Published in final edited form as:

Nature. 2023 April ; 616(7958): 755–763. doi:10.1038/s41586-023-05806-1.

Aberrant activation of TCL1A promotes stem cell expansion in clonal hematopoiesis

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

Abstract

Mutations in a diverse set of driver genes increase fitness of hematopoietic stem cells (HSCs), leading to outgrowths termed 'clonal hematopoiesis' (CH)[1]. These lesions are precursors for blood cancers[2-6], but the reasons for their fitness advantage remain largely unknown, partially due to a paucity of cohorts where clonal expansion rate has been assessed by longitudinal sampling. To circumvent this limitation, we developed a method to infer expansion rate from single timepoint data called PACER (passenger-approximated clonal expansion rate) and applied it to 5,071 persons with CH. A genome-wide association study revealed that a common inherited polymorphism in the *TCL1A* promoter associated with slower expansion rate in CH overall, but the effect

Please address correspondence to: Siddhartha Jaiswal, MD PhD, Department of Pathology, Stanford, 240 Pasteur Dr Rm 4654, Stanford, CA 94304, Phone: 650-723-7211, sjaiswal@stanford.edu; Alexander G. Bick, MD PhD, Division of Genetic Medicine, Vanderbilt, 2200 Pierce Ave, RRB 550, Nashville TN 37232, Phone: 615-322-4153, alexander.bick@vumc.org.

[*]These authors contributed equally to this work

AUTHOR CONTRIBUTIONS

JSW, AGB, and SJ conceived of the study. JSW performed somatic variant calling. JSW, AGB, and SJ conceived of PACER. JSW developed the PACER implementation. JSW performed the GWAS, RVAS, fine-mapping, and colocalization analyses. JKG and SJ functionally characterized the *TCL1A* locus. JKG performed murine model experiments. JSW and SJ performed WHI validation analyses. JSW, JKG, AGB, and SJ wrote the manuscript with input from all authors. BBB, MMU, NK, JAB, HB, BD, ZM, NL, TMM, SEL, KPP, SRM, and ATS performed additional bioinformatic analyses. JK and APR contributed WHI amplicon sequencing data. CAL, JGB, KDT, XG, MFS, ASV, SK, ARS, JRO, JPL, EB, KCB, NC, EEK, RJL, MF, LH, DML, SR, BEC, BMP, JCB, JAB, EKS, JHY, DQ, NDP, BIF, DWB, MHC, DLD, VSR, LRY, LCB, SK, PAP, JH, MR, PVH, RK, SRH, NLS, KLW, DKA, MRI, HT, MJC, SK, JBM, AC, LMR, YG, MA, JIR, SSR, RPT, BAK, JMJ, MMW, JGS, OM, PMN, BSC, RD, JEC, JB, SM, LKW, SX, MY, CCG, YIC, WL, GMM, JPK, CRP, MBS, DD, DR, CA, CK, YZ, JEM, PK, ADJ, RAM, TWB, GRA, AVS, HMK, PN, EW, APR, and the NHLBI TOPMed Consortium contributed to sequencing and phenotyping of the included NHLBI TOPMed cohorts. TWB, GRA, AVS, and HMK contributed TOPMed computing infrastructure and bioinformatics advice.

ETHICS 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 #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.

COMPETING INTERESTS

S. Jaiswal is on advisory boards for Novartis, AVRO Bio, and Roche Genentech, is a paid consultant for Foresite Labs, reports speaking fees/honorarium from GSK, and is on the scientific advisory board to Bitterroot Bio. P. Natarajan reports grants support from Amgen, Apple, and Boston Scientific, and is a paid consultant to Apple and Foresite Labs. A. Bick is a paid consultant to Foresite Labs. G.R.A. is an employee of Regeneron Pharmaceuticals and owns stock and stock options for Regeneron Pharmaceuticals. S. Jaiswal, A. Bick, and P. Natarajan are 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. Jaiswal, A. Bick, and J. Weinstock are inventors (US patent 63/141,333). The patent has been licensed to TenSixteen Bio. Dr. Shuldiner is an employee of Regeneron Pharmaceuticals. Psaty serves on the Steering Committee of the Yale Open Data Access Project funded by Johnson & Johnson. E. K. Silverman reports grant support from Bayer and GSK. J. H. Yun reports consulting fees from Bridgebio Therapeutics. M.H. Cho reports grant support from Bayer and GSK, Consulting and speaking fees from Illumina and AstraZeneca. D.L. DeMeo reports grant support from Bayer. L.M. Raffield is a consultant for the TOPMed Administrative Coordinating Center (through WeStat). P. Desai reports grant support from Janssen Research. H.M. Kang is an employee of Regeneron Pharmaceuticals. A.T. Satpathy is a founder of Immunai and Cartography Biosciences and receives research funding from Allogene Therapeutics and Merck Research Laboratories. J. Kitzman is on the scientific advisory board of MyOme Inc.

varied by driver gene. Those carrying this protective allele had markedly reduced growth rate or prevalence of clones with driver mutations in *TET2*, *ASXL1*, *SF3B1*, and *SRSF2*, but not *DNMT3A*. *TCL1A* was not expressed in normal or *DNMT3A*-mutated HSCs, but the introduction of mutations in *TET2* or *ASXL1* led to TCL1A protein expression and expansion of HSCs in vitro. The protective allele restricted TCL1A expression and expansion of mutant HSCs, as did *TCL1A* shRNA knockdown. Forced expression of *TCL1A* promoted 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 hematopoiesis may be mediated by *TCL1A* activation.

Aging is characterized by the accumulation of somatic mutations, nearly all of which are “passengers” that have little fitness consequence. However, infrequent fitness-increasing mutations, called “drivers”, may result in an expanded lineage of cells, termed a clone. Clonal hematopoiesis of indeterminate potential (**CHIP**) is defined by the acquisition of specific, cancer-associated driver mutations in HSCs from persons without a blood cancer[1]. The genes commonly mutated in CHIP include regulators of DNA methylation (*TET2*, *DNMT3A*), chromatin remodeling (*ASXL1*), and RNA splicing (*SF3B1*, *SRSF2*, *U2AF1*). CHIP carriers have a risk of hematologic malignancy, coronary heart disease, and mortality in proportion to the variant allele fraction (**VAF**), a measure of clone size[2-8]. In contrast to low VAF clones, 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, partially due to a lack of sizable cohorts with serially sampled blood over decades which would otherwise enable studies on genetic and environmental correlates of clonal expansion. Here, we used PACER to investigate the germline determinants of clonal expansion in 5,071 CHIP carriers from the NHLBI Trans-Omics for Precision Medicine (**TOPMed**) program[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 (Figure 1a). Prior studies have enumerated passenger burden in HSCs by performing WGS on colonies derived from single cells[15,16]. We theorized that the passenger burden in the founding cell for a CHIP clone could instead 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 prior to the acquisition of the driver event (ancestral passengers) and mutations acquired after the driver event (sub-clonal passengers). Because the limit of detection for mutations from WGS at ~38X coverage depth is ~8-10% VAF, 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 due to the limited sensitivity of detection at 38X depth, high fitness clones will harbor more detectable passengers than lower fitness clones that arose at the same time. Based on 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 to 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.

We identified CHIP in 5,071 out of 127,946 participants in TOPMed by analyzing blood DNA whole genome sequencing (**WGS**) data with Mutect2[17] at pre-specified loci (Methods, Supplementary Table 1). CHIP was strongly associated with age at blood draw and >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 artifacts, 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% CI: 51%-57%) (Extended Data Fig 1a) compared to the controls after adjusting for age and study cohort using a negative binomial regression. Greater than 98% of the passengers were non-coding. We presumed the detected passengers in those without CHIP were reflective of clonal hematopoiesis with unknown driver mutations[18], though some of these could have been incompletely removed artifacts. The passengers were also positively associated with age, on average increasing by 13.7% (95% CI: 13.0%-14.3%) each decade. While 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 likely due 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 only 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 Text 1). The simulation indicated that founding passengers were associated with driver fitness (spearman ρ =0.09, pvalue < 2×10^{-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 hematopoietic colonies, in part because we limited the base substitutions in our analysis to C>T or T>C (Methods), but also likely due to the lower sensitivity of detecting true passengers in whole blood WGS compared to 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 Text 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 where clone growth is assessed longitudinally. A prediction of high importance 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 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 ‘passenger-approximated clonal expansion rate’, or PACER. Mutations in splicing factors (*SF3B1*, *SRSF2*, *U2AF1*) and *JAK2* V617F mutations were the fastest growing according to PACER, while *DNMT3A* R882− was among the slowest (Figure 1b, 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 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 prior empirical estimates of variant fitness derived from longitudinal sequencing of samples with clonal hematopoiesis[6,16,20-22]. When driver gene fitness estimates from PACER were directly compared to estimates from a large longitudinal dataset of clonal hematopoiesis[16], the Rsq was 80% (Figure 1c, Methods).

To further validate the utility of passenger count, we asked whether PACER could also predict future clone growth within individuals. We performed targeted sequencing for driver variants from two blood samples taken approximately 13-19 years apart in 55 CHIP carriers with a single driver mutation from the Women’s Health Initiative (**WHI**). WGS from the first time point was used to determine passenger count and the change in VAF of driver variants divided by the change in time ($\frac{dVAF}{dT}$) was used to approximate the empirical growth rate (Figure 1d). We constructed a simple estimator of $\frac{dVAF}{dT}$ using only the passengers, VAF, 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, hence these latter two variables were also considered in the model. A model that included age and VAF in addition to passenger count was superior for predicting $\frac{dVAF}{dT}$ (Rsq = 32.5%, Adjusted Rsq = 28.6%) than models only including passengers (Rsq = 12.6 %, Adjusted Rsq = 11%), age (Rsq = 13.9%, Adjusted Rsq = 12.3%), or VAF (Rsq = 0.3 %, Adjusted Rsq = −1.6%). In all models, the passenger count variable was significantly associated with $\frac{dVAF}{dT}$ (Figure 1e, Extended Data Fig 1c).

To contextualize its performance, we compared PACER to fitness estimators derived from longitudinal datasets (102 individuals with clonal hematopoiesis from Fabre et al. 2022[16] as well as 24 individuals from WHI) (see Supplementary Text 2 and Supplementary Table 3-4). Each individual had 3-5 assessments of VAF over several years, and fitness estimates derived from the first 2-4 measurements were used to predict $\frac{dVAF}{dT}$ between the penultimate and final timepoints. We observed that the point estimates of Rsq for the correlation of $\frac{dVAF}{dT} \sim$ 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 2-4 serial measurements.

To consider alternative statistical approaches, we compared the PACER derived fitness estimates to our hierarchical Bayesian estimator of clone fitness (**PACER-HB**, Methods), and observed strong correspondence between the two fitness estimates (Supplementary Text 1), suggesting that the relative simplicity of PACER does not clearly reduce its performance compared to 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* (Figure 2a), and genetic fine-mapping further narrowed down the associated region to a credible set containing a single variant, rs2887399 (Extended Data Fig. 1d, Methods). 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 alternative (alt) allele of rs2887399 is common, occurring in 26% of haplotypes sequenced in TOPMed, and each additional alt-allele associated with a 0.15 decrease in PACER z-score (pvalue = 4.5 x10^{-12}). rs2887399 lies in the core promoter of *TCL1A* as defined by the Ensembl regulatory build 108[23], 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 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 (pvalue = 2.9 x 10^{-5}, Supplementary Table 5-6), including a distal enhancer for *TNFAIP3* (pvalue = 5.4 x 10^{-7}) (GeneHancer[26]).

Stratified associations with *rs2887399*

We asked 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 alt-allele, pvalue = 9.6 x 10^{-4}, Supplementary Table 7). Stratification of PACER score by rs2887399 genotype revealed that the alt-allele slowed growth of *TET2* clones but had little effect on *DNMT3A* clones (Figure 2b).

Clones with a decreased expansion rate may never grow large enough to be detected, so we also performed association tests between rs2887399 and presence of a CHIP-associated driver mutation stratified by gene. In our previous analysis[11], we reported that the alt-allele was associated with increased risk for *DNMT3A* mutations. Prior reports have also identified that the alt-allele of rs2887399 decreases risk for mosaic loss of the Y chromosome (**LOY**)[27]. Here, we observed that rs2887399 was associated with significantly reduced odds of mutations in *TET2*, *ASXL1*, *SF3B1*, and *SRSF2* (Figure 2c, Supplementary Table 8-9). The effect size of rs2887399 was large, as 2 copies of the alt-allele conferred odds ratios for having a driver mutation from 0.22 to 0.63. The risk reduction was particularly strong for mutations in *SF3B1* and *SRSF2*, as well as for having >1 non-*DNMT3A* driver mutations. In sum, these results indicate that the alt-allele at rs2887399 is protective against CHIP due to driver mutations in several genes that have higher risk of progression to frank hematologic malignancy[6,28].

Our analysis predicts that the alt-allele of rs2887399 should reduce expansion rate of several -non-*DNMT3A* mutant clones. We performed targeted sequencing in 900 additional participants in WHI at two timepoints 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 asked whether the alt-allele was associated with the expansion rate of CH clones. We defined clonal expansion as the percent growth per year of the CH clones as estimated by a Bayesian logistic growth model (Methods). We observed that each alt-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 prediction of PACER (Figure 2d, Supplementary Table 10). *TET2* and *ASXL1* clones with the alt-homozygous rs2887399 genotype had very slow rates of clonal expansion (0.5% mean percent growth per year) compared to clones with the ref-homozygous genotypes (8.3% mean percent growth per year). These results provide further validation that PACER can accurately identify correlates of clonal expansion.

We sought to understand why the alt-allele of rs2887399 was associated with increased prevalence of *DNMT3A*-CHIP but had little effect on *DNMT3A* clonal expansion rate. Recent work has demonstrated that hematopoiesis becomes increasingly oligoclonal during aging as competition between clones with varying degrees of fitness intensifies[13]. We hypothesized that carrying the alt-allele of rs2887399 would lead to increased likelihood of *DNMT3A*-mutant clones growing to detectable levels due 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 *DNMT3A* clone fitness was kept constant but the relative fitness of the *TET2* clone was 20% higher relative to *DNMT3A* in one setting, but 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 hematopoietic cells

We sought to establish how rs2887399 alters clonal expansion. We first asked if rs2887399 was associated with *TCL1A* expression in any cell type. As identified in the GTEx v8[29], the alt-allele reduces expression of *TCL1A* in whole blood (normalized effect size = −0.13, pvalue = 1.4 x 10^{-5}). The GWAS of PACER 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 driven by B-cells, as *TCL1A* is highly expressed in B-cells but appears to have absent or low expression in all other cell types in blood except for rare plasmacytoid dendritic cells (Supplementary Table 11, Extended Data Fig 2c, 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 hematopoietic stem and progenitor cells (**HSPCs**) using publicly available single-cell RNA sequencing (**scRNA-seq**) and ATAC-sequencing (**ATAC-seq**) datasets of normal and malignant hematopoiesis. *TCL1A* was expressed in fewer than 1 in 1000 cells identified as HSC/MPPs in scRNA-seq data from 6 normal human marrow samples (range 0-0.17%)[32,33]. In contrast, *TCL1A* was expressed in a much higher 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, Supplementary Table 12). Next, using a dataset of ATAC-seq in normal and pre-leukemic HSCs (**pHSCs**)[34], which are residual non-leukemic HSCs present in patients with AML that often harbor only the initiating driver mutations, we evaluated chromatin accessibility at the *TCL1A* promotor. Consistent with the lack of *TCL1A* transcripts in normal HSCs, we observed that the promoter was not accessible in normal human donor HSCs, in HSCs from patients with AML that carried no driver mutations, or in pHSCs with *DNMT3A* mutations. In 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 log2 fold-change of any differentially accessible TSS peak in *TET2*-mutant versus control samples (Supplementary Table 13).

We next asked if the neighboring genes *TCL6* or *TCL1B* either 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 (Supplementary Table 12, Extended Data Fig 3c), further supporting *TCL1A* as the causal gene for clonal expansion.

Functional effect of rs2887399 on HSCs

Based on these observations, we proposed the following mechanistic model: Normally, the *TCL1A* promoter is inaccessible and gene expression is repressed in HSCs. In the presence of driver mutations in *TET2*, *ASXL1*, *SF3B1*, *SRSF2*, or LOY, *TCL1A* is aberrantly expressed and drives clonal expansion of the mutated HSCs. The presence of the alt-allele of rs2887399 restricts accessibility of chromatin 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 (homozygous reference), G/T (heterozygous), or T/T (homozygous alternate) at rs2887399. The three donors were healthy and between 29-32 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 (Figure 3a, 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 multipotent progenitors (**MPPs**) (Lineage− CD34+ CD38− CD45RA−), cultured them in cytokine-supported media, and then performed ATAC-seq. Consistent with the pHSC data, we detected increased accessibility at the *TCL1A* promoter in *TET2*-edited, but not *DNMT3A*-edited, cells from the rs2887399 G/G donor relative to AAVS1-edited cells (Figure 3b, Extended Data Fig 6, Supplementary Table 14). However, accessibility was decreased in samples from carriers of the alt-allele in a dose-dependent manner, indicating that the protective effect of the alt-allele of rs2887399 is mediated by blocking *TCL1A* promoter accessibility.

Next, we asked if the alt-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. ~1% of HSCs/MPPs from AAVS1 or *DNMT3A* edited samples were positive for TCL1A, which did not vary by rs2887399 genotype. In 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 alt-allele (Figure 3c-d, 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 >90% (Extended Data Fig 5), suggesting 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 single-cell RNA sequencing data from hematological malignancy samples (Extended Data Fig 3a).

To test if rs2887399 genotype had an effect on expansion of HSPCs in vitro, we edited the CD34+ cells from GG and TT donors, sorted HSCs (Lin− CD34+ CD38− CD45RA− CD90+), and analyzed for 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 (Figure 3e). A population of cells that was Lin−/lo CD34+ CD38− CD45RA dim (CD45RA^{dim} 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 to the T/T donor, indicating either less retention of stem/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 alt-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 shRNA targeting *TCL1A* or scramble control. The *TCL1A* shRNA construct we used was validated to knockdown TCL1A protein by ~90% (Extended Data Fig 8a). We then sorted GFP+ HSCs 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 (Figure 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) (Figure 4a) and performed in vitro clonal expansion assays on purified HSCs. The per-cell level of TCL1A protein expression in *TCL1A*-eGFP transduced HSCs was similar to *TET2*-mutant HSCs (Extended Data Fig 8b). After 14 days, cultures from HSCs that received *TCL1A*-eGFP virus had ~4-fold higher counts of phenotypic HSC/MPPs and colony forming cells compared to cultures from HSCs that received control-eGFP virus (Figure 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 lineage negative (Lin−) fraction of the resulting cell mixture totaling ~4% in each group (Methods, 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 post-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 (Figure 4c, Extended Data Fig 9b). After 22 weeks post-transplant, we assessed chimerism in the marrow. For our primary analysis, we examined the Lin− c-Kit+ Sca-1+ compartment that contains all relevant mouse HSC and MPP subsets and found a marked increase in percent GFP+ donor cells in the mice given

TCL1A-eGFP transduced cells compared to mice given control cells (mean 23.8% versus 3.9%, p=0.0054) (Figure 4d). For secondary analyses, we also looked at the different subsets of HSC/MPPs (LT-HSC, ST-HSC, MPP2, MPP3, 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 cell cycle status of cultured human HSC/MPPs and observed that *TCL1A* expressing cells were ~2-fold more likely to be cycling compared to control cells (Figure 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 CITE-seq after 7 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 (Figure 4f, Extended Data Fig 10a, Supplementary Table 16). Pseudotime[36] analysis supported a trajectory of progression from HSC/MPP 1 (initial state) to 4 (most ‘differentiated’ state) (Extended Data Fig 10b). HSC/MPP 1 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* (Figure 4g). In contrast, HSC/MPP 2-4 expressed lower levels of stem cell identity genes and proliferative markers. Cell cycle analysis confirmed these clusters contained cells that were predominantly in G0 or G1 phase (Extended Data Fig 10c). HSC/MPP 3-4 also displayed a progressive increase in genes associated with the integrated stress response such as *PPP1R15A* (GADD34), *DDIT3* (CHOP), and *ATF4*, as well as FOXO target genes such as *CDKN1A* (p21), *CDKN1B* (p27), *SOD2*, *CCNG2*, and *TXNIP* (Figure 4g, Extended Data Fig 10d and 11a). TCL1A has been reported to bind to and increase kinase activity of all AKT isoforms via 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/MPP 4 also expressed the highest levels of apoptosis effector genes *BAD*, *BCL2L11* (BIM), and *BBC3* (PUMA). Strikingly, we found that *TCL1A* expression led to a significant increase in the proportion of cells in the HSC/MPP 1 cluster, and a significant decrease in the proportion of cells in the HSC/MPP 3 and 4 clusters, an effect that was consistent in both donors (Figure 4h, Extended Data Fig 11b-c). When considered in aggregate, the HSC/MPP clusters from *TCL1A* expressing samples had reduced expression of FOXO target genes/gene sets and increased expression of cell cycle associated genes/gene sets compared to control samples (Supplementary Tables 17-18). This indicates that *TCL1A* may function to preserve HSCs in a proliferative state by avoiding prolonged, deleterious stress responses.

DISCUSSION

We developed a novel approach for inferring clonal expansion rate from a single time point and used it to perform a GWAS for this trait (see also Supplementary Note 3). Remarkably, a common variant of large effect in the promoter of *TCL1A* was associated with slower expansion rate and markedly reduced prevalence of several common driver mutations in CHIP. This variant likely blocks the aberrant de-repression of *TCL1A* which normally occurs in HSCs downstream of mutations in *TET2*, *ASXL1*, *SF3B1*, *SRSF2*, LOY, and possibly 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 for clonal expansion of HSCs. Importantly, our results suggest that pharmacologically targeting TCL1A may suppress growth of CHIP and hematological 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 where pre-malignant clones exist[40-42].

METHODS

Study Samples

Whole genome sequencing (WGS) was performed on 127,946 samples as part of 51 studies contributing to Freeze 8 NHLBI TOPMed program as previously described[10,11]. None of the TOPMed studies included selected individuals for sequencing because of hematologic malignancy. Each of the included studies provided informed consent. Information on the included cohorts, sequencing centers, and ethical approvals is included in Supplementary Tables 19-21. Age was obtained for 82,807 of the samples, and the median age was 55, 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).

WGS Processing, Variant Calling and CHIP annotation

BAM files were remapped and harmonized through the functionally equivalent pipeline[43]. SNPs and indels were discovered across TOPMed and were jointly genotyped across samples using the GotCloud pipeline[44]. An SVM filter was trained to discriminate between high- and low-quality variants. Variants were annotated with snpEff 4.3[45]. Sample quality was assessed through mendelian discordance, contamination estimates, sequencing converge, and among other quality control metrics.

Putative somatic single nucleotide variants and indels were called with GATK Mutect2[17], which searches for sites where there is evidence for alt-reads that support evidence for variation, and then performs local haplotype assembly. We used a panel of normals to filter sequencing artifacts and used an external reference of germline variants to exclude germline calls. We deployed this pipeline on Google Cloud using Cromwell[46].

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 leukemogenic 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[47], 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[48]. 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 5 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 persons 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 artifacts. 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 call set (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[49] to parse the Mutect2 VCFs and encoded each variant in an int64 value using the variant key encoding[50]. 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[51]. 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[52,53] probabilistic programming language. We used the negative binomial likelihood with a mean and overdispersion parameterization to facilitate interpretation. We used the identity function to link the passenger counts to the predictors as we modeled the effects on an additive scale. We modeled the expectation and overdispersion of the passenger counts observed at time (t_i) as

$$E(counts_i(t_i)) = \mu T_i + s_i(t_i - T_i) + \alpha_k$$
$$counts_i(t_i) \sim NB(E(counts_i(t_i)), I(i \in CHIP)\theta_0 + (1 - I(i \in CHIP))\theta_1)$$

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 overdispersion 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 overdispersion 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 $T_i = \psi_i * age_i$, 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 supplement.

To estimate the posterior, we used the Stan Hamiltonian Monte-Carlo (HMC) 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 great 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[54].

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[55]. 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 Fabre et al.[16], Supplementary Table 6 (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 $1 / Fabre_{SE}$, 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[11,56]) 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 Women’s Health Initiative (WHI). Women aged 50–79 years were enrolled from forty WHI clinical centers in the United States between 1993 and 1998. All WHI participants had a blood sample collected at the time of enrollment, and a subset had subsequent blood sample collected 14-19 years later. Reads were aligned with bwa-mem and processed with the mimips pileline[57]. We called somatic variants using an ensemble of VarScan[58], Mutect2[17], and manual inspection with IGV[59] as previously described[60]. Including the 55 individuals previously known to have CHIP, a total of 455 individuals were identified to have CH at a VAF threshold for inclusion of variants of >0.005.

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 (dVAF/dT), we divided the change in VAF at the two timepoints (from smMIPS) by the change in age in years. Of the 55 CHIP carriers, 15 had clones which had negative dVAF/dT. 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 dVAF/dT 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 dVAF/dT 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-squared and Akaike information criterion (AIC) 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 CH carriers (VAF>0.005) with mutations in *DNMT3A*, *TET2*, *ASXL1*, or *SF3B1*, as identified using smMIPS described above, to test whether the alt-allele at rs2887399 altered clonal expansion rate. To estimate the rate of clonal expansion in the CHIP carriers in units of percent 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 modeled the observed number of mutated alleles using a beta-binomial likelihood, and included a random intercept and slope for each individual donor:

$$x_i = (Gene_{ij} + R_i * Gene_{ij} + U_{i1}) * age + U_{i2}$$
$$q_i = \frac{0.5}{1 + e^{-x_i}}$$
$$P(Y_i = y) = P(BetaBinomial(q_i, \beta, D_i) = y)$$

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 alt-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:

$$Gene_{ij} \sim Normal(0, 0.20)$$
$$R_i \sim Normal(0, 0.05)$$
$$U_{i1} \sim Normal(0, 0.05)$$

We performed inference using the MCMC sampler implementation available in the RStan probabilistic programming language[52,53].

Single Variant Association

Single variant association for each variant in the TOPMed Freeze 8 germline genetic variant call set[10] with a MAC > 20 was performed with SAIGE[61] 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. 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 declared variants from this analysis as significant if their p-value was less than 5 x 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 alt-homozygotes using the ref-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 200kb region surrounding *TCL1A*. We used the same covariates as the single variant association analysis. We used the posterior inclusion probabilities (PIP) and credible sets identified by SuSIE to identify the putative causal variant. We used LD directly calculated on the genotypes as opposed to an external reference.

Rare Variant Analyses

We performed 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 four and less than 300 (MAF of 3.7% in the analyzed samples). We parsed the genotypes using cyvcf2[49] 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 x 10^{-4}. We did not group the variants by annotation and we declared regions as significant if their pvalue was less than 2.9 x10^{-5} (.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 analyzed samples was less than 200.

Re-analysis of single-cell RNA sequencing 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 10000 by the NormalizeData command. We then used the FindVariableFeatures command with 'vst' selection method and 2000 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 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 5 healthy marrow samples (id01, id06, id09, id13, id17) and 4 MPN samples (id2, id7, id11, 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 single-cell RNAseq data from Velten et al.[32], generated using MutaSeq, was downloaded from[69] 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 preleukemic ‘HSC/MPP’ and preleukemic ‘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, 7256), or preleukemic HSCs (pHSC), defined as Lin− CD34+ CD38− TIM3− CD99− by the authors. For the pHSC samples, we selected 3 where there were no detectable driver mutations in the pHSC compartment (SU336, SU306, SU623), 2 where there were founding *DNMT3A* mutations only (SU444, SU575), and 3 where there were founding *TET2* mutations only (SU070, SU501, SU048).

Fastq files for these samples were downloaded, and ATAC-seq data analysis was performed as previously described[70]. Briefly, 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 chrX were retained -- chrY 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 transcription start sites 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 3 *TET2* mutant samples to the 7 control samples using the DESeq function in the *DESeq2*[71] 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 TSS 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 media (StemSpanII + 10% CD34+ Expansion Supplement + 0.1% Penicillin/Streptomycin) for 48 hours before CRISPR editing. Editing of *AAVS*, *TET2*, *DNMT3A*, and *ASXL1* was performed by electroporation of Cas9 ribonucleoprotein complex (RNP). For each combination of rs2887399 genotype and gRNA (Supplementary Table 22), 100,000 cells were incubated with 3.2 ug of Synthego synthetic sgRNA guide and 8.18 ug of IDT Alt-R S.p. Cas9 Nuclease V3 for 15 minutes at room temperature before electroporation. CD34+ cells were resuspended in 18 uL 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 media and allowed to recover for 24 hours. CRISPR editing efficiency was measured using Sanger Sequencing and ICE Analysis.

ATAC-seq

24 hours post electroporation, Lineage− 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 media before 40,000 cells were harvested, and bulk Omni-ATAC[70] was performed on them. Briefly, cells were lysed with ATAC-Resuspension Buffer containing 0.1% NP40, 0.1% Tween-20, and 0.01% Digitonin for 3 minutes, and then the transposition was performed for 30 minutes 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 (cat# D4014). The transposed fragments were amplified and indexed using NEBNext 2x 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 2x75 bp reads on an Illumina NextSeq550 instrument using the High Output Kit.

ATAC-seq data analysis was performed as described above. Briefly, reads were trimmed and filtered using fastp and mapped to the hg38 reference genome using hisat2[72] with the --no-spliced-alignment option. BAM files were deduplicated using Picard. Only reads mapping to chromosomes 1-22 and chrX were retained -- chrY 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 transcription start sites 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 IGV. 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 7 days in culture, from which point the Omni-ATAC protocol was followed. We used the top 1000 most accessible TSSes genome-wide to perform normalization. We devised this strategy based on our observation that some inaccessible TSSes 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 alt-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

Lineage− CD34+ CD38− CD90+ CD45RA− cells were sorted on a BD FACS Aria III from the electroporated CD34+ cells. All cells were harvested and stained with the extracellular HSPC marker panel in 100 uL of PBS + 2% FBS + 1 mm EDTA (Supplementary Table 23). Four to eight replicates of 500-1,000 Lineage− CD34+ CD38− CD90+ CD45RA− cells were sorted into 100 uL of HSC Expansion media 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 uL of HSC Expansion media was added to each well. 10 days post sort, the samples were transferred from the 96 well plate to a 48 well plate and an additional 400 uL of HSPC Expansion media was added. Fourteen days post sort, the cells were harvested, and live cells were counted using trypan blue and hemocytometer. 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 Figure 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 media were harvested and intracellularly stained 11 days following electroporation. Cells were first stained with the Live/Dead and extracellular surface markers simultaneously for 30 minutes in the dark on ice. After a PBS wash, cells were stained with 100 uL of IC Fixation Buffer for 30 minutes in the dark at room temperature. Cells were then washed twice with 1X Permeabilization Buffer. Next, cells were resuspended in 100 uL of 1X Permeabilization Buffer, and blocked with 2 uL of goat serum and 2.5 uL of TruStain FcX for 15 minutes in the dark at room temperature. Next, 1 ug of e450 antibodies (anti-TCL1A or isotype control) was added to each sample tube and stained for 30 minutes in the dark at room temperature (Supplementary Table 24). Cells were then washed twice with 1X 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 (CAT#: TL301172V). The shRNA constructs were validated to knockdown TCL1A protein by flow cytometry in NALM6 cells (from Ronald Levy, Stanford University). 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 EF1a promoter, as well as a control sequence composed of GFP under the EF1a promoter, was synthetized 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 media 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% confluency 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 media was exchanged 16 hours later, and the viral supernatant was collected at 72h post-transfection. The collected viral supernatant was filtered via a 0.45 μm filtration unit, and concentrated using the LentiX concentrator (Takara) for 2 hours at 4 C and then spun down at 1500 x g for 45 minutes at 4 C. The concentrated supernatant was subsequently aliquoted, flash frozen, and stored in −80°C until use.

Combined CRISPR and shRNA Assay

CD34+ cells were thawed and cultured in HSPC Expansion media (StemSpanII + 10% CD34+ Expansion Supplement + 0.1% Penicillin/Streptomycin) for 48 hours before CRISPR editing. Editing of AAVS, TET2, DNMT3A, and ASXL1 was performed by electroporation of Cas9 ribonucleoprotein complex (RNP). For each combination of rs2887399 genotype and gRNA, 100,000 cells were incubated with 3.26 ug of Synthego synthetic sgRNA guide and 8.332 ug of IDT Alt-R S.p. Cas9 Nuclease V3 for 15 minutes at room temperature before electroporation. CD34+ cells were resuspended in 18 uL 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 500,000 cells was transferred to 2 mLs of HSPC Expansion media and allowed to recover for 8 hours. Later that same day, 250,000 CRISPR edited cells were collected, spun down, and resuspended in a final volume of HSPC Lentivirus Media (StemSpanII + 10% CD34+ Expansion Supplement + 0.1% Penicillin/Streptomycin + 10 uM prostaglandin E2 + 100 ng/uL poloxamer 407) with virus added at an MOI of 20. Cells were plated in a 96 well u-bottom plate for 16 hours. shRNA-A and the scramble-shRNA from Origene CAT#: TL301172V were used for this experiment. Following a 16-hour incubation, cells were washed in PBS, and then plated in 2 mL of HSPC Expansion media. After 72 hours, previously described liquid culture expansion assay was done on sorted Lineage− CD34+ CD38− CD90+ CD45RA− GFP+ cells.

Lentiviral *TCL1A* Expression in Human HSPCs

CD34+ cells were thawed and cultured in HSPC Expansion media (StemSpanII + 10% CD34+ Expansion Supplement + 0.1% Penicillin/Streptomycin) for 48 hours before lentivirus transduction. 750,000 cells were collected, spun down, and resuspended in a final volume of HSPC Lentivirus Media (StemSpanII + 10% CD34+ Expansion Supplement + 0.1% Penicillin/Streptomycin + 10 uM prostaglandin E2 + 100 ng/uL poloxamer 407) with virus added at an MOI of 100. Cells were plated in a 96 well u-bottom plate for 16 hours. eGFP control was purchased from Origene (CAT#: PS100093V) or produced in house as described above, and the TCL1A-eGFP was purchased from Origene (CAT#: RC204243L4V) or produced in house as described above. Following 16-hour incubation, cells were washed in PBS, and then plated in 2 mL of HSPC Expansion media. After 72 hours, previously described liquid culture expansion assay was done on sorted Lineage− CD34+ CD38− CD90+ CD45RA− GFP+ cells. After 14 days, cells were harvested 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 of Methocult + 0.1% P/S, and plated in 35 mm dishes. Eight 35 mm dishes were placed in one 245 x 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 CFU count in the day 14 liquid culture was determined by multiplying the number of CFU 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 media. Cells were first stained with the Alexa-700 Live/Dead and extracellular surface markers simultaneously for 30 minutes in the dark on ice (Supplementary Table 25). After a PBS wash, cells were stained with 100 uL of IC Fixation Buffer for 30 minutes in the dark at room temperature. Cells were then washed twice with 1X Permeabilization Buffer. Next, cells were resuspended in 100 uL of 1X Permeabilization Buffer, and blocked with 2 uL of goat serum for 15 minutes in the dark at room temperature. Cells were then washed twice with 1X Permeabilization Buffer and then resuspended in 75 uL of 1 ug/mL DAPI diluted in 1X Permeabilization buffer. After 10 minutes, 75 uL of 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 Figure 4b.

Mouse Bone Marrow Competitive Transplant

Mice were obtained from The Jackson Laboratory and housed at the Research Animal Facility **(RAF)** 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 68–73 °F under 40–60% humidity. All animal procedures were performed in accordance with protocols approved by Stanford University’s Administrative Panel on Laboratory Animal Care (APLAC).

Bone marrow from 10-week-old female CD45.2+ C57BL/6 mice was harvested, and c-Kit cells were enriched for using the EasySep Mouse cKIT Positive Selection Kit (Catalog #18757) according to manufacture protocol. 2.8 million c-KIT enriched cells were transduced with 45 uL of the previously described Control-eGFP or TCL1A-eGFP and cultured overnight in U-bottom plates in mouse HSC transduction media (StemSpan II, 10 ng/mL SCF, 100 ng/mL TPO, 10 uM PGE2, 100 ng/uL P407, 0.1% P/S) 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 to achieve chimeric donor bone marrow graft. Sorting of GFP+ cells pre-transplant was not conducted because anecdotal evidence from several labs suggests that culture of transduced HSCs for >24 hours diminishes their potency for *in vivo* reconstitution. Post-hoc analysis of stored aliquots from the input cells confirmed ~4% of Lineage− 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$^+$ 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 uL 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 uL of the mouse peripheral blood antibody cocktail (Supplementary Table 26). Twenty-two weeks post-transplant, mice were euthanized and bone marrow was harvested from femurs. Following RBC lysis, bone marrow was stained with 50 uL 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 Figure 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 media (StemSpanII + 10% CD34+ Expansion Supplement + 0.1% Penicillin/Streptomycin) for 48 hours before lentiviral transduction. 72 hours after lentivirus addition, Lineage− 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 harvested and resuspended in 50 uL of PBS + 1% BSA. Cells were then blocked with 5 uL of TruStain FX for 10 minutes. Next, cells were stained with 0.5 uL of each TotalSeq-B antibody (CD34, CD38, CD45RA, CD90, CD49f, CD35, CD11a, CD59, CD117) for 30 minutes. Following 4 washes with PBS + 1% BSA, 10,000 cells were loaded onto a Chromium Next GEM Chip G. GEM generation & barcoding, post GEM–RT cleanup & cDNA amplification, 3’ gene expression library construction, and cell surface protein library construction were performed as described in CG000317_ChromiumNextGEMSingleCell3'v3.1_CellSurfaceProtein_RevC (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-barcodingtechnology-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

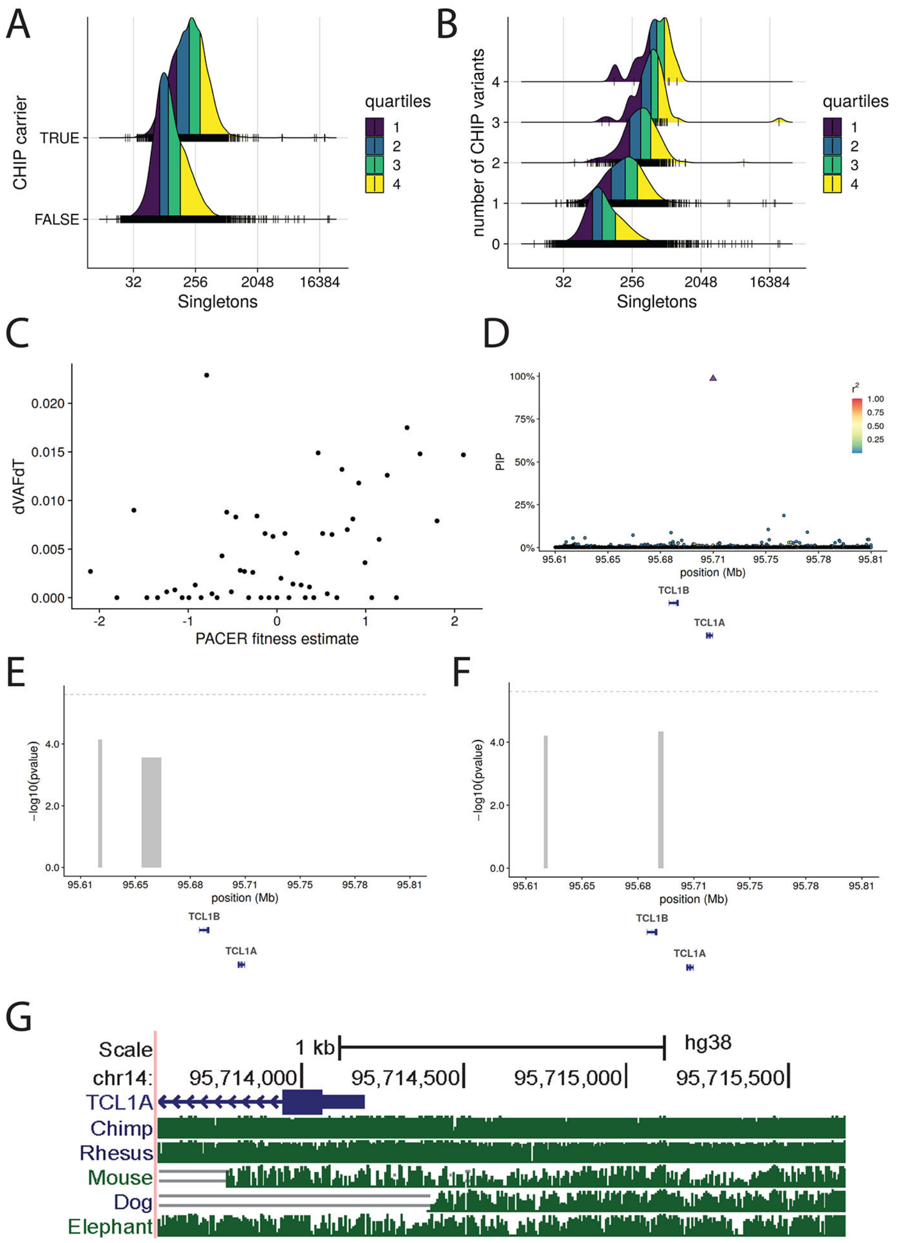
The BCL files were demultiplexed using 8 base pair 10X sample indexes and cellranger mkfastq to generate paired-end FASTQ. We ran cellranger count to align the reads to the hg38 reference genome from GenBank using STAR[73] aligner as well as perform filtering, barcode counting, and UMI counting. The 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 centered log-ratio (CLR) normalization, computed independently for each feature.

The 4 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/MPP 1-4 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/MPP 4 versus HSC/MPP 1 was performed using the FindMarkers function in Seurat with no thresholds for min.pct or logfc.threshold.

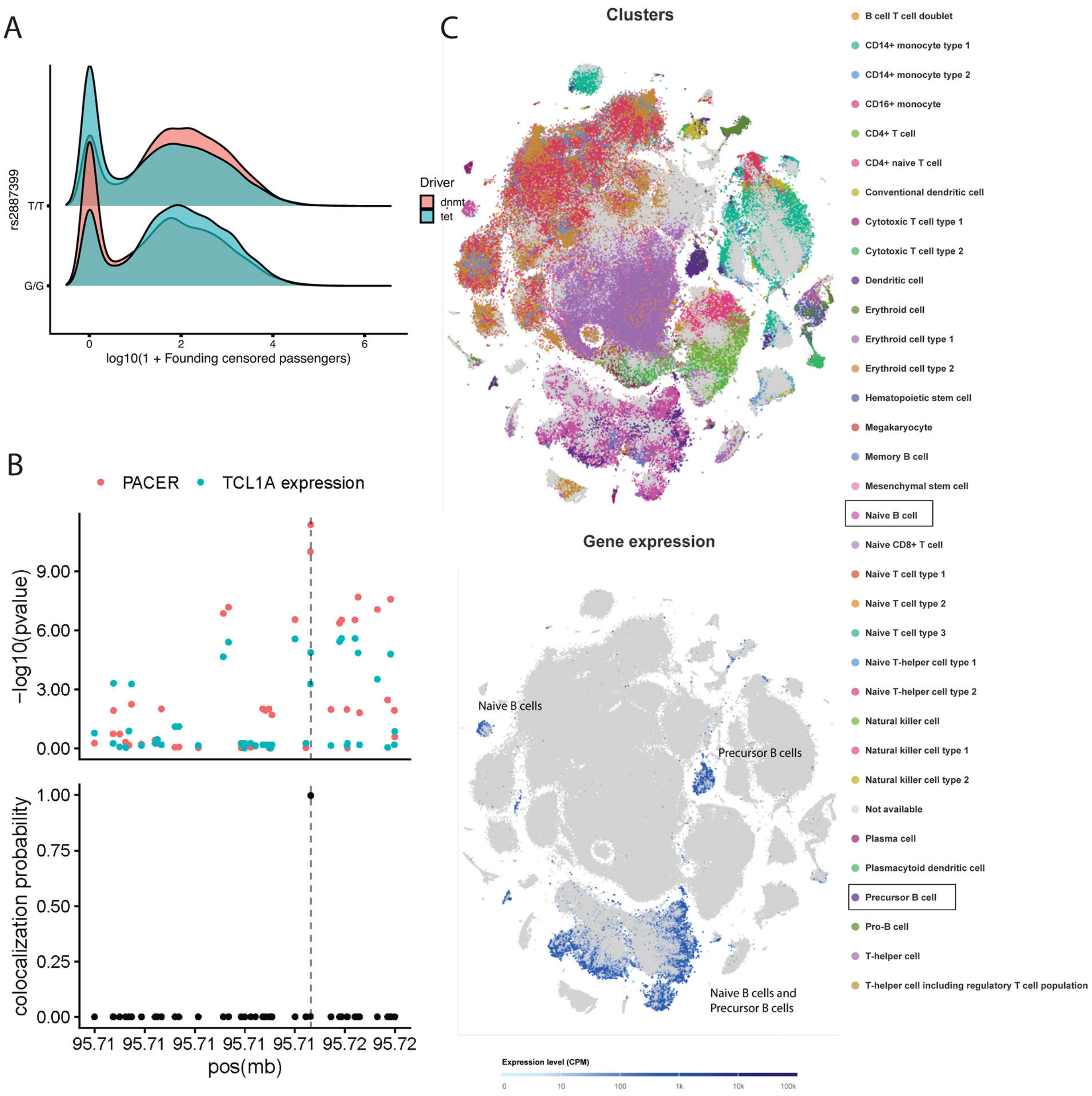
Gene-set enrichment analysis was performed using the fgsea package (https://github.com/ctlab/fgsea) and the REACTOME gene sets using the following parameters for the fgsea function: nperm = 1000, scoreType = "std", minSize=5. Results of differential expression analysis and GSEA can be found in Supplementary Tables 17-18.

Extended Data



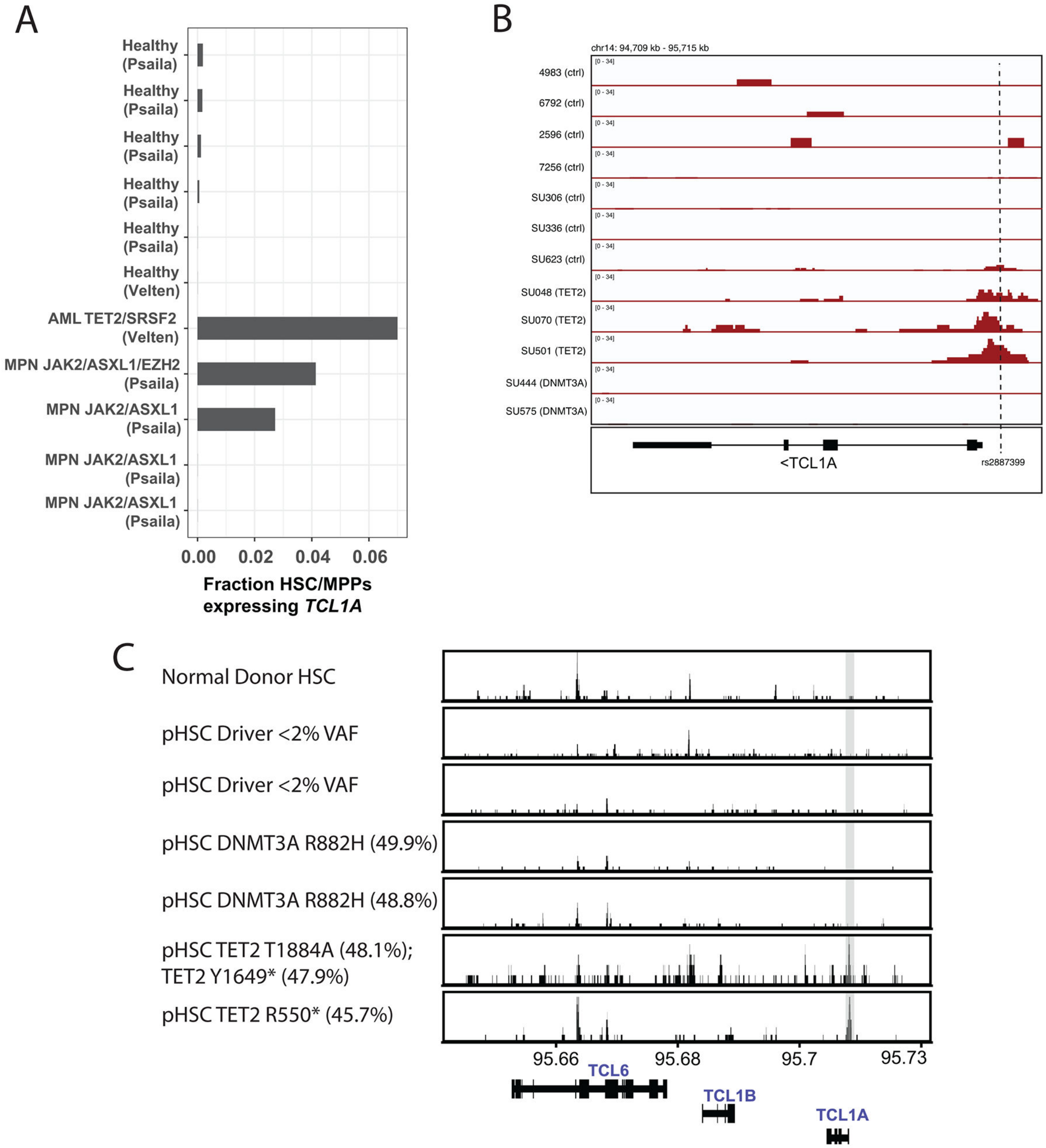
Extended Data Fig 1|. PACER Estimates Clonal Expansion Rate

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[45] 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.



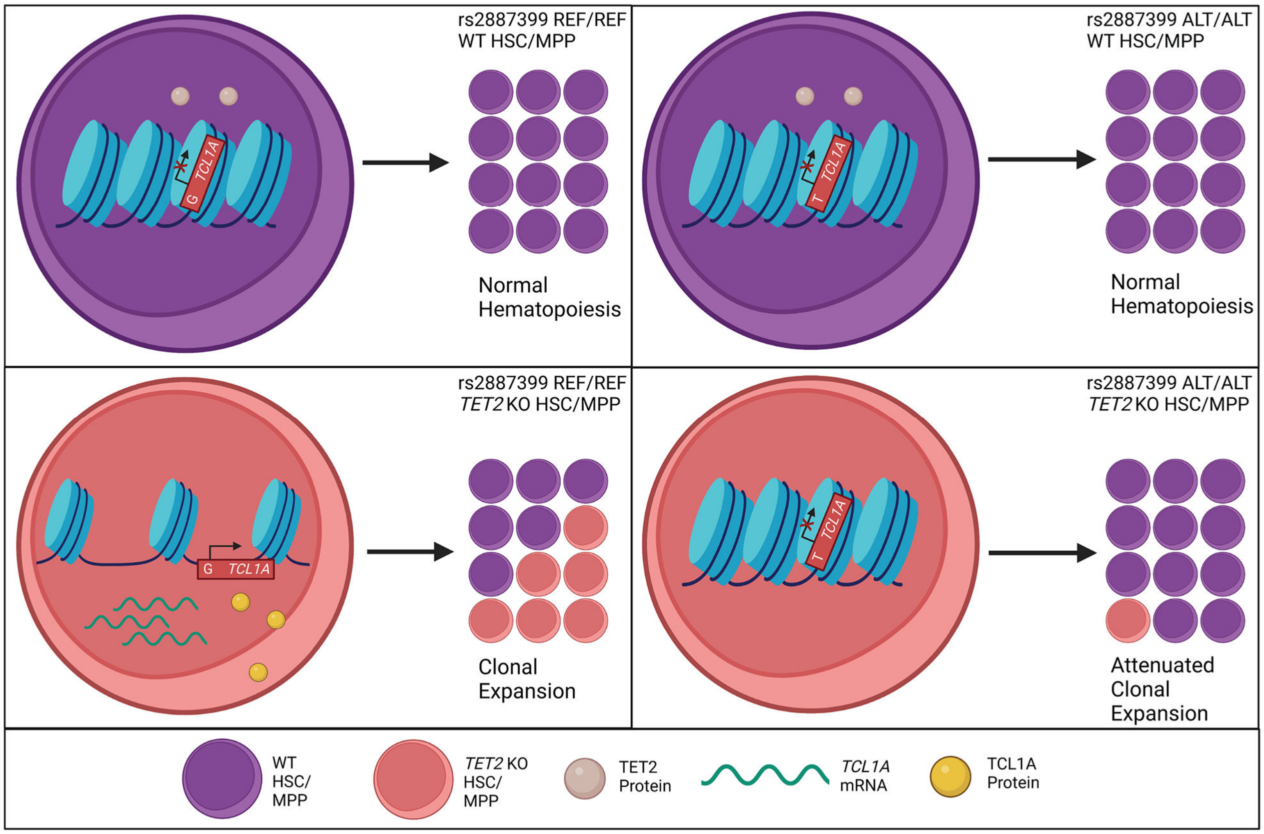
Extended Data Fig 2|. GWAS Implicates rs2887399 as a Modifier of Clonal Expansion Rate

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.



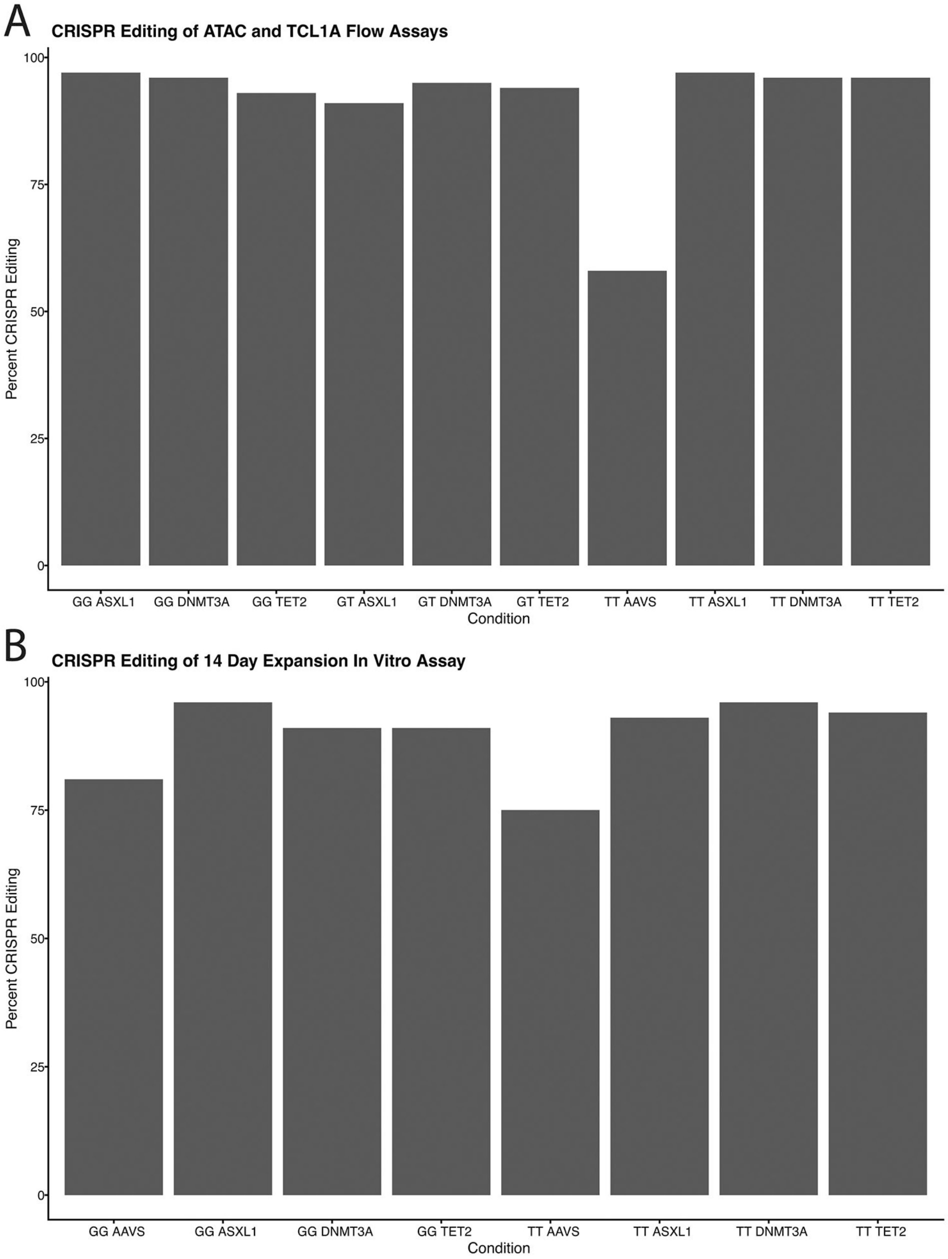
Extended Data Fig 3|. Chromatin Accessibility and Transcript Expression of TCL1A

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[36]. 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].



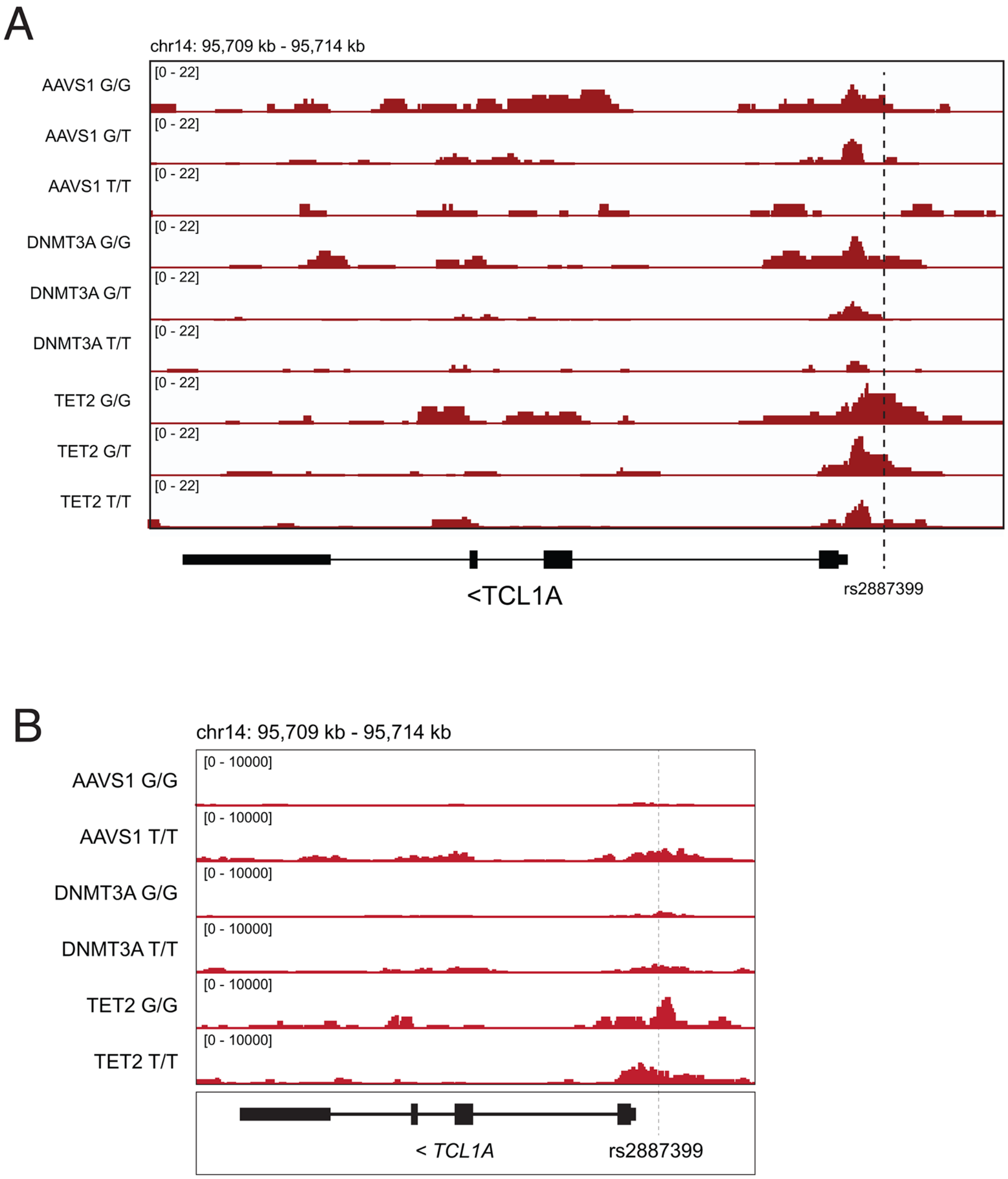
Extended Data Fig 4|. Schematic of rs2887399 Effect on TET2 Clonal Expansion

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.



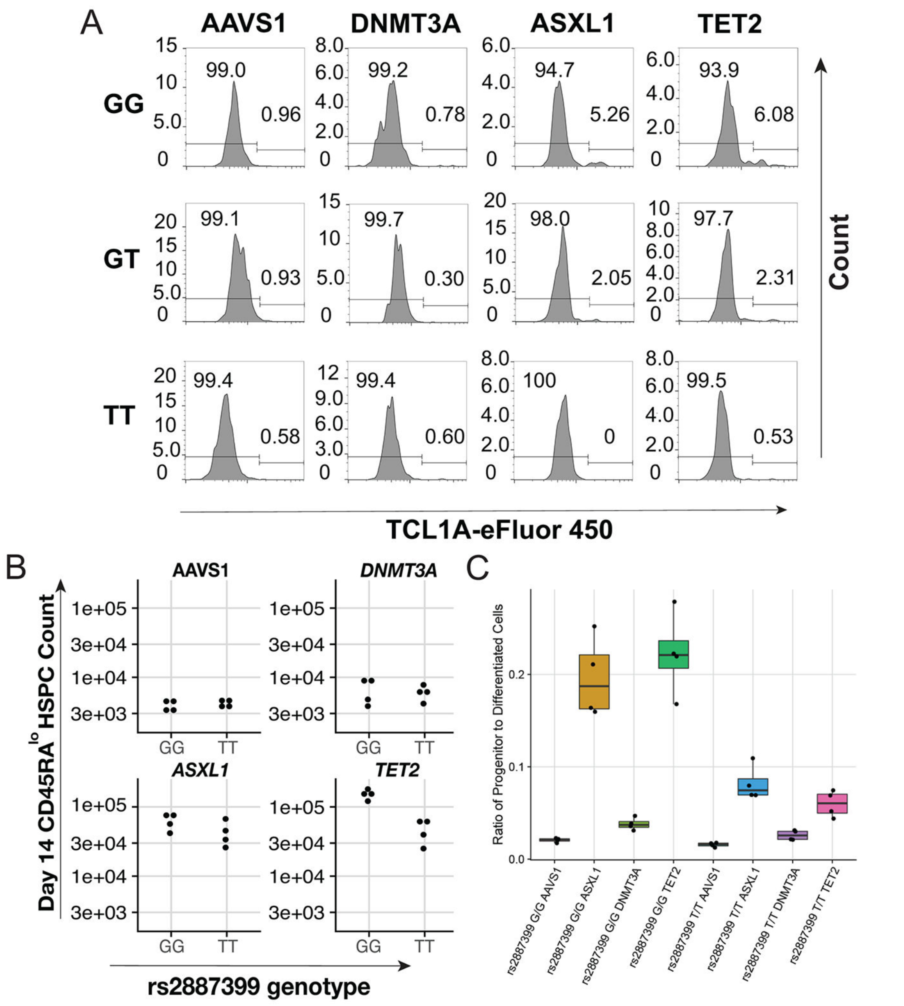
Extended Data Fig 5|. CRISPR Editing Efficiency

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 Figure 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.

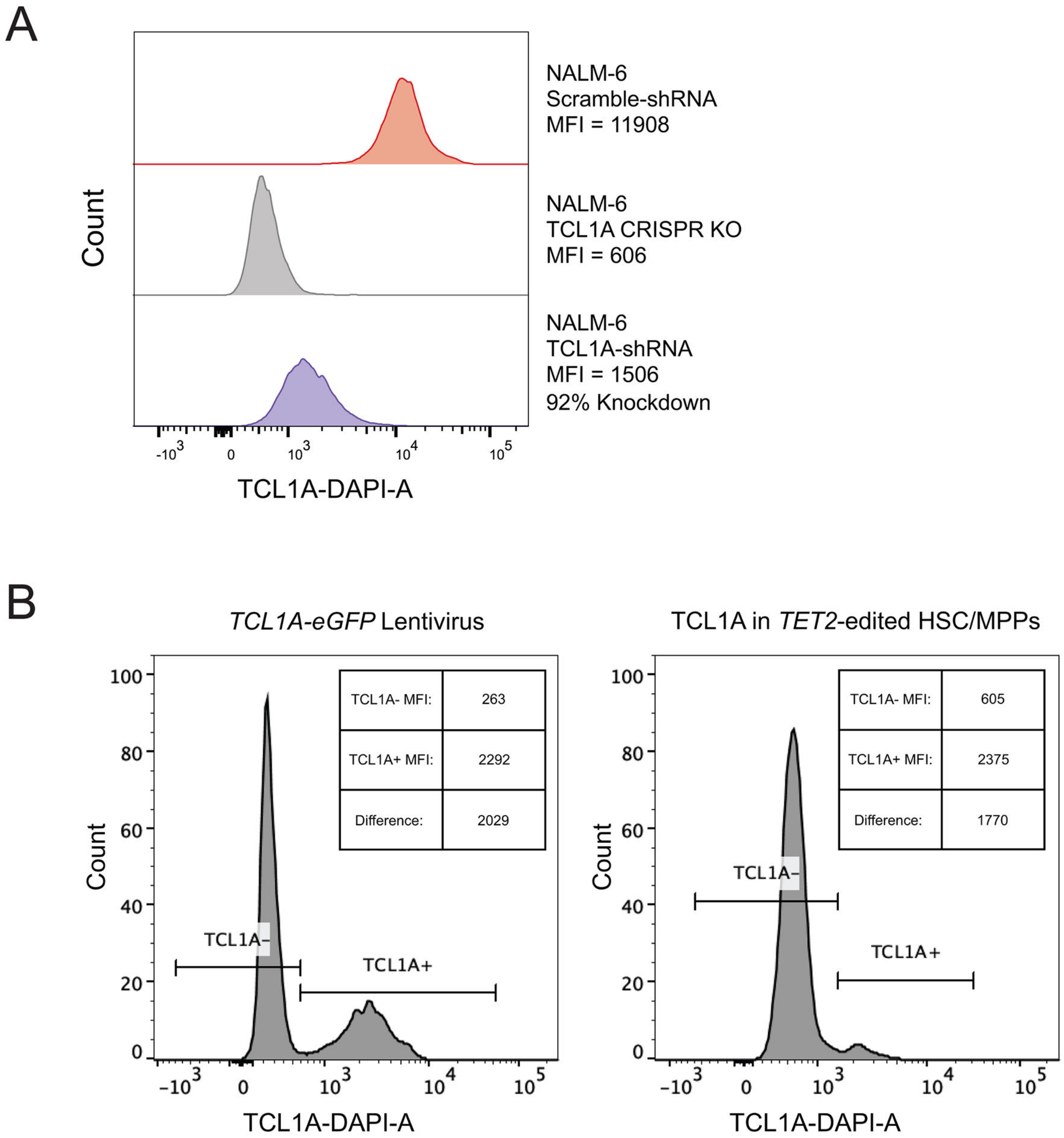


Variable	Est.	S.E.	p. val.
AAVS	(referent)		
DNMT3A	2695	9923	0.79
ASXL1	58573	9923	4.3×10^{-6}
TET2	148772	9923	1.1×10^{-13}
rs2887399 G/G	(referent)		
rs2887399 T/T	277	9923	0.98
DNMT3A:T/T	-849	14033	0.95
ASXL1:T/T	-19771	14033	0.17
TET2:T/T	-106080	14033	8.5×10^{-8}

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}

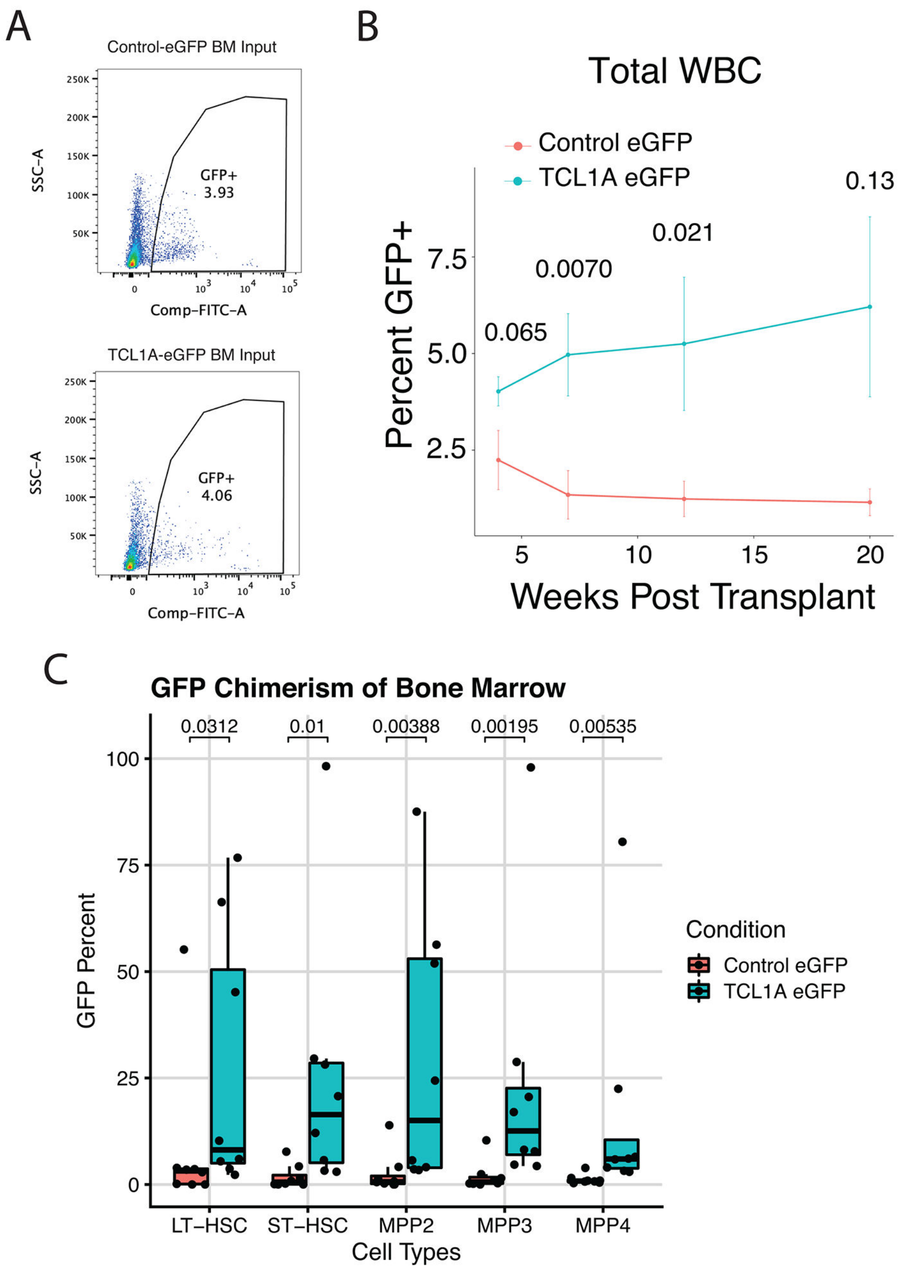
Extended Data Fig 7|. Interaction of CHIP Mutations and rs2887399 in human HSPC phenotypes

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.



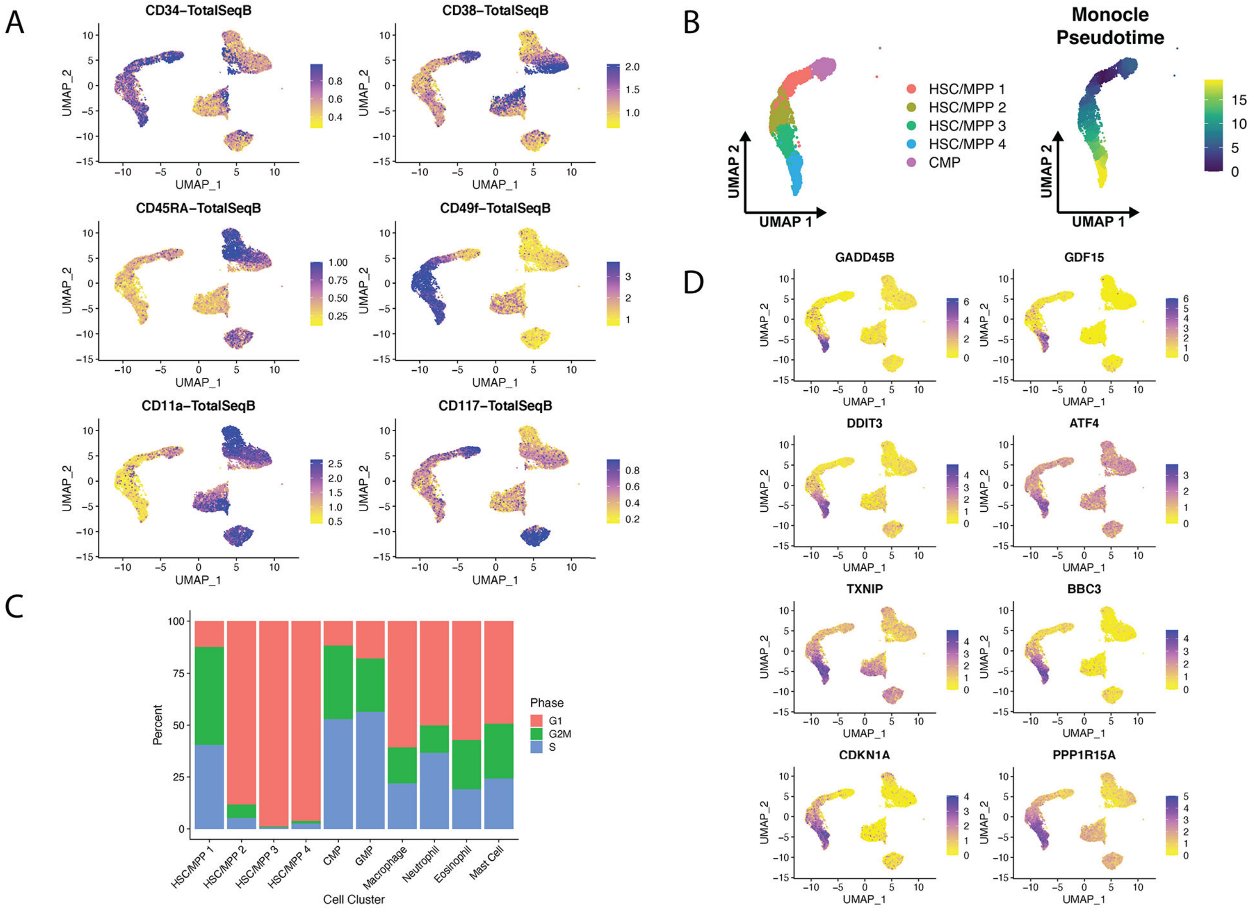
Extended Data Fig 8|. Validation of TCL1A shRNA and Expression Lentivirus

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.



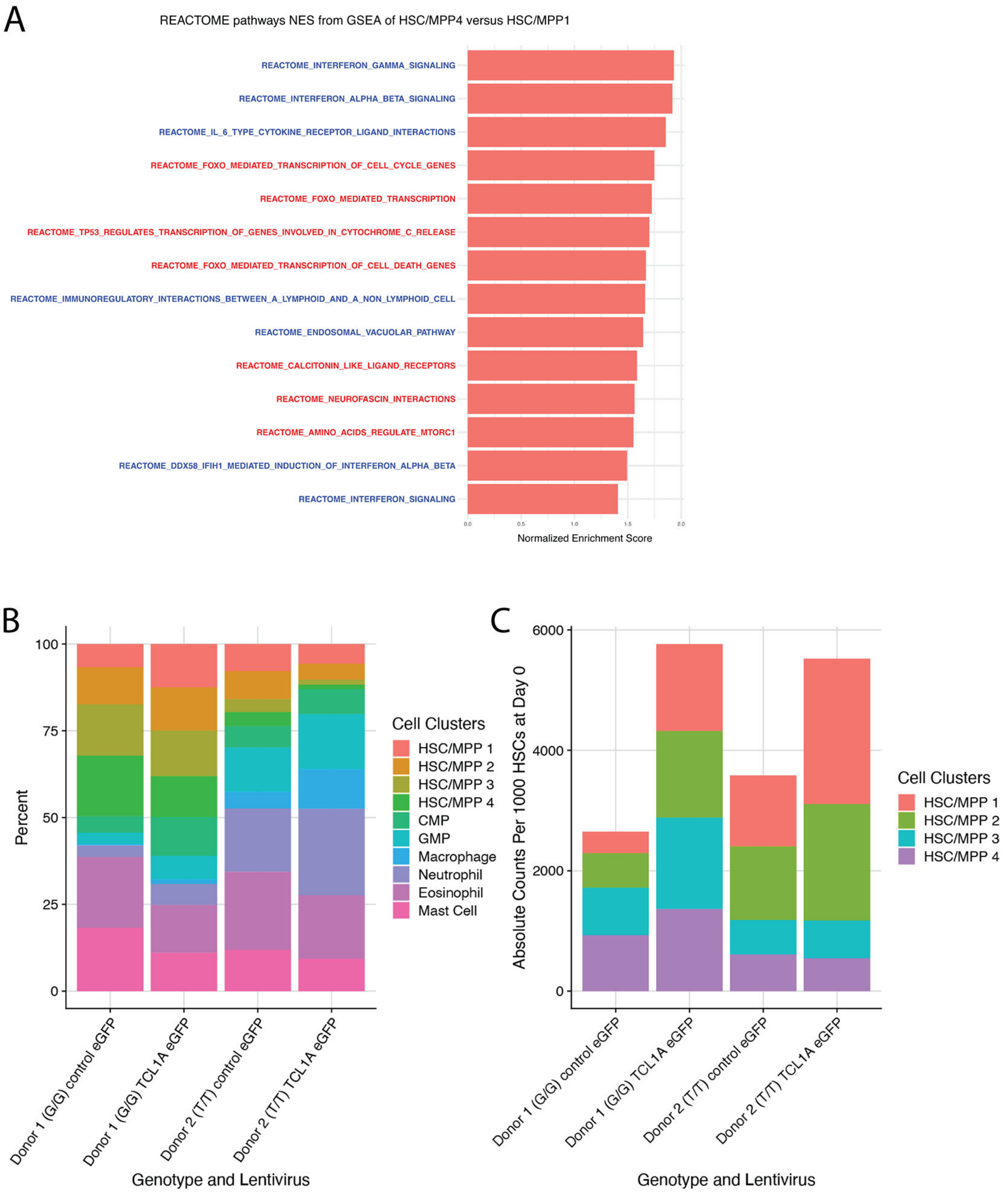
Extended Data Fig 9|. TCL1A Expression Promotes HSC Fitness in Mice

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 timepoint. 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

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

Affiliations

[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 & 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 Washington, Seattle, WA

[10] -Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA

[11] -The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA USA

[12] -Department of Medicine I, University Hospital, LMU Munich, Munich, Germany

[13] -German Centre for Cardiovascular Research (DZHK), partner site: Munich Heart Alliance, Munich, Germany

[14] -Department of Medicine, University of Maryland, Baltimore, Baltimore, MD, USA

[15] -Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA

[16] -Division of Biomedical Informatics and Personalized Medicine, Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

[17] -The Charles Bronfman Institute of Personalized Medicine

[18] -The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[19] -Institute for Genomic Health

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

[21] -Department of Preventive Medicine, Northeastern University, Chicago, IL

[22] -Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA

[23] -Harvard Medical School, Boston, MA USA

[24] -Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA

[25] -Department of Epidemiology, University of Washington, Seattle, WA, USA

[26] -Department of Medicine, University of Washington, Seattle, WA, USA

[27] -Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA USA

[28] -Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC USA

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

[30] -Channing Division of Network Medicine and Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, MA USA

[31] -National Heart Lung and Blood Institute's, Boston University's Framingham Heart Study, Framingham, MA, USA

[32] -National Heart, Lung and Blood Institute, Population Sciences Branch, Framingham, MA USA

[33] -Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA

[34] -Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI, USA

[35] -Department of Epidemiology, Tulane University School of Public Health and Tropical Medicine, New Orleans, LA, USA

[36] -Department of Cardiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

[37] -Albert Einstein College of Medicine, Department of Epidemiology and Population Health, Bronx, NY USA

[38] -Department of Epidemiology, University of Washington, Seattle WA 98195, USA

[39] -Kaiser Permanente Washington Health Research Institute, Kaiser Permanente Washington, Seattle WA 98101, USA

[40] -Â· Seattle Epidemiologic Research and Information Center, Department of Veterans Affairs Office of Research and Development, Seattle WA 98108, USA

[41] -Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA 98101, USA

[42] -Dean's Office, College of Public Health, University of Kentucky, Lexington, KY, USA

[43] -University of Alabama at Birmingham, Birmingham, AL, USA

[44] -Intermountain Heart Institute, Intermountain Medical Center

[45] -Department of Medicine, Jackson Heart Study, University of Mississippi Medical Center, Jackson, MS

[46] -Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[47] -Department of Medicine, University of Mississippi Medical Center, Jackson, MS, USA

[48] -Mayo Clinic, Department of Health Sciences Research, Rochester, MN, USA

[49] -Department of Public Health Sciences, Center for Public Health Genomics, University of Virginia, Charlottesville, VA USA

[50] -Department of Pathology & Laboratory Medicine and Biochemistry, Larner College of Medicine at the University of Vermont, Colchester, VT, USA

[51] -Department of Medicine, University of Washington, Seattle, WA USA

[52] -Research Institute, Bloodworks Northwest, Seattle, WA USA

[53] -Genome Science, University of Washington, Seattle, WA USA

[54] -Department of Cardiology, Clinical Sciences, Lund University and Skåne University Hospital, Lund, Sweden

[55] -The Wallenberg Laboratory/Department of Molecular and Clinical Medicine, Institute of Medicine, Gothenburg University and the Department of Cardiology, Sahlgrenska University Hospital, Gothenburg, Sweden

[56] -Wallenberg Center for Molecular Medicine and Lund University Diabetes Center, Lund University, Lund, Sweden

[57] -Department of Internal Medicine, Clinical Sciences, Lund University and Skane University Hospital, Malmo, Sweden

[58] -Vitalant Research Institute, San Francisco, CA, USA

[59] -Department of Human Genetics, University of Texas Rio Grande Valley School of Medicine, Brownsville, TX USA

[60] -South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley School of Medicine, Brownsville, TX USA

[61] -Department of Epidemiology and International Health Institute, Brown University School of Public Health, Providence, RI, USA

[62] -Center for Individualized and Genomic Medicine Research (CIGMA), Department of Internal Medicine, Henry Ford Health System, Detroit, MI, USA

[63] -Division of Biostatistics, Campus Box 8067 Washington University School of Medicine 660 S. Euclid Avenue St. Louis, MO 63110 USA

[64] -Department of Medical Research, Taichung Veterans General Hospital, Taichung, Taiwan; 1650, Sec. 4, Taiwan Boulevard, Taichung City, Taiwan

[65] -Division of Cardiology, University of California, San Francisco, San Francisco, CA

[66] -Department of Medicine, Cardiovascular Research Institute, University of California, San Francisco, San Francisco, CA

[67] -Cardiovascular Research Institute, University of California, San Francisco

[68] Division of Cardiology, Vanderbilt University Medical Center, Nashville, TN, USA

[69] Division of Cardiology, University of Illinois at Chicago, Chicago, IL, USA

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

[71] Department of Cardiology, Cedars-Sinai, Los Angeles, CA, USA

[72] Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

[73] Department of Medicine, Brigham and Women's Hospital and Harvard Medical School

[74] Division of Hematology and Oncology, Weill Cornell Medicine, New York, NY, USA

[75] Englander Institute of Precision Medicine, Weill Cornell Medicine, NY

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

[77] Regeneron Pharmaceuticals, Tarrytown, NY, USA

[78] Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA

[79] Department of Medicine, Harvard Medical School, Boston, MA, USA

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

[81] Department of Epidemiology, University of Washington, Seattle, WA 98195

[82] Institute for Stem Cell Biology and Regenerative Medicine, Stanford University School of Medicine, Stanford, CA

ACKNOWLEDGEMENTS

WGS for the TOPMed program was supported by the National Heart, Lung and Blood Institute (NHLBI). Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Phenotype harmonization, data management, sample-identity quality control and general study coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We thank the studies and participants who provided biological samples and data for TOPMed. The full study-specific acknowledgments are included in Supplementary 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, 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 and the NIH Director's Early Independence Award (DP5-OD029586). We thank Ravi Majeti, Thomas Koehnke, and Benjamin Ebert for helpful discussions. The Fred Hutchinson Cooperative Center of Excellence in Hematology cell collection and processing is supported by NIDDK Grant # DK106829.

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. Whole-genome sequences are mapped to GRCh38. dbGaP accession numbers are included in the Supplementary Tables 19-20. GWAS summary statistics are deposited to dbGaP at accession phs001974. Amplicon sequencing data from WHI is deposited in dbGaP (parent study phs000200 and substudy phs003206.v1) 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/s41586-022-04785-z/MediaObjects/41586_2022_4785_MOESM9_ESM.xlsx).

CODE AVAILABILITY

https://github.com/weinstockj/longitudinal_clonal_expansion_analysis

https://github.com/weinstockj/hsc_simulation

https://github.com/weinstockj/pileup_region

https://github.com/weinstockj/passenger_count_variant_calling

https://github.com/weinstockj/PACER_analyses

https://github.com/jkgopa/HSC_TCL1A_overexpression_scRNAseq

https://github.com/weizhouUMICH/SAIGE

https://github.com/zilinli1988/SCANG

https://dockstore.org/workflows/github.com/broadinstitute/gatk/mutect2:4.1.8.1?tab=info

https://stephenslab.github.io/susieR/index.html

10.5281/zenodo.7474678

10.5281/zenodo.7474719

NHLBI TOPMed Consortium Membership

Namiko Abe[83], Gonçalo Abecasis[1], Francois Aguet[84], Christine Albert[85], Laura Almasy[86], Alvaro Alonso[87], Seth Ament[88], Peter Anderson[89], Pramod Anugu[90], Deborah Applebaum-Bowden[91], Kristin Ardlie[84], Dan Arking[92], Donna K Arnett[93], Allison Ashley-Koch[94], Stella Aslibekyan[95], Tim Assimes[96], Paul Auer[97], Dimitrios Avramopoulos[92], Najib Ayas[98], Adithya Balasubramanian[15], John Barnard[99], Kathleen Barnes[100], R. Graham Barr[101], Emily Barron-Casella[92], Lucas Barwick[102], Terri Beaty[92], Gerald Beck[103], Diane Becker[104], Lewis Becker[92], Rebecca Beer[105], Amber Beitelshees[88], Emelia Benjamin[106], Takis Benos[107], Marcos Bezerra[108], Larry Bielak[109], Joshua Bis[110], Thomas Blackwell[1], John Blangero[59], Eric Boerwinkle[111], Donald W. Bowden[112], Russell Bowler[113], Jennifer Brody[89], Ulrich Broeckel[97], Jai Broome[89], Deborah Brown[114], Karen Bunting[83], Esteban Burchard[115], Carlos Bustamante[116], Erin Buth[9], Brian Cade[117], Jonathan Cardwell[118], Vincent Carey[119], Julie Carrier[120], Cara Carty[121], Richard Casaburi[122], Juan P Casas Romero[123], James Casella[92], Peter Castaldi[22], Mark Chaffin[84], Christy Chang[88], Yi-Cheng Chang[124], Daniel Chasman[125], Sameer Chavan[118], Bo-Juen Chen[83], Wei-Min Chen[126], Yii-Der Ida Chen[127], Michael Cho[119], Seung Hoan Choi[84], Lee-Ming Chuang[128], Mina Chung[129], Ren-Hua Chung[130], Clary Clish[131], Suzy Comhair[132], Matthew Conomos[9], Elaine Cornell[133], Adolfo Correa[134], Carolyn Crandall[122], James Crapo[135], L. Adrienne Cupples[136], Joanne Curran[137], Jeffrey Curtis[109], Brian Custer[58], Coleen Damcott[88], Dawood Darbar[138], Sean David[139], Colleen Davis[89], Michelle Daya[118], Mariza de Andrade[140], Lisa de las Fuentes[141], Paul de Vries[142], Michael DeBaun[143], Ranjan Deka[144], Dawn DeMeo[119], Scott Devine[88], Huyen Dinh[15], Harsha Doddapaneni[15], Qing Duan[145], Shannon Dugan-Perez[15], Ravi Duggirala[146], Jon Peter Durda[133], Susan K. Dutcher[147], Charles Eaton[148], Lynette Ekunwe[90], Adel El Boueiz[149], Patrick Ellinor[150], Leslie Emery[89], Serpil Erzurum[99], Charles Farber[126], Jesse Farek[15], Tasha Fingerlin[151], Matthew Flickinger[1], Myriam Fornage[111], Nora Franceschini[152], Chris Frazar[89], Mao Fu[88], Stephanie M. Fullerton[89], Lucinda Fulton[153], Stacey Gabriel[84], Weiniu Gan[105], Shanshan Gao[118], Yan Gao[90], Margery Gass[154], Heather Geiger[155], Bruce Gelb[156], Mark Geraci[107], Soren Germer[83], Robert Gerszten[157], Auyon Ghosh[119], Richard Gibbs[15], Chris Gignoux[96], Mark Gladwin[107], David Glahn[158], Stephanie Gogarten[89], Da-Wei Gong[88], Harald Goring[159], Sharon Graw[100], Kathryn J. Gray[160], Daniel Grine[118], Colin Gross[1], C. Charles Gu[153], Yue Guan[88], Xiuqing Guo[127], Namrata Gupta[84], David M. Haas[161], Jeff Haessler[154], Michael Hall[162], Yi Han[15], Patrick Hanly[163], Daniel Harris[164], Nicola L. Hawley[165], Jiang He[166], Ben Heavner[9], Susan Heckbert[25], Susan Heckbert[38], Susan Heckbert[81], Ryan Hernandez[115], David Herrington[167], Craig Hersh[168], Bertha Hidalgo[95], James Hixson[111], Brian Hobbs[119], John Hokanson[118], Elliott Hong[88], Karin Hoth[169], Chao (Agnes) Hsiung[170], Jianhong Hu[15], Yi-Jen Hung[171], Haley Huston[172], Chii Min Hwu[173], Marguerite Ryan Irvin[95], Rebecca Jackson[174], Deepti Jain[89], Cashell Jaquish[105], Jill Johnsen[175], Andrew Johnson[105], Craig Johnson[89], Rich Johnston[87], Kimberly Jones[92], Hyun Min Kang[1], Robert Kaplan[176], Sharon Kardia[109], Shannon Kelly[115], Eimear Kenny[156], Michael Kessler[88], Alyna Khan[89], Ziad Khan[15], Wonji Kim[177], John Kimoff[178], Greg Kinney[179], Barbara Konkle[172], Charles Kooperberg[154], Holly Kramer[180], Christoph Lange[181], Ethan Lange[118], Leslie Lange[118], Cathy Laurie[89], Cecelia Laurie[89], Meryl LeBoff[119], Jiwon Lee[119], Sandra Lee[15], Wen-Jane Lee[173], Jonathon LeFaive[1], David Levine[89], Dan Levy[105],

Joshua Lewis[88], Xiaohui Li[127], Yun Li[145], Henry Lin[127], Honghuang Lin[182], Xihong Lin[183], Simin Liu[184], Yongmei Liu[185], Yu Liu[186], Ruth J.F. Loos[187], Steven Lubitz[150], Kathryn Lunetta[182], James Luo[105], Ulysses Magalang[188], Michael Mahaney[137], Barry Make[92], Ani Manichaikul[126], Alisa Manning[189], JoAnn Manson[119], Lisa Martin[190], Melissa Marton[155], Susan Mathai[118], Rasika Mathias[92], Susanne May[9], Patrick McArdle[88], Merry-Lynn McDonald[95], Sean McFarland[177], Stephen McGarvey[191], Daniel McGoldrick[192], Caitlin McHugh[9], Becky McNeil[193], Hao Mei[90], James Meigs[194], Vipin Menon[15], Luisa Mestroni[100], Ginger Metcalf[15], Deborah A Meyers[195], Emmanuel Mignot[196], Julie Mikulla[105], Nancy Min[90], Mollie Minear[197], Ryan L Minster[107], Braxton D. Mitchell[88], Matt Moll[22], Zeineen Momin[15], May E. Montasser[88], Courtney Montgomery[198], Donna Muzny[15], Josyf C Mychaleckyj[126], Girish Nadkarni[156], Rakhi Naik[92], Take Naseri[199], Pradeep Natarajan[84], Sergei Nekhai[200], Sarah C. Nelson[9], Bonnie Neltner[118], Caitlin Nessner[15], Deborah Nickerson[192], Osuji Nkechinyere[15], Kari North[145], Jeff O'Connell[201], Tim O'Connor[88], Heather Ochs-Balcom[202], Geoffrey Okwuonu[15], Allan Pack[203], David T. Paik[204], Nicholette Palmer[112], James Pankow[205], George Papanicolaou[105], Cora Parker[206], Gina Peloso[136], Juan Manuel Peralta[146], Marco Perez[96], James Perry[88], Ulrike Peters[207], Patricia Peyser[109], Lawrence S Phillips[87], Jacob Pleiness[1], Toni Pollin[88], Wendy Post[208], Julia Powers Becker[209], Meher Preethi Boorgula[118], Michael Preuss[156], Bruce Psaty[89], Pankaj Qasba[105], Dandi Qiao[119], Zhaohui Qin[87], Nicholas Rafaels[118], Laura Raffield[210], Mahitha Rajendran[15], Vasan S. Ramachandran[182], D.C. Rao[153], Laura Rasmussen-Torvik[211], Aakrosh Ratan[126], Susan Redline[22], Robert Reed[88], Catherine Reeves[212], Elizabeth Regan[135], Alex Reiner[213], Muagututi‘a Sefuiva Reupena[214], Ken Rice[89], Stephen Rich[126], Rebecca Robillard[215], Nicolas Robine[155], Dan Roden[216], Carolina Roselli[84], Jerome Rotter[217], Ingo Ruczinski[92], Alexi Runnels[155], Pamela Russell[118], Sarah Ruuska[172], Kathleen Ryan[88], Ester Cerdeira Sabino[218], Danish Saleheen[101], Shabnam Salimi[88], Sejal Salvi[15], Steven Salzberg[92], Kevin Sandow[219], Vijay G. Sankaran[220], Jireh Santibanez[15], Karen Schwander[153], David Schwartz[118], Frank Sciurba[107], Christine Seidman[79], Jonathan Seidman[23], Frédéric Sériès[221], Vivien Sheehan[222], Stephanie L. Sherman[223], Amol Shetty[88], Aniket Shetty[118], Wayne Hui-Heng Sheu[173], M. Benjamin Shoemaker[224], Brian Silver[225], Edwin Silverman[119], Robert Skomro[226], Albert Vernon Smith[1], Jennifer Smith[109], Josh Smith[89], Nicholas Smith[25], Nicholas Smith[38], Nicholas Smith[81], Tanja Smith[83], Sylvia Smoller[176], Beverly Snively[227], Michael Snyder[96], Tamar Sofer[119], Nona Sotoodehnia[89], Adrienne M. Stilp[89], Garrett Storm[228], Elizabeth Streeten[88], Jessica Lasky Su[119], Yun Ju Sung[153], Jody Sylvia[119], Adam Szpiro[89], Daniel Taliun[1], Hua Tang[229], Margaret Taub[92], Kent D. Taylor[230], Matthew Taylor[100], Simeon Taylor[88], Marilyn Telen[94], Timothy A. Thornton[89], Machiko Threlkeld[231], Lesley Tinker[154], David Tirschwell[89], Sarah Tishkoff[232], Hemant Tiwari[233], Catherine Tong[9], Russell Tracy[234], Michael Tsai[205], Dhananjay Vaidya[92], David Van Den Berg[235], Peter VandeHaar[1], Scott Vrieze[205], Tarik Walker[118], Robert Wallace[169], Avram Walts[118], Fei Fei Wang[89], Heming Wang[236], Jiongming Wang[1], Karol Watson[122], Jennifer Watt[15], Daniel E. Weeks[107], Joshua Weinstock[1], Bruce Weir[89], Scott T Weiss[237], Lu-Chen Weng[150], Jennifer Wessel[238], Cristen Willer[239], Kayleen Williams[9], L. Keoki Williams[240], Carla Wilson[119], James Wilson[241], Lara Winterkorn[155], Quenna Wong[89], Joseph Wu[204], Huichun Xu[88], Lisa Yanek[92], Ivana Yang[118], Ketian Yu[109], Seyedeh Maryam Zekavat[84], Yingze Zhang[242], Snow Xueyan Zhao[135], Wei Zhao[243], Xiaofeng Zhu[244], Michael Zody[83], Sebastian Zoellner[1]

83 - New York Genome Center, New York, NY, USA; 84 - Broad Institute, Cambridge, MA, USA; 85 - Cedars Sinai, Boston, MA, USA; 86 - Children's Hospital of Philadelphia, University of Pennsylvania, Philadelphia, PA, USA; 87 - Emory University, Atlanta, GA, USA; 88 - University of Maryland, Baltimore, MD, USA; 89 - University of Washington, Seattle, WA, USA; 90 - University of Mississippi, Jackson, MS, USA; 91 - National Institutes of Health, Bethesda, MD, USA; 92 - Johns Hopkins University, Baltimore, MD, USA; 93 - University of Kentucky, Lexington, KY, USA; 94 - Duke University, Durham, NC, USA; 95 - University of Alabama, Birmingham, AL, USA; 96 - Stanford University, Stanford, CA, USA; 97 - Medical College of Wisconsin, Milwaukee, WI, USA; 98 - Department of Medicine, Providence Health Care, Vancouver, CA; 99 - Cleveland Clinic, Cleveland, OH, USA; 100 - University of Colorado Anschutz Medical Campus, Aurora, CO, USA; 101 - Columbia University, New York, NY, USA; 102 - Department of LTRC, The Emmes Corporation, Rockville, MD, USA; 103 - Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland, OH, USA; 104 - Department of Medicine, Johns Hopkins University, Baltimore, MD, USA; 105 - National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD, USA; 106 - Department of Boston University School of Medicine, Boston University, Massachusetts General Hospital, Boston, MA, USA; 107 - University of Pittsburgh, Pittsburgh, PA, USA; 108 - Fundação de Hematologia e Hemoterapia de Pernambuco - Hemope, Recife, BR; 109 - University of Michigan, Ann Arbor, MI, USA; 110 - Department of Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA; 111 - University of Texas Health at Houston, Houston, TX, USA; 112 - Department of Biochemistry, Wake Forest Baptist Health, Winston-Salem, NC, USA; 113 - Department of National Jewish Health, National Jewish Health, Denver, CO, USA; 114 - Department of Pediatrics, University of Texas Health at Houston, Houston, TX, USA; 115 - University of California, San Francisco, San Francisco, CA, USA; 116 - Department of Biomedical Data Science, Stanford University, Stanford, CA, USA; 117 - Department of Brigham and Women's Hospital, Brigham & Women's Hospital, Boston, MA, USA; 118 - University of Colorado at Denver, Denver, CO, USA; 119 - Brigham & Women's Hospital, Boston, MA, USA; 120 - University of Montreal,, USA; 121 - Washington State University, Pullman, WA, USA; 122 - University of California, Los Angeles, Los Angeles, CA, USA; 123 - Brigham & Women's Hospital,, USA; 124 - National Taiwan University, Taipei, TW; 125 - Department of Division of Preventive Medicine, Brigham & Women's Hospital, Boston, MA, USA; 126 - University of Virginia, Charlottesville, VA, USA; 127 - Lundquist Institute, Torrance, CA, USA; 128 - Department of National Taiwan University Hospital, National Taiwan University, Taipei, TW; 129 - Department of Cleveland Clinic, Cleveland Clinic, Cleveland, OH, USA; 130 - National Health Research Institute Taiwan, Miaoli County, TW; 131 - Department of Metabolomics Platform, Broad Institute, Cambridge, MA, USA; 132 - Department of Immunity and Immunology, Cleveland Clinic, Cleveland, OH, USA; 133 - University of Vermont, Burlington, VT, USA; 134 - Department of Population Health Science, University of Mississippi, Jackson, MS, USA; 135 - National Jewish Health, Denver, CO, USA; 136 - Department of Biostatistics, Boston University, Boston, MA, USA; 137 - University of Texas Rio Grande Valley School of Medicine, Brownsville, TX, USA; 138 - University of Illinois at Chicago, Chicago, IL, USA; 139 - University of Chicago, Chicago, IL, USA; 140 - Department of Health Quantitative Sciences Research, Mayo Clinic, Rochester, MN,

USA; 141 - Department of Medicine, Cardiovascular Division, Washington University in St Louis, St. Louis, MO, USA; 142 - Department of Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, University of Texas Health at Houston, Houston, TX, USA; 143 - Vanderbilt University, Nashville, TN, USA; 144 - University of Cincinnati, Cincinnati, OH, USA; 145 - University of North Carolina, Chapel Hill, NC, USA; 146 - University of Texas Rio Grande Valley School of Medicine, Edinburg, TX, USA; 147 - Department of Genetics, Washington University in St Louis, St Louis, MO, USA; 148 - Brown University, Providence, RI, USA; 149 - Department of Channing Division of Network Medicine, Harvard University, Cambridge, MA, USA; 150 - Massachusetts General Hospital, Boston, MA, USA; 151 - Department of Center for Genes, Environment and Health, National Jewish Health, Denver, CO, USA; 152 - Department of Epidemiology, University of North Carolina, Chapel Hill, NC, USA; 153 - Washington University in St Louis, St Louis, MO, USA; 154 - Fred Hutchinson Cancer Research Center, Seattle, WA, USA; 155 - New York Genome Center, New York City, NY, USA; 156 - Icahn School of Medicine at Mount Sinai, New York, NY, USA; 157 - Beth Israel Deaconess Medical Center, Boston, MA, USA; 158 - Department of Psychiatry, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA; 159 - University of Texas Rio Grande Valley School of Medicine, San Antonio, TX, USA; 160 - Department of Obstetrics and Gynecology, Mass General Brigham, Boston, MA, USA; 161 - Department of OB/GYN, Indiana University, Indianapolis, IN, USA; 162 - Department of Cardiology, University of Mississippi, Jackson, MS, USA; 163 - Department of Medicine, University of Calgary, Calgary, CA; 164 - Department of Genetics, University of Maryland, Philadelphia, PA, USA; 165 - Department of Chronic Disease Epidemiology, Yale University, New Haven, CT, USA; 166 - Tulane University, New Orleans, LA, USA; 167 - Wake Forest Baptist Health, Winston-Salem, NC, USA; 168 - Department of Channing Division of Network Medicine, Brigham & Women's Hospital, Boston, MA, USA; 169 - University of Iowa, Iowa City, IA, USA; 170 - Department of Institute of Population Health Sciences, NHRI, National Health Research Institute Taiwan, Miaoli County, TW; 171 - Tri-Service General Hospital National Defense Medical Center,, TW; 172 - Blood Works Northwest, Seattle, WA, USA; 173 - Taichung Veterans General Hospital Taiwan, Taichung City, TW; 174 - Department of Internal Medicine, DIvision of Endocrinology, Diabetes and Metabolism, Oklahoma State University Medical Center, Columbus, OH, USA; 175 - Department of Research Institute, Blood Works Northwest, Seattle, WA, USA; 176 - Albert Einstein College of Medicine, New York, NY, USA; 177 - Harvard University, Cambridge, MA, USA; 178 - McGill University, Montréal, CA; 179 - Department of Epidemiology, University of Colorado at Denver, Aurora, CO, USA; 180 - Department of Public Health Sciences, Loyola University, Maywood, IL, USA; 181 - Department of Biostats, Harvard School of Public Health, Boston, MA, USA; 182 - Boston University, Boston, MA, USA; 183 - Harvard School of Public Health, Boston, MA, USA; 184 - Department of Epidemiology and Medicine, Brown University, Providence, RI, USA; 185 - Department of Cardiology, Duke University, Durham, NC, USA; 186 - Department of Cardiovascular Institute, Stanford University, Stanford, CA, USA; 187 - Department of The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 188 - Department of Division of Pulmonary, Critical Care and Sleep Medicine, Ohio State University, Columbus, OH, USA; 189 - Broad Institute, Harvard University, Massachusetts General Hospital,,; 190 - Department of cardiology, George Washington University, Washington, USA; 191 - Department of Epidemiology, Brown University, Providence, RI, USA; 192 - Department of Genome Sciences, University of Washington, Seattle, WA, USA; 193 - RTI International,, USA; 194 - Department of Medicine, Massachusetts General Hospital, Boston, MA, USA; 195 - University of Arizona, Tucson, AZ, USA; 196 - Department of Center For Sleep Sciences and Medicine, Stanford University, Palo Alto, CA, USA; 197 - National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD, USA; 198 - Department of Genes and Human Disease, Oklahoma Medical Research Foundation, Oklahoma City, OK, USA; 199 - Ministry of Health, Government of Samoa, Apia, WS; 200 - Howard University, Washington, USA; 201 - University of Maryland, Balitmore, MD, USA; 202 - University at Buffalo, Buffalo, NY, USA; 203 - Department of Division of Sleep Medicine/Department of Medicine, University of Pennsylvania, Philadelphia, PA, USA; 204 - Department of Stanford Cardiovascular Institute, Stanford University, Stanford, CA, USA; 205 - University of Minnesota, Minneapolis, MN, USA; 206 - Department of Biostatistics and Epidemiology Division, RTI International, Research Triangle Park, NC, USA; 207 - Department of Fred Hutch and UW, Fred Hutchinson Cancer Research Center, Seattle, WA, USA; 208 - Department of Cardiology/Medicine, Johns Hopkins University, Baltimore, MD, USA; 209 - Department of Medicine, University of Colorado at Denver, Denver, CO, USA; 210 - Department of Genetics, University of North Carolina, Chapel Hill, NC, USA; 211 - Northwestern University, Chicago, IL, USA; 212 - Department of New York Genome Center, New York Genome Center, New York City, NY, USA; 213 - Fred Hutchinson Cancer Research Center, University of Washington, Seattle, WA, USA; 214 - Lutia I Puava Ae Mapu I Fagalele, Apia, WS; 215 - Department of Sleep Research Unit, University of Ottawa Institute for Mental Health Research, University of Ottawa, Ottawa, CA; 216 - Department of Medicine, Pharmacology, Biomedicla Informatics, Vanderbilt University, Nashville, TN, USA; 217 - Department of Pediatrics, Lundquist Institute, Torrance, CA, USA; 218 - Department of Faculdade de Medicina, Universidade de Sao Paulo, Sao Paulo, BR; 219 - Department of TGPS, Lundquist Institute, Torrance, CA, USA; 220 - Department of Division of Hematology/Oncology, Harvard University, Boston, MA, USA; 221 - Université Laval, Quebec City, CA; 222 - Department of Pediatrics, Emory University, Atlanta, GA, USA; 223 - Department of Human Genetics, Emory University, Atlanta, GA, USA; 224 - Department of Medicine/Cardiology, Vanderbilt University, Nashville, TN, USA; 225 - UMass Memorial Medical Center, Worcester, MA, USA; 226 - University of Saskatchewan, Saskatoon, CA; 227 - Department of Biostatistical Sciences, Wake Forest Baptist Health, Winston-Salem, NC, USA; 228 - Department of Genomic Cardiology, University of Colorado at Denver, Aurora, CO, USA; 229 - Department of Genetics, Stanford University, Stanford, CA, USA; 230 - Department of Institute for Translational Genomics and Populations Sciences, Lundquist Institute, Torrance, CA, USA; 231 - Department of University of Washington, Department of Genome Sciences, University of Washington, Seattle, WA, USA; 232 - Department of Genetics, University of Pennsylvania, Philadelphia, PA, USA; 233 - Department of Biostatistics, University of Alabama, Birmingham, AL, USA; 234 - Department of Pathology & Laboratory Medicine, University of Vermont, Burlington, VT, USA; 235 - Department of USC Methylation Characterization Center, University of Southern California, University of Southern California, CA, USA;

236 - Brigham & Women's Hospital, Mass General Brigham, Boston, MA, USA; 237 - Department of Channing Division of Network Medicine, Department of Medicine, Brigham & Women's Hospital, Boston, MA, USA; 238 - Department of Epidemiology, Indiana University, Indianapolis, IN, USA; 239 - Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA; 240 - Henry Ford Health System, Detroit, MI, USA; 241 - Department of Cardiology, Beth Israel Deaconess Medical Center, Cambridge, MA, USA; 242 - Department of Medicine, University of Pittsburgh, Pittsburgh, PA, USA; 243 - Department of Epidemiology, University of Michigan, Ann Arbor, MI, USA; 244 - Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH, USA;

WORKS CITED MAIN TEXT

1. Steensma DP et al. Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. Blood 126, 9–16 (2015). [PubMed: 25931582]
2. Jaiswal S. et al. Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes A BS TR AC T. NEJM.org. N Engl J Med 26, 2488–98 (2014).
3. Genovese G. et al. Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. New England Journal of Medicine 371, 2477–2487 (2014). [PubMed: 25426838]
4. Xie M. et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. Nature Medicine 20, 1472–1478 (2014).
5. Abelson S. et al. Prediction of acute myeloid leukaemia risk in healthy individuals. Nature 559, 400–404 (2018). [PubMed: 29988082]
6. Desai P. et al. Somatic mutations precede acute myeloid leukemia years before diagnosis. Nature Medicine 24, 1015–1023 (2018).
7. Jaiswal S. et al. Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. New England Journal of Medicine (2017) doi:10.1056/NEJMoa1701719.
8. Bick Alexander G. et al. Genetic Interleukin 6 Signaling Deficiency Attenuates Cardiovascular Risk in Clonal Hematopoiesis. Circulation 141, 124–131 (2020). [PubMed: 31707836]
9. Young AL, Challen GA, Birmann BM & Druley TE Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. Nat Commun 7, 12484 (2016). [PubMed: 27546487]
10. Taliun D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature 590, 290–299 (2021). [PubMed: 33568819]
11. Bick AG et al. Inherited causes of clonal haematopoiesis in 97,691 whole genomes. Nature 586, 763–768 (2020). [PubMed: 33057201]
12. Osorio FG et al. Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. Cell Reports 25, 2308–2316.e4 (2018). [PubMed: 30485801]
13. Mitchell E. et al. Clonal dynamics of haematopoiesis across the human lifespan. Nature 606, 1–8 (2022).
14. Williams N. et al. Life histories of myeloproliferative neoplasms inferred from phylogenies. Nature 602, 162–168 (2022). [PubMed: 35058638]
15. Lee-Six H. et al. Population dynamics of normal human blood inferred from somatic mutations. Nature 561, 473–478 (2018). [PubMed: 30185910]
16. Fabre MA et al. The longitudinal dynamics and natural history of clonal haematopoiesis. Nature 606, 335–342 (2022). [PubMed: 35650444]
17. Cibulskis K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature Biotechnology 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). [PubMed: 28483762]

19. Watson CJ et al. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. Science 367, 1449–1454 (2020). [PubMed: 32217721]

20. van Deuren RC et al. Clone expansion of mutation-driven clonal hematopoiesis is associated with aging and metabolic dysfunction in individuals with obesity. 2021.05.12.443095 https://www.biorxiv.org/content/10.1101/2021.05.12.443095v1 (2021) doi:10.1101/2021.05.12.443095.

21. Robertson NA et al. Longitudinal dynamics of clonal hematopoiesis identifies gene-specific fitness effects. 2021.05.27.446006 https://www.biorxiv.org/content/10.1101/2021.05.27.446006v3 (2021) doi:10.1101/2021.05.27.446006.

22. van Zeventer IA et al. Mutational spectrum and dynamics of clonal hematopoiesis in anemia of older individuals. Blood 135, 1161–1170 (2020). [PubMed: 32243522]

23. Dr Z, Sp W, N J, T J & Pr F The ensembl regulatory build. Genome Biol 16, 56–56 (2015). [PubMed: 25887522]

24. Carvalho-Silva D. et al. Open Targets Platform: new developments and updates two years on. Nucleic Acids Res 47, D1056–D1065 (2019). [PubMed: 30462303]

25. Narducci MG et al. TCL1 Is Overexpressed in Patients Affected by Adult T-Cell Leukemias. Cancer Res 57, 5452–5456 (1997). [PubMed: 9407948]

26. Fishilevich S. et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database (Oxford) 2017, (2017).

27. Thompson DJ et al. Genetic predisposition to mosaic Y chromosome loss in blood. Nature 575, 652–657 (2019). [PubMed: 31748747]

28. Malcovati L. et al. Clinical significance of somatic mutation in unexplained blood cytopenia. Blood 129, 3371–3378 (2017). [PubMed: 28424163]

29. Consortium, T. Gte. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 369, 1318–1330 (2020). [PubMed: 32913098]

30. Giambartolomei C. et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLOS Genetics 10, e1004383 (2014). [PubMed: 24830394]

31. Regev A. et al. The Human Cell Atlas. eLife 6, e27041 (2017). [PubMed: 29206104]

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). [PubMed: 33649320]

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). [PubMed: 32386542]

34. Corces MR et al. The chromatin accessibility landscape of primary human cancers. Science 362, eaav1898 (2018). [PubMed: 30361341]

35. Pietras EM 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). [PubMed: 26095048]

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). [PubMed: 24658644]

37. Laine J, Künstle G, Obata T, Sha M & Noguchi M The Protooncogene TCL1 Is an Akt Kinase Coactivator. Molecular Cell 6, 395–407 (2000). [PubMed: 10983986]

38. Brunet A. et al. Akt promotes cell survival by phosphorylating and inhibiting a Forkhead transcription factor. Cell 96, 857–868 (1999). [PubMed: 10102273]

39. Eijkelenboom A & Burgering BMT FOXOs: signalling integrators for homeostasis maintenance. Nat Rev Mol Cell Biol 14, 83–97 (2013). [PubMed: 23325358]

40. Kakiuchi N & Ogawa S Clonal expansion in non-cancer tissues. Nature Reviews Cancer 21, 239–256 (2021). [PubMed: 33627798]

41. Martincorena I. et al. High burden and pervasive positive selection of somatic mutations in normal human skin. Science 348, 880–886 (2015). [PubMed: 25999502]

42. Martincorena I. et al. Somatic mutant clones colonize the human esophagus with age. Science 362, 911–917 (2018). [PubMed: 30337457]

WORKS CITED METHODS

43. Regier AA et al. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. Nature Communications 9, 1–8 (2018).
44. Jun G, Wing MK, Abecasis GR & Kang HM An efficient and scalable analysis framework for variant extraction and refinement from population scale DNA sequence data. Genome Res. gr.176552.114 (2015) doi:10.1101/gr.176552.114.
45. 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). [PubMed: 22728672]
46. Voss K, Gentry J & Van der Auwera G Full-stack genomics pipelining with GATK4 + WDL + Cromwell. in (F1000 Research, 2017). doi:10.7490/f1000research.1114631.1.
47. Beauchamp EM et al. ZBTB33 Is Mutated in Clonal Hematopoiesis and Myelodysplastic Syndromes and Impacts RNA Splicing. Blood Cancer Discov (2021) doi:10.1158/2643-3230.BCD-20-0224.
48. Miller CA et al. Failure to detect mutations in U2AF1 due to changes in the GRCh38 reference sequence. 2021.05.07.442430 www.biorxiv.org/content/10.1101/2021.05.07.442430v1 (2021) doi:10.1101/2021.05.07.442430.
49. Pedersen BS & Quinlan AR cyvcf2: fast, flexible variant analysis with Python. Bioinformatics 33, 1867–1869 (2017). [PubMed: 28165109]
50. VariantKey: A Reversible Numerical Representation of Human Genetic Variants | bioRxiv. https://www.biorxiv.org/content/10.1101/473744v3.
51. Alexandrov LB et al. Clock-like mutational processes in human somatic cells. Nature Genetics 47, 1402–1407 (2015). [PubMed: 26551669]
52. Stan Development Team. Stan Modeling Language Users Guide and Reference Manual, 2.17. (2020).
53. Stan Development Team. RStan: The R interface to Stan. (2020).
54. Bezanson J, Edelman A, Karpinski S & Shah VB Julia: A fresh approach to numerical computing. (2017).
55. Venables WN & Ripley BD Modern Applied Statistics with S. (Springer-Verlag, 2002). doi:10.1007/978-0-387-21706-2.
56. Hiatt JB, Pritchard CC, Salipante SJ, O'Roak BJ & Shendure J Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. Genome Res. 23, 843–854 (2013). [PubMed: 23382536]
57. mimips. (2020).
58. Koboldt DC et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 22, 568–576 (2012). [PubMed: 22300766]
59. Robinson JT et al. Integrative genomics viewer. Nature Biotechnology 29, 24–26 (2011).
60. Uddin MM et al. Longitudinal profiling of clonal hematopoiesis provides insight into clonal dynamics. Immun Ageing 19, 23 (2022). [PubMed: 35610705]
61. Zhou W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nature Genetics 50, 1335–1341 (2018). [PubMed: 30104761]
62. Ma C, Blackwell T, Boehnke M & Scott LJ Recommended Joint and Meta-Analysis Strategies for Case-Control Association Testing of Single Low-Count Variants. Genetic Epidemiology 37, 539–550 (2013). [PubMed: 23788246]
63. Wang G, Sarkar A, Carbonetto P & Stephens M A simple new approach to variable selection in regression, with application to genetic fine mapping. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82, 1273–1300 (2020). [PubMed: 37220626]
64. Li Z. et al. Dynamic Scan Procedure for Detecting Rare-Variant Association Regions in Whole-Genome Sequencing Studies. The American Journal of Human Genetics 104, 802–814 (2019). [PubMed: 30982610]
65. Bates D. et al. Matrix: Sparse and Dense Matrix Classes and Methods. (2019).
66. R Core Team. R: A Language and environment for statistical computing. (2020).

67. Gogarten SM et al. Genetic association testing using the GENESIS R/Bioconductor package. Bioinformatics 35, 5346–5348 (2019). [PubMed: 31329242]

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). [PubMed: 29608179]

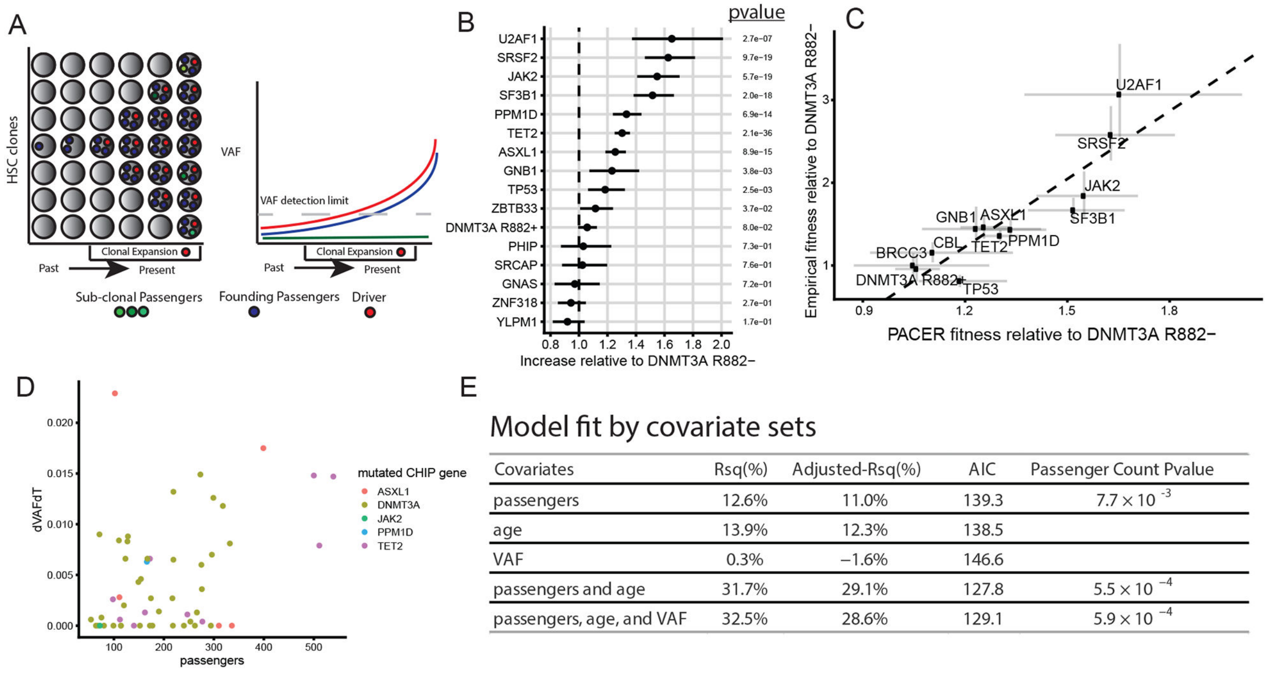
69. Dataset for the manuscript 'Identification of leukemic and pre-leukemic stem cells by clonal tracking from single-cell transcriptomics'. (2021) doi:10.6084/m9.figshare.12382685.v1.

70. Omni-ATAC-seq: Improved ATAC-seq protocol. https://www.researchsquare.com (2017) doi:10.1038/protex.2017.096.

71. Love MI, Huber W & Anders S Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15, 550 (2014). [PubMed: 25516281]

72. Kim D, Paggi JM, Park C, Bennett C & Salzberg SL Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37, 907–915 (2019). [PubMed: 31375807]

73. Dobin A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013). [PubMed: 23104886]



Model fit by covariate sets

Covariates	Rsq(%)	Adjusted-Rsq(%)	AIC	Passenger Count Pvalue
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 hematopoietic stem cell (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. Unadjusted, two-sided p-values are reported. Error bars indicate 95 percent confidence intervals. **C,** The relative abundances of passenger counts are plotted against the empirical estimates of gene fitness derived from the longitudinal deep sequencing in Fabre et al.[16]. Error bars indicate 95 percent confidence intervals. The estimate of the association from weighted least squares (slope = 2.7, p-value = 9.6 x 10^{-5}, R^2 = 80%) is plotted as a dashed line. **D**, The observed clonal expansion rates (dVAFdT), as expressed in the change in variant allele frequency (VAF) over time (years), were associated with increased passenger counts in 55 CHIP carriers from the Women's Health Initiative. Colors 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 is Akaike information criteria, where smaller values indicate better model fit. Unadjusted, two-sided p-values are reported for the passengers variable in the respective models.

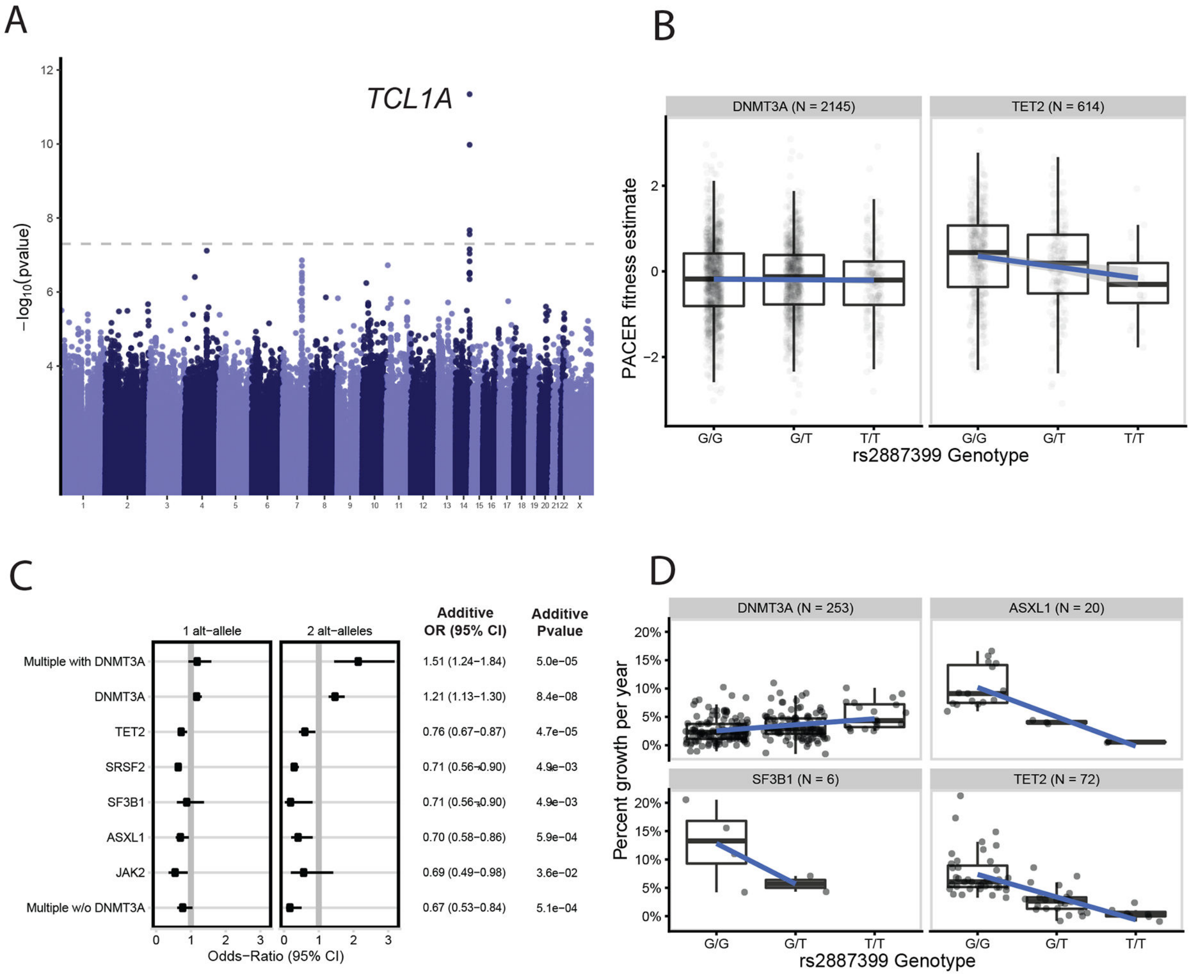
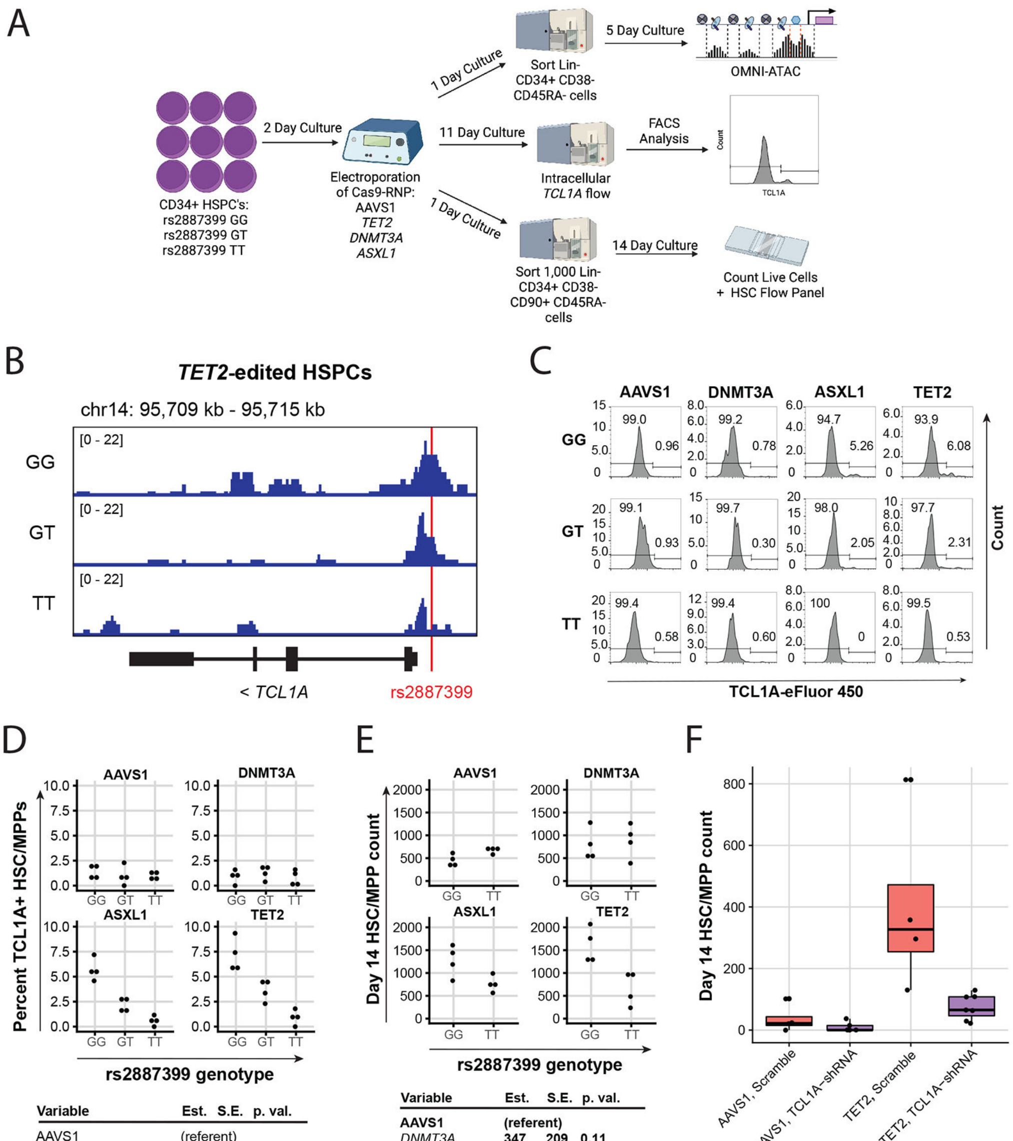


Fig 2|. GWAS of PACER Identifies Germline Determinants of Clonal Expansion in Blood

A, A genome-wide association study (GWAS) of passenger counts identifies *TCL1A* as a genome-wide significant locus. Test statistics were estimated with SAIGE[61]. **B**, The association between the genotypes of rs2887399 and PACER varied between *TET2* and *DNMT3A*. Alt-alleles were associated with decreased PACER score in *TET2* mutation carriers, but no association was observed in *DNMT3A* carriers. **C**, The association between alt-alleles at rs2887399 and presence of specific CHIP mutations varies by CHIP mutations (n = 5,071 CHIP carriers). Forest plot shows the odds ratios for having specific mutations in those carrying a single T-allele and two T-alleles, respectively. Odds ratios were estimated using Firth logistic regression, with error bars representing 95 percent confidence intervals. On the right of the forest plot, effect estimates and p-values are included from SAIGE, which uses an additive coding of the alt-alleles for hypothesis testing and uses a generalized linear mixed model to estimate test statistics. Unadjusted, two-sided p-values are reported. In the additive tests, *SF3B1* and *SRSF2* were grouped together to aid convergence. **D,** The association between the genotypes of rs2887399 and percent growth per year of CHIP clones from 351 carriers in the Women's Health Initiative. Percent growth per year is estimated using a Bayesian logistic growth model of clonal expansion.

For box and whisker plots in 2b and 2d, the horizontal line indicates the median, the tops and bottoms of the boxes indicate the interquartile range, and top and bottom error bars indicate maxima and minima, respectively.



Variable	Est.	S.E.	p. val.
AAVS1	(referent)		
DNMT3A	-0.24	0.57	0.67
ASXL1	4.07	0.57	1.1×10^{-8}
TET2	5.69	0.57	1.8×10^{-12}
Per T-allele	-0.19	0.31	0.55
DNMT3A:Per T-allele	0.12	0.44	0.79
ASXL1:Per T-allele	-2.36	0.44	3.6×10^{-6}
TET2:Per T-allele	-2.91	0.44	6.4×10^{-8}

Variable	Est.	S.E.	p. val.
AAVS1	**(referent)**		
DNMT3A	**347**	**209**	**0.11**
ASXL1	**818**	**209**	**6.5×10^{-4}**
TET2	**1155**	**209**	**1.1×10^{-5}**
rs2887399 GG	**(referent)**		
rs2887399 TT	**229**	**209**	**0.28**
DNMT3A:**TT**	**-145**	**296**	**0.63**
ASXL1:**TT**	**-735**	**296**	**0.020**
TET2:**TT**	**-1170**	**296**	**5.9×10^{-4}**

Variable	Est.	S.E.	p. val.
AAVS1	(referent)		
TET2	363	92.4	1.2×10^{-3}
Scramble	(referent)		
TCL1A-shRNA	-26.1	87.7	0.77
TET2:*TCL1A*-shRNA	-298	120	0.02

Figure 3|. Effect of rs2887399 on TCL1A Expression and Clonal Expansion

A, Schematic of experimental workflow. **B**, ATAC-sequencing tracks illustrating chromatin accessibility at rs2887399 in *TET2*-edited HSPCs from donors of the GG, GT, and TT genotypes after 5 days liquid culture. Red line indicates location of rs2887399. See also Extended Data Figure 8 and Table S14. **C**, Percent Lin− CD34+ CD38− CD45RA− cells expressing TCL1A by flow cytometry after 11 days liquid culture of edited HSPCs, stratified by edited gene and rs2887399 genotype. Results of a linear regression model for the effect of edited gene (referent to AAVS1), number of T-alleles at rs2887399, and the interaction term of edited gene with T-alleles are presented below. Est. = estimate, S.E. = standard error, p. val. = p-value. Unadjusted p-values from a two-sided test are reported. n=4 biologically independent replicates for each group. **D**, Lin− CD34+ CD38− CD45RA− cell counts after

14 days liquid culture of edited HSCs. 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 a two-sided test are reported. n=4 biologically independent replicates for each group. **E,** Lin− CD34+ CD38− CD45RA− cell counts after 14 days liquid culture of edited and shRNA transduced HSCs. Results of a linear regression model for the effect of edited gene (referent to AAVS1), shRNA (referent to scramble control), and the interaction term of edited gene with shRNA are presented below. Unadjusted p-values from a two-sided test are reported. 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. n=4 for AAVS1 gRNA/scramble, n=5 for AAVS1 gRNA/*TCL1A* shRNA, n=4 for *TET2* gRNA/scramble, and n=7 for *TET2* gRNA/*TCL1A* shRNA, which represent biologically independent replicates.

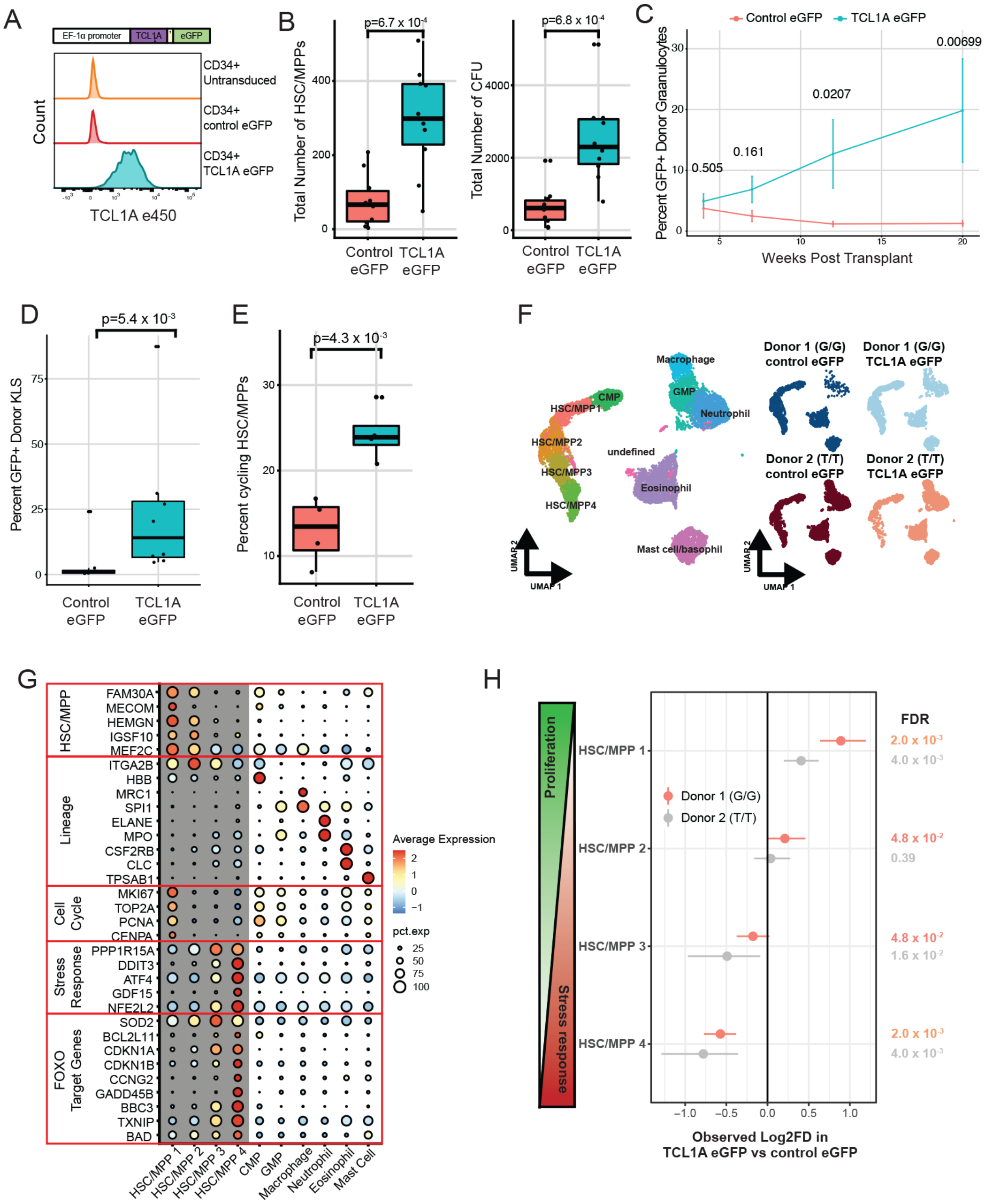


Figure 4|. *TCL1A* Expression is Sufficient for HSC Expansion

A, Schematic of *TCL1A*-eGFP lentivirus construct (top) and effect of viral transduction on *TCL1A* expression in human CD34+ HSPCs (bottom). **B,** Lin−CD34+CD38−CD45RA− cell counts after 14 days liquid culture of transduced HSCs (left), and quantification of colony forming units in methylcellulose after 14 days of liquid culture of transduced HSCs (right); p-values were estimated using a 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. Shown are mean percent GFP+ donor granulocytes and standard errors for each time point. Hypothesis testing was performed using two-sided Wilcoxon rank sum tests and p-values are indicated above each timepoint. n=8 mice for each group. **D**, Percent GFP+ donor cells in Lin− c-Kit+ Sca-1+ (KLS) marrow at 22 weeks post-transplant. P-value obtained from a two-sided Wilcoxon rank sum test. n=8 mice for each group. **E,** Percent Lin−CD34+CD38− cells in cycle by DAPI staining after 10 days liquid culture of transduced HSC/MPPs; p-values were calculated using a two-sided Wilcoxon rank sum test. n=4 biologically independent replicates for each group. **F,** UMAP of clusters identified after 7 days liquid culture of transduced HSC/MPPs; all samples combined (left) and split by the 4 individual samples (right). G/G or T/T refers to the donor rs2887399 genotype. **G,** Dot plot illustrating expression of representative marker genes across different cell clusters arranged by functional group. **H,** Forest plot of log2 fold-difference (Log2FD) in proportion of cells within each HSC/MPP cluster in TCL1A-eGFP versus control-eGFP transduced cells using a permutation test. Each donor represents an independent experiment and the false discovery rate (FDR) for each comparison is shown to the right.

For box and whisker plots in 4b, 4d, and 4e, horizontal lines indicate the median, the tops and bottoms of the boxes indicate the interquartile range, and top and bottom error bars indicate maxima and minima, respectively.