Characterizing the landscape of gene expression variance in humans

Scott Wolf[†,1] Diogo Melo[†,1,2] Kristina M. Garske[1] Luisa F. Pallares[3]
Julien F. Ayroles[1,2,*]

Abstract

Gene expression variance has been linked to organismal function and fitness but remains a commonly neglected aspect of gene expression research. We lack a comprehensive view of the patterns of variance across genes, and how this variance is linked to context-specific gene regulation and gene function. Here, we use large publicly available human RNA-seq data sets to investigate the landscape of gene expression variance. In particular, we ask if there are consistently more or less variable genes across tissues and across data sets and what mechanisms drive these patterns. We show that gene expression variance is broadly similar across tissues and studies. We use this similarity to create both global and within-tissue rankings of variation, which we use to show that function, sequence variation, and gene regulatory signatures contribute to gene expression variance. Gene expression variance is strongly predictive of gene function, with low-variance genes being associated with fundamental cell processes, and high-variance genes being linked to responding to the environment. Our results show differences in the regulatory mechanisms of high and low gene expression variance, in addition to a clear link between function and gene expression variance, suggesting that these differences are adaptive. We expect these results will help to place the pattern of variation at the center of our understanding of molecular phenotypes.

† These authors contributed equally to this work.

[1] Lewis-Sigler Institute for Integrative Genomics, Princeton University
[2] Department of Ecology and Evolutionary Biology, Princeton University
[3] Friedrich Miescher Laboratory, Max Planck Society

* Correspondence: Julien F. Ayroles <jayroles@princeton.edu>

Introduction

Molecular phenotypes such as gene expression are powerful tools for understanding physiology, disease, and evolutionary adaptations. In this context, average trait values are usually the focus of investigation, while variation is treated as a nuisance [1]. However, gene expression variance can be directly involved in determining fitness [2,3], and changes in the associations between gene expression can be indicative of disease, even in the absence of changes in mean expression [4]. From an evolutionary perspective, the availability of gene expression variation is what allows evolutionary change, and the genetic architecture that controls gene expression variance can also evolve [5]. Focusing on the landscape of gene expression variance, and how variable it is across genes and across human populations is then a neglected avenue for understanding biological evolution and our relation to the environment. In particular, we lack a clear picture of which genes show relatively high or low variance in gene expression, or even if the pattern of gene expression variance is consistent across populations and tissues.

Several competing forces act to shape gene expression variance [5,6], and the outcome of the interaction between these processes is still poorly understood [7]. From a genomic perspective, we expect the influx of new mutations to increase observed variation, while the selective removal of polymorphisms, via purifying selection or selective sweeps, would decrease variation. From a trait-centric perspective, stabilizing selection should decrease variation around an optimal value, and directional selection can lead to a transient increase in variance while selected alleles sweep to fixation, followed by a reduction in variance as these alleles become

fixed. This simple picture is complicated by epistatic interactions between loci and other aspects of genetic architecture. For example, pleiotropic effects allow selection on one trait to influence the variance of other traits, potentially limiting the direct response to selection [8,9]. The indirect effect of directional selection on variance opens the possibility that the main driver of gene expression variance is not direct selection on variance but indirect effects due to selection on trait means [7]. Furthermore, gene-by-environment (GxE) interactions can also lead to changes in the observed phenotypic variance of gene expression, further complicating the landscape of variation. To what extent these processes shape gene expression variance is an open question. If homogeneous selection across groups is the main driver of gene expression variance, we would expect to have consistently more or less variable genes. If idiosyncratic selection patterns and context-specific environmental interactions are more important, we could observe large differences in gene expression variance across groups.

Even within individuals, gene expression is also variable across tissues [10]. To what extent differential expression (i.e., differences in mean expression level) translate into differences in expression variance is not clear. Of course, genes that are exclusively expressed in a single cell type or tissue are necessarily more variable in that particular tissue, but differentially expressed genes could also be more variable in specific contexts. For example, stabilizing selection on gene expression could be more intense depending on the role of that gene in a particular tissue, leading to a local reduction in variation that causes differences in variance across tissues. Alternatively, expression variation across tissues could be tightly coupled, and in this example, selection in one tissue would lead to a reduction in variance across tissues, resulting in a consistent pattern of variation. Alemu et al. [11] used microarray data from several human tissues to show that epigenetic markers were linked to gene expression variation and that these markers were variable across tissues and between high- and low-variance genes.

Here, we use publicly available human gene expression data sets to evaluate how the differences in gene expression variance are structured across independent samples. By comparing the gene expression variance measured across many studies, we show that the patterns of gene expression variance are broadly similar across studies and tissues. We used the observed similarities across these studies to create an across-study gene expression variance ranking, which orders genes from least variable to most variable. We then integrate various functional annotations as well as sequence variation to probe the drivers of this across-study ranking. Finally, we explore the link between gene expression variance and biological function by leveraging gene ontology and disease annotations.

Results

We use 57 publicly available human gene expression data sets which are derived from the studies listed in table 1 of the Methods section. For each study, gene expression standard deviations (SDs) were calculated using a unified pipeline that normalized the mean-variance relation in count data, controlled for batch effects, and removed outliers (see Methods for details). Spearman correlations (ρ_s) between gene expression SDs reveal a broadly similar rank of gene expression variance, such that genes that are most variable in one study tend to be most variable in all studies (fig. 1 A and B). Several data sets were derived from two large research projects: GTEx [10] and TCGA [12], and we note these study origins in the figures. (We refer to data sets and studies interchangeably.) A principal coordinate analysis [13] using $|1 - \rho_s|$ as a distance measure does not show clearly delineated groups, but GTEx and TCGA studies are clustered among themselves and close together (fig. 1 C). This clustering indicates some effect of study source on the similarity between gene expression SD across studies, which we explore in detail below. The observed range of gene expression SD across genes is variable across studies but can be normalized so that the distributions are comparable (fig. 1 D). Given that the correlations across studies are mostly positive and high (75% of correlations are between 0.45 and 0.9), indicating similar ordering of the genes, we seek to summarize the differences in variance across genes by using a single cross-study rank, averaging the ordering across all studies. To create this rank, we used the score of each gene in the first principal component of the Spearman correlation matrix. These scores generate a ranked list of genes, with most variable genes having the highest rank. We create a similar across-study rank for mean expression. The red and blue ticks at the bottom of fig. 1 D show the positions on the SD distributions of the least and most variable genes in our variance rank. The position of these genes in the SD distributions illustrates how the extremes of the rank are indeed some of the least and most variable genes across all studies. We also create a set of tissue-specific SD ranks, which use the same procedure outlined above but using only studies which were performed on the same tissue. This creates a series of gene ranks, one for each sampled tissue, which describes

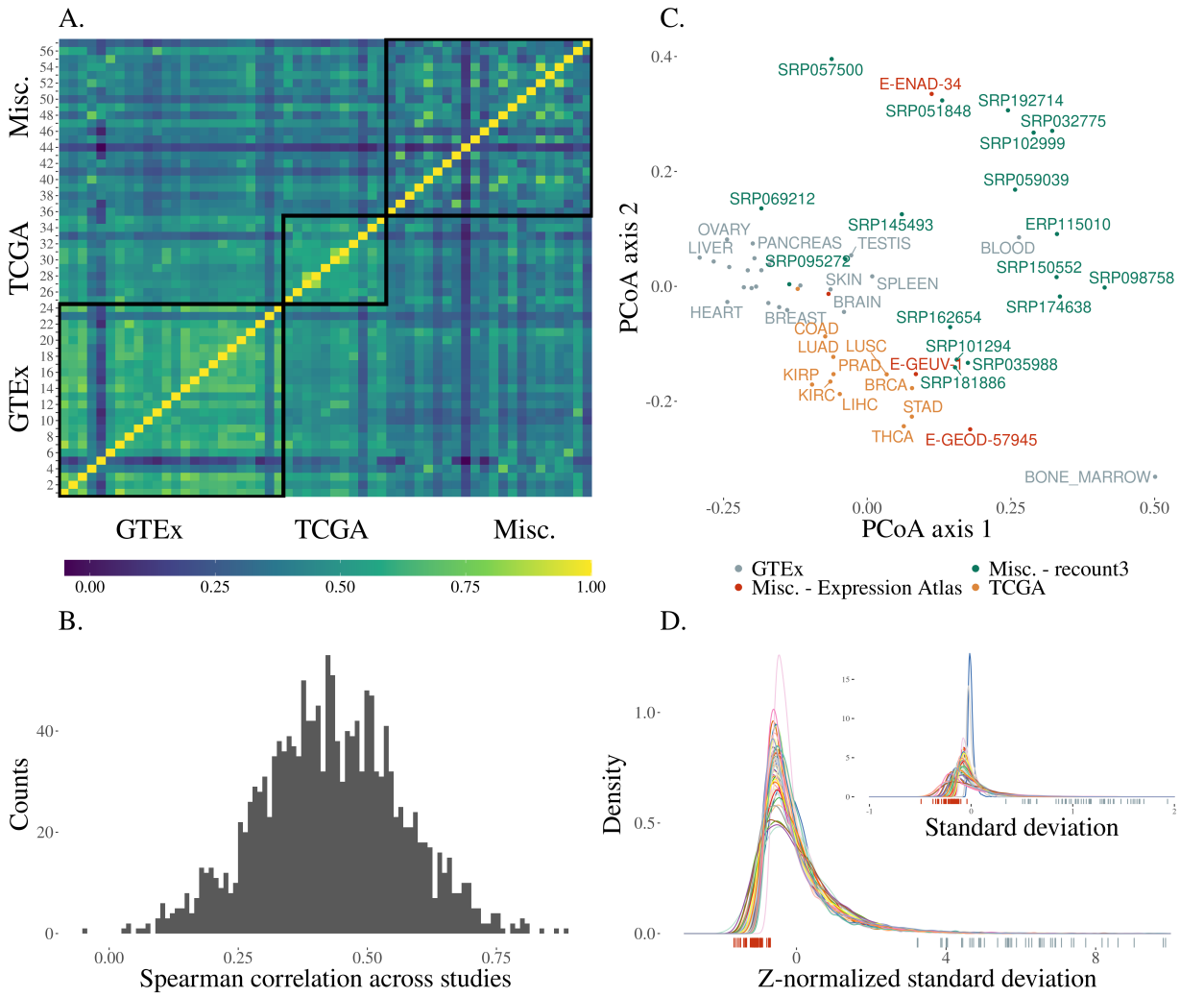

Figure 1: A. Correlation heatmap showing the across-study Spearman correlation of standard deviations. Pairs of studies with more similar patterns of gene expression variance have higher correlations. Studies are shown in the same order as in fig. 2, panel A; B. Histogram of the correlations shown in the previous panel; C. Standard deviation correlation PCoA. There is no clear structuring of the studies with respect to their source, which is indicated by the colors; D. Density plot of standard deviations after z-normalization. The inset plot shows the distribution of mean-centered standard deviations grouped by study without normalization. The corresponding rug plots show the location of the highest-ranking gene in standard deviation rank (right, blue) and lowest (left, red).

the gene expression SD rank in that particular tissue. Both tissue-specific and across-study ranks are available in the Supporting Information.

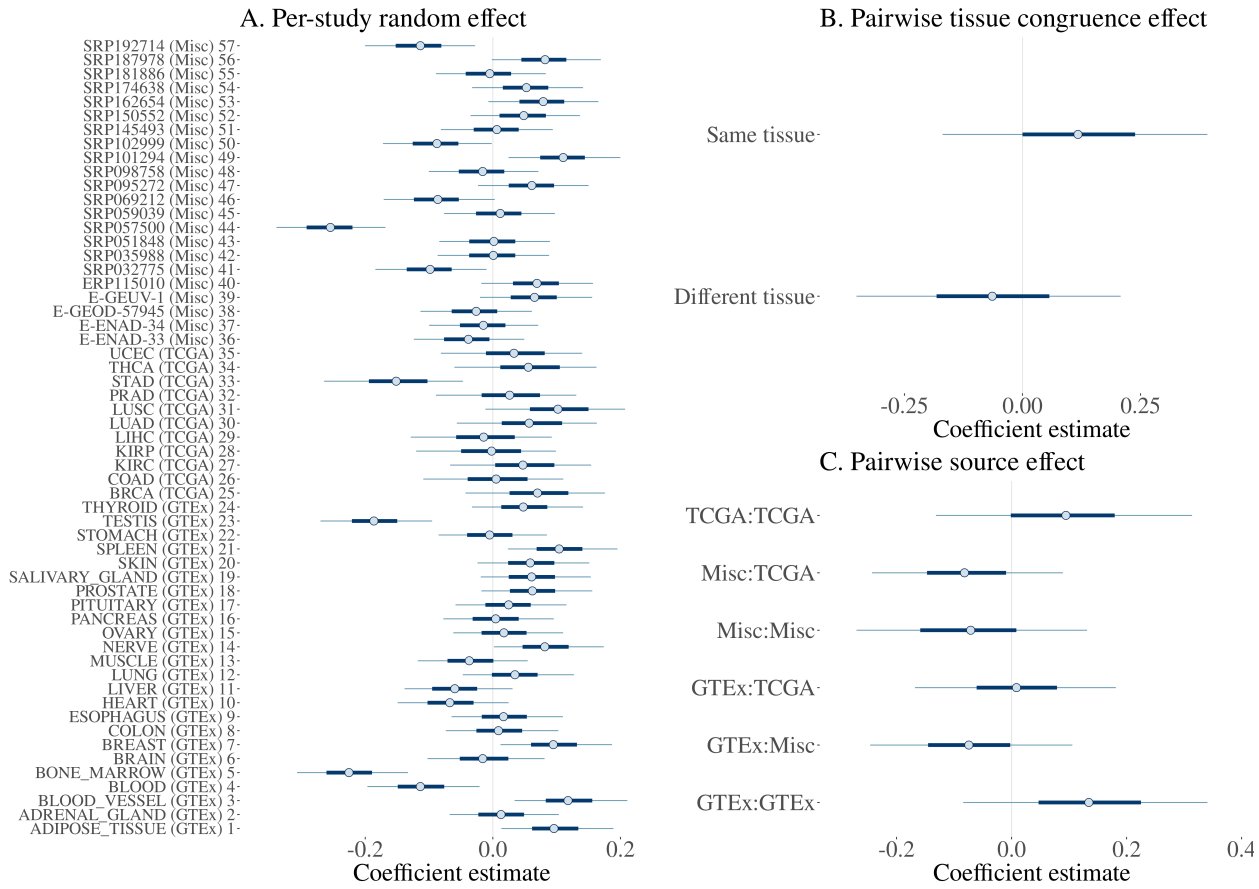

Figure 2: Coefficient estimates from a linear model using the among studies Spearman correlations as the response variable. These correlations are shown in fig. 1 A and B. In the linear model (see Methods for model equation), correlations are Fisher z-transformed. Study source and tissue are added as fixed effects. Coefficient estimates are shown with 50% and 95% credibility intervals. Panel A: The per-study random effect which accounts for the non-independence between the pairwise correlation values and estimates the characteristic contribution of each study to these correlations. For example, the lowest estimate among these parameters, which corresponds to the study BONE MARROW (from GTEx), indicates that correlations involving this study tend to be lower than the others. Panels B and C: Fixed effect estimates for the effects of tissue congruence and study-source effect. In (B) we see that correlations among studies that use the same tissue are slightly higher; and (C) correlations involving studies in the "Misc." category (non-GTEx and non-TCGA) tend to be lower, while comparisons involving GTEx and TCGA are higher.

What drives differences in gene expression variance?

To characterize the drivers of across-study similarity, we directly modeled the correlations across-study using a mixed-effect linear model [14,15]. In this model, we use individual study, pairwise tissue congruence (whether a comparison is within the same tissue or different tissue), and pairwise study source (GTEx, TCGA, and miscellaneous) as predictors of the correlations (see Methods). This modeling (fig. 2) shows that comparisons of studies within GTEx and TCGA have on average higher values of ρ_s, but also that comparing studies across GTEx and TCGA also shows a mild increase in the average correlation (fig. 2 C). Correlations involving studies that are not from TCGA or GTEx (marked as "Misc.") are on average lower (fig. 2 C). Since these two sources are independent, this mild effect on the similarities could be due to the quality of the data coming from these two large projects. Tissue also affects the similarity between gene expression SD, with studies using the same tissue being, on average, more similar (fig. 2 B). However, all of these pairwise effects are mild, and the

largest effects on the correlations are those associated with individual studies, in particular some specific tissues, i.e., comparisons involving BONE MARROW (from GTEx) and study SRP057500 (which used platelets) are on average lower (fig. 2 A). The only negative correlation we observe is between these two studies, which also appear further away in the PCoA plot in fig. 1 C.

Does biological function explain variance in expression?

To explore the relationship between variance and function, we took the top 5% most variable and the bottom 5% least variable genes in our ranking (560 genes in each group) and performed a Gene Ontology (GO) enrichment analysis within each group. This analysis allowed us to establish the representative functions of these consistently high and low-variance genes. In total, using a hypergeometric test and Benjamini-Hochberg (BH) adjusted p-value threshold of 10^{-3}, we found 59 enriched terms in the low variance genes, and 738 enriched terms in the high variance genes (see S2 Table for a complete listing). Among the 5% most variable genes we observe enrichment for biological processes like immune function, response to stimulus, maintenance of homeostasis, and tissue morphogenesis (fig. 3 A). In line with this GO term enrichment, the top 5% most variable genes are enriched 7.7-fold for genes that encode secreted proteins, relative to all other genes (hypergeometric test, $p < 10^{-3}$). Among the 5% least variable genes we see enrichment for housekeeping functions like mRNA processing, cell cycle regulation, methylation, histone modification, translation, transcription, and DNA repair (fig. 3 B); and accordingly, we find that previously characterized human housekeeping genes [16] are enriched within the 5% least variable genes 2.0-fold relative to all other genes (hypergeometric test, $p < 10^{-3}$). The genes exhibiting the lowest variance (lowest 5%) are also enriched for those that have been previously shown to have a high probability of being loss-of-function intolerant (pLI) [17] (1.2-fold enrichment, hypergeometric test, $p < 10^{-3}$). Genes with a high pLI have been shown to be important in housekeeping functions, and have higher mean expression values across a broad set of tissues and cell types [17]. Our result that genes with low variance are enriched for both housekeeping genes and genes with high pLI is consistent with this previous report; and we further see that the mean expression of genes positively correlates with pLI (Partial Spearman correlation $\rho_s = 0.32$, $p < 10^{-3}$), showing the opposite relationship between variance and mean expression when considering pLI.

We also explored the distribution of expression variance among the genes associated with GO terms. To do this, we gathered all biological process GO terms in level 3 (i.e. terms that are at a distance of 3 from the top of the GO hierarchy). Using only the set of genes that are associated with at least one of these level-3 terms, we separated the genes into expression variance deciles, with the first decile having the lowest variance. We then counted how many genes in each decile have been associated with each term. If variance rank is not linked to the GO annotations, terms should have an equal proportion of genes in each decile. We measured how far from this uniform allocation each term is by measuring the Shannon entropy of the proportion of genes in each decile. Higher entropy is associated with a more uniform distribution of genes across deciles. GO terms with low entropy indicate some decile is over-represented in the genes associated with that term. We also measured skewness for each term, which should be zero if no decile is over-represented, negative if high-variance terms are over-represented, and positive if low-variance deciles are over-represented. Skewness by entropy for each GO term can be seen in fig. 4 . Positive-skew low-entropy terms, those enriched with low-variance genes, are associated with housekeeping functions, like RNA localization, translation initiation, methylation and chromosome segregation (fig. 5 A). Likewise, terms with negative skew and low entropy, enriched for high-variance genes, are related to immune response, tissue morphogenesis, chemotaxis—all dynamic biological functions related to interacting with the environment (fig. 5 B).

Both GO analyses suggest a strong influence of biological function in determining gene expression variance. Genes associated with baseline fundamental functions, expected to be under strong stabilizing selection, are also low-variance; high-variance genes are associated with responding to external stimuli (i.e., tissue reorganization and immune response).

Sequence variation and gene expression connectivity

We use gene-level statistics capturing evolutionary and population variation to link processes that potentially influence variation in gene expression to the observed variance rank. We focus on three gene-level measures: nucleotide diversity (π), gene expression connectivity, and the proportion of substitutions that are adaptive (α). Nucleotide diversity is used as a proxy for cis-regulation sites, and we expect variation to increase with

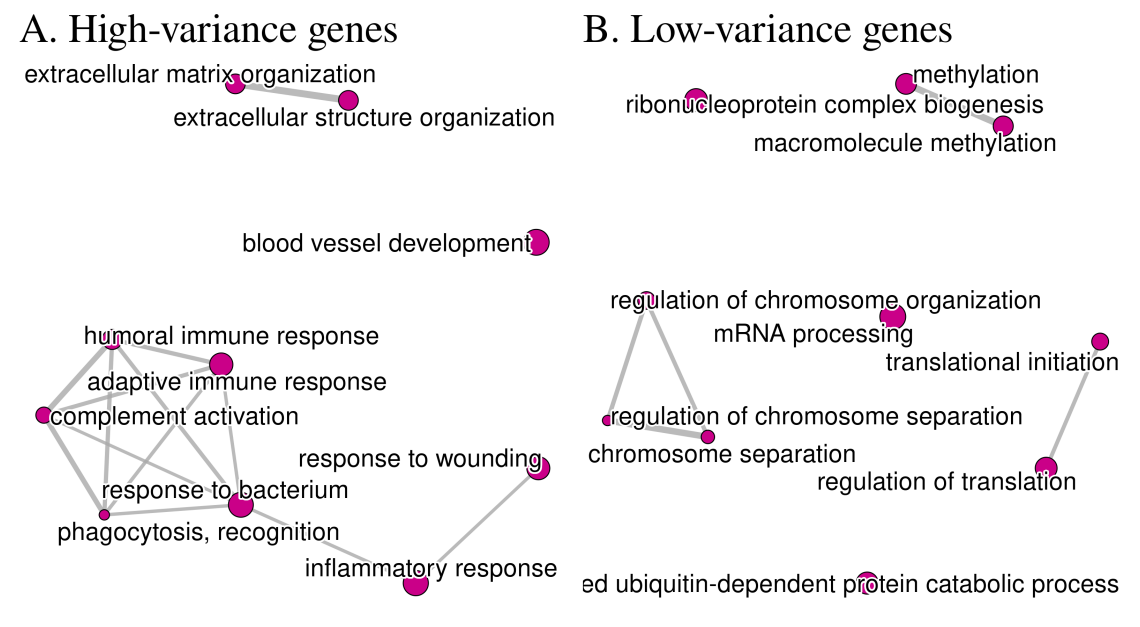

Figure 3: Gene set enrichment analyses testing for over-representation of gene ontology categories in the upper and lower 5% quantiles of the gene variance rank. (A) High-variance genes are enriched for terms related to immune function, response to wounding, blood vessel morphogenesis, and inflammatory response. In contrast, (B) low-variance genes are associated with translation, control of methylation, RNA processing, chromosome separation, and other cell housekeeping functions. All displayed terms are significant with a 5% FDR corrected p-value below 10^{-3}.

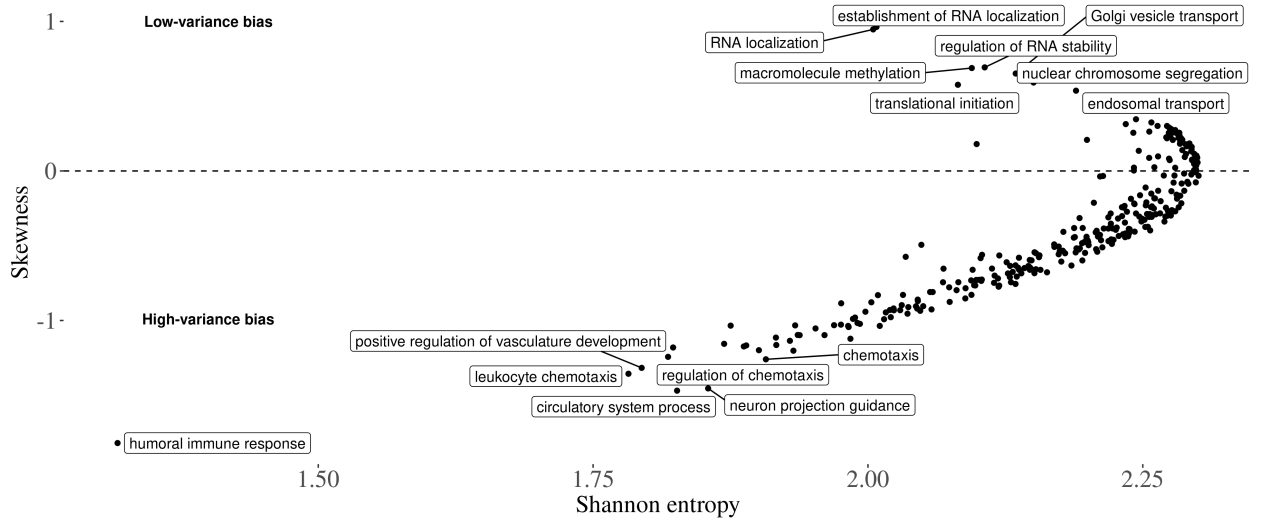

Figure 4: Relationship between skew and entropy of rank decile distributions for each GO term. High entropy terms, to the right of the plot, are associated with a more egalitarian proportion of genes in each of the SD rank deciles. Terms on the left of the plot are associated with more genes in some particular decile. The skewness in the y-axis measures if the high- or low-variance deciles are more represented for a particular term. Terms on the positive side of the y-axis are associated with low-variance genes, and terms on the negative side of the y-axis are associated with high-variance genes. The GO terms are filtered for gene counts greater than 100, as in fig. 5. Some of the top high- and low-skewness terms are labeled for illustration.

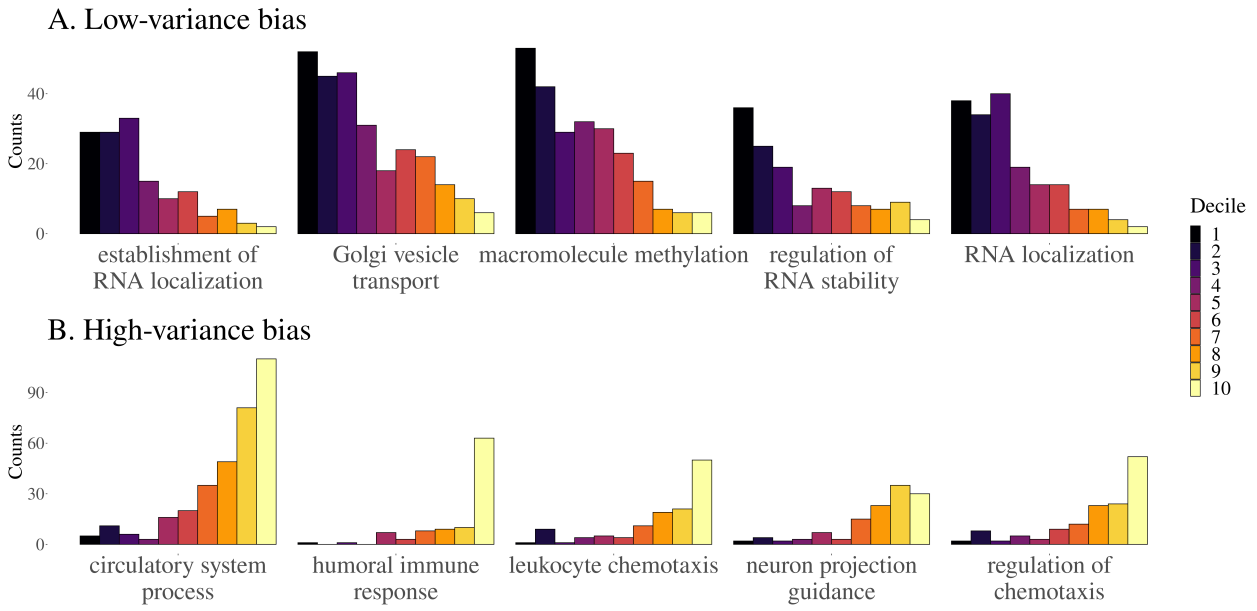

Figure 5: Distributions of decile ranks of level 3 GO terms. Each plot shows the count of genes in each decile of the rank. These GO terms are filtered for gene counts greater than 100 and sorted by the skewness of the distribution. The top panel shows the 5 most positively skewed terms and the bottom panel shows the 5 most negatively skewed terms.

diversity. Here, we use a partial Spearman correlation to account for any potential residual effect of mean expression and find a partial correlation between gene expression variance and π of 0.184 ($p < 10^{-3}$). Connectivity, a proxy for regulatory interactions with other genes and of selective constraints [18], in turn, should be negatively correlated with variation, as highly connected genes are expected to be more constrained in their variability. The resulting partial Spearman correlation is -0.024 ($p \approx 6 \times 10^{-3}$). Finally, we find a partial Spearman correlation of -0.046 ($p \approx 10^{-3}$) for the proportion of substitutions that are adaptive. Although all associations are significant and in the expected direction, their effect sizes are very small, suggesting a weak link between these broad measures and gene expression variance.

How do molecular signatures of gene regulation relate to gene expression variance?

We assess how local epigenetic features relate to gene expression variance. We use each gene, including the surrounding 10 kb on both ends, to calculate the proportion of gene regions that correspond to epigenetic signatures of gene regulation defined through ChromHMM [19] chromatin states. Chromatin states associated with distal (i.e., non-promoter) gene regulation are positively correlated with the across-study variance rank, regardless of whether the regulatory effect on gene expression is positive or negative (fig. 6; see also "across-study" correlations in supplementary fig. 1A). For example, both the proportion of gene regions made up of enhancers and repressed genomic states are positively correlated with gene expression variance (BH adjusted spearman correlation $p < 0.05$). Histone modifications associated with active promoters, as well as transcribed states, are inversely correlated with gene expression variance (supplementary fig. 1A), whereas they are positively correlated with the mean rank (supplementary fig. 1B). Taken together, these results are compatible with gene expression variance being more associated with distal (i.e., non-promoter) gene regulation, rather than the overall active transcriptional state of a gene region, as is the case with mean gene expression.

We also explore the relationship between tissue-specific ChromHMM chromatin states and SD rank and contrast these tissue-level analyses to the across-study analysis outlined above. Many of the across-study correlations are recapitulated at the tissue-specific level, including a strong and highly consistent positive correlation between the proportion of gene regions made up of enhancer states and that gene's expression variance, and an inverse relationship between gene expression variance and histone marks associated with gene transcription (supplementary fig. 1A). Two blood associations stand out as being different from the consistent effects across the other tissue-level and across-study associations. First, the weak (i.e., histone marks associated with

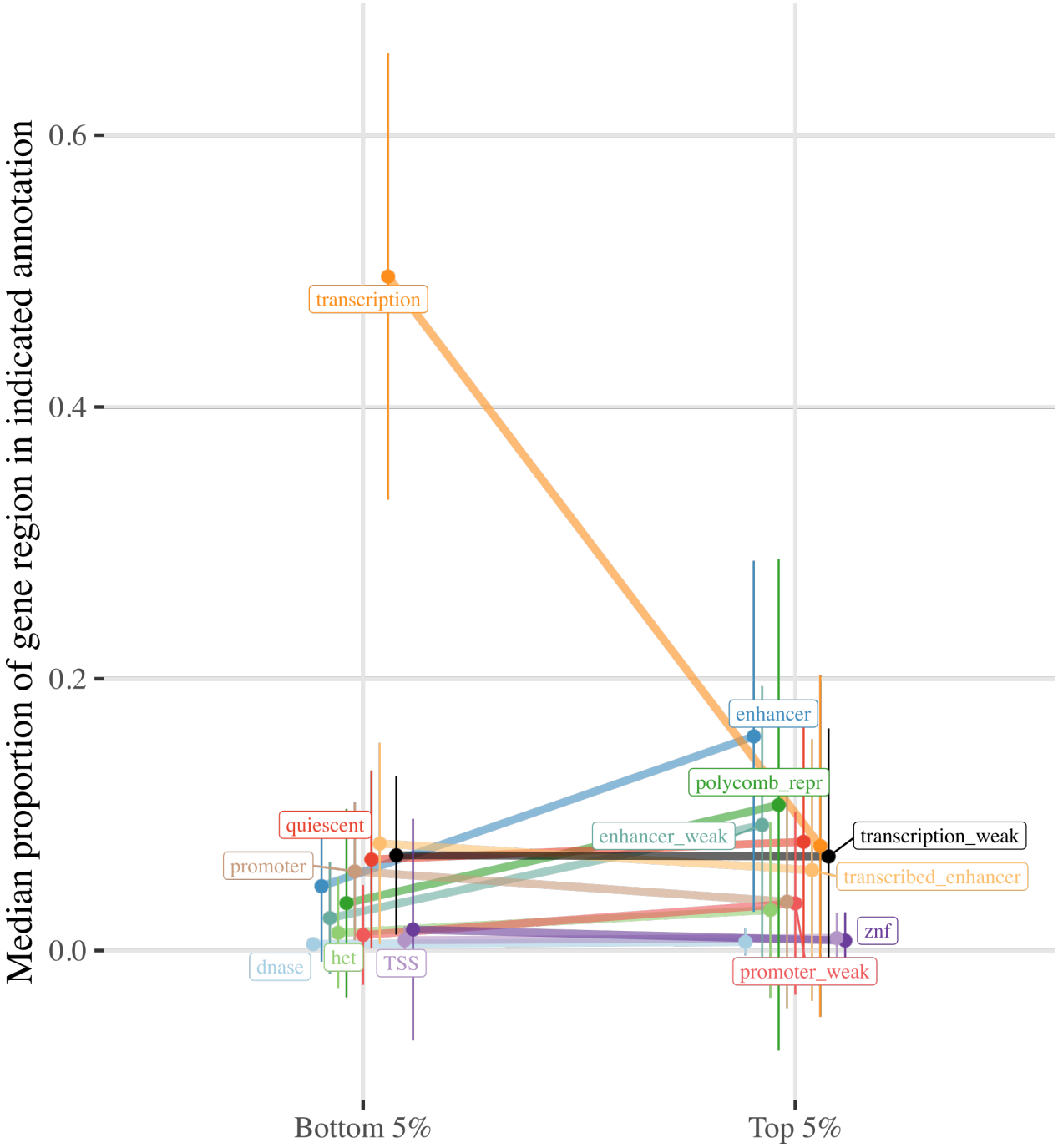

Figure 6: Proportion of gene regions made up of ChromHMM chromatin states for genes in the top and bottom 5% of the across-study variance rank metric. Line plot contrasts the proportion of gene regions made up of the indicated chromatin states for genes in the top and bottom 5% of the across-study variance rank metric. Ends denote the median proportion of gene regions made up of the chromatin state, and error bars represent the standard error of the mean. States colored black are not significant, all others exhibit significant differences in gene region made up of the chromatin state for genes in the top and bottom 5% of the variance rank metric (BH adjusted Wilcoxon signed-rank test, $p < 0.05$). Het indicates heterochromatin; TSS, transcription start sites; znf, zinc finger genes.

both activating and repressive functions) promoter state is positively correlated with gene expression variance in all comparisons except blood. Second, the consistent inverse correlation of gene expression variance with weak transcription is reversed in blood, such that there is a positive correlation between histone marks associated with weak transcription and blood gene expression variance (supplementary fig. 1A). Taken together, these results suggest that, rather than genes with a bivalent promoter state (i.e., poised genes) exhibiting more expression variance, blood high-variance genes are more likely already expressed at basal levels (i.e., weakly transcribed), as discussed previously [20].

Immediate early genes (IEGs) respond quickly to external signals without requiring *de novo* protein synthesis, and a bivalent state has been reported to be associated with IEG promoters [reviewed in 21]. Given our results that genes with high expression variance are enriched for cellular signaling and response mechanisms (fig. 3 A), and bivalent promoter states are correlated with the gene expression variance rank (supplementary fig. 1A), we hypothesized that IEGs would be enriched within genes in the top expression variance ranks. This was the case for all tissue-level gene expression variance ranks (enrichment ratios range from 3.3-8.8, Bonferroni-adjusted hypergeometric test, $p < 0.05$), except for in blood (enrichment ratio = 1.2, hypergeometric test, $p = 0.3$). Thus, once again blood stands out when attempting to understand genomic regulatory drivers of expression variance. In all, while high-variance genes are generally shared across tissues and enriched for immune and environmental signaling pathways, it seems that the gene regulatory mechanisms governing their expression are distinct between immune cell types and other tissues studied here.

Linking expression variance and disease

To explore the link between expression variance and disease, we use the gene annotations derived from a probabilistic transcriptome-wide association study (PTWAS) [22]. Using the list of significant gene-trait pairs at 5% FDR provided by Zhang et al. [22], we performed a hypergeometric enrichment test for the top 5% high- and low-variance genes in our across-study rank and in all tissue-specific gene variance ranks. Despite their overall high similarity, we use both across-study and tissue-specific ranks because some genes only appear in the tissue-specific rank due to their limited tissue-specific gene expression. In the high-variance group, we find no enrichment in the across-study rank, but we do find enrichment of genes annotated for allergy, immune disease, and endocrine system disease among the high-variance genes in several tissue-specific variance ranks. Among high-variance genes in the colon rank, we see enrichment for endocrine system disease (1.77-fold, hypergeometric test, $p < 10^{-4}$). Among high-variance genes in the immune cell tissue rank, we see enrichment for endocrine system disease (1.67-fold, hypergeometric test, $p < 10^{-3}$), allergy (1.7-fold, hypergeometric test, $p < 10^{-3}$), and immune disease (1.32-fold, hypergeometric test, $p < 10^{-2}$). Among high-variance genes in the thyroid rank, we see enrichment for endocrine system disease (1.9-fold, hypergeometric test, $p < 10^{-5}$), allergy (1.85-fold, hypergeometric test, $p < 10^{-4}$), and immune disease (1.45-fold, hypergeometric test, $p < 10^{-4}$). These are all quite similar and suggest a stable pattern of high-variance gene expression across these tissues, with enrichment for these three classes of diseases. The link with immune diseases is expected given the high enrichment for immune-related genes in the high-variance group [23]. As for the low-variance group, we found strong enrichment for genes associated with psychiatric and neurological disorders in the across-study rank and in some tissue-specific ranks (breast, liver, and stomach; ~1.2-fold enrichment, hypergeometric test, $p < 0.05$, for all cases). The psychiatric disease link is consistent with previous work [24] and is discussed below; however, the enrichment among the low-variance genes is weaker.

Discussion

By using large publicly available data sets, we were able to probe the landscape of gene expression variance in several human tissues. Differences in gene expression variance were driven by technical aspects of gene expression measurement, with data derived from large consortia showing more similar patterns of variance across genes; and by tissue, with studies using the same tissues also showing higher similarities. This would suggest that careful consideration of sample sizes and experimental design are fundamental to the study of gene expression variance, and the usual small samples of RNA-seq studies might be underpowered for the study of this particular aspect of gene expression. However, both the effects of study origin and tissue were small, and the largest drivers of differences across studies were idiosyncratic differences related to single data sets, with tissues known to have divergent gene expression patterns (i.e. bone marrow, blood, testis, and platelets) also showing the largest differences in gene expression variance. Understanding the consequences

of these differences in variance for specific tissues is still an open field. It is clear, however, that differences in variance are informative beyond the differences in mean expression. Even after we account for differences in mean expression, differences in gene expression variance carry information about tissue origin and function.

While these observed differences are notable, we also find a broadly similar pattern of gene expression variance across studies, with high correlations between gene expression variance across most studies, consistent with measurements of expression variance in single cells and in populations of cells for various tissues [11,25,26]. Leveraging this similarity between gene expression variance, we used a multivariate strategy to create a single rank of expression variance, which allowed us to order almost 13k genes according to their expression variance. This rank is associated with within-gene sequence variation, with more polymorphic genes being more variable. Furthermore, genes with high connectivity, those with higher levels of gene expression correlations with other genes, are less variable.

Functional analysis using GO enrichment indicated a clear link between function and gene expression variance. First, genes with high gene expression variance were enriched for biological functions related to reacting to environmental pressures, like immune function and tissue reconstruction. Likewise, low-variance genes were enriched for basic cell functions, like RNA processing, translation, DNA methylation, and cell duplication. These results are consistent with previous analysis of gene expression variance on a tissue-by-tissue basis [11]. This pattern of enrichment is also observed when we look at enrichment for high- or low-variance genes within the genes associated with each term in the GO hierarchy. Basic cell function terms are enriched for low variance genes, and terms involved in response to external stimulus are enriched for high variance genes.

While indirect, all these patterns point to a selective structuring of gene expression variance. Stabilizing and purifying selection are consistent: genes expected to be under strong stabilizing selection, those linked with fundamental baseline biological processes, are indeed overrepresented in the least variable genes. These same genes are also expected to be under strong purifying selection and to show low levels of polymorphisms, which we observe. Likewise, genes whose function is constrained by myriad interactions with several other genes, those with high connectivity, are less variable. Furthermore, genes involved with direct interaction with the environment, which must change their pattern of expression depending on external conditions, are expected to be more variable, and again we see a strong enrichment of genes related to interacting with the environment among the most variable. Given this strong functional linkage between function and variance, it is not surprising that the gene variance ranking is similar across studies, allowing us to create our across-study ranking in the first place.

One interesting aspect of the GO term analysis shown in fig. 4 and fig. 5 is that there is no biological process term associated with enrichment for intermediate variance genes: the low-entropy terms have either positive or negative skew, never zero skew. In other words, there is no annotated biological process for which the associated genes are kept at some intermediary level of variation. For the GO terms we used, either there is no relation between the gene expression variance and the biological process, or there is a strong bias toward high or low-variance genes. This suggests that selective shaping of gene expression has two modes, corresponding with (1) biological processes under strong stabilizing selection (i.e., variance-reducing selection) or (2) biological processes under disruptive selection (i.e., varianc-increasing selection). In short, we find strong support for the idea that there are genes with consistently more (or less) variable expression levels, and that these differences in variance are the result of different patterns of selection.

Following Alemu et al. [11], we observe that epigenetic signatures of gene regulation, such as enhancer histone marks, make up a higher proportion of the surrounding genomic regions of genes that exhibit higher variance in expression. In contrast, an accumulation of strong promoter elements and overall transcriptional activation is associated with genes with lower expression variance. These results suggest the presence of distinct modes of regulation for genes with high vs. low variance. Combined, the differences in the types of genomic regulatory features surrounding the high- and low-variance genes and their distinct functional annotations suggest different mechanisms of regulation of their gene expression variance [11]. This heterogeneity could lead to detectable differences in selection signatures between distal regulatory elements and promoters depending on the gene expression variance. This heterogeneity in regulation for high and low-variance genes is also notable due to the usual focus on gene expression robustness, in the sense of reducing variation [27–30]. For example, Siegal and Leu [27] provide several examples of known regulatory mechanisms for reducing gene expression variance, but no examples for the maintenance of high gene expression variance. We posit that it should be possible to go beyond the usual characterization of strategies of gene expression robustness, in the sense of

reducing variation, and to explore mechanisms for the *robustness of plasticity*, that is, the maintenance of high levels of gene expression variation given environmental cues.

Given the broad consistency of gene expression variance in healthy tissues, a natural question is how do these well-regulated levels of variation behave in perturbed or disease conditions. We find some suggestive links between tissue-specific variance ranks and disease, but these links need to be better explored using more specific methods. Comparing two HapMap populations, Li et al. [25] showed that gene expression variance was similar in both populations and that high variance genes were enriched for genes related to HIV susceptibility, consistent with our observation of enrichment for immune-related genes among those with more variable expression. In a case-control experiment, Mar et al. [24] showed that expression variance was related to disease status in Schizophrenia and Parkinson's disease patients, with altered genes being non-randomly distributed across signaling networks. These authors also find a link between gene network connectivity and expression variance, in agreement with the effect we find using the gene expression variance rank. Also, the pattern of variance alteration differed across diseases, with Parkinson's patients showing increased expression variance, and Schizophrenia patients showing more constrained patterns of expression. The authors hypothesize that the reduced variance in Schizophrenia patients reduces the robustness of their gene expression networks, what we refer to as a loss of plasticity. This suggests several types of shifts in gene expression variation are possible, with different outcomes. We highlight three distinct possibilities: First, low variance genes, under strong stabilizing selection, could become more variable under stress, indicating a reduced capacity for maintaining homeostasis. Second, high-variance genes, expected to be reactive to changes in the environment, could become less variable, indicating a reduced capacity to respond to external stimuli. Third, the covariance between different genes could be altered, leading to decoherence between interdependent genes [4]. Any one of these changes in expression variance patterns could have physiological consequences, and exploring these differences should be a major part of linking gene expression to cell phenotypes and function (see Hagai et al. [23] for example). Genes are also expected to differ in their capacity to maintain an optimal level of gene expression variance [29]. Variation in robustness is linked to gene regulatory networks and epigenetic gene expression regulation [28,31], and therefore, should differ across high- and low-variance genes. Our results suggest that low- and high-variance genes could use different strategies in order to maintain their optimal levels of variation and that this variability in strategies is the result of different patterns of selection.

Methods

Data sources

We selected 57 human RNA-seq studies with large sample sizes from the public gene expression repositories recount3 [32] and Expression Atlas [33]. Because we are interested in population-level variation of gene expression, we exclude single-cell studies and focus only on studies derived from tissue samples. We only used studies for which raw read count data was available, and for which we could parse the metadata for batch effects. We use studies to refer to independent data sets, which could have been generated by the same consortium. For example, the GTEx data are separated by tissue, and we refer to each tissue as a separate study. We divide our studies into three categories depending on their origin: GTEx, TCGA, and Miscellaneous.

Table 1: Data Source Table

Study ID	Citation
ADIPOSE_TISSUE (Fat), ADRENAL_GLAND (Adrenal), BLOOD (Blood), BLOOD_VESSEL (Blood_vessel), BONE_MARROW (Marrow), BRAIN (Neuron), HEART (Heart), BREAST (Breast), SALIVARY_GLAND (Salivary), COLON (Colon), LIVER (Liver), NERVE (Neuron), LUNG (Lung), PANCREAS (Pancreas), MUSCLE (Muscle), THYROID (Thyroid), OVARY (Ovary), STOMACH (Stomach), ESOPHAGUS (Esophagus), SPLEEN (Spleen), PROSTATE (Prostate), SKIN (Skin), PITUITARY (Pituitary), TESTIS (Testis)	The GTEx Consortium, 2020 - [34]
LUSC (Lung), STAD (Stomach), COAD (Colon), LUAD (Lung), BRCA (Breast), KIRC (Kidney), KIRP (Kidney), LIHC (Liver), THCA (Thyroid), PRAD (Prostate), UCEC (Uterus)	The Cancer Genome Atlas Research Network et al., 2013 - [12]
SRP150552 (Blood)	Altman et al., 2019 - [35]
SRP101294 (Fat)	Armenise et al., 2017 - [36]
SRP057500 (Platelets)	Best et al., 2015 - [37]
SRP051848 (Immune)	Breen et al., 2015 - [38]
SRP187978 (Liver)	Çalışkan et al., 2019 - [39]
E-ENAD-34 (Immune)	Chen et al., 2016 - [40]

Study ID	Citation
SRP059039 (Blood)	DeBerg et al., 2018 - [41]
SRP174638 (Immune)	Dufort et al., 2019 - [42]
E-GEOD-57945 (Colon)	Haberman et al., 2014 - [43]
SRP162654 (Blood)	Harrison et al., 2019 - [44]
SRP095272 (Blood)	Jadhav et al., 2019 - [45]
SRP102999 (Blood)	Kuan et al., 2017 - [46]
SRP145493 (Immune)	Kuan et al., 2019 - [47]
E-GEUV-1 (Immune)	Lappalainen et al., 2013 - [48]
SRP035988 (Skin)	Li et al., 2014 - [49]
SRP192714 (Blood)	Michlmayr et al., 2020 - [50]
ERP115010 (Blood)	Roe et al., 2020 - [51]
E-ENAD-33 (Neuron)	Schwartzentruber et al., 2018 - [52]
SRP181886 (Neuron)	Srinivasan et al., 2020 - [53]
SRP098758 (Blood)	Suliman et al., 2018 - [54]
SRP032775 (Blood)	Tran et al., 2016 - [55]
SRP069212 (Liver)	Yang et al., 2017 - [56]

Processing pipeline: We use a standardized pipeline to measure gene expression variance while removing extraneous sources of variation. Data from case-control studies were filtered to keep only control samples. Technical replicates were summed. For each study, we filtered genes that did not achieve a minimum of 1 count per million (cpm) reads in all samples and a mean of 5 cpm reads. To account for the mean-variance relation in count data, the remaining genes were subjected to the variance stabilizing transformation implemented in DESeq2 [57]. Fixed effects were manually curated from the metadata for all studies and removed using a linear fixed-effects model. Outlier individuals in the residual distribution were removed using a robust Principal Component Analysis (PCA) approach of automatic outlier detection [58]. Gene expression standard deviation is measured as the residual standard deviation after fixed effect correction and outlier removal. The full annotated pipeline is available at the github repository ayroles-lab/ExpressionVariance.

Gene expression variance across-study correlation

We assessed the similarity in gene expression variance across studies by using a between-study Spearman correlation matrix of the measured SDs. Only genes present in all studies were used to calculate the Spearman correlation matrix, ~4200 genes in total. Using Spearman correlations avoids problems related to overall scaling or coverage differences, and allows us to assess if the same genes are usually more or less variable across studies. To investigate the factors involved in determining correlations between studies, we used a Bayesian varying effects model to investigate the effect of study origin and tissue on the correlations across studies. This model is designed to take the non-independent nature of a set of correlations into account when modeling the correlation between gene expression SDs. This is accomplished by adding a per-study random effect, see [15] for details. The Fisher z-transformed Spearman correlations across studies ($z(\rho_{ij})$) are modeled as:

$$
\begin{aligned}
z(\rho_{ij}) &\sim N(\mu_{ij}, \sigma) \\
\mu_{ij} &= \mu_0 + \alpha_i + \alpha_j + \beta X \\
\alpha_i &\sim N(0, \sigma_\alpha)
\end{aligned}
$$

The α terms account for the non-independence between the pairs of correlations and estimate the idiosyncratic contribution of each study to all the correlations it is involved in. The fixed effects encoded in the design matrix X measure the effects of tissue congruence and study-origin congruence. All fixed effect parameters (β) and per-study parameters (α) receive weakly informative normal priors with a standard deviation of one quarter. For the overall variance (σ) we use a unit exponential prior, and for the intercept (μ_0) a unit normal prior. This model was fit in Stan [59] via the rethinking R package [60], using eight chains, with 4000 warm-up iterations and 2000 sampling iterations. Convergence was assessed using R-hat diagnostics [61], and we observed no warnings or divergent transitions.

Gene expression SD rank: Given that most of the variation in the Spearman correlation across studies is explained by a single principal component, we use the ranked projections of gene expression SDs in this principal component (PC1) to create an across-study rank of gene variation. The higher the rank, the higher the expression SD of a given gene. Genes that were expressed in at least 50% of the studies were included in the rank. In order to project a particular gene onto the PC1 of the between-study correlation matrix, we impute missing values using a PCA-based imputation [62]. The imputation procedure has minimal effect on the ranking, and imputing missing SD ranks at the beginning or at the end of the ranks produces similar results. We also create a tissue-specific variance ranking, using the same ranking procedure but joining studies done in the same tissue type. For this tissue-level ranking, we only use genes that are expressed in all studies of a given tissue. For tissues that are represented by a single study, we use the SD ranking for that study as the tissue rank. We further investigate the tissue-level expression variance ranks as they relate to genomic regulation.

Gene expression mean rank: We also use the same strategy to create a mean gene expression rank, repeating the process but using mean expression instead of standard deviation. All ranks are available in the supporting information.

Gene level statistics

Genetic variation: Genetic variation measures were obtained from the PopHuman project, which provides a comprehensive set of genomic information for human populations derived from the 1000 Genomes Project. Gene-level metrics were used when available. If only window-based metrics are available, we assembled gene-level information from 10 kb window tracks where each window that overlaps

with a given gene was assigned to the gene and the mean metric value is reported. In parallel, we use the PopHumanScan data set, which expands PopHuman by compiling and annotating regions under selection. Similarly, we used gene-level information when possible, and for tracks with only window-based metrics, gene-level information was assembled from the 10 kb windows using the same assignment method described above. Nucleotide diversity (π), the average pairwise number of differences per site among the chromosomes in a population [63], provides insight into the genetic diversity within a population, in this case, the CEU population within 1000 genomes. The nucleotide diversity can also be used as an estimator of the central population genetic parameter, normally given as θ.

Gene connectivity: We calculated the average weighted connectivity for all genes by creating a fully connected gene-by-gene graph in which each edge is weighted by the Spearman correlation between gene expression levels. We then trimmed this graph by keeping only edges for which the Spearman correlation is significant at a BH false discovery rate of 1%. In this trimmed network, we then took the average of the Spearman correlation of all remaining edges for each gene. So, for each study, we have a measure of the average correlation of each gene with every other gene. The average connectivity for each gene is the average across all studies in which that gene is expressed.

Cross-tissue vs. tissue-level chromatin states: We use the universal [64] and tissue-specific [65] ChromHMM [19] chromatin states to compare the non-overlapping genome segmentation to cross-tissue and tissue-level gene expression variance metrics. We use the proportion of the gene regions (gene +/- 10 kb) made up of each of the chromHMM chromatin states.

Correlations: We use the ppcor R package v1.1 [66] to run the pairwise partial Spearman correlations between gene-level statistics and the gene expression variance rank while controlling for the mean expression rank. P-values are corrected using the Benjamini-Hochberg procedure and comparisons with an adjusted p<0.05 are considered significant.

Gene function assessment

GO term enrichment: All gene ontology (GO) analyses were done using the clusterProfiler R package v4.2.2 [67] and the Org.Hs.eg.db database package v3.14.0 [68]. GO and all further enrichment analysis used the hypergeometric test to assess the significance of the enrichment.

Secreted genes: We use The Protein Atlas [69] to extract information on which proteins are secreted [70] and test for enrichment of genes with secreted products in the genes within the highest and lowest 5% of gene expression variance rank.

Housekeeping genes: Human housekeeping genes were identified as genes that are expressed with low variance in all 52 human cell and tissue types, assessed in over 10,000 samples [16]. We test for enrichment of housekeeping genes in the genes within the highest and lowest 5% of gene expression variance rank.

Immediate early genes (IEGs): Human IEGs were curated from the literature in [71] as genes that respond to experimental stimulation through up-regulation within the first 60 minutes of the experiment. We use the hypergeometric test to assess the significance of the enrichment. Immediate early genes (IEGs): Human IEGs were curated from the literature in [71] as genes that respond to experimental stimulation through up-regulation within the first 60 minutes of the experiment.

Probability of being loss-of-function intolerant (pLI): Genes that are likely haploinsufficient (i.e., intolerant of heterozygous loss-of-function variants) were detected as those with fewer than expected protein-truncating variants (PTVs) in ExAC [72]. We use genes with a pLI > 0.9 to test for the enrichment of loss-of-function intolerant genes in the genes exhibiting the highest and lowest 5% gene expression variance estimates.

Disease annotations: We use the gene annotations for involvement with diseases provided by the supporting information Table S2 from Zhang et al. [22].

Code availability

All code for reproducing all analysis and figures, along with a walk-through, is available at github.com/ayroles-lab/ExpressionVariance.

Supporting information

Supporting information is available at github.com/diogro/expVarManuscript.

1. SI figure 1 - Across-study and tissue-specific gene expression variance and mean correlations with non-overlapping chromatin states through ChromHMM.
2. SI figure 2 - Proportion of gene regions made up of ChromHMM chromatin states for genes in the top and bottom 5% of the across-study mean rank metric.
3. SI table 1 - Variance and mean rank metrics and the corresponding ChromHMM annotations used.
4. SI data - Study metadata - Metadata file describing the data used in the study as well as some intermediate processing information.
5. SI data - Gene ranks - Gene expression mean and variance ranks, across-study and tissue-specific.
6. SI data - GO enrichment - Combined table describing gene ontology enrichment in the top 5% and bottom 5% of genes as ranked by variance.

References

1. Jong TV de, Moshkin YM, Guryev V. Gene expression variability: The other dimension in transcriptome analysis. Physiol Genomics. 2019 May;51(5):145–58.
2. Fraser HB, Hirsh AE, Giaever G, Kumm J, Eisen MB. Noise minimization in eukaryotic gene expression. PLoS Biol. 2004 Jun;2(6):e137.
3. Wang Z, Zhang J. Impact of gene expression noise on organismal fitness and the efficacy of natural selection. Proc Natl Acad Sci U S A. 2011 Apr;108(16):E67–76.
4. Lea A, Subramaniam M, Ko A, Lehtimäki T, Raitoharju E, Kähönen M, et al. Genetic and environmental perturbations lead to regulatory decoherence. Elife. 2019 Mar;8.
5. Bruijning M, Metcalf CJE, Jongejans E, Ayroles JF. The evolution of variance control. Trends Ecol Evol. 2020 Jan;35(1):22–33.
6. Houle D. How should we explain variation in the genetic variance of traits? Genetica. 1998;102-103(1-6):241–53.
7. Hansen TF. Epigenetics: Adaptation or contingency. In: Benedikt Hallgrímsson BKH, editor. Epigenetics: Linking genotype and phenotype in development and evolution. University of California press Berkeley, CA; 2011. p. 357–76.
8. Wagner GP, Booth G, Bagheri-Chaichian H. A POPULATION GENETIC THEORY OF CANALIZATION. Evolution. 1997 Apr;51(2):329–47.
9. Pavlicev M, Hansen TF. Genotype-Phenotype Maps Maximizing Evolvability: Modularity Revisited. Evol Biol. 2011 Dec;38(4):371–89.
10. GTEx Consortium. Genetic effects on gene expression across human tissues. Nature. 2017 Oct;550(7675):204–13.
11. Alemu EY, Carl JW Jr, Corrada Bravo H, Hannenhalli S. Determinants of expression variability. Nucleic Acids Res. 2014 Apr;42(6):3503–14.
12. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The cancer genome atlas Pan-Cancer analysis project. Nat Genet. 2013 Oct;45(10):1113–20.
13. Gower JC. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika. 1966 Dec;53(3-4):325–38.
14. Dias FS, Betancourt M, Rodríguez-González PM, Borda-de-Água L. Analysing the distance decay of community similarity in river networks using bayesian methods. Sci Rep. 2021 Nov;11(1):21660.
15. Dias FS, Betancourt M, Rodríguez-González PM, Borda-de-Água L. BetaBayes-A bayesian approach for comparing ecological communities. 2021;
16. Hounkpe BW, Chenou F, Lima F de, De Paula EV. HRT atlas v1.0 database: Redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. Nucleic Acids Res. 2020 Jul;49(D1):D947–55.
17. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285–91.
18. Mähler N, Wang J, Terebieniec BK, Ingvarsson PK, Street NR, Hvidsten TR. Gene co-expression network connectivity is an important determinant of selective constraint. PLoS Genet. 2017 Apr;13(4):e1006402.
19. Ernst J, Kellis M. ChromHMM: Automating chromatin-state discovery and characterization. Nature methods. 2012;9(3):215–6.
20. Rogatsky I, Adelman K. Preparing the first responders: Building the inflammatory transcriptome from the ground up. Mol Cell. 2014 Apr;54(2):245–54.
21. Bahrami S, Drabløs F. Gene regulation in the immediate-early response process. Vol. 62, Advances in Biological Regulation. 2016. p. 37–49.
22. Zhang Y, Quick C, Yu K, Barbeira A, GTEx Consortium, Luca F, et al. PTWAS: Investigating tissue-relevant causal molecular mechanisms of complex traits using probabilistic TWAS analysis. Genome Biol. 2020 Sep;21(1):232.
23. Hagai T, Chen X, Miragaia RJ, Rostom R, Gomes T, Kunowska N, et al. Gene expression variability across cells and species shapes innate immunity. Nature. 2018 Nov;563(7730):197–202.
24. Mar JC, Matigian NA, Mackay-Sim A, Mellick GD, Sue CM, Silburn PA, et al. Variance of gene expression identifies altered network constraints in neurological disease. PLoS Genet. 2011 Aug;7(8):e1002207.
25. Li J, Liu Y, Kim T, Min R, Zhang Z. Gene expression variability within and between human populations and implications toward disease susceptibility. PLoS Comput Biol. 2010 Aug;6(8).
26. Dong D, Shao X, Deng N, Zhang Z. Gene expression variations are predictive for stochastic noise. Nucleic Acids Res. 2011 Jan;39(2):403–13.
27. Siegal ML, Leu JY. On the nature and evolutionary impact of phenotypic robustness mechanisms. Annu Rev Ecol Evol Syst. 2014 Nov;45:496–517.
28. Payne JL, Wagner A. Mechanisms of mutational robustness in transcriptional regulation. Front Genet. 2015 Oct;6(October):1–10.
29. Macneil LT, Walhout AJM. Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. Genome Res. 2011 May;21(5):645–57.
30. Denby CM, Im JH, Yu RC, Pesce CG, Brem RB. Negative feedback confers mutational robustness in yeast transcription factor regulation. Proc Natl Acad Sci U S A. 2012 Mar;109(10):3874–8.
31. Chalancon G, Ravarani CNJ, Balaji S, Martinez-Arias A, Aravind L, Jothi R, et al. Interplay between gene expression noise and regulatory network architecture. Trends Genet. 2012 May;28(5):221–32.

32. Wilks C, Zheng SC, Chen FY, Charles R, Solomon B, Ling JP, et al. recount3: Summaries and queries for large-scale RNA-seq expression and splicing. bioRxiv. 2021. p. 2021.05.21.445138.

33. Papatheodorou I, Moreno P, Manning J, Fuentes AMP, George N, Fexova S, et al. Expression atlas update: From tissues to single cells. Nucleic Acids Res. 2020 Jan;48(D1):D77–83.

34. GTEx Consortium. The GTEx consortium atlas of genetic regulatory effects across human tissues. Science. 2020 Sep;369(6509):1318–30.

35. Altman MC, Gill MA, Whalen E, Babineau DC, Shao B, Liu AH, et al. Transcriptome networks identify mechanisms of viral and nonviral asthma exacerbations in children. Nat Immunol. 2019 May;20(5):637–51.

36. Armenise C, Lefebvre G, Carayol J, Bonnel S, Bolton J, Di Cara A, et al. Transcriptome profiling from adipose tissue during a low-calorie diet reveals predictors of weight and glycemic outcomes in obese, nondiabetic subjects. Am J Clin Nutr. 2017 Sep;106(3):736–46.

37. Best MG, Sol N, Kooi I, Tannous J, Westerman BA, Rustenburg F, et al. RNA-Seq of Tumor-Educated platelets enables Blood-Based Pan-Cancer, multiclass, and molecular pathway cancer diagnostics. Cancer Cell. 2015 Nov;28(5):666–76.

38. Breen MS, Maihofer AX, Glatt SJ, Tylee DS, Chandler SD, Tsuang MT, et al. Gene networks specific for innate immunity define post-traumatic stress disorder. Mol Psychiatry. 2015 Dec;20(12):1538–45.

39. Çalışkan M, Manduchi E, Rao HS, Segert JA, Beltrame MH, Trizzino M, et al. Genetic and epigenetic fine mapping of complex trait associated loci in the human liver. Am J Hum Genet. 2019 Jul;105(1):89–107.

40. Chen L, Ge B, Casale FP, Vasquez L, Kwan T, Garrido-Martín D, et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. Cell. 2016 Nov;167(5):1398–1414.e24.

41. DeBerg HA, Zaidi MB, Altman MC, Khaenam P, Gersuk VH, Campos FD, et al. Shared and organism-specific host responses to childhood diarrheal diseases revealed by whole blood transcript profiling. PLoS One. 2018 Jan;13(1):e0192082.

42. Dufort MJ, Greenbaum CJ, Speake C, Linsley PS. Cell type-specific immune phenotypes predict loss of insulin secretion in new-onset type 1 diabetes. JCI Insight. 2019 Feb;4(4).

43. Haberman Y, Tickle TL, Dexheimer PJ, Kim MO, Tang D, Karns R, et al. Pediatric crohn disease patients exhibit specific ileal transcriptome and microbiome signature. J Clin Invest. 2014 Aug;124(8):3617–33.

44. Harrison GF, Sanz J, Boulais J, Mina MJ, Grenier JC, Leng Y, et al. Natural selection contributed to immunological differences between hunter-gatherers and agriculturalists. Nat Ecol Evol. 2019 Aug;3(8):1253–64.

45. Jadhav B, Monajemi R, Gagalova KK, Ho D, Draisma HHM, Wiel MA van de, et al. RNA-Seq in 296 phased trios provides a high-resolution map of genomic imprinting. BMC Biol. 2019 Jun;17(1):50.

46. Kuan PF, Waszczuk MA, Kotov R, Clouston S, Yang X, Singh PK, et al. Gene expression associated with PTSD in world trade center responders: An RNA sequencing study. Transl Psychiatry. 2017 Dec;7(12):1297.

47. Kuan PF, Yang X, Clouston S, Ren X, Kotov R, Waszczuk M, et al. Cell type-specific gene expression patterns associated with posttraumatic stress disorder in world trade center responders. Transl Psychiatry. 2019 Jan;9(1):1.

48. Lappalainen T, Sammeth M, Friedländer MR, Hoen PAC 't, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013 Sep;501(7468):506–11.

49. Li B, Tsoi LC, Swindell WR, Gudjonsson JE, Tejasvi T, Johnston A, et al. Transcriptome analysis of psoriasis in a large case-control sample: RNA-seq provides insights into disease mechanisms. J Invest Dermatol. 2014 Jul;134(7):1828–38.

50. Michlmayr D, Kim EY, Rahman AH, Raghunathan R, Kim-Schulze S, Che Y, et al. Comprehensive immunoprofiling of pediatric zika reveals key role for monocytes in the acute phase and no effect of prior dengue virus infection. Cell Rep. 2020 Apr;31(4):107569.

51. Roe J, Venturini C, Gupta RK, Gurry C, Chain BM, Sun Y, et al. Blood transcriptomic stratification of short-term risk in contacts of tuberculosis. Clin Infect Dis. 2020 Feb;70(5):731–7.

52. Schwartzentruber J, Foskolou S, Kilpinen H, Rodrigues J, Alasoo K, Knights AJ, et al. Molecular and functional variation in iPSC-derived sensory neurons. Nat Genet. 2018 Jan;50(1):54–61.

53. Srinivasan K, Friedman BA, Etxeberria A, Huntley MA, Brug MP van der, Foreman O, et al. Alzheimer's patient microglia exhibit enhanced aging and unique transcriptional activation. Cell Rep. 2020 Jun;31(13).

54. Suliman S, Thompson EG, Sutherland J, Weiner J 3rd, Ota MOC, Shankar S, et al. Four-Gene Pan-African blood signature predicts progression to tuberculosis. Am J Respir Crit Care Med. 2018 May;197(9):1198–208.

55. Tran TM, Jones MB, Ongoiba A, Bijker EM, Schats R, Venepally P, et al. Transcriptomic evidence for modulation of host inflammatory responses during febrile plasmodium falciparum malaria. Sci Rep. 2016 Aug;6:31291.

56. Yang Y, Chen L, Gu J, Zhang H, Yuan J, Lian Q, et al. Recurrently deregulated lncRNAs in hepatocellular carcinoma. Nat Commun. 2017 Feb;8:14421.

57. Love M, Anders S, Huber W. Differential analysis of count data–the DESeq2 package. Genome Biol. 2014;15(550):10–1186.

58. Chen X, Zhang B, Wang T, Bonni A, Zhao G. Robust principal component analysis for accurate outlier sample detection in RNA-Seq data. BMC Bioinformatics. 2020 Jun;21(1):269.

59. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: A probabilistic programming language. Journal of statistical software. 2017;76(1).

60. McElreath R. Statistical rethinking: A bayesian course with examples in r and stan. Chapman; Hall/CRC; 2020.

61. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis, third edition. CRC Press; 2013.

62. Husson F, Josse J, Narasimhan B, Robin G. Imputation of mixed data with multilevel singular value decomposition. J Comput Graph Stat. 2019 Jul;28(3):552–66.

63. Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci U S A. 1979 Oct;76(10):5269–73.

64. Vu H, Ernst J. Universal annotation of the human genome through integration of over a thousand epigenomic datasets. Genome biology. 2022;23(1):1–37.

65. Ernst J, Kellis M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nat Biotechnol. 2015 Apr;33(4):364–76.

66. Kim S. Ppcor: An r package for a fast calculation to semi-partial correlation coefficients. Communications for statistical applications and methods. 2015;22(6):665.

67. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation (N Y). 2021 Aug;2(3):100141.

68. Carlson M. Org.hs.eg.db: Genome wide annotation for human. 2021.

69. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science. 2015;347(6220):1260419.

70. Uhlén M, Karlsson MJ, Hober A, Svensson AS, Scheffel J, Kotol D, et al. The human secretome. Science signaling. 2019;12(609):eaaz0274.

71. Arner E, Daub CO, Vitting-Seerup K, Andersson R, Lilje B, Drabløs F, et al. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. Science. 2015 Feb;347(6225):1010–4.

72. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 Aug;536(7616):285–91.