Tissue-specific impacts of aging and genetics on gene expression patterns in humans

Ryo Yamamoto[1,2,A], Ryan Chung[3,A], Juan Manuel Vazquez[1], Huanjie Sheng[1], Philippa Steinberg[1], Nilah M Ioannidis[3,4], and Peter H Sudmant[1,3]

[1]Department of Integrative Biology, University of California, Berkeley
[2]Bioinformatics Interdepartmental Program, University of California, Los Angeles
[3]Center for Computational Biology, University of California, Berkeley
[4]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley

Age is the primary risk factor for many common human diseases including heart disease, Alzheimer's dementias, cancers, and diabetes. Determining how and why tissues age differently is key to understanding the onset and progression of such pathologies. Here, we set out to quantify the relative contributions of genetics and aging to gene expression patterns from data collected across 27 tissues from 948 humans. We show that gene expression patterns become more erratic with age in several different tissues reducing the predictive power of expression quantitative trait loci. Jointly modelling the contributions of age and genetics to transcript level variation we find that the heritability (h^2) of gene expression is largely consistent among tissues. In contrast, the average contribution of aging to gene expression variance varied by more than 20-fold among tissues with $R^2_{\text{age}} > h^2$ in 5 tissues. We find that the coordinated decline of mitochondrial and translation factors is a widespread signature of aging across tissues. Finally, we show that while in general the force of purifying selection is stronger on genes expressed early in life compared to late in life as predicted by Medawar's hypothesis, a handful of highly proliferative tissues exhibit the opposite pattern. In contrast, gene expression variation that is under genetic control is strongly enriched for genes under relaxed constraint. Together we present a novel framework for predicting gene expression phenotypes from genetics and age and provide insights into the tissue-specific relative contributions of genes and the environment to phenotypes of aging.

[A] RY and RC contributed equally to this work

aging | genetics| eQTL | Medawar

Correspondence: ***psudmant@berkeley.edu, nilah@berkeley.edu***

Introduction

Organismal survival requires molecular processes to be carried out with the utmost precision. However, as individuals age many biological processes deteriorate resulting in impaired function and disease. Such increases in the overall variance of molecular processes are predicted by Medawar's germline mutation accumulation theory (1), which states that because older individuals are less likely to contribute their genetic information to the next generation, there is reduced selection to eliminate deleterious phenotypes that appear late in life (2). This theory also predicts that genes expressed early in life should be under increased selective constraint compared to genes expressed late in life. However, a key challenge remains in both quantifying age-associated changes in biological processes across tissues and identifying how genetic variation influences such changes.

At the organismal level, age-associated changes in the heterogeneity of gene expression between individuals have been observed for a handful of genes in humans (3). In an analysis of gene expression in monozygotic (identical) twins, 42 genes showed age-associated differences in gene expression, suggesting a role for the environment in modulating gene expression with age (2, 3). Similarly, the proportion of expression quantitative trait loci (eQTLs) detected from blood in 70 year olds declined by 2.7% when they were resampled at 80 years old (4). However, the extent of this phenomenon, both across genes and tissues, remains unclear (5). Age-associated increases in the heterogeneity of gene expression have also been observed at the level of individual cell-to-cell variation; however, only some cell types appear to be impacted (6). In a recent study of immune T-cells from young and aged individuals, no difference in cell-to-cell variability was observed in unstimulated cells, however, upon immune activation the older cells appeared more heterogeneous (7). It is not known why some cell-types and not others may be more likely to exhibit increased cellular variability.

The relationship between the age at which a specific gene is expressed and the force of purifying selection has also recently been explored across a number of species (8, 9). These analyses have broadly confirmed that, on average, genes expressed later in life are under less constraint compared to those expressed early in life. However, how these patterns vary across different tissues and are impacted by genetic variation has not been systematically explored.

Here we set out to understand how aging affects the molecular heterogeneity of gene expression and to model the relative impact of age and genetic variation on this phenotype across tissues. First, using gene expression data from 948 individuals in GTEx V8 (10), we show that eQTLs are less predictive in older individuals, however to a different extent across various tissues. We show that gene expression heterogeneity between individuals increases with age in these tissues. Using a regularized linear model-based approach to jointly model the impact of both age and genetic variation on gene expression we find that while the average heritability of gene expression is consistent across tissues, the average contribution of age varies substantially. Furthermore, while the genetic regulation of gene expression is similar across tissues,

age-associated changes in gene expression are highly tissue-specific in their action. We use this joint model to identify each gene's age of expression and show that while in most tissues late-expressed genes do tend to be under more relaxed selective constraint, among a handful of highly proliferative tissues the opposite trend holds.

Results

Expression quantitative trait loci exhibit varying predictive power in *old* and *young* individuals across several different tissues. To gain insight into how gene regulatory programs might be impacted by aging we analyzed transcriptomic data collected across multiple tissues from 948 humans (GTEx version 8) (10). We hypothesized that aging might dampen the effect of expression quantitative trait loci (eQTLs) due to factors such as increased environmental variance or molecular infidelity (Fig. 1A). To test this hypothesis we first classified individuals into *young* and *old* age groups conservatively grouping individuals above and below the median age (55 years old, Fig. S1), respectively, restricting our analyses to tissues with at least 100 individuals in both groups (27 tissues in total, Fig. S2, Table S1). In each tissue we down-sampled to match the sample size of old and young individuals while additionally controlling for co-factors such as ancestry and technical confounders (methods). Of note, a common approach to controlling for unobserved confounders in large gene expression experiments is to probabilistically infer hidden factors using statistical tools such as PEER (11). We noticed that many of the GTEx PEER factors were significantly correlated with sample age, with the top three correlated PEER factors having a Pearson r of 0.33, -0.21, and -0.15 (Fig. S3). To prevent loss of age related variation, we recalculated a corrected set of PEER factors that were independent of sample age (Methods). We then assessed the significance of GTEx eQTLs in the young and old cohorts respectively, comparing the distribution of P-values over all genes between old and young individuals (Fig. 1A). In 20 out of 27 (74%) of the assessed tissues, the P-value distribution was significantly different between young and old individuals with genotypes more predictive of expression in younger individuals in 12/20 cases (e.g. Fig. 1D). These results were largely identical when the analyses were performed with the original non-corrected PEER factors (18/27 tissues, Fig. S4). These results suggest that the predictive power of eQTLs is impacted by the sample age across the vast majority of tissues. Furthermore this effect is more pronounced in older samples compared with younger samples.

Age-associated increases in gene expression heterogeneity reduce gene expression heritability. We hypothesized that the reduced predictive power of eQTLs in some older tissues might be in part due to an overall increase in expression heterogeneity in these tissues, potentially as a result of increased environmental variance. To test if such an effect would broadly affect expression across all genes in a tissue (Fig. 2A) we calculated the distribution of pairwise distances among individual's tissue-specific gene expression profiles using the Jensen-Shannon Divergence (JSD) (12, 13) as a distance metric. The JSD is a robust distance which is less impacted by outliers compared to other methods (e.g. Euclidean distance) (13). Comparing the distribution of pairwise differences in transcriptional profiles within distinct age groups allows us to determine if gene expression signatures are more similar among *younger* individuals versus among *older* individuals.

We compared the mean difference in gene expression distances among old and young individuals as well as the slope of the inter-individual JSD and when grouping individuals into six bins spanning 20-80 years old (see methods, Fig. 2B, 2C). These two strategies yielded highly similar results (Fig. 2B R=0.8) and identified a cluster of 12 tissues exhibiting robust increases in the average inter-individual expression distance as a function of age (e.g. Fig. 2C). Our JSD analysis of old and young individuals was also negatively correlated with the results from our analysis of eQTLs across old and young individuals (Fig. S5, R=-0.48, P=0.01) highlighting that tissues with age-associated increases in inter-individual heterogeneity were likely to also exhibit reductions in the proportion of variance described by eQTLs. Conversely, tissues in which eQTLs explained a higher proportion of gene expression variance in older individuals exhibited a decrease in inter-individual gene expression variation.

To expand our eQTL analyses to account for the combined impact of nearby SNPs, we utilized the multi-SNP regularized linear model of PrediXcan (14). This model has the benefit of combining genetic effects across many loci, instead of examining just a single eQTL variant. This combined genetic contribution to gene expression variance results in an estimate of the heritability (h^2) for each gene. We applied this model independently in old and young individuals to quantify h^2 and found that the average per-gene difference in h^2 between old and young individuals was strongly negatively correlated with the difference in JSD between samples (R=0.6, P=9.9e-4, Fig. 2D, Fig. S6). Together these results suggest that across numerous tissues aging is associated with an overall increase in gene expression heterogeneity. This increased expression variance drives a reduction in the average heritability of gene expression across these tissues.

Jointly modeling the impact of age and genetics on gene expression identifies distinct, tissue-specific patterns of aging. A more powerful approach to understand how both genetics and age impact gene expression variation is to jointly model these factors simultaneously. We set out to extend the regularized linear model to incorporate an age factor (Fig. 3A) allowing us to parse apart the individual contributions of genetics (R^2_{genetics} or h^2), age (R^2_{age}), and the environment ($R^2_{\text{environment}}$), to the expression variance of each gene (e.g. Fig. 3B, Fig. 3C). We define $R^2_{\text{environment}}$ as all sources of variation not captured by h^2 and R^2_{age}. Estimates of h^2 in our extended model were highly consistent with those in the original PrediXcan approach (Fig. S 7).

Employing our model across each tissue independently we find that average heritability of gene expression was largely

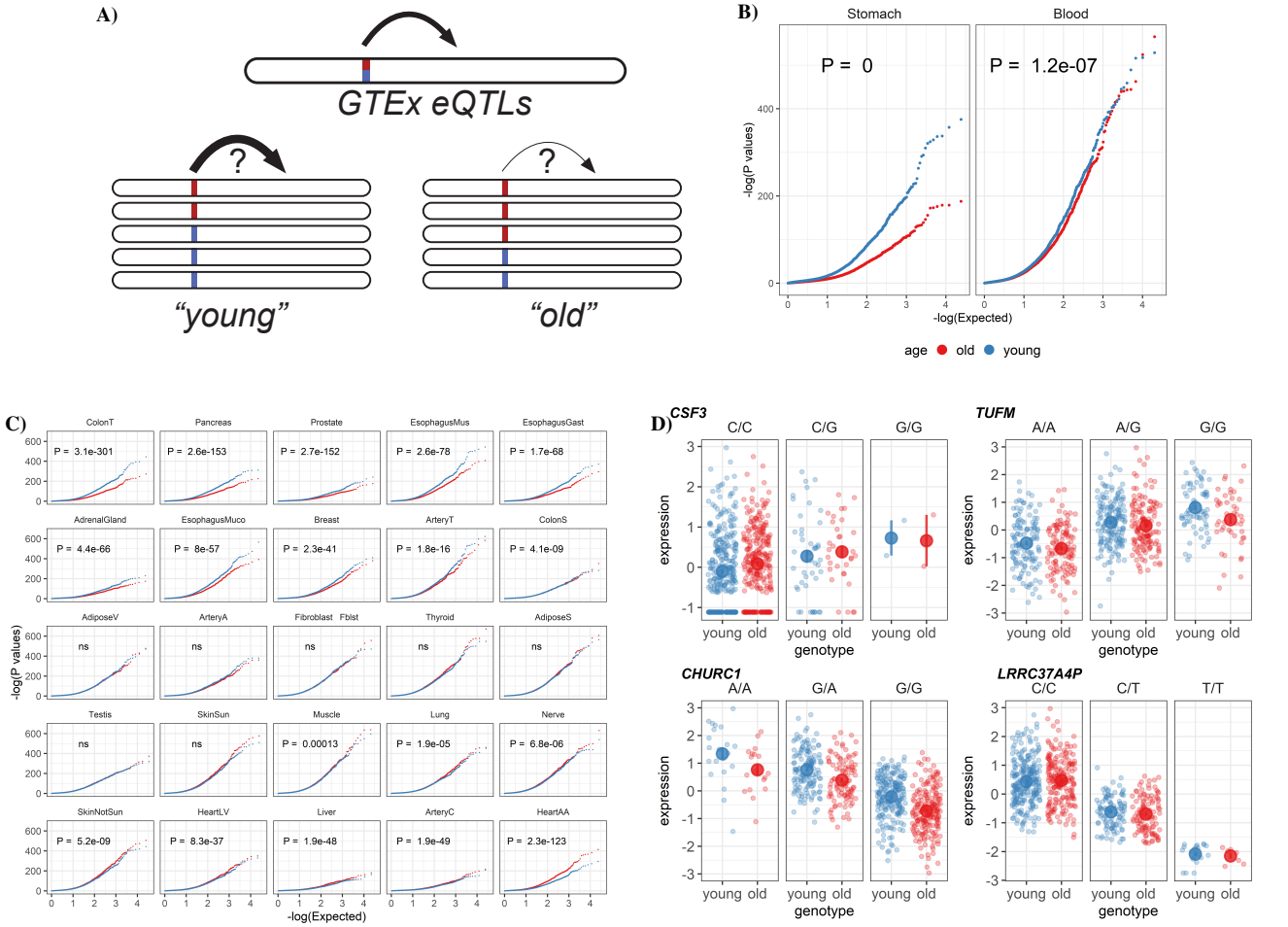

Fig. 1. eQTLs become less explanatory with age in many tissues. A) A hypothetical model of the differing power to detect eQTLs in old and young cohorts. **B)** QQ plots of eQTL p-values (plotted as -log(P)) for old (red) and young (blue) individuals from a linear model correlating expression with the lead SNP for each gene in blood and stomach tissue and **C)** all other tissues (Table S2). **D)** Examples of gene expression binned by genotype and age for four genes in which eQTLs are less explanatory in older individuals in whole blood.

consistent among tissues ranging from 2.9%-5.7% with 40% of genes having an h^2>10% in at least one tissue (Fig. 3D, S8). Thus, while the variation in expression of many individual genes is strongly influenced by genetics, on average, genetics explains a small proportion of overall gene expression variation. In contrast, the average contribution of aging to gene expression varied more than 20-fold among tissues from 0.4%-7.9% with the average R^2_{age} greater than the average h^2 in 5 tissues. Among these 5 tissues the expression of 39-54% of genes was more influenced by age than by genetics (i.e. $R^2_{\text{age}} > h^2$, Fig. S9) and across all tissues 45% of genes had an R^2_{age}>10% in at least one tissue. Assessing the tissue-specificity of these trends on a per-gene basis we found while the estimated heritability of gene expression tended to be similar among different tissues, the age-associated component exhibited significantly more tissue specificity (P<2.2e-16, Fig. 3E). We note that the widespread signatures of age-associated gene expression variance that we identify are virtually undetectable when using the GTEx-provided PEER factors. Just 1.84% of the age-associated genes we identify have nonzero age coefficient when using these GTEx PEER factors (Fig. S10). Our model thus widely expands the utility of the GTEx dataset and exploration of critical biological signatures of aging. Together these results imply that age-associated patterns of gene expression exhibit substantially more tissue specificity than those that are influenced by genetics and among several tissues age plays a much stronger role in driving gene expression patterns than genetics.

Coordinated decline of mitochondrial and translation factors is a widespread signature of aging across tissues. To understand the underlying biological implications of age-associated gene expression changes we applied gene set enrichment analysis (GSEA)(15) to each tissue independently, ranking genes either by the relative contribution of genetics (h^2) or aging (R^2_{age}). Comparing the distribution of P-values from enriched GO-annotations we found that pathways enriched for age-associated variance were substantially more enriched for significance than pathways associated with genetic-associated variance (e.g. Fig. 4A). We found more age-associated pathway enrichment even in tissues for which the average age-associated contribution to gene expression was low (e.g. Pancreas, Fig. S11). This implies that while age-associated changes in gene expression vary widely in their magnitude among tissues, these changes

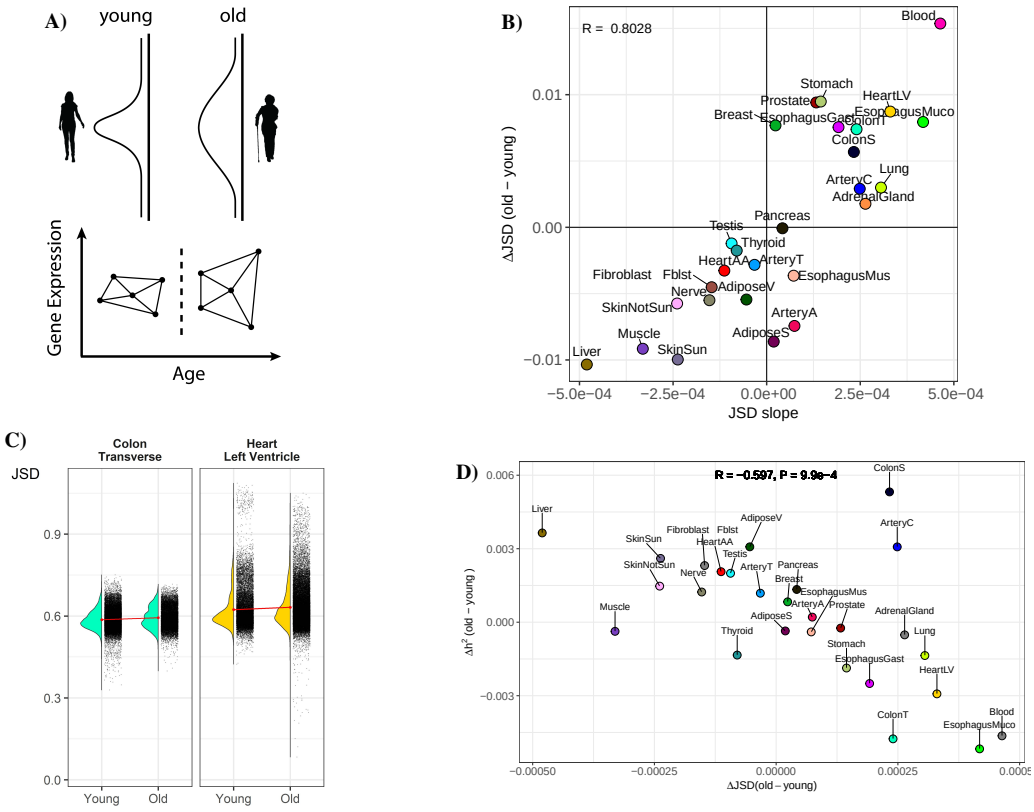

Fig. 2. Inter-individual gene expression heterogeneity increases with age for a subset of tissues. A) Hypothesized age-associated increase in gene expression heterogeneity (top) and our approach for quantifying the inter-individual expression distance with age using the Jensen Shannon Divergence metric (JSD) for age-binned individuals (bottom). **B)** Consistency of measuring the average age-associated change in gene expression heterogeneity across a tissue using a binary binning strategy (y-axis, JSD_{old}-JSD_{young}) or a 6 bin strategy (x-axis, slope of JSD across 6 bins). **C)** The distributions of JSD distances for two tissues in old and young bins. **D)** The relationship between gene expression heterogeneity and the difference in expression heritability between old and young individuals.

consistently impact critical biological processes. A GSEA enrichment analysis of genes ranked by the tissue-averaged slope of the age-associated trend (β_{age}) highlighted several key aging-associated pathways. Pathways associated with various mitochondrial and metabolic processes and translation were enriched for having $-\beta_{\text{age}}$ values, implying age-associated decreases (Fig. 4B). A single immune pathway, the interferon-gamma response was enriched having $+\beta_{\text{age}}$ values (Fig. 4B). An additional 18 immune pathways were identified as having age-associated increases in gene expression using a more lenient significance threshold (FDR<0.05) (Fig. S12). In contrast, no pathways were significantly enriched when genes were ranked by average h^2.

To further explore the functional impact of age-associated gene expression changes we compared the R^2_{age} of all nuclear-encoded mitochondrial genes (n=1120, (16)), and translation initiation, elongation, and termination factors, across tissues (Fig. 4C, Fig. S13). Genes in these pathways were exceptionally enriched for age-associated gene expression across several tissues. In some cases >10% of the average expression variation of mitochondrial or translation factor genes could be explained by age. β_{age} was consistently negative in these mitochondrial and translation factor genes (Fig. 4D) highlighting that genes in these pathways exhibit a systematic decrease in expression as a function of age. Overall across tissues an average of 36% of all mitochondrial genes (406/1120), and 35% of translation factors (119/337) exhibited age-associated declines, however in some tissues these proportions exceeded 60%. In contrast, the only pathway associated with age-associated increases in expression, interferon-gamma response genes, was largely specific to blood and arterial tissue (Fig. 4C), likely due to the role of this pathway in immune cells. Together these results demonstrate that the coordinated decline of mitochondrial genes and translation factors is a widespread phenomenon of aging across several tissues with potential phenotypic consequences.

Distinct evolutionary signatures of gene expression patterns influenced by aging and genetics. Evolutionary theory predicts that due to the increased impact of selection in younger individuals, genes that increase as a function of age ($\beta_{\text{age}} > 0$) should be under reduced selective constraint compared to genes that are highly expressed in young individuals ($\beta_{\text{age}} < 0$), a theory of aging known as *Medawar's hypothesis* (1) (Fig. 5A). Several recent studies have demonstrated the generality of this trend across species (8, 9, 17) however the tissue-specificity of this theoretical prediction has not been explored. We sought to test the generality of this trend across different tissues by comparing β_{age} with the level of constraint on genes, quantified as the probability loss of function intolerance (pLI) score from gnomAD

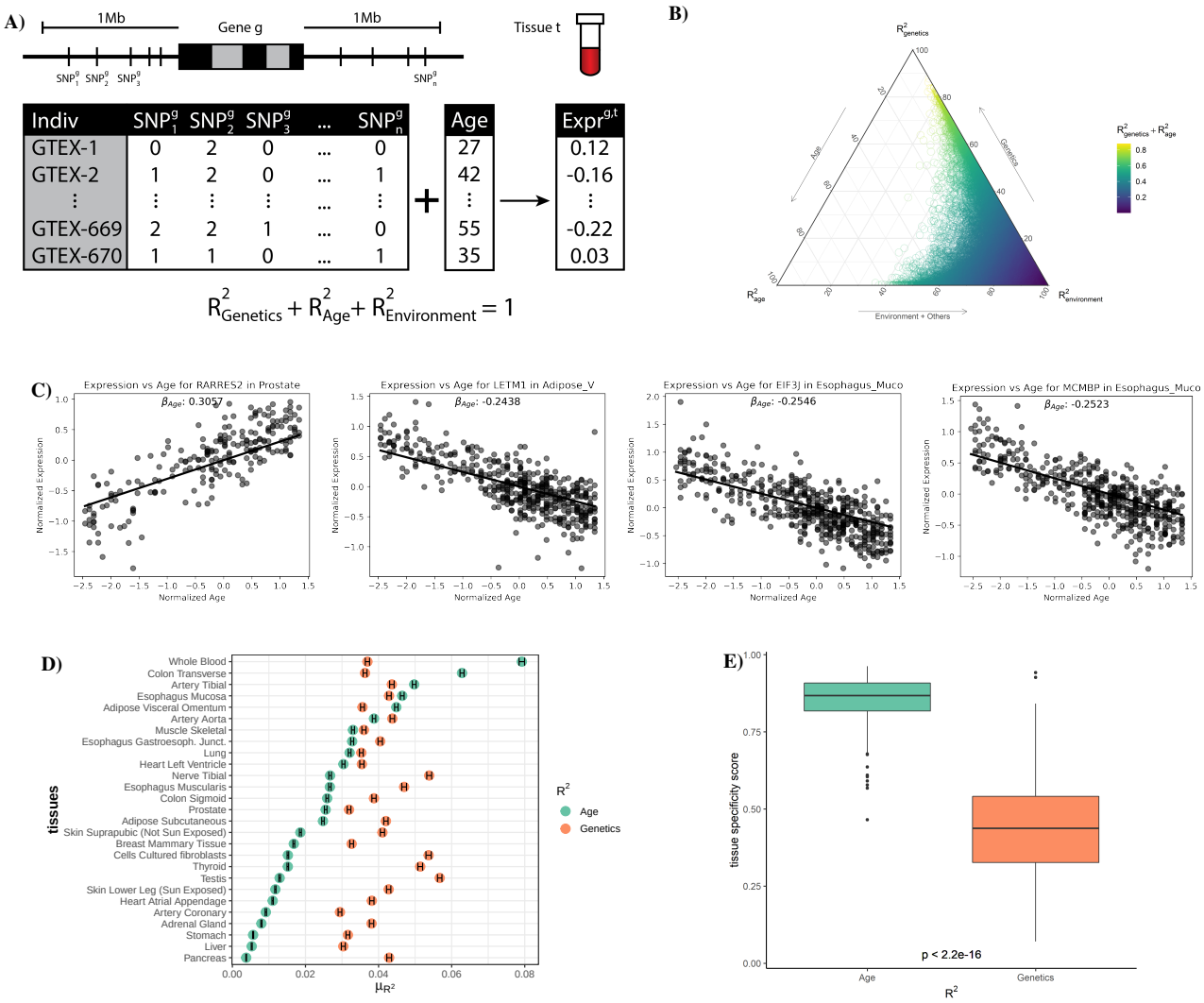

Fig. 3. A joint predictive model of gene expression identifies tissue-specific contributions of age and genetics to transcript levels. A) A schematic of our multi-SNP gene expression association model incorporating sample age. Common SNPs around each gene *g* are used in combination with an individual's age to predict expression within tissue *t*. Using this trained model, variation in gene expression can be separated into three parts: the components explained by genetics (R^2_{genetics} or h^2) , by age (R^2_{age}) and by all other factors ($R^2_{\text{environment}}$). **B)** Ternary plot of the proportion of each genes expression variance explained by age , genetics and the environment. **C)** Plot of normalized expression vs age for four genes with significant age-correlated expression. Line shows fitted β_{age} from regularized linear model. **D)** Point estimates of the mean R^2_{age} and h^2 for each tissue, error bar indicating confidence interval for the estimate. **E)** The tissue specificity score of R^2 across 27 tissues for each gene from either age genetics.

(18). As expected, across the vast majority of tissues β_{age} was significantly negatively correlated with pLI (Fig. 5B, 5C), in line with Medawar's hypothesis. However, five tissues exhibited significant signatures in the opposite direction including prostate, transverse colon, breast, whole blood, and lung tissue ($P < 10^{-3}$). These five tissues with *non-Medawarian* trends are driven by highly constrained, functionally important genes being expressed at a higher rate in older individuals (Fig. S14). Using dN/dS (19) as an alternative metric of gene constraint yielded highly correlated results (R=-0.72, P=2.5e-5 Fig. S15, S16).

To explore why these five tissues might exhibit distinctive evolutionary signatures of aging we compared the distribution of significant β_{age} parameters between *Medawarian* and *non-Medawarian* tissues among different *hallmark pathways* (20). We found 11 signatures exhibiting significantly increased β_{age} (FDR<0.01) compared to *non-Medawarian* tissues (Fig. 5D, 5E) including DNA-damage, TGF-β signalling, MYC targets, and epithelial-to-mesenchymal transition pathways most prominently. All of these signatures are broadly correlated with cellular proliferation, differentiation, and cancer. These results highlight that gene expression patterns in tissues and cell-types that proliferate throughout the course of an individuals life may be subjected to distinct evolutionary pressures.

We also explored the relationship between gene expression heritability and constraint. Across all tissues h^2 was significantly negatively correlated with pLI (27/27 tissues, P-value $< 10^{-3}$) (Fig. 5F, S17). While this trend was consistent across tissues, intriguingly it was strongest in heart tissues. The exception was liver, which also had the highest average $R^2_{\text{environment}}$ among all tissues, which was only nominally significant after multiple test correction (P<0.00185). These result indicate that genes in which the variance in expression

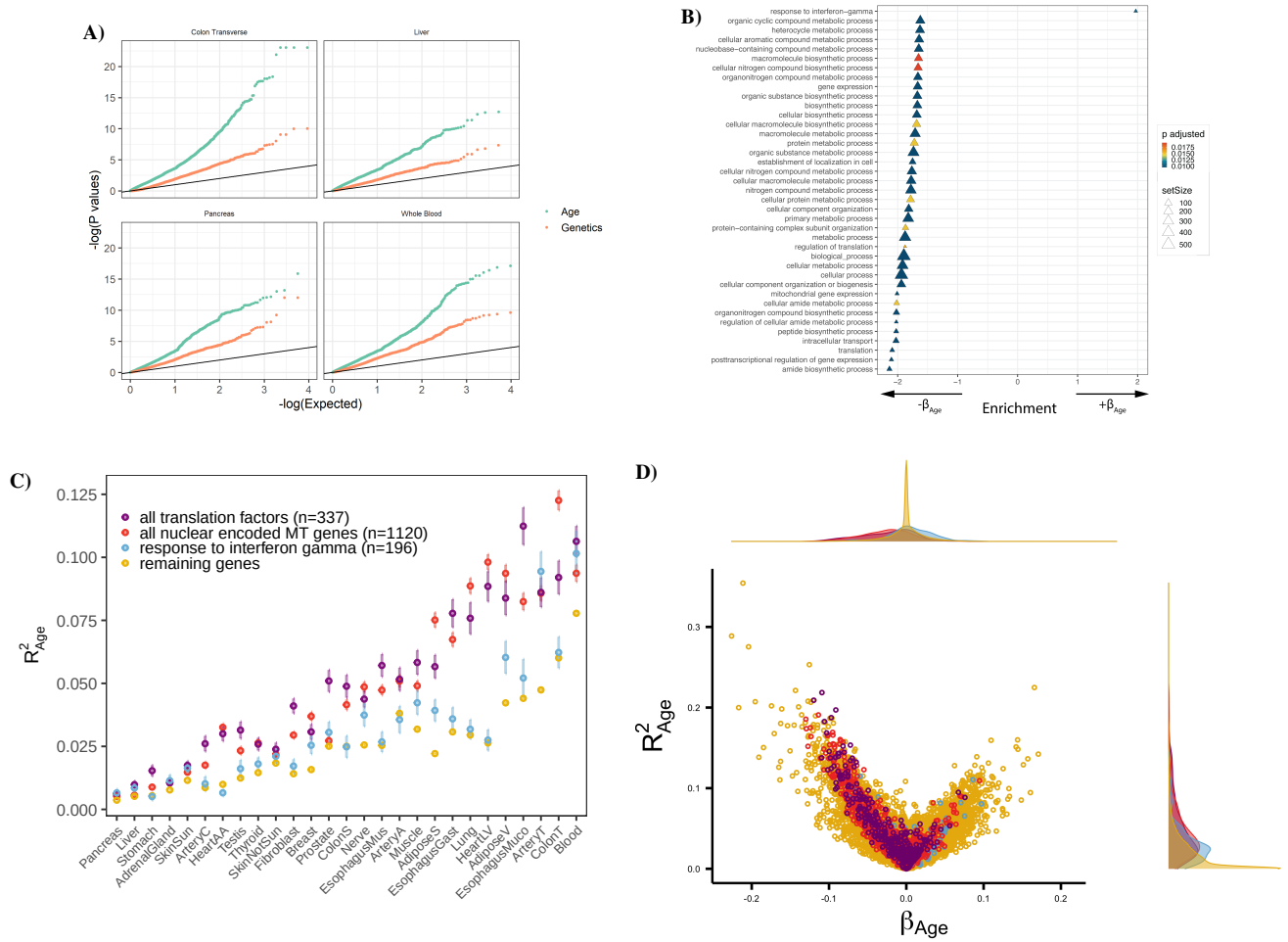

Fig. 4. Functional analysis of age-related genes reveals enriched biological processes. A) A QQ plot of p-values for pathways tested for enrichment using gene set enrichment analysis (GSEA) with genes ranked by either h^2 or R^2_{age} in four example tissues. **B)** GSEA enrichments from genes ranked by the mean β_{age} across tissues. Pathways with a correct $P < 0.02$ are shown. **C)** Average gene expression variance explained by age for mitochondrial (MT) genes (red), translation factor genes (purple), interferon gamma genes (blue) and remaining genes (yellow) across all tissues. **D)** Volcano plot of the variance explained by age vs β_{age} for mitochondrial, translation factors, interferon gamma factors, and remaining genes. Density plot of each axis show on top and right.

is heritable tend to be under significantly less functional constraint. In contrast, highly conserved genes that are intolerant to mutation are significantly less likely to exhibit heritable variation in gene expression, likely because their expression levels are additionally under constraint.

Discussion

Studying age-associated changes in gene expression provides critical insights into the underlying biological processes of aging. Here, we set out to quantify the relative contributions of aging and genetics to gene expression phenotypes across different human tissues. Our study finds that the predictive power of eQTLs is significantly impacted by age across several different tissues and that his effect is more pronounced in older individuals. These results extend upon previous work examining blood tissue (4) and highlight the varied impact of aging on eQTLs among different tissues. We show that this result is likely to be in part due to an increase in the inter-individual heterogeneity of gene expression patterns among older individuals, potentially as a result of the increased impact of the environment. However, our study is limited in it's focus on bulk-tissue transcriptomic data. Early evidence from single cell studies already suggests that differences in gene expression heterogeneity vary among cell types of tissues as a function of age (6, 7, 21, 22). While these studies lack sufficient individual sample sizes and genetic diversity for the statistical approaches used herein, it is possible that in the future the availability of larger datasets will facilitate studying these phenomena at the single-cell level. The extensive tissue heterogeneity we observe suggests that patterns of aging will exhibit substantial cell-type specificity.

We also present a novel approach to jointly model the impact of genetics and aging on gene expression variance to parse out the individual contributions of each of these factors. The increased complexity of our model has little impact on its accuracy with our expression heritability estimates strongly correlated with previous heritability measures across all tissues (mean Pearson's r 0.89, Fig. S7). Using this model we show that age exhibits exceptionally varied affects on different tissues, and indeed, in several tissues age contributes more to gene expression variance on average than genetics. These results also highlight a widespread coordinated signature of age-associated decline in mitochondrial and trans-

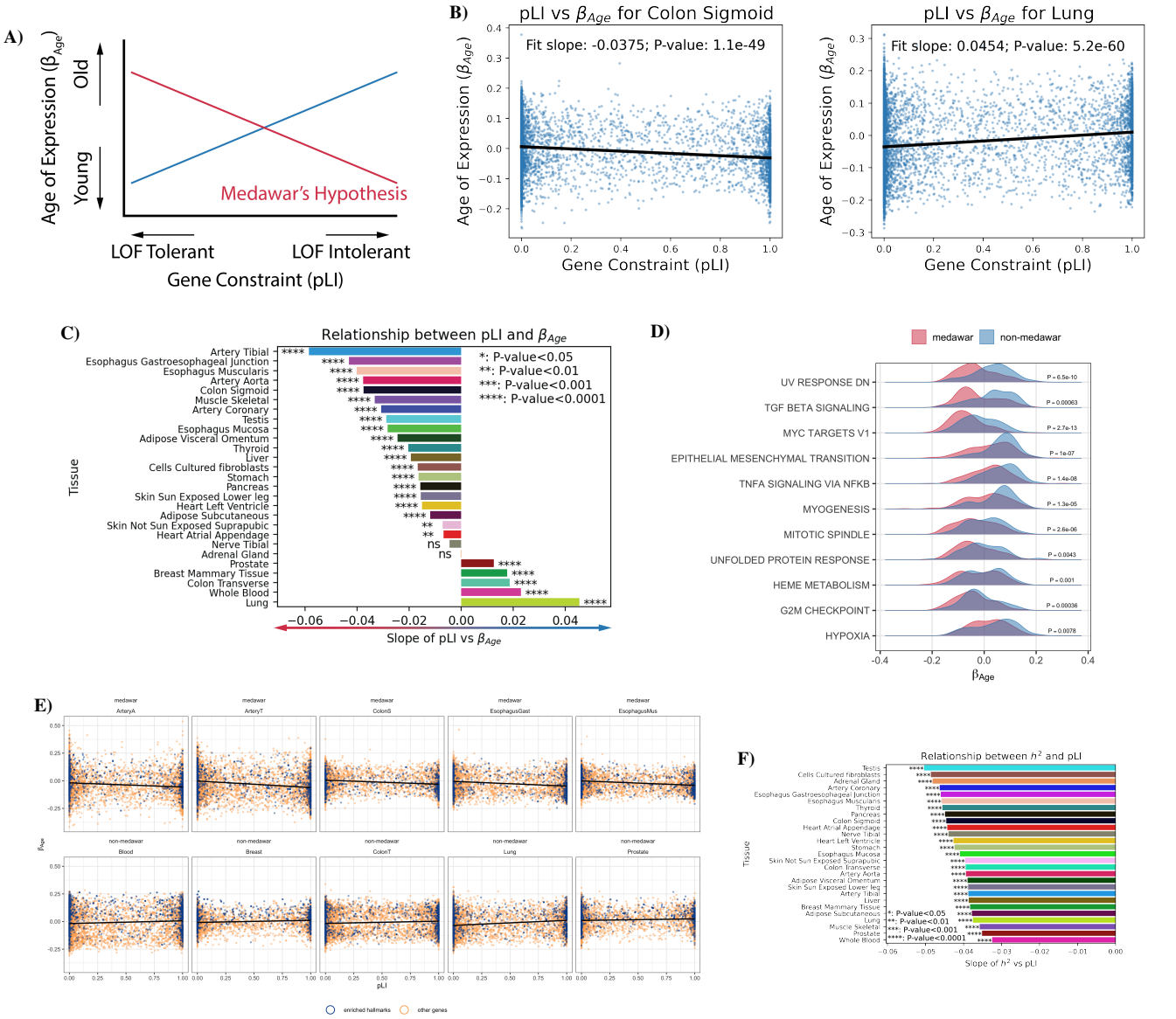

Fig. 5. Tissue-specific evolutionary signatures of aging. **A)** The expected relationship across genes between the per-gene age-associated slope of gene expression (β_{age}) and a genes level of constraint (measured by probability loss of function intolerance - pLI) . Medawar's hypothesis predicts a negative relationship (shown in red) between the time of expression and the level of constraint. **B)** β_{age} across genes plotted as a function of pLI for a tissue exhibiting a *Medawarian* signature, and a *non-Medawarian* signature. **C)** The slope of the relationship between β_{age} and constraint across all tissues. **D)** Hallmark pathways in which the β_{age} was significantly different between *Medawarian* and *non-Medawarian* tissues. **E)** The relationship between β_{age} and pLI among tissues showing the strongest *Medawarian* and *non-Medawarian* signatures with genes in pathways from (**D**) highlighted in blue. **F)** The slope of the relationship between constraint (pLI) and heritability (h^2) across tissues.

lation factors. Dysregulation in mitochondrial function and ribosome biogenesis have been documented as key players in aging, (23, 24), however our results highlight the tissue-specificity of these trends. Our model also allows us to quantify the tissue-specific evolutionary context of age-associated gene expression changes. We corroborate the inverse relationship between age-at-expression and constraint, as predicted by Medawar's hypothesis and recently documented by others (8, 9, 17) across the vast majority of tissues. However, we also surprisingly identify five tissues which exhibit the opposite pattern and show that age-associated signatures of increased proliferation and cancer are enriched in these tissues. These results highlight the distinct evolutionary forces that act on late-acting genes expressed in highly proliferative cell-types. Future work extending these analyses to the single-cell level will provide further insights into the cell-type-specific age-associated patterns of constraint, both in terms of gene expression levels and at the protein-coding level.

Overall this work has several important implications. Our results shed light on recent work on the prediction accuracy of polygenic risk scores (PRS) (25) which found that numerous factors, including age, sex, and socioeconomic status can profoundly impact the prediction accuracy of such scores even in individuals with the same genetic ancestry. Our results highlight that genetics are less predictive of expression phenotypes in several different tissues in older individuals, potentially playing a role in differential PRS accuracy between young and old individuals. This also has important implica-

tions for disease association and prediction approaches that leverage expression quantitative trait loci to prioritize variants (e.g. TWAS (26)). If a significant proportion of eQTLs exhibit age-associated biases in their effect size in a tissue of interest, then these approaches may be less powerful when applied to diseases for which age is a primary risk factor such as heart disease, Alzheimer's dementias, cancers, and diabetes.

The critical role of aging as a risk factor for many common human diseases underscores the importance of understanding its impact on cellular systems at the molecular level. Together our analyses provide novel insights into tissue-specific patterns of aging and the relative impact of genetics and aging on gene expression. We anticipate that future studies across tissues and cells of gene expression, chromatin structure, and epigenetics will further elucidate how both programmed and stochastic processes of aging drive human disease.

Supplementary Note 1: Methods

Data collection age groupings. We downloaded gene expression data for multiple individuals and tissues from GTEx V8 (10), which were previously aligned and processed against the hg19 human genome. Tissues were included in the analysis if they had >100 individuals in both the age ≥ 55 and <55 cohorts described below (Fig. S2). To compare gene expression heritability across individuals of different ages, for some analyses we split the GTEx data for each tissue into two age groups, "young" and "old," based on the median age of individuals in the full dataset, which was 55 (Fig. S1). Within each tissue dataset, we then equalized the number of individuals in the young and old groups by randomly downsampling the larger group, to ensure that our models were equally powered for the two age groups.

PEER factor analysis. We analyzed existing precomputed PEER factors available from GTEx to check for correlations between these hidden covariates and age. In particular, we fit a linear regression between age and each hidden covariate and identified significant age correlations using an F-statistic (Fig. S3). Because some of the covariates were correlated with age, we generated new age-independent hidden covariates of gene expression to remove batch and other confounding effects on gene expression while retaining age related variation. In particular, we first removed age contributions to gene expression by regressing gene expression on age and then ran PEER on the age-independent residual gene expression to generate 15 age-independent hidden PEER factors.

Quantifying the effect of eQTLs on gene expression in different age groups. Using the binary age groups defined above, we assessed the relative significance of eQTLs in old and young individuals by carrying out separate assessment of eQTLs identified by GTEx. For each gene in each tissue and each age group, we regressed the GTEx pre-normalized expression levels on the genotype of the lead SNP (identified by GTEx) using 5 PCs, 15 PEER factors, sex, PCR protocol and sequencing platform as covariates, following the GTEx best practices. We confirmed our results using both our recomputed PEER factors as well as the PEER factors provided by GTEx (Fig. S 4). To test for significant differences in genetic associations with gene expression between the old and young age groups, we compared the p-value distributions between these groups for all genes and all SNPs in a given tissue using Welch's t-test.

Jensen-Shannon Divergence as a distance metric between transcriptome profiles. To quantify differences in gene expression between individuals, we computed the pairwise distance for all pairs of individuals in an age group using the square root of Jensen-Shannon Divergence (JSD) distance metric, which measures the similarity of two probability distributions. Here we applied JSD between pairs of individuals' transcriptome vectors containing the gene expression values for each gene, which we converted to a distribution by normalizing by the sum of the entries in the vector. For two individuals' transcriptome distributions, the JSD can be calculated as:

$$\mathrm{JSD}(P_1, P_2) = H(\frac{1}{2}P_1 + \frac{1}{2}P_2) - \frac{1}{2}(H(P_1) + H(P_2)) \quad \textbf{(1)}$$

where P_i is the distribution for individual i and H is the Shannon entropy function:

$$H(X) = -\sum_{i=1}^{n} P(x_i)\log_2(P(x_i)) \quad \textbf{(2)}$$

JSD is known to be a robust metric that is less sensitive to noise when calculating distance compared to traditional metrics such as Euclidean distance and correlation. It has been shown that JSD metrics and other approaches yield similar results but that JSD is more robust to outliers (12). The square root of the raw JSD value follows the triangle inequality, enabling us to treat it as a distance metric.

Slope of JSD versus age. In addition to comparing JSD between the two age groups defined above, "young" and "old", we also binned all GTEx individuals into 6 age groups, from 20 to 80 years old with an increment of 10 years. We then computed pairwise distance and average age for each pair of individuals within each bin using the square root of JSD as the distance metric. We applied a linear regression model of JSD versus age to obtain slopes, confidence intervals, and p-values.

Multi-SNP gene expression prediction. We used a multi-SNP gene expression prediction model based on PrediXcan (14) to corroborate our findings from the eQTL and JSD analyses on the two age groups, "young" and "old". For each gene in each tissue, we trained a multi-SNP model separately within each age group to predict individual-level gene expression.

$$Y_{g,t} = \sum_i \beta_{i,g,t} X_i + \epsilon \quad \textbf{(3)}$$

Where $\beta_{i,g,t}$ is the coefficient or effect size for SNP X_i in gene g and tissue t and ϵ includes all other noise and environmental effects. The regularized linear model for each gene

considers dosages of all common SNPs within 1 megabase of the gene's TSS as input, where common SNPs are defined as MAF > 0.05 and Hardy-Weinberg equilibrium P > 0.05. We removed covariate effects on gene expression prior to model training by regressing out both GTEx covariates and age-independent PEER factors (described above). Coefficients were fit using an elastic net model which solves the problem ((27)):

$$\min_{\beta_0,\beta} \frac{1}{2N}\sum_{j=1}^{N}(Y_j-\beta_0-X_j^T\beta)^2+\lambda(\frac{1-\alpha}{2}||\beta||_2^2+\alpha||\beta||_1) \quad \textbf{(4)}$$

The minimization problem contains both the error of our model predictions $(Y_j-\beta_0-X_j^T\beta)^2$ and a regularization term $\lambda(\frac{1-\alpha}{2}||\beta||_2^2+\alpha||\beta||_1)$ to prevent model overfitting. The elastic net regularization term incorporates both L1 ($||\beta||_1$)) and L2 ($||\beta||_2^2$) penalties. Following PrediXcan, we weighted the L1 and L2 penalties equally using $\alpha = 0.5$ (14). For each model, the regularization parameter λ was chosen via 10-fold cross validation. The elastic net models were fit using Python's glmnet package and R^2 was evaluated using scikit-learn. From the trained models for each gene, we evaluated training set genetic R^2 (or h^2) for the two age groups and subtracted $h^2_{\text{young}} - h^2_{\text{old}}$ to get the difference in gene expression heritability between the groups. We compared this average difference in heritability to the mean $\text{JSD}_{old} - \text{JSD}_{young}$ and $\log(P_{old}) - \log(P_{young})$ using P-values from the eQTL analyses across genes.

Joint model for expression prediction using SNPs and age. To uncover linear relationships between gene expression and both age and genetics, we built a set of gene expression prediction models using both common SNPs and standardized age as input. An individual's gene expression level Y for a gene g and tissue t is modeled as:

$$Y_{g,t} = \sum_i \beta_{i,g,t}X_i + \beta_{\text{age},g,t}A + \epsilon \quad \textbf{(5)}$$

Where A is the normalized age of an individual. Coefficients were fit using elastic net regularization, as above, which sets coefficients for non-informative predictors to zero. The sign of the fitted age coefficient ($\beta_{age,g,t}$), when nonzero, reflects whether the gene in that tissue is expressed more in young (negative coefficient) or old (positive coefficient) individuals. We also evaluated the training set R^2 using the fit model separately for genetics (across all SNPs in the model) and age. To check consistency of tissue-specific gene expression heritability estimates from our model and the original PrediXcan model trained on GTEx data, we evaluate Pearson's r between our heritability estimates and those of PrediXcan (Fig. S7), using heritability estimates from the original PrediXcan model available in PredictDB.

Tissue specificity of age and genetic associations. We evaluated the variability of age and genetic associations across tissues using a measure of tissue specificity for age and genetic R^2 (28). We measured the tissue-specificity of a gene g's variance explained R^2_g using the following metric:

$$S_g = \frac{\sum_{t=1}^{n}(1-\frac{R^2_{g,t}}{R^2_{g,\max}})}{n-1} \quad \textbf{(6)}$$

Where n is the total number of tissues, $R^2_{g,t}$ is the variance explained by either age or genetics for the gene g in tissue t and $R^2_{g,\max}$ is the maximum variance explained for g over all tissues. This metric can be thought of as the average reduction in variance explained relative to the maximum variance explained across tissues for a given gene. The metric ranges from 0 to 1, with 0 representing ubiquitously high genetic or age R^2 and 1 representing only one tissue with nonzero genetic or age R^2 for a given gene. We calculate S_g separately for R^2_{age} and R^2_{genetics} across all genes.

Functional constraint analysis. We quantified gene constraint using probability of loss of function intolerance (pLI) from gnomAD 2.1.1 (18). We analyzed the relationships between pLI vs β_{age} and pLI vs heritability across genes. For these analyses, genes were only included if age or genetics were predictive of gene expression ($R^2 > 0$) for that gene. For genes with $R^2 > 0$, we used linear regression to determine the direction of the relationship between pLI and β_{age} and heritability for each tissue. The F-statistic was used to determine whether pLI was significantly related to these two model outputs. For pLI vs β_{age}, a significant negative slope was considered a Medawar trend (consistent with the Medawar hypothesis) and a significant positive slope a non-Medawar trend. We also analyzed the evolutionary constraint metric dN/dS (19) and its tissue-specific relationship with β_{age} by determining the slope and significance of the linear regression, as above.

Non-Medawar tissue analysis. To explore the non-Medawar trend in some tissues, we assessed the distribution of β_{age} across Medawar and non-Medawar tissues for genes within each of the 50 MSigDB hallmark pathways (20). Significant differences between the distributions were called using a t-test, and p-values were adjusted for multiple hypothesis testing using a Benjamini-Hochberg correction.

Code availability. All analyses were performed in R version 4.0.2 and Python 3.6. All code is available online at https://github.com/sudmantlab/gene_expression_aging.

ACKNOWLEDGEMENTS

This work was supported by the National Institute of General Medical Sciences grant R35GM142916 to P.H.S. and the National Human Genome Research Institute grant R00HG009677 to N.M.I.

AUTHOR CONTRIBUTIONS

RY, RC, JMV, HS, PS, and PHS performed all analysis. RY, RC, NMI, and PHS wrote the manuscript. PHS and NMI supervised the project. PHS conceived of the project.

Bibliography

1. B Charlesworth, Fisher, medawar, hamilton and the evolution of aging. *Genetics* **156**, 927–931 (2000).
2. MR Rose, CL Rauser, G Benford, M Matos, LD Mueller, HAMILTONS FORCES OF NATURAL SELECTION AFTER FORTY YEARS. *Evolution* **61**, 1265–1276 (2007).

3. A Viñuela, et al., Age-dependent changes in mean and variance of gene expression across tissues in a twin cohort. *Hum. Mol. Genet.* **27**, 732–741 (2017).
4. B Balliu, et al., Genetic regulation of gene expression and splicing during a 10-year period of human aging. *Genome Biol.* **20** (2019).
5. M Somel, P Khaitovich, S Bahn, S Pääbo, M Lachmann, Gene expression becomes heterogeneous with age. *Curr. Biol.* **16**, R359–R360 (2006).
6. S Wang, et al., Single-cell transcriptomic atlas of primate ovarian aging. *Cell* **180**, 585–600.e19 (2020).
7. CP Martinez-Jimenez, et al., Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science* **355**, 1433–1436 (2017).
8. C Cheng, M Kirkpatrick, Molecular evolution and the decline of purifying selection with age. *Nat. Commun.* **12** (2021).
9. K Jia, C Cui, Y Gao, Y Zhou, Q Cui, An analysis of aging-related genes derived from the genotype-tissue expression project (GTEx). *Cell Death Discov.* **4** (2018).
10. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
11. O Stegle, L Parts, R Durbin, J Winn, A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol.* **6**, e1000770 (2010).
12. PH Sudmant, MS Alexis, CB Burge, Meta-analysis of RNA-seq expression data across species, tissues and studies. *Genome Biol.* **16** (2015).
13. P Sen, PP Shah, R Nativio, SL Berger, Epigenetic mechanisms of longevity and aging. *Cell* **166**, 822–839 (2016).
14. ER Gamazon, et al., A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
15. A Subramanian, et al., Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).
16. S Rath, et al., Mitocarta3.0: an updated mitochondrial proteome now with sub-organelle localization and pathway annotations. *Nucleic Acids Res.* **49**, D1541–D1547 (2020).
17. R Cui, et al., Relaxed selection limits lifespan by increasing mutation load. *Cell* **178**, 385–399.e20 (2019).
18. M Lek, et al., Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
19. M Gayà-Vidal, M Albà, Uncovering adaptive evolution in the human lineage. *BMC Genomics* **15**, 599 (2014).
20. A Liberzon, et al., The molecular signatures database hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
21. N Almanzar, et al., A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* **583**, 590–595 (2020).
22. P Cheung, et al., Single-cell chromatin modification profiling reveals increased epigenetic variations with aging. *Cell* **173**, 1385–1397.e14 (2018).
23. S Srivastava, The mitochondrial basis of aging and age-related disorders. *Genes* **8**, 398 (2017).
24. S Tahmasebi, A Khoutorsky, MB Mathews, N Sonenberg, Translation deregulation in human disease. *Nat. Rev. Mol. Cell Biol.* **19**, 791–807 (2018).
25. H Mostafavi, et al., Variable prediction accuracy of polygenic scores within an ancestry group. *eLife* **9** (2020).
26. M Wainberg, et al., Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* **51**, 592–599 (2019).
27. JH Friedman, T Hastie, R Tibshirani, Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 122 (2010).
28. I Yanai, et al., Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2004).

Supplementary Figures

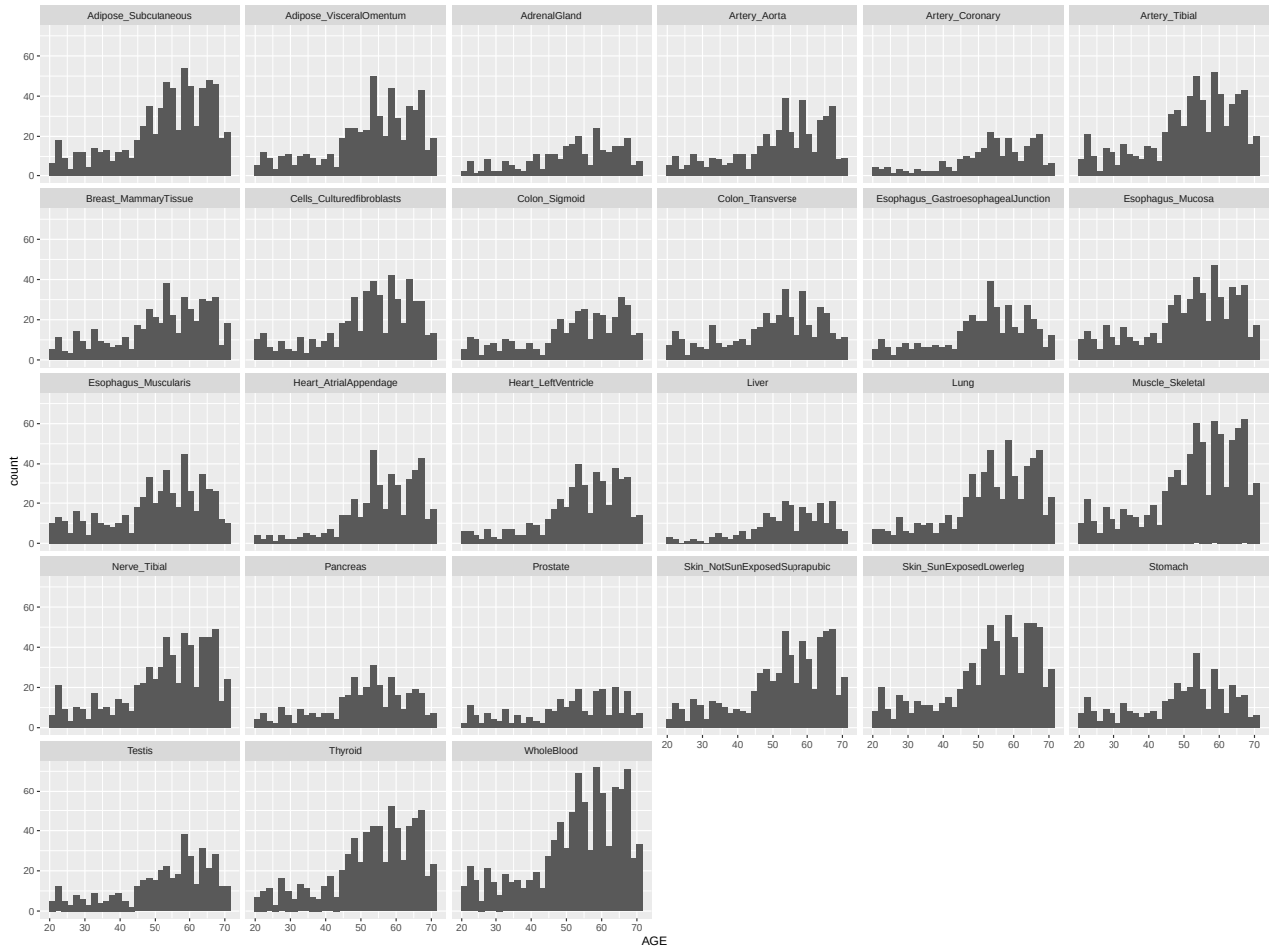

Fig S 1. Sample distribution of each GTEx tissue by age

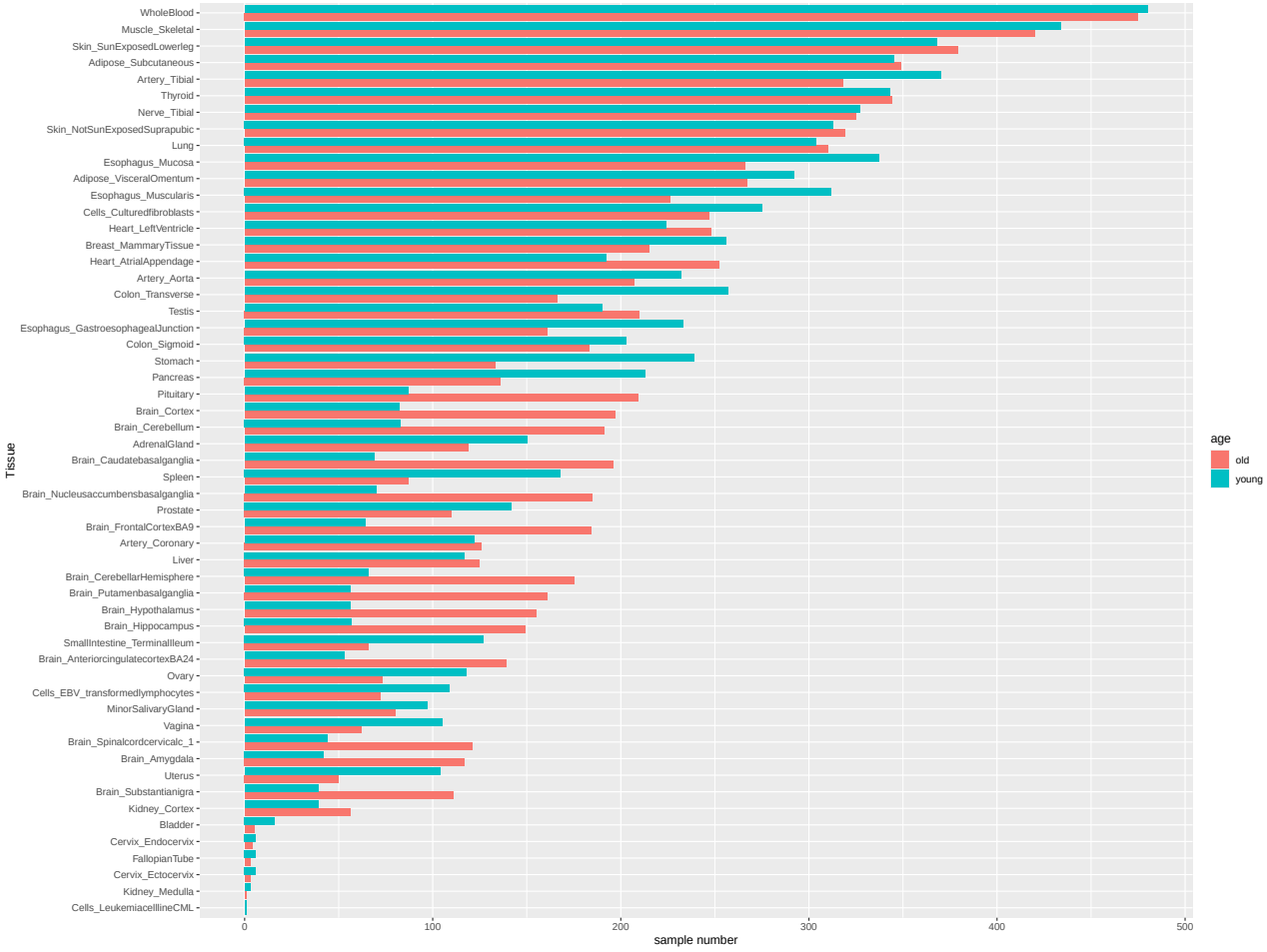

Fig S 2. Sample number for individuals above and below the median age by tissue for 47 GTEx tissues

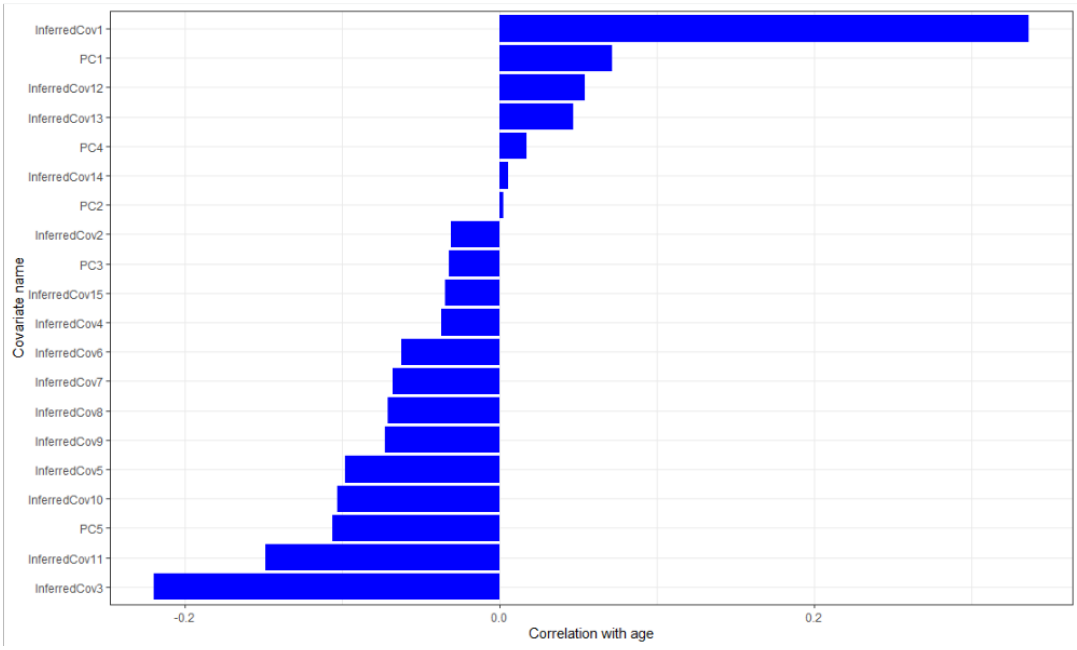

Fig S 3. Correlation of GTEx covariates with age.

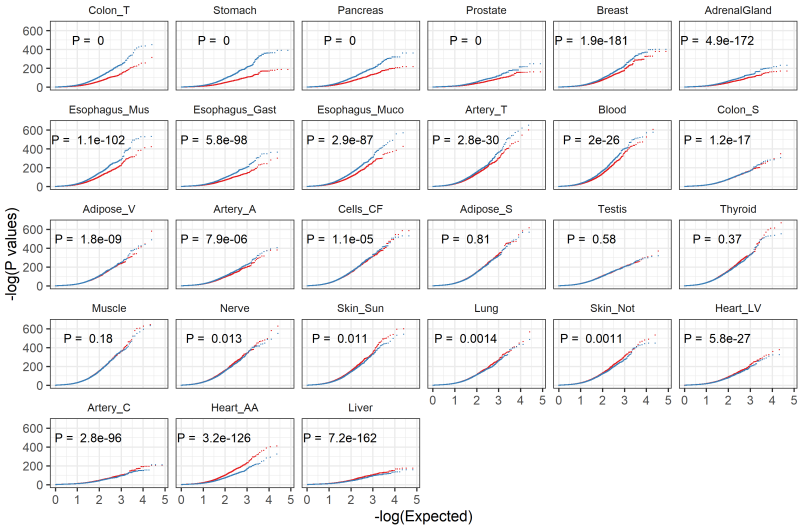

Fig S 4. QQ plot of eQTL p-values for old (red) and young (blue) cohorts across 27 tissues using GTEx PEER factors. Significant differences between p-value distributions annotated for each tissue.

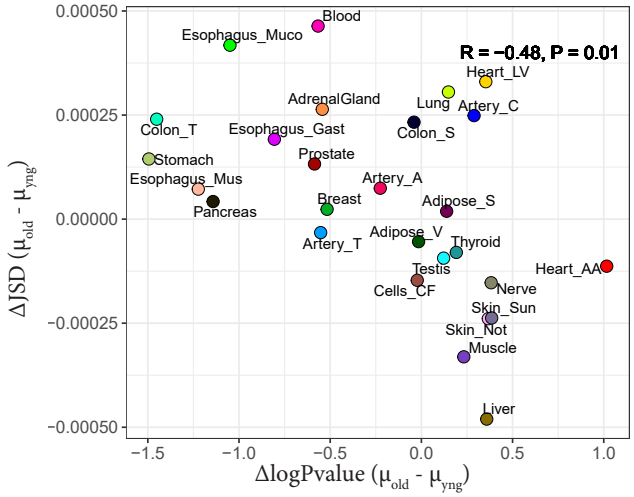

Fig S 5. Scatter plot showing the correlation between the difference in average JSD between young and old individuals and difference in eQTL log(p-value) between young and old

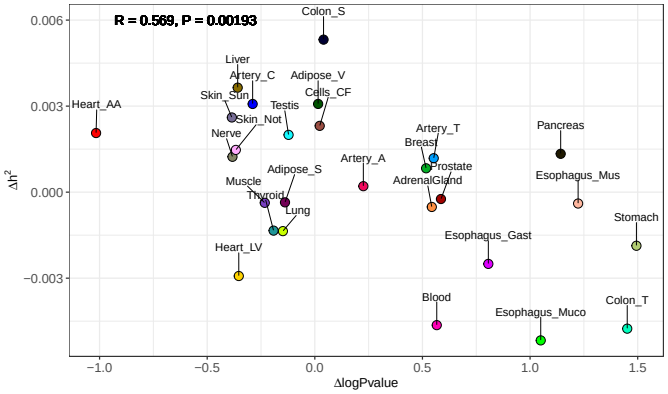

Fig S 6. High correspondence of age-related changes in gene expression associations for single SNP and multi-SNP models. Scatter plot showing the correlation between the difference in heritability estimated by the multi-SNP linear model (PrediXcan) in young and old individuals and difference in eQTL log(p-value) in young and old individuals

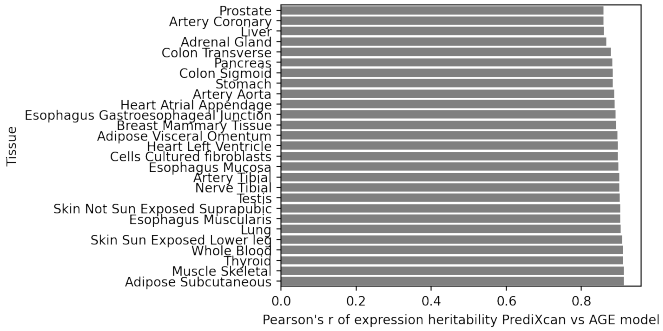

Fig S 7. Pearson's r of heritability estimate from PrediXcan (PredictDB) vs our model for each tissue

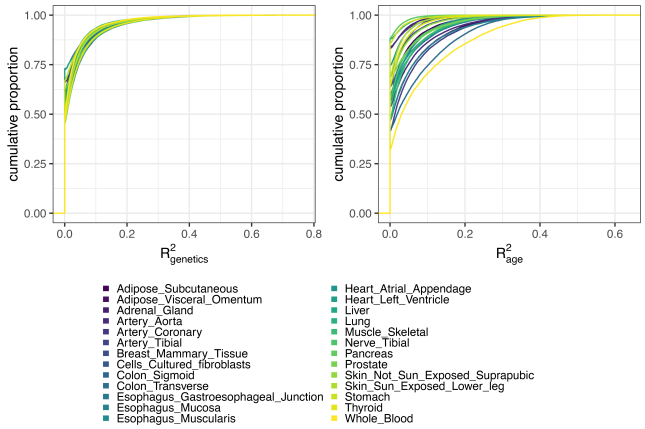

Fig S 8. Cumulative distribution of R^2_{age} and h^2 for all modeled genes within 27 tissues.

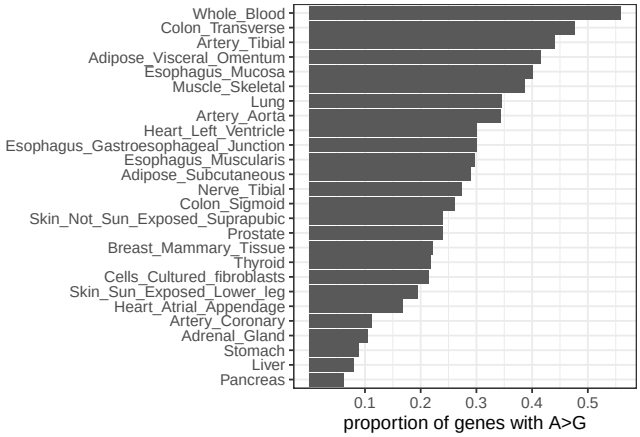

Fig S 9. Proportion of genes within a tissue that have $R^2_{\text{age}} > h^2$

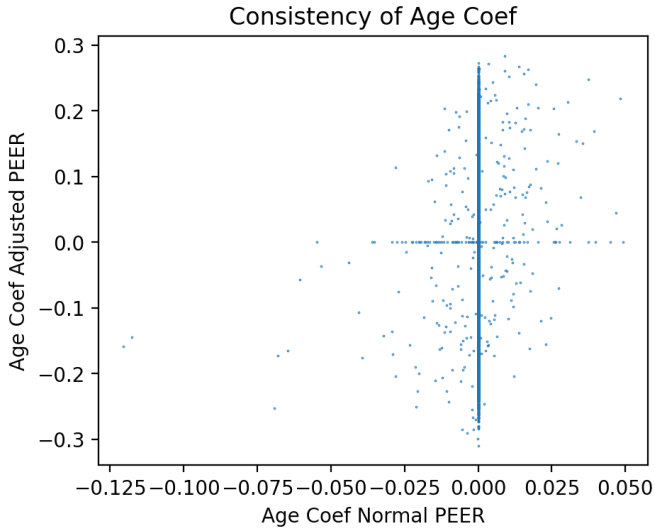

Fig S 10. Scatter plot of each gene's β_{age} of multiSNP model using GTEx PEER factors vs age-independent PEER factors for Whole Blood

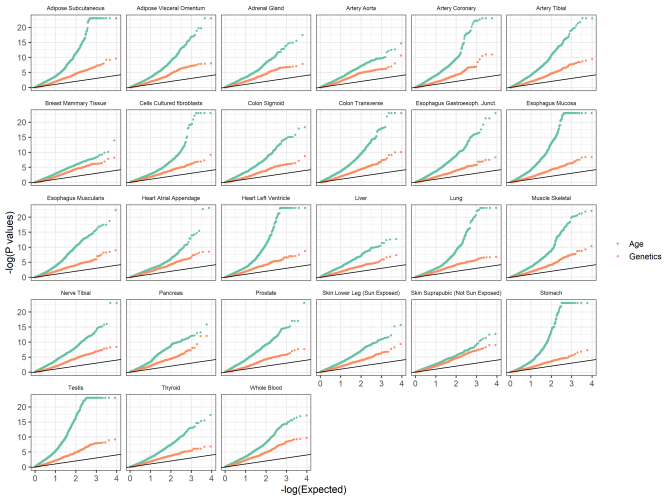

Fig S 11. GO gene set enrichment P-values across all tissue for genes ranked by either h^2 or R^2_{age}

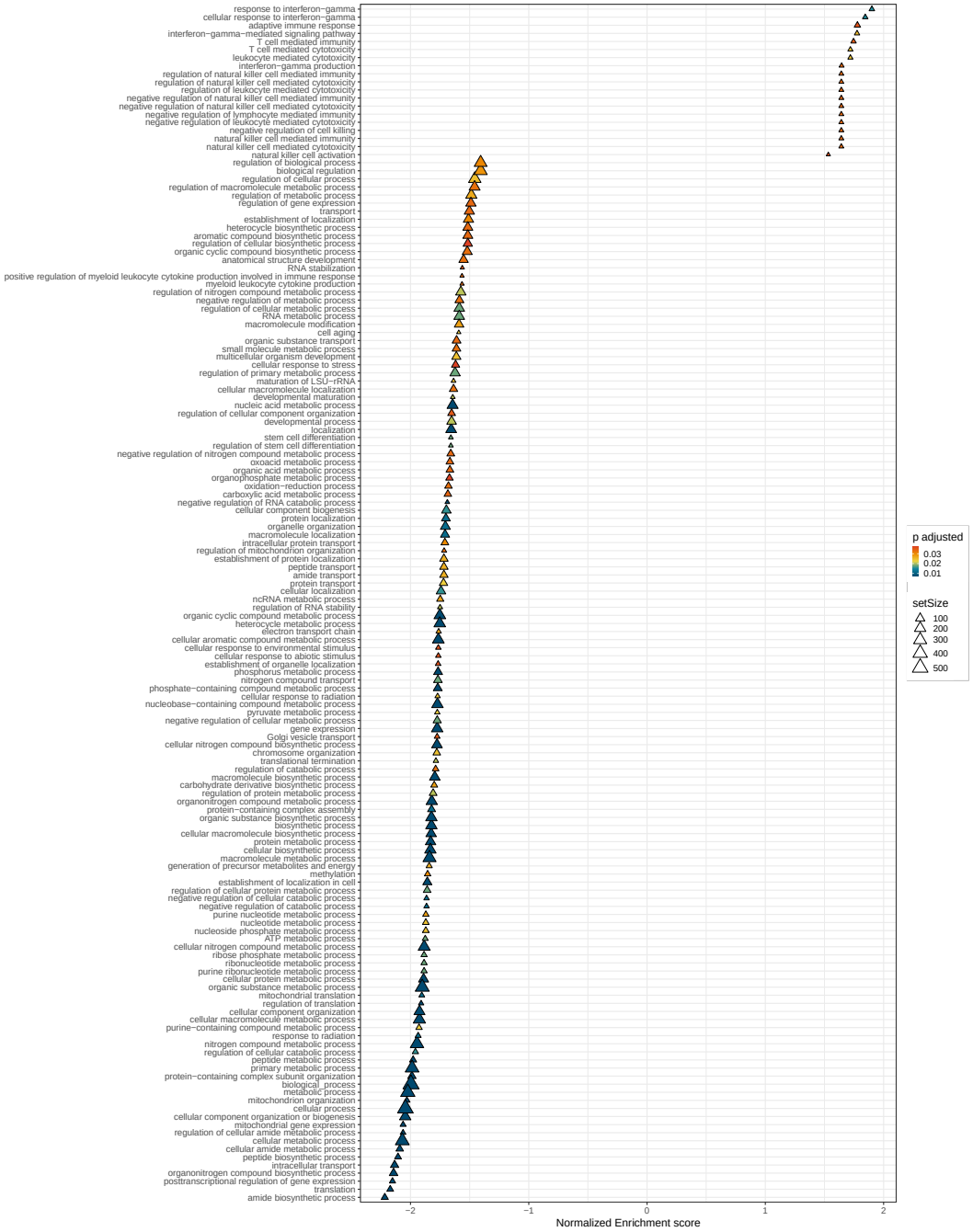

Fig S 12. GO Biological Processes common enrichment in all tissues with genes ranked by average β_{age} (FDR 0.05)

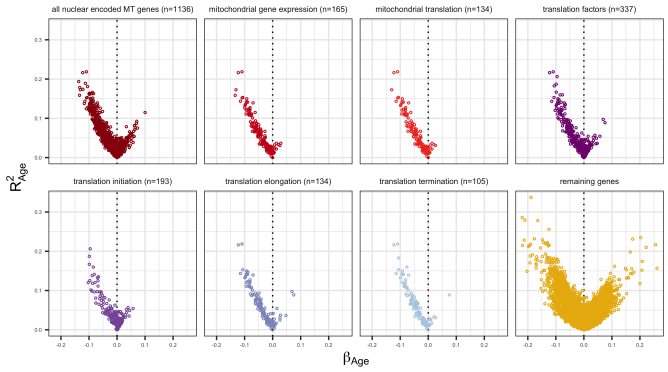

Fig S 13. Relationship between the average β_{age} and average R^2 across tissues for genes associated with specific mitochondrial and translation pathways

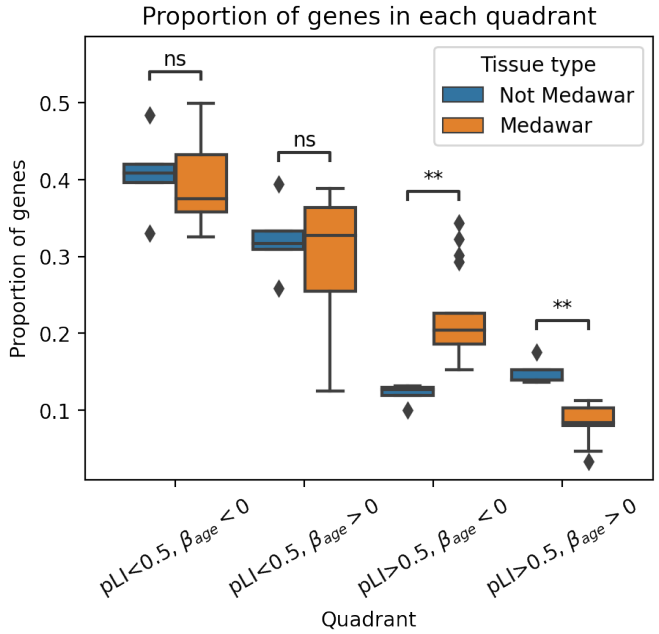

Fig S 14. More highly constrained, late expressed genes in non-Medawar tissues than in Medawar tissues. We plot the proportion of genes within each quadrant of the gene constraint (pLI) vs. age of expression (β_{age}) plots stratified by whether the tissue showed a significant Medawar or Non-Medawar trend.

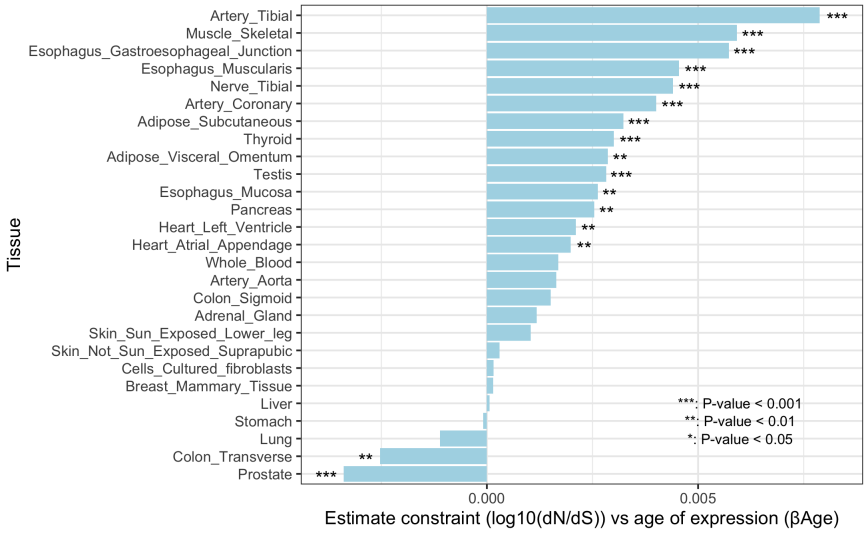

Fig S 15. Linear regression estimates of evolutionary constraint (dN/dS data from ortholog comparison between 8,175 human and chimpanzee genes) vs age-associated gene expression (β_{age}) across tissues for R^2_{age} > 0. 10/27 tissues have significant p-values (p-value < 0.001)

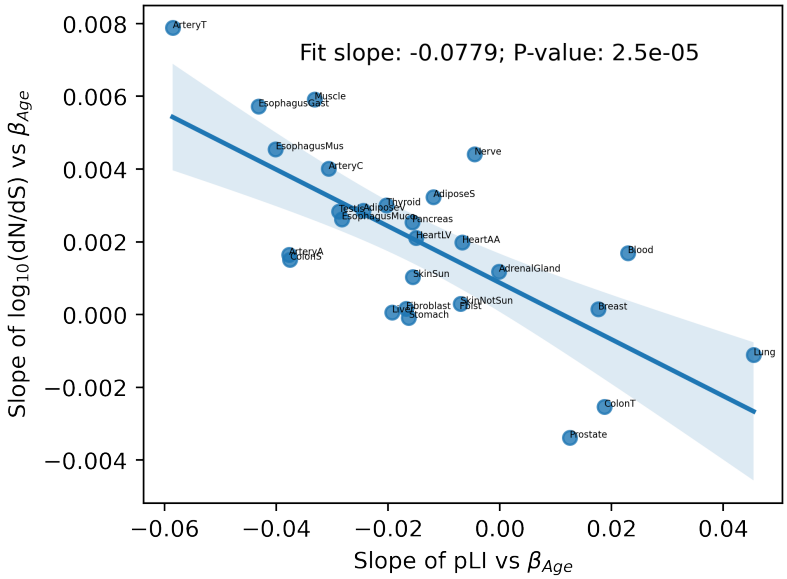

Fig S 16. Consistency of Medawarian trend measures. We plot the slope of gene constraint metrics (pLI and dN/dS) vs β_{age} for each tissue

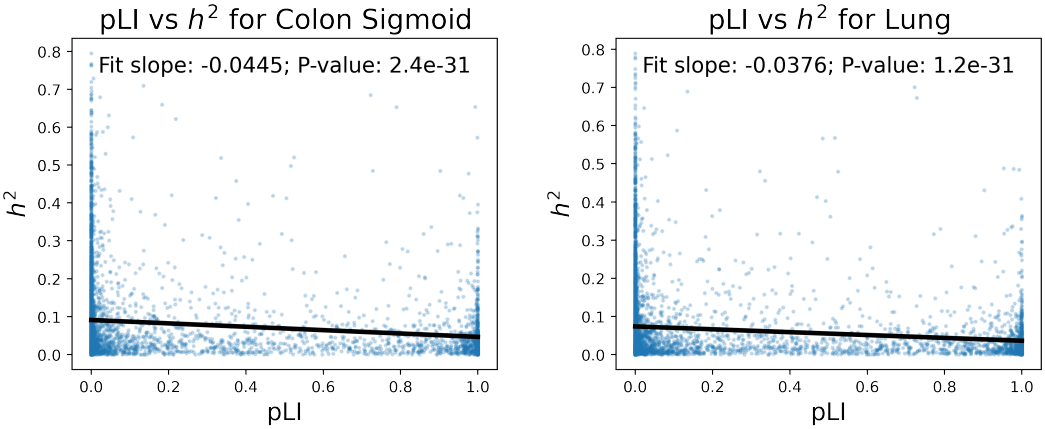

Fig S 17. Gene expression heritability vs pLI for two tissues

TISSUE	#(old individuals)	#(young individuals)
Adipose_Subcutaneous	349	345
Adipose_VisceralOmentum	267	292
AdrenalGland	119	150
Artery_Aorta	207	232
Artery_Coronary	126	122
Artery_Tibial	318	370
Breast_MammaryTissue	215	256
Cells_Culturedfibroblasts	247	275
Colon_Sigmoid	183	203
Colon_Transverse	166	257
Esophagus_GastroesophagealJunction	161	233
Esophagus_Mucosa	266	337
Esophagus_Muscularis	226	312
Heart_AtrialAppendage	252	192
Heart_LeftVentricle	248	224
Liver	125	117
Lung	310	304
Muscle_Skeletal	420	434
Nerve_Tibial	325	327
Pancreas	136	213
Prostate	110	142
Skin_NotSunExposedSuprapubic	319	313
Skin_SunExposedLowerleg	379	368
Stomach	133	239
Testis	210	190
Thyroid	344	343
WholeBlood	475	480

Table S 1. Number of old and young individuals per tissue

Tissue	Short name
Adipose_Subcutaneous	AdiposeS
Adipose_VisceralOmentum	AdiposeV
AdrenalGland	AdrenalGland
Artery_Aorta	ArteryA
Artery_Coronary	ArteryC
Artery_Tibial	ArteryT
Breast_Mammary_Tissue	Breast
Cells_Cultured_fibroblasts	Fibroblast
Colon_Sigmoid	ColonS
Colon_Transverse	ColonT
Esophagus_Gastroesophageal_Junction	EsophagusGast
Esophagus_Mucosa	EsophagusMuco
Esophagus_Muscularis	EsophagusMus
Heart_Atrial_Appendage	HeartAA
Heart_Left_Ventricle	HeartLV
Liver	Liver
Lung	Lung
Muscle_Skeletal	Muscle
Nerve_Tibial	Nerve
Pancreas	Pancreas
Prostate	Prostate
Skin_Not_Sun_Exposed_Suprapubic	SkinNotSun
Skin_Sun_Exposed_Lower_leg	SkinSun
Stomach	Stomach
Testis	Testis
Thyroid	Thyroid
Whole_Blood	Blood

Table S 2. Tissues with short names