Characterizing the landscape of gene expression variance in humans

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

[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

† These authors contributed equally to this work.
* Correspondence: Julien Ayroles <jayroles@princeton.edu>

Intro

Molecular phenotypes such as gene expression are a powerful tool 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.(Jong, Moshkin, and Guryev 2019) However, gene expression variance can be directly involved in determining fitness,(Fraser et al. 2004; Wang and Zhang 2011) and changes in the associations between gene expression can be indicative of disease, even in the absence of changes in mean expression.(Lea et al. 2019) From an evolutionary perspective, the availability of gene expression variance is what allows evolutionary change, and the genetic architecture of gene expression variance can also evolve.(Bruijning et al. 2020) Understanding the landscape of gene expression variance, and how variable it is across genes and across human populations is then a neglected avenue to understand biological evolution and our relation to the environment. In particular, we lack a clear picture of which genes show more gene expression variance, or even if the pattern of gene expression variance is consistent across populations.

Several competing forces act to shape gene expression variance,(Houle 1998; Bruijning et al. 2020) and the outcome of the interaction between these processes is still poorly understood.(Hansen 2011) 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 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.(Wagner, Booth, and Bagheri-Chaichian 1997; Pavlicev and Hansen 2011) 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.(Hansen 2011) 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 different processes shape gene expression variance is an open question. If consistent selection across populations 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 populations.

Even within individuals, gene expression is also variable across tissues.(GTEx Consortium 2017) To what extent differences in mean expression level translate to 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 a particular context. 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 reduced variation and differences in variation across tissue. 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.

Here, we use public gene expression data sets to evalu-

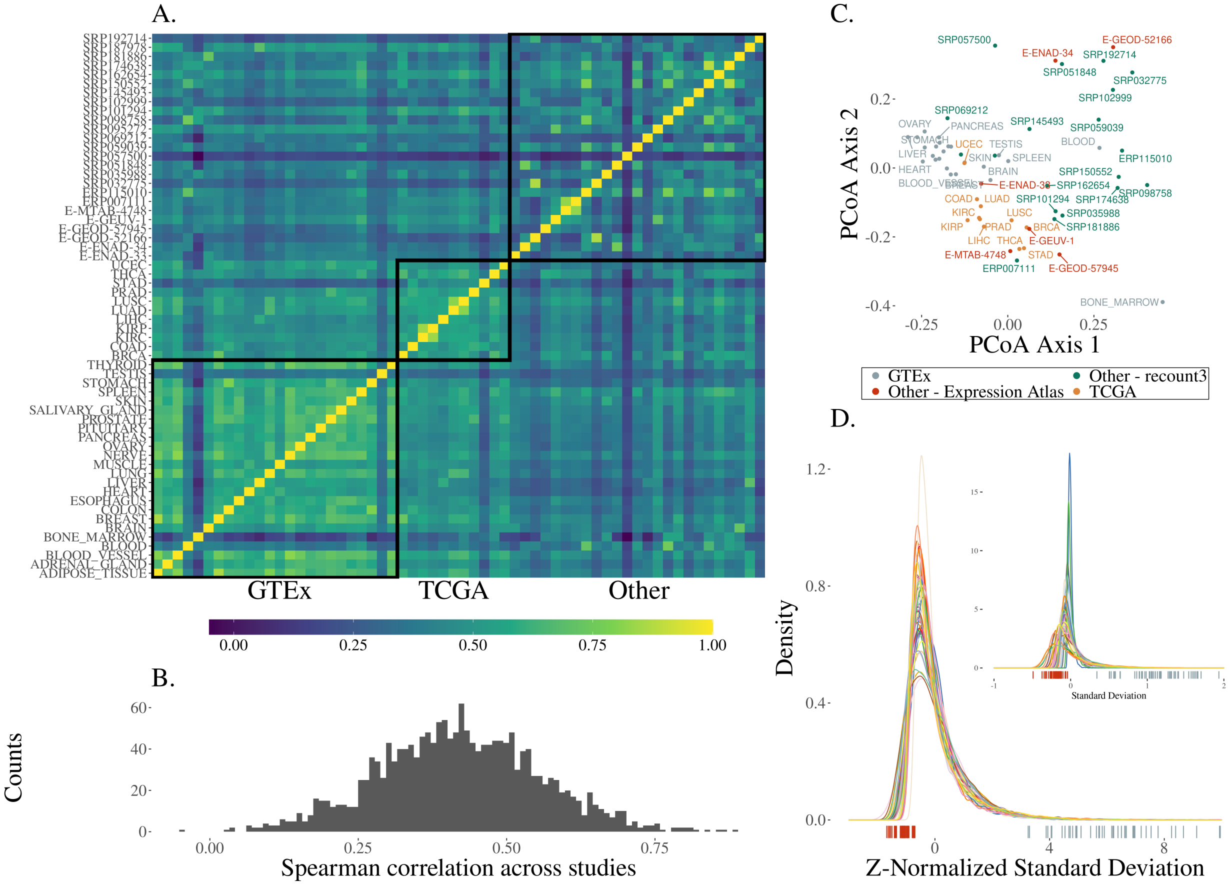

Figure 1: A. Correlation heatmap showing the across study Spearman correlation of standard deviations. Pairs of studies with more similar patterns of gene expression variance have higher correlations; B. Histogram of the correlations shown in the previous panel; C. Standard deviation correlation PCoA, with colors ; D. Density plot of standard deviations after z-normalization. Inset plot shows 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).

ate how the differences in gene expression variance is structured across independent samples. We collected and compared the gene expression variance across many studies and used the similarities across these studies to create a gene expression variance ranking, which orders genes from least variable to most variable. We then explore the expected drivers of this gene expression ranking, showing that both cis and trans regulation are involved with the determination of gene expression variance. Finally, we explored the link between gene expression variance and biological function by leveraging gene ontology annotations.

Results

Gene expression standard deviations (SDs) were calculated for each data set 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 reveals a broadly similar rank of gene expression variance, so genes that are most variable in one study tend to be most variable in all studies (fig. 1A and B). A principal coordinate analysis(Gower 1966) 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. 1C). This indicates some effect of study source on the similarity between gene expression SD across studies, which we explore in detail below. Observed range of gene expression SD across genes is variable across studies, but can be normalized so that the distributions are comparable (fig. 1D). Given that the correlations across studies are broadly high, 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 use the score of each gene in the first principal component

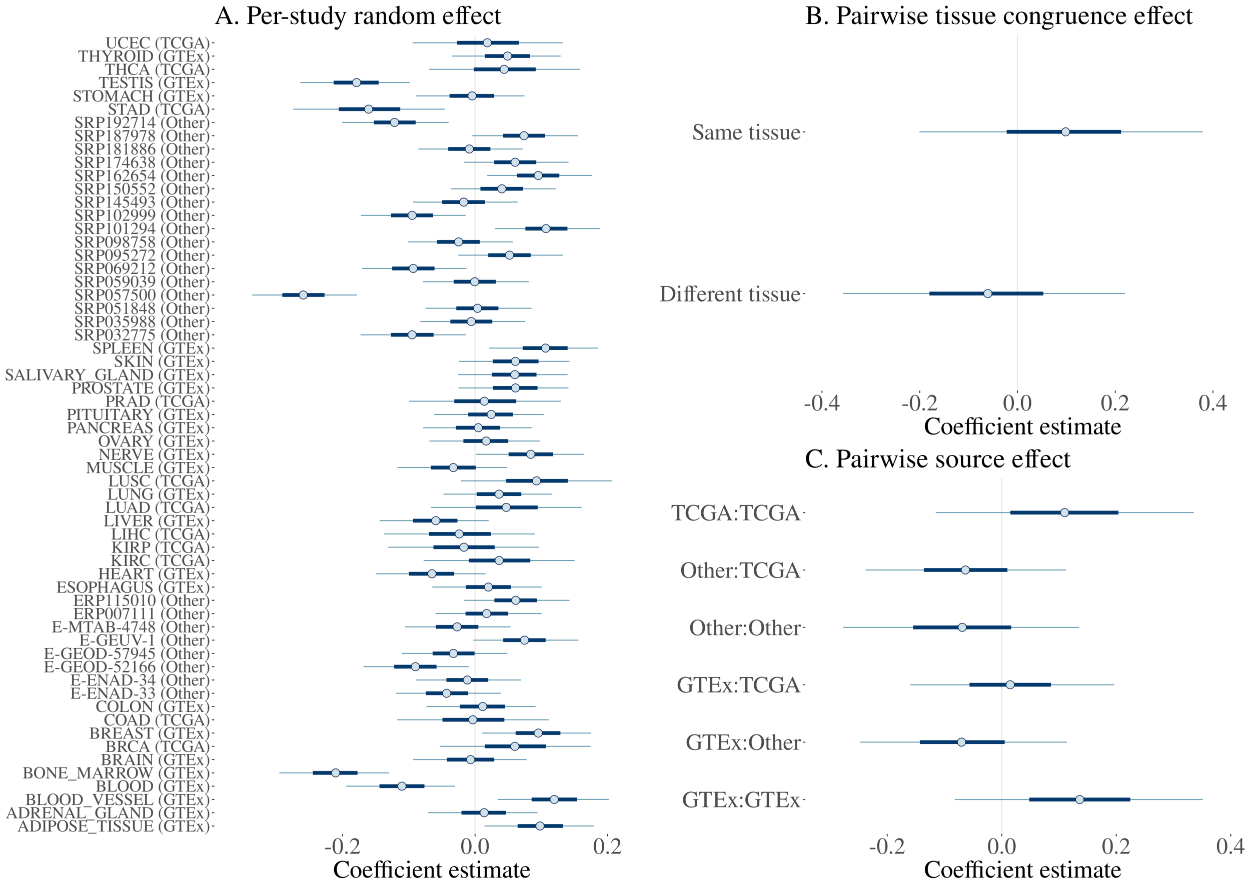

Figure 2: Coefficients estimates from a linear model using the among studies Spearman correlations as the response variable. These correlations are shown in fig. 1A and B. In the linear model, correlations are Fisher z-transformed. Study source and tissue are added as fixed effects. Coefficient estimates are shown with 50% and 95% credibility intervals. Panel A: The per-study random effect captures the non-independence of the correlation values and estimates the characteristic contribution of each study to the correlation. For example: comparisons involving bone marrow (from GTEx) tend to be lower than the others. Panels B and C: Fixed effect estimates: correlations among studies that use the same tissue are higher, and correlations involving studies in the "Other" category (non GTEx and TCGA) tend to be lower, while comparison involving GTEx and TCGA are higher.

of the Spearman correlation matrix. This generates a ranked list of genes, with most variable genes having highest rank. The red and blue ticks at the bottom of fig. 1D show the positions on the SD distributions of the least and most variable gene in our variance rank.

What drives differences in gene expression variance?

To characterize the drivers of across study similarity, we directly model the correlations across studies using a mixed effect linear model.(Dias et al. 2021a; 2021b) In this mode, we use study, sampled tissue, and study origin as predictors of the pairwise 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. 2C). Correlation involving studies that are not from TCGA and GTEx (marked as "Other") are on average lower (fig. 2C). Since these two sources are independent, this 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. 2B). 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. 2A). These studies also show up further away in the PCoA plot in fig. 1C.

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 (about ~560 genes in each group) and performed a Gene Ontology (GO) enrichment analysis within each group. This allows us to establish the representative functions of these consistently high and low-variance genes. In total, using a Benjamini-Hochberg 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 supporting table 1 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, left). In line with this GO term enrichment, the top 5% most variable genes are enriched 8-fold for genes that encode secreted proteins, relative to all other genes ($p < 10^{-3}$). 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, right); and accordingly we find that previously characterized human housekeeping genes(Hounkpe et al. 2020) are enriched within the 5% least variable genes 2-fold relative to all other genes ($p < 10^{-3}$).

We also explore the distribution of expression variance among the genes associated with GO terms. For this, we gather all biological process GO terms in level 3 (i.e. terms that are at a distance of 3 for 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 separate the genes into expression variance deciles, with the first decile having the lowest variance. We then count how many genes in each decile has 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 measure 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 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 measure 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 house keeping 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).

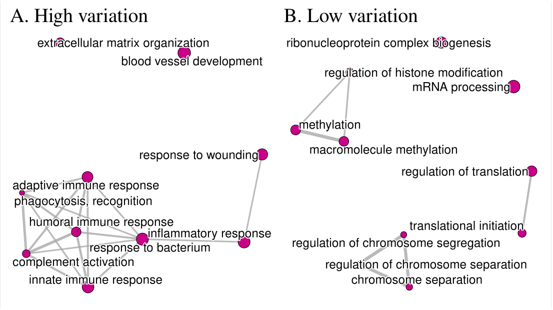

Figure 3: Gene set enrichment analyses testing for over representation of gene ontology categories in the upper and lower 5% quantiles of the gene variance rank. High-variance gene are enriched for terms related to immune function, response to wounding, blood vessel morphogenesis and inflammatory response. In contrast, 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}.

Both GO analyses suggests 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 diversity. Here, we find a partial Spearman's correlation of 0.184 ($p < 10^{-3}$). Connectivity, a proxy for regulatory interactions with other genes, 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's correlation is -0.024 ($p \approx 6 \times$

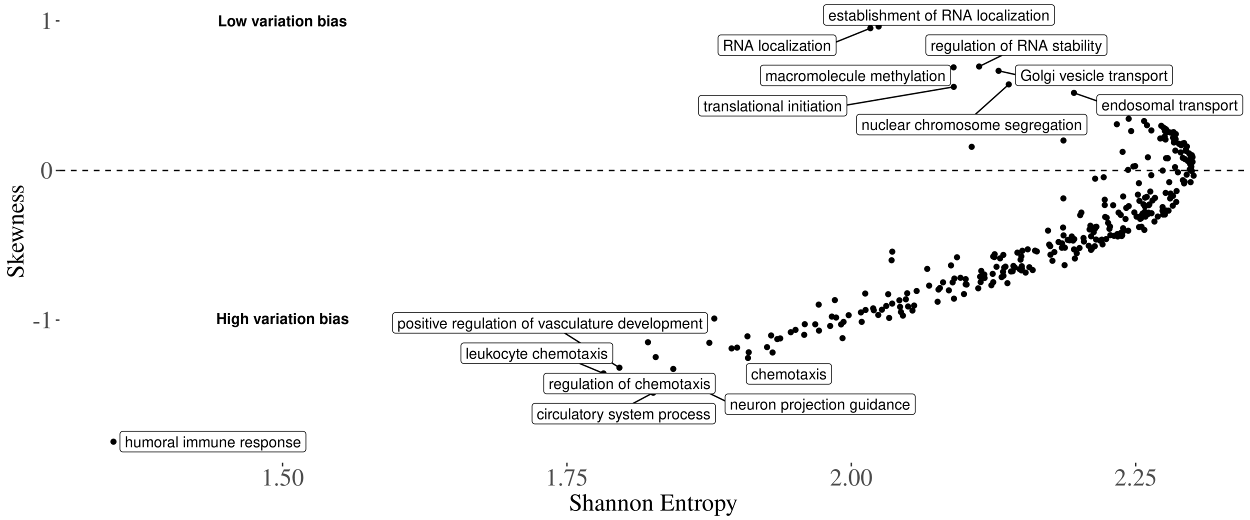

Figure 4: Relationship between skew and entropy of rank decile distributions for each GO term. The GO terms are filtered for gene counts greater than 100 as in fig. 5.

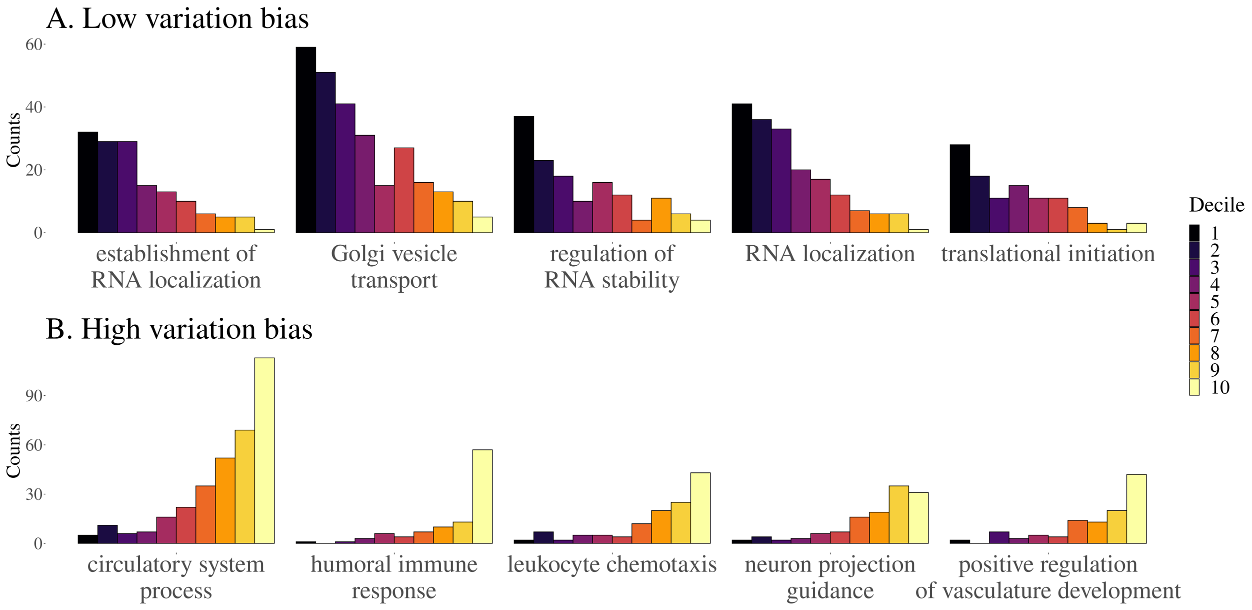

Figure 5: Distributions of decile ranks of second level GO terms. Each plot shows the count of genes in each decile of the rank. These GO terms are filtered for gene counts greater than 100 and sorted by the skewness of the distribution. The top panel shows the top 5 and the bottom panel shows the bottom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

10^{-3}). Finally, we find a partial Spearman's correlation of -0.046 ($p \approx 1 \times 10^{-3}$) for the proportion of substitutions that are adaptive.

Discussion

By using large publicly available data sets, we were able to show that gene expression variance is reasonably consistent across studies. Differences in gene expression variance were driven by technical aspects of gene expression measurement, with data derived from large consortia showing more similar patters of variance across genes; and by tissue, with studies using the same tissues also showing higher similarities. However, the largest driver of differences across studies was idiosyncratic differences related to single data sets, with tissues know to have divergent gene expression patterns (i.e. bone marrow, blood, testis, and platelets) also showing the largest differences in gene expression variance. Differences in variance are informative beyond the differences in mean expression:

it is not just that more expressed genes are more variable, residual differences in gene expression variance also carry information about tissue specific patterns.

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 (75% of correlations are between 0.45 and 0.9), consistent with measurements of expression variance in single cells and in populations of cells for various tissues.(J. Li et al. 2010; Dong et al. 2011; Alemu et al. 2014) 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 function, like RNA processing, translation, DNA methylation, and cell duplication. This pattern of enrichment is also observed when we look at enrichment for high or low variance genes within the genes associated with each terms 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 variance reducing stabilizing selection, those linked with fundamental baseline biological processes, are indeed over represented in the least variable genes. These same genes are also expected to be under strong purifying selection and show low levels of substitution and polymorphisms, which we observe. Likewise, genes whose function is constrained by myriad interactions with several other genes, those with high connectivity, also less variable. Furthermore, genes involved with direct interaction to 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 be somewhat consistent across studies, allowing us to create our ranking in the first place. We find strong support for the idea that there are indeed genes with consistently more (or less) variable expression levels.

Given this consistency, the natural question is then how do these well regulated levels of gene expression variance behave in perturbed or disease conditions. Comparing two HapMap populations, Li et al(2010) 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.(2011) 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, consistent 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 patters of expression. The authors hypothesizes that the reduced variance in Schizophrenia patients reduces the robustness of their gene expression networks.

Presumably genes will differ in their capacity to maintain their baseline variation levels, and changes in the variation level of some genes could have major physiological consequences.

Drafts:

- Variation in perturbed conditions
- Differences in robustness

Methods

Data sources

We selected 60 RNA-seq studies with large sample sizes from public gene expression repositories recount3(Wilks et al. 2021) and Expression Atlas.(Papatheodorou et al. 2020) Because we are interested in population variation of gene expression, we exclude single-cell studies and focus only on studies derived from populational 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.

Data processing pipeline

We use a standardized pipeline to measure gene expression variance while removing extraneous sources of variation. Data from case-control studies was filtered to keep only control samples.

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 5 cpm reads. To account for the mean variance relation in count data, remaining genes were subjected to the variance stabilizing transformation implemented in DESeq2.(Love, Anders, and Huber 2014) Fixed effects were manually curated from the metadata for all studies and removed using a linear fixed effect model. Outlier individuals in the residual distribution were removed using a robust Principal Component Analysis (PCA) approach of automatic outlier detection.(Chen et al. 2020) Gene expression standard deviation is measured as the residual standard deviation after fixed effect correction and outlier removal.

Variance 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, 4300 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 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 variance. This is accomplished by adding a per-study random effect, see(Dias et al. 2021b) for details. 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 variance in this principal component (PC1) to create an across study rank of gene variation. The higher the rank, the higher the gene 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.(Husson et al. 2019) 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.

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 5kb 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 10kb 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,(Nei and W. H. Li 1979) provides insight in the genetic diversity within a population, in this case 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: As a proxy for the degree of trans regulation that each gene is subjected to, we calculate the average weighted connectivity for all genes. To do this, for each study, we create a fully connected gene-by-gene graph in which each edge is weighted by the Spearman correlation between gene expression. We then trim this graph by keeping only edges for which the Spearman correlation is significant at a false discovery rate of 1%. In this trimmed network, we then take 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. As a proxy for the degree of trans regulation that each gene is subjected to, we calculate the average weighted connectivity for all genes. To do this, for each study, we create a fully connected gene-by-gene graph in which each edge is weighted by the Spearman correlation between gene expression. We then trim this graph by keeping only edges for which the Spearman correlation is significant at a false discovery rate of 1%. In this trimmed network, we then take 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.

Code availability

All code for reproducing all analysis and figures, along with a walthrough, is available at github.com/Wolfffff/exp_var.

References

Alemu, Elfalem Y et al. (Apr. 2014). “Determinants of expression variability”. en. In: *Nucleic Acids Res.* 42.6, pp. 3503–3514. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkt1364 (page 6).

Bruijning, Marjolein et al. (Jan. 2020). “The Evolution of Variance Control”. en. In: *Trends Ecol. Evol.* 35.1, pp. 22–33. ISSN: 0169-5347, 1872-8383. DOI: 10.1016/j.tree.2019.08.005 (page 1).

Chen, Xiaoying et al. (June 2020). “Robust principal component analysis for accurate outlier sample detection in RNA-Seq data”. en. In: *BMC Bioinformatics* 21.1, p. 269. ISSN: 1471-2105. DOI: 10.1186/s12859-020-03608-0 (page 7).

Dias, Filipe S et al. (Nov. 2021a). “Analysing the distance decay of community similarity in river networks using Bayesian methods”. en. In: *Sci. Rep.* 11.1, p. 21660. ISSN: 2045-2322. DOI: 10.1038/s41598-021-01149-x (page 3).

– (2021b). “BetaBayes-A Bayesian approach for comparing ecological communities”. In: (pages 3, 7).

Dong, Dong et al. (Jan. 2011). “Gene expression variations are predictive for stochastic noise”. en. In: *Nucleic Acids Res.* 39.2, pp. 403–413. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkq844 (page 6).

Fraser, Hunter B et al. (June 2004). “Noise minimization in eukaryotic gene expression”. en. In: *PLoS Biol.* 2.6, e137. ISSN: 1544-9173, 1545-7885. DOI: 10.1371/journal.pbio.0020137 (page 1).

Gower, J C (Dec. 1966). “Some distance properties of latent root and vector methods used in multivariate analysis”. In: *Biometrika* 53.3-4, pp. 325–338. ISSN: 0006-3444. DOI: 10.1093/biomet/53.3-4.325 (page 2).

GTEx Consortium (Oct. 2017). “Genetic effects on gene expression across human tissues”. en. In: *Nature* 550.7675, pp. 204–213. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature24277 (page 1).

Hansen, Thomas F (2011). “Epigenetics: adaptation or contingency”. In: *Epigenetics: linking genotype and phenotype in development and evolution*. Ed. by Brian K Hall Benedikt Hallgrímsson. University of California press Berkeley, CA, pp. 357–376 (page 1).

Houle, D (1998). “How should we explain variation in the genetic variance of traits?” en. In: *Genetica* 102-103.1-6, pp. 241–253. ISSN: 0016-6707, 1573-6857 (page 1).

Hounkpe, Bidossessi Wilfried et al. (July 2020). “HRT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets”. en. In: *Nucleic Acids Res.* 49.D1, pp. D947–D955 (page 4).

Husson, François et al. (July 2019). “Imputation of Mixed Data With Multilevel Singular Value Decomposition”. In: *J. Comput. Graph. Stat.* 28.3, pp. 552–566. ISSN: 1061-8600. DOI: 10.1080/10618600.2019.1585261 (page 7).

Jong, Tristan V de, Yuri M Moshkin, and Victor Guryev (May 2019). “Gene expression variability: the other dimension in transcriptome analysis”. en. In: *Physiol. Genomics* 51.5, pp. 145–158. ISSN: 1094-8341, 1531-2267. DOI: 10.1152/physiolgenomics.00128.2018 (page 1).

Lea, Amanda et al. (Mar. 2019). “Genetic and environmental perturbations lead to regulatory decoherence”. en. In: *Elife* 8. ISSN: 2050-084X. DOI: 10.7554/eLife.40538 (page 1).

Li, Jingjing et al. (Aug. 2010). “Gene expression variability within and between human populations and implications toward disease susceptibility”. en. In: *PLoS Comput. Biol.* 6.8. ISSN: 1553-734X, 1553-7358. DOI: 10.1371/journal.pcbi.1000910 (page 6).

Love, Michael, Simon Anders, and Wolfgang Huber (2014). “Differential analysis of count data–the DESeq2 package”. In: *Genome Biol.* 15.550, pp. 10–1186. ISSN: 1465-6906 (page 7).

Mar, Jessica C et al. (Aug. 2011). “Variance of gene expression identifies altered network constraints in neurological disease”. en. In: *PLoS Genet.* 7.8, e1002207. ISSN: 1553-7390, 1553-7404. DOI: 10.1371/journal.pgen.1002207 (page 6).

Nei, M and W H Li (Oct. 1979). “Mathematical model for studying genetic variation in terms of restriction endonucleases”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 76.10, pp. 5269–5273 (page 7).

Papatheodorou, Irene et al. (Jan. 2020). “Expression Atlas update: from tissues to single cells”. en. In: *Nucleic Acids Res.* 48.D1, pp. D77–D83. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkz947 (page 6).

Pavlicev, Mihaela and Thomas F Hansen (Dec. 2011). “Genotype-Phenotype Maps Maximizing Evolvability: Modularity Revisited”. In: *Evol. Biol.* 38.4, pp. 371–389. ISSN: 0071-3260, 1934-2845. DOI: 10.1007/s11692-011-9136-5 (page 1).

Wagner, Günter P, Ginger Booth, and Homayoun Bagheri-Chaichian (Apr. 1997). “A POPULATION GENETIC THEORY OF CANALIZATION”. en. In: *Evolution* 51.2, pp. 329–347. ISSN: 0014-3820, 1558-5646. DOI: 10.1111/j.1558-5646.1997.tb02420.x (page 1).

Wang, Zhi and Jianzhi Zhang (Apr. 2011). “Impact of gene expression noise on organismal fitness and the efficacy of natural selection”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 108.16, E67–76. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1100059108 (page 1).

Wilks, Christopher et al. (Oct. 2021). “recount3: summaries and queries for large-scale RNA-seq expression and splicing”. en (page 6).