The landscape of gene expression variation in human populations

Intro

Gene expression variation drives phenotypic difference.

Structure of genetic variation can affect .

Here we use public gene expression data sets to evaluate how the differences in gene expression variation is structures across several independent studies. We collected and compared the gene expression variation across several studies, and used the similarities across these studies to create a gene expression variation 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.

Methods

Data sources

We selected 60 studies with large sample sizes from public gene expression repositories recount3 (Wilks et al. 2021) and Expression Atlas (Papatheodorou et al. 2020).

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. The same applies for TCGA data.

Data processing pipeline

We use a standardized pipeline to measure gene expression variation while removing extraneous sources of variation.

Case-control studies were filtered to keep only control samples.

Filtering by min cpm and mean cpm. Variance stabilizing transformation from DESeq2 (Love et al. 2014). Fixed effects correction. Outlier removal using (Chen et al. 2020). Gene expression variance is measured in the residuals after fixed effect correction and outlier removal.

Gene connectivity

To estimate 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.

Variance correlation

We assessed the similarity in gene expression variation across studies by using a between study Spearman correlation matrix. using Spearman correlations avoid 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.

PCoA of studies using the Spearman correlation matrix.

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.

Gene level statistics

PopHuman stuff: pi and … (Casillas et al. 2018)

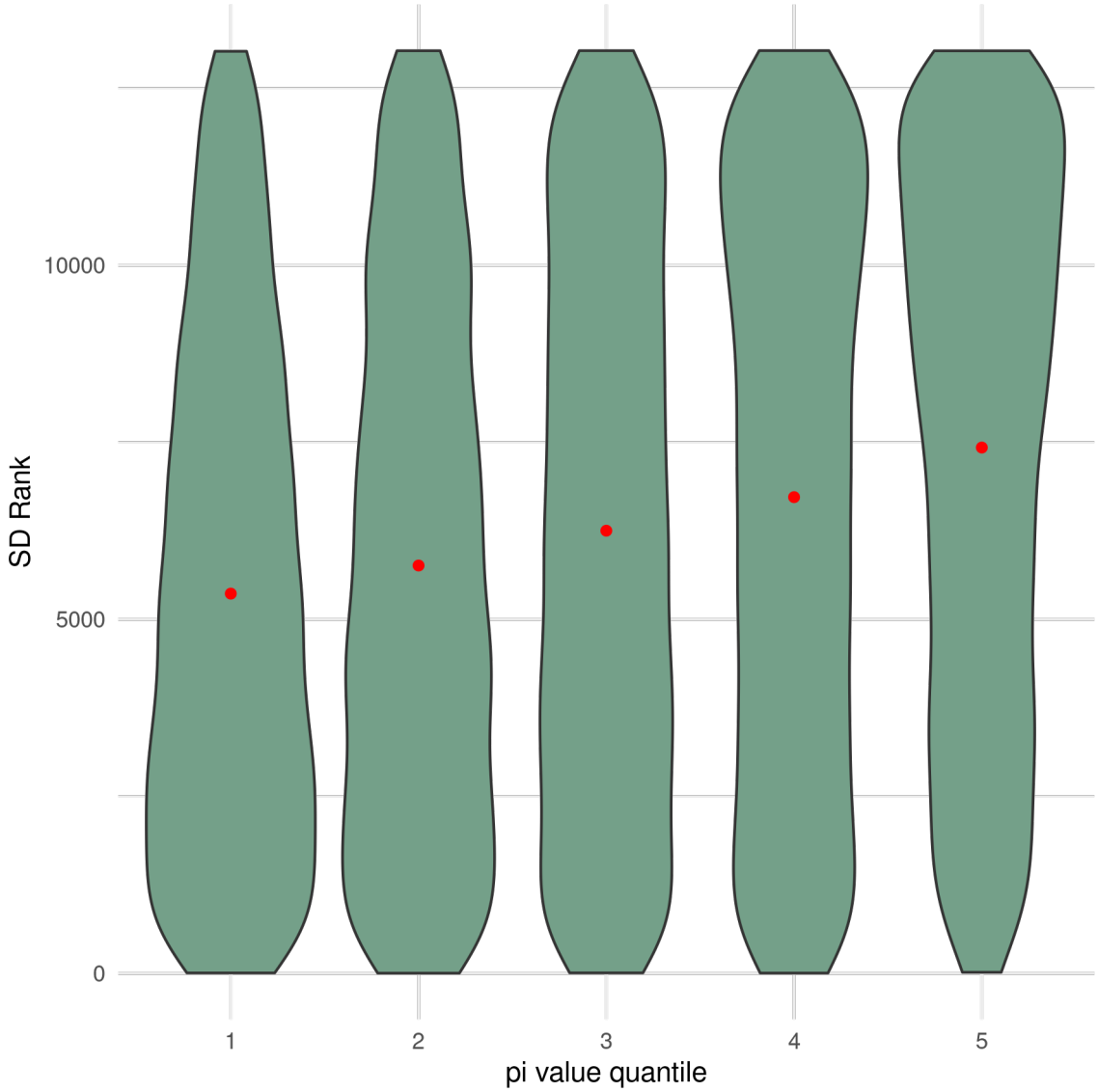

Figure 1: Violin plot showing the relationship between SD rank and mean value for genes

Gene Ontology enrichment

Results

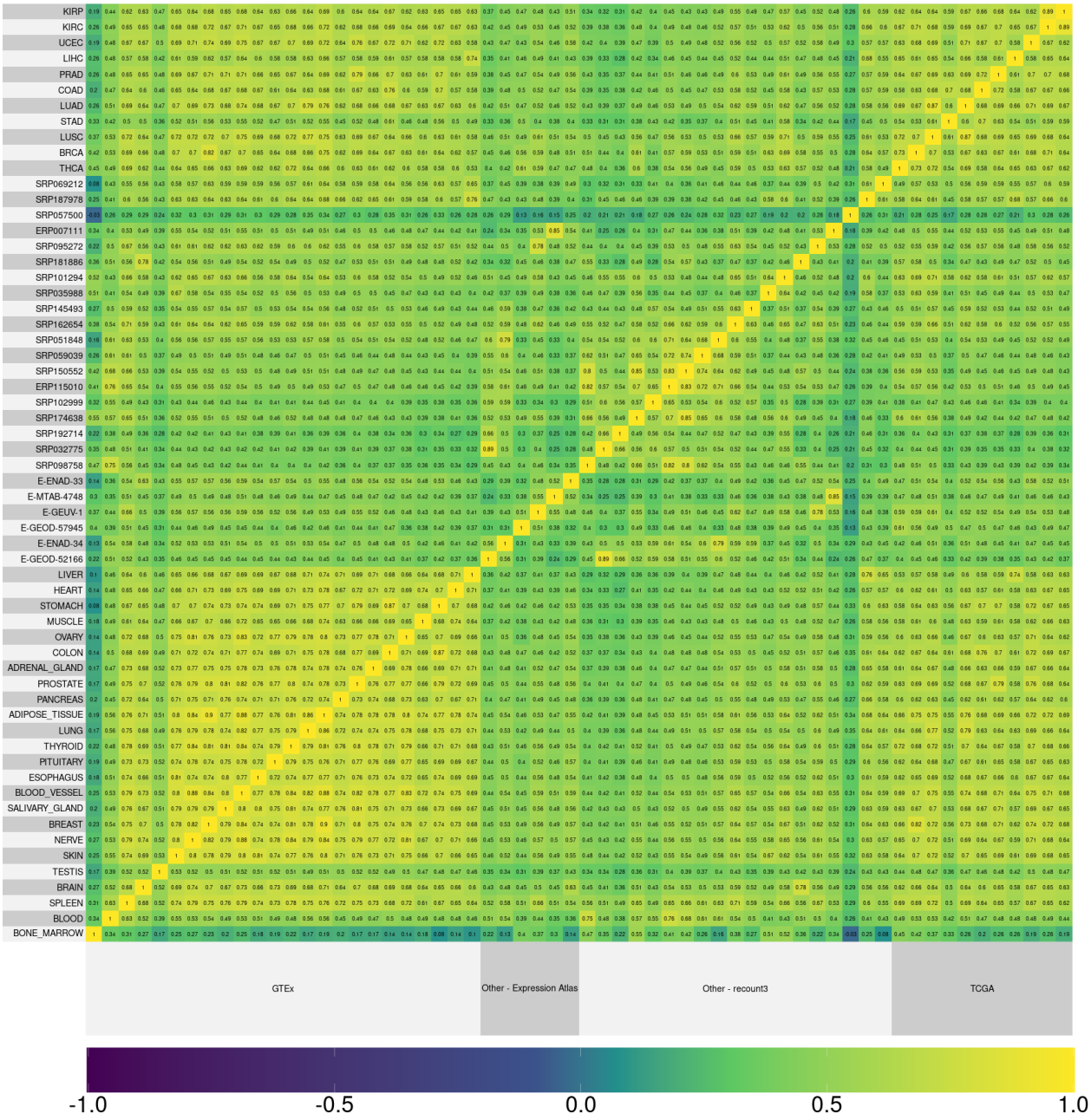

Figure 2: Correlation plot showing the cross study Spearman rank correlation of standard deviations after filtering and batch correction

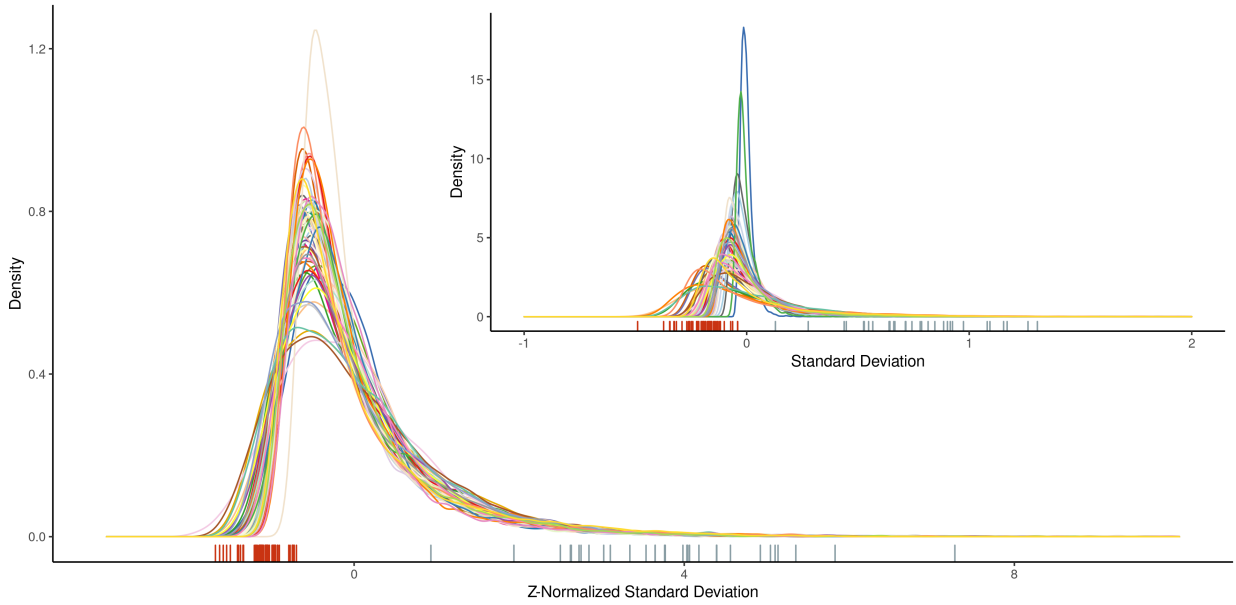

Figure 3: Density plot ofstandard deviations after z-normalization. Inset plot shows distribution of mean centered standard deviations grouped by study without normalization. The corresponding rug plots show the location of the highest ranking gene in standard deviation rank (blue) and lowest (red).

Discussion

Gene expression variance is reasonably conserved across studies. Gene expression variance is predictive of biological function. Gene expression variance can be partially explained by genetic variation and genetic associations between gene expression.

References

Casillas, S., R. Mulet, P. Villegas-Mirón, S. Hervas, E. Sanz, D. Velasco, J. Bertranpetit, H. Laayouni, and A. Barbadilla. 2018. PopHuman: The human population genomics browser. Nucleic Acids Res. 46:D1003–D1010.

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

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

Papatheodorou, I., P. Moreno, J. Manning, A. M.-P. Fuentes, N. George, S. Fexova, N. A. Fonseca, A. Füllgrabe, M. Green, N. Huang, L. Huerta, H. Iqbal, M. Jianu, S. Mohammed, L. Zhao, A. F. Jarnuczak, S. Jupp, J. Marioni, K. Meyer, R. Petryszak, C. A. Prada Medina, C. Talavera-López, S. Teichmann, J. A. Vizcaino, and A. Brazma. 2020. Expression atlas update: From tissues to single cells. Nucleic Acids Res. 48:D77–D83.

Wilks, C., S. C. Zheng, F. Y. Chen, R. Charles, B. Solomon, J. P. Ling, E. L. Imada, D. Zhang, L. Joseph, J. T. Leek, A. E. Jaffe, A. Nellore, L. Collado-Torres, K. D. Hansen, and B. Langmead. 2021. recount3: Summaries and queries for large-scale RNA-seq expression and splicing.

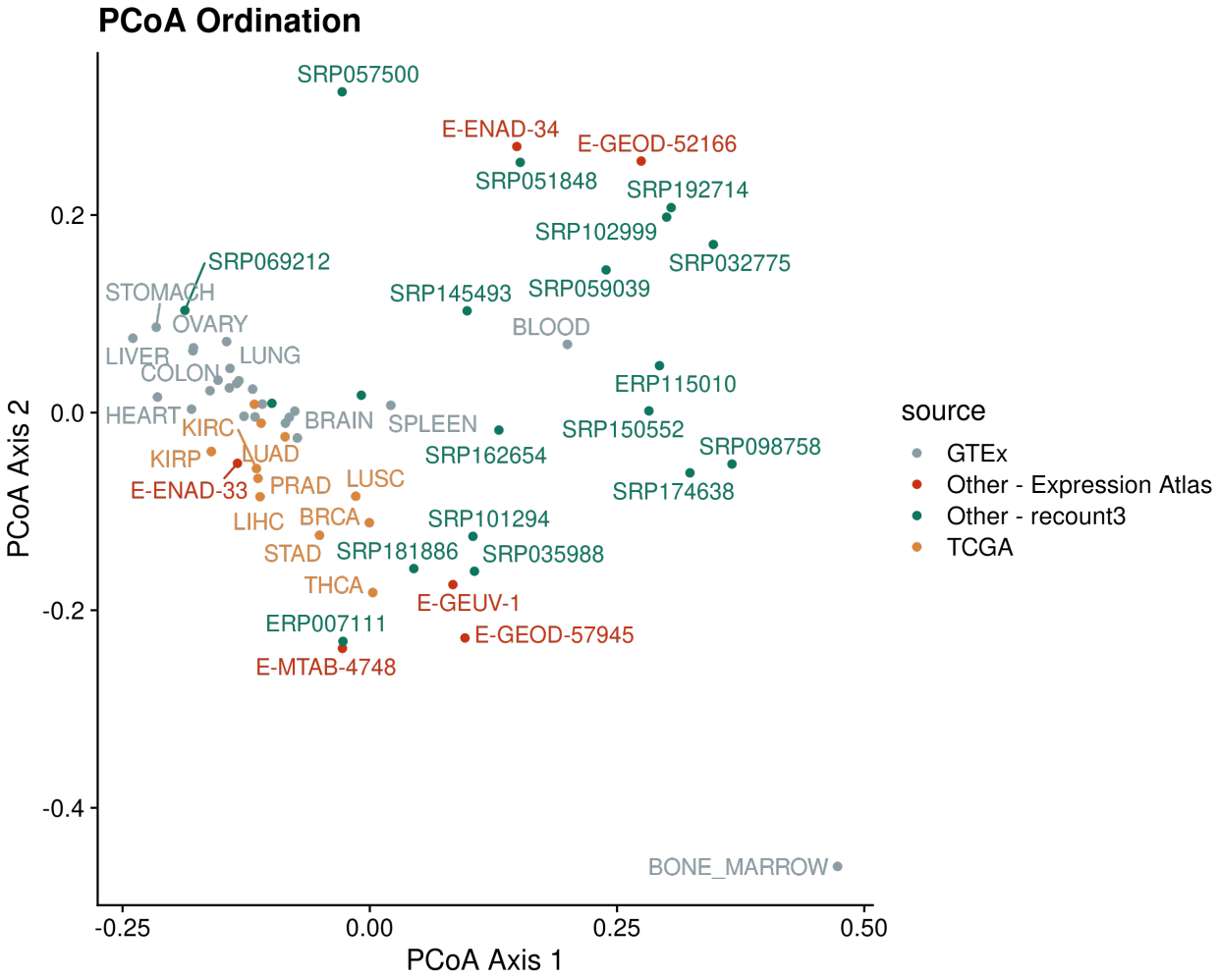

Figure 4: Standard deviation correlation PCoA