Aberrant activation of TCL1A promotes stem cell expansion in clonal haematopoiesis

NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium; Weinstock, Joshua S.; Gopakumar, Jayakrishnan; Burugula, Bala Bharathi; Uddin, Md Mesbah; Jahn, Nikolaus; Belk, Julia A.; Bouzid, Hind; Daniel, Bence; Miao, Zhuang

Published in:
Nature

DOI:
10.1038/s41586-023-05806-1

Publication date:
2023

Article

Aberrant activation of TCL1A promotes stem cell expansion in clonal haematopoiesis

https://doi.org/10.1038/s41586-023-05806-1

Received: 3 September 2021

Accepted: 8 February 2023

Published online: 12 April 2023

Check for updates

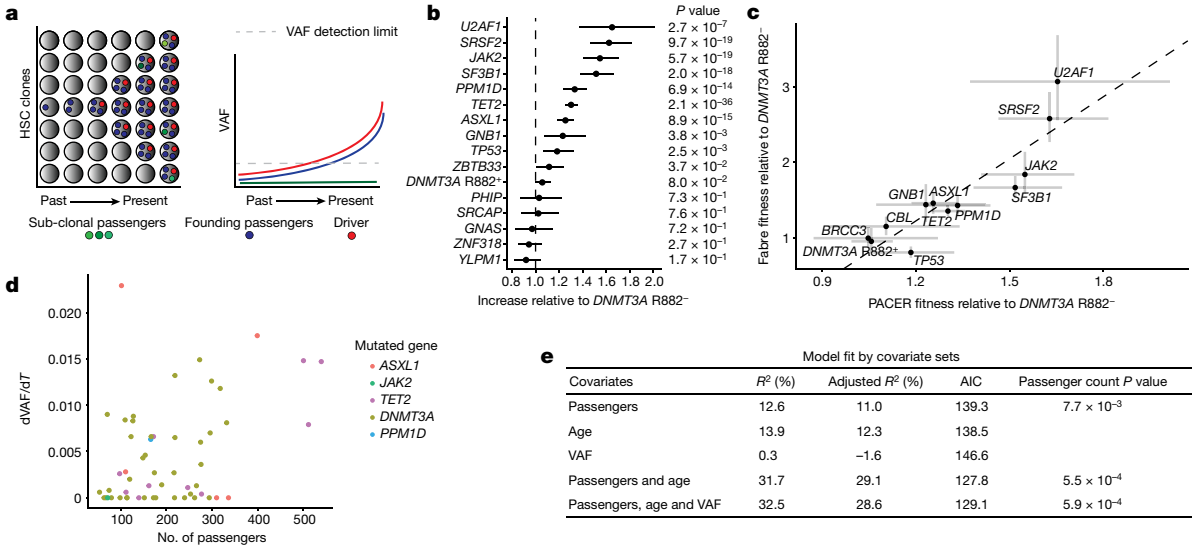
Mutations in a diverse set of driver genes increase the fitness of haematopoietic stem cells (HSCs), leading to clonal haematopoiesis[1]. These lesions are precursors for blood cancers[2–6], but the basis of their fitness advantage remains largely unknown, partly owing to a paucity of large cohorts in which the clonal expansion rate has been assessed by longitudinal sampling. Here, to circumvent this limitation, we developed a method to infer the expansion rate from data from a single time point. We applied this method to 5,071 people with clonal haematopoiesis. A genome-wide association study revealed that a common inherited polymorphism in the *TCL1A* promoter was associated with a slower expansion rate in clonal haematopoiesis overall, but the effect varied by driver gene. Those carrying this protective allele exhibited markedly reduced growth rates or prevalence of clones with driver mutations in *TET2*, *ASXL1*, *SF3B1* and *SRSF2*, but this effect was not seen in clones with driver mutations in *DNMT3A*. *TCL1A* was not expressed in normal or *DNMT3A*-mutated HSCs, but the introduction of mutations in *TET2* or *ASXL1* led to the expression of TCL1A protein and the expansion of HSCs in vitro. The protective allele restricted TCL1A expression and expansion of mutant HSCs, as did experimental knockdown of *TCL1A* expression. Forced expression of *TCL1A* promoted the expansion of human HSCs in vitro and mouse HSCs in vivo. Our results indicate that the fitness advantage of several commonly mutated driver genes in clonal haematopoiesis may be mediated by *TCL1A* activation.

Aging is characterized by the accumulation of somatic mutations, nearly all of which are 'passengers' that have little consequence for fitness. However, infrequent fitness-increasing mutations—'drivers'—may result in an expanded lineage of cells—that is, a clone. Clonal haematopoiesis of indeterminate potential (CHIP) is defined by the acquisition of specific, cancer-associated driver mutations in HSCs from people without a blood cancer[1]. Genes commonly mutated in CHIP include regulators of DNA methylation (*TET2* and *DNMT3A*), chromatin remodelling (*ASXL1*) and RNA splicing (*SF3B1*, *SRSF2* and *U2AF1*). CHIP carriers have a risk of haematologic malignancy, coronary heart disease and mortality in proportion to the variant allele fraction (VAF), a measure of clone size[2–8]. In contrast to clones with small VAF, which are ubiquitous in older individuals[9], large-VAF clones are less common. The factors driving the expansion of these mutant clones are largely unknown, partly owing to a lack of large cohorts with serially sampled blood over decades, which would otherwise enable studies on genetic and environmental correlates of clonal expansion. Here we developed an approach called passenger-approximated clonal expansion rate (PACER) to investigate the germline determinants of clonal expansion in 5,071 CHIP carriers from the NHLBI Trans-Omics for Precision Medicine (TOPMed) programme[10,11], which revealed activation of *TCL1A* as an event driving clonal expansion downstream of multiple driver genes in CHIP.

Development of PACER

HSCs accrue passenger mutations at a rate that is constant over time and that is similar across individuals[12–14]. Thus, the number of passengers in the founding cell of a CHIP clone can be used to approximate the date of acquisition of the driver mutation (Fig. 1a). Previous studies have enumerated the passenger burden in HSCs by performing whole-genome sequencing (WGS) on colonies derived from single cells[15,16]. We theorized that the passenger burden in the founding cell of a CHIP clone could be approximated from WGS of whole-blood DNA without isolation of single cells. As a mutant clone expands, the VAF of both the driver and passenger mutations increases. The number of passengers in any given cell is simply the sum of the mutations present before the acquisition of the driver event (ancestral passengers) and the mutations acquired after the driver event (sub-clonal passengers). Because the limit of detection for mutations with WGS at approximately ×38 coverage depth is equivalent to a VAF of around 8–10%, the detectable passengers in whole-blood DNA are far more likely to be ancestral passengers than sub-clonal passengers. This is because the sub-clonal passengers are private to each subsequent division of the original mutant cell, and, in the absence of a second driver event, quickly fall below the limit of detection in WGS data from bulk tissue (Supplementary Text 1). Furthermore, as the size of the clone also determines the number of detectable passengers from WGS owing to the limited sensitivity of detection at ×38 depth, clones with high fitness will harbour more detectable passengers than those with lower fitness that arose at the same time. On the basis of these observations, we used the detectable passengers as a composite measure of clone fitness (defined as relative yearly growth rate of mutant HSC clones compared with HSCs without drivers) and birth date. For two individuals of the same age and with clones of the same size, we expect the clone with more passengers to be more fit, as it must have expanded to the same size in less time.

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



e

Model fit by covariate sets

Covariates	R^2 (%)	Adjusted R^2 (%)	AIC	Passenger count *P* value
Passengers	12.6	11.0	139.3	7.7×10^{-3}
Age	13.9	12.3	138.5	
VAF	0.3	−1.6	146.6	
Passengers and age	31.7	29.1	127.8	5.5×10^{-4}
Passengers, age and VAF	32.5	28.6	129.1	5.9×10^{-4}

Fig. 1 | PACER enables estimation of clonal expansion rate from a single blood draw. a, A schematic depiction of using passenger counts to estimate the rate of expansion of a HSC clone after the acquisition of a driver mutation. The passengers (blue) that precede the driver (red) can be used to date the acquisition of the driver. **b**, The relative abundances of passenger counts were estimated for CHIP driver genes with at least 30 cases using a negative binomial regression, adjusting for age at blood draw, driver VAF and study. The total number of CHIP carriers included is 4,536. The coefficients are relative to DNMT3A R882$^-$ CHIP. Data are mean ± 95% confidence intervals; unadjusted, two-sided *P* values. **c**, The relative abundances of passengers are plotted against the empirical estimates of gene fitness derived from longitudinal deep sequencing in Fabre et al.[16]. Data are mean ± 95% confidence intervals. The estimate of the association from weighted least squares (slope = 2.7, $P = 9.6 \times 10^{-5}$, $R^2 = 80\%$) is plotted as a dashed line. **d**, Observed clonal expansion rate (dVAF/d*T*), defined as the change in VAF over time (in years), was associated with increased passenger counts in 55 CHIP carriers from the Women's Health Initiative (WHI) dataset. Colours indicate the mutated driver gene. **e**, A multivariable model including passenger counts, age at blood draw and VAF indicates the relative contributions of age and VAF over baseline models. AIC, Akaike information criteria—smaller values indicate better model fit. Unadjusted, two-sided *P* values are reported for the passengers variable in the respective models.

We identified CHIP in 5,071 out of 127,946 TOPMed participants by analysing blood DNA WGS data with Mutect2 (ref. 17) at pre-specified loci (Methods and Supplementary Table 1). CHIP was strongly associated with age at blood draw and more than 75% of these mutations were in *DNMT3A*, *TET2* or *ASXL1*, similar to our previous report from TOPMed[11]. To estimate the number of passenger mutations, we performed genome-wide somatic variant calling for the 5,071 CHIP carriers and 23,320 controls without CHIP using Mutect2. As these variant calls contain a combination of true somatic variants, germline variants and sequencing artefacts, we implemented a series of filters to enrich for the detection of true passengers (Methods). CHIP carriers had on average 271 passengers per genome after filtering (interquartile range: 142–317), representing an increase of 54% (95% confidence interval: 51%–57%) (Extended Data Fig. 1a) compared with the controls after adjusting for age and study cohort using a negative binomial regression. More than 98% of the passengers were non-coding. We presumed the detected passengers in those without CHIP were reflective of clonal haematopoiesis with unknown driver mutations[18], although some of these could have been incompletely removed artefacts. The passengers were also positively associated with age, increasing by 13.7% on average (95% confidence interval: 13.0–14.3%) each decade. Although 89% of CHIP carriers had a single driver mutation, each additional driver mutation was associated with an increment in passenger mutation counts (Extended Data Fig. 1b). This is probably owing to the presence of cooperating driver mutations within a clone, as each successive expansion caused by a new driver captures additional passengers that accumulated in the time between the last driver event and the newer one. For this reason, we limited further analyses to the 4,536 CHIP carriers with a single driver event. In summary, the detected variants in our callset had several characteristics to suggest that they were highly enriched for bona fide passengers.

We first validated the passenger count as an estimator of fitness theoretically, by constructing a simulation of HSC dynamics to characterize the relationship between fitness and detectable passenger counts (Supplementary Note 1). The simulation indicated that founding passengers were associated with driver fitness (Spearman's $\rho = 0.09$, $P < 2 \times 10^{-16}$). We estimated a passenger mutation rate per diploid genome per year of 2.3, or a per-base pair rate of 3.83×10^{-10}. This number is substantially lower than previous estimates using WGS from single haematopoietic colonies, in part because we limited the base substitutions in our analysis to C>T or T>C (Methods), but also probably owing to the lower sensitivity of detecting true passengers in whole-blood WGS compared with single-cell-derived colonies. Nonetheless, we were able to use these data to derive a hierarchical Bayesian estimator of clone fitness, which adjusts for age at blood draw and cohort effects and confirmed its correspondence to the observed passenger counts (Supplementary Note 1).

PACER estimates mutation fitness

An important test for the accuracy of our fitness estimator is a comparison of its predictions with those from empirical datasets in which clone growth is assessed longitudinally. An important prediction is fitness estimates of different driver mutations. Building on recent computational estimates of variant fitness[19], we estimated the distribution of passenger counts for the most common CHIP driver genes as a measure of fitness. We used non-R882 *DNMT3A* mutations (*DNMT3A* R882$^-$) as a reference point and estimated the relative abundances of passengers in other genes using negative binomial regression adjusting for age, VAF, sex and study cohort. We termed the approach of using age- and VAF-adjusted passenger mutations to estimate fitness in regression models PACER. According to PACER, mutations in splicing factors (*SF3B1*, *SRSF2* and *U2AF1*) and *JAK2*^{V617F} were the fastest growing, whereas *DNMT3A* R882$^-$ were among the slowest (Fig. 1b and Supplementary Table 2). Mutations in *TET2*, *ASXL1*, *PPM1D*, *TP53*, *ZBTB33* and *GNB1* were in the next tier and had approximately the same level

of fitness, as estimated from PACER. Relative to the R882^{-} carriers, we observed a modest increase in fitness in *DNMT3A*^{R882} mutant clones. These observations are concordant with previous empirical estimates of variant fitness derived from longitudinal sequencing of samples with clonal haematopoiesis[6,16,20–22]. When driver gene fitness estimates from PACER were directly compared to estimates from a large longitudinal dataset of clonal haematopoiesis[16], the coefficient of determination (R^2) was 80% (Fig. 1c and Methods).

To further validate the utility of the passenger count, we tested whether PACER could also predict future clone growth within individuals. We performed targeted sequencing in 55 CHIP carriers from the Women's Health Initiative (WHI) with a single driver mutation. Each individual had two blood samples taken at an interval ranging from 13 to 19 years apart, which allowed us to determine the rate of change in VAF of the driver variant ($\frac{dVAF}{dT}$) (Fig. 1d). WGS was used to determine passenger count at the first time point. We constructed a simple estimator of $\frac{dVAF}{dT}$ using only the passengers, VAF from the first blood draw and age from the first blood draw (Methods). Our theoretical framework considered passengers to be an estimate of clone fitness after accounting for age and VAF; thus age and VAF variables were also considered in the model. A model that included age and VAF in addition to passenger count was superior for predicting $\frac{dVAF}{dT}$ (R^2 = 32.5%, adjusted R^2 = 28.6%) than models only including passengers (R^2 = 12.6%, adjusted R^2 = 11%), age (R^2 = 13.9%, adjusted R^2 = 12.3%) or VAF (R^2 = 0.3%, adjusted R^2 = −1.6%). In all models, the passenger count variable was significantly associated with $\frac{dVAF}{dT}$ (Fig. 1e and Extended Data Fig. 1c).

To contextualize its performance, we compared PACER with fitness estimators derived from longitudinal datasets (102 individuals with clonal haematopoiesis from Fabre et al.[16] as well as 24 individuals from WHI) (Supplementary Note 2 and Supplementary Tables 3 and 4). Each individual had between three and five assessments of VAF over several years, and fitness estimates derived from the first two to four measurements were used to predict $\frac{dVAF}{dT}$ between the penultimate and final time points. We observed that the point estimates of R^2 for the correlation of $\frac{dVAF}{dT}$ with fitness in these datasets ranged from 4.5% to 20%. These results indicate that PACER, which is derived from a single blood draw, predicted future clone growth comparably to, if not better than, fitness estimators derived from longitudinal data with two to four serial measurements.

To consider alternative statistical approaches, we compared the fitness estimates derived with PACER with our hierarchical Bayesian estimator of clone fitness (PACER-HB; Methods), and observed strong correspondence between the two fitness estimates (Supplementary Note 1), suggesting that the relative simplicity of PACER does not clearly reduce its performance compared with more sophisticated approaches.

GWAS of PACER

We performed a genome-wide association study (GWAS) of PACER in CHIP carriers to identify inherited genetic variation that associates with clonal expansion rate (Methods). In this analysis, we refer to the PACER score as the residuals from the linear regression of passenger counts with age at blood draw, study, VAF and the first ten genetic ancestry principal components included as covariates.

The GWAS identified a single locus at genome-wide significance overlapping *TCL1A* (Fig. 2a), and genetic fine-mapping further narrowed down the associated region to a credible set containing a single variant, rs2887399 (Extended Data Fig. 1d and Methods). The reference allele at this variant is a guanine (G) and the alternate allele is a thymine (T). We did not find any association between PACER and rare variants near rs2887399, suggesting that rs2887399 is not tagging other genetic variants and is the causal variant at this locus (Extended Data Fig. 1e,f). The T allele of rs2887399 is common, occurring in 26% of haplotypes sequenced in TOPMed, and each additional T allele was associated with a decrease of 0.15 in the PACER *z*-score ($P = 4.5 \times 10^{-12}$). rs2887399 is located in the core promoter of *TCL1A* as defined by the Ensembl[23] regulatory build 108,162 base-pairs from the canonical transcription start site (TSS) and was nominated as the causal gene by the Open Targets[24] variant-to-gene prediction algorithm. *TCL1A* has been implicated in lymphoid malignancies[25], but to our knowledge, it has not been studied in the context of HSC biology. Of note, the region in the *TCL1A* promoter where rs2887399 resides is poorly conserved with non-primate species (Extended Data Fig. 1g).

We next performed a genome-wide search of rare variation associated with the passengers and identified 15 windows associated with passenger counts at Bonferroni significance ($P = 2.9 \times 10^{-5}$, Supplementary Tables 5 and 6), including a distal enhancer for *TNFAIP3* ($P = 5.4 \times 10^{-7}$) (GeneHancer[26]).

Stratified associations with rs2887399

We tested whether the association between rs2887399 and PACER varied by CHIP driver gene. Using *DNMT3A* as the reference, we observed that rs2887399 was more protective against clonal expansion in *TET2* than *DNMT3A* CHIP (beta = −0.24 per T allele, $P = 9.6 \times 10^{-4}$, Supplementary Table 7). Stratification of PACER score by rs2887399 genotype revealed that the T allele slowed growth of *TET2* clones but had little effect on *DNMT3A* clones (Fig. 2b).

Clones with a decreased expansion rate may never grow large enough to be detected, so we also performed association tests between rs2887399 and the presence of a CHIP-associated driver mutation stratified by gene. In our previous analysis[11], we reported that the T allele was associated with increased risk for *DNMT3A* mutations. Previous reports have also identified that the T allele of rs2887399 decreases risk for mosaic loss of the Y chromosome[27] (LOY). We observed that rs2887399 was associated with significantly reduced odds of mutations in *TET2*, *ASXL1*, *SF3B1* and *SRSF2* (Fig. 2c and Supplementary Tables 8 and 9). The effect size of rs2887399 was large, as carrying the T/T genotype conferred odds ratios for having a driver mutation in these genes from 0.22 to 0.63. The risk reduction was particularly marked for mutations in *SF3B1* and *SRSF2*, as well as for having more than one non-*DNMT3A* driver mutation. In sum, these results indicate that the T allele at rs2887399 is protective against CHIP owing to driver mutations in several genes that have higher risk of progression to frank haematologic malignancy[6,28].

Our analysis predicts that the T allele of rs2887399 should reduce the expansion rate of several non-*DNMT3A* mutant clones. We performed targeted sequencing in 900 additional participants in the WHI dataset at 2 time points taken a mean of 16.2 years apart and identified those with mutations in *DNMT3A, TET2, ASXL1* or *SF3B1* (n = 351, including 53 previously identified from the PACER validation). Using this dataset, we tested whether the T allele was associated with the expansion rate of clonal haematopoiesis clones. We defined clonal expansion as the per cent growth per year of the clonal haematopoiesis clones as estimated by a Bayesian logistic growth model (Methods). We observed that each T allele of rs2887399 was associated with reduced expansion in *TET2* and *ASXL1* mutant clones by 4% but not in *DNMT3A*-mutant clones, concordant with the PACER prediction (Fig. 2d and Supplementary Table 10). *TET2* and *ASXL1* clones with the T/T rs2887399 genotype had very slow rates of clonal expansion (0.5% mean growth per year) compared to clones with the G/G genotype (8.3% mean growth per year). These results provide further validation that PACER can accurately identify correlates of clonal expansion.

We sought to understand why the T allele of rs2887399 was associated with an increased prevalence of *DNMT3A* CHIP but had little effect on *DNMT3A* clonal expansion rate. Recent work has demonstrated that haematopoiesis becomes increasingly oligoclonal during aging as competition between clones with varying degrees of fitness intensifies[13]. We hypothesized that carrying the T allele of rs2887399 would lead to an increased likelihood of *DNMT3A*-mutant clones growing to

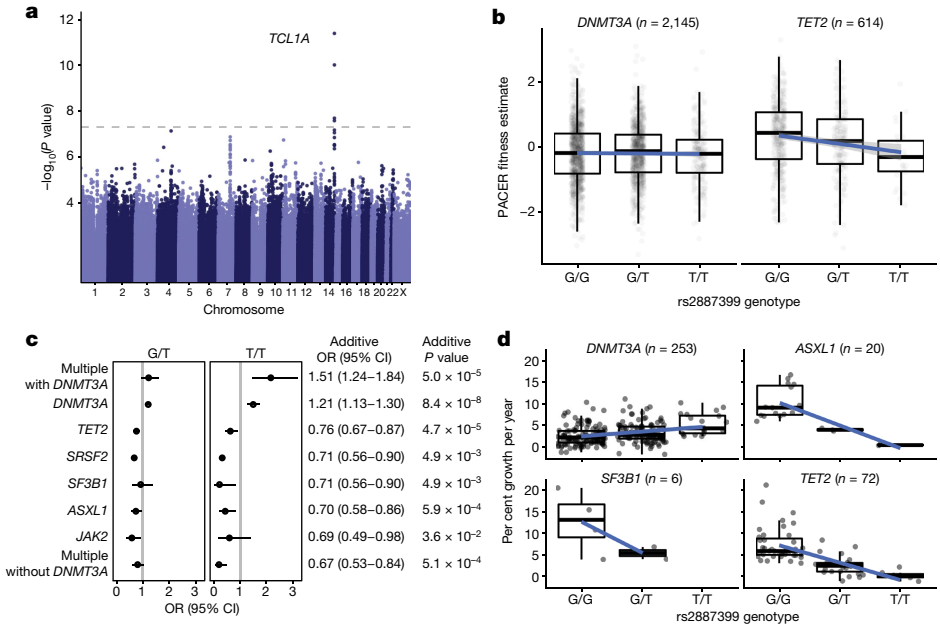


Fig. 2 | GWAS of PACER identifies germline determinants of clonal expansion in blood. a, A GWAS of passenger counts identifies *TCL1A* as a genome-wide significant locus. Test statistics were estimated with SAIGE[43]. **b**, The association between the genotypes of rs2887399 and PACER fitness estimates varied between *TET2* and *DNMT3A*. T alleles were associated with decreased PACER score in *TET2* mutation carriers, but no association was observed in *DNMT3A* carriers. **c**, The association between T alleles at rs2887399 and presence of specific CHIP mutations varied by CHIP mutation (n = 5,071 CHIP carriers). Left, Forest plots show the odds ratios (OR) for having specific mutations in individuals who were G/T or T/T, relative to those who were G/G. OR were estimated using Firth logistic regression; error bars represent 95% confidence intervals (CI). Right, effect estimates and P values from SAIGE, which uses an additive coding of the T alleles for hypothesis testing and uses a generalized linear mixed model to estimate test statistics. Unadjusted, two-sided P values. In the additive tests, *SF3B1* and *SRSF2* were grouped together to aid convergence. **d**, The association between the genotypes of rs2887399 and per cent growth per year of CHIP clones from 351 carriers in the WHI dataset. Per cent growth per year was estimated using a Bayesian logistic growth model of clonal expansion. In box plots in **b**,**d**, the centre line indicates the median, box edges indicate the interquartile range, and error bars indicate maximum and minimum values.

detectable levels owing solely to reduced fitness of other competing clones. To test this hypothesis, we performed a simulation of clonal expansion with two competing clones carrying *DNMT3A* and *TET2* mutations, respectively. The fitness of the *DNMT3A* clone remained constant but the fitness of the *TET2* clone was 20% higher relative to *DNMT3A* in one setting and 20% lower in the other setting, similar to the estimates from PACER for relative fitness of *TET2* clones from those with G/G versus T/T genotype at rs2887399. Reducing the fitness of *TET2* was sufficient to increase the likelihood of the *DNMT3A* clone expanding to detectable levels (Extended Data Fig. 2a).

TCL1A expression in haematopoietic cells

We sought to establish how rs2887399 alters clonal expansion. We first tested whether rs2887399 was associated with *TCL1A* expression in any cell type. As identified in GTEx v8 (ref. 29), the T allele reduces expression of *TCL1A* in whole blood (normalized effect size = −0.13, $P = 1.4 \times 10^{-5}$). The PACER GWAS colocalized[30] with *cis*-expression quantitative trait loci (eQTLs) for *TCL1A* in whole blood (posterior probability of a single shared causal variant = 97.1%; Extended Data Fig. 2b). This association is likely to be driven by B cells, as *TCL1A* is highly expressed in B cells but appears to have absent or exhibit low expression in all other cell types in blood except in rare plasmacytoid dendritic cells (Extended Data Fig. 2c, Supplementary Table 11 and the Human Cell Atlas[31]).

Little is known about *TCL1A* expression in HSCs. We examined whether CHIP-associated mutations altered the regulation of the *TCL1A* locus in human haematopoietic stem and progenitor cells (HSPCs) using publicly available single-cell RNA-sequencing (scRNA-seq) and transposase-accessible chromatin high-throughput sequencing (ATAC-seq) datasets of normal and malignant haematopoiesis. *TCL1A* was expressed in less than 1 in 1,000 cells identified as HSCs or multipotent progenitors (MPPs) (HSC/MPPs) in scRNA-seq data from six normal human marrow samples[32,33] (range 0–0.17%). By contrast, *TCL1A* was expressed in a much larger fraction of HSC/MPPs in 3 out of 5 patients with *TET2*- or *ASXL1*-mutated myeloid malignancies (range 2.7–7%) (Extended Data Fig. 3a and Supplementary Table 12). Next, using a dataset of ATAC-seq in normal and pre-leukaemic HSCs[34] (pHSCs), which are residual non-leukaemic HSCs present in patients with acute myeloid leukaemia (AML) that often harbour only the initiating driver mutations, we evaluated chromatin accessibility at the *TCL1A* promoter. Consistent with the lack of *TCL1A* transcripts in normal HSCs, we observed that the promoter was not accessible in HSCs from healthy donors, in HSCs from patients with AML that carried no driver mutations, or in pHSCs with *DNMT3A* mutations. By contrast, the patients with *TET2*-mutated pHSCs had clearly accessible chromatin at the *TCL1A* promoter (Extended Data Fig. 3b), and this locus had the greatest $\log_2$ fold-change of any differentially accessible TSS peak in *TET2*-mutant versus control samples (Supplementary Table 13).

We next tested whether the neighbouring genes *TCL6* or *TCL1B* became expressed or had accessible chromatin in HSCs carrying CHIP mutations in these same datasets. In contrast to the result for *TCL1A*, no RNA expression or accessible promoter chromatin could be found at these genes in HSCs (Extended Data Fig. 3c and Supplementary Table 12), further supporting *TCL1A* as the causal gene for clonal expansion.

Functional effect of rs2887399 on HSCs

On the basis of these observations, we proposed the following mechanistic model: normally, the *TCL1A* promoter is inaccessible and gene

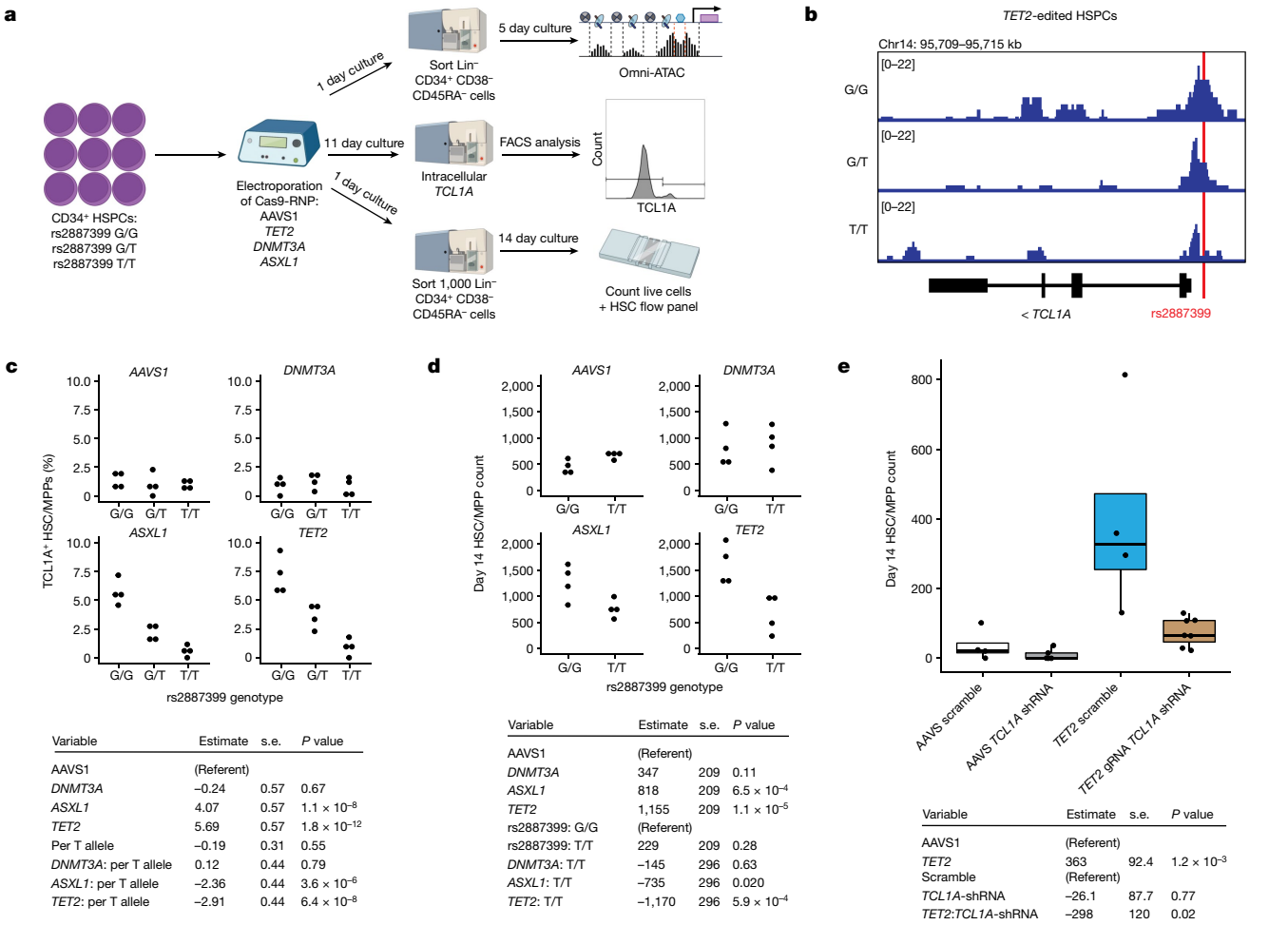


Fig. 3 | Effect of rs2887399 on TCL1A expression and clonal expansion. **a**, Schematic of experimental workflow. FACS, fluorescence-activated cell sorting; RNP, ribonuclear protein complex. **b**, ATAC-seq tracks illustrating chromatin accessibility at rs2887399 in *TET2*-edited HSPCs from donors with G/G, G/T and T/T genotypes after 5 days in liquid culture. The red line indicates the location of rs2887399. See also Extended Data Fig. 8 and Supplementary Table 14. **c**, Top, the percentage of Lin$^-$CD34$^+$CD38$^-$CD45RA$^-$ cells expressing TCL1A, as determined by flow cytometry of edited HSPCs after 11 days in liquid culture, stratified by edited gene and rs2887399 genotype. Bottom, results of a linear regression model for the effect of the edited gene (relative to AAVS1), the number of T alleles at rs2887399 and the interaction term of edited genes with T alleles. Unadjusted *P* values from a two-sided test. $n = 4$ biologically independent replicates for each group. **d**, Top, Lin$^-$CD34$^+$CD38$^-$CD45RA$^-$ cell counts of edited HSCs after 14 days in liquid culture. Bottom, results of a linear regression model for the effect of the edited gene (relative to AAVS1), rs2887399 genotype (relative to G/G) and the interaction term of the edited gene with rs2887399 genotype. Unadjusted *P* values from a two-sided test. $n = 4$ biologically independent replicates for each group. **e**, Top, Lin$^-$CD34$^+$CD38$^-$CD45RA$^-$ cell counts after 14 days liquid culture of edited and shRNA-transduced HSCs. Bottom, results of a linear regression model for the effect of the edited gene (relative to AAVS1), shRNA (relative to scramble control) and the interaction term of the edited gene with shRNA. Unadjusted *P* values from a two-sided test. The centre line indicates the median, box edges indicate the interquartile range, and error bars indicate maximum and minimum values. $n = 4$ for AAVS1 gRNA and scramble, $n = 5$ for AAVS1 gRNA and *TCL1A* shRNA, $n = 4$ for *TET2* gRNA and scramble, and $n = 7$ for *TET2* gRNA and *TCL1A* shRNA, where *n* represents the number of biologically independent replicates. gRNA, CRISPR guide RNA.

expression is repressed in HSCs. In the presence of driver mutations in *TET2*, *ASXL1*, *SF3B1*, *SRSF2* or with LOY, *TCL1A* is aberrantly expressed and drives clonal expansion of the mutated HSCs. The presence of the T allele of rs2887399 restricts chromatin accessibility at the *TCL1A* promoter, leading to reduced expression of *TCL1A* RNA and protein and abrogation of the clonal advantage due to the mutations (Extended Data Fig. 4).

To test our model experimentally, we obtained human CD34$^+$ mobilized peripheral blood cells from donors who were G/G, G/T, or T/T at rs2887399. The three donors were healthy and between 29 and 32 years old at the time of donation. We used CRISPR to introduce insertion–deletion mutations with high efficiency in *DNMT3A*, *TET2* or *ASXL1* to mimic CHIP variants, or at the adeno-associated virus integration site 1 (AAVS1) as a control (Fig. 3a and Extended Data Fig. 5).

First, we examined whether chromatin accessibility at the *TCL1A* promoter was altered by rs2887399 genotype. We edited CD34$^+$ cells from each genotype for *TET2*, sorted cells with a marker profile of HSCs and MPPs (Lin$^-$CD34$^+$CD38$^-$CD45RA$^-$), cultured them in cytokine-supported medium, and then performed ATAC-seq. Consistent with the pHSC data, we detected increased chromatin accessibility at the *TCL1A* promoter in *TET2*-edited, but not *DNMT3A*-edited, cells from the rs2887399 G/G donor relative to AAVS1-edited cells (Fig. 3b, Extended Data Fig. 6 and Supplementary Table 14). However, chromatin accessibility was decreased in samples from carriers of the T allele in a dose-dependent manner, indicating that the protective effect of the T allele of rs2887399 is mediated by blocking *TCL1A* promoter accessibility.

Next, we tested whether the T allele of rs2887399 altered TCL1A protein expression in HSC/MPPs. We edited CD34$^+$ cells with the three rs2887399 genotypes at AAVS1, *DNMT3A*, *TET2* and *ASXL1* and performed a flow cytometry-based assay for TCL1A protein expression after culturing the cells for 11 days. Around 1% of HSCs/MPPs from AAVS1- or

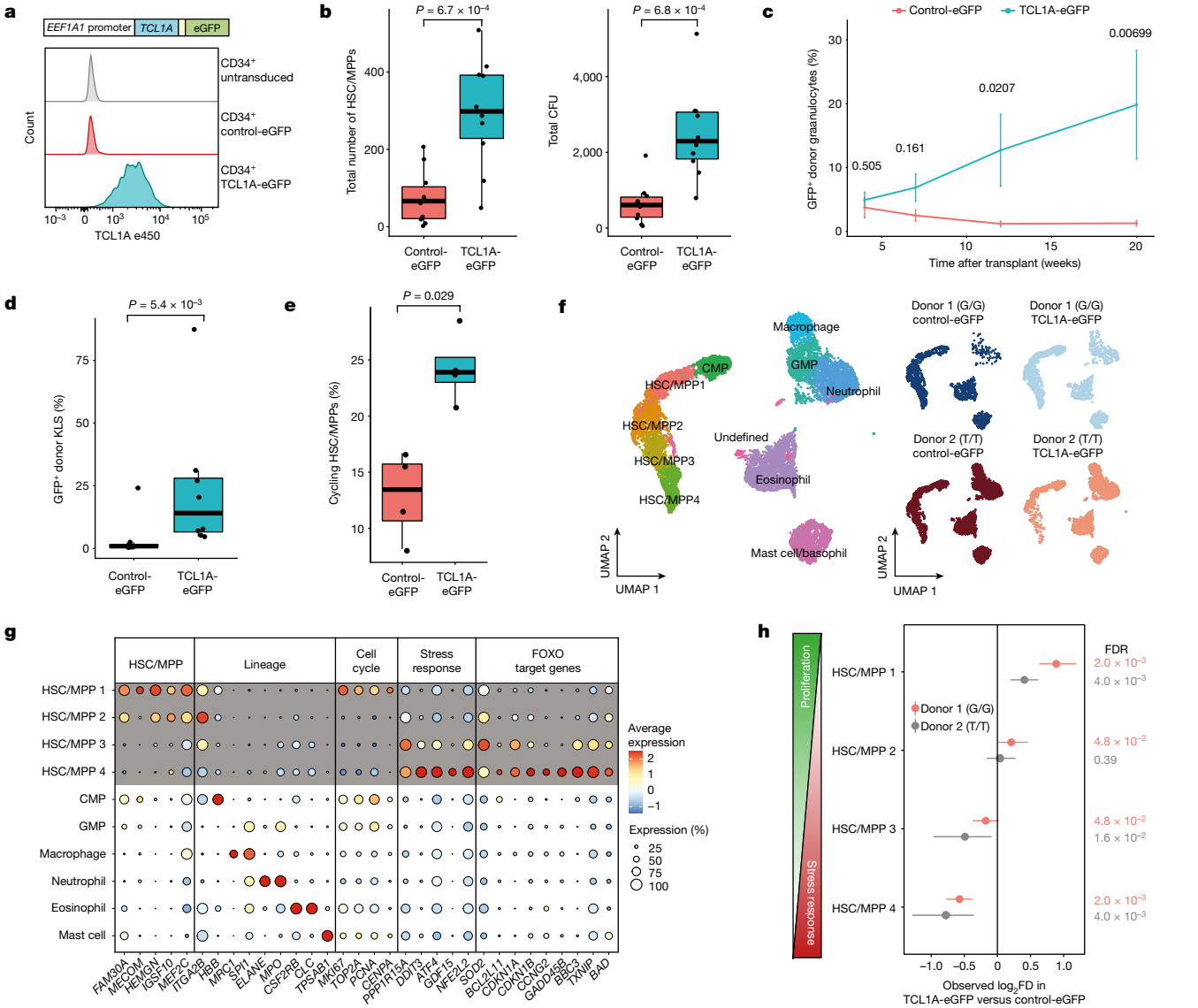


Fig. 4 | *TCL1A* expression is sufficient for HSC expansion. a, Schematic of TCL1A-eGFP lentivirus construct (top) and the effect of viral transduction on *TCL1A* expression in human CD34$^+$ HSPCs (bottom). **b**, Number of Lin$^-$CD34$^+$CD38$^-$CD45RA$^-$ cells (left) and quantification of colony-forming units in methylcellulose (right) after 14 days of liquid culture of transduced human HSCs. Two-sided *t*-test. $n = 10$ biologically independent replicates for each group. **c**, Donor granulocyte chimerism of mice transplanted with TCL1A-eGFP- or control-eGFP-transduced c-Kit$^+$ marrow cells plus GFP$^-$ competitor marrow. Data are mean ± s.e.m. for each time point. Two-sided Wilcoxon rank sum tests. $n = 8$ mice for each group. **d**, The percentage of GFP$^+$ donor cells in Lin$^-$c-Kit$^+$Sca-1$^+$ (KLS) marrow 22 weeks after transplant. Two-sided Wilcoxon rank sum test. $n = 8$ mice for each group. **e**, The percentage of cycling Lin$^-$CD34$^+$CD38$^-$ cells, determined by DAPI staining after 10 days of liquid culture of transduced human HSC/MPPs. Two-sided Wilcoxon rank sum test. $n = 4$ biologically independent replicates for each group. **f**, Uniform manifold approximation and projection (UMAP) of clusters identified after 7 days liquid culture of transduced human HSC/MPPs. Left, all samples combined. Right, the four individual samples. G/G and T/T refer to the donor rs2887399 genotype. **g**, Dot plot illustrating the average expression level (colour) and percentage of cells (circle size) expressing representative marker genes across different cell clusters arranged by functional group. **h**, Forest plot of $\log_2$ fold difference (FD) in the proportion of cells within each HSC/MPP cluster in cells transduced with TCL1A-eGFP versus control-eGFP (permutation test). Each donor represents an independent experiment and the false discovery rate (FDR) for each comparison is shown on the right. In box plots **b**,**d**,**e**, the centre line indicates the median, box edges indicate the interquartile range, and error bars indicate maximum and minimum values.

DNMT3A-edited samples were positive for TCL1A, which did not vary by rs2887399 genotype. By contrast, 4.6–9.3% of HSC/MPPs from the G/G donor that had been edited for *ASXL1* or *TET2* expressed TCL1A, and the proportion of TCL1A-positive HSC/MPPs decreased in donor samples with each additional T allele (Fig. 3c,d and Extended Data Fig. 7a). There was minimal expression of TCL1A in any non-HSC/MPP CD34$^+$ population in any of the samples. Notably, less than 10% of HSC/MPPs expressed TCL1A in any sample, even though the proportion of mutant cells was greater than 90% (Extended Data Fig. 5), suggesting that only a fraction of HSC/MPPs express TCL1A at any given time, even in the presence of *TET2* or *ASXL1* mutations. This is consistent with scRNA-seq data from haematological malignancy samples (Extended Data Fig. 3a).

To test whether the rs2887399 genotype had an effect on expansion of HSPCs in vitro, we edited the CD34$^+$ cells from G/G and T/T donors, sorted HSCs (Lin$^-$CD34$^+$CD38$^-$CD45RA$^-$CD90$^+$), and analysed HSPC counts after 14 days. There was a notable expansion of cells bearing markers of HSC/MPPs in the *ASXL1*- and *TET2*-edited samples from the rs2887399 G/G donor compared to the AAVS1-edited sample, but this

effect was abrogated in edited samples from the rs2887399 T/T donor (Fig. 3e). A population of cells that was $Lin^{-/lo}CD34^{+}CD38^{-}CD45RA^{lo}$ ($CD45RA^{lo}$ HSPCs), presumably progenitors descended from the HSC/MPP population, was also markedly expanded in the *ASXL1*- and *TET2*-edited samples from the G/G donor, but the degree of expansion was partially reversed in the edited samples from the T/T donor (Extended Data Fig. 7b). The ratio of $CD34^{+}CD45RA^{-/lo}$ progenitors to $CD34^{-}$ cells was also increased in the *ASXL1*- and *TET2*-edited samples from the G/G donor compared with the T/T donor, indicating either less retention of stem or progenitor cell activity or faster differentiation in the absence of *TCL1A* expression (Extended Data Fig. 7c). There was no effect on HSPC expansion in the AAVS1- or *DNMT3A*-edited samples based on rs2887399 genotype. Furthermore, we were unable to detect any significant differences in expansion of *DNMT3A*-edited HSCs based on rs2887399 genotype even when older donors were used (Supplementary Table 15). Thus, carrying the T allele of rs2887399 abrogates the clonal expansion of HSPCs with *ASXL1* and *TET2* mutations in an experimental system, but has minimal direct effect on fitness of mutant *DNMT3A* clones, consistent with the PACER analysis.

To orthogonally validate the necessity of *TCL1A* for clonal expansion, we edited $CD34^{+}$ cells from a rs2887399 G/G donor with AAVS1 or *TET2* guides, followed by lentiviral delivery of short hairpin RNA (shRNA) targeting *TCL1A* or scramble control. The *TCL1A* shRNA construct we used was validated to reduce TCL1A protein expression by around 90% (Extended Data Fig. 8a). We then sorted GFP^{+} HSC/MPPs and performed the same in vitro expansion assay. The increase in *TET2*-mutated HSC/MPP counts seen after 14 days was nearly completely attenuated by *TCL1A* knockdown (Fig. 3f), indicating that *TCL1A* expression is necessary for expansion of *TET2*-mutant HSCs in this assay.

TCL1A expression promotes HSC expansion

If aberrant *TCL1A* expression is the major reason for positive selection of *TET2*-, *ASXL1*-, *SF3B1*- and *SRSF2*-mutant HSCs, then forced expression of *TCL1A* in unmutated HSCs should be sufficient to recapitulate clonal expansion phenotypes. To test this hypothesis, we transduced human $CD34^{+}$ cells with lentivirus containing the *TCL1A* open reading frame (TCL1A-eGFP) or empty vector control (control-eGFP) (Fig. 4a) and performed in vitro clonal expansion assays on purified HSC/MPPs. The per-cell level of TCL1A protein expression in TCL1A-eGFP-transduced HSCs was similar to that in *TET2*-mutant HSCs (Extended Data Fig. 8b). After 14 days, cultures from HSCs that received TCL1A-eGFP virus had approximately fourfold higher counts of phenotypic HSC/MPPs and colony-forming cells compared with cultures from HSCs that received control-eGFP virus (Fig. 4b), indicating that *TCL1A* expression was sufficient for HSC clonal expansion.

To assess whether *TCL1A* expression was sufficient to promote HSPC fitness in vivo, we infected c-Kit^{+} bone marrow cells from CD45.2 mice with TCL1A-eGFP or control-eGFP lentivirus and admixed these cells with competitor GFP^{-} CD45.2 whole bone marrow, with the proportion of GFP^{+} cells in the Lin^{-} fraction of the resulting cell mixture totalling around 4% in each group (Methods and Extended Data Fig. 9a). Following transplantation of these cells into lethally irradiated CD45.1 recipient mice, we tracked the proportion of GFP^{+} donor cells in blood over time ($n = 8$ per group). At 4 weeks after transplant, the proportion of donor GFP^{+} granulocytes and total leukocytes was similar in both groups, but over the subsequent 16 weeks the proportion of GFP^{+} blood cells increased in the mice that received TCL1A-eGFP-transduced cells but not in the mice that received control-eGFP transduced cells (Fig. 4c and Extended Data Fig. 9b). Twenty-two weeks after transplant, we assessed chimerism in the marrow. For our primary analysis, we examined the Lin^{-}c-$Kit^{+}SCA\text{-}1^{+}$ compartment that contains all relevant mouse HSC and MPP subsets and found a marked increase in the percentage of GFP^{+} donor cells in the mice given TCL1A-eGFP-transduced cells compared with mice given control cells (mean 23.8% versus 3.9%, $P = 0.0054$) (Fig. 4d). For secondary analyses, we also examined the different subsets of HSC/MPPs (LT-HSC, ST-HSC, MPP2, MPP3 and MPP4, as defined in Pietras et al.[35]) and found significant increases in the percentage of GFP^{+} cells in all these compartments in the mice receiving TCL1A-eGFP cells compared to mice receiving control cells (Extended Data Fig. 9c). These results provide in vivo confirmation of stem and progenitor cell expansion due to *TCL1A* expression.

To further characterize the effect of *TCL1A*, we assessed the cell cycle status of cultured human HSC/MPPs and observed that *TCL1A*-expressing cells were about twofold more likely to be cycling compared with control cells (Fig. 4e). To uncover the mechanism by which *TCL1A* promotes proliferation of HSCs, we transduced TCL1A-eGFP or control-eGFP into $CD34^{+}$ cells from two normal donors that were G/G or T/T at rs2887399, cultured GFP^{+} HSC/MPPs, and then performed cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) after seven days. After integration, dimensionality reduction and clustering (Methods), we annotated four clusters of HSC/MPPs as well as two populations of myeloid progenitors using the cell surface markers CD34, CD38, CD45RA, CD49f and CD11a (Fig. 4f, Extended Data Fig. 10a and Supplementary Table 16). Pseudotime[36] analysis supported a trajectory of progression from HSC/MPP1 (initial state) to 4 (most 'differentiated' state) (Extended Data Fig. 10b). HSC/MPP1 expressed stem-cell identity genes such as *MECOM*, *FAM30A* and *HEMGN*, as well as high levels of proliferative markers such as *MKI67*, *TOP2A*, *PCNA* and *CENPA* (Fig. 4g). By contrast, HSC/MPP2–4 expressed lower levels of stem cell identity genes and proliferative markers. Cell cycle analysis confirmed that these clusters contained cells that were predominantly in G0 or G1 phase (Extended Data Fig. 10c). HSC/MPP3–4 also displayed a progressive increase in genes associated with the integrated stress response such as *PPP1R15A* (also known as *GADD34*), *DDIT3* (also known as *CHOP*) and *ATF4*, as well as FOXO target genes such as *CDKN1A* (which encodes p21), *CDKN1B* (encoding p27), *SOD2*, *CCNG2* and *TXNIP* (Fig. 4g and Extended Data Figs. 10d and 11a). TCL1A has been reported to bind to and increase kinase activity of all AKT isoforms through an unknown mechanism[37], and one well-studied downstream consequence of active AKT is inhibition of FOXO-mediated transcription[38]. FOXO transcription factors can drive downstream target gene expression in an adaptive response to stressors to preserve cell viability, but prolonged activation of this response can lead to a terminal state of cell cycle arrest or apoptosis[39]. Indeed, cells in HSC/MPP4 also expressed the highest levels of apoptosis effector genes *BAD*, *BCL2L11* (encoding BIM) and *BBC3* (encoding PUMA). Of note, we found that *TCL1A* expression led to a significant increase in the proportion of cells in the HSC/MPP1 cluster, and a significant decrease in the proportion of cells in the HSC/MPP3 and HSC/MPP4 clusters, an effect that was consistent in both donors (Fig. 4h and Extended Data Fig. 11b,c). When considered in aggregate, the HSC/MPP clusters from *TCL1A*-expressing samples had reduced expression of FOXO target genes or gene sets and increased expression of cell cycle associated genes or gene sets compared with control samples (Supplementary Tables 17 and 18). This indicates that *TCL1A* may function to preserve HSCs in a proliferative state by avoiding prolonged, deleterious stress responses.

Discussion

We have developed an approach for inferring clonal expansion rate from a single time point and used it to perform a GWAS for CHIP clonal expansion rate (see also Supplementary Note 3). We found that a common variant with a large effect in the promoter of *TCL1A* was associated with a slower expansion rate and a markedly reduced prevalence of several common driver mutations in CHIP. This variant is likely to block the aberrant de-repression of *TCL1A*, which normally occurs in HSCs downstream of mutations in *TET2*, *ASXL1*, *SF3B1*, *SRSF2*, LOY and possibly mutations in other driver genes, thus implicating *TCL1A* expression as a dominant reason for positive selection of these clones. Necessity

and sufficiency experiments further supported *TCL1A* expression as a causal factor in clonal expansion of HSCs. Notably, our results suggest that pharmacologically targeting TCL1A may suppress the growth of CHIP and haematological cancers associated with mutations in these genes. PACER is a powerful approach for identifying the genetic and environmental factors mediating clonal expansion in humans at population scale and may be applied to any tissue in which pre-malignant clones exist[40–42].

Online content

Any methods, additional references, Nature Portfolio 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/s41586-023-05806-1.

1. Steensma, D. P. et al. Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood* **126**, 9–16 (2015).
2. Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **26**, 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. Abelson, S. et al. Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* **559**, 400–404 (2018).
6. Desai, P. et al. Somatic mutations precede acute myeloid leukemia years before diagnosis. *Nat. Med.* **24**, 1015–1023 (2018).
7. Jaiswal, S. et al. Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease. *N. Engl. J. Med.* **377**, 111–121 (2017).
8. Bick Alexander, G. et al. Genetic interleukin 6 signaling deficiency attenuates cardiovascular risk in clonal hematopoiesis. *Circulation* **141**, 124–131 (2020).
9. 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).
10. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *Nature* **590**, 290–299 (2021).
11. Bick, A. G. et al. Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature* **586**, 763–768 (2020).
12. Osorio, F. G. et al. Somatic mutations reveal lineage relationships and age-related mutagenesis in human hematopoiesis. *Cell Reports* **25**, 2308–2316.e4 (2018).
13. Mitchell, E. et al. Clonal dynamics of haematopoiesis across the human lifespan. *Nature* **606**, 343–350 (2022).
14. Williams, N. et al. Life histories of myeloproliferative neoplasms inferred from phylogenies. *Nature* **602**, 162–168 (2022).
15. Lee-Six, H. et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).
16. Fabre, M. A. et al. The longitudinal dynamics and natural history of clonal haematopoiesis. *Nature* **606**, 335–342 (2022).
17. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
18. Zink, F. et al. Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* **130**, 742–752 (2017).
19. Watson, C. J. et al. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science* **367**, 1449–1454 (2020).
20. Deuren, R. C. V. et al. Clone expansion of mutation-driven clonal hematopoiesis is associated with aging and metabolic dysfunction in individuals with obesity. Preprint at *bioRxiv* https://doi.org/10.1101/2021.05.12.443095 (2021).
21. Robertson, N. A. et al. Longitudinal dynamics of clonal hematopoiesis identifies gene-specific fitness effects. *Nat. Med.* **28**, 1439–1446 (2022).
22. van Zeventer, I. A. et al. Mutational spectrum and dynamics of clonal hematopoiesis in anemia of older individuals. *Blood* **135**, 1161–1170 (2020).
23. Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T. & Flicek, P. R. The ensembl regulatory build. *Genome Biol.* **16**, 56 (2015).
24. Carvalho-Silva, D. et al. Open Targets Platform: new developments and updates two years on. *Nucleic Acids Res.* **47**, D1056–D1065 (2019).
25. Narducci, M. G. et al. TCL1 is overexpressed in patients affected by adult T-cell leukemias. *Cancer Res.* **57**, 5452–5456 (1997).
26. Fishilevich, S. et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database* **2017**, bax028 (2017).
27. Thompson, D. J. et al. Genetic predisposition to mosaic Y chromosome loss in blood. *Nature* **575**, 652–657 (2019).
28. Malcovati, L. et al. Clinical significance of somatic mutation in unexplained blood cytopenia. *Blood* **129**, 3371–3378 (2017).
29. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
30. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
31. Regev, A. et al. The Human Cell Atlas. *eLife* **6**, e27041 (2017).
32. Velten, L. et al. Identification of leukemic and pre-leukemic stem cells by clonal tracking from single-cell transcriptomics. *Nat. Commun.* **12**, 1366 (2021).
33. Psaila, B. et al. Single-cell analyses reveal megakaryocyte-biased hematopoiesis in myelofibrosis and identify mutant clone-specific targets. *Mol. Cell* **78**, 477–492.e8 (2020).
34. Corces, M. R. et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
35. Pietras, E. M. et al. Functionally distinct subsets of lineage-biased multipotent progenitors control blood production in normal and regenerative conditions. *Cell Stem Cell* **17**, 35–46 (2015).
36. Trapnell, C. et al. Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. *Nat. Biotechnol.* **32**, 381–386 (2014).
37. Laine, J., Künstle, G., Obata, T., Sha, M. & Noguchi, M. The protooncogene *TCL1* is an akt kinase coactivator. *Mol. Cell* **6**, 395–407 (2000).
38. Brunet, A. et al. Akt promotes cell survival by phosphorylating and inhibiting a Forkhead transcription factor. *Cell* **96**, 857–868 (1999).
39. Eijkelenboom, A. & Burgering, B. M. T. FOXOs: signalling integrators for homeostasis maintenance. *Nat. Rev. Mol. Cell Biol.* **14**, 83–97 (2013).
40. Kakiuchi, N. & Ogawa, S. Clonal expansion in non-cancer tissues. *Nat. Rev. Cancer* **21**, 239–256 (2021).
41. Martincorena, I. et al. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
42. Martincorena, I. et al. Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
43. 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).

Joshua S. Weinstock[1,233], Jayakrishnan Gopakumar[2,233], Bala Bharathi Burugula[3], Md Mesbah Uddin[4], Nikolaus Jahn[2], Julia A. Belk[2], Hind Bouzid[2], Bence Daniel[2], Zhuang Miao[5], Nghi Ly[2], Taralynn M. Mack[6], Sofia E. Luna[7], Katherine P. Prothro[8], Shaneice R. Mitchell[2], Cecelia A. Laurie[9,10], Jai G. Broome[9,10,11], Kent D. Taylor[12,13], Xiuqing Guo[12,14], Moritz F. Sinner[15,16], Aenne S. von Falkenhausen[15,16], Stefan Kääb[15,16], Alan R. Shuldiner[17], Jeffrey R. O'Connell[17], Joshua P. Lewis[17,18], Eric Boerwinkle[19,20], Kathleen C. Barnes[21,22], Nathalie Chami[23,24], Eimear E. Kenny[25,26], Ruth J. F. Loos[23,24,27], Myriam Fornage[20,28], Lifang Hou[29], Donald M. Lloyd-Jones[29], Susan Redline[30,31], Brian E. Cade[4,30,31,32], Bruce M. Psaty[10,33,34,35], Joshua C. Bis[33], Jennifer A. Brody[10,33], Edwin K. Silverman[32,36], Jeong H. Yun[36], Dandi Qiao[32,36], Nicholette D. Palmer[37,38], Barry I. Freedman[39], Donald W. Bowden[37,38], Michael H. Cho[32,40], Dawn L. DeMeo[32,40], Ramachandran S. Vasan[41], Lisa R. Yanek[42,43], Lewis C. Becker[42,43], Sharon L. R. Kardia[44,45], Patricia A. Peyser[44,45], Jiang He[46,47], Michiel Rienstra[48], Pim Van der Harst[48], Robert Kaplan[49,50], Susan R. Heckbert[34,51], Nicholas L. Smith[34,51,52,53], Kerri L. Wiggins[33], Donna K. Arnett[54,55], Marguerite R. Irvin[56], Hemant Tiwari[57], Michael J. Cutler[58], Stacey Knight[58], J. Brent Muhlestein[58], Adolfo Correa[59,60], Laura M. Raffield[61], Yan Gao[62,63], Mariza de Andrade[64], Jerome I. Rotter[12,65], Stephen S. Rich[66,67], Russell P. Tracy[68,69], Barbara A. Konkle[70,71], Jill M. Johnsen[70,72], Marsha M. Wheeler[73], J. Gustav Smith[70,74,75,76], Olle Melander[77], Peter M. Nilsson[77], Brian S. Custer[78], Ravindranath Duggirala[79,80], Joanne E. Curran[79,80,81], John Blangero[79,80], Stephen McGarvey[82,83], L. Keoki Williams[84,85], Shujie Xiao[84], Mao Yang[84], C. Charles Gu[86,87], Yii-Der Ida Chen[12,14], Wen-Jane Lee[88,89], Gregory M. Marcus[90], John P. Kane[91], Clive R. Pullinger[92], M. Benjamin Shoemaker[93,94], Dawood Darbar[95,96], Dan M. Roden[97], Christine Albert[98,99], Charles Kooperberg[100,101], Ying Zhou[100], JoAnn E. Manson[32,102], Pinkal Desai[103,104], Andrew D. Johnson[105,106,107], Rasika A. Mathias[42,43], NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium*, Thomas W. Blackwell[1], Goncalo R. Abecasis[1,109], Albert V. Smith[1], Hyun M. Kang[1], Ansuman T. Satpathy[2], Pradeep Natarajan[4,53,192,193], Jacob O. Kitzman[3], Eric A. Whitsel[231], Alexander P. Reiner[53,100,206], Alexander G. Bick[6] ✉ & Siddhartha Jaiswal[2,232] ✉

[1]Center for Statistical Genetics, Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI, USA. [2]Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA. [3]Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA. [4]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA. [5]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA. [6]Division of Genetic Medicine, Department of Medicine, Vanderbilt University, Nashville, TN, USA. [7]Department of Pediatrics, Stanford University School of Medicine, Stanford, CA, USA. [8]Department of Biochemistry, Stanford University School of Medicine, Stanford, CA, USA. [9]Department of Biostatistics, University of Wash ington, Seattle, WA, USA. [10]University of Washington, Seattle, WA, USA. [11]Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA, USA. [12]Department of Pediatrics, The Institute for Translational Genomics and Population Sciences, The Lundquist Institute for Biomedical Innovation at Harbor–UCLA Medical Center, Torrance, CA, USA. [13]Institute for Translational Genomics and Populations Sciences, Lundquist Institute, Torrance, CA, USA. [14]Lundquist Institute, Torrance, CA, USA. [15]Department of Medicine I, University Hospital, LMU Munich, Munich, Germany. [16]German Centre for Cardiovascular Research (DZHK), partner site: Munich Heart Alliance, Munich, Germany. [17]Department of Medicine, University of Maryland, Baltimore, Baltimore, MD, USA. [18]University of Maryland,

Baltimore, MD, USA. [19]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA. [20]University of Texas Health at Houston, Houston, TX, USA. [21]Division of Biomedical Informatics and Personalized Medicine, Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [22]University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [23]The Charles Bronfman Institute of Personalized Medicine, New York, NY, USA. [24]The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [25]Institute for Genomic Health, New York, NY, USA. [26]Icahn School of Medicine at Mount Sinai, New York, NY, USA. [27]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [28]Brown Foundation Institute of Molecular Medicine, McGovern Medical School, University of Texas Health Science Center at Houston, Houston, TX, USA. [29]Department of Preventive Medicine, Northeastern University, Chicago, IL, USA. [30]Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. [31]Harvard Medical School, Boston, MA, USA. [32]Brigham and Women's Hospital, Boston, MA, USA. [33]Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA. [34]Department of Epidemiology, University of Washington, Seattle, WA, USA. [35]Department of Medicine, University of Washington, Seattle, WA, USA. [36]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, USA. [37]Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC, USA. [38]Department of Biochemistry, Wake Forest Baptist Health, Winston-Salem, NC, USA. [39]Department of Internal Medicine, Section on Nephrology, Wake Forest School of Medicine, Winston-Salem, NC, USA. [40]Channing Division of Network Medicine and Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, MA, USA. [41]National Heart Lung and Blood Institute's, Boston University's Framingham Heart Study, Framingham, MA, USA. [42]Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [43]Johns Hopkins University, Baltimore, MD, USA. [44]Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI, USA. [45]University of Michigan, Ann Arbor, MI, USA. [46]Department of Epidemiology, Tulane University School of Public Health and Tropical Medicine, New Orleans, LA, USA. [47]Tulane University, New Orleans, LA, USA. [48]Department of Cardiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands. [49]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY, USA. [50]Albert Einstein College of Medicine, New York, NY, USA. [51]Kaiser Permanente Washington Health Research Institute, Kaiser Permanente Washington, Seattle, WA, USA. [52]Seattle Epidemiologic Research and Information Center, Department of Veterans Affairs Office of Research and Development, Seattle, WA, USA. [53]Broad Institute, Cambridge, MA, USA. [54]College of Public Health, University of Kentucky, Lexington, KY, USA. [55]University of Kentucky, Lexington, KY, USA. [56]University of Alabama at Birmingham, Birmingham, AL, USA. [57]Department of Biostatistics, University of Alabama, Birmingham, AL, USA. [58]Intermountain Heart Institute, Intermountain Medical Center, Salt Lake City, UT, USA. [59]Department of Medicine, Jackson Heart Study, University of Mississippi Medical Center, Jackson, MS, USA. [60]Department of Population Health Science, University of Mississippi, Jackson, MS, USA. [61]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [62]Department of Medicine, University of Mississippi Medical Center, Jackson, MS, USA. [63]University of Mississippi, Jackson, MS, USA. [64]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, USA. [65]Department of Pediatrics, Lundquist Institute, Torrance, CA, USA. [66]Department of Public Health Sciences, Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA. [67]University of Virginia, Charlottesville, VA, USA. [68]Department of Pathology and Laboratory Medicine and Biochemistry, Larner College of Medicine at the University of Vermont, Colchester, VT, USA. [69]Department of Pathology and Laboratory Medicine, University of Vermont, Burlington, VT, USA. [70]Department of Cardiology, Clinical Sciences, Lund University and Skåne University Hospital, Lund, Sweden. [71]Blood Works Northwest, Seattle, WA, USA. [72]Research Institute, Bloodworks Northwest, Seattle, WA, USA. [73]Genome Science, University of Washington, Seattle, WA, USA. [74]The Wallenberg Laboratory, Department of Molecular and Clinical Medicine, Institute of Medicine, Gothenburg University, Gothenburg, Sweden. [75]Wallenberg Center for Molecular Medicine and Lund University Diabetes Center, Lund University, Lund, Sweden. [76]Department of Cardiology, Sahlgrenska University Hospital, Gothenburg, Sweden. [77]Department of Internal Medicine, Clinical Sciences, Lund University and Skane University Hospital, Malmo, Sweden. [78]Vitalant Research Institute, San Francisco, CA, USA. [79]Department of Human Genetics, University of Texas Rio Grande Valley School of Medicine, Brownsville, TX, USA. [80]South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley School of Medicine, Brownsville, TX, USA. [81]University of Texas Rio Grande Valley School of Medicine, Brownsville, TX, USA. [82]Department of Epidemiology and International Health Institute, Brown University School of Public Health, Providence, RI, USA. [83]Department of Epidemiology, Brown University, Providence, RI, USA. [84]Center for Individualized and Genomic Medicine Research (CIGMA), Department of Internal Medicine, Henry Ford Health System, Detroit, MI, USA. [85]Henry Ford Health System, Detroit, MI, USA. [86]Division of Biostatistics, Washington University School of Medicine, St Louis, MO, USA. [87]Washington University in St Louis, St Louis, MO, USA. [88]Department of Medical Research, Taichung Veterans General Hospital, Taichung, Taiwan. [89]Taichung Veterans General Hospital Taiwan, Taichung City, Taiwan. [90]Division of Cardiology, University of California, San Francisco, San Francisco, CA, USA. [91]Department of Medicine, Cardiovascular Research Institute, University of California, San Francisco, San Francisco, CA, USA. [92]Cardiovascular Research Institute, University of California, San Francisco, USA. [93]Division of Cardiology, Vanderbilt University Medical Center, Nashville, TN, USA. [94]Department of Medicine and Cardiology, Vanderbilt University, Nashville, TN, USA. [95]Division of Cardiology, University of Illinois at Chicago, Chicago, IL, USA. [96]University of Illinois at Chicago, Chicago, IL, USA. [97]Departments of Medicine, Pharmacology and Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA. [98]Department of Cardiology, Cedars-Sinai, Los Angeles, CA, USA. [99]Cedars-Sinai, Boston, MA, USA. [100]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [101]Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [102]Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [103]Division of Hematology and Oncology, Weill Cornell Medicine, New York, NY, USA. [104]Englander Institute of Precision Medicine, Weill Cornell Medicine, New York, NY, USA. [105]National Heart, Lung and Blood Institute, Population Sciences Branch, Framingham, MA, USA. [106]Population Sciences Branch, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, MD, USA. [107]National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, MD, USA. [109]Regeneron Pharmaceuticals, Tarrytown, NY, USA. [192]Department of Medicine, Harvard Medical School, Boston, MA, USA. [193]Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA. [206]Fred Hutchinson Cancer Research Center, University of Washington, Seattle, WA, USA. [231]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC, USA. [232]Institute for Stem Cell Biology and Regenerative Medicine, Stanford University School of Medicine, Stanford, CA, USA. [233]These authors contributed equally: Joshua S. Weinstock and Jayakrishnan Gopakumar. *A list of authors and their affiliations appears online. [✉]e-mail: alexander.bick@vumc.org; sjaiswal@stanford.edu

NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium

Namiko Abe[108], Gonçalo R. Abecasis[1,109], Francois Aguet[53], Christine Albert[98,99], Laura Almasy[110], Alvaro Alonso[111], Seth Ament[18], Peter Anderson[10], Pramod Anugu[63], Deborah Applebaum-Bowden[112], Kristin Ardlie[53], Dan Arking[43], Donna K. Arnett[54,55], Allison Ashley-Koch[113], Stella Aslibekyan[56], Tim Assimes[114], Paul Auer[115], Dimitrios Avramopoulos[43], Najib Ayas[116], Adithya Balasubramanian[19], John Barnard[117], Kathleen C. Barnes[21,22], R. Graham Barr[118], Emily Barron-Casella[43], Lucas Barwick[119], Terri Beaty[43], Gerald Beck[120], Diane Becker[121], Lewis C. Becker[42,43], Rebecca Beer[107], Amber Beitelshees[18], Emelia Benjamin[122], Takis Benos[123], Marcos Bezerra[124], Larry Bielak[45], Joshua Bis[33], Thomas W. Blackwell[1], John Blangero[79,80], Eric Boerwinkle[19,20], Donald W. Bowden[37,38], Russell Bowler[125], Jennifer A. Brody[10,33], Ulrich Broeckel[115], Jai G. Broome[9,10,11], Deborah Brown[126], Karen Bunting[108], Esteban Burchard[127], Carlos Bustamante[128], Erin Buth[9], Brian E. Cade[4,30,31,32], Jonathan Cardwell[129], Vincent Carey[32], Julie Carrier[130], Cara Carty[131], Richard Casaburi[132], Juan P. Casas Romero[32], James Casella[43], Peter Castaldi[30], Mark Chaffin[53], Christy Chang[18], Yi-Cheng Chang[133], Daniel Chasman[134], Sameer Chavan[129], Bo-Juen Chen[108], Wei-Min Chen[67], Yii-Der Ida Chen[12,14], Michael H. Cho[32,40], Seung Hoan Choi[53], Lee-Ming Chuang[135], Mina Chung[136], Ren-Hua Chung[137], Clary Clish[138], Suzy Comhair[139], Matthew Conomos[9], Elaine Cornell[140], Adolfo Correa[59,60], Carolyn Crandall[132], James Crapo[141], L. Adrienne Cupples[142], Joanne E. Curran[79,80,81], Jeffrey Curtis[45], Brian S. Custer[78], Coleen Damcott[18], Dawood Darbar[95,96], Sean David[143], Colleen Davis[10], Michelle Daya[129], Mariza de Andrade[64], Lisa de las Fuentes[144], Paul de Vries[145], Michael DeBaun[146], Ranjan Deka[147], Dawn L. DeMeo[32,40], Scott Devine[18], Huyen Dinh[19], Harsha Doddapaneni[19], Qing Duan[148], Shannon Dugan-Perez[19], Ravi Duggirala[149], Jon Peter Durda[140], Susan K. Dutcher[150], Charles Eaton[151], Lynette Ekunwe[63], Adel El Boueiz[152], Patrick Ellinor[153], Leslie Emery[10], Serpil Erzurum[117], Charles Farber[67], Jesse Farek[19], Tasha Fingerlin[154], Matthew Flickinger[1], Myriam Fornage[20,28], Nora Franceschini[155], Chris Frazar[10], Mao Fu[18], Stephanie M. Fullerton[10], Lucinda Fulton[87], Stacey Gabriel[53], Weiniu Gan[107], Shanshan Gao[129], Yan Gao[62,63], Margery Gass[101], Heather Geiger[108], Bruce Gelb[26], Mark Geraci[123], Soren Germer[108], Robert Gerszten[156], Auyon Ghosh[32], Richard Gibbs[19], Chris Gignoux[114], Mark Gladwin[123], David Glahn[157], Stephanie Gogarten[10], Da-Wei Gong[18], Harald Goring[158], Sharon Graw[22], Kathryn J. Gray[159], Daniel Grine[129], Colin Gross[1], C. Charles Gu[86,87], Yue Guan[18], Xiuqing Guo[12,14], Namrata Gupta[53], David M. Haas[160], Jeff Haessler[101], Michael Hall[161], Yi Han[19], Patrick Hanly[162], Daniel Harris[163], Nicola L. Hawley[164], Jiang He[46,47], Ben Heavner[9], Susan R. Heckbert[34,51], Ryan Hernandez[127], David Herrington[165], Craig Hersh[166], Bertha Hidalgo[56], James Hixson[20], Brian Hobbs[32], John Hokanson[129], Elliott Hong[18], Karin Hoth[167], Chao Agnes Hsiung[168], Jianhong Hu[19], Yi-Jen Hung[169], Haley Huston[71], Chii Min Hwu[89], Marguerite R. Irvin[56], Rebecca Jackson[170], Deepti Jain[10], Cashell Jaquish[107], Jill M. Johnsen[70,72], Andrew D. Johnson[105,106,107], Craig Johnson[10], Rich Johnston[111], Kimberly Jones[43], Hyun Min Kang[1], Robert Kaplan[49,50], Sharon L. R. Kardia[44,45], Shannon Kelly[127], Eimear E. Kenny[25,26], Michael Kessler[18], Alyna Khan[10], Ziad Khan[19], Wonji Kim[171], John Kimoff[172], Greg Kinney[173], Barbara A. Konkle[70,71], Charles Kooperberg[100,101], Holly Kramer[174], Christoph Lange[175], Ethan Lange[129], Leslie Lange[129], Cathy Laurie[10], Cecelia A. Laurie[9,10], Meryl LeBoff[32], Jiwon Lee[32], Sandra Lee[19], Wen-Jane Lee[88,89], Jonathon LeFaive[1], David Levine[10], Dan Levy[107], Joshua P. Lewis[17,18], Xiaohui Li[14], Yun Li[148], Henry Lin[14], Honghuang Lin[176], Xihong Lin[177], Simin Liu[178], Yongmei Liu[179], Yu Liu[180], Ruth J. F. Loos[23,24,27], Steven Lubitz[153], Kathryn Lunetta[176], James Luo[107], Ulysses Magalang[181], Michael Mahaney[81], Barry Make[43], Ani Manichaikul[67], Alisa Manning[182], JoAnn E. Manson[32,102], Lisa Martin[183], Melissa Marton[108], Susan Mathai[129], Rasika A. Mathias[42,43], Susanne May[9], Patrick McArdle[18], Merry-Lynn McDonald[56], Sean McFarland[171], Stephen McGarvey[82,83], Daniel McGoldrick[184], Caitlin McHugh[9], Becky McNeil[185], Hao Mei[63], James Meigs[186], Vipin Menon[19], Luisa Mestroni[22], Ginger Metcalf[19], Deborah A. Meyers[187], Emmanuel Mignot[188], Julie Mikulla[107], Nancy Min[63], Mollie Minear[189], Ryan L. Minster[123], Braxton D. Mitchell[18], Matt Moll[30], Zeineen Momin[19], May E. Montasser[18], Courtney Montgomery[190], Donna Muzny[19], Josyf C. Mychaleckyj[67], Girish Nadkarni[26], Rakhi Naik[43], Take Naseri[191], Pradeep Natarajan[4,53,192,193], Sergei Nekhai[194], Sarah C. Nelson[9], Bonnie Neltner[129], Caitlin Nessner[19], Deborah Nickerson[184], Osuji Nkechinyere[19], Kari North[148], Jeff O'Connell[195], Tim O'Connor[18], Heather Ochs-Balcom[196], Geoffrey Okwuonu[19], Allan Pack[197], David T. Paik[198], Nicholette D. Palmer[37,38], James Pankow[199], George Papanicolaou[107], Cora Parker[200], Gina Peloso[142], Juan Manuel Peralta[149], Marco Perez[114], James Perry[18], Ulrike Peters[201], Patricia A. Peyser[44,45], Lawrence S. Phillips[111], Jacob Pleiness[1], Toni Pollin[18], Wendy Post[202], Julia Powers Becker[203], Meher Preethi Boorgula[129], Michael Preuss[26], Bruce M. Psaty[10,33,34,35], Pankaj Qasba[107], Dandi Qiao[32,36], Zhaohui Qin[111], Nicholas Rafaels[129], Laura M. Raffield[61], Mahitha Rajendran[19], Vasan S. Ramachandran[176], D. C. Rao[87], Laura Rasmussen-Torvik[204], Aakrosh Ratan[67], Susan Redline[30,31], Robert Reed[18], Catherine Reeves[205], Elizabeth Regan[141], Alexander P. Reiner[53,100,206], Muagututi'a Sefuiva Reupena[207], Ken Rice[10], Stephen S. Rich[66,67], Rebecca Robillard[208], Nicolas Robine[108], Dan M. Roden[97], Carolina Roselli[53], Jerome I. Rotter[12,65], Ingo Ruczinski[43], Alexi Runnels[108], Pamela Russell[129], Sarah Ruuska[71], Kathleen Ryan[18], Ester Cerdeira Sabino[209], Danish Saleheen[118], Shabnam Salimi[18], Sejal Salvi[19], Steven Salzberg[43], Kevin Sandow[210], Vijay G. Sankaran[211], Jireh Santibanez[19], Karen Schwander[87], David Schwartz[129], Frank Sciurba[123], Christine Seidman[192], Jonathan Seidman[31], Frédéric Sériès[212], Vivien Sheehan[213], Stephanie L. Sherman[214], Amol Shetty[18], Aniket Shetty[129], Wayne Hui-Heng Sheu[89], M. Benjamin Shoemaker[93,94], Brian Silver[215], Edwin K. Silverman[32,36], Robert Skomro[216], Albert Vernon Smith[1], Jennifer Smith[45], Josh Smith[10], Nicholas L. Smith[34,51,52,53], Tanja Smith[108], Sylvia Smoller[50], Beverly Snively[217], Michael Snyder[114], Tamar Sofer[32], Nona Sotoodehnia[10], Adrienne M. Stilp[10], Garrett Storm[218], Elizabeth Streeten[18], Jessica Lasky Su[32], Yun Ju Sung[87], Jody Sylvia[32], Adam Szpiro[10], Daniel Taliun[1], Hua Tang[219], Margaret Taub[43], Kent D. Taylor[12,13], Matthew Taylor[22], Simeon Taylor[18], Marilyn Telen[113], Timothy A. Thornton[10], Machiko Threlkeld[220], Lesley Tinker[101], David Tirschwell[10], Sarah Tishkoff[221], Hemant Tiwari[57], Catherine Tong[9], Russell P. Tracy[68,69], Michael Tsai[199], Dhananjay Vaidya[43], David Van Den Berg[222], Peter VandeHaar[1], Scott Vrieze[199], Tarik Walker[129], Robert Wallace[167], Avram Walts[129], Fei Fei Wang[10], Heming Wang[223], Jiongming Wang[1], Karol Watson[132], Jennifer Watt[19], Daniel E. Weeks[123], Joshua S. Weinstock[1], Bruce Weir[10], Scott T. Weiss[224], Lu-Chen Weng[153], Jennifer Wessel[225], Cristen Willer[226], Kayleen Williams[9], L. Keoki Williams[84,85], Carla Wilson[32], James Wilson[227], Lara Winterkorn[108], Quenna Wong[10], Joseph Wu[198], Huichun Xu[18], Lisa R. Yanek[42,43], Ivana Yang[129], Ketian Yu[45], Seyedeh Maryam Zekavat[53], Yingze Zhang[228], Snow Xueyan Zhao[141], Wei Zhao[229], Xiaofeng Zhu[230], Michael Zody[108] & Sebastian Zoellner[1]

[108]New York Genome Center, New York, NY, USA. [110]Children's Hospital of Philadelphia, University of Pennsylvania, Philadelphia, PA, USA. [111]Emory University, Atlanta, GA, USA. [112]National Institutes of Health, Bethesda, MD, USA. [113]Duke University, Durham, NC, USA. [114]Stanford University, Stanford, CA, USA. [115]Medical College of Wisconsin, Milwaukee, WI, USA. [116]Department of Medicine, Providence Health Care, Vancouver, British Columbia, Canada. [117]Cleveland Clinic, Cleveland, OH, USA. [118]Columbia University, New York, NY, USA. [119]LTRC, The Emmes Corporation, Rockville, MD, USA. [120]Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland, OH, USA. [121]Department of Medicine, Johns Hopkins University, Baltimore, MD, USA. [122]Boston University School of Medicine, Boston University, Massachusetts General Hospital, Boston, MA, USA. [123]University of Pittsburgh, Pittsburgh, PA, USA. [124]Fundação de Hematologia e Hemoterapia de Pernambuco–Hemope, Recife, Brazil. [125]National Jewish Health, Denver, CO, USA. [126]Department of Pediatrics, University of Texas Health at Houston, Houston, TX, USA. [127]University of California, San Francisco, San Francisco, CA, USA. [128]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [129]University of Colorado at Denver, Denver, CO, USA. [130]University of Montreal, Montreal, Quebec, Canada. [131]Washington State University, Pullman, WA, USA. [132]University of California, Los Angeles, Los Angeles, CA, USA. [133]National Taiwan University, Taipei, Taiwan. [134]Division of Preventive Medicine, Brigham and Women's Hospital, Boston, MA, USA. [135]National Taiwan University Hospital, National Taiwan University, Taipei, Taiwan. [136]Cleveland Clinic, Cleveland, OH, USA. [137]National Health Research Institute Taiwan, Miaoli County, Taiwan. [138]Metabolomics Platform, Broad Institute, Cambridge, MA, USA. [139]Department of Immunity and Immunology, Cleveland Clinic, Cleveland, OH, USA. [140]University of Vermont, Burlington, VT, USA. [141]National Jewish Health, Denver, CO, USA. [142]Department of Biostatistics, Boston University, Boston, MA, USA. [143]University of Chicago, Chicago, IL, USA. [144]Department of Medicine, Cardiovascular Division, Washington University in St Louis, St Louis, MO, USA. [145]Human Genetics Center, Department of Epidemiology, Human Genetics and Environmental Sciences, University of Texas Health at Houston, Houston, TX, USA. [146]Vanderbilt University, Nashville, TN, USA. [147]University of Cincinnati, Cincinnati, OH, USA. [148]University of North Carolina, Chapel Hill, NC, USA. [149]University of Texas Rio Grande Valley School of Medicine, Edinburg, TX, USA. [150]Department of Genetics, Washington University in St Louis, St Louis, MO, USA. [151]Brown University, Providence, RI, USA. [152]Channing Division of Network Medicine, Harvard University, Cambridge, MA, USA. [153]Massachusetts General Hospital, Boston, MA, USA. [154]Center for Genes, Environment and Health, National Jewish Health, Denver, CO, USA. [155]Department of Epidemiology, University of North Carolina, Chapel Hill, NC, USA. [156]Beth Israel Deaconess Medical Center, Boston, MA, USA. [157]Department of Psychiatry, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA. [158]University of Texas Rio Grande Valley School of Medicine, San Antonio, TX, USA. [159]Department of Obstetrics and Gynecology, Mass General Brigham, Boston, MA, USA. [160]Department of Obstetrics and Gynecology, Indiana University, Indianapolis, IN, USA. [161]Department of Cardiology, University of Mississippi, Jackson, MS, USA. [162]Department of Medicine, University of Calgary, Calgary, Alberta, Canada. [163]Department of Genetics, University of Maryland, Philadelphia, PA, USA. [164]Department of Chronic Disease Epidemiology, Yale University, New Haven, CT, USA. [165]Wake Forest Baptist Health, Winston-Salem, NC, USA. [166]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, USA. [167]University of Iowa, Iowa City, IA, USA. [168]Institute of Population Health Sciences, NHRI, National Health Research Institute Taiwan, Miaoli County, Taiwan. [169]Tri-Service General Hospital National Defense Medical Center, New Taipei, Taiwan. [170]Department of Internal Medicine, DIvision of Endocrinology, Diabetes and Metabolism, Oklahoma State University Medical Center, Columbus, OH, USA. [171]Harvard University, Cambridge, MA, USA. [172]McGill University, Montreal, Quebec, Canada. [173]Department of Epidemiology, University of Colorado at Denver, Aurora, CO, USA. [174]Department of Public Health Sciences, Loyola University, Maywood, IL, USA. [175]Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA. [176]Boston University, Boston, MA, USA. [177]Harvard School of Public Health, Boston, MA, USA. [178]Department of Epidemiology and Medicine, Brown University, Providence, RI, USA. [179]Department of Cardiology, Duke University, Durham, NC, USA. [180]Cardiovascular Institute, Stanford University, Stanford, CA, USA. [181]Division of Pulmonary, Critical Care and Sleep Medicine, Ohio State University, Columbus, OH, USA. [182]Broad Institute, Harvard University, Massachusetts General Hospita l, Boston, MA, USA. [183]Department of Cardiology, George Washington University, Washington, DC, USA. [184]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [185]RTI International, Research Triangle Park, NC, USA. [186]Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. [187]University of Arizona, Tucson, AZ, USA. [188]Center For Sleep Sciences and Medicine, Stanford University, Palo Alto, CA, USA. [189]National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD, USA. [190]Department of Genes and Human Disease, Oklahoma Medical Research Foundation, Oklahoma City, OK, USA. [191]Ministry of Health, Government of Samoa, Apia, Samoa. [194]Howard University, Washington, DC, USA. [195]University of Maryland, Balitmore, MD, USA. [196]University at Buffalo, Buffalo, NY, USA. [197]Division of Sleep Medicine and Department of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [198]Stanford Cardiovascular Institute, Stanford University, Stanford, CA, USA. [199]University of Minnesota, Minneapolis, MN, USA. [200]Biostatistics and Epidemiology Division, RTI International, Research Triangle Park, NC, USA. [201]Department of Fred Hutch and UW, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [202]Department of Cardiology and Medicine, Johns Hopkins University, Baltimore, MD, USA. [203]Department of Medicine, University of Colorado at Denver, Denver, CO, USA. [204]Northwestern University, Chicago, IL, USA. [205]New York Genome Center, New York Genome Center, New York City, NY, USA. [207]Lutia I Puava Ae Mapu I Fagalele, Apia, Samoa. [208]Sleep Research Unit, University of Ottawa Institute for Mental Health Research, University of Ottawa, Ottawa, Ontario, Canada. [209]Faculdade de Medicina, Universidade de Sao Paulo, Sao Paulo, Brazil. [210]TGPS, Lundquist Institute, Torrance, CA, USA. [211]Division of Hematology and Oncology, Harvard University, Boston, MA, USA. [212]Université Laval, Quebec City, Quebec, Canada. [213]Department of Pediatrics, Emory University, Atlanta,

GA, USA. [214]Department of Human Genetics, Emory University, Atlanta, GA, USA. [215]UMass Memorial Medical Center, Worcester, MA, USA. [216]University of Saskatchewan, Saskatoon, Saskatchewan, Canada. [217]Department of Biostatistical Sciences, Wake Forest Baptist Health, Winston-Salem, NC, USA. [218]Department of Genomic Cardiology, University of Colorado at Denver, Aurora, CO, USA. [219]Department of Genetics, Stanford University, Stanford, CA, USA. [220]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [221]Department of Genetics, University of Pennsylvania, Philadelphia, PA, USA. [222]USC Methylation Characterization Center, University of Southern California, Los Angeles, Los Angeles, CA, USA. [223]Brigham and Women's Hospital, Mass General Brigham, Boston, MA, USA. [224]Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. [225]Department of Epidemiology, Indiana University, Indianapolis, IN, USA. [226]Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA. [227]Department of Cardiology, Beth Israel Deaconess Medical Center, Cambridge, MA, USA. [228]Department of Medicine, University of Pittsburgh, Pittsburgh, PA, USA. [229]Department of Epidemiology, University of Michigan, Ann Arbor, MI, USA. [230]Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH, USA.

Methods

Ethical approval

Informed consent was obtained by each of the participating TOPMed cohorts for all participants. The participating cohorts and institutional review boards are described in Supplementary Table 19, 'TOPMed studies included'. Blood DNA samples from WHI participants were obtained with informed consent and the CHIP longitudinal assessment study was reviewed and approved by the Fred Hutchinson Cancer Center Institutional Review Board (IRB no. 10186). Mobilized peripheral blood was obtained from donors by Fred Hutchinson Cooperative Center of Excellence in Hematology using protocols approved by the Fred Hutchinson Cancer Center Institutional Review Board.

Study samples

WGS was performed on 127,946 samples as part of 51 studies contributing to the Freeze 8 NHLBI TOPMed programme as previously described[10,11]. None of the TOPMed studies included individuals selected for sequencing because of haematologic malignancy. Each of the included studies provided informed consent. Information on the included cohorts, sequencing centres and ethical approvals is included in Supplementary Tables 19–21. Age was obtained for 82,807 of the samples: the median age was 55 years, the mean age was 52.5 years, and the maximum age was 98 years. The samples have diverse reported ethnicity (40% European, 32% African, 16% Hispanic/Latino and 10% Asian).

WGS processing, variant calling and CHIP annotation

BAM files were remapped to hg38 and harmonized through the functionally equivalent pipeline[44]. SNPs and indels were discovered across TOPMed and were jointly genotyped across samples using the GotCloud pipeline[45]. An SVM filter was trained to discriminate between high- and low-quality variants. Variants were annotated with snpEff 4.3 (ref. 46). Sample quality was assessed through Mendelian discordance, contamination estimates and sequencing convergence, among other quality control metrics.

Putative somatic single nucleotide variants and indels were called with GATK Mutect2 (ref. 17), which searches for sites where there is evidence for alternative reads that support evidence for variation, and then performs local haplotype assembly. We used a panel of normals to filter sequencing artefacts and used an external reference of germline variants to exclude germline calls. We deployed this pipeline on Google Cloud using Cromwell[47].

As described in our previous report[11], samples were annotated as having CHIP if the Mutect2 output contained at least one variant in a curated list of leukaemogenic driver mutations with at least three alt-reads supporting the call. We expanded the list of driver mutations to include those in recently identified CHIP genes[48], increasing the number of CHIP cases from our previous report. A special approach was required to identify somatic variants in *U2AF1* since an erroneous segmental duplication in the region of the gene in the hg38 reference genome resulted in a mapping score of zero during alignment of the FASTQ file[49]. We developed a Rust-HTSLIB binary (https://github.com/weinstockj/pileup_region) to specifically identify reads associated with the *U2AF1* variants S34F, S34Y, R156H, Q157P and Q157R. A minimum of five alternate reads was required to include a variant in the somatic set of CHIP calls. The variant set was judged to have a high likelihood of being somatic based on the strong age association for people carrying mutations as well as a high rate of co-mutation with other known drivers. The VAF was estimated by dividing the alternate read count by the total read count for *U2AF1*.

True passengers should very rarely be recurrent in a dataset, unlike many germline variants or technical artefacts. Therefore, we pruned our callset by identifying Mutect2 variants that appeared in only a single individual among the CHIP carriers and 23,320 additional controls for a total of 28,391 individuals. We excluded any variant that appeared in the TOPMed Freeze 5 germline callset (463 million variants). We excluded variants with a depth below 25 or above 100 and excluded any variants in low-complexity regions or segmental duplications, as these are challenging for variant calling. We only included somatic singletons that were aligned to the primary chromosomal contigs. We excluded any variant with a VAF exceeding 35% as these may be enriched for germline variants that were not included in our other filters. We used cyvcf2 (ref. 50) to parse the Mutect2 VCFs and encoded each variant in an int64 value using the variant key encoding[51]. Since different base substitutions varied in their association with age at blood draw, we selected only C>T and T>C mutations, as these were the most strongly age-associated in our data, consistent with prior work identifying such mutations as essential elements of the 'clock-like' signature[52]. We developed a bespoke Python application to perform the singleton identification and filtering.

Estimation of passenger mutation rate, clone fitness and clone birth date with PACER-HB

We developed a hierarchical Bayesian latent variable model using the Stan[53,54] probabilistic programming language. We used the negative binomial likelihood with a mean and over-dispersion parameterization to facilitate interpretation. We used the identity function to link the passenger counts to the predictors as we modelled the effects on an additive scale. We modelled the expectation and over-dispersion of the passenger counts observed at time t_i as

$$E(\text{counts}_i(t_i)) = \mu T_i + s_i(t_i - T_i) + \alpha_k$$

$$\text{counts}_i(t_i) \sim \text{NegativeBinomial}(E(\text{counts}_i(t_i)), I(i \in \text{CHIP})\theta_0 + (1 - I(i \in \text{CHIP}))\theta_1)$$

where T_i is the time of the driver acquisition for sample i with a blood draw at time t_i, μ is the mutation rate per diploid genome per year for the HSC population, s_i is the fitness of the clone, and α_k represents a study-specific random intercept for sample i included in study k. We can interpret $t_i - T_i$ as the lifetime of the clone in years. We used a negative binomial likelihood as there was over-dispersion relative to a Poisson distribution.

We included several constraints and priors on the parameters to make them identifiable. We constrained T_i to be positive but exceeded by t_i such that the parameter would be in yearly units. We included case-control specific over-dispersion terms θ_0 and θ_1 as the CHIP carriers had greater dispersion. To adjust for batch effects, we included a random intercept, as the amount of singletons in controls varied by study.

To include the constraint on T_i, we defined $T_i = \psi_i \times \text{age}_i$, with ψ_i constrained between 0 and 1, and age_i is the age at blood draw. We placed an uninformative Beta(1, 1.3) prior on ψ_i, which is equivalent to the supposition that the driver mutation is twice as likely to be acquired in the second half of life (at the time of blood draw) then the first. We assumed the study-specific deviations were exchangeable with respect to a $N(0, 20)$ prior, providing some shrinkage on the study-specific intercepts. We placed a $N(0, 1)$ prior on the s_i parameter to aid identification. Further details are described in the supplementary information.

To estimate the posterior, we used the Stan Hamiltonian Monte Carlo sampler with four separate chains, and used 400 samples of burn-in. We assessed convergence using the Rhat and effective sample size statistics. We tried multiple parameterizations to reduce the number of divergent transitions. We performed posterior predictive checks to assess the model fit.

Simulation of HSC dynamics

We simulated the number of cells within an HSC clone as a birth–death continuous time Markov chain, which models the size of an HSC clone as the composite of simultaneous Poisson birth and Poisson death point processes (Supplementary Note 1). Following Watson et al.[19], HSCs

could transition to one of three states: asymmetric renewal, symmetric self-renewal and symmetric differentiation. The rate of transition was determined by the symmetric differentiation rate of the cell per year, which was set to five. The symmetric self-renewal and symmetric differentiation increase and decrease the size of the HSC clone respectively. As asymmetric division does not affect the size of the clone, we did not explicitly simulate transition to this state. The proclivity towards self-renewal was determined by the fitness of the clone. We set the entire HSC population to acquire a single driver mutation during the 'lifetime' of the simulation.

Passengers were accumulated over time using a birth Poisson point process. We then calculated the number of 'detectable' passengers that preceded the acquisition of the driver based on whether the underlying clone had expanded to a large enough proportion of HSC cells. We examined the association between the number of detectable passengers and the fitness of the underlying HSC clone. We implemented this simulation in the Julia programming language 1.4 (ref. 55).

Fitness estimates for driver genes

We determined the association between the driver genes and the passenger counts using *DNMT3A* non-R882 mutations as the reference in a negative binomial regression using the glm.nb function from the MASS R package[56]. We included age, study cohort, VAF and sex as covariates. We included the genes that had at least 30 carriers in the dataset, excluding those with multiple driver genes mutated. To benchmark PACER, we compared the fitness estimate from our model (the coefficient for each gene using *DNMT3A* non-R882 mutations as the referent group) with the fitness estimates from supplementary Table 6 of Fabre et al.[16] (GeneEffect_mean + SiteEffect_mean variable). To transform the Fabre et al. gene level estimates to a scale comparable to the PACER estimates, we performed a linear regression of the log transformed fitness estimate against an independent variable indicating the driver gene, with *DNMT3A* non-R882 mutations as the reference level. To estimate the association between these fitness estimates and the PACER estimates, we performed weighted least squares regression of the Fabre et al. fitness estimates against the PACER gene fitness estimates, with the weights defined as $1/\text{Fabre}_{SE}$, where Fabre_{SE} is defined as the standard error of the Fabre et al. driver gene fitness estimate. For this comparison, we included genes that were reported in our PACER gene fitness estimates.

Amplicon sequencing of longitudinal samples in WHI

We performed targeted sequencing of the CHIP driver genes using single-molecule molecular inversion probe sequencing (smMIPS[11,57]) on two blood DNA samples taken approximately 14–19 years apart from 900 individuals not previously assessed for CHIP as well as 55 individuals known to have a single CHIP mutation from TOPMed WGS from the WHI. Women aged 50–79 years were enroled from 40 WHI clinical centres in the USA between 1993 and 1998. All WHI participants had a blood sample collected at the time of enrolment, and a subset had subsequent blood sample collected 14–19 years later. Reads were aligned with bwa-mem to hg38 and processed with the mimips pileline[58]. We called somatic variants using an ensemble of VarScan[59], Mutect2 (ref. 17) and manual inspection with IGV[60] as previously described[61]. Including the 55 individuals previously known to have CHIP, a total of 455 individuals were identified to have clonal haematopoiesis at a VAF threshold for inclusion of variants of >0.005, and 351 of these had mutations in *DNMT3A*, *TET2*, *ASXL1*, or *SF3B1*.

Prediction of future growth in WHI

We used longitudinal sequencing data from the 55 CHIP carriers from WHI with WGS done at baseline to assess whether passengers could predict future clone growth rate. To determine the change in clone size over time ($\text{dVAF}/\text{d}T$), we divided the change in VAF at the two time points (from smMIPS) by the change in age in years. Of the 55 CHIP carriers, 15 had clones which had negative $\text{dVAF}/\text{d}T$. It was unlikely that these driver mutations had negative fitness since they had expanded to detectable levels in the blood starting from a single mutant cell. For these 15 carriers, we set the $\text{dVAF}/\text{d}T$ to 0, since we presumed the negative change in clone size observed was due to short-term factors not related to intrinsic fitness of the clone, such as a change in blood cell differential across time leading to an apparently lower VAF at the second time point or stochastic drift. We then performed a series of linear models with inverse normal transformed $\text{dVAF}/\text{d}T$ as the dependent variable and age at first blood draw, VAF, and passenger count as the independent variables. Model performance was assessed with adjusted R^2 and Akaike information criterion for each model. We performed hypothesis testing of the passenger count coefficient using a Wald test.

Bayesian logistic growth model of clonal expansion

We used longitudinal sequencing data from 351 clonal haematopoiesis carriers (VAF > 0.005) with mutations in *DNMT3A*, *TET2*, *ASXL1* or *SF3B1*, as identified using smMIPS described above, to test whether the T allele at rs2887399 altered clonal expansion rate. To estimate the rate of clonal expansion in the CHIP carriers in units of per cent growth per year, we developed a Bayesian logistic growth model. The model includes four terms that encode the growth rate of *DNMT3A, TET2, ASXL1* and *SF3B1* carriers with the rs2887399 G/G genotype, and four interaction terms that estimate how the rate of clonal expansion is modified for each additional T allele at rs2887399. We modelled the observed number of mutated alleles using a beta-binomial likelihood, and included a random intercept and slope for each individual donor:

$$x_i = (\text{Gene}_{ij} + R_i \times \text{Gene}_{ij} + U_{i1}) \times \text{age} + U_{i2}$$

$$q_i = \frac{0.5}{1 + e^{-x_i}}$$

$$P(Y_i = y) = P(\text{beta binomial}(q_i, \beta, D_i) = y)$$

We defined Gene_{ij} as an indicator matrix that describes the mutation type of the donor. We defined R_i as the number of rs2887399 T alleles in the ith individual. β is included as an over-dispersion term for the likelihood, and D_i indicates the sequencing depth of the CHIP mutation. We included the following priors:

$$\text{Gene}_{ij} \sim \text{Normal}(0, 0.20)$$

$$R_i \sim \text{Normal}(0, 0.05)$$

$$U_{i1} \sim \text{Normal}(0, 0.05)$$

We performed inference using the MCMC sampler implementation available in the RStan probabilistic programming language[53,54].

Single variant association

Single variant association for each variant in the TOPMed Freeze 8 germline genetic variant callset[10] with a MAC > 20 was performed with SAIGE[43] using the TOPMed Encore analysis server. To identify associations between rs2887399 and the presence of specific CHIP mutations, we used the same methods as our previous report on an analysis set of 74,974 individuals, including 4,697 cases and 70,277 controls[11]. Age, genotype inferred sex, the first ten genetic ancestry principal components, and study were included as covariates.

We performed SAIGE single variant association analyses on the passengers including age at blood draw, sex, VAF, study, and the first ten genetic ancestry principal components as covariates. We applied an inverse normal transformation to the passenger counts. We included 3,931 CHIP carriers with a single driver mutation and available age at blood draw. We declared variants from this analysis as significant if their P value was less than 5×10^{-8}.

Estimation of association between rs2887399 genotypes and CHIP mutation acquisition

We coded the rs2887399 genotypes as a categorical variable rather than a linear quantitative coding to estimate effects separately for the heterozygotes and the T-homozygotes using the G-homozygotes as the reference level. We estimated the associations using firth logistic regression to reduce bias in estimation resulting from low cell counts[62], and included age, genotype inferred sex and the first ten genetic ancestry components as covariates.

Fine-mapping of the *TCL1A* region

We applied the SuSIE[63] algorithm to the genotypes included in a 200-kb region surrounding *TCL1A*. We used the same covariates as the single variant association analysis. We used the posterior inclusion probabilities and credible sets identified by SuSIE to identify the putative causal variant. We used linkage disequilibrium directly calculated on the genotypes as opposed to an external reference.

Rare variant analyses

We performed a rare variant association study from gene-based tests on 1,698 cancer-associated genes and their flanking regions using the SCANG[64] procedure. We identified these genes by downloading the targets associated with cancer in Open Targets[24], and then filtered to include only genes with an association score of 1.0. The most prevalent CHIP driver genes were included among this list. We used the inverse normal transformed passenger counts as the phenotype with the same covariates as before. We specified the minimum size of the grouped regions as 30 variants and the maximum as 200. We included all PASS variants with a minor allele count greater than 4 and less than 300 (MAF of 3.7% in the analysed samples). We parsed the genotypes using cyvcf2 (ref. 50) and stored them as dgCMatrix using the Matrix[65] package from the R 4.1.2 programming language[66].

We set the *P* value filter to calculate SKAT test statistics at 5×10^{-4}. We did not group the variants by annotation and we declared regions as significant if their *P* value was less than 2.9×10^{-5} (0.05/1,698). We controlled for relatedness by incorporating a sparse kinship matrix as estimated by the PC-AiR method from the GENESIS R package[67]. We specified separate residual variance terms for each study to control for heterogeneous residual variance. We grouped together all studies where the number of analysed samples was less than 200.

Re-analysis of scRNA-seq data

The cell-by-gene count matrix data for each sample from Psaila et al.[33], generated using the 10X Genomics platform, was downloaded from Gene Expression Omnibus (GSE144568). Each matrix was loaded in Seurat[68] with the read10X command, and only cells with a minimum of 200 features were retained using the CreateSeuratObject command. Data was log normalized using a scale factor of 10,000 by the NormalizeData command. We then used the FindVariableFeatures command with the vst selection method and 2,000 features. The data was scaled using ScaleData using all genes as features. We then used the RunPCA command with VariableFeatures identified earlier. For clustering, we used FindNeighbors set to the first 10 principal component analysis (PCA) dimensions and FindClusters using a resolution of 0.5. We excluded samples that did not have a distinct cluster of HSC/MPPs, defined as clusters enriched for cells that were $CD34^+CD38^{-/lo}THY1^+$. This left five healthy marrow samples (id01, id06, id09, id13 and id17) and four myoproliferative neoplasm samples (id2, id7, id11 and id14). For each of these samples, we assessed the number of cells with *TCL1A*, *TCL1B* or *TCL6* transcripts within the cluster or clusters that contained HSC/MPPs, as defined above.

Additional preprocessed scRNA-seq data from Velten et al.[32], generated using MutaSeq, was downloaded from https://doi.org/10.6084/m9.figshare.12382685.v1 as an RDS file. We utilized data from one patient with AML (P1) and the healthy control (H1). We then determined the number of cells containing *TCL1A*, *TCL1B* or *TCL6*, transcript in the pre-leukaemic 'HSC/MPP' and pre-leukaemic '$CD34^+$ blasts and HSPCs' clusters for the P1 sample and the 'HSC/MPP' cluster for the H1 sample, in both cases as defined by the original study authors.

Re-analysis of ATAC-seq data

We obtained ATAC-seq data for AML samples as well as healthy controls from Corces et al.[34] available at Gene Expression Omnibus (GSE74912). For our analysis, we used data from HSCs, defined as $Lin^-CD34^+CD38^-CD90^+CD10^-$ by the authors, from 4 healthy donors (4983, 6792, 2596 and 7256), or pHSCs, defined as $Lin^-CD34^+CD38^-TIM3^-CD99^-$ by the authors. For the pHSC samples, we selected three where there were no detectable driver mutations in the pHSC compartment (SU336, SU306 and SU623), two where there were founding *DNMT3A* mutations only (SU444 and SU575), and three where there were founding *TET2* mutations only (SU070, SU501 and SU048).

Fastq files for these samples were downloaded, and ATAC-seq data analysis was performed as previously described[69]. In brief, reads were trimmed and filtered using fastp and mapped to the hg38 reference genome using hisat2 with the --no-spliced-alignment option. Bam files were deduplicated using Picard. Only reads mapping to chromosomes 1–22 and the X chromosome were retained; Y chromosome reads, mitochondrial reads and other reads were discarded. Genome track files were created by loading the fragments for each sample into R, and exporting bigwig files normalized by reads in TSSs using 'rtracklayer::export'. Coverage files were visualized using the Integrative Genomics Viewer. A counts matrix was created as described previously[34]. Peaks were called individually for each sample using MACS2 and then iteratively merged into a union peak set of high confidence disjoint fixed width peaks of 500 bp encompassing all peaks in all samples. Then, bias-corrected Tn5 insertions in each sample overlapping each peak location were counted, and the resulting counts matrix was imported into DESeq2 for statistical analysis. For differential accessibility analysis, we compared all peaks in the three *TET2*-mutant samples to the seven control samples using the DESeq function in the DESeq2 (ref. 70) R package (https://bioconductor.org/packages/release/bioc/html/DESeq2.html). Adjusted *P* values were calculated on the full set of peaks, and those with a FDR *q* value of <0.10 were retained for further analysis. The peaks that overlap with TSSs of protein coding genes are supplied in Supplementary Table 13.

CRISPR–Cas9 editing of $CD34^+$ human HSPCs

$CD34^+$ HSPCs from adult donors were purchased from the Cooperative Center of Excellence in Hematology (CCEH) at the Fred Hutch Cancer Research Center, Seattle, USA. TCL1A rs2887399 genotyping was performed using ThermoFisher SNP assay (assay ID: C_15842295_20). $CD34^+$ cells were thawed and cultured in HSPC expansion medium (StemSpanII + 10% $CD34^+$ expansion supplement + 0.1% penicillin-streptomycin) for 48 h before CRISPR editing. Editing of *AAVS*, *TET2*, *DNMT3A* and *ASXL1* was performed by electroporation of Cas9 RNP. For each combination of rs2887399 genotype and gRNA (Supplementary Table 22), 100,000 cells were incubated with 3.2 µg Synthego synthetic sgRNA guide and 8.18 µg of IDT Alt-R *Streptococcus pyogenes* Cas9 Nuclease V3 for 15 min at room temperature before electroporation. $CD34^+$ cells were resuspended in 18 µl of Lonza P3 solution and mixed with the ribonucleoprotein complex, and then transferred to Nucleocuvette strips for electroporation with program DZ-100 (Lonza 4D Nucleofector). Immediately following electroporation, each condition of 100,000 cells was transferred to 2 ml of HSPC expansion medium and allowed to recover for 24 h. CRISPR editing efficiency was measured using Sanger sequencing and ICE analysis. Statistical methods to predetermine sample sizes were not used and investigators were not blinded to experimental conditions for all experiments using human HSPCs.

ATAC-seq

Twenty-four hours after electroporation, $Lin^-CD34^+CD38^-CD45RA^-$ cells were sorted from the electroporated $CD34^+$ cells using a BD FACS Aria III. Cells were allowed to culture for 5–7 days in HSPC medium before 40,000 cells were collected, and bulk Omni-ATAC[69] was performed on them. In brief, cells were lysed with ATAC resuspension buffer containing 0.1% NP-40, 0.1% Tween-20, and 0.01% digitonin for 3 min, and then the transposition was performed for 30 min at 37 C using 100 nM of Illumina Tagment DNA TDE1 Enzyme and Buffer Kit per 50,000 cells. The fragmented DNA was then cleaned up using a Zymo DNA Clean and Concentrator-5 Kit (D4014). The transposed fragments were amplified and indexed using NEBNext 2× Master Mix. The final PCR product was purified using the Zymo DNA Clean and Concentrator-5 Kit. Prior to sequencing, the quality of the libraries was evaluated via DNA High Sensitivity Bioanalyzer assays. The sequencing was performed using 2 × 75 bp reads on an Illumina NextSeq550 instrument using the High Output Kit.

ATAC-seq data analysis was performed as described above. In brief, reads were trimmed and filtered using fastp and mapped to the hg38 reference genome using hisat2 (ref. 71) with the --no-spliced-alignment option. BAM files were deduplicated using Picard. Only reads mapping to chromosomes 1–22 and X chromosome were retained—Y chromosome reads, mitochondrial reads and other reads were discarded. Genome track files were created by loading the fragments for each sample into R, and exporting bigwig files normalized by reads in TSSs using 'rtracklayer::export'. Coverage files were visualized using the Integrative Genomics Viewer. ATAC-seq tracks were normalized based on counts in TSS and were visualized using the same scale for all tracks in Integrated Genome Viewer. For the tracks shown in Extended Data Fig. 6b, the same experimental strategy was used as above, except cells were sorted based on the markers $CD34^+CD38^-CD45RA^-Lin^-$ after seven days in culture, from which point the Omni-ATAC protocol was followed. We used the top 1,000 most accessible TSSs genome-wide to perform normalization. We devised this strategy based on our observation that some inaccessible TSSs were prone to noise, which confounded the normalization. Differential accessibility analysis was done as described above except the *TCL1A* TSS peak was manually defined as the 300-base pair region around rs2887399 (chr14:95714209-95714508, and DESeq2 was used in a model that included edit (AAVS1, *TET2* or *DNMT3A*) and number of rs2887399 T alleles (0, 1 or 2). Results for nominally significant TSS peaks in the *TET2*-edited versus AAVS1-edited samples can be found in Supplementary Table 14.

Liquid culture expansion assay

$Lin^-CD34^+CD38^-CD90^+CD45RA^-$ cells were sorted on a BD FACS Aria III from the electroporated $CD34^+$ cells. All cells were collected and stained with the extracellular HSPC marker panel in 100 µl PBS + 2% FBS + 1 mm EDTA (Supplementary Table 23). For each replicate, 500–1,000 $Lin^-CD34^+CD38^-CD90^+CD45RA^-$ cells were sorted into 100 µl HSC expansion medium and cells were plated into a 96-well plate. The wells on the edges of the 96-well plate were filled with water to keep the cultures hydrated. Four days post sort, another 100 µl of HSC expansion medium was added to each well. Ten days post sort, the samples were transferred from the 96-well plate to a 48-well plate and an additional 400 µl of HSPC expansion medium was added. Fourteen days post sort, the cells were collected and live cells were counted using trypan blue and haemocytometer. Additionally, the cells were stained with the extracellular HSPC marker panel, and flow cytometry analysis was performed using FlowJo v10.8.1. Absolute number of HSC/MPPs (defined as $Lin^-CD34^+CD38^-CD45RA^-$) and $CD45RA^{lo}$ progenitors (defined as $Lin^{-/lo}CD34^+CD38^-CD45RA^{lo}$) were determined by multiplying the total cell count at 14 days by the percentage of cells in each compartment as determined by flow cytometry. Example gating for the HSC stain is shown in Supplementary Fig. 4a.

Flow cytometry for TCL1A staining

Anti-human TCL1A antibody clone eBio1-21 was obtained from ThermoFisher. The specificity of the antibody was assessed by staining NALM6 cells that had been CRISPR-edited for complete loss of *TCL1A* with the antibody, which confirmed only a very low level of non-specific binding.

To assess for TCL1A expression in edited human $CD34^+$ HSPCs, cells in HSPC expansion medium were grown using culture conditions as described above, then collected and intracellularly stained 11 days following electroporation. Cells were first stained with the Live/Dead and extracellular surface markers simultaneously for 30 min in the dark on ice. After a PBS wash, cells were stained with 100 µl of IC fixation buffer for 30 min in the dark at room temperature. Cells were then washed twice with 1× permeabilization buffer. Next, cells were resuspended in 100 µl of 1× permeabilization buffer, and blocked with 2 µl of goat serum and 2.5 µl of TruStain FcX for 15 min in the dark at room temperature. Next, 1 µg of e450 antibodies (anti-TCL1A or isotype control) was added to each sample tube and stained for 30 min in the dark at room temperature (Supplementary Table 24). Cells were then washed twice with 1× permeabilization buffer and then resuspended in PBS before flow cytometry was performed. Analysis was performed using FlowJo v10.8.1.

Lentivirus plasmids for TCL1A knockdown and expression

For knockdown of *TCL1A*, we obtained plasmids for 4 separate shRNAs targeting *TCL1A*, as well as scramble control shRNA, from Origene (TL301172V). The shRNA constructs were validated to knockdown TCL1A protein by flow cytometry in NALM6 cells (from R. Levy). NALM6 cells were tested for mycoplasma prior to use and not further authenticated.

An insert containing the *TCL1A* coding region followed in frame with GFP (TCLA1-T2A linker-GFP) under the control of mammalian *EEF1A1* promoter, as well as a control sequence composed of GFP under the *EEF1A1* promoter, was synthesized by Gene Universal. The insert was cloned into a second-generation lentivirus backbone, adapted from the Addgene vector pMH0001, using enzymatic cloning. Briefly both the insert and backbone were digested with MluI and SbfI enzymes (NEB) and ligated using the T4 ligase (NEB). NEB DH5a competent bacteria were transformed with the ligation product. The transformed bacteria were screened by Ampilicin resistance and grown in liquid culture in LB medium to amplify the plasmid. Maxiprep plasmid purification (Macherey-Nagel NucleoBond Xtra Maxi) was performed to obtain the final purified plasmid used for lentivirus production.

Lentivirus production

Plasmids were transfected into 293T HEK cells (ATCC CRL-3216) at roughly 80% confluence in 10 cm tissue culture plates coated with poly-D-lysine using Lipofectamine 3000. 293T HEK cells were not further authenticated or tested for mycoplasma. The lipofectamine medium was exchanged 16 h later, and the viral supernatant was collected at 72 h post-transfection. The collected viral supernatant was filtered via a 0.45 µm filtration unit, and concentrated using the LentiX concentrator (Takara) for 2 h at 4 C and then spun down at 1,500*g* for 45 min at 4 C. The concentrated supernatant was subsequently aliquoted, flash frozen and stored at −80 °C until use.

Combined CRISPR and shRNA assay

$CD34^+$ cells were thawed and cultured in HSPC expansion medium (StemSpanII + 10% $CD34^+$ expansion supplement + 0.1% penicillin-streptomycin) for 48 h before CRISPR editing. Editing of AAVS, *TET2*, *DNMT3A* and *ASXL1* was performed by electroporation of Cas9 RNP. For each combination of rs2887399 genotype and gRNA, 100,000 cells were incubated with 3.26 µg of Synthego synthetic sgRNA guide and 8.332 ug of IDT Alt-R *S. pyogenes* Cas9 Nuclease V3 for 15 min at room temperature before electroporation. $CD34^+$ cells were resuspended

in 18 µl Lonza P3 solution and mixed with the ribonucleoprotein complex, and then transferred to Nucleocuvette strips for electroporation with program DZ-100 (Lonza 4D Nucleofector). Immediately following electroporation, each condition of 500,000 cells was transferred to 2 ml HSPC expansion medium and allowed to recover for 8 h. Later that same day, 250,000 CRISPR-edited cells were collected, spun down, and resuspended in a final volume of HSPC lentivirus medium (StemSpanII + 10% CD34+ expansion supplement + 0.1% penicillin-streptomycin + 10 µM prostaglandin E2 + 100 ng µl^{-1} poloxamer 407) with virus added at a multiplicity of infection (MOI) of 20. Cells were plated in a 96-well U-bottom plate for 16 h. shRNA-A and the scramble shRNA from Origene TL301172V were used for this experiment. Following a 16-h incubation, cells were washed in PBS, and then plated in 2 ml HSPC expansion medium. After 72 h, previously described liquid culture expansion assay was done on sorted Lin$^-$CD34$^+$CD38$^-$CD45RA$^-$GFP$^+$ cells, with assessment of counts and flow cytometry after 14 days.

Lentiviral *TCL1A* expression in human HSPCs

CD34$^+$ cells were thawed and cultured in HSPC expansion medium (StemSpanII + 10% CD34$^+$ expansion supplement + 0.1% penicillin-streptomycin) for 48 h before lentivirus transduction. In total, 750,000 cells were collected, spun down, and resuspended in a final volume of HSPC lentivirus medium (StemSpanII + 10% CD34$^+$ expansion supplement + 0.1% penicillin-streptomycin + 10 µM prostaglandin E2 + 100 ng µl^{-1} poloxamer 407) with virus added at an MOI of 100. Cells were plated in a 96-well U-bottom plate for 16 h. eGFP control was purchased from Origene (PS100093V) or produced in house as described above, and the TCL1A-eGFP nucleotide was purchased from Origene (RC204243L4V) or produced in house as described above. Following 16-h incubation, cells were washed in PBS, and then plated in 2 ml HSPC expansion medium. After 72 h, previously described liquid culture expansion assay was done using sorted Lin$^-$CD34$^+$CD38$^-$CD45RA$^-$GFP$^+$ cells. After 14 days, cells were collected and assessed for HSC/MPP frequency using flow cytometry as previously described. The total HSC/MPP count was determined by multiplying the percentage of live cells that were in the HSC/MPP gate by the total live cell count for each replicate.

After 14 days of in vitro liquid culture expansion, 800 live cells were sorted, resuspended in 1.1 ml Methocult + 0.1% penicillin-streptomycin, and plated in 35 mm dishes. Eight 35 mm dishes were placed in one 245 × 245 mm square dish along with four open 35 mm dishes of water and one 120 mm dish of water. After 14 days in Methocult, the number of colony-forming units was counted. The total colony-forming unit count in the day 14 liquid culture was determined by multiplying the number of colony-forming units in each replicate by the total live cell count after 14 days of liquid culture and dividing by 800.

For cell cycle analysis, sorted HSCs were cultured for 10 days in liquid culture expansion medium. Cells were first stained with the Alexa-700 Live/Dead and extracellular surface markers simultaneously for 30 min in the dark on ice (Supplementary Table 25). After a PBS wash, cells were stained with 100 µl of IC fixation buffer for 30 min in the dark at room temperature. Cells were then washed twice with 1× permeabilization buffer. Next, cells were resuspended in 100 µl of 1× permeabilization buffer, and blocked with 2 µl of goat serum for 15 min in the dark at room temperature. Cells were then washed twice with 1× permeabilization buffer and then resuspended in 75 µl of 1 µg ml^{-1} DAPI diluted in 1× permeabilization buffer. After 10 min, 75 µl PBS was added, and then flow cytometry was performed. HSC/MPPs were defined as CD34$^+$CD38$^-$Lin$^-$. Example gating for the DAPI HSPC analysis is shown in Supplementary Fig. 4b.

Mouse bone marrow competitive transplant

Mice were obtained from The Jackson Laboratory and housed at the Research Animal Facility of the Stanford School of Medicine. All experiments used female mice. The mice were housed under a 12-h light:12-h dark cycle with dark hours from 18:30–06:30 and housed at 20–23 °C under 40–60% humidity. All animal procedures were performed in accordance with protocols approved by Stanford University's Administrative Panel on Laboratory Animal Care. Statistical methods to predetermine samples size were not used and investigators were not blinded to experimental conditions.

Bone marrow from 10-week-old female CD45.2$^+$ C57BL/6 mice was collected, and c-Kit cells were enriched for using the EasySep Mouse cKIT Positive Selection Kit (18757) according to the manufacturer's protocol. 2.8 million c-KIT enriched cells were transduced with 45 µl of the previously described control-eGFP or TCL1A-eGFP and cultured overnight in U-bottom plates in mouse HSC transduction medium (StemSpan II, 10 ng ml^{-1} SCF, 100 ng ml^{-1} TPO, 10 µM prostaglandin E2, 100 ng µl P407, 0.1% penicillin-streptomycin) with an expected transduction efficiency of ~10%. Following overnight transduction, transduced c-KIT cells were washed with PBS and admixed with fresh CD45.2$^+$GFP$^-$ competitor whole bone marrow at a 1:3 ratio to achieve chimeric donor bone marrow graft. Sorting of GFP$^+$ cells pre-transplant was not conducted because anecdotal evidence from several laboratories suggests that culture of transduced HSCs for >24 h diminishes their potency for in vivo reconstitution. Post hoc analysis of stored aliquots from the input cells confirmed ~4% of Lin$^-$ cells were GFP$^+$ for both conditions, mimicking a CHIP clone of ~2% VAF (Extended Data Fig. 9a).

For the bone marrow transplant, recipient 9-week-old female CD45.1$^+$ C57BL/6 mice were lethally irradiated with one 950 cGy dose of γ-irradiation. Post-irradiation, recipients were transplanted with 1×10^6 of the previously described chimeric bone marrow in suspension via retro-orbital injection, $n = 8$ per group. Following transplantation, recipient mice were fed with Envigo Uniprim diet for four weeks.

The proportion of GFP$^+$ donor cells was tracked by collecting 100 µl of peripheral blood retro-orbitally at 4 weeks, 7 weeks, 12 weeks and 20 weeks post-transplant. Following RBC lysis, peripheral blood was stained with 100 µl of the mouse peripheral blood antibody cocktail (Supplementary Table 26). Twenty-two weeks post-transplant, mice were euthanized and bone marrow was collected from femurs. Following RBC lysis, bone marrow was stained with 50 µl of the mouse bone marrow antibody cocktail to determine the proportion of GFP$^+$ HSC or MPP donor cells (Supplementary Table 27).

Flow cytometry gating schema are shown in Supplementary Fig. 5a,b. Flow cytometry analysis was performed using FlowJo v10.8.1.

CITE-seq cell preparation and 10X workflow

Human CD34$^+$ cells were thawed and cultured in HSPC expansion medium (StemSpanII + 10% CD34$^+$ expansion supplement + 0.1% penicillin-streptomycin) for 48 h before lentiviral transduction. Seventy-two hours after lentivirus addition, Lin$^-$CD34$^+$CD38$^-$CD45RA$^-$GFP$^+$ were sorted and plated. Seven days after sort, 10X 3′ v3.1 with feature barcoding was performed. 60,000–120,000 cells were collected and resuspended in 50 µl of PBS + 1% BSA. Cells were then blocked with 5 µl TruStain FX for 10 min. Next, cells were stained with 0.5 µl of each TotalSeq-B antibody (CD34, CD38, CD45RA, CD90, CD49f, CD35, CD11a, CD59 and CD117) for 30 min. Following 4 washes with PBS + 1% BSA, 10,000 cells were loaded onto a Chromium Next GEM Chip G. GEM generation and barcoding, post GEM–RT cleanup and cDNA amplification, 3′ gene-expression library construction, and cell surface protein library construction were performed as described at https://support.10xgenomics.com/single-cell-gene-expression/index/doc/user-guide-chromium-single-cell-3-reagent-kits-user-guide-v31-chemistry-dual-index-with-feature-barcoding-technology-for-cell-surface-protein. Gene-expression and cell surface protein libraries were pooled together at a ratio of 4:1 and sequenced on an Illumina NovaSeq S4 flowcell (Supplementary Table 28).

Computational analysis of scRNA-seq sequencing data

The BCL files were demultiplexed using eight base pair 10X sample indexes and cellranger mkfastq to generate paired-end FASTQ. We ran cellranger count to align the reads to the hg38 reference genome from GenBank using STAR[72] aligner as well as perform filtering, barcode counting, and UMI counting. The alignment results were used to quantify the expression level of human genes and generation of gene-barcode matrix.

Each sample's cellranger matrix was then loaded in a SeuratObject_4.1.0 using Seurat[68] (version 4.1.1, https://github.com/satijalab/seurat). Low-quality cells, doublets and potential dead cells were removed according to the percentage of mitochondrial genes and number of genes and UMIs expressed in each cell (nFeature_RNA > 200 & nFeature_RNA < 10000 & nCount_RNA > 2500 & percent.mt < 10). Clean count matrices from each sample were then combined using Seurat's merge function. The merged gene-expression data was normalized using sctransform based normalization while removing confounding variables, percentage of mitochondrial genes and sample origin. Then, cell cycle scores were assigned using Seurat's CellCycleScoring function. The difference between the G2M and S phase scores was then calculated and regressed out using sctransform based normalization to minimize differences due to differences in cell cycle phase among proliferating cells. The cell surface feature output was normalized using centred log-ratio (CLR) normalization, computed independently for each feature.

The four datasets were integrated using Harmony (https://github.com/immunogenomics/harmony) on sctransform normalized gene counts to group cells by cell type while correcting for sample origin. Dimensionality reduction via PCA and UMAP embedding was performed on the integrated dataset. Identities of the cell clusters were determined using canonical RNA cell type markers and cell surface feature expression patterns. HSC/MPP clusters were identified by staining positively for CD34 and CD49f, and negatively for CD38, CD45RA and CD11a. The common myeloid progenitor cluster was identified by staining positively for CD34 and CD38, and negatively for CD45RA and CD49f. The granulocyte macrophage progenitor cluster was identified by staining positively for CD34, CD38 and CD45RA, and negatively for CD49f. The difference between the proportion of cells in HSC/MPP1–4 clusters between control-eGFP and TCL1A-eGFP transduced cells was calculated by a proportion test using the Single Cell Proportion Test R package (https://github.com/rpolicastro/scProportionTest). To reconstruct the pseudotime trajectory of the HSC/MPP and CMP clusters, Monocle 3 pseudotime analysis was performed using the central node of the HSC/MPP1 Cluster as the root node (https://satijalab.org/signac/articles/monocle.html). Differential gene-expression analysis of TCL1A-eGFP versus control-eGFP HSC/MPPs was performed using the FindMarkers function in Seurat with the LR test and rs2887399 genotype as the latent variable, and with min.pct = 0.05 and logfc.threshold = 0.1. Differential gene-expression analysis of HSC/MPP4 versus HSC/MPP1 was performed using the FindMarkers function in Seurat with no thresholds for min.pct or logfc.threshold. Gene-set enrichment analysis (GSEA) was performed using the fgsea package (https://github.com/ctlab/fgsea) and the REACTOME gene sets using the following parameters for the fgsea function: nperm = 1000, scoreType = "std", minSize = 5. Results of differential expression analysis and GSEA can be found in Supplementary Tables 17 and 18.

Reporting summary

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

Data availability

Individual whole-genome sequence data for TOPMed whole genomes, individual-level harmonized phenotypes and the CHIP variant call sets used in this analysis are available through restricted access via the dbGaP TOPMed Exchange Area available to TOPMed investigators. Controlled-access release to the general scientific community via dbGaP is ongoing. dbGaP accession numbers are included in the Supplementary Tables 19 and 20. GWAS summary statistics have been deposited to dbGaP at accession phs001974. CHIP amplicon sequencing data from WHI have been deposited in dbGaP (parent study phs000200.v12.p3 and sub-study phs003206.v1). Data from scRNA-seq and ATAC-seq generated for this study are deposited under Gene Expression Omnibus (GEO) accession GSE205637. Source data are provided with this paper.

Code availability

Code developed for this study is available at: Re-analysis of Fabre et al.[16] data https://github.com/weinstockj/longitudinal_clonal_expansion_analysis; Simulation of mutation counts in HSCs https://github.com/weinstockj/hsc_simulation; Rust binary used to call U2AF1 mutations https://github.com/weinstockj/pileup_region; Passenger count variant calling pipeline https://github.com/weinstockj/passenger_count_variant_calling; Analyses using PACER estimates https://github.com/weinstockj/PACER_analyses; Analysis code for TCL1A over-expression CITE-seq data https://github.com/jkgopa/HSC_TCL1A_overexpression_scRNAseq; Mutect2 WDL pipeline c https://dockstore.org/workflows/github.com/broadinstitute/gatk/mutect2:4.1.8.1?tab=info; Zenodo archives https://doi.org/10.5281/zenodo.7474678 and https://doi.org/10.5281/zenodo.7474719.

44. 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).
45. 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).
46. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
47. Voss, K., Gentry, J. & Van der Auwera, G. Full-stack genomics pipelining with GATK4 + WDL + Cromwell. *F1000 Research* https://doi.org/10.7490/f1000research.1114631.1 (2017).
48. Beauchamp, E. M. et al. ZBTB33 is mutated in clonal hematopoiesis and myelodysplastic syndromes and impacts RNA splicing. *Blood Cancer Discov.* **2**, 500–517 (2021).
49. Miller, C. A. et al. Failure to detect mutations in U2AF1 due to changes in the GRCh38 reference sequence. *J. Mol. Diagn.* **24**, 219–223 (2022).
50. Pedersen, B. S. & Quinlan, A. R. cyvcf2: fast, flexible variant analysis with Python. *Bioinformatics* **33**, 1867–1869 (2017).
51. Asuni, N. & Wilder, S. VariantKey: a reversible numerical representation of human genetic variants. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/473744v3 (2019).
52. Alexandrov, L. B. et al. Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).
53. *Stan Modeling Language Users Guide and Reference Manual* version 2.17 (Stan Development Team, 2020).
54. Stan Development Team. RStan: The R interface to Stan v.2.21.5. https://mc-stan.org/ (2020).
55. Bezanson, J., Edelman, A., Karpinski, S. & Shah, V. B. Julia: A fresh approach to numerical computing v.1.4 (2017).
56. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S* (Springer-Verlag, 2002).
57. 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).
58. mimips. https://github.com/kitzmanlab/mimips (2020).
59. Koboldt, D. C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
60. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
61. Uddin, M. M. et al. Longitudinal profiling of clonal hematopoiesis provides insight into clonal dynamics. *Immun. Ageing* **19**, 23 (2022).
62. Ma, C., Blackwell, T., Boehnke, M. & Scott, L. J. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genet. Epidemiol.* **37**, 539–550 (2013).
63. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. B* **82**, 1273–1300 (2020).
64. Li, Z. et al. Dynamic scan procedure for detecting rare-variant association regions in whole-genome sequencing studies. *Am. J. Hum. Genet.* **104**, 802–814 (2019).
65. Bates, D. et al. Matrix: Sparse and Dense Matrix Classes and Methods v.1.4-1 (2019).

66. R Core Team. R: A Language and Environment for Statistical Computing. http://www.R-project.org/ (R Foundation for Statistical Computing, 2020).
67. Gogarten, S. M. et al. Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics* **35**, 5346–5348 (2019).
68. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
69. Corces, M. R. et al. Omni-ATAC-seq: improved ATAC-seq protocol. *Protocol Exchange* https://doi.org/10.1038/protex.2017.096 (2017).
70. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
71. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
72. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

Acknowledgements WGS for the 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 quality control and general study coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). The authors thank the studies and participants who provided biological samples and data for TOPMed. The full study-specific acknowledgments are included in Supplementary Note 4. The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the US Department of Health and Human Services. The authors wish to acknowledge the contributions of the consortium working on the development of the NHLBI BioData Catalyst ecosystem. S.J. is supported by the Burroughs Wellcome Foundation Career Award for Medical Scientists, Foundation Leducq (TNE-18CVD04), Ludwig Center for Cancer Stem Cell Research, the American Society of Hematology Scholar Award, the NIH Director's New Innovator Award (DP2-HL157540), and a Leukemia and Lymphoma Society Discovery Grant. A.G.B. is supported by a Burroughs Wellcome Foundation Career Award for Medical Scientists, the NIH Director's Early Independence Award (DP5-OD029586), and the Pew-Stewart Scholar for Cancer Research award, supported by the Pew Charitable Trusts and the Alexander and Margaret Stewart Trust. WHI CHIP amplicon sequencing was supported by the NHLBI (R01 HL148565). The Fred Hutchinson Cooperative Center of Excellence in Hematology cell collection and processing is supported by NIDDK Grant # DK106829. The authors thank R. Majeti, T. Koehnke and B. Ebert for helpful discussions.

Author contributions J.S.W., A.G.B. and S.J. conceived of the study and conceived of PACER. J.S.W. performed somatic variant calling, developed the PACER implementation and performed the human genetic association analyses. J.G. and S.J. functionally characterized the *TCL1A* locus. J.G. performed mouse model experiments. J.S.W. and S.J. performed WHI validation analyses. J.S.W., J.G., A.G.B. and S.J. wrote the manuscript with input from all authors. B.B.B., M.M.U., N.K., J.A.B., H.B., B.D., Z.M., N.L., T.M.M., S.E.L., K.P.P., S.R.M. and A.T.S. performed additional bioinformatic analyses. J.O.K., E.A.W. and A.P.R. contributed WHI amplicon sequencing data. C.A.L., J.G.B., K.D.T., X.G., M.F.S., A.S.V., S.K., A.R.S., J.R.O., J.P.L., E.B., K.C.B., N.C., E.E.K., R.J.L., M.F., L.H., D.M.L., S.R., B.E.C., B.M.P., J.C.B., J.A.B., E.K.S., J.H.Y., D.Q., N.D.P., B.I.F., D.W.B., M.H.C., D.L.D., V.S.R., L.R.Y., L.C.B., S.K., P.A.P., J.H., M.R., P.V.H., R.K., S.R.H., N.L.S., K.L.W., D.K.A., M.R.I., H.T., M.J.C., S.K., J.B.M., A.C., L.M.R., Y.G., M.A., J.I.R., S.S.R., R.P.T., B.A.K., J.M.J., M.M.W., J.G.S., O.M., P.M.N., B.S.C., R.D., J.E.C., J.B., S.M., L.K.W., S.X., M.Y., C.C.G., Y.I.C., W.L., G.M.M., J.P.K., C.R.P., M.B.S., D.D., D.R., C.A., C.K., Y.Z., J.E.M., P.K., A.D.J., R.A.M., T.W.B., G.R.A., A.V.S., H.M.K., P.N., E.A.W., A.P.R. and the NHLBI TOPMed Consortium contributed to sequencing and phenotyping of the included NHLBI TOPMed cohorts. T.W.B., G.R.A., A.V.S. and H.M.K. contributed TOPMed computing infrastructure and bioinformatics advice.

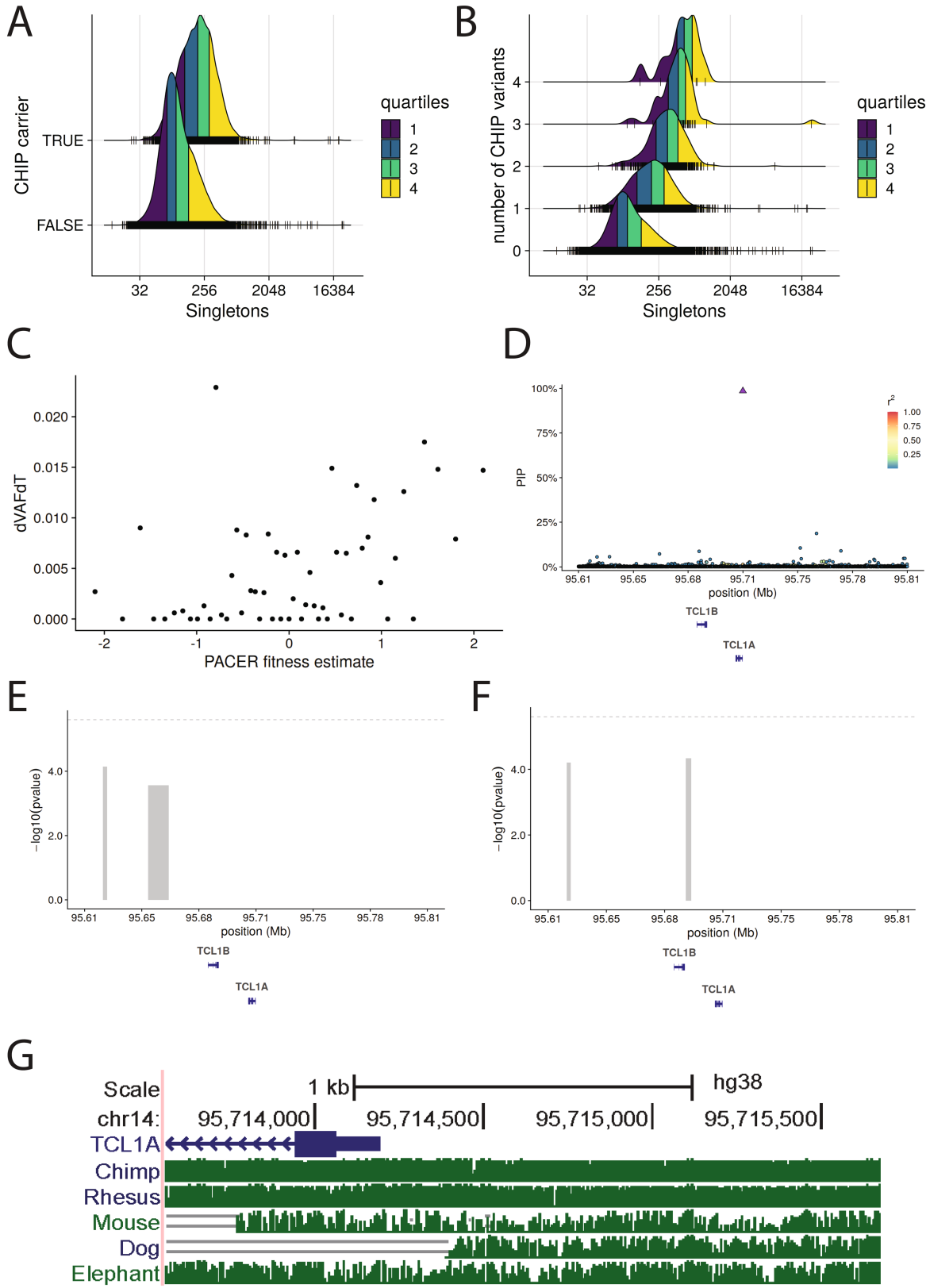
Competing interests S.J. is on advisory boards for Novartis, AVRO Bio, and Roche Genentech, reports speaking fees and a honorarium from GSK, and is on the scientific advisory board of Bitterroot Bio. P.N. reports grants support from Amgen, AstraZeneca, Apple, Novartis and Boston Scientific, is a paid consultant for Apple, AstraZeneca, Novartis, Genentech, Blackstone Life Sciences and spousal employment at Vertex, all unrelated to the present work. S.J., A.G.B. and P.N. are paid consultants for Foresite Labs and co-founders, equity holders, and on the scientific advisory board of TenSixteen Bio. Stanford University has filed a patent application for the use of PACER to identify therapeutic targets on which S.J., A.G.B. and J.S.W. are inventors (US patent 63/141,333). The patent has been licensed to TenSixteen Bio. B.M.P. serves on the Steering Committee of the Yale Open Data Access Project funded by Johnson & Johnson. E.K.S. reports grant support from Bayer and GSK. J.H.Y. reports consulting fees from Bridgebio Therapeutics. M.H.C. reports grant support from Bayer and GSK, and consulting and speaking fees from Illumina and AstraZeneca. D.L.D. reports grant support from Bayer. L.M.R. is a consultant for the TOPMed Administrative Coordinating Center (through Westat). P.D. reports grant support from Janssen Research. H.M.K., G.R.A. and A.R.S. are employees of Regeneron Pharmaceuticals and recieve salary, stock and stock options as compensation. A.T.S. is a founder of Immunai and Cartography Biosciences and receives research funding from Allogene Therapeutics and Merck Research Laboratories. J.O.K. is on the scientific advisory board of MyOme Inc.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-05806-1.

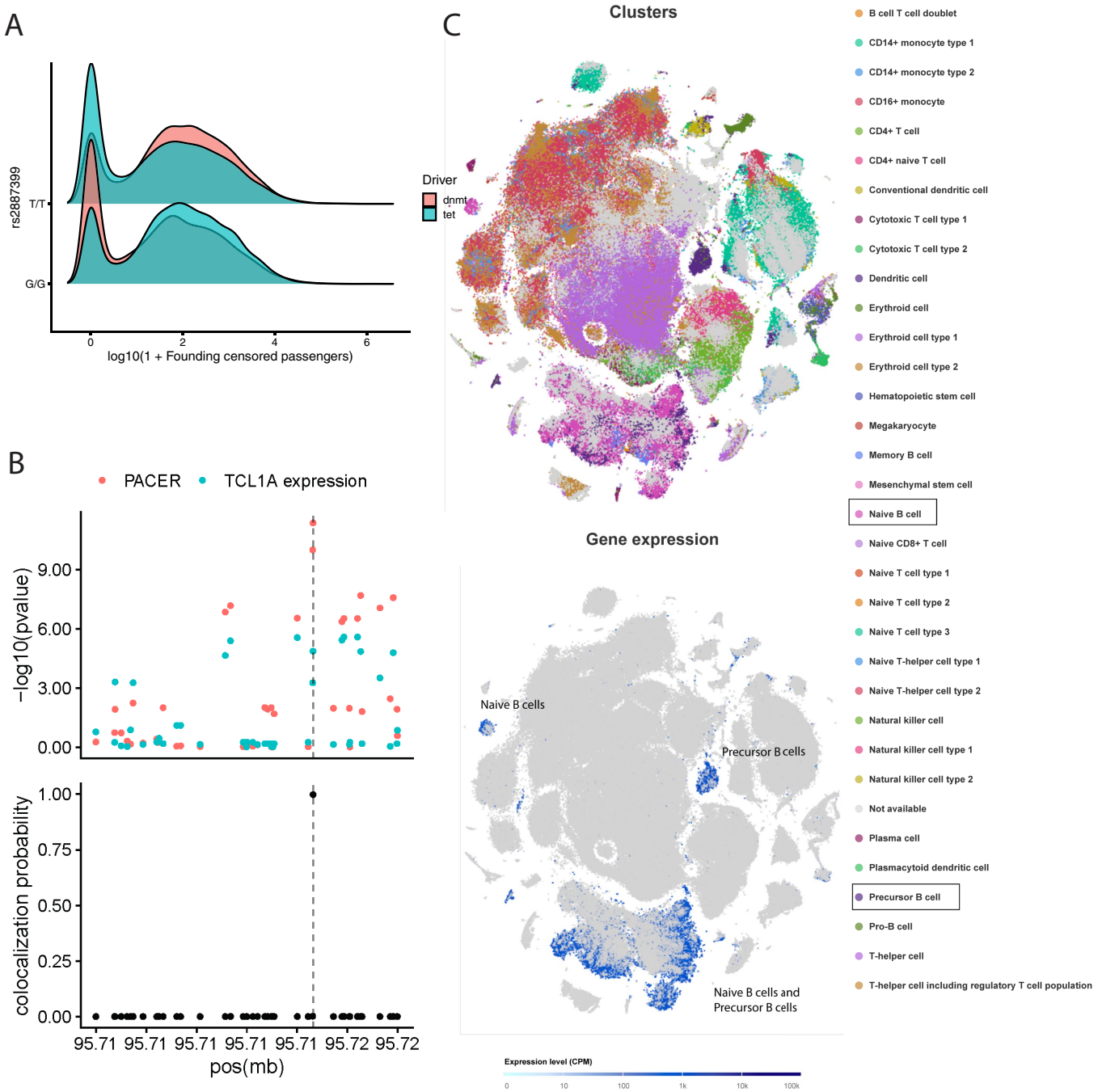
Correspondence and requests for materials should be addressed to Alexander G. Bick or Siddhartha Jaiswal.

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

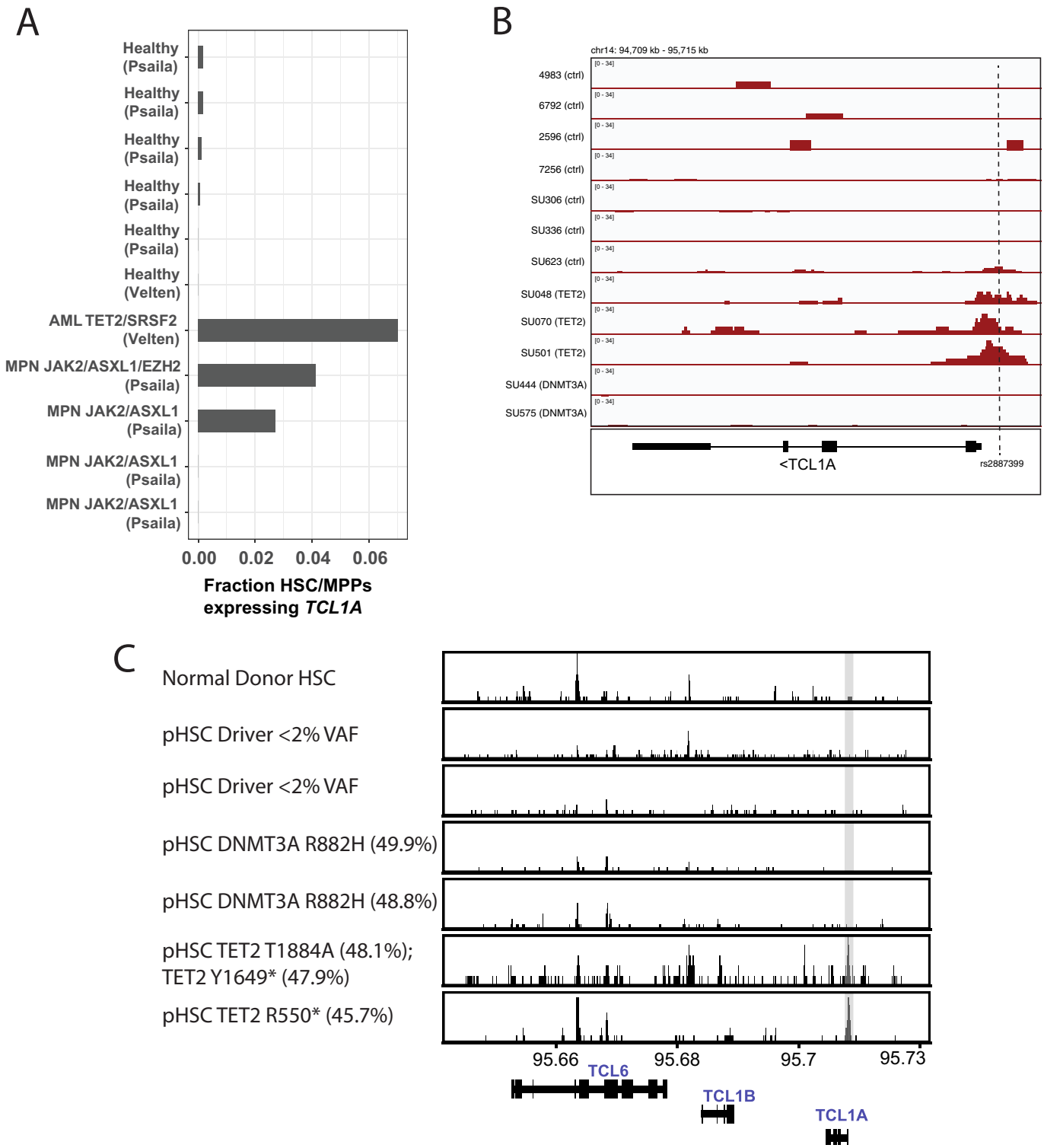


Extended Data Fig. 1 | See next page for caption.

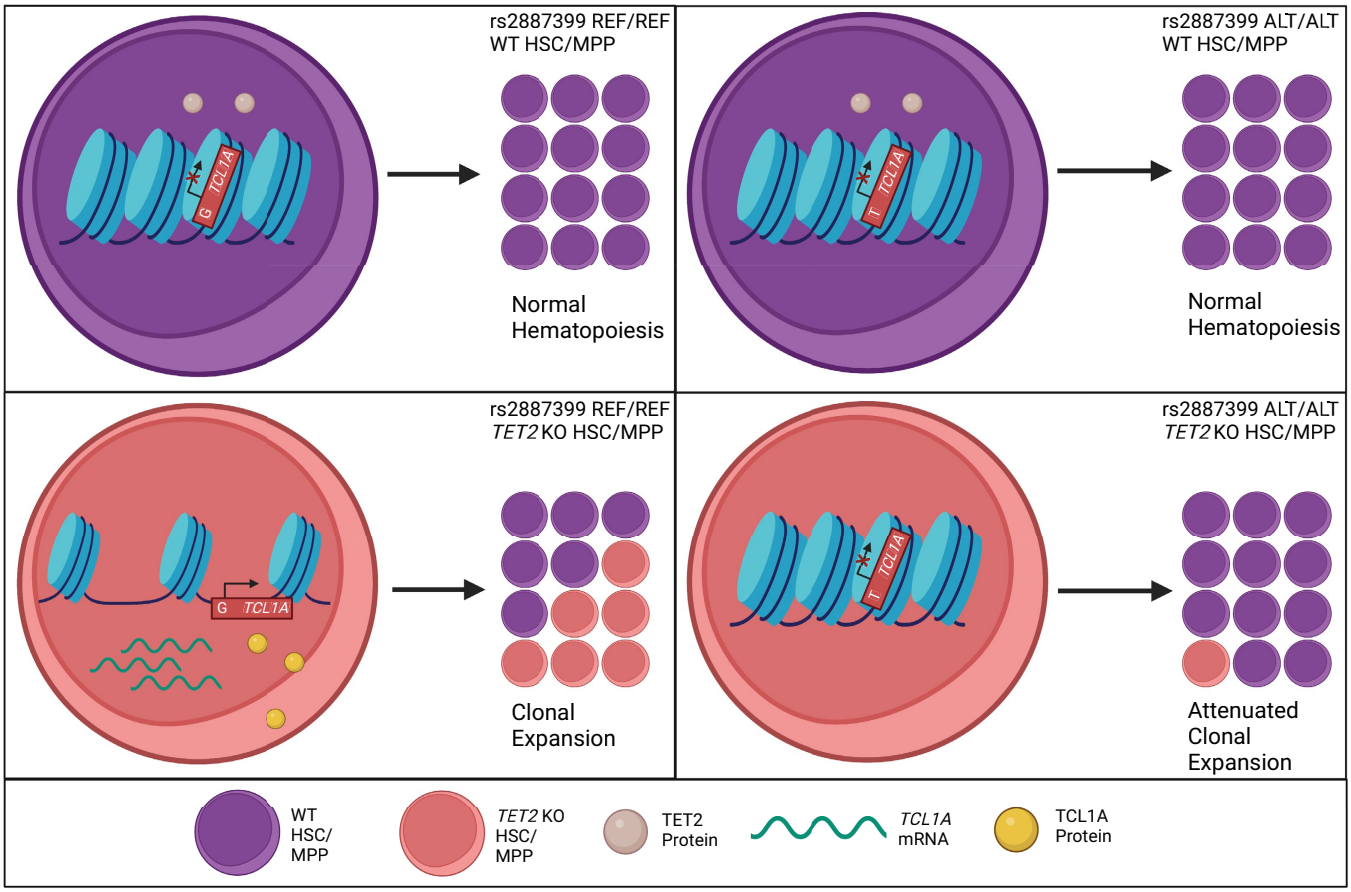
Extended Data Fig. 1 | PACER Estimates Clonal Expansion Rate. A. The passenger counts are enriched by 54% (95% CI: 51%-57%) after adjusting for age and study using a negative binomial regression. The different colors in the density plots correspond to quartiles of the marginal probability distributions. As the density estimates are smoothed, the underlying data points are indicated with hash marks. **B**. The distributions of passenger counts are stratified by the number of CHIP driver variants acquired. The different colors in the density plots correspond to quartiles of the marginal probability distributions. **C**. The observed clonal expansion rates (dVAFdT), as expressed in the change in variant allele frequency (VAF) over time (years), were associated with increased PACER fitness estimates in 55 CHIP carriers from the Women's Health Initiative. The PACER fitness estimates have been inverse normal transformed. **D**. The posterior inclusion probabilities (PIP) as estimated by SuSIE[63] are plotted on the y-axis, and the genomic position of a 0.8 Mb region including TCL1A is plotted on the x-axis. The linkage disequilibrium (LD) estimates are plotted on a color scale and are estimated on the genotypes used for association analyses. **E**. Rare variant analyses were performed using the SCANG[46] rare variant scan procedure including all variants with a minor allele count less than 300. Identified rare variant windows are plotted as gray rectangles where the width corresponds to the size of the genomic region and the height corresponds to the pvalue of the SCANG[64] test statistic for the window. **F**. Rare variant analyses were performed including the rs2887399 genotypes as covariate. Hypothesis testing was performed using the SCANG rare variant scan procedure. **G**. Multiz alignments across multiple species are shown for the TCL1A locus.



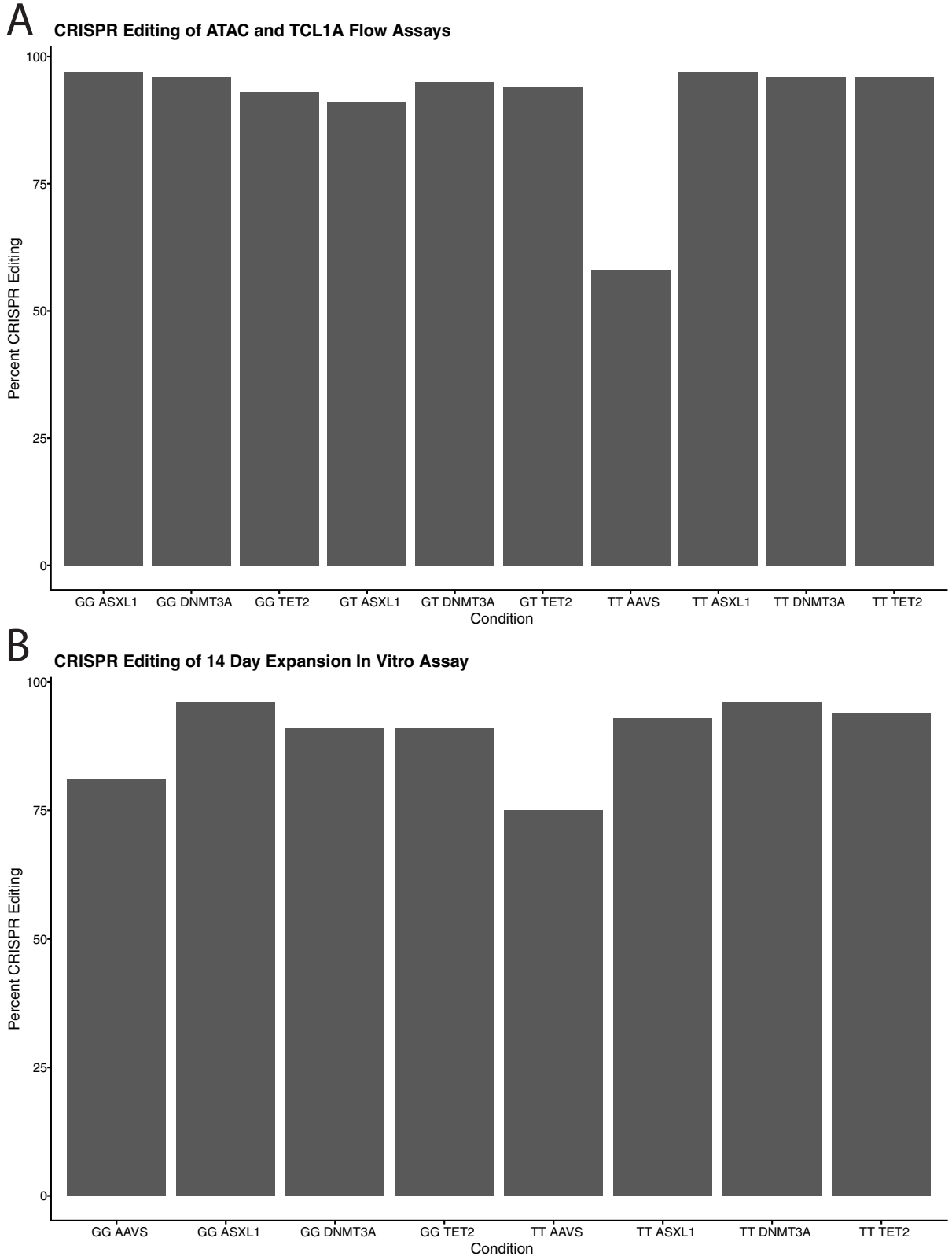
Extended Data Fig. 2 | GWAS Implicates rs2887399 as a Modifier of Clonal Expansion Rate. A. The distributions of the four conditions – *DNMT3A* and *TET2* mutant clones stratified by homozygous genotype of rs2887399. The y-axis indicates the density of the distributions and the x-axis indicates the log10 founding censored passengers, which are the simulated equivalent to the singleton mutations observed in the real data analysis. Simulated *DNMT3A* mutations out-compete *TET2* when rs2887399 is set to the protective T/T allele even though its fitness is unchanged by rs2887399. **B**. The top panel includes the -log10 pvalues from both the PACER GWAS and TCL1A cis-eQTLs in whole blood from GTEx v8[29]. The GWAS p-values are estimated with SAIGE. In the bottom panel, posterior probability of colocalization from COLOC[30] identifies rs2887399 as the likely shared causal variant. **C**. UMAP plot of scRNA-seq data from immune cells in the Human Cell Atlas[31]. TCL1A expression is highlighted on the bottom plot. UMAP plot was generated in the EMBL-EBI Single Cell Expression Atlas.



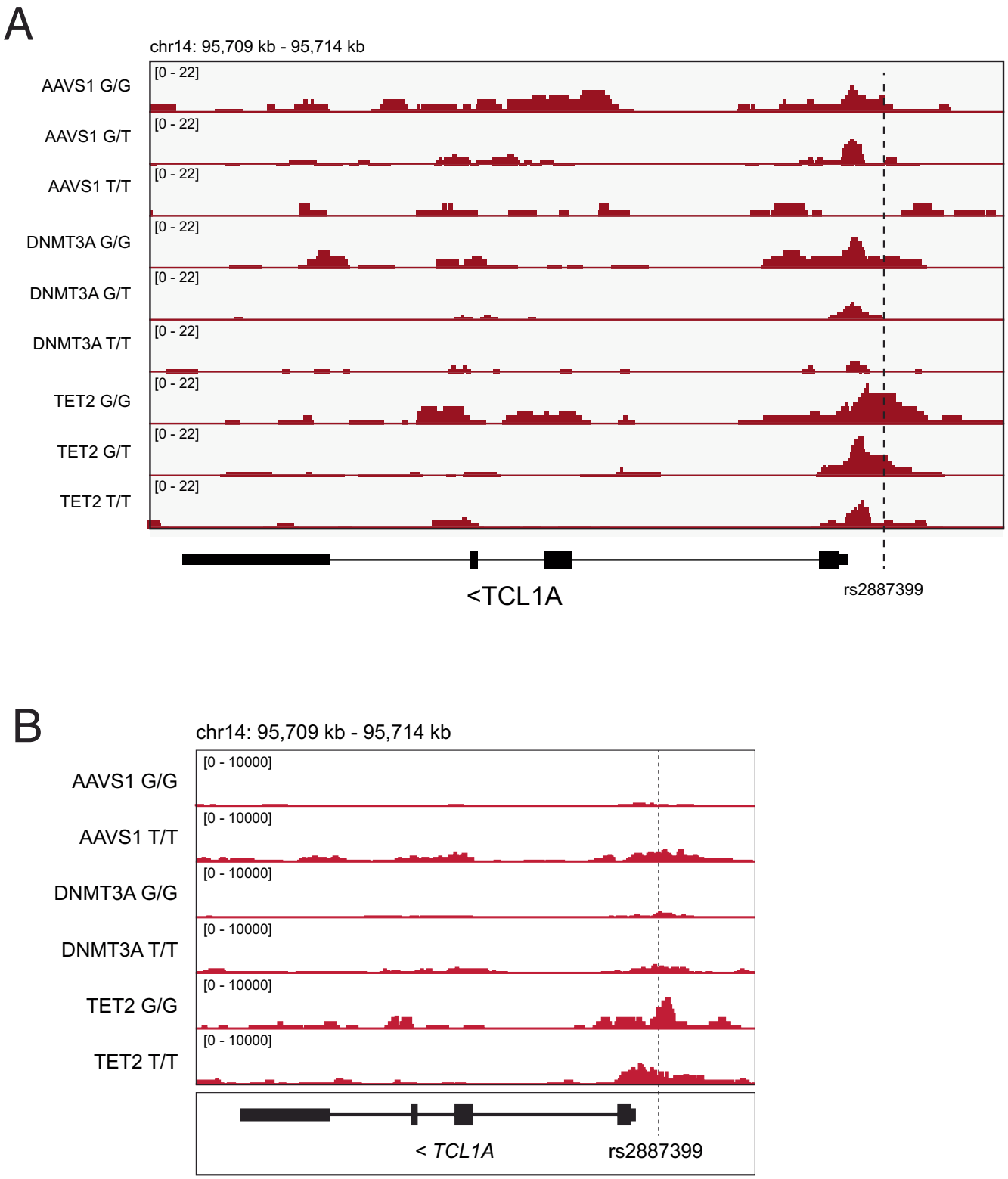
Extended Data Fig. 3 | Chromatin Accessibility and Transcript Expression of TCL1A. A. Quantification of fraction of HSC/MPPs expressing *TCL1A* transcripts in patients with *TET2* or *ASXL1* driven acute myeloid leukemia (AML) or myeloproliferative neoplasm (MPN) compared to healthy donors. Data is from single-cell RNA sequencing generated in Psaila[33] et al. and Velten[32] et al. **B**. ATAC-sequencing tracks of the *TCL1A* locus near rs2887399 in HSCs from healthy donors (row 1-4), pre-leukemic hematopoietic stem cells (pHSCs) from patients with AML but no detected driver mutations (rows 5-7), in pHSCs with *TET2* mutations (rows 8-10), and pHSCs with *DNMT3A* mutations (rows 11-12). Data is from Corces et al.[34]. Vertical dashed line indicates location of the rs2887399 SNP. **C**. ATAC-sequencing tracks of the *TCL6-TCL1A* locus in HSCs from healthy donors (row 1), pre-leukemic hematopoietic stem cells (pHSCs) from patients with AML but no detected driver mutations (rows 2-3), pHSCs with *DNMT3A* mutations (rows 4-5), and in pHSCs with *TET2* mutations (rows 6-7). Amino acid change and variant allele fraction (VAF) for the driver mutations are shown. Data is from Corces et al.[34].



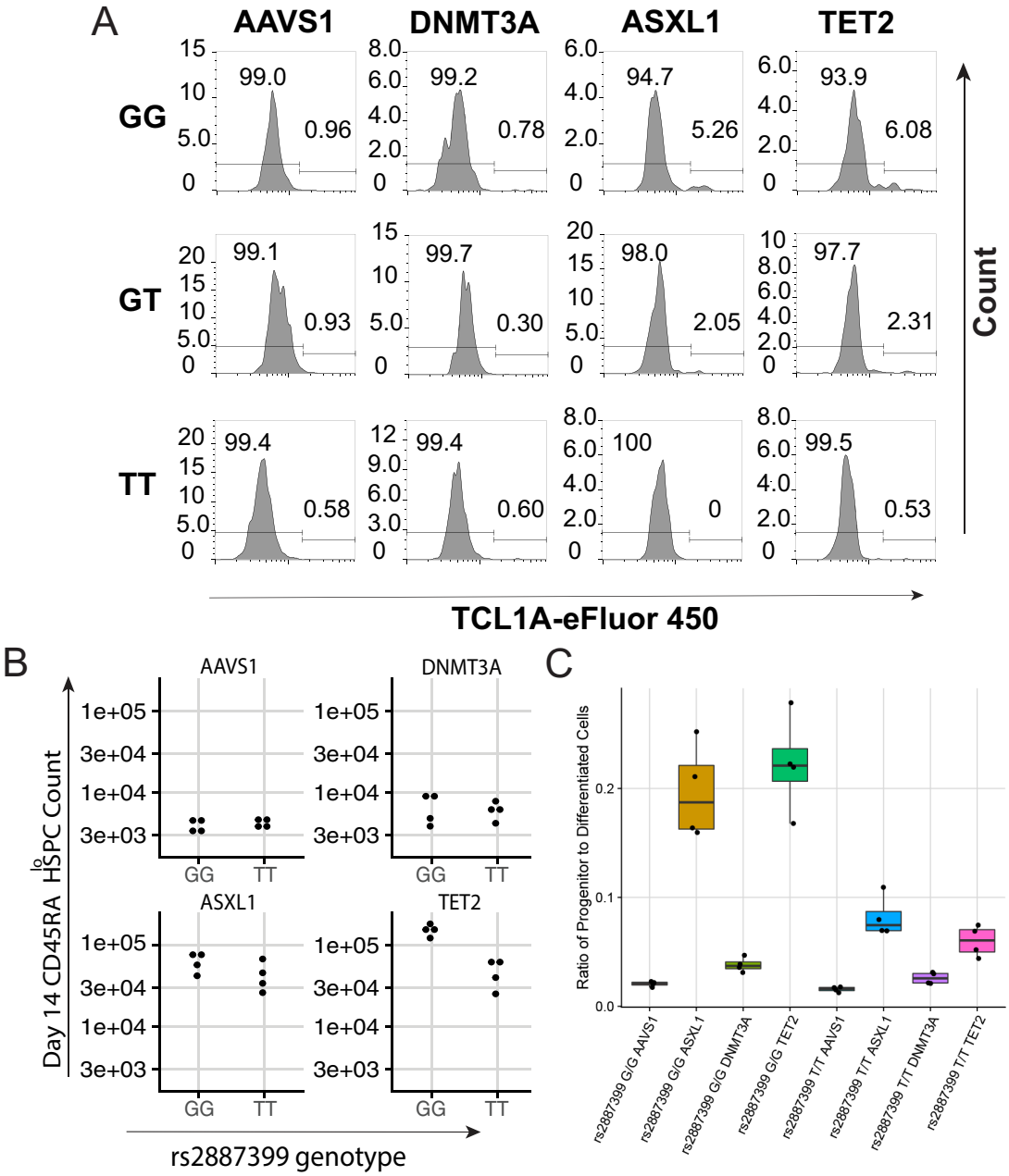
Extended Data Fig. 4 | Schematic of rs2887399 Effect on TET2 Clonal Expansion. Proposed model for clonal advantage due to mutations in *TET2*. In cells with the rs2887399 REF/REF genotype, loss of *TET2* function leads to an accessible *TCL1A* locus, aberrant *TCL1A* RNA and protein expression in hematopoietic stem cells (HSC's) and multi-potent progenitors (MPP's), and subsequent clonal expansion. The presence of rs2887399 ALT alleles diminishes the *TET2* clonal expansion phenotype by limiting *TCL1A* locus accessibility and downstream protein expression. Figure created with BioRender under a paid license.



Extended Data Fig. 5 | CRISPR Editing Efficiency. A. ICE analysis of Sanger traces to determine targeted CRISPR editing efficiency. Bar plots display percent of CD34+ CD38− CD45RA− cells with indel formation in gene of interest. These cells were used for the OMNI-ATAC and intracellular TCL1A flow assays. **B**. ICE analysis of Sanger traces to determine targeted CRISPR editing efficiency. Bar plots display percent of CD34+ CD38− CD45RA− cells with indel formation in gene of interest. These cells were used for the 14-day expansion assay.



Extended Data Fig. 6 | ATAC Sequencing Tracks of TCL1A. A. ATAC-sequencing tracks illustrating chromatin accessibility at rs2887399 in *TET2* or *DNMT3A*-edited HSC/MPPs cultured for 5 days from donors of the GG, GT, and TT genotypes. Red line indicates location of rs2887399. *TET2* edited samples are the same as in Fig. 4, shown here for comparison. **B**. ATAC-sequencing tracks illustrating chromatin accessibility at rs2887399 in AAVS, *TET2* or *DNMT3A*-edited HSC/MPPs cultured for 7 days from donors of the GG and TT genotypes, and then sorted for CD34hi CD38− CD45RA− Lin− cells prior to nuclei preparation. Red line indicates location of rs2887399.

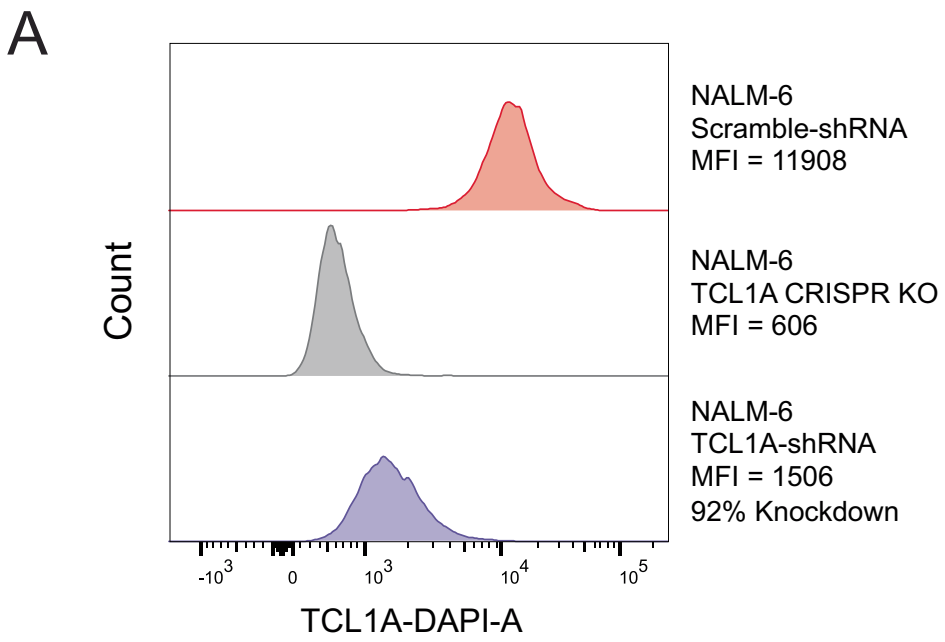


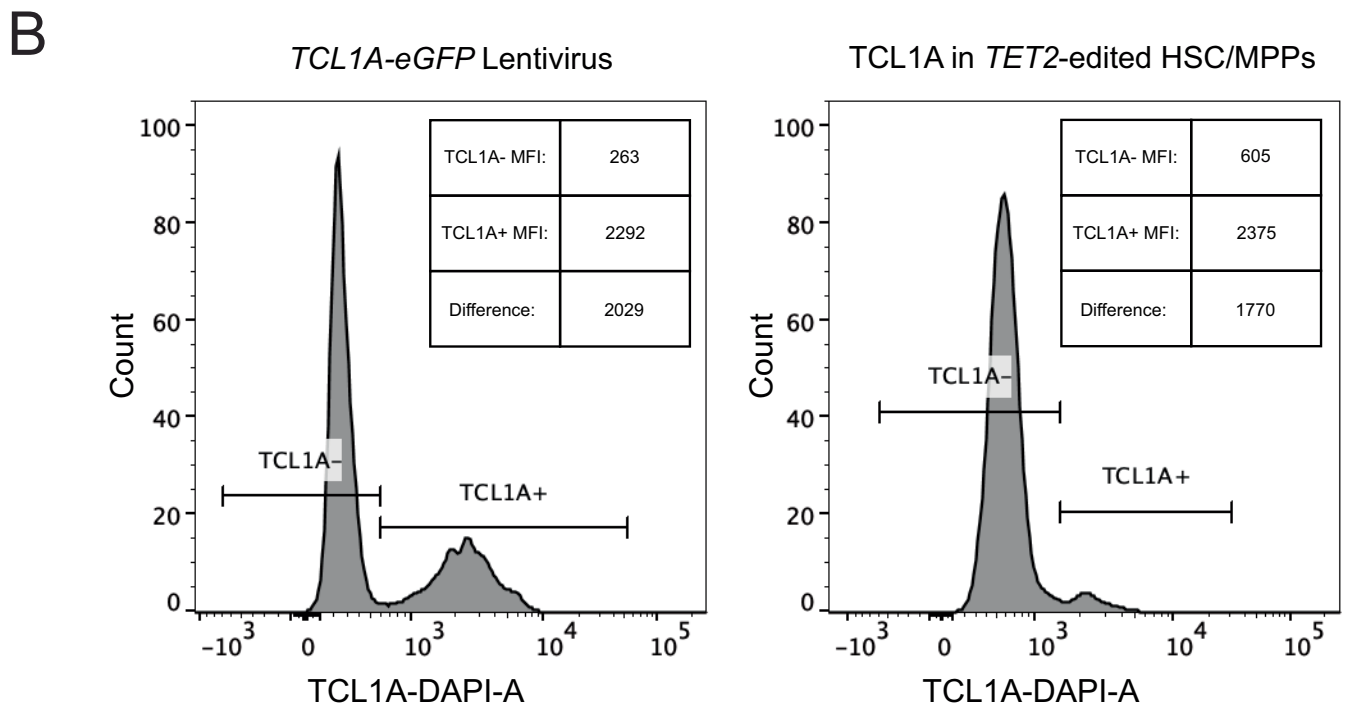
Variable	Est.	S.E.	p. val.
AAVS	(referent)		
DNMT3A	2695	9923	0.79
ASXL1	58573	9923	4.3×10^{-6}
TET2	148772	9923	1.1×10^{-13}
rs2887399 G/G	(referent)		
rs2887399 T/T	277	9923	0.98
DNMT3A:T/T	-849	14033	0.95
ASXL1:T/T	-19771	14033	0.17
TET2:T/T	-106080	14033	8.5×10^{-8}

Variable	Est.	S.E.	p. val
AAVS	(referent)		
ASXL1	0.176	0.017	2.37×10^{-10}
DNMT3A	0.018	0.017	0.313
TET2	0.202	0.017	1.53×10^{-11}
rs2887399 G/G	(referent)		
rs2887399 T/T	-0.005	0.017	0.774
ASXL1:T/T	-0.110	0.024	1.23×10^{-4}
DNMT3A:T/T	-0.007	0.024	0.765
TET2:T/T	-0.157	0.024	8.72×10^{-7}

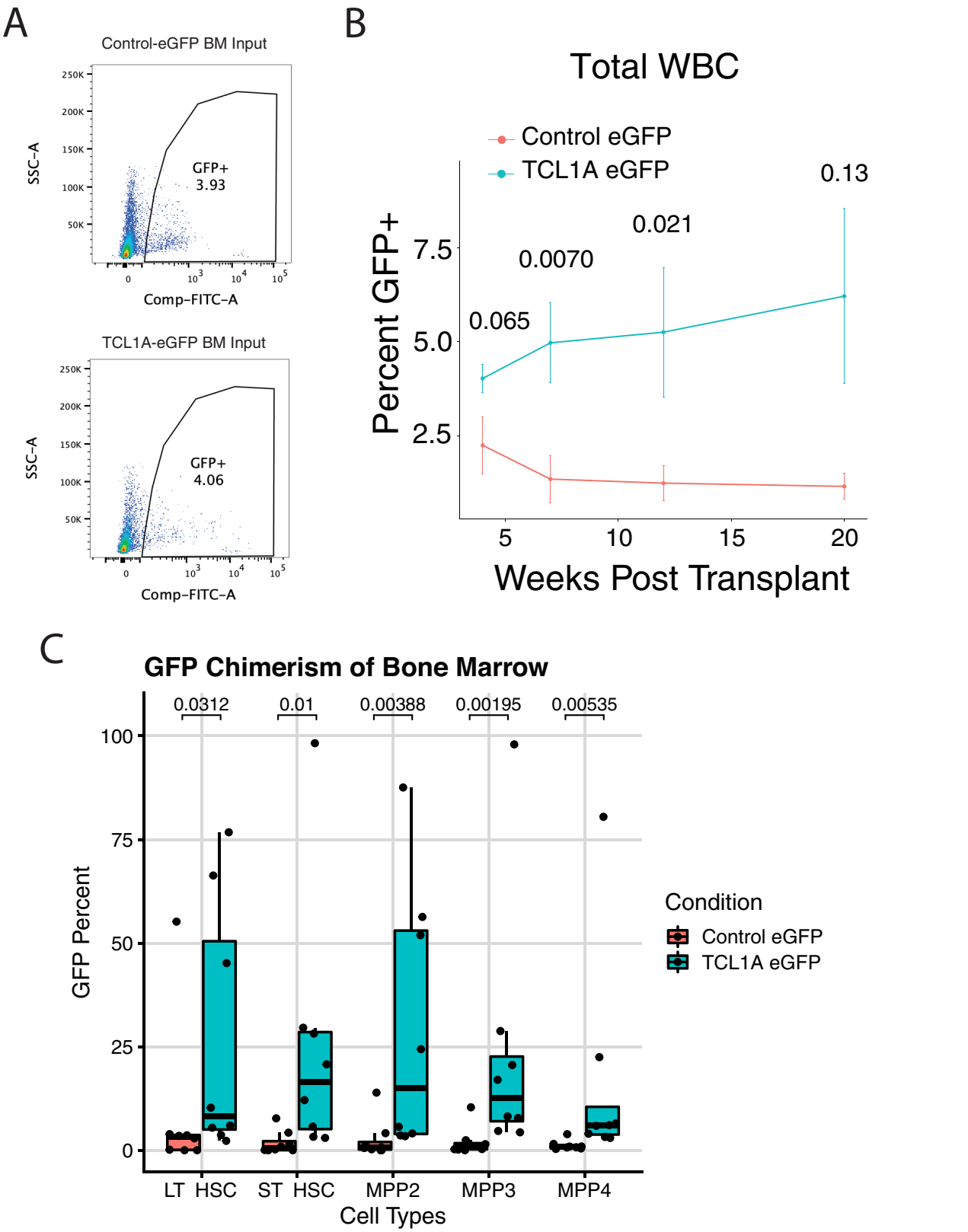
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Interaction of CHIP Mutations and rs2887399 in human HSPC phenotypes. A. Representative intracellular flow plots of TCL1A protein expression in edited HSC/MPPs from each rs2887399 donor after 11 days in culture. **B**. Quantification of Lin−/lo CD34+ CD38− CD45RAlo HSPCs (CD45RAlo HSPCs) after 14 days of *in vitro* expansion stratified by edited gene and rs2887399 genotype. Results of a linear regression model for the effect of edited gene (referent to AAVS1), rs2887399 genotype (referent to GG), and the interaction term of edited gene with rs2887399 genotype are presented below. Unadjusted p-values from two-sided tests are reported. n = 4 for each group. **C**. Ratio of CD34+CD45RA- cells to CD34- cells after 14 days of in vitro expansion stratified by edited gene and rs2887399 genotype. Results of a linear regression model for the effect of edited gene (referent to AAVS1), rs2887399 genotype (referent to GG), and the interaction term of edited gene with rs2887399 genotype are presented below. The horizontal line in each box indicates the median, the tops and bottoms of the boxes indicate the interquartile range, and the top and bottom error bars indicate maxima and minima, respectively. Unadjusted p-values from two-sided tests are reported. n = 4 for each group.

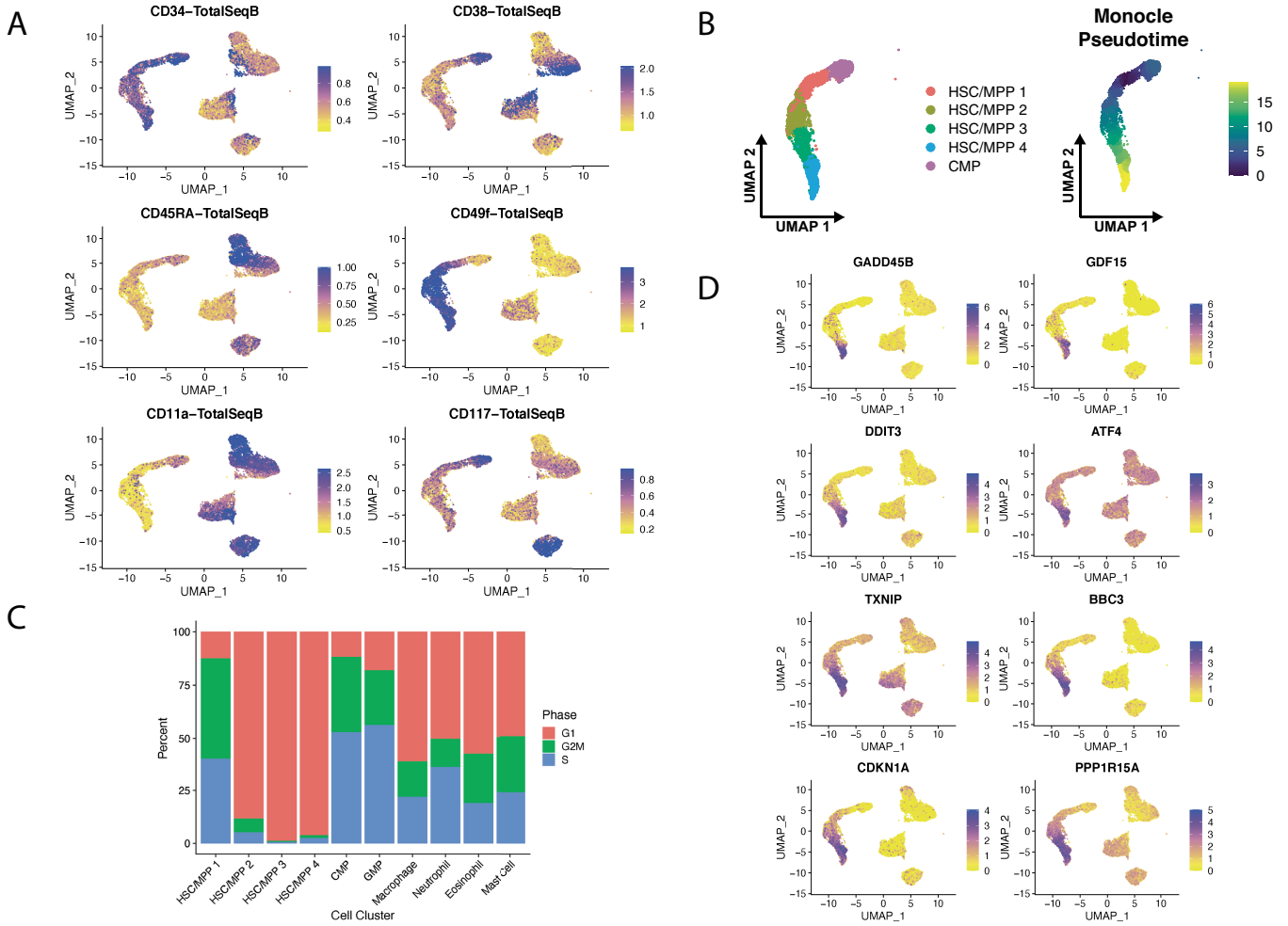




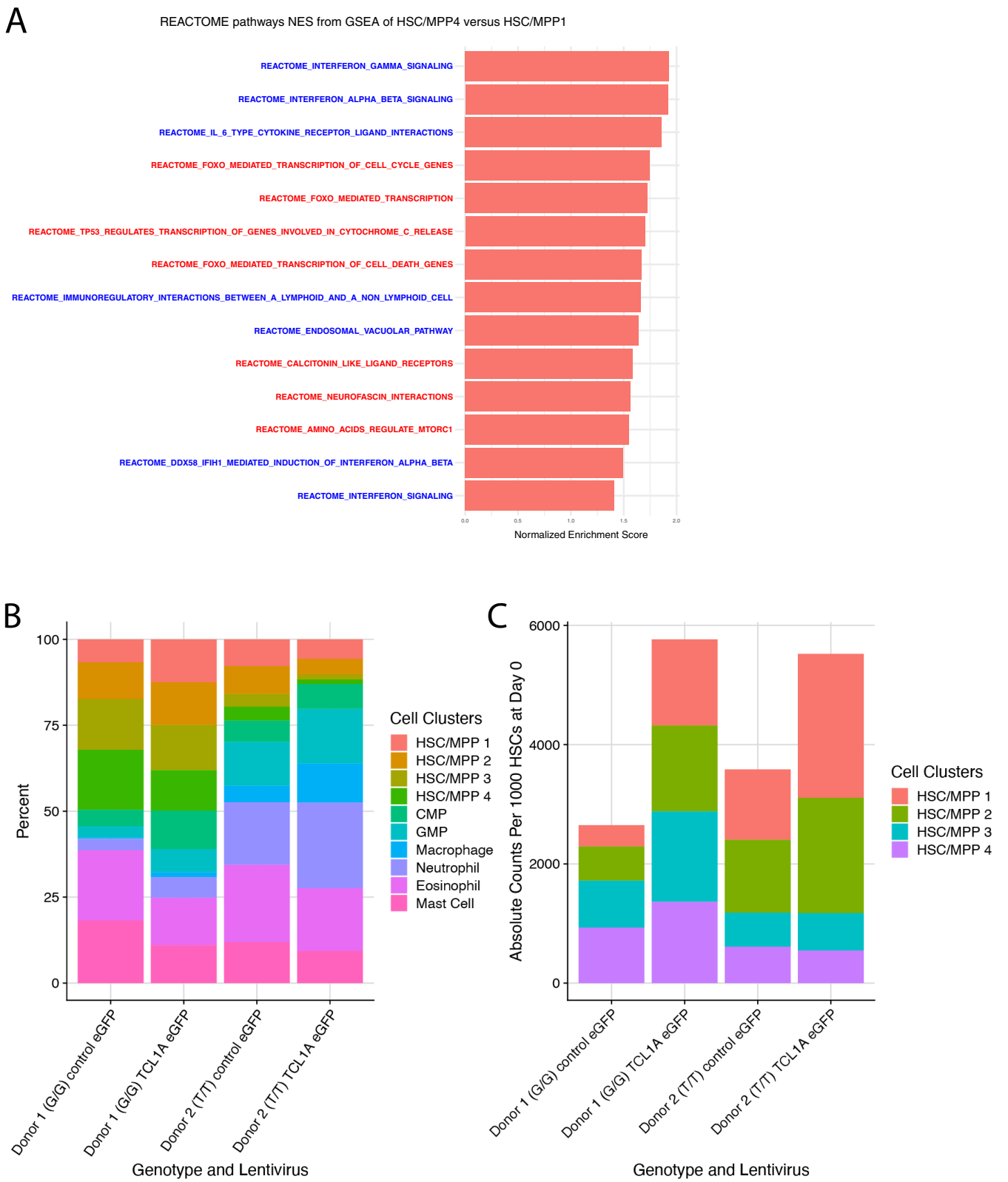
Extended Data Fig. 8 | Validation of TCL1A shRNA and Expression Lentivirus. A. Histogram of TCL1A-DAPI in wild-type, *TCL1A* CRISPR knockout, and *TCL1A* shRNA knockdown in NALM-6 cell line. **B**. Histogram of TCL1A-DAPI in human HSC/MPPs transduced with *TCL1A-eGFP* lentivirus or *TET2*-edited HSC/MPPs. MFI = geometric mean fluorescence intensity.



Extended Data Fig. 9 | TCL1A Expression Promotes HSC Fitness in Mice. **A.** Post-hoc analysis of percent GFP+ cells in the lineage negative fraction of the input cell mixture used for transplant. **B.** GFP+ chimerism over 20 weeks post-transplant as a fraction of total donor white blood cells. Shown are mean percent GFP+ cells and error bars represent standard errors for each time point. Hypothesis testing was performed with a two-sided Wilcoxon rank sum test and unadjusted p-values are shown above each time point. n = 8 for each group. **C.** Percent GFP+ cells in donor HSC/MPP subsets at 22 weeks post-transplant. The horizontal line in each box indicates the median, the tops and bottoms of the boxes indicate the interquartile range, and the top and bottom error bars indicate maxima and minima, respectively. Unadjusted p-values obtained from two-sided Wilcoxon rank sum tests are reported. n = 8 for each group.



Extended Data Fig. 10 | CITE-seq of TCL1A Expressing Human HSPCs. **A**. UMAP feature plots of Antibody Derived Tags (ADTs) for cell surface markers for HSPC identification. **B**. UMAP clustering of HSC/MPP populations colored by cell subtype clusters next to UMAP clustering of HSC/MPP populations colored by Monocle Pseudotime values. **C**. Stacked bar plot of percent of cells in each cell cycle phase as determined by Seurat cell cycle scoring module for each cell cluster. **D**. UMAP feature plot of select stress response and FOXO target genes.



Extended Data Fig. 11 | Effect of TCL1A Expression on Human HSC/MPP Phenotypes. A. Normalized enrichment scores (NES) of REACTOME pathways upregulated in HSC/MPP cluster 4 compared to HSC/MPP cluster 1 and filtered for those with FDR < 0.1 and NES > 1. Pathways printed in blue contain interferon response genes and pathways printed in red contain FOXO response genes. **B**. Stacked bar plot of all clusters in each analyzed sample dataset as a percentage of total cells in that sample. G/G or T/T refers to the genotype at rs2887399 in the donor. **C**. Stacked bar plot of absolute counts for each HSC/MPP cluster from each sample. Counts are shown as number of output cells at Day 7 per 1000 HSC/MPPs plated at Day 0.

nature portfolio

Corresponding author(s): Siddhartha Jaiswal, Alexander Bick

Last updated by author(s): 01/16/2023

Reporting Summary

Nature Portfolio 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 Portfolio 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

Provide a description of all commercial, open source and custom code used to collect the data in this study, specifying the version used OR state that no software was used.

Data analysis

R 4.1.2 and Python 3.7 were both used for analysis. In addition, several other open source packagers were used, linked here:
SAIGE v0.36.3.1 R package: https://github.com/weizhouUMICH/SAIGE
SCANG v1.0.3.1 R package: https://github.com/zilinli1988/SCANG
Mutect2 v4.1.8 : https://dockstore.org/workflows/github.com/broadinstitute/gatk/mutect2:4.1.8.1?tab=info
susieR R package v0.11.92 : https://stephenslab.github.io/susieR/index.html
RStan v2.21.5

Julia 1.4

Bespoke applications developed for analysis:
https://github.com/weinstockj/hsc_simulation
https://github.com/weinstockj/pileup_region
https://github.com/weinstockj/PACER_analyses (10.5281/zenodo.7474678)
https://github.com/weinstockj/passenger_count_variant_calling (10.5281/zenodo.7474719)

A full Renv lock file is available here: https://github.com/weinstockj/PACER_analyses/blob/master/renv.lock

FlowJo v10.8.1

R packages used for scRNAseq: scCATCH, Seurat, CytoTRACE, ggplot2, sctransform, tidyverse, RColorBrewer, gprofiler2, clusterProfiler, enrichplot, DOSE, scProportionTest, harmony, cowplot, data.table, RColorBrewer, colorspace, scater, TSCAN, SingleCellExperiment,

SeuratWrappers, monocle3, Matrix, ggplot2, patchwork, msigdbr, fgsea

Commented R Script for scRNAseq: https://github.com/jkgopa/HSC_TCL1A_overexpression_scRNAseq/blob/main/scRNAseq%20Final%20Analysis%20Commented.R

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 Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Individual whole-genome sequence data for TOPMed whole genomes, individual-level harmonized phenotypes and the CHIP variant call sets used in this analysis are available through restricted access via the dbGaP TOPMed Exchange Area available to TOPMed investigators. Controlled-access release to the general scientific community via dbGaP is ongoing. The accession numbers for the TOPMed WGS are as follows:
phs001543
phs000956
phs001211
phs001211
phs001435
phs001143
phs001644
phs001644
phs001644
phs001624
phs001624
phs001612
phs001612
phs001600
phs001600
phs001189
phs000954
phs000954
phs001368
phs001368
phs001368
phs000951
phs000951
phs000951
phs000951
phs000951
phs001546
phs001412
phs001472
phs001606
phs000946
phs000974
phs000974
phs000974
phs000974
phs000974
phs000974
phs000974
phs001218
phs001218
phs001218
phs001345
phs001345
phs001217
phs001725
phs001359
phs001395
phs000993
phs000993
phs001293
phs001545
phs000964
phs001598
phs001402

phs001416
phs001416
phs001416
phs001416
phs001416
phs001416
phs001416
phs001062
phs001062
phs001434
phs001515
phs001515
phs001515
phs001544
phs001024
phs001601
phs001601
phs001468
phs001468
phs001215
phs001215
phs000972
phs000972
phs001467
phs001387
phs001933
phs000997
phs000997
phs001032
phs001040
phs001237
phs001237
phs001237
phs001237

GWAS summary statistics are deposited to dbGaP at accession phs001974. Amplicon sequencing data from WHI is deposited in dbGaP (phs000200.v12.p3) and is mapped to GRCh38. Data from single-cell RNAseq and ATACseq generated for this study are deposited under Gene Expression Omnibus accession GSE205637 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE205637).

Previously published data used in this study are cellranger files from GEO accession GSE144568 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE144568), a seurat RDS file from https://doi.org/10.6084/m9.figshare.12382685.v1, ATAC-seq FASTQ files from GEO accession GSE74912 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE74912), and variant calling data from Fabre et al. Table S6 (https://static-content.springer.com/esm/art%3A10.1038%2Fs41586-022-04785-z/MediaObjects/41586_2022_4785_MOESM9_ESM.xlsx).

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	5,071 CHIP carriers were included based on convenience sampling. For mouse experiments, the sample size was estimated using a power calculation based on expected change in clonal expansion seen in prior experiments with mice having Tet2 knocked out.
Data exclusions	No data was excluded.
Replication	The experimental data in Figure 3 and 4 had sufficiently high n to perform statistical comparison. The CITE-seq analysis of TCL1A transduced HSCs was done on two donor samples, which each represent an independent biological replicate.
Randomization	Sample recruitment included no randomization mechanism, but relevant covariates are controlled for in all statistical analyses. Mice were randomized to receive either control or TCL1A-eGFP transduced cells, and were identical in age and sex for both groups.
Blinding	Investigators were blinded with respect to generation of whole genome sequencing data from TOPMed participants and generation of longitudinal sequencing data from WHI. Experimental data in Figures 3 and 4 were not blinded due lack of feasibility for the investigators conducting the 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

Antibodies

Antibodies used

e450-TCL1A ThermoFisher eBio1-21 48-6699-42
APC-CD34 BioLegend 561 343608
PE/Cy7-CD38 BioLegend HIT2 303515
FITC-CD90 BioLegend 5E10 562556
BV605-CD45RA BioLegend HI100 304134
PE/Cy5-CD2 BioLegend RPA-2.10 300210
PE/Cy5-CD3 BioLegend HIT3a 300310
PE/Cy5-CD4 BioLegend SK3 344654
PE/Cy5-CD8a BioLegend HIT8a 300910
PE/Cy5-CD16 BioLegend 3G8 302010
PE/Cy5-CD19 BioLegend HIB19 302210
PE/Cy5-CD20 BioLegend 2H7 302308
PE/Cy5-CD56 BioLegend 5.1H11 362516
PE/Cy5-CD235a BioLegend HIR2 306606
PE/Cy5-CD14 BioLegend M5E2 301864
Fixable Viability Stain 700 BD 564997
APC-CD45.1 BioLegend A20 110714
BV605-CD45.2 BioLegend 104 109841
PE-CD3 BioLegend 17A2 100206
PE/Cy5-CD19 BioLegend HIB19 302210
BV421-CD115 BioLegend AFS98 135513
PE/Cy7-CD11b BioLegend M1/70 101216
APC/Fire 750-Ly-6G BioLegend 1A8 127651
APC-Lineage Cocktail R&D Systems FLC001A
PE/Cy5-CD117 BioLegend 2B8 105809
BV421-Sca1 BioLegend D7 108127
APC/Cy7-CD45.2 BioLegend 104 109823
PE-Flk2 BioLegend A2F10 135305
PE/Cy7-CD150 BioLegend TC15-12F12.2 115913
BV605-CD48 BioLegend HM48-1 103439

Validation

Biolegend Validation (https://www.biolegend.com/en-us/quality/product-development): "Clones of hybridomas are carefully selected based on a number of criteria including robust growth and efficient production of a single clone of antibody that is specific to the intended target. The best clones move on to applications testing. Antibody clones are tested in a variety of assays to see which applications they are suited for. As an example, clone 13A3-1 for phosphorylated STAT3 (Tyr705) demonstrated excellent performance in flow cytometry, western blot, and chromatin immunoprecipitation. Thus, the clone cross-validates itself by demonstrating functionality across orthogonal testing methods. Additionally, the biological induction of the phosphorylated state using IL-6 further validates the specificity of the antibody. Knockout or knockdown of gene expression, such as with siRNA, is also an excellent tool for target validation. Our SIRT5 antibody, clone O91G9, was verified by western blot using HeLa cells treated with SIRT5 targeting siRNA. Lane 1 indicates untreated HeLa cells, lane 2 contains scrambled siRNA control treated cells, and lane 3 contains SIRT5-specific siRNA treated cells."

R&D Validation (https://www.rndsystems.com/quality/antibodies-built-for-reproducibility): "R&D Systems® takes rigorous steps towards antibody validation and reproducibility. We have been since the beginning. For 30 years, we have used our industry-leading production standards and quality control specifications to develop antibodies that can be relied on for specificity and reproducibility. By developing and testing our products in-house, we can ensure a validated and specific antibody. We are confident in our antibodies and provide 100% guarantee for our products. With R&D Systems® antibodies your results will stand the test of time."

BD Validation (https://www.bdbiosciences.com/en-us/products/reagents/flow-cytometry-reagents/research-reagents/quality-and-reproducibility): "BD Biosciences identifies key targets of interest in scientific research and develops its own specific antibodies or collaborates with top research scientists around the world to license their antibodies. We then transform these antibodies into flow

cytometry reagents by conjugating them to a broad portfolio of high-performing dyes, including our vastly popular portfolio of BD Horizon Brilliant™ Dyes. A world-class team of research scientists helps ensure that these reagents work reliably and consistently for flow cytometry applications. The specificity is confirmed using multiple methodologies that may include a combination of flow cytometry, immunofluorescence, immunohistochemistry or western blot to test staining on a combination of primary cells, cell lines or transfectant models. All flow cytometry reagents are titrated on the relevant positive or negative cells. To save time and cell samples for researchers, test size reagents are bottled at an optimal concentration with the best signal-to-noise ratio on relevant models during the product development. To ensure consistent performance from lot-to-lot, each reagent is bottled to match the previous lot MFI."

Thermo Validation: This eBio1-21 antibody has been pre-diluted and tested by intracellular staining followed by flow cytometric analysis of Daudi cells using the Intracellular Fixation & Permeabilization Buffer Set (Product # 88-8824-00) and protocol. Please refer to "Staining Intracellular Antigens for Flow Cytometry, Protocol A: Two step protocol for intracellular (cytoplasmic) proteins" located at www.thermofisher.com/flowprotocols. This may be used at 5 µL (0.5 µg) per test. A test is defined as the amount (µg) of antibody that will stain a cell sample in a final volume of 100 µL. Cell number should be determined empirically but can range from 10^5 to 10^8 cells/test.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HEK293T cells from ATCC (CRL-3216) were used for virus generation. NALM6 cells were obtained from Ron Levy (Stanford) and were used to test shRNA constructs for TCL1A knockdown efficiency.
Authentication	Additional authentication of HEK293T or NALM6 cells was not performed after acquiring the cells.
Mycoplasma contamination	Mycoplasma testing of HEK293T cells was not performed. NALM6 cells were tested for mycoplasma prior to using for experiments.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	B6.SJL (Strain #:002014) from Jackson Laboratory were used as recipient transplant mice for murine bone marrow from C57BL/6J (Strain #:000664) from Jackson Laboratory. The mice were housed under a 12-h light/12-h dark cycle with dark hours from 18:30–06:30 and housed at 68–73°F under 40–60% humidity.
Wild animals	*Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	Mouse study protocol was approved by Stanford APLAC (APLAC protocol-32928).

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	In TOPMed, the median age was 55 years, the mean age 52.5, and the maximum age 98. The samples have diverse reported ethnicity (40% European, 32% African, 16% Hispanic/Latino, 10% Asian). For the Women's Health Initiative (WHI) longitudinal assessment of clonal hematopoiesis, all participants were women and the median age at baseline was 64 years.
Recruitment	In TOPMed, 49 cohorts were included, each with differing sample recruitment protocols. More information can be found at the dbGaP page for each of the constituent cohorts. See supplementary tables for more information. Women aged 50–79 years were enrolled from forty WHI clinical centers in the United States between 1993 and 1998.
Ethics oversight	Each of the 49 cohorts obtained IRB permission from their respective institutions and are supported by NHLBI. Each of the studies obtained informed consent from each participant. Blood samples from WHI participants were obtained with informed consent and the CHIP longitudinal assessment study was reviewed and approved by the Fred Hutchinson Cancer Center Institutional Review Board (IRB #10186). For the purchased CD34 cells, donors were recruited and consented, and mobilized peripheral blood was obtained using protocols approved by the Fred Hutchinson Cancer Center Institutional Review Board.

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

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Cells in suspension were harvested from plates, washed with PBS, stained with extracellular or intracellular antibodies, washed, and finally strained before acquistion
Instrument	Aria III
Software	FlowJo v10.8.1
Cell population abundance	500-1000 human HSC/MPPs were sorted per well for the liquid culture expansion assays as described in the methods for each experiment. Sorted cell purity >95% was confirmed by performing flow on an aliquot of sorted cells.
Gating strategy	Human HSC/MPP: FSC vs SSC, Doublet Exclusion, Live/Dead, Lineage Exclusion, CD34+, CD38-, CD45RA-; TCL1A+/-; DAPI DNA staining +/- Mouse HSC/MPP subsets: FSC vs SSC,Doublet Exclusion, Live/Dead, Lineage Exclusion, c-Kit+, Sca-1+, Flk2+/-, CD48+/-, CD150+/- Mouse granulocytes: FSC vs SSC, Doublet Exclusion, Live/Dead, CD3- CD19-, CD11b+ Ly6G+

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.