nature portfolio

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

Supplementary information

Aberrant activation of TCL1A promotes stem cell expansion in clonal haematopoiesis

In the format provided by the authors and unedited

Supplementary Guide

Supplementary Note 1: PACER Simulations and Methodology

Table of Contents

PACER Simulation Parameters

We assume that the accumulation of passenger mutations is described by a Poisson birth-death stochastic process. As the birth and death rates scale with the number of HSCs, we assume a linear birth-death process.

We assume that the birth rate for a given hematopoietic stem cell (HSC) i at time t with fitness $s_i(t)$) is $\lambda_i(t) \sim Poisson(\omega * X_i(t) * (1 + s_i(t)) * dt)$, where dt represents the amount of time in years, and ω represents the number of stem cell divisions per year. We assume that the death rate can be described as

$$\psi_i(t) \sim Poisson(\omega * X_i(t) * (1 - s_i(t)) * dt)$$

. The death rate is the rate at which an HSC divides into two differentiated cells, and the birth rate is the rate at which an HSC divides into two HSCs. We don't consider asymmetric HSC differentiation as this would not change the clone size. The HSC clone cell count is defined as $X_i(t) = \sum_{l \leq t} \lambda_i\,(l) - \psi_i(l)$, and the HSC clone size (a fraction of the total cell population) is $VAF_i(t) = \frac{X_i(t)}{\sum_j X_j(t)}$.

We start with 500 HSC clones, each with 200 identical cells in each clone $X_i(t = 0) = 200$. Each cell divides once every three years (= 1/3), and each clone with an initial $s_i(t = 0) = 0$. At each iteration, we also center the $s_i(t)$ such that $\overline{s_i(t)} = 0$. This means that there are 100,000 total HSCs at the start of the simulation.

For each clone, we set the passenger mutation rate:

1. μ_p, the passenger accumulation rate, $A_i(t) \sim Poisson(X_i(t) * \mu_p * dt)$

Where $A_i(t)$ is the number of passengers accumulated in a given clone through time t. We set $\mu_p = 0.006$, which is the passenger mutation rate of a diploid genome for a single HSC per year. This implies a mutation rate of 6 passengers per year for a clone with 1000 cells, and a mutation rate of 600 passengers per year across the entire population of 100,000 HSCs. We will later consider the effects of an insensitive sequencing assay that captures a small fraction of the passengers.

We assign a single driver to one of the HSC clones, which is randomly selected among the HSC clones. The time of acquisition is uniformly drawn from each cell division after 10 years, such that are driver is equally likely to be acquired at either 10 years or 78 years. We simulate the HSC population across a lifetime of 90 years. We refer to the time of driver acquisition as T_d.

We assume that each HSC clone can at most acquire a single driver, which represents a similar HSC population to the TOPMed CHIP driver carriers.

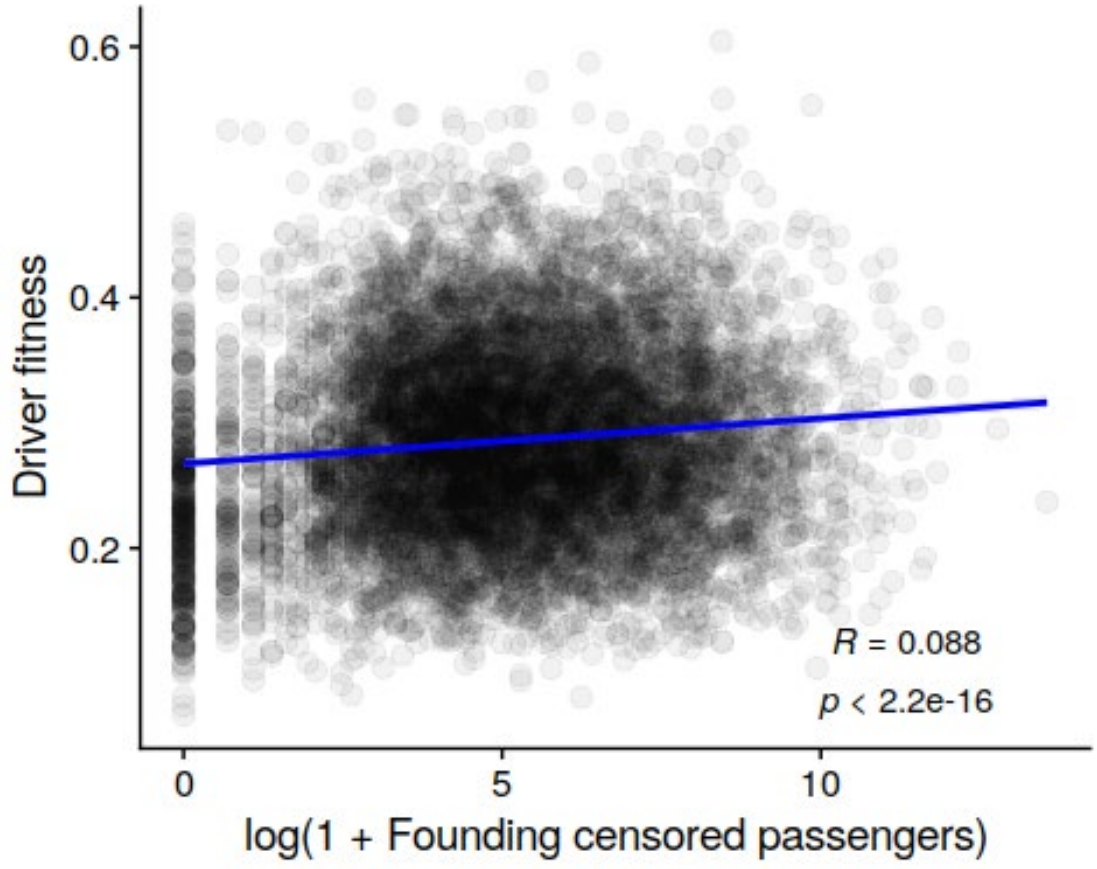
If an HSC clone i acquires a driver at time t, we set $s_i(t) = Beta(4,16)$. A $Beta(4,16)$ random variable is bounded between 0 and 1 and has an expectation of 0.20. An HSC with $_i(t) = 0.20$ will self-renew 60% of the time, and terminally differentiate 40% of the time.

For a given HSC population, we simulate 90 years, and track the accumulation of passengers and drivers. To incorporate the censoring from using 38x sequencing coverage, we simulate whether a given passenger would be observed at 38x coverage by sampling the number of alt-reads from $R \sim Binomial(38, VAF_i(t))$ and comparing $R \geq 2$, since two reads are required by our variant calling process. We refer to $P(R \geq 2|VAF = vaf) = P(Binomial(38, vaf) \geq 2)$ and we refer to the passengers that would be detected at 38x coverage as the censored founding passengers, $AC_i(t)$, where $t = T_d$.

We ran the simulation 10,000 times, where at most a single HSC clone acquires a driver mutation. We then compared $AC_i(T_d)$ to the fitness of the clone at the end of each simulation.

PACER Simulation Results

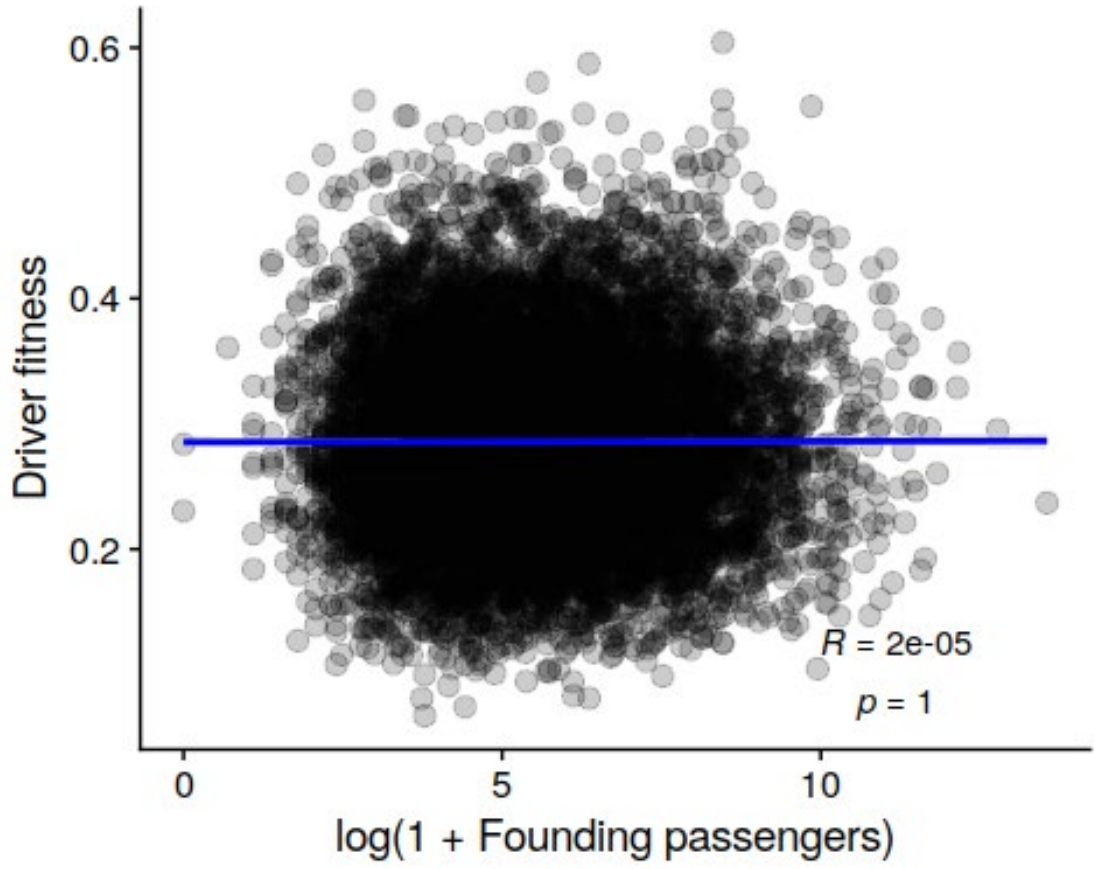
Supplementary Figure S6



We observed a modest concordance between the founding censored passengers with the fitness of each HSC at the end of the simulation (spearman = 0.09, pvalue < 2.2e-16, Supplementary Figure S6). This suggests that the censored founding passengers are proportional to the fitness of the driver mutations. Stochastic drift of the HSC clone sizes contributes substantial residual variance.

Supplementary Figure S7

We observe no concordance between the uncensored founding passengers with fitness (Supplementary Figure S7). This suggests that if PACER were applied to WGS with extraordinarily deep coverage, the mutation count may no longer associate with driver fitness (without further adjustments).



PACER-HB Methodology

PACER-HB (PACER-Hierarchical Bayes) is described in the Methods section, but for convenience we also describe the model here. PACER-HB is a Bayesian Hierarchical model of the observed passenger counts, which are censored due to our read depth (an average of 38x).

To motivate our specification of the expected number of observed passengers, we used the following theoretical formulation. We define the birth-death stochastic process governing the number of HSC cells in a clone as:

$$E(X(t)) = X(0)e^{(\lambda-\psi)*t}$$

where λ represents the "birth" rate and ψ represents the "death" rate. In our case, we set $\lambda - \psi = \omega * X(0) * 2s$, where ω is the number of stem cell divisions per year.

We can approximate the number of the number of cells in a given clone as:

$$E(X_i(t)) \simeq \sum_{z=10}^{t} P\,(T_d <= z) * E(X_i(t; s > 0)) + P(T_d > z) * E(X_i(t; s = 0))$$

Which assumes that mutations can be acquired at any time *z* between 10 years old and time t. If, for simplicity, we assume that the time of driver acquisition is distributed uniformly over this range,

$$E(X_i(t)) \simeq \sum_{z=10}^{t} \frac{z-10}{t-10} * X_i(0) e^{\omega * X(0) * 2s * (t-z)} + \frac{t-z}{t-10} * X_i(0)$$

Next, we consider how to relate the number of HSC cells in a clone to the number of uncensored passengers: $A_i(t) \sim Poisson\big(X_i(t) * \mu_p * dt\big)$, where μ_p is the passenger accumulation rate.

We denote the number of censored passengers at time t as $AC_i(t)$:

$$\begin{aligned} E\big(AC_i(t)\big) &= \int_{vaf=0}^{0.5} a\, c_i(t) * P\big(AC_i(t) = ac_i(t) \big| VAF_i(t) = vaf_i(t)\big) \\ &\quad * P\big(VAF_i(T) = vaf_i(t)\big) dvaf_i(t) \\ &= \int_{vaf=0}^{0.5} E\left(AC_i(t) \big| VAF_i(t) = vaf_i(t)\right) * P\big(VAF_i(T) = vaf_i(t)\big) dvaf_i(t) \end{aligned}$$

Where we are integrating the number of censored passengers conditional on a given VAF across the VAF distribution at time t.

This conditional expectation can be approximated as the product of the number of uncensored passengers, $A_i(t)$ and the probability that a given passenger is observed conditional on the VAF:

$$E\left(AC_i(t)\middle|VAF_i(t) = vaf_i(t)\right) \simeq E\left(A_i(t)\right) * P\left(R \geq 2\middle|VAF = vaf_i(t)\right) = E\left(X_i(t)\right) * \mu_p * P(R \geq 2|VAF = vaf_i(t))$$

The latter term, $P(R \geq 2\ |VAF = vaf_i(t))$ clearly is proportional to $s_i(t)$, as more fit clones will comprise a greater fraction of peripheral blood cells. Thus, we observe that $AC_i(t)$ is proportional both to the time of driver acquisition and fitness of the driver mutation.

Thus, we modeled the expectation of the observed censored passenger counts in CHIP carriers with a single driver as simple linear function of these quantities: $E\left(AC_i(t)\right) = \mu_p T_i + s_i(t_i - T_i)$, where T_i denotes the time of driver acquisition. Although a multiplicative relationship is likely more realistic, the additive model is somewhat more tractable computationally.

We assume that the expectation of the observed passenger counts is linear with respect to three terms: 1. The number of ancestral passenger mutations 2. A fitness term of the driver, which affects the number of observed passengers through increasing the likelihood of detection of mutations 3. a study specific batch correction. In addition, we assume heterogenous overdispersion between the CHIP carriers and controls, and we assume a negative binomial distribution over the observed passenger counts:

$$E\left(\mathrm{AC}_i(t_i)\right) = \left(\mu T_i + s_i(t_i - T_i)\right) I(i \in CHIP) + \alpha_k$$

$$AC_i(t_i) \sim NB(E(counts_i(t_i)), I(i \in CHIP)\theta_0 + \left(1 - I(i \in CHIP)\right)\theta_1)$$

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 define $CHIP$ as the set of CHIP carriers. We can interpret $t_i - T_i$ as the lifetime of the clone in years. We constrain T_i such that $0 \leq T_i \leq t_i$ by parametrizing $T_i = \psi_i * age_i$, with ψ_i assumed to be a Beta random variable with support between 0 and 1. T_i is interpreted as the expected time of driver acquisition.

We place the following priors on these parameters:

$$\psi_i \sim Beta(1.0, 1.3)$$

$$\alpha_k \sim N(0, 20)$$

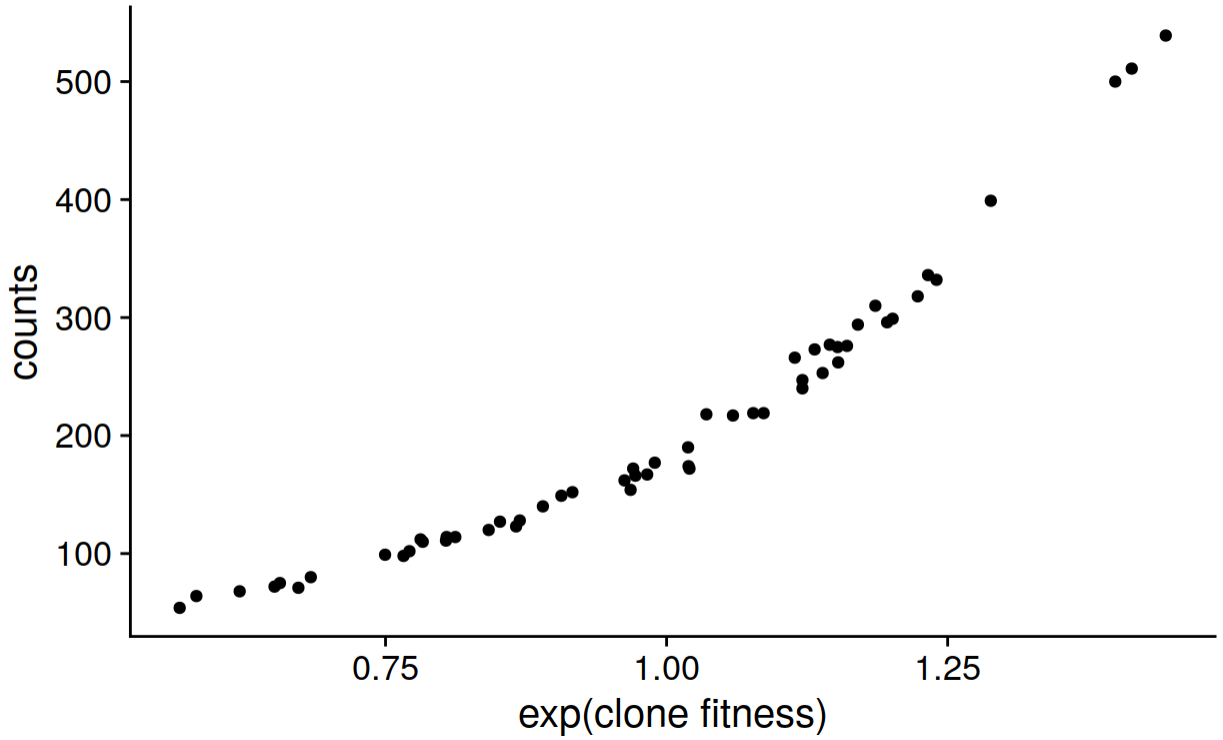
$$s_i \sim N(0, 1)$$

This model is implemented in the Stan probabilistic programming language, and the posterior is estimated using the default MCMC sampler with 400 samples for burn-in. Convergence was assessed using Rhat statistics and effective sample size calculation.

We then asked the extent to which PACER-HB fitness estimates correlated with the passenger counts when applying these two models to the WHI CHIP carriers (described in the

main text). We observed very strong concordance between the PACER-HB fitness estimates with the observed passenger counts, indicating strong concordance between PACER-HB and PACER (Supplementary Figure S8).

Supplementary Figure S8



PACER Methodology

We next define the definition of PACER and describe its interpretation. We begin with the observed passenger counts $AC_i(t_i)$ measured in individual i at blood draw t_i. Next, the observed passenger counts are normalized using an inverse quantile transformation, which is a rank transformation that assumes the quantiles are drawn from a standard gaussian. If a rate of artifacts α has been estimated – for example, using passenger counts in individuals who are known *apriori* to not have clonal hematopoiesis – this should be subtracted from the observed passenger counts prior to normalization. In addition – if sensitivity to detect passengers is heterogenous across individuals, $AC_i(t_i)$ should be adjusted for varying sensitivity. We again emphasize based on Supplementary Text Supplementary Figure 7 that the passenger counts should be identified using a sequencing assay that only detects mutation in >5-10% of blood cells.

We then fit the following linear model to the normalized passenger counts:

$$\widetilde{AC_i(t_i)} \sim N(X\beta, \sigma^2)$$

Where X is a matrix of covariates including the age at blood draw, the variant allele fraction of the driver variant, and fixed effects which describe any variation in batch. If loss of heterozygosity has been detected in the CHIP driver mutation, the driver VAF should be corrected.

We then define PACER fitness $\widehat{s_i(t_i)} = \widehat{AC_i(t_i)} - X\hat{\beta}$, such that fitness is defined as the excess passenger counts relative to our expectation. These estimates are proportional to the yearly growth rate of a given clone, but the PACER estimates do not have these units. Rather, within a given cohort, the fitness estimates should be interpreted as ranking the fitness of different CHIP clones. We remark that for downstream association analyses, quantile transformation of the fitness estimates is likely useful. We also remark that the fitness estimates are relative estimates within a cohort, as they are mean centered and standardized, and thus cannot be compared across cohorts without fitting the model jointly.

PACER-Gene Methodology

We next define the application of PACER for inference of gene-level fitness estimates. We recommend specification of a "reference" gene such that passenger count abundances can be interpreted relative to a reference gene or mutation. We fit the following negative binomial regression:

$$g\left(E\left(AC_i(t_i)\right)\right) = X\beta$$

$$AC_i(t_i) \sim NegativeBinomial\left(g^{-1}\left(E\left(AC_i(t_i)\right)\right), \phi\right)$$

Where ϕ is an over-dispersion term, and g is the log function. The same covariates as before should be included in X, in addition to an intercept for each gene of interest. The gene level coefficients, once exponentiated, can be interpreted as the relative increase in passenger counts for a given gene relative to the reference gene after adjustment for the other included covariates.

Supplementary Note 2

Re-analysis of deep longitudinal sequencing of clonal hematopoiesis carriers contextualizes prediction of future clone growth

As described in Results, the PACER model for prediction of future clone growth in WHI (dVAF/dT regressed onto inverse-normal transformed passenger count, age at blood draw, sex, and VAF) had an R^2= 32.5% (95% CI 15.0-56.3%). To contextualize PACER's performance as an estimator of future clonal expansion, we compared its performance to estimators derived from longitudinal sequencing. A cohort of 305 donors with clonal hematopoiesis mutations was recently described in Fabre et al., 2021. Each donor was assayed up to 6 times over a median of 13 years. Given the size of the cohort, sequencing depth, and longitudinal assessment, these data are substantially richer than the data used to develop PACER (a single blood draw sequenced to 38x). Despite these differences in data quality, we observe that predictors of clonal expansion derived from the Fabre et al. data perform similarly to PACER.

Out-of-sample performance of the Fabre. et al. fitness estimator

We considered three separate approaches using the Fabre et al. data to assess prediction performance of clonal expansion, including two separate statistical methods and two filtering criteria. Firstly, we used the 11 donors that were included in the original Fabre et al. analysis as a separate validation cohort for the assessment of their clonal expansion predictions. To predict clonal expansion of the 15 mutations in these 11 donors, we then used the posterior mean fitness estimates from the Fabre et al. hierarchical Bayesian model, which we will refer to as the FHBM. FHBM is a mixed effects beta-binomial model of the number of mutated allele reads $counts_{ij}(t)$ from the ith mutation of the jth individual at time t such that:

$$counts_{ij}(t) \sim BetaBinomial\big(coverage_{ij}(t), \alpha(t), \beta\big),$$

where $\alpha(t) = \frac{\beta q(t)}{1-q(t)}$ and $q(t) = \sigma\left(\left(b_{gene_i} + b_{site_i} + b_{c_i}\right) * \mathrm{t} + \mathrm{u_{ij}}\right)$, and $\sigma(x) = 1/\big(1 + exp(-x)\big)$.

We define $FHBM = E\left(\pi\big(b_{gene_i} + b_{site_i} + b_{c_i} | X\big)\right)$, where π represents the posterior of the model.

To estimate the uncertainty associated with the prediction performance of FHBM, we computed the Rsquared between FHBM predictions with the inverse normal transformed dVAFdT values across 1,000 bootstraps. We observed that the FHBM had a similar association (R^2= 36.8%, 95% CI: [4.4%, 73.6%], Supplementary Note 2 Figure 1, Supplementary Note 2 Table 1) with future clone growth as PACER did in the TOPMed cohort (R^2= 32.5%). Due to the limited number of donors, these point estimates have substantial uncertainty. Furthermore, these 11 donors were highly enriched for splicing factor mutations, which exhibited more stable growth than other clones in Fabre. Thus, it is expected that the predictive ability in the dataset as a whole would be lower than in these selected individuals, as we describe below.

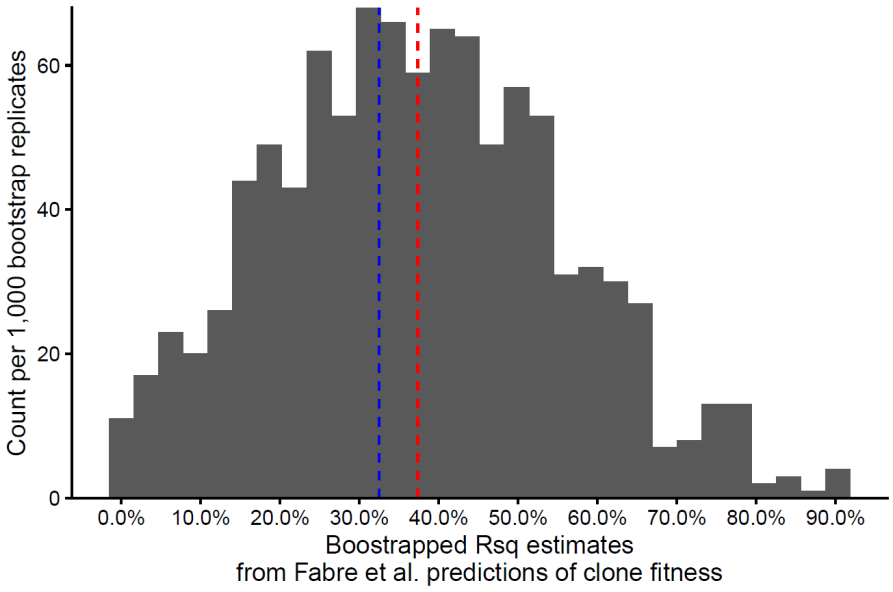


Figure 1: Bootstrapped Rsq estimates of the FBHM estimator of future clone growth. The histogram is comprised of 1,000 Rsq estimates from 1,000 bootstrap resamples of the data. The blue line, 32.5%, is the Rsq between PACER and future clone growth on the WHI samples included in the main text. The red line is the mean Rsq estimate from the FBHM estimator.

Development of an expanded validation cohort from the Fabre. et al data

Due to the limited sample size of the original Fabre et al. validation set, we developed an expanded validation sample derived from the same data. Among the 385 donors in the data, we filtered to those with at least three separate mutational assessments. As the criteria for clonal hematopoiesis mutations varies between the Fabre et al. mutation calls and our TOPMed analysis, we then filtered the mutations present in the Fabre et al. data to a high confidence subset (Beauchamp et al., Blood Cancer Discovery 2021). As PACER is developed for application to donors with a single mutation, we then filtered the Fabre et al. data to donors with a single mutation. Finally, we separated the last mutational assessment from each donor from the previous timepoints (1st through penultimate) to form training (102 donors and 316 mutations) and testing strata (102 donors and 102 mutations). We then estimated mutation fitness using our own implementation of the FHBM on the training set. We estimated mutation fitness using FBHM on the training data, and then estimated its association with inverse normal transformed dVAFdT, where dVAFdT is computed as $\left(\frac{VAF_2 - VAF_1}{t_2 - t_1}\right)$ between the last and penultimate timepoints. This is an analogous analysis to what we performed to evaluate PACER on the TOPMed data. On this greatly expanded validation set, we observed that the FHBM estimator was only modestly associated with future clone growth (R^2= 4.0%, 95% CI: [0.03%, 13.3%], Supplementary Note 2 Figure 2, Supplementary Note 2 Table 1).

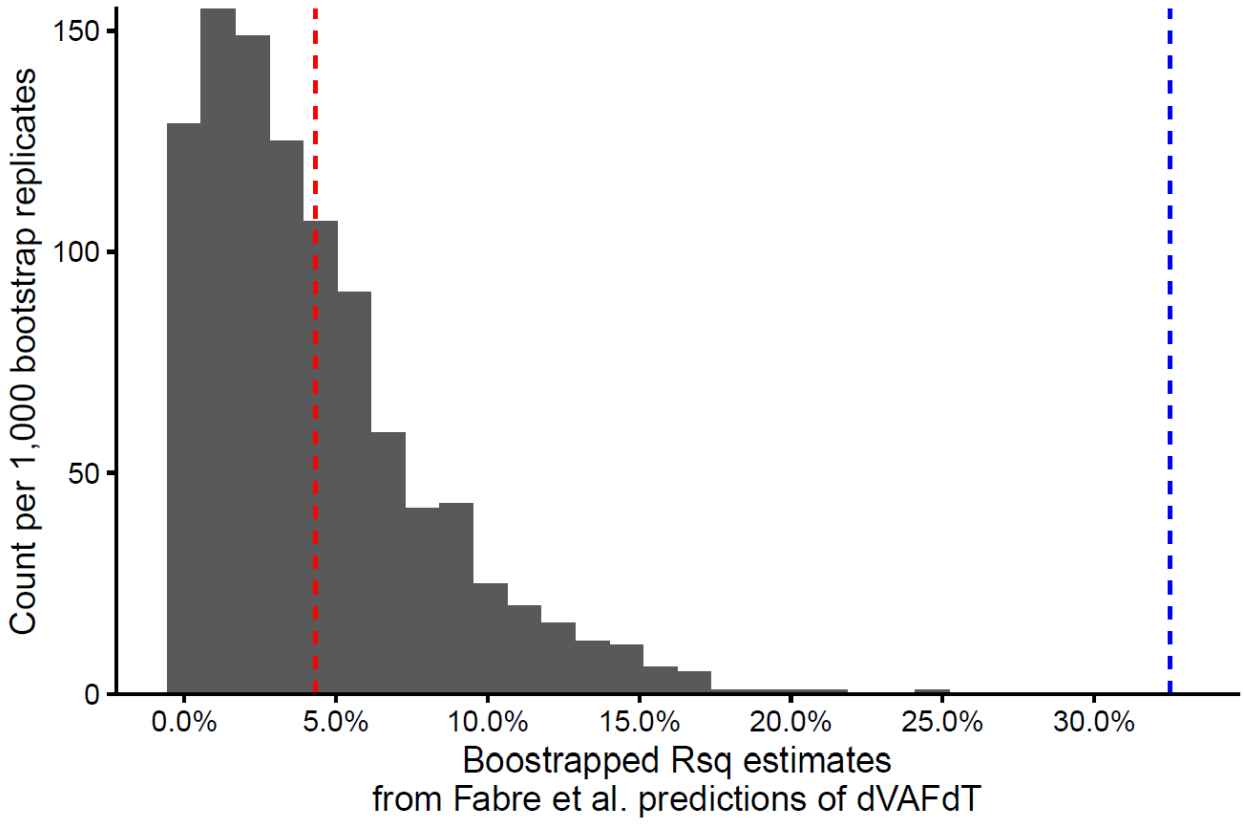


Figure 2: Bootstrapped Rsq estimates of the FBHM estimator of future clone growth applied to the expanded validation cohort. The histogram is comprised of 1,000 Rsq estimates from 1,000 bootstrap resamples of the data. The blue line, 32.5%, is the Rsq between PACER and future clone growth on the WHI samples included in the main text. The red line is the mean Rsq estimate from the FBHM estimator.

Out-of-sample performance of UFE fitness estimates

In addition to the FHBM estimator, we sought to assess the performance an alternative approach which estimates fitness via a series of individual specific linear models that regress VAF onto age at mutational assessment:

$$VAF_{ij}(t) \sim N\left(\beta_{0j} + \beta_{1j} * age(t), \sigma_j^2\right)$$

We define the estimator $\widehat{\beta_{1j}}$ as the unpooled fitness estimator (UFE) for the jth individual. This estimator performed similarly to FHBM (R^2= 11.0%, 95% CI: [2.5%, 20.2%], Supplementary Note 2 Figure 3, Supplementary Note 2 Table 1).

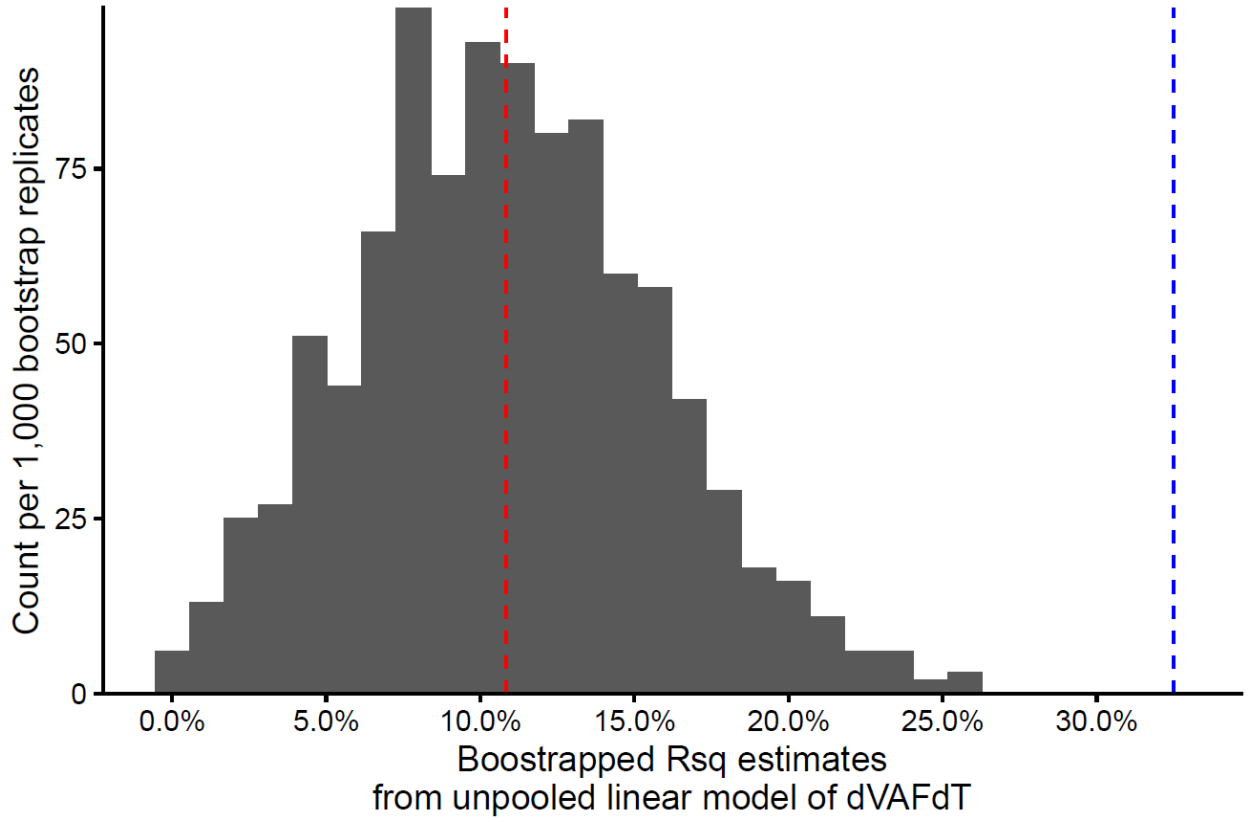


Figure 3: Bootstrapped Rsq estimates of the UFE estimator of future clone growth applied to the expanded validation cohort. The histogram is comprised of 1,000 Rsq estimates from 1,000 bootstrap resamples of the data. The blue line, 32.5%, is the Rsq between PACER and future clone growth on the WHI samples included in the main text. The red line is the mean Rsq estimate from the UFE estimator.

Analysis of 24 clonal hematopoiesis carriers from the Women's Health Initiative across multiple timepoints

As a secondary validation cohort, we utilized 1,181 mutational assessments of clonal hematopoiesis mutations across 156 donors (mean age = 70.7) in the Women's Health Initiative. As described in the main text, a targeted panel of clonal hematopoiesis mutations was sequenced using molecular inversion probes. Among these mutations, we filtered to the 988 that were high confidence mutations. We then filtered to donors with a single driver and at least three separate mutational assessments, resulting in 102 mutational assessments across 24 donors with 3-6 mutational assessments. We performed a parallel analysis to the Fabre data: we stratified the first through penultimate mutational assessments into a training set, and the last mutational timepoint into a test set. We then trained the UFE estimator as above and evaluated its performance as a predictor of future dVAFdT. We observed highly similar predictive performance between UFE and PACER (R^2= 20.0%, 95% CI: [0.8%, 62.2%], Supplementary Note Figure 4, Supplementary Note Table 1).

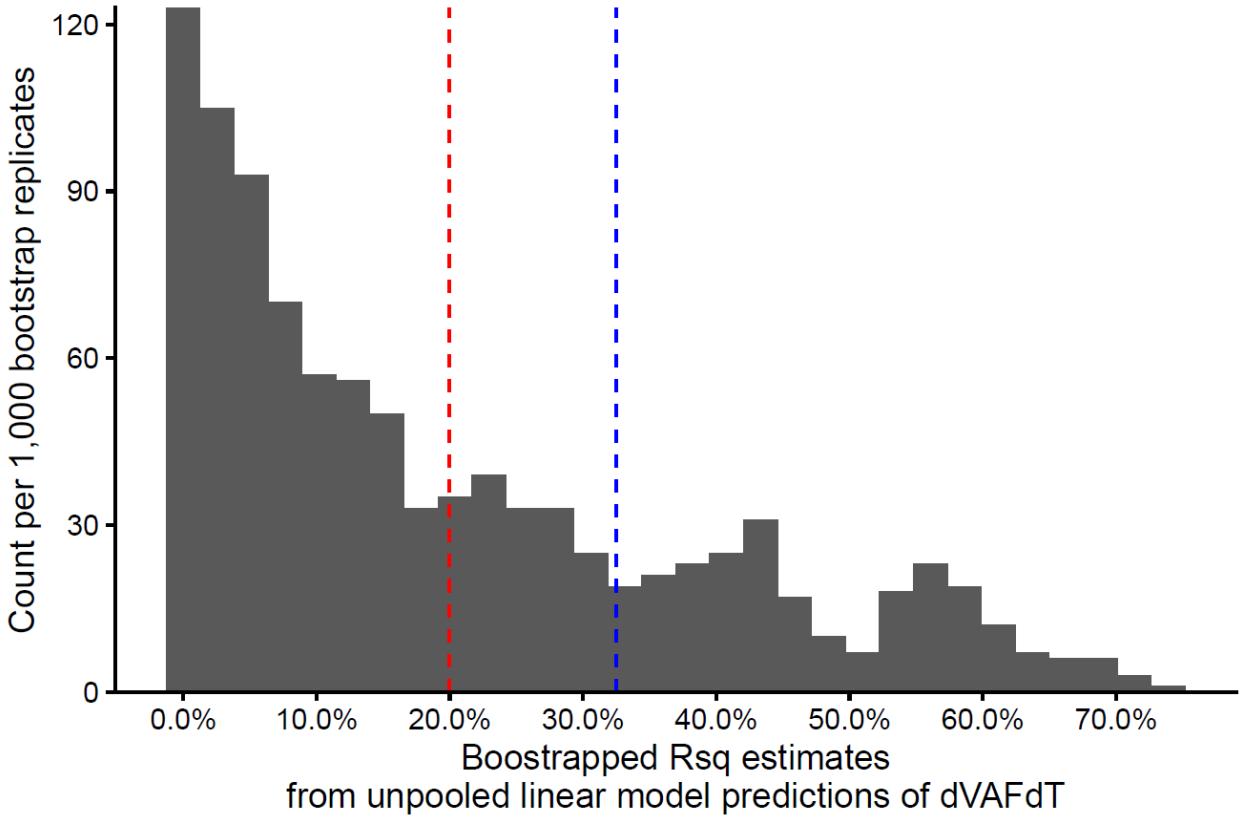


Figure 4: Bootstrapped Rsq estimates of the UFE estimator of future clone growth applied to the WHI cohort. The histogram is comprised of 1,000 Rsq estimates from 1,000 bootstrap resamples of the data. The blue line, 32.5%, is the Rsq between PACER and future clone growth on the WHI samples included in the main text. The red line is the mean Rsq estimate.

Table 1: R^2 estimates across filtering criteria

Cohort	Minimum number of mutational assessments	Donors	FHBM Estimator R^2 (95% CI)	UFE Estimator R^2 (95% CI)
Fabre et al.	3	102	4.0% (0.03% - 13.33%)	11.0% (2.47%-20.24%)
Fabre et al.	4	82	5.5% (0.03% - 19.70%)	11.4% (1.55% - 21.96%)
Fabre et al.	5	52	10.0% (0.04% - 31.70%)	14.4% (0.60% - 30.99%)
WHI	3	24		20.0% (0.8%, 62.2%)

Discussion

Here, we perform a series of systematic analyses to contextualize the ability of multiple statistical methods predict future clonal expansion. We observed that across two independent cohorts, prediction of clonal expansion is challenging, likely due to the noise induced by stochastic drift in the sizes of HSC clones. We observed that PACER, which estimates fitness from a single mutational assessment, performs similarly in prediction future clone growth to estimators derived from longitudinal mutational assessments. Due to the unavailability of WGS in the Fabre et al. validation set, we were unable to apply PACER to these samples. These analyses highlight the utility of larger cohorts for benchmarking mutation fitness estimators – due to the limited sample sizes, many of our estimates of out-of-sample Rsq have substantial uncertainty. This uncertainty inhibits more precise comparisons of these estimators.

Supplementary Note 3: Supplementary Discussion

PACER strengths and weaknesses

Here, we have developed a novel method that allows us to infer clonal expansion rate from a single time point. Our results extend and apply recently developed theory on the evolutionary fitness of clones to permit estimation of fitness within a single individual[20]. Unlike prior methods which used the VAFs of driver variants to estimate fitness[20], our development of a fitness estimator based on passenger mutation counts permits us to perform association tests for other factors associated with clonal expansion, such as inherited genetic variation and environmental exposures. PACER performed similarly to datasets where clones were empirically tracked over time with regards to predicting fitness of different driver genes and predicting future growth of individual clones. The major strength of PACER is that it requires just three variables—passenger count, age, and VAF—to provide accurate fitness estimates from a single timepoint measurement. It is likely that additional information would improve PACER's predictive ability, but this would be expected to limit the number of cohorts to which PACER can be applied if that information is not universally available.

PACER performed well at estimating driver gene fitness (Rsq 80%), but less well at predicting individual future clone growth in the WHI sample (Rsq ~30%), though as noted above it was better or comparable to empirically assessed time series data in this regard. Why would PACER perform relatively better at gene fitness estimation compared to future growth prediction in the WHI validation? We believe that there are two potential explanations for this. 1) It is possible that PACER has large standard errors on an individual level, but the average estimates from many people approximate the true population value for driver gene fitness (or other associations). This follows from first statistical principles – a statistical estimator derived from multiple samples will in general have a much smaller root mean squared error than an estimator derived from a single sample. This suggests that even if PACER has considerable standard errors for individual samples, our large sample size is sufficient to produce aggregated estimates that are close to the true fitness values. 2) An alternative explanation is that much of the difference could be due to the time scales involved in these two assessments. It is now established that the mutations that give rise to clonal expansions that become evident later in life arise much earlier (perhaps between the age of 10-30 years for most clones). In TOPMed, CHIP carriers were close to 70 years old, on average, at the time of WGS. This means that we assessed average clone growth over ~40-60 years in the gene fitness analysis. Contrast this with the situation for the future growth assessment in WHI, where future clone growth was predicted over timepoints spanning ~10-15 years. The shorter time frame is more likely to be subject to random fluctuations/stochastic effects. This can be somewhat overcome with repeated VAF sampling over a shorter time course to generate a more reliable growth trajectory, as Fabre et al. showed, but prediction of clone growth between any two short-term timepoints is still subject to stochastic effects in their dataset, as we show.

Do either of these explanations suggest that PACER is not precise enough to be useful for identifying correlates of clonal expansion, which is the premise of our study? If 1) is true, there is some risk of not being able to identify such correlates if the datasets are too small. This is one reason why we believe that it is important to use PACER in larger datasets, which should overcome imprecision at the individual level. The procedure for determining gene fitness

estimates is conceptually similar to other associational analyses like GWAS, so if one accepts that PACER is valid tool for gene fitness estimation, then it is logical to expect that it is valid for GWAS or other association studies to identify correlates of clonal expansion. If 2) is true, then it is possible that factors that modulate clone growth primarily at short time scales (chemotherapy, for example), could be missed in a PACER association study, but factors that are expected to be relevant lifelong or long-term (like germline variants or frequent environmental exposures) would still be able to be assessed using PACER. These considerations are not unique to PACER and would also be relevant to any method for assessment of clonal growth, including recently published longitudinal analyses of clonal hematopoiesis.

Some other limitations of PACER bear mentioning. The sequencing coverage in TOPMed WGS was 38x, which does not enable detection of mutations with VAFs below 5%[11]. More sensitive assays would increase detection of both driver and passenger mutations and would reduce error in estimation of clonal expansion using PACER. Unlike WGS of single-HSC derived colonies, PACER does not generate phylogenetic trees depicting relationships of clones which could enable more precise determination of clone birth date and fluctuations in growth rate over time[14,16,17]. However, this precision comes at a cost—WGS of single colonies is not easily scalable to thousands of people, unlike PACER. We anticipate that larger cohorts of CHIP carriers with longitudinal sequencing will provide fertile substrate for more sophisticated computational analyses and benchmarking of PACER. We observed that PACER performed similarly to the more sophisticated PACER-HB implementation, but it is challenging to distinguish between similar predictive models in small cohorts. Larger cohorts will likely provide power to utilize additional informative features for clonal expansion and will likely facilitate machine learning motivated approaches. Lastly, we were underpowered to identify germline determinants of clonal expansion in less frequently mutated genes in CHIP, but the addition of WGS data from other cohorts in the coming years should enable additional gene-level analyses of PACER.

Additionally, the use of bulk-WGS precludes analysis of the clonal expansion rate in those with more than one driver mutation without much deeper sequencing coverage. PACER is currently optimized for use in pre-malignant states with a single driver present and would likely have limited utility in datasets of established malignancy. Recent work has demonstrated that hematopoiesis in the elderly becomes increasingly oligoclonal, with many expanded clones occurring in the absence of a known driver mutation[14]. Therefore, there is the potential for passenger mutations from these “driverless” clones to be present in our analysis of CHIP carriers (and could be one reason for some imprecision in the individual clone growth prediction in WHI). However, as long as the probability of carrying a large, “driverless” clone is randomly distributed with respect to carrying clones with known drivers, this should not introduce confounding for the types of analyses that we propose for CHIP clones (but would add noise and make the power to detect associations lower by increasing the background level of passengers randomly across all samples). Indeed, the ability of PACER to infer driver gene fitness accurately at the population level suggests that this is not a critical shortcoming in our approach of performing association studies in large datasets. Nonetheless, the ability to identify and account for hidden clones in the data would improve PACER’s performance.

In summary, we believe PACER performs well for its intended purpose, which is elucidations of correlates of clonal expansion in large datasets with available WGS.

The role of TCL1A in clonal expansion

We performed the first ever GWAS for determinants of clonal expansion rate and identified a common variant of large effect in the promoter of *TCL1A* as the top hit. Remarkably, this single variant, which has previously been linked to reduced risk of LOY[32,42], was also associated with protection from driver mutations in *TET2*, *ASXL1*, *SF3B1*, *SRSF2,* and possibly *JAK2*. We also demonstrated with experimental work that *TCL1A* was normally lowly expressed in HSCs, but that the introduction of mutations in *TET2* or *ASXL1* led to expression of the protein, possibly by permitting promoter chromatin accessibility and hence transcription of the gene. This was completely prevented by the alt-allele of rs2887399, explaining the reduction in predicted clonal expansion rate by PACER and decreased prevalence of these driver mutations in those carrying the allele. These results strongly implicate *TCL1A* expression as being a causal factor promoting clonal expansion downstream of several driver mutations in CHIP. To our knowledge, *TCL1A* itself is not somatically mutated in CHIP, perhaps because gain-of-function point mutations are not directly possible. Importantly, our results suggest that pharmacologically targeting TCL1A may suppress growth of CHIP and hematological cancers associated with mutations in these genes.

The large protective effect seen with rs2887399 suggests that *TCL1A* expression is likely a dominant factor mediating clonal expansion due to these mutations. This was especially the case for driver mutations in *SRSF2* and *SF3B1* which were very rare in those homozygous for the alt-allele, suggesting that activation of *TCL1A* expression may be a near requirement for clonal expansion due to these mutations. We do not explore how splicing factor mutations are mechanistically linked to *TCL1A* activation in this study, but one possibility is that mutations in *SF3B1* or *SRSF2* lead to a cryptic splice junction[43,44] within the *TCL1A* 3' UTR, which may lead to increased stability of the transcript in HSCs[45]. These results may also potentially explain why mutations in *Asxl1*, *Sf3b1*, and *Srsf2* in mouse HSCs do not lead to robust clonal expansion, as the regulatory elements and non-coding regions of the mouse *Tcl1* gene are not well conserved with human *TCL1A*. We previously reported that the alt-allele of rs2887399 was associated with increased risk of *DNMT3A* mutations[11], but here we found that carrying the alt-allele did not increase expansion rate of *DNMT3A* clones by PACER or result in increased expansion of *DNMT3A* edited HSPCs in vitro. One explanation for these discordant results is that the relative reduction in fitness advantage for other non-*DNMT3A* drivers in carriers of the alt-allele permits more opportunity for low fitness *DNMT3A* mutant clones to expand as hematopoiesis becomes more oligoclonal with aging[14], as we demonstrated in a simulation. However, further analyses and experimentation would be needed to fully validate this hypothesis and to rule out competing models that we have not considered.

We observed that expression of *TCL1A*, even in the absence of CHIP driver mutations, was sufficient to promote clonal expansion of HSCs in vitro. Several studies have shown that HSCs are especially sensitive to various stressors, such as the integrated stress response[46], unfolded protein response[47], and loss of autophagic capacity[48]. *TCL1A* has a known role in promoting AKT oligomerization and activation[49], which could promote downstream growth

pathways such as mTORC1 while also inhibiting FOXO transcriptional activity[50]. We found that stress response and FOXO target genes, including genes involved in cell cycle arrest and apoptosis, were suppressed in HSCs expressing *TCL1A*, consistent with AKT mediated inhibition of FOXO in these cells. Studies in mice have shown that constitutive activation of PI3K/AKT[51,52] or deletion of FOXO[53,54] can push HSCs into cycle, but also lead to loss of functional HSCs. In accordance, directly activating mutations in genes of the PI3K/AKT pathway are not seen in CHIP, possibly due to the same deleterious effect in human HSCs. Our results indicate that expression of *TCL1A* does not lead to these deleterious effects, but the reasons for this are unclear. It is possible that *TCL1A* expression activates AKT in moderation, which allows for clonal expansion without long-term negative effects on HSCs. Alternatively, *TCL1A* may alter other aspects of AKT activity that preserve stemness, or even have AKT-independent effects. We anticipate that additional functional analyses of *TCL1A* will further elucidate its key role in promoting clonal expansion of HSCs.

Supplementary Note 4: Supplementary Acknowledgements

African-American Coronary Artery Calcium consortium study

MESA Family is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support is provided by grants and contracts R01HL071051, R01HL071205, R01HL071250, R01HL071251, R01HL071258, R01HL071259, by the National Center for Research Resources, Grant UL1RR033176. Also supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR001881, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center.

Atrial Fibrillation Biobank Ludwig Maximilian University (AFLMU)

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: Defining time-dependent genetic and transcriptomic responses to cardiac injury among patients with arrhythmias" (phs001543) was performed at the Broad Institute of MIT and Harvard (UM1HG008895). This funding source was an NHLBI supplement to NHGRI's Centers for Common Disease Genomics (CCDG). Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1). Phenotype harmonization, data management, sample-identity QC, and general study coordination, were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

This study is part of the Centers for Common Disease Genomics (CCDG) program, a large-scale genome sequencing effort to identify rare risk and protective alleles that contribute to a range of common disease phenotypes. The CCDG program is funded by the National Human Genome Research Institute (NHGRI) and the National Heart, Lung, and Blood Institute (NHLBI). Sequencing was completed at the Human Genome Sequencing Center at Baylor College of Medicine under NHGRI grant UM1 HG008898. The cohort participants were recruited at the University Hospital of the LMU Munich, Munich, Germany, additionally facilitated through past funding by the German Federal Ministry of Research (BMBF) and the German Competence Network on Atrial Fibrillation (AFNET).

Genetics of Cardiometabolic Health in the Amish

The TOPMed component of the Amish Research Program was supported by NIH grants R01 HL121007, U01 HL072515, and R01 AG18728.

Atherosclerosis Risk in Communities

The Atherosclerosis Risk in Communities study has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services (contract numbers HHSN268201700001I,

HHSN268201700002I, HHSN268201700003I, HHSN268201700004I and HHSN268201700005I). The authors thank the staff and participants of the ARIC study for their important contributions.

Australian Familial Atrial Fibrillation Study
The NHGRI supplements were awarded with funds from the NHLBI.

Barbados Asthma Genetics Study

We gratefully acknowledge the contributions of Pissamai and Trevor Maul, Paul Levett, Anselm Hennis, P. Michele Lashley, Raana Naidu, Malcolm Howitt and Timothy Roach, and the numerous health care providers, and community clinics and co-investigators who assisted in the phenotyping and collection of DNA samples, and the families and patients for generously donating DNA samples to the Barbados Asthma Genetics Study (BAGS). Funding for BAGS was provided by National Institutes of Health (NIH) R01HL104608, R01HL087699, and HL104608 S1.

Mount Sinai BioMe Biobank

The Mount Sinai BioMe Biobank has been supported by The Andrea and Charles Bronfman Philanthropies and in part by Federal funds from the NHLBI and NHGRI (U01HG00638001; U01HG007417; X01HL134588). We thank all participants in the Mount Sinai Biobank. We also thank all our recruiters who have assisted and continue to assist in data collection and management and are grateful for the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai.

Duke Cardiac CATHeterization (CATHGEN) study

Funding support for the Genetic Mediators of Metabolic CVD Risk was provided by NHLBI grant RC2 HL101621 (William E. Kraus)

The Cleveland Clinic Atrial Fibrillation Study

The research reported in this article was supported by grants from the National Institutes of Health (NIH) National Heart, Lung, and Blood Institute grants R01 HL090620 and R01 HL111314, the NIH National Center for Research Resources for Case Western Reserve University and Cleveland Clinic Clinical and Translational Science Award (CTSA) UL1-RR024989, the Department of Cardiovascular Medicine philanthropic research fund, Heart and Vascular Institute, Cleveland Clinic, the Fondation Leducq grant 07-CVD 03, and The Atrial Fibrillation Innovation Center, State of Ohio.

Whole Genome Sequence Analysis in Early Cerebral Small Vessel Disease

The Coronary Artery Risk Development in Young Adults Study (CARDIA) is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with the University of Alabama at Birmingham (HHSN268201800005I & HHSN268201800007I),

Northwestern University (HHSN268201800003I), University of Minnesota (HHSN268201800006I), and Kaiser Foundation Research Institute (HHSN268201800004I). CARDIA was also partially supported by the Intramural Research Program of the National Institute on Aging (NIA) and an intra-agency agreement between NIA and NHLBI (AG0005).

Cleveland Family Study

The Cleveland Family Study has been supported in part by National Institutes of Health grants [R01-HL046380, KL2-RR024990, R35-HL135818, and R01-HL113338].

Cardiovascular Health Study

Cardiovascular Health Study: This research was supported by contracts HHSN268201200036C, HHSN268200800007C, HHSN268201800001C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, 75N92021D00006, and grants U01HL080295 and U01HL130114 from the National Heart, Lung, and Blood Institute (NHLBI), with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided by R01AG023629 from the National Institute on Aging (NIA). A full list of principal CHS investigators and institutions can be found at CHS-NHLBI.org. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Genetic Epidemiology of COPD

The COPDGene project described was supported by Award Number U01 HL089897 and Award Number U01 HL089856 from the National Heart, Lung, and Blood Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health. The COPDGene project is also supported by the COPD Foundation through contributions made to an Industry Advisory Board that has included AstraZeneca, Bayer Pharmaceuticals, Boehringer Ingelheim, Genentech, GlaxoSmithKline, Novartis, Pfizer, and Sunovion. A full listing of COPDGene investigators can be found at: http://www.copdgene.org/directory

Determining the association of chromosomal variants with non-PV triggers and ablation-outcome in AF

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). The TOPMed acknowledgement statement can be found at: https://www.nhlbiwgs.org/acknowledgements. WGS for “NHLBI TOPMed: Determining the association of chromosomal variants with non-PV

triggers and ablation-outcome in AF-DECAF" (phs001546.v1.p1) was performed at the Broad Institute of MIT and Harvard (3U54HG003067-12S2, 3U54HG003067-13S1). Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1). Phenotype harmonization, data management, sample-identity QC, and general study coordination, were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. The NHGRI supplements were awarded with funds from the NHLBI.

Diabetes Heart Study

This work was supported by R01 HL92301, R01 HL67348, R01 NS058700, R01 AR48797, R01 DK071891, R01 AG058921, the General Clinical Research Center of the Wake Forest University School of Medicine (M01 RR07122, F32 HL085989), the American Diabetes Association, and a pilot grant from the Claude Pepper Older Americans Independence Center of Wake Forest University Health Sciences (P60 AG10484).

Early-onset Atrial Fibrillation in the Estonian Biobank

The TOPMed acknowledgement statement can be found at: https://www.nhlbiwgs.org/acknowledgements. Please include: "Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: Defining time-dependent genetic and transcriptomic responses to cardiac injury among patients with arrhythmias" (phs# with version extension) was performed at the Broad Institute of MIT and Harvard (3R01HL092577-06S1). Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1). Phenotype harmonization, data management, sample-identity QC, and general study coordination, were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed."

Evaluation of COPD Longitudinally to Identify Predictive Surrogate Endpoints

The ECLIPSE study (NCT00292552) was sponsored by GlaxoSmithKline. The ECLIPSE investigators included: ECLIPSE Investigators — Bulgaria: Y. Ivanov, Pleven; K. Kostov, Sofia. Canada: J. Bourbeau, Montreal; M. Fitzgerald, Vancouver, BC; P. Hernandez, Halifax, NS; K. Killian, Hamilton, ON; R. Levy, Vancouver, BC; F. Maltais, Montreal; D. O'Donnell, Kingston, ON. Czech Republic: J. Krepelka, Prague. Denmark: J. Vestbo, Hvidovre. The Netherlands: E. Wouters, Horn-Maastricht. New Zealand: D. Quinn, Wellington. Norway: P. Bakke, Bergen. Slovenia: M. Kosnik, Golnik. Spain: A. Agusti, J. Sauleda, P. de Mallorca. Ukraine: Y. Feschenko, V. Gavrisyuk, L. Yashina, Kiev; N. Monogarova, Donetsk. United Kingdom: P. Calverley,

Liverpool; D. Lomas, Cambridge; W. MacNee, Edinburgh; D. Singh, Manchester; J. Wedzicha, London. United States: A. Anzueto, San Antonio, TX; S. Braman, Providence, RI; R. Casaburi, Torrance CA; B. Celli, Boston; G. Giessel, Richmond, VA; M. Gotfried, Phoenix, AZ; G. Greenwald, Rancho Mirage, CA; N. Hanania, Houston; D. Mahler, Lebanon, NH; B. Make, Denver; S. Rennard, Omaha, NE; C. Rochester, New Haven, CT; P. Scanlon, Rochester, MN; D. Schuller, Omaha, NE; F. Sciurba, Pittsburgh; A. Sharafkhaneh, Houston; T. Siler, St. Charles, MO; E. Silverman, Boston; A. Wanner, Miami; R. Wise, Baltimore; R. ZuWallack, Hartford, CT. ECLIPSE Steering Committee: H. Coxson (Canada), C. Crim (GlaxoSmithKline, USA), L. Edwards (GlaxoSmithKline, USA), D. Lomas (UK), W. MacNee (UK), E. Silverman (USA), R. Tal-Singer (Co-chair, GlaxoSmithKline, USA), J. Vestbo (Co-chair, Denmark), J. Yates (GlaxoSmithKline, USA). ECLIPSE Scientific Committee: A. Agusti (Spain), P. Calverley (UK), B. Celli (USA), C. Crim (GlaxoSmithKline, USA), B. Miller (GlaxoSmithKline, USA), W. MacNee (Chair, UK), S. Rennard (USA), R. Tal-Singer (GlaxoSmithKline, USA), E. Wouters (The Netherlands), J. Yates (GlaxoSmithKline, USA).

Boston Early-Onset COPD Study

The Boston Early-Onset COPD Study was supported by R01 HL113264 and U01 HL089856 from the National Heart, Lung, and Blood Institute.

Framingham Heart Study

The Framingham Heart Study (FHS) acknowledges the support of contracts NO1-HC-25195, HHSN268201500001I and 75N92019D00031 from the National Heart, Lung and Blood Institute and grant supplement R01 HL092577-06S1 for this research. We also acknowledge the dedication of the FHS study participants without whom this research would not be possible. Dr. Vasan is supported in part by the Evans Medical Foundation and the Jay and Louis Coffman Endowment from the Department of Medicine, Boston University School of Medicine.

Genetic Epidemiology Network of Arteriopathy

Support for GENOA was provided by the National Heart, Lung and Blood Institute (HL054457, HL054464, HL054481, HL119443, and HL087660) of the National Institutes of Health. GENOA may also require study-specific acknowledgements depending on the phenotypes analyzed (please consult with the GENOA co-author on the paper).

Genetic Epidemiology Network of Salt Sensitivity

The Genetic Epidemiology Network of Salt-Sensitivity (GenSalt) was supported by research grants (U01HL072507, R01HL087263, and R01HL090682) from the National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, MD.

GeneSTAR

GeneSTAR was supported by the National Institutes of Health/National Heart, Lung, and Blood Institute (U01 HL72518, HL087698, HL112064) and by a grant from the National Institutes of Health/National Center for Research Resources (M01-RR000052) to the Johns Hopkins General Clinical Research Center.

Groningen Genetics of Atrial Fibrillation

The TOPMed acknowledgement statement can be found at: https://www.nhlbiwgs.org/acknowledgements. In addition, please include: "Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for “NHLBI TOPMed - Centers for Common Disease Genomics (CCDG): Groningen Genetics of Atrial Fibrillation (GGAF)” (phs# with version extension) was performed at the Baylor College of Medicine. This funding source was an NHLBI supplement to NHGRI's Centers for Common Disease Genomics (CCDG). Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1). Phenotype harmonization, data management, sample-identity QC, and general study coordination, were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed."

Genetics of Lipid Lowering Drugs and Diet Network

GOLDN biospecimens, baseline phenotype data, and intervention phenotype data were collected with funding from National Heart, Lung and Blood Institute (NHLBI) grant U01 HL072524. Whole-genome sequencing in GOLDN was funded by NHLBI grant R01 HL104135 and supplement R01 HL104135-04S1.

Hispanic Community Health Study - Study of Latinos

The Hispanic Community Health Study/Study of Latinos is a collaborative study supported by contracts from the National Heart, Lung, and Blood Institute (NHLBI) to the University of North Carolina (HHSN268201300001I / N01-HC-65233), University of Miami (HHSN268201300004I / N01-HC-65234), Albert Einstein College of Medicine (HHSN268201300002I / N01-HC-65235), University of Illinois at Chicago – HHSN268201300003I / N01-HC-65236 Northwestern Univ), and San Diego State University (HHSN268201300005I / N01-HC-65237). The following Institutes/Centers/Offices have contributed to the HCHS/SOL through a transfer of funds to the NHLBI: National Institute on Minority Health and Health Disparities, National Institute on Deafness and Other Communication Disorders, National Institute of Dental and Craniofacial Research, National Institute of Diabetes and Digestive and Kidney Diseases, National Institute of Neurological Disorders and Stroke, NIH Institution-Office of Dietary Supplements.

Heart and Vascular Health Study

The Heart and Vascular Health Study was supported by grants HL068986, HL085251, HL095080, and HL073410 from the National Heart, Lung, and Blood Institute.

Hypertension Genetic Epidemiology Network

The HyperGEN Study is part of the National Heart, Lung, and Blood Institute (NHLBI) Family Blood Pressure Program; collection of the data represented here was supported by grants U01 HL054472 (MN Lab), U01 HL054473 (DCC), U01 HL054495 (AL FC), and U01 HL054509 (NC FC). The HyperGEN: Genetics of Left Ventricular Hypertrophy Study was supported by NHLBI grant R01 HL055673 with whole-genome sequencing made possible by supplement -18S1.

Jackson Heart Study

The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson State University (HHSN268201800013I), Tougaloo College (HHSN268201800014I), the Mississippi State Department of Health (HHSN268201800015I) and the University of Mississippi Medical Center (HHSN268201800010I, HHSN268201800011I and HHSN268201800012I) contracts from the National Heart, Lung, and Blood Institute (NHLBI) and the National Institute on Minority Health and Health Disparities (NIMHD). The authors also wish to thank the staffs and participants of the JHS.

The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the U.S. Department of Health and Human Services.

Mayo Clinic Venous Thromboembolism Study

Funded, in part, by grants from the National Institutes of Health, National Heart, Lung and Blood Institute (HL66216 and HL83141). the National Human Genome Research Institute (HG04735, HG06379), and research support provided by Mayo Foundation.

Multi-Ethnic Study of Atherosclerosis

The MESA project is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts 75N92020D00001, HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, UL1-TR-001420. Also supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR001881, and the National Institute of Diabetes and Digestive and Kidney

Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center. We gratefully acknowledge the studies and participants who provided biologic samples and data for MESA and TOPMed.

Massachusetts General Hospital (MGH) Atrial Fibrillation Study

The research reported in this article was supported by NIH grants K23HL071632, K23HL114724, R21DA027021, R01HL092577, R01HL092577S1, R01HL104156, K24HL105780, and U01HL65962. The research has also been supported by an Established Investigator Award from the American Heart Association (13EIA14220013) and by support from the Fondation Leducq (14CVD01). This manuscript was not prepared in collaboration with MGH AF Study investigators and does not necessarily reflect the opinions or views of the MGH AF Study investigators or the NHLBI.

Defining the time-dependent genetic and transcriptomic responses to cardiac injury among patients with arrhythmias

The NHGRI supplements were awarded with funds from the NHLBI.

My Life, Our Future: Genotyping for Progress in Hemophilia

The My Life, Our Future samples and data are made possible through the partnership of Bloodworks Northwest, the American Thrombosis and Hemostasis Network, the National Hemophilia Foundation, and Bioverativ. We gratefully acknowledge the hemophilia treatment centers and their patients who provided biological samples and phenotypic data.

Malmo Preventive Project

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: Defining time-dependent genetic and transcriptomic responses to cardiac injury among patients with arrhythmias" (phs001544) was performed at the Broad Institute of MIT and Harvard (UM1HG008895). This funding source was an NHLBI supplement to NHGRI's Centers for Common Disease Genomics (CCDG). Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1). Phenotype harmonization, data management, sample-identity QC, and general study coordination, were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. This study is part of the Centers for Common Disease Genomics (CCDG) program, a large-scale genome sequencing effort to identify rare risk and protective alleles that contribute to a range of common disease phenotypes. The CCDG program is funded by the National Human Genome Research Institute (NHGRI) and the National Heart, Lung, and Blood Institute (NHLBI). Sequencing was completed at the Human Genome Sequencing Center at Baylor College of Medicine under NHGRI grant UM1 HG008898.

Partners HealthCare Biobank

We thank the Broad Institute for generating high-quality sequence data supported by the NHLBI grant 3R01HL092577-06S1 to Dr. Patrick Ellinor

San Antonio Family Studies

Collection of the San Antonio Family Study data was supported in part by National Institutes of Health (NIH) grants R01 HL045522, MH078143, MH078111 and MH083824; and whole genome sequencing of SAFS subjects was supported by U01 DK085524 and R01 HL113323. We are very grateful to the participants of the San Antonio Family Study for their continued involvement in our research programs.

Study of Asthma Phenotypes and Pharmacogenomic Interactions by Race-Ethnicity

Grant funding sources must be acknowledged in any study using SAPPHIRE data. These include the Fund for Henry Ford Hospital, the American Asthma Foundation, the National Heart Lung and Blood Institute (R01HL118267, X01HL134589), the National Institute of Allergy and Infectious Diseases (R01AI079139), and the National Institute of Diabetes and Digestive and Kidney Diseases (R01DK113003).

Study of Adiposity in Samoans

Date collection was funded by NIH grant R01-HL093093. We thank the Samoan participants of the study and local village authorities. We acknowledge the support of the Samoan Ministry of Health and the Samoa Bureau of Statistics for their support of this research.

Taiwan Study of Hypertension using Rare Variants

The Rare Variants for Hypertension in Taiwan Chinese (THRV) is supported by the National Heart, Lung, and Blood Institute (NHLBI) grant (R01HL111249) and its participation in TOPMed is supported by an NHLBI supplement (R01HL111249-04S1). THRV is a collaborative study between Washington University in St. Louis, LA BioMed at Harbor UCLA, University of Texas in Houston, Taichung Veterans General Hospital, Taipei Veterans General Hospital, Tri-Service General Hospital, National Health Research Institutes, National Taiwan University, and Baylor University. THRV is based (substantially) on the parent SAPPHIRe study, along with additional population-based and hospital-based cohorts. SAPPHIRe was supported by NHLBI grants (U01HL54527, U01HL54498) and Taiwan funds, and the other cohorts were supported by Taiwan funds.

The Women's Genome Health Study

The WGHS is supported by the National Heart, Lung, and Blood Institute (HL043851 and HL080467) and the National Cancer Institute (CA047988 and UM1CA182913). The most recent cardiovascular endpoints were supported by ARRA funding HL099355.

Women's Health Initiative

The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts 75N92021D00001, 75N92021D00002, 75N92021D00003, 75N92021D00004, 75N92021D00005.

Supplementary Figures

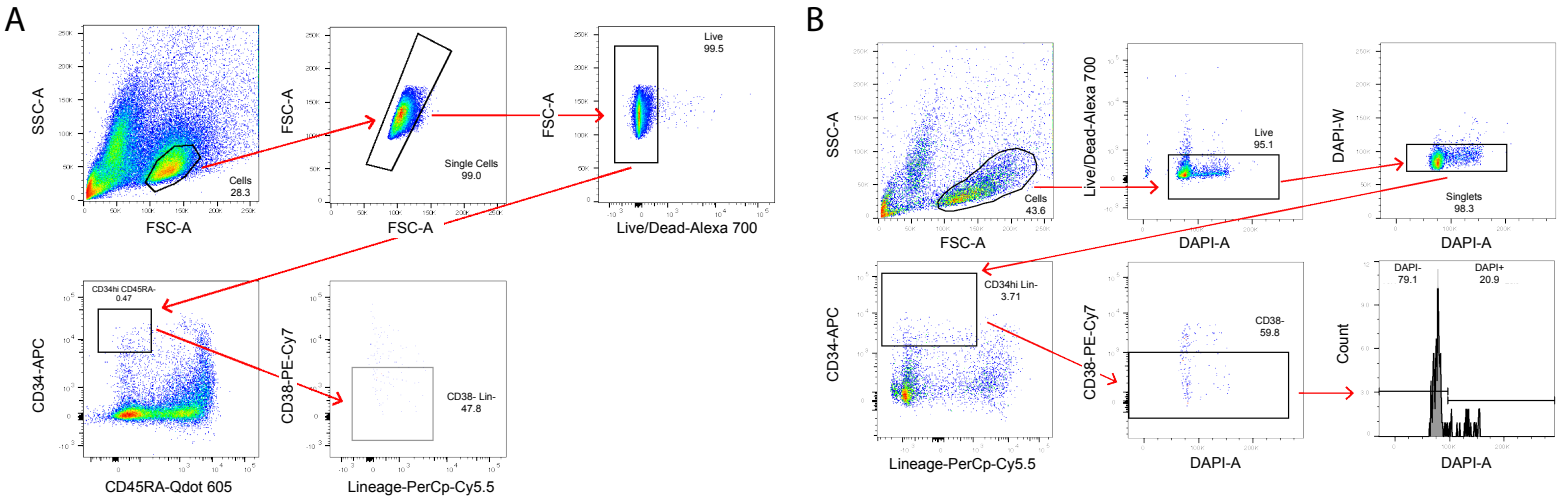
Supplementary Figure 1

A Flow gating scheme for identifying and sorting Lineage-CD34+ CD38-CD45RA-human hematopoietic stem cells (HSCs) and multi-potent progenitors (MPPs) presented on Figure 3D,E,F, and 4B. 500-1000HSC/MPPs were sorted per well for the liquid culture expansion assays. Purity of sorted cells was confirmed by performing flow on sorted cells. **B** Flow gating scheme for identifying DAPI+ Lineage-CD34+ CD38-cycling hematopoietic stem cells (HSCs) and multi-potent progenitors (MPPs) presented on Figure 4E.

Supplementary Figure 2

A Flow gating scheme for identifying KLS (Lin-Sca1+ c-Kit+), LT-HSC (Lin-Sca1+ c-Kit+ Flk2-CD150+CD48-), ST-HSC (Lin-Sca1+ c-Kit+ Flk2-CD150-CD48-), MPP2 (Lin-Sca1+ c-Kit+ Flk2-CD150+CD48+), MPP3 (Lin-Sca1+ c-Kit+ Flk2-CD150-CD48+), and MPP4 (Lin-Sca1+ c-Kit+ Flk2+ CD150-CD48+) populations in mouse bone marrow presented in Figure 4D and Extended Data Figure 11C. **B** Flow gating scheme for identifying GFP+ granulocytes (CD3-CD19-CD11b+ Ly6G+ GFP+) in mouse peripheral blood presented in Figure 4C.

Supplementary Figure 1



Supplementary Figure 2