10-14-2020

Inherited causes of clonal haematopoiesis in 97,691 whole genomes

Alexander G. Bick

Joshua S. Weinstock

Satish K. Nandakumar

Charles P. Fulco

Erik L. Bao

See next page for additional authors

Recommended Citation

Bick, A.G., Weinstock, J.S., Nandakumar, S.K. et al. Inherited causes of clonal haematopoiesis in 97,691 whole genomes. Nature (2020). https://doi.org/10.1038/s41586-020-2819-2

Authors

Alexander G. Bick, Joshua S. Weinstock, Satish K. Nandakumar, Charles P. Fulco, Erik L. Bao, Seyedeh M. Zekavat, Mindy D. Szeto, Juan M. Peralta, Joanne E. Curran, and John Blangero

Inherited causes of clonal haematopoiesis in 97,691 whole genomes

https://doi.org/10.1038/s41586-020-2819-2

Received: 25 September 2019

Accepted: 3 July 2020

Published online: 14 October 2020

Check for updates

A list of authors and affiliations appears at the end of the paper.

Age is the dominant risk factor for most chronic human diseases, but the mechanisms through which ageing confers this risk are largely unknown[1]. The age-related acquisition of somatic mutations that lead to clonal expansion in regenerating haematopoietic stem cell populations has recently been associated with both haematological cancer[2–4] and coronary heart disease[5]—this phenomenon is termed clonal haematopoiesis of indeterminate potential (CHIP)[6]. Simultaneous analyses of germline and somatic whole-genome sequences provide the opportunity to identify root causes of CHIP. Here we analyse high-coverage whole-genome sequences from 97,691 participants of diverse ancestries in the National Heart, Lung, and Blood Institute Trans-omics for Precision Medicine (TOPMed) programme, and identify 4,229 individuals with CHIP. We identify associations with blood cell, lipid and inflammatory traits that are specific to different CHIP driver genes. Association of a genome-wide set of germline genetic variants enabled the identification of three genetic loci associated with CHIP status, including one locus at *TET2* that was specific to individuals of African ancestry. In silico-informed in vitro evaluation of the *TET2* germline locus enabled the identification of a causal variant that disrupts a *TET2* distal enhancer, resulting in increased self-renewal of haematopoietic stem cells. Overall, we observe that germline genetic variation shapes haematopoietic stem cell function, leading to CHIP through mechanisms that are specific to clonal haematopoiesis as well as shared mechanisms that lead to somatic mutations across tissues.

The US National Heart, Lung, and Blood Institute (NHLBI) TOPMed project seeks to use high-coverage (>35×) whole-genome sequencing (WGS) and molecular profiling to improve the fundamental understanding of heart, lung, blood and sleep disorders[7]. Within the TOPMed programme, we designed a study to detect CHIP from WGS of blood DNA in 97,691 individuals across 51 largely observational epidemiological 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 leukaemogenic driver mutations[5] (see Methods, Supplementary Table 2).

In total, we identified 4,938 CHIP mutations in 4,229 individuals (Supplementary Table 3). The median variant allele fraction (VAF) of the observed CHIP mutations was 16%. Consistent with previous reports, more than 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*) (Fig. 1). Among these eight genes, there was marked heterogeneity in the clonal fraction. For example, the *DNMT3A* and *TET2* CHIP clonal fractions of the peripheral blood were about 25% smaller ($P = 1.3 \times 10^{-15}$) and about 14% smaller ($P = 2.1 \times 10^{-4}$), respectively, than the *ASXL1* clonal fraction, implicating the presence of driver mutation gene-specific differences in clonal selection (Extended Data Fig. 1a). Ninety percent of individuals with CHIP driver mutations had only one identified mutation (Extended Data Fig. 1b).

Phenotypic associations with CHIP

CHIP prevalence was strongly correlated with age at the time of blood drawing ($P < 10^{-300}$, Fig. 1b). CHIP prevalence was highly consistent across studies and resembled those found in previous reports[2–4] using whole-exome sequencing (Extended Data Fig. 1c, d). Consistent with previous studies, a history of smoking was associated with increased probability of CHIP (odds ratio (OR) = 1.18, $P = 5 \times 10^{-5}$) whereas Hispanic and East Asian ancestry were both associated with reduced probability of CHIP (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 had a higher mean age than carriers of single-nucleotide CHIP mutations (Wilcoxon rank sum test, $P = 0.01$). Similarly, in the subset of individuals with *ASXL1* CHIP mutations, which are exclusively loss-of-function single-nucleotide stop-gain or frameshift mutations, individuals with *ASXL1* frameshift mutations were older on average (Wilcoxon rank sum test: $P = 0.009$, Extended Data Fig. 2a).

Carriers of *JAK2* CHIP mutations had the lowest mean age among CHIP mutation 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, and *PPM1D*, *SF3B1* and *SRSF2* carriers were 5.0 ($P = 5.7 \times 10^{-4}$), 6.9 ($P = 1.8 \times 10^{-6}$) and 7.7 ($P = 1.3 \times 10^{-4}$) years older, respectively (Extended Data Fig. 2b).

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 Fig. 2c).

CHIP is distinguished from other clonal haematological disorders on the basis of 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 haemoglobin ($P = 0.04$) among patients with CHIP mutations compared to those without such mutations (Extended Data Fig. 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 previous observations[2]. Notably, RDW is a haematological parameter that increases with age and predicts overall mortality and poor clinical outcomes in the setting of cardiovascular disease and in older adults[10].

Given the previous association of CHIP with atherosclerotic cardiovascular disease[5,11], we investigated whether CHIP carriers had altered lipid profiles. Consistent with previous 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 Fig. 3b, Supplementary Table 6).

We characterized the inflammatory profile of CHIP carriers (Extended Data Fig. 3c, Supplementary Table 7). In aggregate, CHIP was associated with an increased level of interleukin 6 (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 (area under the curve = 0.55; for cut-off of CRP > 2 mg l^{-1}, positive predictive value = 6.3%, sensitivity = 60%). Driver-gene-specific analyses highlighted the association of *TET2* CHIP with increased IL-1β ($P = 2.4 \times 10^{-4}$), whereas *JAK2* and *SF3B1* were associated with increased circulating IL-18 ($P = 1.3 \times 10^{-4}$ and 1.27×10^{-20}, respectively).

To identify underlying determinants of the somatic mutational spectrum, we performed COSMIC mutational signature analysis[12] on passenger somatic mutations identified in CHIP carriers and non-carriers (see Methods). Among CHIP carriers, we observed enrichment of signature 4, which has been associated with smoking, and signature 6, which has been associated with defective DNA mismatch repair (Extended Data Fig. 4).

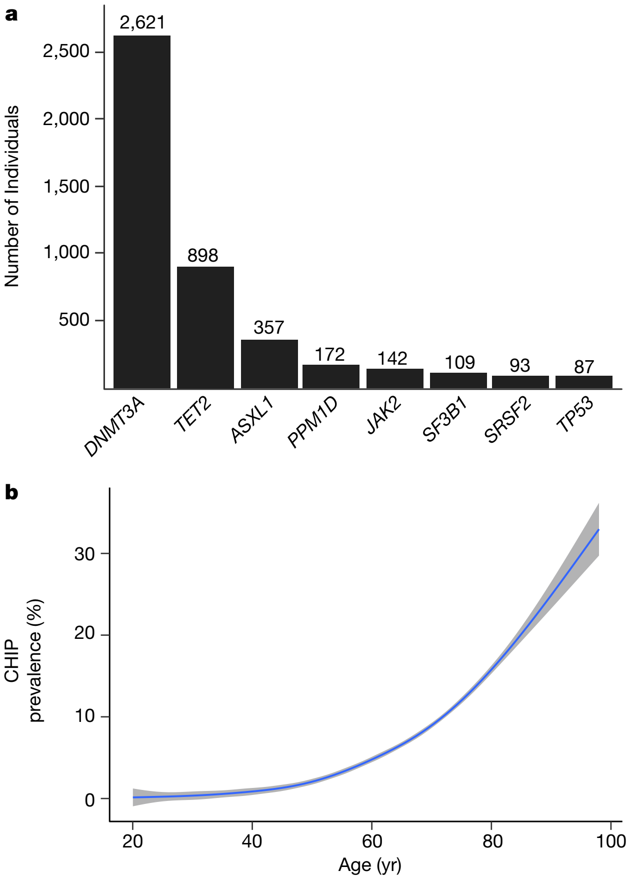


Fig. 1 | Identifying CHIP in TOPMed genomes. a, CHIP was identified in 97,631 peripheral blood WGS samples through the curation of somatic driver mutations. Counts for the eight most common driver genes are plotted. **b**, CHIP prevalence increased with age of donor at time of blood sampling. The centre line represents the general additive model spline and the shaded region is the 95% confidence interval ($n = 82,807$ individuals; two-sided t-test, $P < 10^{-300}$).

Germline genetic determinants of CHIP

Previous genome-wide association analyses in individuals of European ancestry have identified germline genetic variants at the *TERT* locus as predisposing to clonal haematopoiesis, defined either by somatic mosaicism of single-nucleotide variants (SNVs) and indels[13] or by large-scale chromosomal rearrangements[9]. Given the distinct association of clonal haematopoiesis with known leukaemogenic mutations (that is, CHIP) with both cancer[2,14,15] and atherosclerotic cardiovascular disease[5,11], we sought to discover germline genetic variations that confer increased risk of CHIP acquisition. We performed a single-variant genome-wide association analysis in a subset of 65,405 individuals (3,831 CHIP cases) in which the likelihood of having a CHIP mutation was higher than 1% (see Methods). The trait heritability explained by the analysis with linkage disequilibrium 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 haematopoiesis[13] (defined on the basis of somatic mosaicism of SNVs and indels), rs34002450 (OR 1.2, $P = 2.0 \times 10^{-13}$). rs34002450 is in strong linkage disequilibrium ($r^2 = 0.55$) with rs7705526, our lead variant at this locus and a common variant (minor allele frequency (MAF) = 0.29) in the fifth intron of *TERT*, which encodes telomere enzyme reverse transcriptase. In TOPMed, carriers of the rs34002450-A (minor) allele have a 1.3-fold increased risk of developing CHIP ($P = 8.4 \times 10^{-24}$). This variant was previously associated with increased leukocyte telomere length[16], myeloproliferative neoplasms[17] (MPNs) and clonal chromosomal mosaicism[9]. In a phenome-wide association analysis (PheWAS) of rs34002450-A in UK Biobank, we identified a significantly increased risk of MPNs ($P = 2.6 \times 10^{-13}$), 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, $P = 6.1 \times 10^{-10}$; conditional OR 1.1, $P = 4.7 \times 10^{-4}$).

In the TOPMed single-variant association analysis, we additionally identified two other novel genome-wide-significant genetic loci, including one locus on chromosome 3 in an intergenic region spanning *KPNA4* and *TRIM59* and one locus on chromosome 4 near *TET2* (Fig. 2, Extended Data Fig. 5, Supplementary Table 8).

rs1210060191 is a common variant (MAF 0.54) in a locus with an association signal that spans a 300-kb region that includes *KPNA4*, *TRIM59*, *IFT80* and *SMC4*. The lead variant is a 1-base pair (bp) intronic deletion in *TRIM59*. Carriers of the del(T) allele have a 1.16-fold increased risk of CHIP ($P = 5.3 \times 10^{-10}$). Variants in linkage disequilibrium with this variant have been identified as associated with MPNs[17]. No other significant phenotypic associations were noted in UK Biobank PheWAS analyses.

rs144418061 is a variant specific to samples from individuals with African ancestry (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 of CHIP ($P = 4.0 \times 10^{-9}$). 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

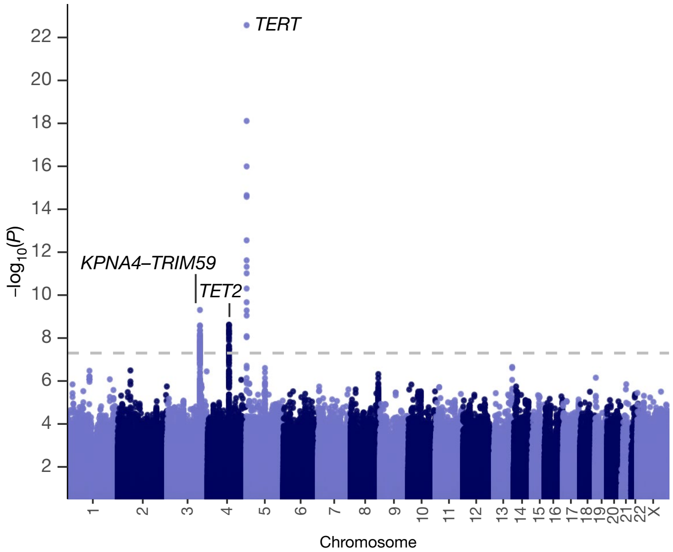


Fig. 2 | Genetic determinants of CHIP. Single-variant genetic-association analyses of CHIP identified three genome-wide-significant loci. Two-sided association testing performed using SAIGE (n = 65,405 individuals; see Methods).

equally robust for *DNMT3A* CHIP, *TET2* CHIP and *ASXL1* CHIP, suggesting that the germline variant does not specifically predispose individuals to *TET2* CHIP. Although other variants in the vicinity of *TET2* have been associated with MPNs[17], to our knowledge, this variant has not previously been identified as associated with any traits, probably owing to the underrepresentation of genomes with African ancestry in published association studies.

We considered whether there might be germline variants that predispose to specific CHIP driver mutations by separately performing a genome-wide association study (GWAS) on samples with *DNMT3A*- and *TET2*-associated CHIP. We identified a single novel locus for *DNMT3A* chip at rs2887399 in an intron of the T-cell leukaemia/lymphoma 1A gene (*TCL1A*). Carriers of the T allele (MAF 0.26) are at 1.23-fold risk of acquiring a *DNMT3A* CHIP mutation ($P = 3.9 \times 10^{-9}$). Of note, 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 the associations between germline loci and CHIP clones were robust across the size spectrum of CHIP clones, using the association between the *JAK2* 46/1 haplotype (tagged by rs1327494) and *JAK2* CHIP[19]. We found that rs1327494 associates with *JAK2* CHIP across VAF thresholds. We evaluated whether this observation generalized beyond *JAK2* CHIP to encompass all CHIP mutations. We found that the *TERT* locus (tagged by rs7705526) is associated with CHIP mutations across all VAF thresholds (Supplementary Table 9). These observations imply that our genetic associations are not dependent on clone size being detectable by deep-coverage WGS.

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 *KPNA4–TRIM59* locus on chromosome 3 and six additional loci: *AHRR*, *ASL*, *KREMN2*, *LEAP2*, *JSRP1* and *RASEF* (Extended Data Figs. 6, 7). *AHRR* directs haematopoietic progenitor cell expansion and differentiation[21].

We also performed gene-based association tests for aggregations of rare (MAF < 0.1%) putative loss-of-function germline variants within genes for CHIP. Although no genes reached exome-wide significance, the top associated gene was the DNA damage repair gene *CHEK2* (OR 1.7, $P = 1.3 \times 10^{-5}$; Supplementary Table 10). Rare germline variants in *CHEK2* are implicated in a diverse set of haematological and solid tumour malignancies[22,23]. Common variants in *CHEK2* are associated with MPNs[19], and a low-frequency frameshift *CHEK2* variant 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[17]. These results suggest that whereas *CHEK2* may ordinarily limit haematopoietic stem cell expansion, loss of *CHEK2* function may promote self-renewal, increasing the risk of CHIP.

We next sought to determine whether rare variants in noncoding regions associate with CHIP acquisition (see Methods). One set of variants in *HAPLN1* enhancers was associated with CHIP acquisition (OR 6.8, $P = 1.96 \times 10^{-5}$; Supplementary Table 11). HAPLN1 is an extracellular matrix protein that is produced in bone marrow stromal cells and has previously been implicated in NF-κB signalling[24].

We also tested whether germline genetic variants were associated with CHIP clone size, but found that no single variant or aggregated rare variants exceeded Bonferroni significance (Supplementary Tables 12, 13).

Characterization of the *TET2* CHIP risk locus

Finally, we bioinformatically and experimentally characterized the mechanism by which the noncoding variant at the *TET2* locus specifically identified in individuals of African ancestry influenced risk of CHIP. First, iterative conditional analyses at the locus suggested that there was most probably only a single causal variant. Fine mapping prioritized 25 variants in the credible set (greater than 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 of *TET2* in haematopoietic stem cells, because heterozygous *Tet2* knockout in mice increases the self-renewal of haematopoietic stem cells in vivo[25] and recapitulates the clonal expansion observed in humans with somatic mutations in *TET2*[2,5]. Accordingly, we used the activity-by-contact (ABC) model to predict which noncoding elements act as enhancers in CD34$^+$ haematopoietic stem and progenitor cells (HSPCs; see Methods). One variant (rs79901204) in this credible set overlapped with an element predicted to regulate any gene, and this element was indeed predicted to regulate *TET2* expression. (Fig. 3a, Supplementary Table 14) The T risk allele disrupts a consensus GATA–E-Box motif, probably resulting in reduced binding of the activating transcription factors GATA1 and GATA2 (Fig. 3b, c).

We then evaluated whether rs79901204 affected *TET2* expression in vivo in human peripheral blood samples. We used whole-blood RNA sequencing (RNA-seq) from 247 African American individuals, 16 of whom were heterozygous and one of whom was homozygous for rs79901204. In these samples, the T risk allele led to a dose-dependent decrease in whole-blood *TET2* expression (Fig. 3d; $\beta = -0.27 \pm 0.11$, mean ± s.e.m.; two-sided linear mixed model $P = 0.012$). Therefore, we sought to test our hypothesis 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-bp region containing the regulatory element using a plasmid-based luciferase enhancer assay in haematopoietic cells. The reference sequence activated luciferase expression by 118-fold versus control constructs with no enhancer sequence, whereas the T risk allele activated expression by only 27-fold (two-sided t-test $P = 0.007$; Fig. 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 h (Fig. 3f).

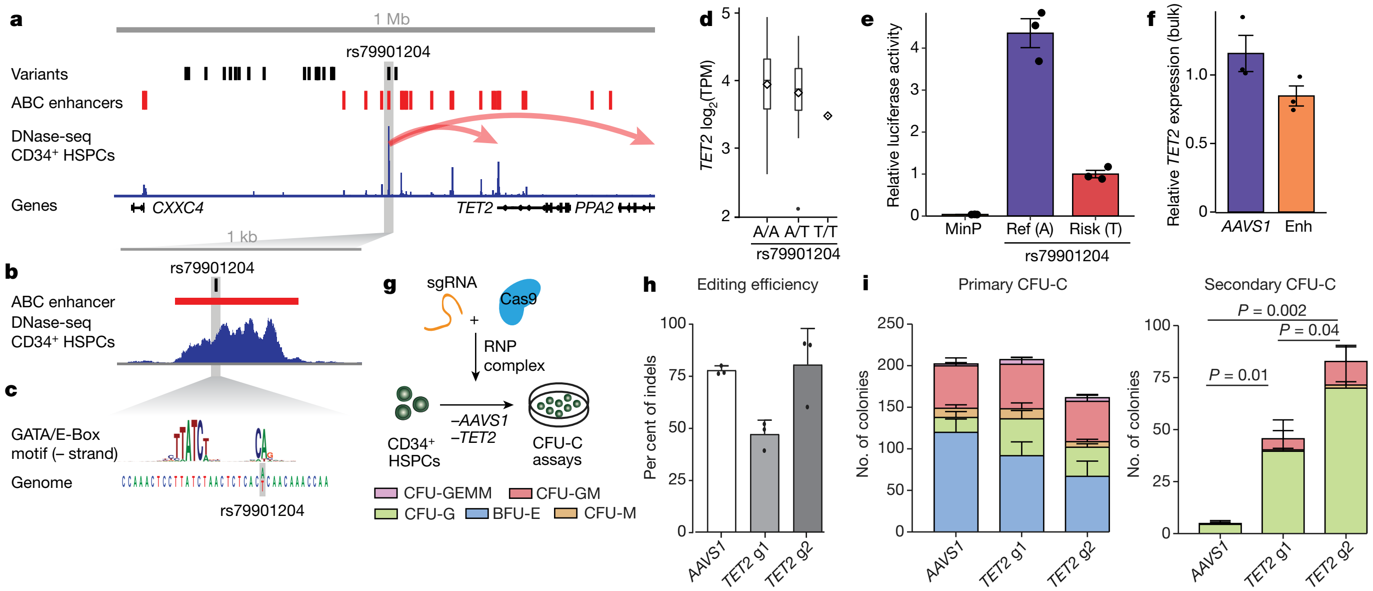


Fig. 3 | A *TET2* locus risk variant specific to donors with African ancestry disrupts the haematopoietic stem cell *TET2* enhancer, decreasing *TET2* expression and increasing self-renewal. a, The *TET2* locus, with fine-mapped risk variants, ABC HSPC enhancers, DNase I-hypersensitive site sequencing (DNase-seq) CD34$^+$ HSPCs and RefSeq genes. The ABC model predicts that rs79901204 disrupts a *TET2* enhancer, resulting in decreased *TET2* expression (see Methods). **b**, Expanded view of the *TET2* enhancer element. **c**, rs79901204 disrupts a GATA–E-Box motif. **d**, rs79901204 is associated with decreased *TET2* expression, as measured by RNA-seq of human peripheral blood ($n = 230$, 16 and 1 for 0, 1 and 2 rs79901204 alternate alleles, respectively; two-sided linear mixed model, $P = 0.012$). Box plot displays median (centre line), 25th and 75th percentiles (box edges), mean (diamond) and outliers (black dots). TPM, transcripts per million. **e**, Luciferase assay in CD34$^+$ primary cells demonstrates fourfold attenuation of enhancer activity by the rs79901204 T risk allele relative to the A reference allele ($n = 3$; two-sided *t*-test, $P = 0.007$). MinP, vector-only control. **f**, Deleting the *TET2* enhancer (Enh) in CD34$^+$ primary cells results in decreased *TET2* expression relative to deletion of the 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* (control locus) and plated for primary and secondary colony-forming assays. RNP, ribonucleoprotein; sgRNA, single guide RNA. **h**, Two *TET2* guides (g1 and g2) had differential editing efficiency. **i**, Coding-disrupted *TET2* leads to expanded secondary colony formation compared with *AAVS1* controls ($n = 3$; two-sided *t*-test, $P = 0.01$ (g1), $P = 0.002$ (g2)) with greater expansion identified in the *TET2* guide with greater editing efficiency (two-sided *t*-test, $P = 0.04$). Data are mean ± s.d. of number of each colony type. BFU-E, burst forming unit-erythroid; CFU-G, colony forming unit-granulocyte; CFU-GEMM, granulocyte erythrocyte macrophage megakaryocyte; CFU-GM, granulocyte macrophage; CFU-M, macrophage. In **e**, **f**, **h**, points represent independent replicates, bars show mean and error bars represent s.e.m.

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 (Fig. 3g). To establish a dose-response relationship, we used two *TET2* guides with differential editing efficiency (Fig. 3h, Extended Data Fig. 8). Disruption of *TET2* resulted in expanded secondary colony formation compared with *AAVS1* controls, with greater expansion identified in the *TET2* guide with greater editing efficiency (Fig. 3i). These results demonstrate that reduction of *TET2* activity promotes self-renewal and proliferation of HSPCs, illustrating how, at this locus, both germline noncoding and somatic coding variation converge to affect *TET2* and influence the development of CHIP.

Given the established role of *TET2* in DNA demethylation and our finding that rs79901204 is associated with decreased *TET2* expression (Fig. 3d), we hypothesized that carriers of the rs79901204 T allele might have altered peripheral blood methylation profiles. We performed a methylation quantitative trait locus (QTL) analysis of 1,747 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. 9, Supplementary Table 15).

Our observations lead to several conclusions. First, our sample size, which is nearly an order of magnitude larger than previous CHIP analyses[2,3,13], enables refinement of CHIP phenotype associations at the level of CHIP driver genes. We find that there is considerable driver-gene-dependent heterogeneity across CHIP phenotypes. For example, IL-1β and IL-18 are both activated through the inflammasome and increase IL-6 production. However, whereas *TET2* CHIP is associated with increased levels of IL-1β, *JAK2* and *SF3B1* CHIP are associated with increased IL-18 levels.

Second, our work highlights multiple mechanisms through which germline genetic variation can shape somatic variation in haematopoietic stem cells. One set of the germline loci is associated with increased propensity to acquire mutations owing to the failure of genes that maintain genome integrity (for example, *TERT* and *CHEK2*) and which have been implicated in stem cell maintenance or self-renewal[17]. These loci are associated with acquisition of somatic mutations resulting in neoplasm in multiple tissues. Other germline loci are associated with increased haematopoietic stem cell self-renewal (for example, *TET2*). Whereas the *TET2* locus is associated with increased risk of acquiring any CHIP driver mutations, it is not associated with cancer outside of the haematopoietic stem cell compartment. A third set of germline loci is associated with the acquisition of CHIP mutations in specific driver genes. This has previously been described in the *JAK2* 46/1 haplotype leading to *JAK2* p.V617F via a *cis*-haplotype effect[26–28]. We now identify a distinct *DNMT3A* CHIP-specific locus at the *TCL1A* promoter that is specifically associated with increased risk of *DNMT3A* CHIP, but not other CHIP subsets.

Thus, we have identified a convergence of common and rare germline genetic predisposition to leukocyte telomere length, MPNs, large-scale somatic chromosomal mosaicism and CHIP, suggesting shared causal mechanisms. So far, only CHIP with leukaemogenic driver mutations (as opposed to somatic chromosomal mosaicism[9] or CHIP with unknown driver mutations[13]) has been robustly associated with non-oncological diseases independently of age. The partially overlapping genetic predisposition that we observe across these three clonal phenomena suggests that although there may be similar genetic architectures 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 haematopoiesis itself.

Third, our work underscores the benefits of studying genomes from individuals of diverse ancestries. The inclusion of a large number of samples from individuals with African ancestry in TOPMed permitted the discovery of the *TET2* locus, which was not present in samples from individuals of other ancestries. Further inclusion of diverse individuals in genomic analyses will probably highlight other biological pathways.

Important limitations of our study include the reduced sensitivity for detecting CHIP with low allele fractions (VAF of 2–5%), even with high-coverage WGS. Ultrasensitive targeted sequencing can facilitate detection of such leukaemogenic mutations at exceedingly low VAFs, but the clinical consequences of this much more pervasive phenomenon, as well as determinants of progression to CHIP, are 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 previous observations indicating increased levels of inflammatory biomarkers as a consequence of CHIP in previous model experiments[2,5]. Finally, given the age dependence of CHIP, many individuals not observed to have CHIP in this study are likely to develop CHIP in future.

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

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2819-2.

1. Kennedy, B. K. et al. Geroscience: linking aging to chronic disease. *Cell* **159**, 709–713 (2014).
2. Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med*. **371**, 2488–2498 (2014).
3. Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med*. **371**, 2477–2487 (2014).
4. Xie, M. et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat. Med*. **20**, 1472–1478 (2014).
5. Jaiswal, S. et al. Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease. *N. Engl. J. Med*. **377**, 111–121 (2017).
6. Steensma, D. P. et al. Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood* **126**, 9–16 (2015).
7. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Preprint at https://doi.org/10.1101/563866 (2019).
8. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol*. **31**, 213–219 (2013).
9. Loh, P. R. et al. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350–355 (2018).
10. Patel, K. V. et al. Red cell distribution width and mortality in older adults: a meta-analysis. *J. Gerontol. A* **65**, 258–365 (2010).
11. Bick, A. G. et al. Genetic interleukin 6 signaling deficiency attenuates cardiovascular risk in clonal hematopoiesis. *Circulation* **141**, 124–131 (2020).
12. Alexandrov, L. B. et al. Clock-like mutational processes in human somatic cells. *Nat. Genet*. **47**, 1402–1407 (2015).
13. Zink, F. et al. Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* **130**, 742–752 (2017).
14. Bowman, R. L., Busque, L. & Levine, R. L. et al. Clonal hematopoiesis and evolution to hematopoietic malignancies. *Cell Stem Cell* **22**, 157–170 (2018).
15. Desai, P. et al. Somatic mutations precede acute myeloid leukemia years before diagnosis. *Nat. Med*. **24**, 1015–1023 (2018).
16. Bojesen, S. E. et al. Multiple independent variants at the *TERT* locus are associated with telomere length and risks of breast and ovarian cancer. *Nat. Genet*. **45**, 371–384 (2013).
17. Bao, E. L., et al. Inherited myeloproliferative neoplasm risk affects haematopoietic stem cells. *Nature* https://doi.org/10.1038/s41586-020-2786-7 (2020).
18. Zhou, W. et al. Mosaic loss of chromosome Y is associated with common variation near *TCL1A*. *Nat. Genet*. **48**, 563–568 (2016).
19. Hinds, D. A. et al. Germ line variants predispose to both *JAK2* V617F clonal hematopoiesis and myeloproliferative neoplasms. *Blood* **128**, 1121–1128 (2016).
20. Hu, Y. et al. A statistical framework for cross-tissue transcriptome-wide association analysis. *Nat. Genet*. **51**, 568–576 (2019).
21. Smith, B. W. et al. The aryl hydrocarbon receptor directs hematopoietic progenitor cell expansion and differentiation. *Blood* **122**, 376–385 (2013).
22. Cybulski, C. et al. CHEK2 is a multiorgan cancer susceptibility gene. *Am. J. Hum. Genet*. **75**, 1131–1135 (2004).
23. Rudd, M. F., Sellick, G. S., Webb, E. L., Catovsky, D. & Houlston, R. S. Variants in the *ATM–BRCA2–CHEK2* axis predispose to chronic lymphocytic leukemia. *Blood* **108**, 638–644 (2006).
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. *J. Biol. Chem*. **293**, 2452–2465 (2018).
25. Moran-Crusio, K. et al. *Tet2* loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. *Cancer Cell* **20**, 11–24 (2011).
26. Kilpivaara, O. et al. A germline *JAK2* SNP is associated with predisposition to the development of $JAK2^{V617F}$-positive myeloproliferative neoplasms. *Nat. Genet*. **41**, 455–459 (2009).
27. Jones, A. V. et al. *JAK2* haplotype is a major risk factor for the development of myeloproliferative neoplasms. *Nat. Genet*. **41**, 446–449 (2009).
28. Olcaydu, D. et al. A common *JAK2* haplotype confers susceptibility to myeloproliferative neoplasms. *Nat. Genet*. **41**, 450–454 (2009).
29. Young, A. L., Challen, G. A., Birmann, B. M. & Druley, T. E. Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. *Nat. Commun*. **7**, 12484 (2016).
30. Regier, A. A. et al. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nat. Commun*. **9**, 4038 (2018).
31. Jun, G., Wing, M. K., Abecasis, G. R. & Kang, H. M. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res*. **25**, 918–925 (2015).
32. Karczewski, K. J. et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *Nature* **581**, 434–443 (2020).
33. Gibson, C. J. et al. Clonal hematopoiesis associated with adverse outcomes after autologous stem-cell transplantation for lymphoma. *J. Clin. Oncol*. **35**, 1598–1605 (2017).
34. Hiatt, J. B., Pritchard, C. C., Salipante, S. J., O'Roak, B. J. & Shendure, J. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res*. **23**, 843–854 (2013).
35. Pérez Millán, M. I. et al. Next generation sequencing panel based on single molecule molecular inversion probes for detecting genetic variants in children with hypopituitarism. *Mol. Genet. Genomic Med*. **6**, 514–525 (2018).
36. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol*. **34**, 816–834 (2010).
37. Vattathil, S. & Scheet, P. Haplotype-based profiling of subtle allelic imbalance with SNP arrays. *Genome Res*. **23**, 152–158 (2013).
38. Fowler, J., San Lucas, F. A. & Scheet, P. System for quality-assured data analysis: flexible, reproducible scientific workflows. *Genet. Epidemiol*. **43**, 227–237 (2019).
39. Natarajan, P. et al. Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. *Nat. Commun*. **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).
41. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med*. **10**, 33 (2018).
42. Zhou, W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet*. **50**, 1335–1341 (2018).
43. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw*. http://doi.org/10.18637/jss.v067.i01 (2015).
44. Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
45. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE blacklist: identification of problematic regions of the genome. *Sci. Rep*. **9**, 9354 (2019).
46. Fulco, C. P. et al. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet*. **51**, 1664–1669 (2019).
47. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
48. Nasser, J. et al. Genome-wide maps of enhancer regulation connect risk variants to disease genes. Preprint at https://doi.org/10.1101/2020.09.01.278093 (2020).
49. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
50. DeLuca, D. S. et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**, 1530–1532 (2012).
51. eGTEx Project. Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease. *Nat. Genet*. **49**, 1664–1670 (2017).
52. Houseman, E. A. et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).
53. Horvath, S. & Levine, A. J. HIV-1 infection accelerates age according to the epigenetic clock. *J. Infect. Dis*. **212**, 1563–1573 (2015).
54. Barfield, R. T., Kilaru, V., Smith, A. K. & Conneely, K. N. CpGassoc: an R function for analysis of DNA methylation microarray data. *Bioinformatics* **28**, 1280–1281 (2012).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Alexander G. Bick[1,2,3,176,177], Joshua S. Weinstock[4,177], 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[32,33], Sharon L. R. Kardia[34], Jiang He[35,36], Robert C. Kaplan[37,38], Nicholas L. Smith[26,28,39], Donna K. Arnett[40], David A. Schwartz[41], Adolfo Correa[42], Mariza de Andrade[43], Xiuqing Guo[44], Barbara A. Konkle[45,46], Brian Custer[47,48], Juan M. Peralta[49], Hongsheng Gui[50], Deborah A. Meyers[51], Stephen T. McGarvey[52], Ida Yii-Der Chen[53], M. Benjamin Shoemaker[54], Patricia A. Peyser[34], 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[55], Kari E. North[56], Lenore J. Launer[57], Brian E. Cade[23,58], Joshua C. Bis[25], Michael H. Cho[3,29], Jessica Lasky-Su[3,29], Donald W. Bowden[30], L. Adrienne Cupples[59], Angel C. Y. Mak[32], Lewis C. Becker[60], Jennifer A. Smith[34,61], Tanika N. Kelly[35,36], Stella Aslibekyan[62], Susan R. Heckbert[26,28], Hemant K. Tiwari[63], Ivana V. Yang[41], John A. Heit[64], Steven A. Lubitz[2,3,65], Jill M. Johnsen[45,46], Joanne E. Curran[49], Sally E. Wenzel[66], Daniel E. Weeks[67], Dabeeru C. Rao[68], Dawood Darbar[69], Jee-Young Moon[37], Russell P. Tracy[70], Erin J. Buth[12], Nicholas Rafaels[19], Ruth J. F. Loos[20,71], Peter Durda[70], Yongmei Liu[72], Lifang Hou[73], Jiwon Lee[23], Priyadarshini Kachroo[3,29], Barry I. Freedman[74], Daniel Levy[75,76], Lawrence F. Bielak[34], James E. Hixson[77], James S. Floyd[25,26,78], Eric A. Whitsel[79,80], Patrick T. Ellinor[2,3,65], Marguerite R. Irvin[62], Tasha E. Fingerlin[81], Laura M. Raffield[82], Sebastian M. Armasu[83], Marsha M. Wheeler[84], Ester C. Sabino[85], John Blangero[49], L. Keoki Williams[50], Bruce D. Levy[3,86], Wayne Huey-Herng Sheu[87], Dan M. Roden[88,89,90], Eric Boerwinkle[22,91], JoAnn E. Manson[3,92,93], Rasika A. Mathias[60], Pinkal Desai[94], Kent D. Taylor[95,96], Andrew D. Johnson[75,76], NHLBI Trans-Omics for Precision Medicine Consortium*, Paul L. Auer[97], Charles Kooperberg[98], Cathy C. Laurie[12], Thomas W. Blackwell[4], Albert V. Smith[4], Hongyu Zhao[99,100], Ethan Lange[9], Leslie Lange[9], Stephen S. Rich[101], Jerome I. Rotter[95,96], James G. Wilson[102,103], Paul Scheet[11], Jacob O. Kitzman[13,104], Eric S. Lander[2,6,105], Jesse M. Engreitz[2,106], Benjamin L. Ebert[2,3,14,107], Alexander P. Reiner[26,98], Siddhartha Jaiswal[108], Gonçalo Abecasis[4,109], Vijay G. Sankaran[2,3,5], Sekar Kathiresan[2,3,110,111,178] ✉ & Pradeep Natarajan[2,3,65,178] ✉

[1]Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. [2]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [3]Harvard Medical School, Boston, MA, USA. [4]Center for Statistical Genetics, Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI, USA. [5]Division of Hematology/Oncology, Boston Children's Hospital and Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [6]Department of Systems Biology, Harvard Medical School, Boston, MA, USA. [7]Health Sciences and Technology Program, Harvard Medical School, Boston, MA, USA. [8]Yale School of Medicine, New Haven, CT, USA. [9]Division of Biomedical Informatics and Personalized Medicine, Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [10]Medical Scientist Training Program, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [11]Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [12]Department of Biostatistics, University of Washington, Seattle, WA, USA. [13]Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA. [14]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [15]Division of Cardiovascular Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. [16]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [17]Department of Medicine, University of Maryland School of Medicine, Baltimore, MD, USA. [18]Geriatrics Research and Education Clinical Center, Baltimore Veterans Administration Medical Center, Baltimore, MD, USA. [19]Colorado Center for Personalized Medicine, School of Medicine, University of Colorado, Aurora, CO, USA. [20]Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [21]Brown Foundation Institute of Molecular Medicine, McGovern Medical School, University of Texas Health Science Center at Houston, Houston, TX, USA. [22]Human Genetics Center, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX, USA. [23]Division of Sleep and Circadian Disorders, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. [24]Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA. [25]Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA. [26]Department of Epidemiology, University of Washington, Seattle, WA, USA. [27]Department of Health Services, University of Washington, Seattle, WA, USA. [28]Kaiser Permanente Washington Health Research Institute, Seattle, WA, USA. [29]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. [30]Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC, USA. [31]Departments of Medicine and Epidemiology, Boston University School of Medicine, Boston, MA, USA. [32]Department of Medicine, University of California, San Francisco, San Francisco, CA, USA. [33]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, USA. [34]Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI, USA. [35]Department of Epidemiology, Tulane University School of Public Health and Tropical Medicine, New Orleans, LA, USA. [36]Tulane University Translational Science Institute, New Orleans, LA, USA. [37]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, New York, NY, USA. [38]Fred Hutchinson Cancer Research Center, Division of Public Health Sciences, Seattle, WA, USA. [39]Seattle Epidemiologic Information and Research Center, Department of Veterans Affairs, Office of Research and Development, Seattle, WA, USA. [40]College of Public Health, University of Kentucky, Lexington, KY, USA. [41]Department of Medicine, University of Colorado, Aurora, CO, USA. [42]Departments of Medicine and Population Health Science, University of Mississippi Medical Center, Jackson, MS, USA. [43]Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA. [44]Institute for Translational Genomics and Population Sciences, Department of Pediatrics, Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center, Torrance, CA, USA. [45]Bloodworks Northwest, Seattle, WA, USA. [46]Department of Medicine, University of Washington, Seattle, WA, USA. [47]Vitalant Research Institute, San Francisco, CA, USA. [48]Department of Laboratory Medicine, University of California, San Francisco, San Francisco, CA, USA. [49]Department of Human Genetics and South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley School of Medicine, Brownsville, TX, USA. [50]Center for Individualized and Genomic Medicine Research, Department of Internal Medicine, Henry Ford Health System, Detroit, MI, USA. [51]Division of Genetics, Genomics and Precision Medicine, University of Arizona, Tucson, AZ, USA. [52]Department of Epidemiology and International Health Institute, Brown University School of Public Health, Providence, RI, USA. [53]Medical Genetics, Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center, Los Angeles, CA, USA. [54]Division of Cardiology, Vanderbilt University Medical Center, Nashville, TN, USA. [55]Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [56]Department of Epidemiology, University of North Carolina, Chapel Hill, NC, USA. [57]Laboratory of Epidemiology, Demography, and Biometry, Intramural Research Program, National Institute on Aging, Bethesda, MD, USA. [58]Division of Sleep Medicine, Department of Medicine, Harvard Medical School, Boston, MA, USA. [59]Departments of Biostatistics and Epidemiology, Boston University School of Public Health, Boston, MA, USA. [60]GeneSTAR Research Program, Department of Medicine, Johns Hopkins School of Medicine, Baltimore, MD, USA. [61]Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI, USA. [62]Department of Epidemiology, University of Alabama at Birmingham, Birmingham, AL, USA. [63]Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL, USA. [64]Division of Cardiovascular Diseases, Mayo Clinic, Rochester, MN, USA. [65]Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA. [66]Department of Environmental and Occupational Health, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA. [67]Departments of Human Genetics and Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA. [68]Division of Biostatistics, Washington University School of Medicine, St Louis, MO, USA. [69]Division of Cardiology, University of Illinois at Chicago, Chicago, IL, USA. [70]Department of Pathology and Laboratory Medicine, Larner College of Medicine, University of Vermont, Burlington, VT, USA. [71]Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [72]Division of Cardiology, Department of Medicine, Duke University Medical Center, Durham, NC, USA. [73]Department of Preventive Medicine, Northwestern Feinberg School of Medicine, Northwestern University, Chicago, IL, USA. [74]Department of Internal Medicine, Section on Nephrology, Wake Forest School of Medicine, Winston-Salem, NC, USA. [75]Framingham Heart Study, Framingham, MA, USA. [76]Population Sciences Branch, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD, USA. [77]Department of Epidemiology, Human Genetics and Environmental Sciences, University of Texas Health Science Center at Houston School of Public Health, Houston, TX, USA. [78]Department of Medicine, University of Washington, Seattle, WA, USA. [79]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC, USA. [80]Department of Medicine, School of Medicine, University of North Carolina, Chapel Hill, NC, USA. [81]Center for Genes Environment and Health, National Jewish Health, Denver, CO, USA. [82]Department of Genetics, University of North Carolina, Chapel Hill, NC, USA. [83]Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA. [84]Department of Genome Science, University of Washington, Seattle, WA, USA. [85]Instituto de Medicina Tropical, Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil. [86]Division of Pulmonary and Critical Care Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. [87]Division of Endocrinology and Metabolism, Department of Internal Medicine, Taichung Veterans General Hospital, Taichung, Taiwan. [88]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA. [89]Department of Pharmacology, Vanderbilt University Medical Center, Nashville, TN, USA. [90]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA. [91]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA. [92]Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. [93]Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA. [94]Department of Medicine, Weill Cornell Medical School, New York, NY, USA. [95]Institute for Translational Genomics and Population Sciences, Lundquist Institute for Biomedical Innovation, Harbor-UCLA Medical Center, Torrance, CA, USA. [96]Department of Pediatrics, Harbor-UCLA Medical Center, Torrance, CA, USA. [97]Joseph J. Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI, USA. [98]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [99]Department of Biostatistics, School of Public Health, Yale University, New Haven, CT, USA. [100]Computational Biology and Bioinformatics Program, Yale University, New Haven, CT, USA. [101]Department of Public Health Sciences, Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA. [102]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS, USA. [103]Department of Cardiology, Beth Israel Deaconess Medical Center, Boston, MA, USA. [104]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA. [105]Department of Biology, MIT, Cambridge, MA, USA. [106]Harvard Society of Fellows, Harvard University, Cambridge, MA, USA. [107]Howard Hughes Medical Institute, Boston, MA, USA. [108]Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA. [109]Regeneron Pharmaceuticals, Tarrytown, NY, USA. [110]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [111]Verve Therapeutics, Cambridge, MA, USA. [176]Present address: Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA. [177]These authors contributed equally: Alexander G. Bick, Joshua S. Weinstock. [178]These authors jointly supervised this work: Sekar Kathiresan, Pradeep Natarajan. *A list of members and their affiliations appears in the online version of the paper. ✉e-mail: sekar@broadinstitute.org; pradeep@broadinstitute.org

NHLBI Trans-Omics for Precision Medicine Consortium

Namiko Abe[112], Christine Albert[92], Laura Almasy[113], Alvaro Alonso[114], Seth Ament[115], Peter Anderson[116], Pramod Anugu[117], Deborah Applebaum-Bowden[118], Dan Arking[119], Allison Ashley-Koch[120], Stella Aslibekyan[121], Tim Assimes[122], Dimitrios Avramopoulos[119], John Barnard[123], R. Graham Barr[124], Emily Barron-Casella[119], Lucas Barwick[125], Terri Beaty[119], Gerald Beck[123], Diane Becker[60], Rebecca Beer[118], Amber Beitelshees[115], Emelia Benjamin[31], Panagiotis Benos[126], Marcos Bezerra[127], Larry Bielak[4], Russell Bowler[81], Jennifer Brody[116], Ulrich Broeckel[128], Karen Bunting[112], Carlos Bustamante[122], Jonathan Cardwell[129], Vincent Carey[92], Cara Carty[130], Richard Casaburi[131], James Casella[119], Peter Castaldi[22,91], Mark Chaffin[2], Christy Chang[115], Yi-Cheng Chang[132], Daniel Chasman[92], Sameer Chavan[129], Bo-Juen Chen[112], Wei-Min Chen[133], Seung Hoan Choi[2], Lee-Ming Chuang[132], Mina Chung[123], Ren-Hua Chung[134], Clary Clish[2], Suzy Comhair[12], Elaine Cornell[70], Carolyn Crandall[131], James Crapo[135], Jeffrey Curtis[4], Coleen Damcott[115], Sayantan Das[4], Sean David[136], Colleen Davis[116], Michael DeBaun[137], Ranjan Deka[138], Dawn DeMeo[139], Scott Devine[115], Qing Duan[140], Ravi Duggirala[141], Susan Dutcher[139], Charles Eaton[142], Lynette Ekunwe[117], Adel El Boueiz[29], Leslie Emery[116], Serpil Erzurum[123], Charles Farber[46], Matthew Flickinger[4], Nora Franceschini[140], Chris Frazar[26], Mao Fu[115], Stephanie M. Fullerton[116], Lucinda Fulton[139], Stacey Gabriel[4], Weiniu Gan[118], Shanshan Gao[129], Yan Gao[117], Margery Gass[38], Bruce Gelb[143], Xiaoqi (Priscilla) Geng[4], Mark Geraci[144], Soren Germer[112], Robert Gerszten[3], Auyon Ghosh[22,91], Richard Gibbs[91], Chris Gignoux[19], Mark Gladwin[126], David Glahn[8], Da-Wei Gong[115], Harald Goring[141], Sharon Graw[129], Daniel Grine[129], C. Charles Gu[139], Yue Guan[115], Namrata Gupta[3], Jeff Haessler[129], Michael Hall[117], Daniel Harris[115], Nicola L. Hawley[145], Ben Heavner[12], Ryan Hernandez[146], David Herrington[147], Craig Hersh[29], Bertha Hidalgo[62], Brian Hobbs[22,91], John Hokanson[129], Elliott Hong[115], Karin Hoth[148], Chao (Agnes) Hsiung[134], Yi-Jen Hung[149], Haley Huston[45], Chii Min Hwu[150], Rebecca Jackson[151], Deepti Jain[12], Cashell Jaquish[118], Min A. Jhun[4], Craig Johnson[116], Rich Johnston[114], Kimberly Jones[119], Hyun Min Kang[12], Shannon Kelly[60], Michael Kessler[115], Alyna Khan[116], Wonji Kim[29], Greg Kinney[129], Holly Kramer[152], Christoph Lange[153], Meryl LeBoff[48], Seunggeun Shawn Lee[4], Wen-Jane Lee[150], Jonathon LeFaive[4], David Levine[116], Joshua Lewis[115], Xiaohui Li[154], Yun Li[140], Henry Lin[154], Honghuang Lin[155], Keng Han Lin[4], Xihong Lin[153], Simin Liu[156], Yu Liu[157], Kathryn Lunetta[155], James Luo[118], Michael Mahaney[141], Barry Make[119], Ani Manichaikul[97], Lauren Margolin[2], Lisa Martin[113], Susan Mathai[129], Susanne May[12], Patrick McArdle[115], Merry-Lynn McDonald[121], Sean McFarland[2,3,5], Daniel McGoldrick[158], Caitlin McHugh[12], Hao Mei[117], Luisa Mestroni[129], Julie Mikulla[118], Nancy Min[117], Mollie Minear[118], Ryan L. Minster[126], Matt Moll[86], Courtney Montgomery[159], Solomon Musani[42], Stanford Mwasongwe[117], Josyf C. Mychaleckyj[133], Girish Nadkarni[160], Rakhi Naik[119], Take Naseri[161], Sergei Nekhai[162], Sarah C. Nelson[12], Bonnie Neltner[129], Deborah Nickerson[158], Jeff O'Connell[115], Tim O'Connor[115], Heather Ochs-Balcom[163], David Paik[157], James Pankow[164], George Papanicolaou[118], Afshin Parsa[115], Marco Perez[122], James Perry[115], Ulrike Peters[98], Patricia Peyser[4], Lawrence S. Phillips[114], Toni Pollin[115], Wendy Post[119], Julia Powers Becker[129], Meher Preethi Boorgula[129], Michael Preuss[20], Pankaj Qasba[118], Dandi Qiao[29], Zhaohui Qin[114], Laura Rasmussen-Torvik[73], Aakrosh Ratan[133], Robert Reed[115], Elizabeth Regan[135], Muagututi'a Sefuiva Reupena[138], Ken Rice[12], Carolina Roselli[2], Ingo Ruczinski[119], Pamela Russell[129], Sarah Ruuska[45], Kathleen Ryan[115], Danish Saleheen[124], Shabnam Salimi[115], Steven Salzberg[119], Kevin Sandow[154], Christopher Scheller[4], Ellen Schmidt[4], Karen Schwander[139], Frank Sciurba[126], Christine Seidman[2,3,15,107], Jonathan Seidman[2,3], Vivien Sheehan[165], Stephanie L. Sherman[166], Amol Shetty[115], Aniket Shetty[129], Brian Silver[167], Josh Smith[116], Tanja Smith[112], Sylvia Smoller[37], Beverly Snively[147], Michael Snyder[122], Tamar Sofer[58], Nona Sotoodehnia[116], Adrienne M. Stilp[116], Garrett Storm[129], Elizabeth Streeten[115], Jessica Lasky Su[29], Yun Ju Sung[139], Jody Sylvia[48], Adam Szpiro[116], Carole Sztalryd[115], Daniel Taliun[4], Hua Tang[122], Matthew Taylor[129], Simeon Taylor[115], Marilyn Telen[120], Timothy A. Thornton[116], Machiko Threlkeld[158], Lesley Tinker[129], David Tirschwell[116], Sarah Tishkoff[168], Hemant Tiwari[63], Catherine Tong[12], Michael Tsai[164], Dhananjay Vaidya[119], David Van Den Berg[169], Peter VandeHaar[4], Scott Vrieze[170], Tarik Walker[129], Robert Wallace[148], Avram Walts[129], Heming Wang[23], Karol Watson[131], Bruce Weir[116], Lu-Chen Weng[2,65], Jennifer Wessel[171], Cristen Willer[172], Kayleen Williams[116], Carla Wilson[135], Joseph Wu[157], Huichun Xu[115], Lisa Yanek[119], Rongze Yang[115], Norann Zaghloul[115], Yingze Zhang[173], Snow Xueyan Zhao[32], Wei Zhao[2], Degui Zhi[174], Xiang Zhou[4], Xiaofeng Zhu[175], Michael Zody[112] & Sebastian Zoellner[4]

[112]New York Genome Center, New York, NY, USA. [113]Children's Hospital of Philadelphia, University of Pennsylvania, Philadelphia, PA, USA. [114]Emory University, Atlanta, GA, USA. [115]University of Maryland, Baltimore, MD, USA. [116]University of Washington, Seattle, WA, USA. [117]University of Mississippi, Jackson, MS, USA. [118]National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD, USA. [119]Johns Hopkins University, Baltimore, MD, USA. [120]Duke University, Durham, NC, USA. [121]University of Alabama, Birmingham, AL, USA. [122]Stanford University, Stanford, CA, USA. [123]Cleveland Clinic, Cleveland, OH, USA. [124]Columbia University, New York, NY, USA. [125]The Emmes Corporation, LTRC, Rockville, MD, USA. [126]University of Pittsburgh, Pittsburgh, PA, USA. [127]Fundação de Hematologia e Hemoterapia de Pernambuco (HEMOPE), Recife, Brazil. [128]Medical College of Wisconsin, Milwaukee, WI, USA. [129]University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [130]Washington State University, Seattle, WA, USA. [131]University of California, Los Angeles, Los Angeles, CA, USA. [132]National Taiwan University Hospital, Taipei, Taiwan. [133]University of Virginia, Charlottesville, VA, USA. [134]National Health Research Institute, Zhunan, Taiwan, USA. [135]National Jewish Health, Denver, CO, USA. [136]University of Chicago, Chicago, IL, USA. [137]Vanderbilt University, Nashville, TN, USA. [138]University of Cincinnati, Cincinnati, OH, USA. [139]Washington University in St Louis, St Louis, MO, USA. [140]University of North Carolina, Chapel Hill, NC, USA. [141]University of Texas Rio Grande Valley School of Medicine, Edinburg, TX, USA. [142]Brown University, Providence, RI, USA. [143]Icahn School of Medicine at Mount Sinai, New York, NY, USA. [144]Indiana University, Medicine, Indianapolis, IN, USA. [145]Department of Chronic Disease Epidemiology, Yale University, New Haven, CT, USA. [146]McGill University, Montreal, Quebec, Canada. [147]Wake Forest Baptist Health, Winston-Salem, NC, USA. [148]University of Iowa, Iowa City, IA, USA. [149]Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan. [150]Taichung Veterans General Hospital Taiwan, Taichung City, Taiwan. [151]Division of Endocrinology, Diabetes and Metabolism, Department of Medicine, Ohio State University Wexner Medical Center, Columbus, OH, USA. [152]Loyola University, Public Health Sciences, Maywood, IL, USA. [153]Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA. [154]Lundquist Institute, Torrance, CA, USA. [155]Boston University, Boston, MA, USA. [156]Department of Epidemiology, Brown University, Providence, RI, USA. [157]Cardiovascular Institute, Stanford University, Palo Alto, CA, USA. [158]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [159]Oklahoma Medical Research Foundation, Genes and Human Disease, Oklahoma City, OK, USA. [160]Division of Nephrology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [161]Ministry of Health, Government of Samoa, Apia, Samoa. [162]Howard University, Washington, DC, USA. [163]University at Buffalo, Buffalo, NY, USA. [164]University of Minnesota, Minneapolis, MN, USA. [165]Department of Pediatrics, Baylor College of Medicine, Houston, TX, USA. [166]Department of Human Genetics, Emory University, Atlanta, GA, USA. [167]UMass Memorial Medical Center, Worcester, MA, USA. [168]Department of Genetics, University of Pennsylvania, Philadelphia, PA, USA. [169]USC Methylation Characterization Center, University of Southern California, Los Angeles, CA, USA. [170]Department of Psychology, University of Minnesota, Minneapolis, MN, USA. [171]Department of Epidemiology, Indiana University, Indianapolis, IN, USA. [172]Department of Medicine, University of Michigan, Ann Arbor, MI, USA. [173]Department of Medicine, University of Pittsburgh, Pittsburgh, PA, USA. [174]Center for Precision Health, School of Biomedical Informatics, University of Texas Health at Houston, Houston, TX, USA. [175]Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH, USA.

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Study samples

WGS was performed on 97,691 samples sequenced as part of 51 studies contributing to the NHLBI TOPMed research programme Freeze 6 release as previously described for discovery analyses[7]. An additional distinct set of 9,389 WGS 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 that have been assessed for many diverse phenotypes. None of the studies that comprise TOPMed selected individuals for sequencing on the basis of haematological malignancy. Approximately 82% of participants are US residents with diverse ancestry and ethnicity (40% European, 32% African, 16% Hispanic or Latino and 10% Asian). Each of the constituent studies provided informed consent for 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 comprising 82,807 of the samples. The median age was 55, the mean age was 52.5, and the maximum age was 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 before 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]. Single nucleotide polymorphisms (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). In brief, Mutect2 searches for sites where there is evidence for variation and then performs local reassembly. It uses an external reference of recurrent sequencing artefacts 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 before 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 reanalysed 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 WGS. 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 Fig. 10).

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, totalling 44.5 kb. Probes were synthesized as a pool by CustomArray, 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 genomic DNA 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 >1 Mb. 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 the database of genomic variants and (2) was not determined to be a deletion or Log-R ratio 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: haemoglobin, haematocrit, red blood cell count, white blood cell count, basophil count, eosinophil count, neutrophil count, lymphocyte count, monocyte count, platelet count, mean corpuscular haemoglobin, mean corpuscular haemoglobin 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 used for this analysis that had one or more blood traits measured concurrently or following the blood draw used for CHIP ascertainment. Traits were first $\log_2$ normalized and then analysed using a general linear regression model with CHIP status, age, sex, study and the first ten 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^{-1} 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 used 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 analysed 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 analysed including: CD-40, CRP, E-Selectin, ICAM-1, IL-1β, IL-6, IL-10, IL-18, 8-epi PGF2a, Lp-PLA2 mass and activity, MCP1, MMP9, MPO, OPG, P-selectin, TNF, TNF receptor 1 and 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 used 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 $\log_2(x+1)$ transformation and then analysed 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 (https://bravo.sph.umich.edu/freeze5/hg38/). Multiallelic variants and indels were excluded. We used the COSMIC signature file (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[12,40,41]. 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 (that is, one of cases or controls contains at least 99% of the signature observations).

Single variant association

Single variant association for each variant 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 centred 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 modelled with a spline to capture the nonlinearity 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 linkage disequilibrium matrices as input. We fine-mapped the TET2 locus using the summary statistics from the African ancestry single variant summary statistics and estimated linkage disequilibrium 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 cut-off of 2.9×10^{-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 loss-of-function variants as identified by snpEff. Given that some variants were present in both the Mutect2 calls and the germline variant calls, we pruned the loss-of-function 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 ABC model[45].

Predicting enhancer-gene regulation for TET2

We used the ABC model[46] to predict which enhancers regulate specific genes in CD34$^+$ haematopoietic progenitor cells, with minor modifications as follows.

In brief, this model predicts the effect of each putative regulatory element (defined as a DNase peak within 5 Mb 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$^+$ haematopoietic 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 $P < 0.1$. 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 centred 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 centred 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 (https://egg2.wustl.edu/roadmap/data/byFileType/alignments/consolidated/E050-DNase.tagAlign.gz, and https://egg2.wustl.edu/roadmap/data/byFileType/alignments/consolidated/E050-H3K27ac.tagAlign.gz).

We estimated enhancer-promoter Contact from the Knight–Ruiz (KR)-normalized Hi-C contact maps in primary CD34$^+$ cells. We then calculated effect of each putative enhancer (E)–gene (G) connection by multiplying the activity (A) and contact (C) for that element (e) and gene. Dividing the effect of each element by the sum of effects for all elements for a given gene yields the ABC score[46]:

$$\mathrm{ABC}_{\mathrm{E-G}} = \frac{A_{\mathrm{E}} \times C_{\mathrm{E-G}}}{\sum_{\mathrm{e\ within\ 5Mb}} A_{\mathrm{e}} \times C_{\mathrm{e-G}}}$$

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% (ref. [48]).

Functional evaluation of *TET2* locus

The genomic region containing risk and non-risk allele of the variant rs79901204 (600 bp) was synthesized as gblocks (IDT Technologies) and cloned into the Firefly luciferase reporter constructs (pGL4.24) using NheI and EcoRV sites. The Firefly constructs (500 ng) were co-transfected with pRL-SV40 *Renilla* luciferase constructs (50 ng) into 100,000 K562 cells (ATCC) using Lipofectamine LTX (Invitrogen) according to manufacturer's protocols. Cells were harvested after 48 h 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 h before electroporation and cultured in hematopoietic stem cell expansion conditions throughout the experiment (Stemspan II medium with CC100 cytokine cocktail from Stem Cell Technologies and thrombopoietin (50 ng μl^{-1}) 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 Lonza 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 were used simultaneously (ENH_sgRNA_1: GGATTCTGTATTCGTCTGTG and ENH_sgRNA_2: TCTACTCACAGGGCCCAATG). For *TET2* coding-disruption experiments single guides were used (TET2_CDS1: TGGAGAAAGACGTAACTTCG and 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

Three days after RNP electroporation, 500 CD34$^+$ HSPCs were plated in 1 ml methylcellulose medium (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 3 pates were pooled, washed with PBS, and the cells were plated in new methylcellulose medium at 25,000 cells per ml for an additional 2 weeks.

RNA-seq and eQTL analysis

RNA-seq 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[49]. Gene Quantification and quality control was performed using RNA-SeQC 1.19[50]. 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 at https://www.nhlbiwgs.org/sites/default/files/TOPMed_RNAseq_pipeline_COREyr2.pdf.

Analysis was performed using samples from 247 African Americans from the MESA cohort Exam 1. Transcript expression was converted to TPM units and $\log_2$ transformed for analysis consistent with the GTEx consortium[51] 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 the *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*[52] using normal-exponential out-of-band (noob) background correction method with dye-bias normalization. Samples (1,747 total: 1,097 women and 650 men) remained after severe outliers were identified and removed. Seventy-one 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 using ComBat[53]. Relative leukocyte cell counts (CD8$^+$ T lymphocytes, CD4$^+$ T lymphocytes, natural killer cells, B cells, monocytes and granulocytes) were estimated as previously described[53,54].

To investigate methylation in the *TET2* locus, a linear mixed effects model was fitted using CpGassoc[54] 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 $P = 5.8 \times 10^{-8}$ was used to establish statistical significance.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Individual WGS 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 programme. A subset of the TOPMed cohorts analysed here is based on sensitive populations, precluding public sharing of full genomic summary results.

Acknowledgements Investigators who conducted this research report individual research support from R35 HL135818 (S.R.), P01 HL132825 (S.T.W.), R01 HL091357 and R01 HL055673 (D.K.A.), W81 XWH-17-1-0597 (D.A.S.), K01 HL135405 (B.E.C.), P01 HL132825 (J.L.-S.), K01HL136700 (S.A.), R01 HL113323 (J.E.C.), R01HL1333040 (D.E.W.), R01 HL138737 (D.D.), P01 HL132825 (P.K.), T32 HL129982 (L.M.R.), R01 HL113323 (J.B.), HHS-N268201800002I (T.W.B. and A.V.S.), U54 GM115428 (J.G.W.), R01 HL148565 and R01 HL148050 (P.N.), F30 HL149180 (S.M.Z.), R01 HL139731 and AHA-18SFRN34250007 (S.L.), DP5 OD029586 (A.G.B.), Claudia Adams Barr Program for Innovative Cancer Research (V.G.S.), R01 142711, MGH Hassenfeld Scholar Award (P.N.), Fondation Leducq TNE-18CVD04 (A.G.B., B.L.E., S.J., P.N. and S.K.), Burroughs Wellcome Foundation (A.G.B. and S.J.), Ludwig Cancer Center (S.J.) and UM1-HG008895 (S.K.). WGS for the Trans-Omics in Precision Medicine (TOPMed) programme 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. 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.

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 WGS 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.O.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.-D.C., 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., C.L., N.R., R.J.F.L., P. Durda, 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.-H.S., D.M.R., E.B., J.E.M., R.A.M., P. Desai, 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 analysed here. A.G.B., J.S.W., S.K. and P.N. wrote the manuscript with input from all authors.

Competing interests B.M.P. serves on the Steering Committee of the Yale Open Data Access Project funded by Johnson & Johnson. E.K.S. and M.H.C. received grant support from GlaxoSmithKlein and Bayer. S.T.W. received royalties from UpToDate. S.A. reports employment and equity in 23andMe. B.I.F. is a consultant for RenalytixAI and AstraZeneca Pharmaceuticals. M.E.M. reports funding from Regeneron Pharmaceuticals, unrelated to this project. M.H.C. has received grant support from GlaxoSmithKlein and Bayer and consulting or speaking fees from AstraZeneca and Illumina. J.S.F. has consulted for Shionogi. B.D.L. 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.S.L. 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.L.E. reports grant support from Celgene and Deerfield. P.T.E. has received grant support from Bayer AG and has served on advisory boards or consulted for Bayer AG, Quest Diagnostics, MyoKardia and Novartis. G.A. is an employee of Regeneron Pharmaceuticals and owns stock and stock options for Regeneron Pharmaceuticals. S.J. is a scientific advisor to Grail. S.L. receives sponsored research support from Bristol Myers Squibb, Pfizer, Bayer AG, Boehringer Ingelheim and Fitbit; has consulted for Bristol Myers Squibb, Pfizer and Bayer AG; and participates in a research collaboration with IBM. P.N. reports grants support from Amgen, Apple and Boston Scientific, and is a scientific advisor to Apple. S.K. is an employee of Verve Therapeutics, and holds equity in Verve Therapeutics, Maze Therapeutics, Catabasis and San Therapeutics; is a member of the scientific advisory boards for Regeneron Genetics Center and Corvidia Therapeutics; and 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 other authors declare no competing interests.

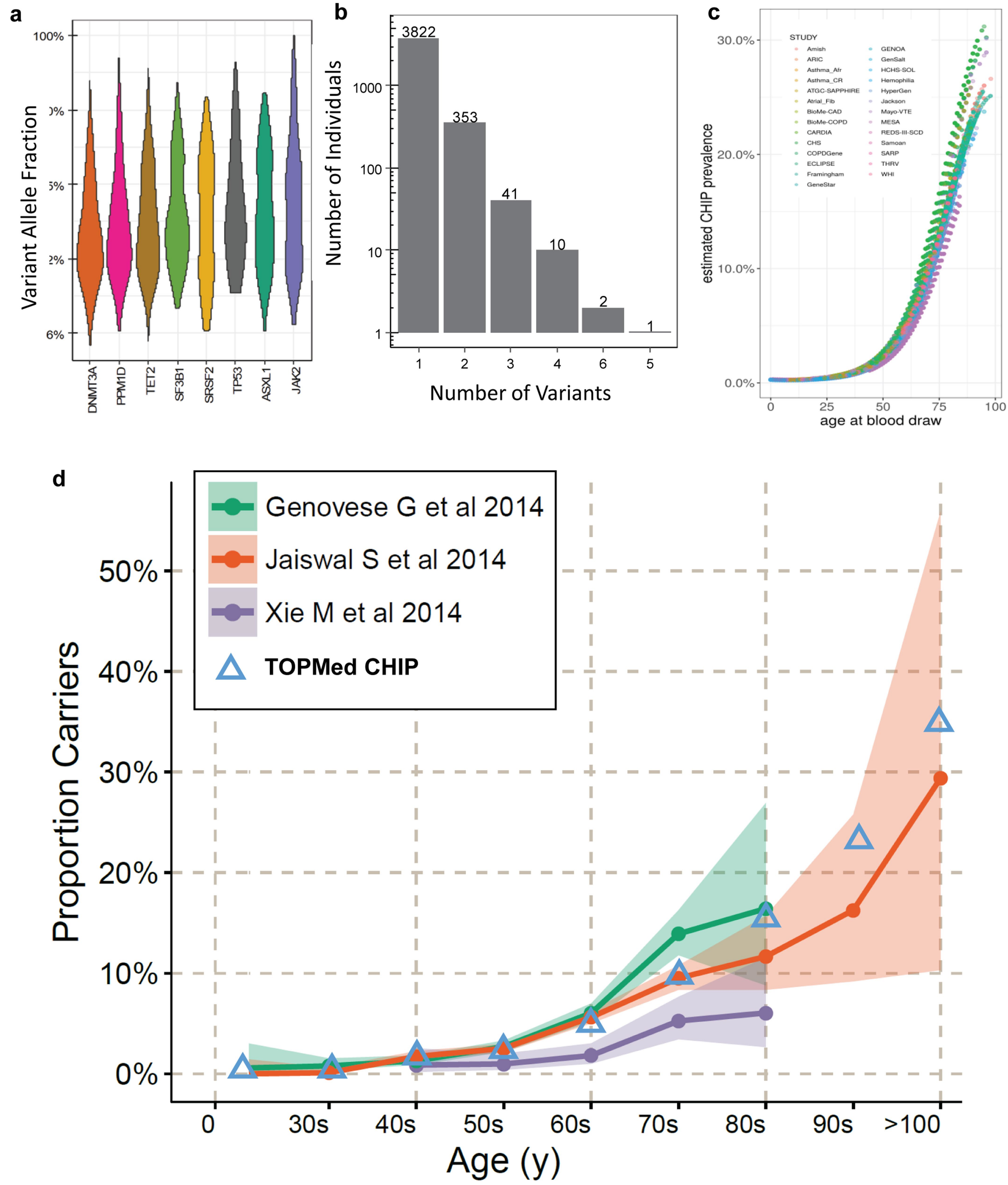
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2819-2.

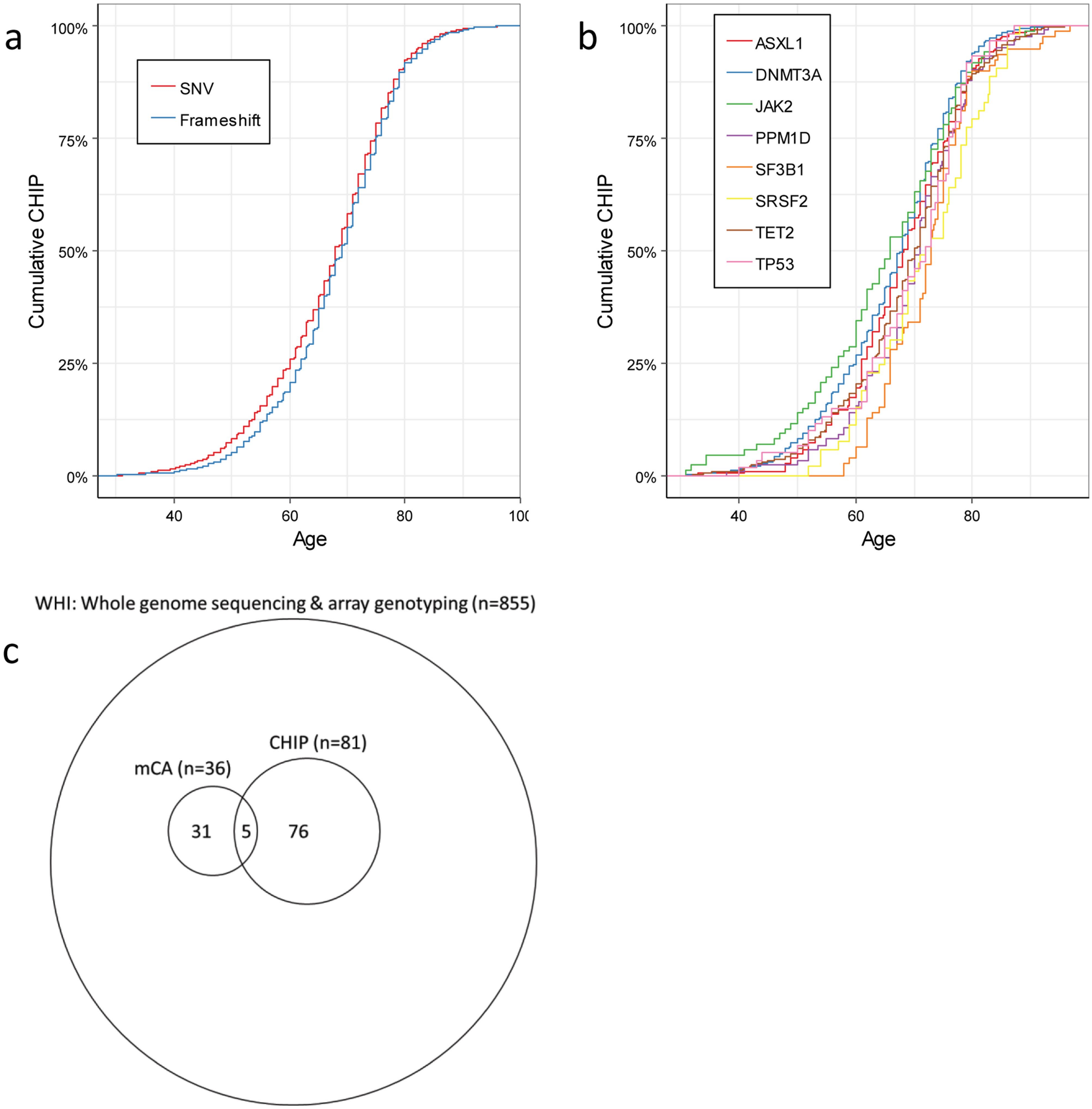
Correspondence and requests for materials should be addressed to S.K. or P.N.

Peer review information *Nature* thanks Stephen Chanock, Ross Levine and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

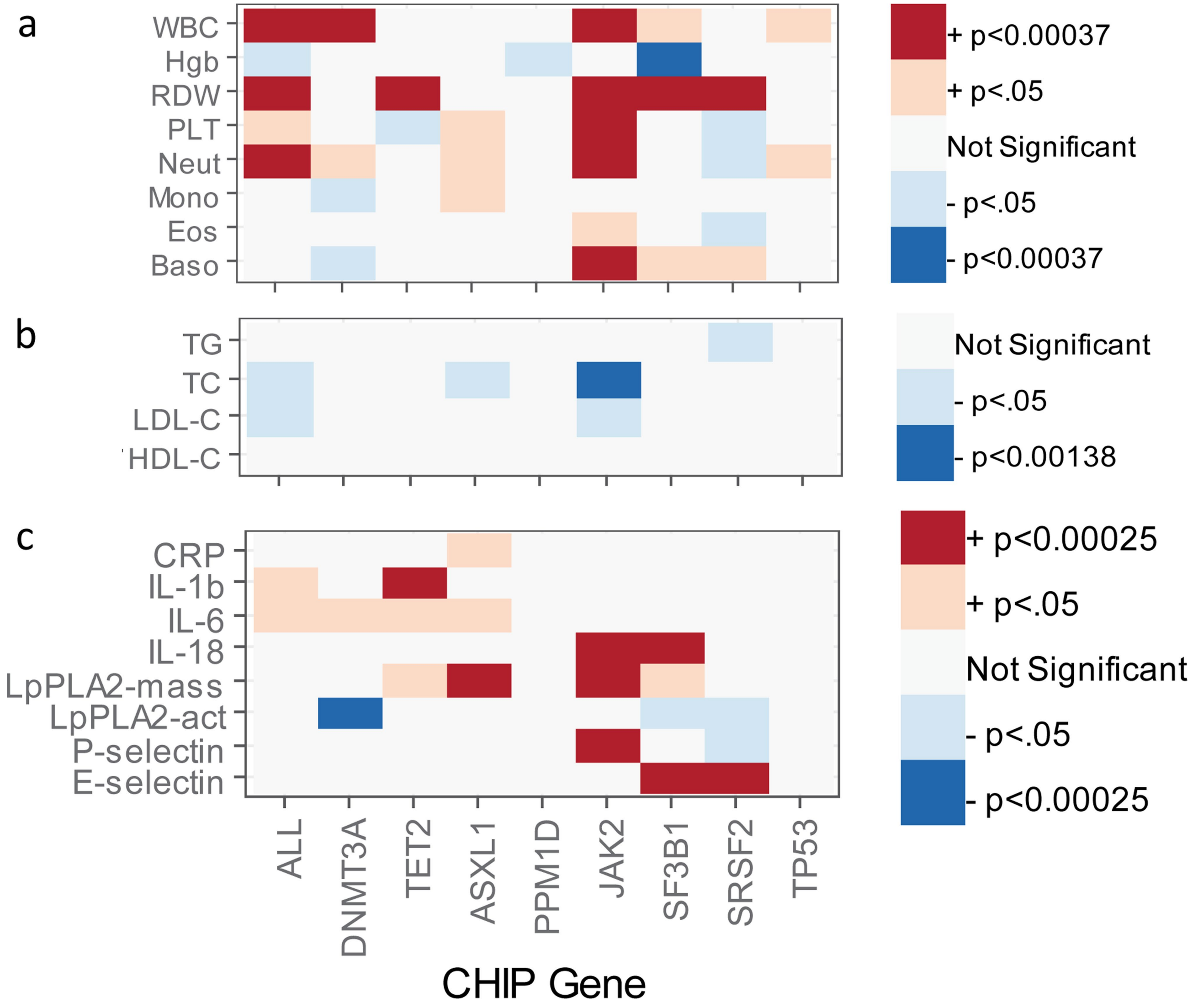
Reprints and permissions information is available at http://www.nature.com/reprints.



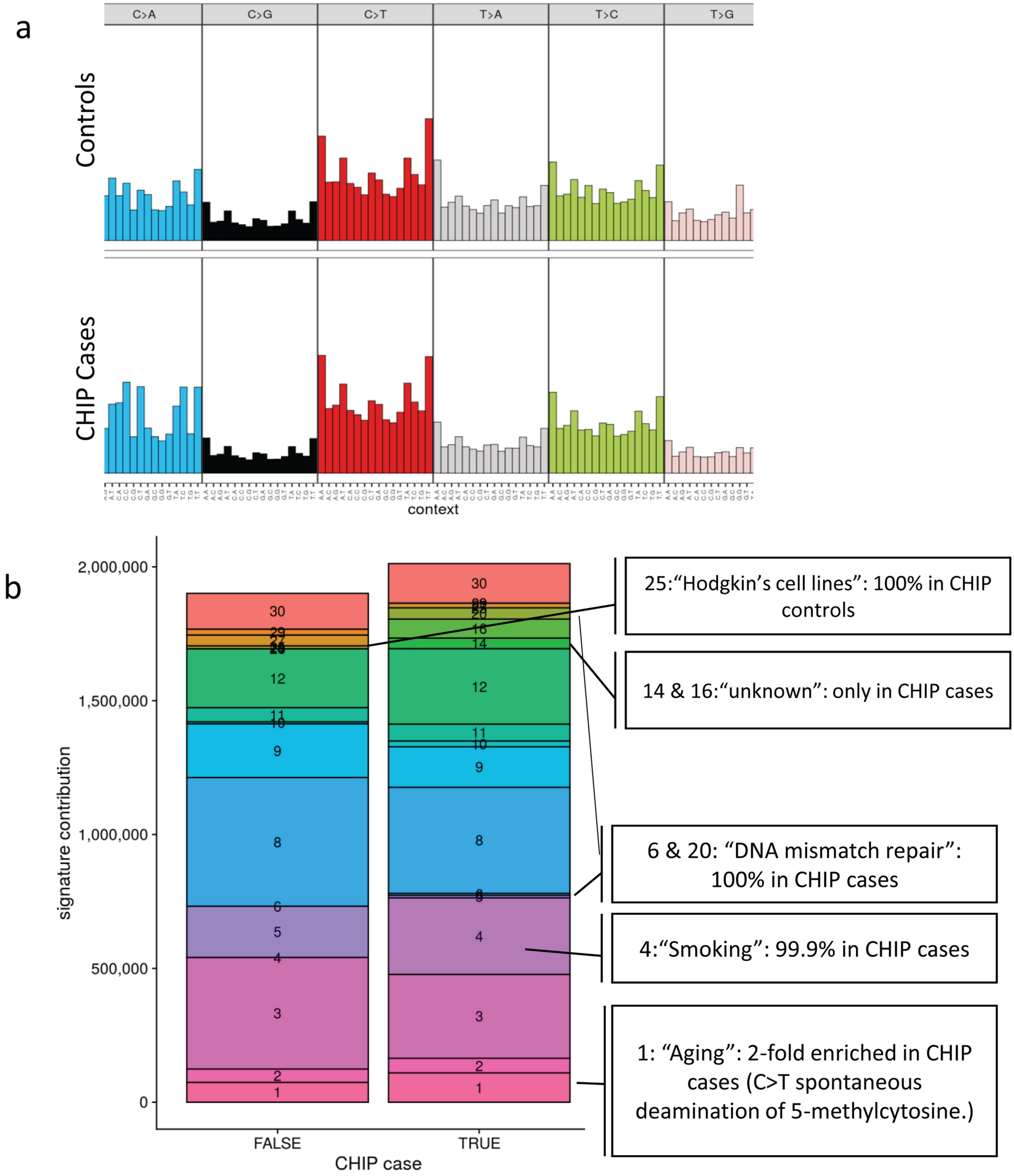
Extended Data Fig. 1 | Characterizing TOPMed CHIP. a, There was marked heterogeneity of CHIP clone size as measured by variant allele fraction by CHIP driver gene. Violin plot spanning minimum and maximum values calculated on full data set (Supplementary Table 3). Sample size for each element in violin plot displayed in Fig. 1. **b**, 90% of individuals with CHIP had only one somatic CHIP driver mutation variant 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; $n_{\mathrm{Genovese}} = 12,380$; $n_{\mathrm{Jaiswal}} = 17,182$; $n_{\mathrm{Xie}} = 2,728$).



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 Wilcoxon 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 somatic 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 RDW. *JAK2*, *SF3B1* and *SRSF2* showed driver gene specific effects on blood traits (see Supplementary Table 5). **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 6). **c**, CHIP status is associated with inflammatory markers, however notable heterogeneity existed across CHIP mutations (see Supplementary Table 7). Associations used a two-sided *t*-test from a multivariate general linear model including age, smoking, race and gender and study centre and were not adjusted for multiple comparisons. Sample sizes and exact p-values for each phenotype are listed in Supplementary Tables 5–7.



25:“Hodgkin’s cell lines”: 100% in CHIP controls

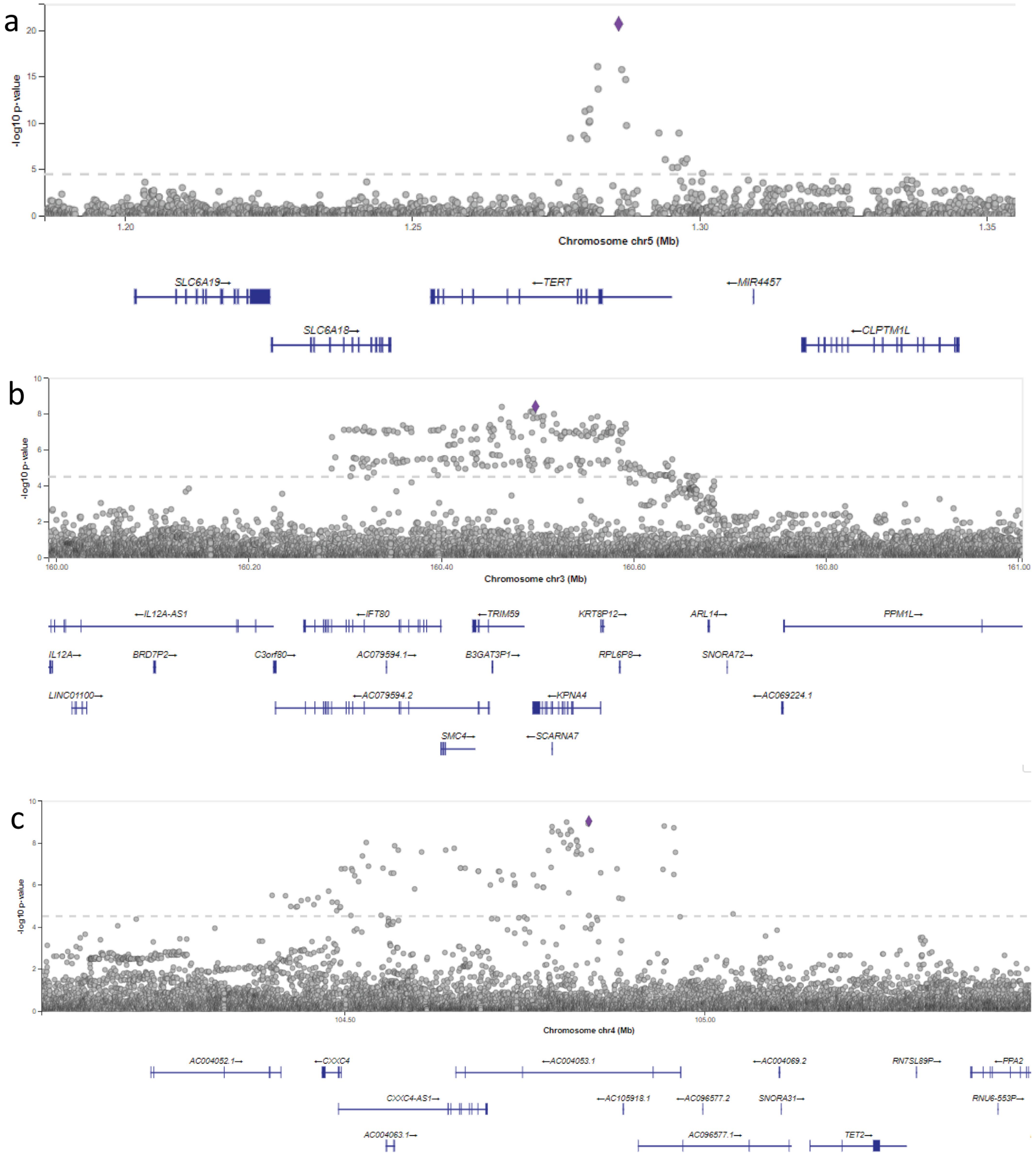
14 & 16:“unknown”: only in CHIP cases

6 & 20: “DNA mismatch repair”: 100% in CHIP cases

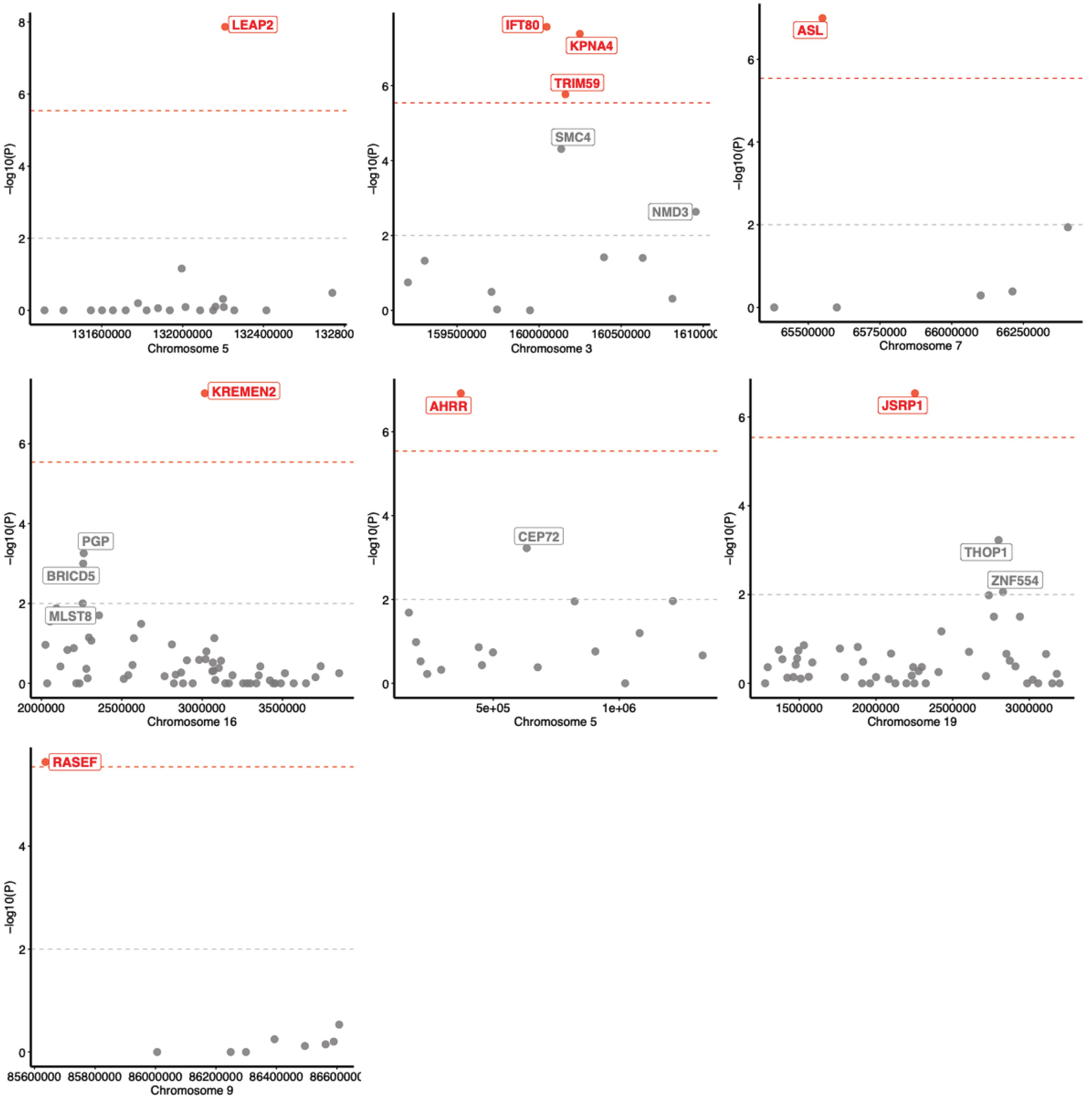
4:“Smoking”: 99.9% in CHIP cases

1: “Aging”: 2-fold enriched in CHIP cases (C>T spontaneous deamination of 5-methylcytosine.)

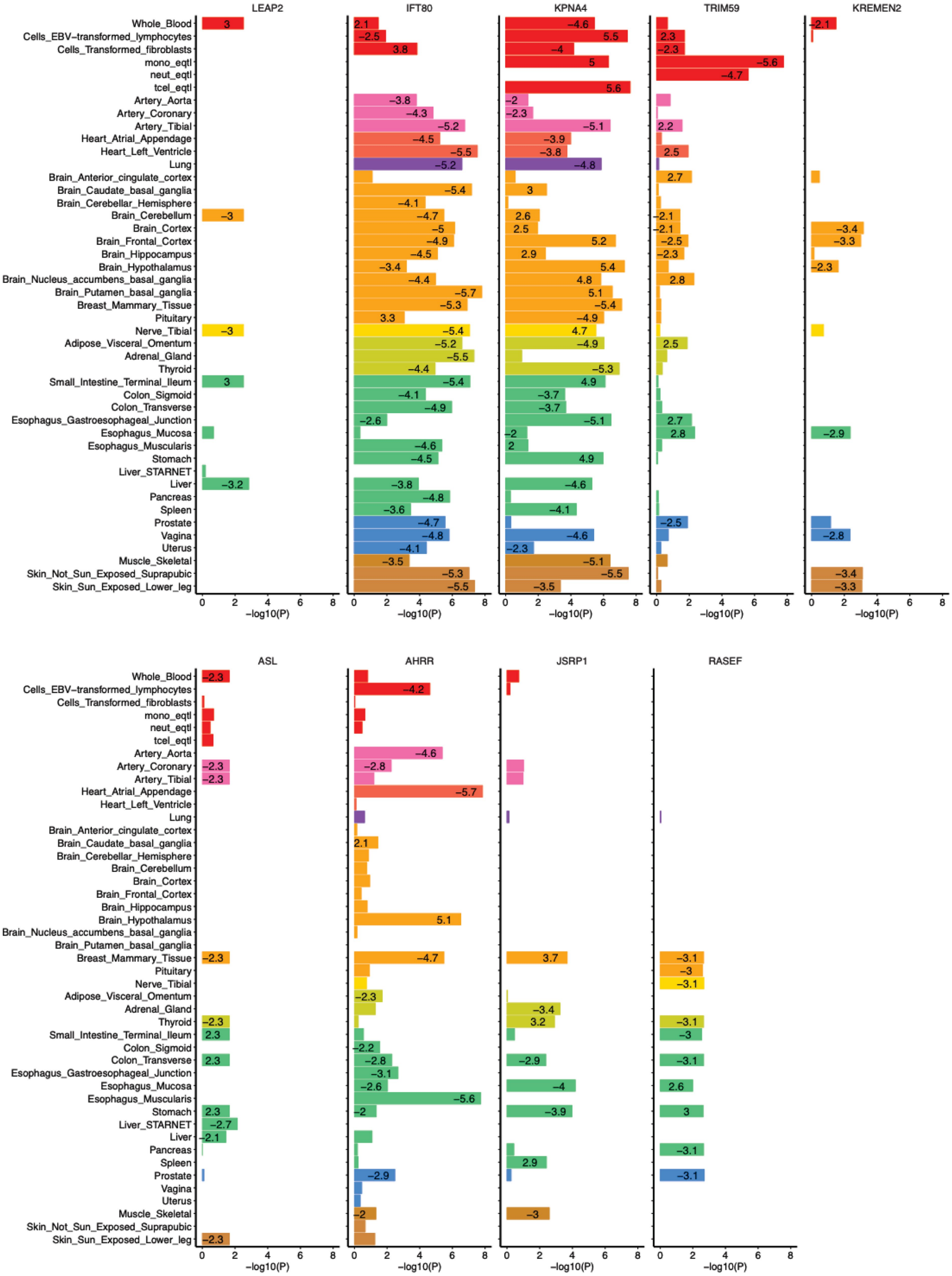
Extended Data Fig. 4 | CHIP passenger somatic mutation spectrum. a, Singleton mutation counts by nucleotide context in CHIP cases and controls. **b**, Signature contribution in CHIP cases and controls identified differential enrichment.



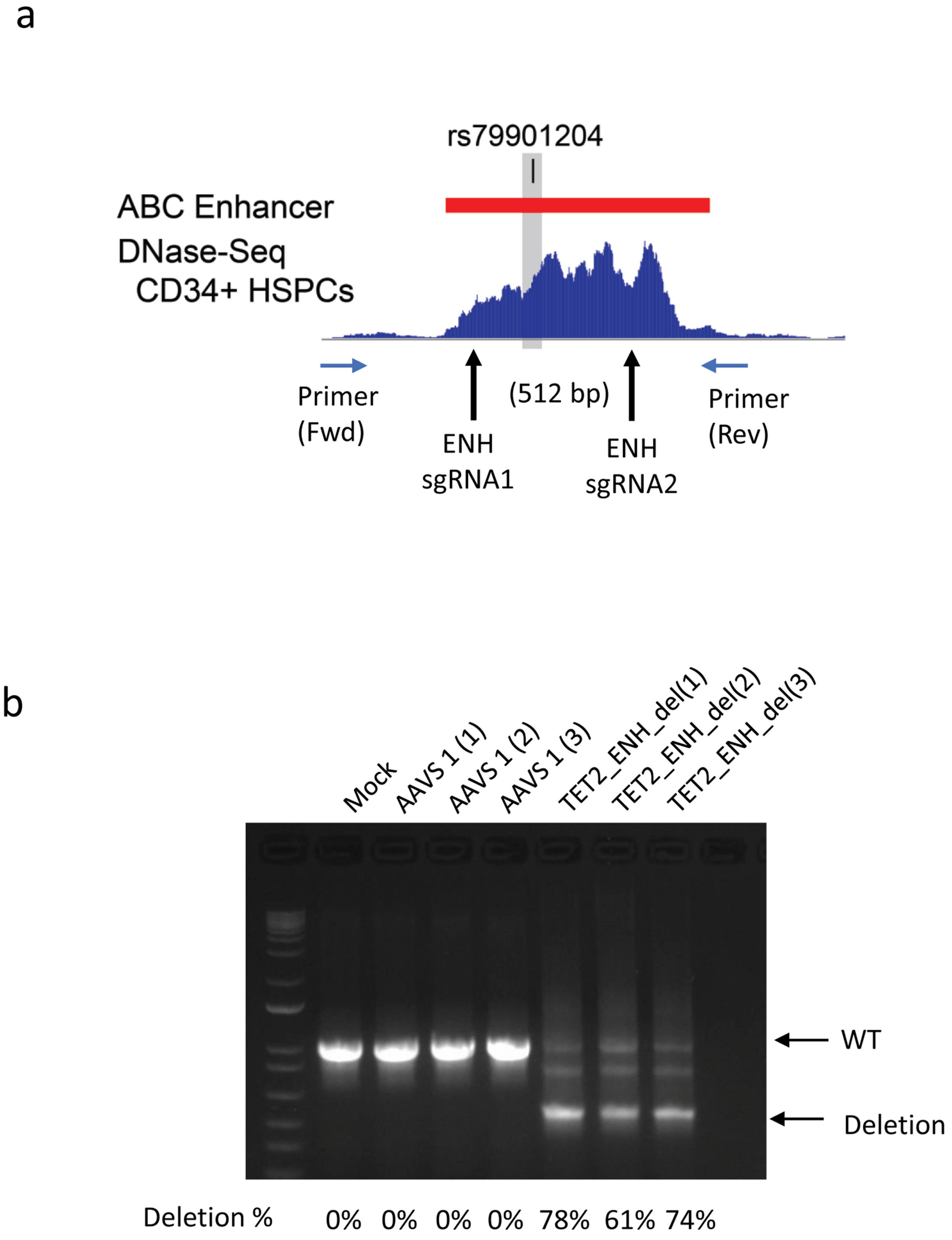
Extended Data Fig. 5 | CHIP single variant association regional association plots. a, *TERT* locus. **b**, *TRIM59*–*KPNA4* locus. **c**, *TET2* locus. Two-sided association testing performed using SAIGE (n = 65,405 individuals, see Methods).



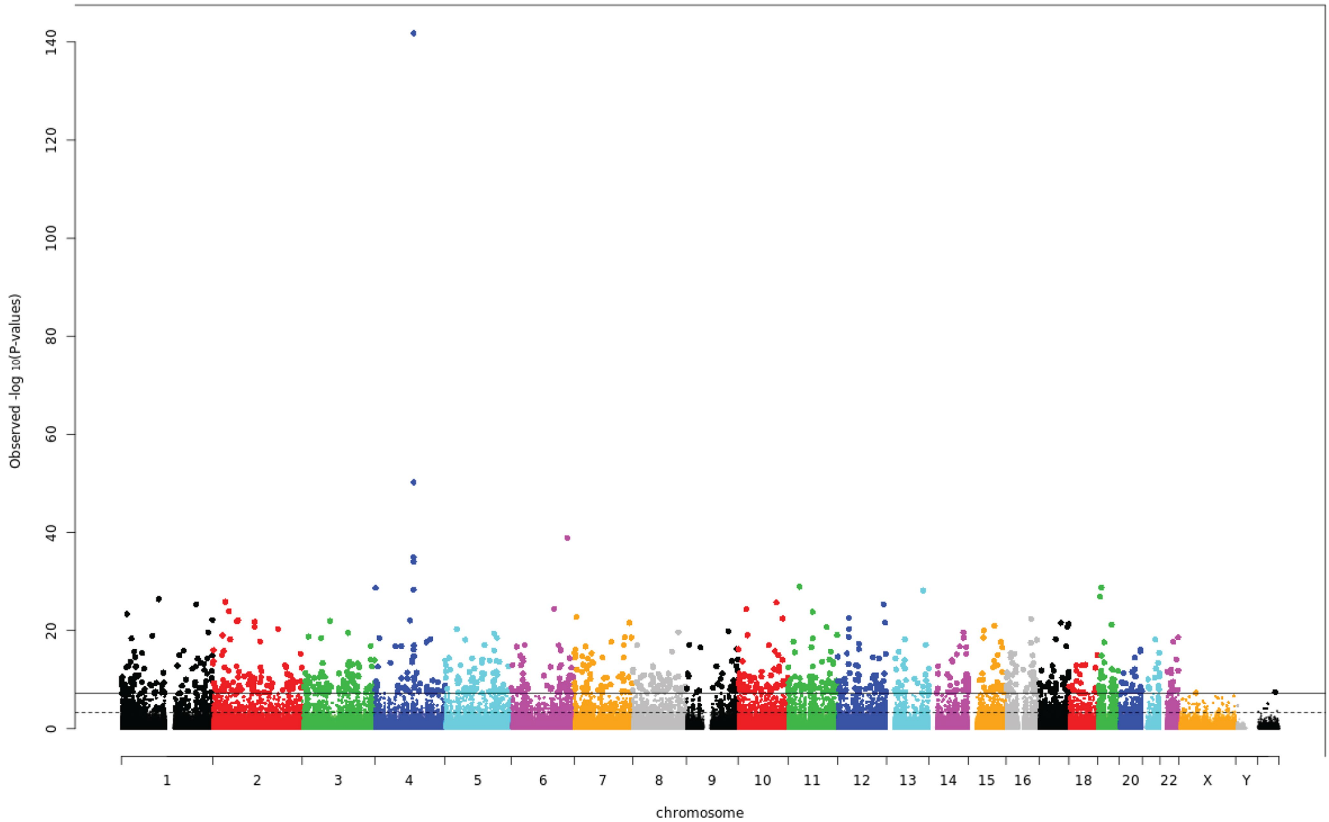
Extended Data Fig. 6 | CHIP transcriptome-wide association study (TWAS) results across 48 tissues identified 7 significant loci. UTMOST algorithm applied to CHIP genome wide association study results from $n = 65,405$ individuals (see Methods). Genomic coordinates listed on x-axis. P value from generalized Berk-Jones test on y-axis. Multiple hypothesis corrected threshold, $P < 2.9 \times 10^{-6}$ displayed as dotted red line.



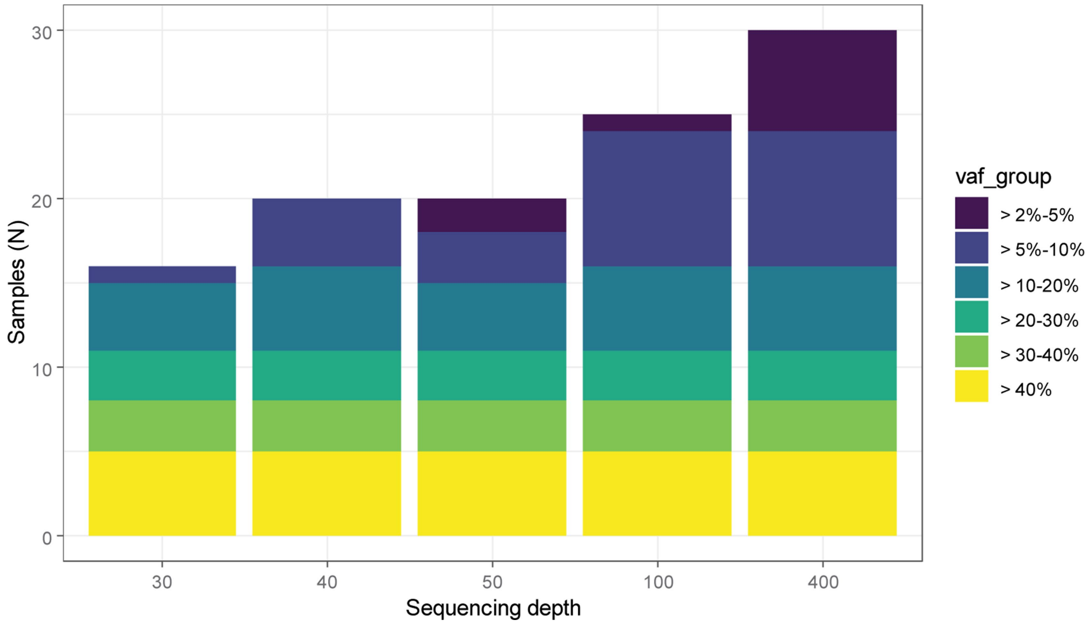
Extended Data Fig. 7 | Tissue-specific results from the top 9 overall UTMOST-significant genes. UTMOST algorithm applied to CHIP genome wide association study results from $n = 65,405$ individuals. P value from generalized Berk-Jones test. eQTL z-scores for associations with $P < 0.05$ are displayed 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 (512 bp) 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 $n = 1{,}747$ 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[33] were computationally down sampled to 30x, 40x, 50x, 100x and 400x sequencing depth. TOPMed WGS data were typically in the 40x depth range across CHIP genes. WGS data have 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%.

nature research

Corresponding author(s): Dr. Sekar Kathiresan
Dr. Pradeep Natarajan

Last updated by author(s): Sep 30, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used
Data analysis	CHIP Identification: putative somatic SNPs and short indels were called with GATK Mutect2 (https://software.broadinstitute.org/gatk). Single variant association analyses were performed with SAIGE version 0.29 (https://github.com/weizhouUMICH/SAIGE) Other statistical analysis was performed with R version 3.5 (https://www.r-project.org/).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on 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 Supplemental Table 1. Summary-level genotype data are available through the BRAVO browser (https://bravo.sph.umich.edu/). Full GWAS summary statistics are available at dbGaP accession phs001974: NHLBI TOPMed: Genomic Summary Results for the Trans-Omics for Precision Medicine Program.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences　□ Behavioural & social sciences　□ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	As this was a genomic discovery effort, we sought to maximize sample size by aggregating a set of samples that was ~10 larger than all prior CHIP analysis efforts. In a post-hoc power calculation, we estimate that we had >80% power to detect variants at a minor allele frequency of >5% that confer at least a 1.15 fold genotype relative risk of CHIP. No statistical methods were used to predetermine sample size.
Data exclusions	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 threshold was pre-established before performing the analysis. This excluded 21,712 samples leading to a final analysis set of 65,405 which was used for downstream association analyses.
Replication	We replicated the association with CHIP at the top loci (TERT) with prior analysis and replicated the TET2 locus using a second cohort of TOPMed samples distinct from our discovery analysis. We found support for all three single variant loci as well as the rare-variant CHEK2 loss of function burden signal in the cosubmitted paper on the closely related myeloproliferative neoplasm phenotype (Bao et al).
Randomization	Not applicable to genetic association studies.
Blinding	Not applicable to genetic association studies.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
□	☒	Eukaryotic cell lines
☒	□	Palaeontology and archaeology
☒	□	Animals and other organisms
□	☒	Human research participants
☒	□	Clinical data
☒	□	Dual use research of concern

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	K562 cell lines were obtained from ATCC
Authentication	Identity validated using STR analysis
Mycoplasma contamination	Mycoplasma testing was routinely performed on all cells used in the study, and confirmed to test negative.
Commonly misidentified lines (See ICLAC register)	None.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Whole genome sequencing (WGS) was performed on 97,691 samples sequenced as part of 51 distinct studies contributing to

Population characteristics	the NHLBI TOPMed research program as previously described. (https://www.biorxiv.org/content/10.1101/563866v1; www.nhlbiwgs.org) Each of the constituent studies used in this analysis provided informed consent on the participating samples. Details on participating cohorts and samples is provided in Supplemental Table S1. Each of the studies contributing to TOPMed has a distinct study design and scientific focus. Study designs included community based prospective cohorts, case-control studies for heart lung, blood and sleep disease, including studies which focused on asthma, COPD, pulmonary fibrosis, hypertension, myocardial infarction, coronary artery disease, stroke, vascular disease, venous thromboembolism, congenital heart disease, atrial fibrillation, adiposity, blood traits, lipids, sleep traits. A subset of the studies contained extended family structures while most contained unrelated individuals. The sequenced individuals were highly diverse including ~40% of European ancestry individuals, ~30% of African ancestry individuals, ~ 15% Hispanic/Latino individuals and ~10% Asian ancestry individuals. Approximately equal proportions of male and female individuals were included. Sequenced individuals spanned the spectrum of ages from birth to >100 years old.
Recruitment	Recruitment of each of the 51 studies contributing to the data analyzed here has been previously described in detail (https://www.biorxiv.org/content/10.1101/563866v1; https://www.nhlbiwgs.org/parent-study-descriptions). Each of the studies contributing to TOPMed has a distinct study design. The most common study design were community based observational epidemiology studies. Recruitment for these most commonly included individuals from a given community who were recruited to participate at random (eg Framingham Heart Study) or through community schools/clinics/hospitals (eg Gene-Environment, Admixture and Latino Asthmatics study); (2) electronic health record/biobank based studies, where individuals volunteered for research studies and samples were later selected for sequencing (eg BioME); (3) disease cohort/registry based studies where individuals with a specific condition were selected (eg Boston Early-Onset COPD).
Ethics oversight	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. Secondary analysis of the TOPMed 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.

Note that full information on the approval of the study protocol must also be provided in the manuscript.