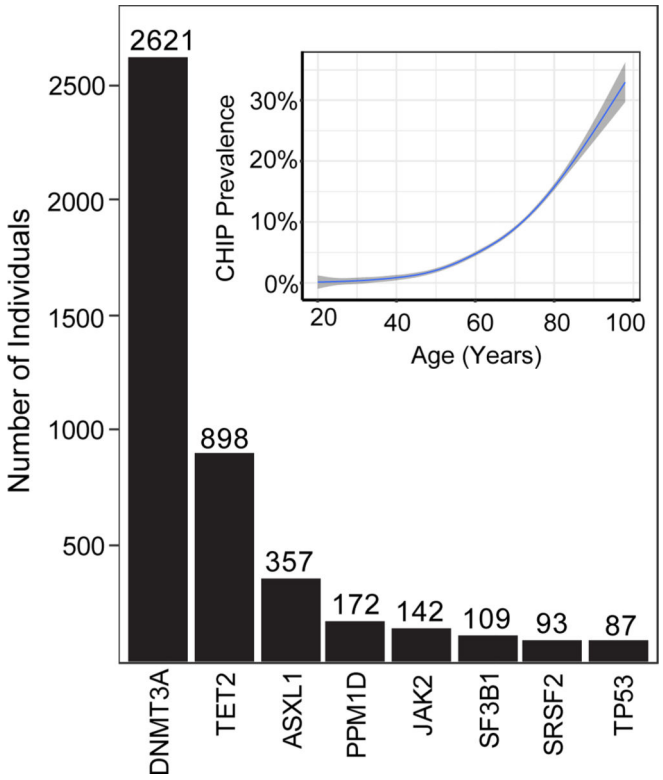



**Fig. 1|. Identifying CHIP in TOPMed Genomes.**

CHIP was identified in 97,631 whole genome sequenced peripheral blood samples through the curation of somatic driver mutations. Counts for 8 most common driver genes plotted. **inset**, CHIP prevalence increased with age. Center line represents general additive model spline, 95% confidence interval is shaded (n=82,807 individuals; two-sided t-test: $p<10^{-300}$).

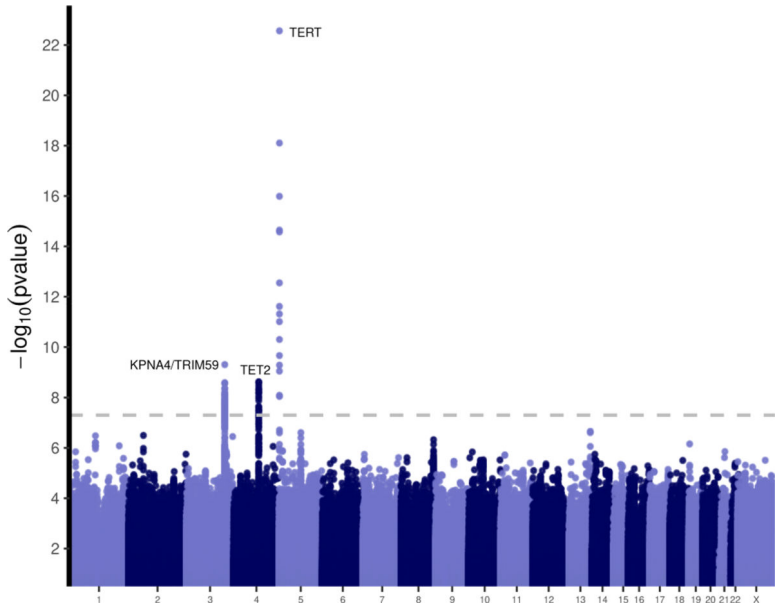



**Fig. 2|. Genetic Determinants of CHIP.**

Single variant genetic association analyses of CHIP identified three genome wide significant loci. Two-sided association testing performed using SAIGE (n=65,405 individuals, see methods)

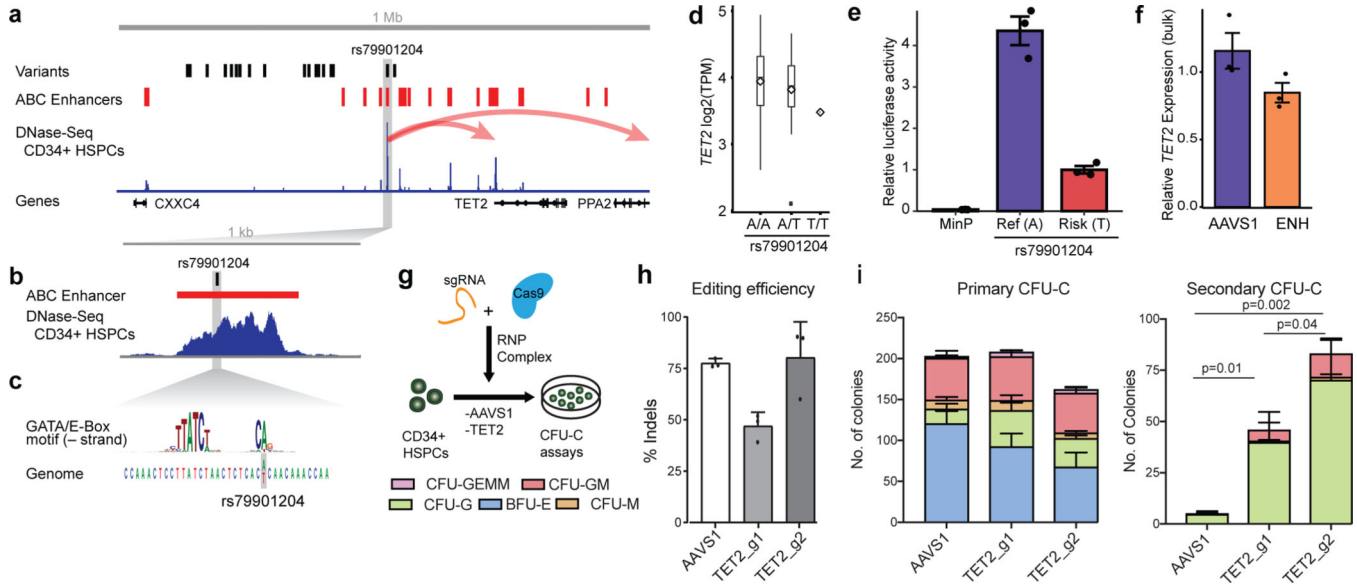



**Fig. 3|. African ancestry specific *TET2* locus risk variant disrupts hematopoietic stem cell *TET2* enhancer decreasing TET2 expression and increasing self-renewal.**

**a,** the TET2 locus with fine-mapped risk variants, Activity-by-Contact (ABC) hematopoietic stem and progenitor cell (HSPC) enhancers, DNase-Seq CD34+ HSPC and RefSeq genes. ABC model predicts that rs79901204 disrupts a TET2 enhancer resulting in decreased TET2 expression (see methods). **b,** expanded view of TET2 enhancer element. **c,** rs79901204 disrupts a GATA motif/E-Box motif. **d,** rs79901204 is associated with decreased TET2 expression in human peripheral blood RNA-seq ($N_{A/A}$=230, $N_{A/T}$=16, $N_{T/T}$=1, two-sided linear mixed model p=0.012). TPM, transcripts per million. Boxplot displays median, 25$^{th}$ and 75$^{th}$ percentiles, mean (diamond symbol) and outlier observations (black dots) **e,** luciferase assay in CD34+ primary cells demonstrates four-fold attenuation of enhancer activity by the rs79901204 T risk allele relative to the A reference allele (N=3, two-sided t-test p=0.007). **f,** deleting the TET2 enhancer (ENH) in CD34+ primary cells results in decreased TET2 expression relative to deletion of control locus AAVS1 (N=3, two-sided t-test, p=0.04). **g,** Human HSPCs were electroporated with Cas9 targeting a coding region of TET2 and AAVS1 (a control locus) and plated for primary and secondary colony-forming assays. **h,** two TET2 guides had differential editing efficiency. **i,** TET2 coding disruption leads to expanded secondary colony formation compared to AAVS1 controls (N=3, two-sided t-test p=0.01, p=0.002 for g1 and g2 respectively, with greater expansion identified in the TET2 guide with greater editing efficiency (two-sided t-test p=0.04). Mean and standard deviation of number of each colony type plotted. CFU-M, colony forming unit-macrophage; CFU-GM, granulocyte macrophage; CFU-GEMM, granulocyte erythrocyte macrophage megakaryocyte; CFU-G, granulocyte; BFU-E, burst forming unit-erythroid. In **e**, **f**, **h**, points represent independent replicates, mean values and error bars represent standard error are plotted.