The relationship between gene expression correlation and 3D genome organization

Jason Yang, Martin Falk, Sameer Abraham

Abstract:

In some organisms such as *E. coli* and *S. cerevisiae* yeast, it is known that there is a relationship between the distance among genes and their coexpression (Pannier et. al., Kruglyak and Tang). It is also known that in general there is a relationship between gene function and genome structure (Szabo et. al). One might also expect to find a relationship between gene expression and TADs, which are domains within the genome where loci inside contact each other more frequently than loci outside. However, by analyzing data from *Mus musculus* brain cells, we do not find a relationship between gene pair correlation of single-cell RNA-seq gene expression and gene pair distance. Furthermore, despite the body of work linking gene expression and TAD structure, we also find no difference between gene pairs within a single TAD and between two TADs in terms of the relationship between gene pair distance and correlation. Additionally, we find that gene pair correlation is not related to the biological functions of the genes. However, there is a relationship between highly negative gene pair correlation and the number of times both genes are expressed 0 times across different cells.

Keywords:

TAD, Hi-C, RNA-seq, correlation

Table of Contents:

1. Introduction
2. Results
    a. Distance and Correlation
    b. Function and Correlation
    c. Shared Zeros and Correlation
3. Conclusion

Introduction:

In some organisms such as *E. coli* and *S. cerevisiae* yeast, it is known that there is a relationship between the distance among genes and their coexpression (Pannier et. al., Kruglyak and Tang). It is also known that in general there is a relationship between gene function and genome structure (Szabo et. al). One might also expect to find a relationship between gene expression and TADs, which are domains within the genome where loci inside contact each other more frequently than loci outside. We investigate if these relationships hold in *Mus musculus* neurons.

Single-cell RNA sequencing measures the activity of all genes in a variety of cell types. The data is collected by isolating the desired cells, converting RNA strands into cDNA strands, and

counting the frequency of genes in those strands. To make this counting practical, the cDNA is amplified, or duplicated many times, using techniques such as PCR (Luecken et. al.)

TADs, or topologically associating domains, are regions of DNA where genes interact with each other more often inside than outside of (Szabo et. al.) They are shown in Hi-C as small, bright squares along the diagonal within a chromosome. For background, Hi-C is a form of chromosome conformation capture that measures how often pairs of DNA loci contact. First, physically close pairs of DNA loci are bonded together, or crosslinked, by chemicals such as formaldehyde. The non-bonded areas are then cut out with enzymes, leaving only the crosslinked areas. The two strands of each piece of crosslinked DNA are joined, or ligated, into a single strand, and the crosslinkers are removed. This leaves small strands of DNA each coming from two separate locations in the original DNA. (Fig. 1) These strands are sequenced to count how often pairs of loci have been crosslinked. The sequencing results estimate how likely any two loci on DNA are to contact each other within the nucleus. (van Berkum et. al.)

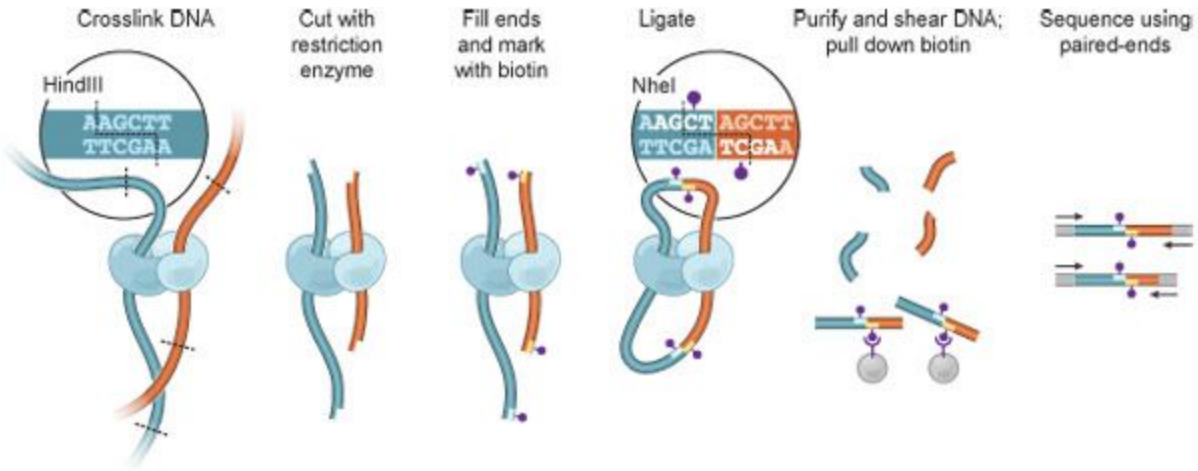

Figure 1: Description of Hi-C as a chromosome conformation capture method, a way of measuring how often any two DNA loci contact each other in the nucleus (van Berkum et. al.). Places on DNA that are close together are first bound, or crosslinked, with chemicals such as formaldehyde. The pieces of DNA not crosslinked are then cut away. The loose ends of each pair of crosslinked DNA loci are ligated together before the crosslinking chemical is removed, and the resulting DNA pieces are sequenced.

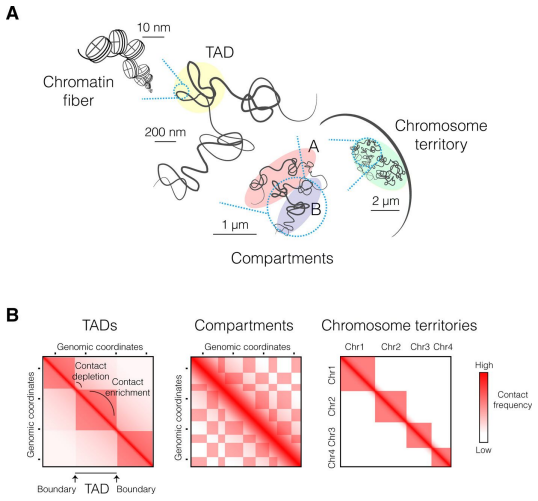

Figure 2: TADs are small areas of DNA where loci within a single TAD contact each other more frequently than loci between two TADs. Compartments are larger groups of loci, spread across the genome, where loci in the same compartment contact each other more frequently than loci from two different compartments (it is theorized that there are mainly two kinds of compartments). Chromosome territories are chromosomes themselves, since loci within a single chromosome contact each other much more frequently than loci across two chromosomes. TADs are generally smaller than compartments, which are smaller than chromosome territories (Szabo et. al.).

Results:

Distance and Correlation:

We acquired the normalized counts of single-cell RNA-seq data from He et. al. and did not remove any outliers. We first investigated the relationship between the distance and correlation of pairs of genes within a single chromosome. Here we define the distance between two genes to be the distance between their centers and define the center of a gene to be the average of the positions of its endpoints. In other words, if gene A has endpoints at a_0 and a_1 and gene B has endpoints at b_0 and b_1, then the distance between the two genes is $|(a_0+a_1)/2 - (b_0+b_1)/2|$. We define the correlation between two genes to be the Pearson correlation of the number of times both genes are expressed across all cells. Figure 3 shows that the average correlation stays relatively constant at around 0 as distance increases. This trend is preserved when gene pairs are plotted separately by chromosome.

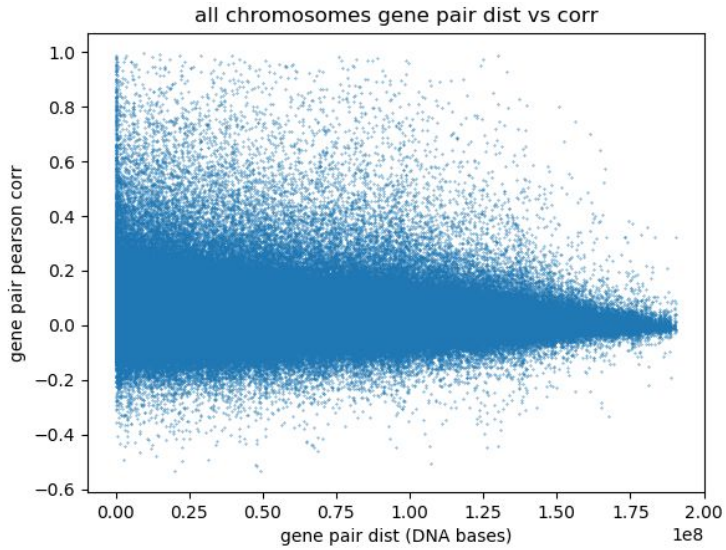

Figure 3: Distance vs. correlation between all pairs of genes within a single chromosome, for all chromosomes. Here the distance between two genes is defined as the distance between their centers, and define the center of a gene to be the average of the positions of its endpoints; the correlation between two genes is defined as the Pearson correlation of the number of times both genes are expressed across all cells.

Since there is a hypothesis that there is a relationship between gene activity and genome folding, specifically that of TADs, we looked to see if accounting for TADs would reveal any interesting results regarding the distance-to-correlation relationships. We got a list of TADs of retinal neurons from Falk et. al. and compared gene pairs that were within a single TAD (intraTAD) to pairs across two different TADs (interTAD). In the violin plot below, which only shows gene pairs with distance less than 500,000 DNA bases, the interTAD distributions of gene pair

correlation are in light blue and the intraTAD distributions of gene pair correlation are in blue. There is little difference between the distance-to-correlation relationship when computed over gene pairs within a single TAD versus gene pairs across two TADs (Fig. 4). This is true even when gene pairs are plotted separately by chromosome.

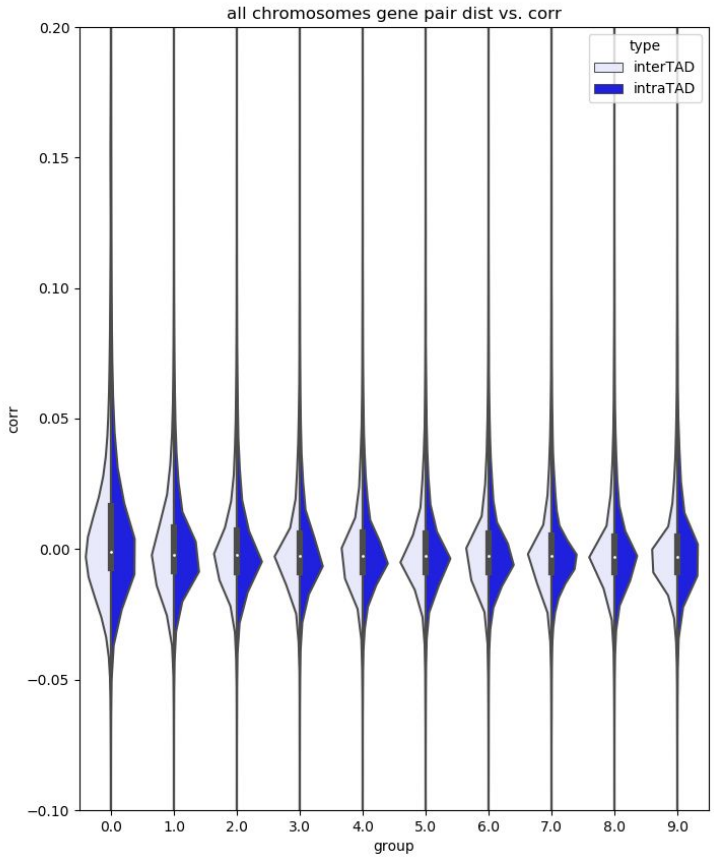

Figure 4: Violin plot of gene pair distance vs. RNA-seq correlation along all cells, with distance binned by 50,000s; only gene pairs with distance <500,000 are shown. Each violin shows the distribution of gene pairs across two TADs on the left (interTAD), and gene pairs within a single TAD on the right (intraTAD).

Finally, we accounted for different cell types among the cells in the RNA-seq data. These cell types were determined by He et. al. using BackSPIN clustering on the RNA-seq data, choosing the 6th split level.

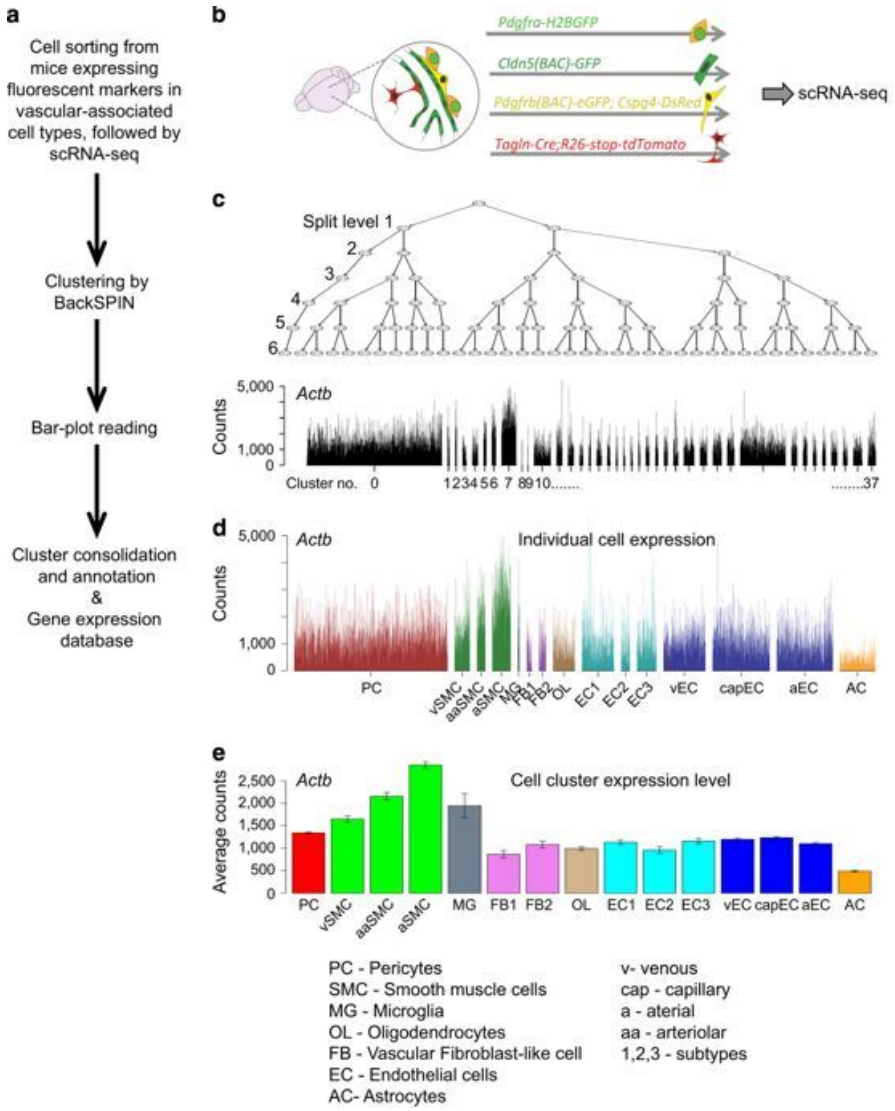

Figure 5: Derivation of cell types from He et. al. using the RNA-seq data. The authors used BackSPIN to cluster the cells based on the gene expression data, decided on choosing the 6th split level, and applied manual inspection to determine that there were 15 cell clusters.

We repeated Figure 4 but instead used correlation values between genes on RNA-seq data on only the cells from a single cluster. We show two such plots based on the two largest cell clusters. Again, there is little difference between gene pairs within a single TAD and across two TADs in terms of the distance-to-correlation relationship (Fig.s 6, 7).

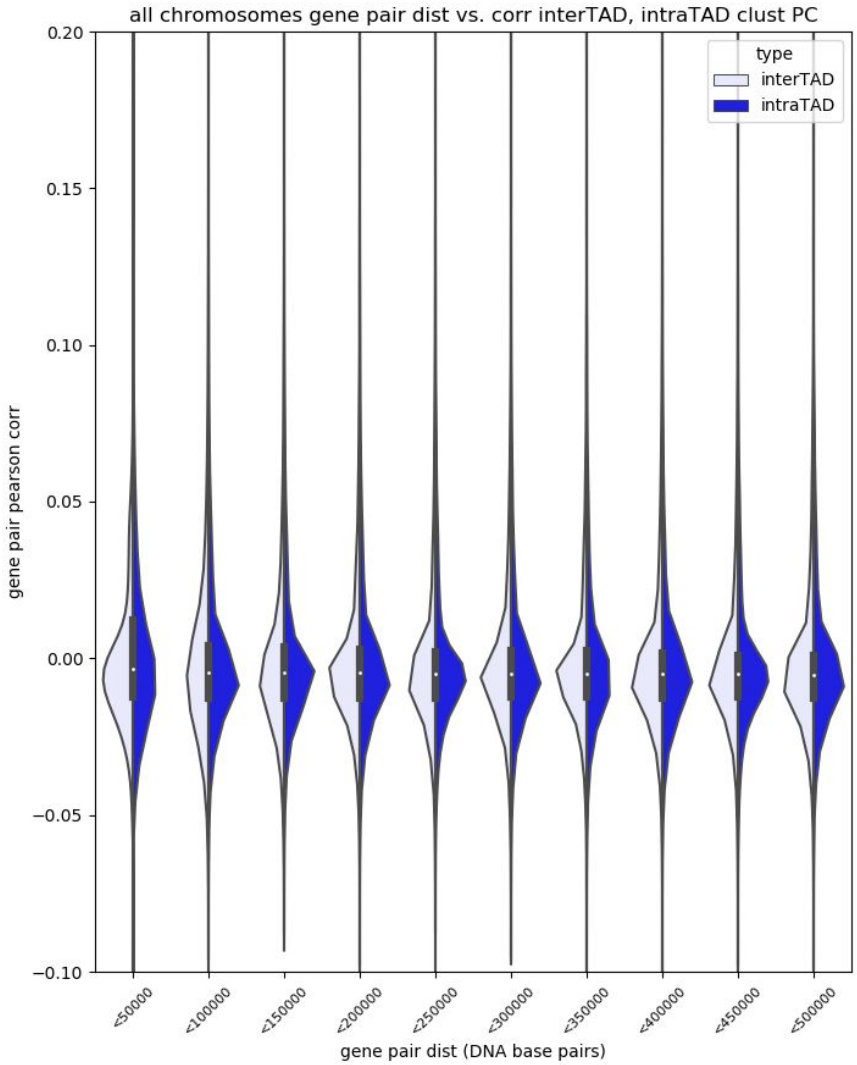

Figure 6: Violin plot of gene pair distance vs. RNA-seq correlation among cells of type PC (pericytes), determined by He et. al., the cell type containing the largest number of cells, with distance binned by 50,000s; only gene pairs with distance <500,000 are shown. Each violin shows the distribution of gene pairs across two TADs on the left (interTAD), and gene pairs within a single TAD on the right (intraTAD).

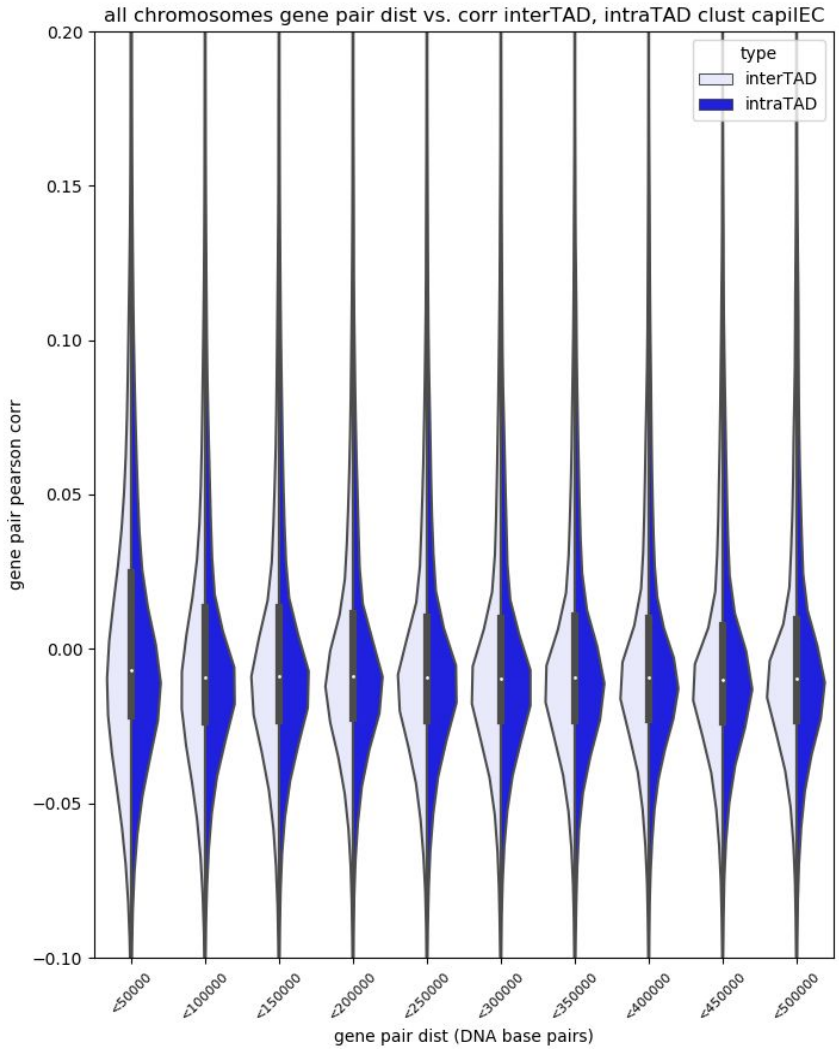

Figure 7: Violin plot of gene pair distance vs. RNA-seq correlation among cells of type capilEC (capillary endothelial cells), determined by He et. al., containing the second largest number of cells, with distance binned by 50,000s; only gene pairs with distance <500,000 are shown. Each violin shows the distribution of gene pairs across two TADs on the left (interTAD), and gene pairs within a single TAD on the right (intraTAD).

Gene Function and Correlation:

We tried finding other signs of statistical significance. We first examined the genes from gene pairs with high correlation, across all chromosomes. Using panther.db, we found that the respective distributions of molecular functions (Fig. 8, 9) and biological processes (Fig. 10, 11) for these genes compared to all genes from He et. al. were similar. We noticed that in molecular function, transporter activity was overrepresented in highly correlated genes (see Fig. 8, 10 captions), whereas in biological processes, response to stimulus was overrepresented and metabolic processes were underrepresented. We are not sure how statistically significant these differences are. Separating the genes by chromosome or by whether the gene pairs they come from are interTAD or intraTAD shows stronger overrepresentation and underrepresentation, but we are also not sure how much this is due to actual biological differences or to noise in our data.

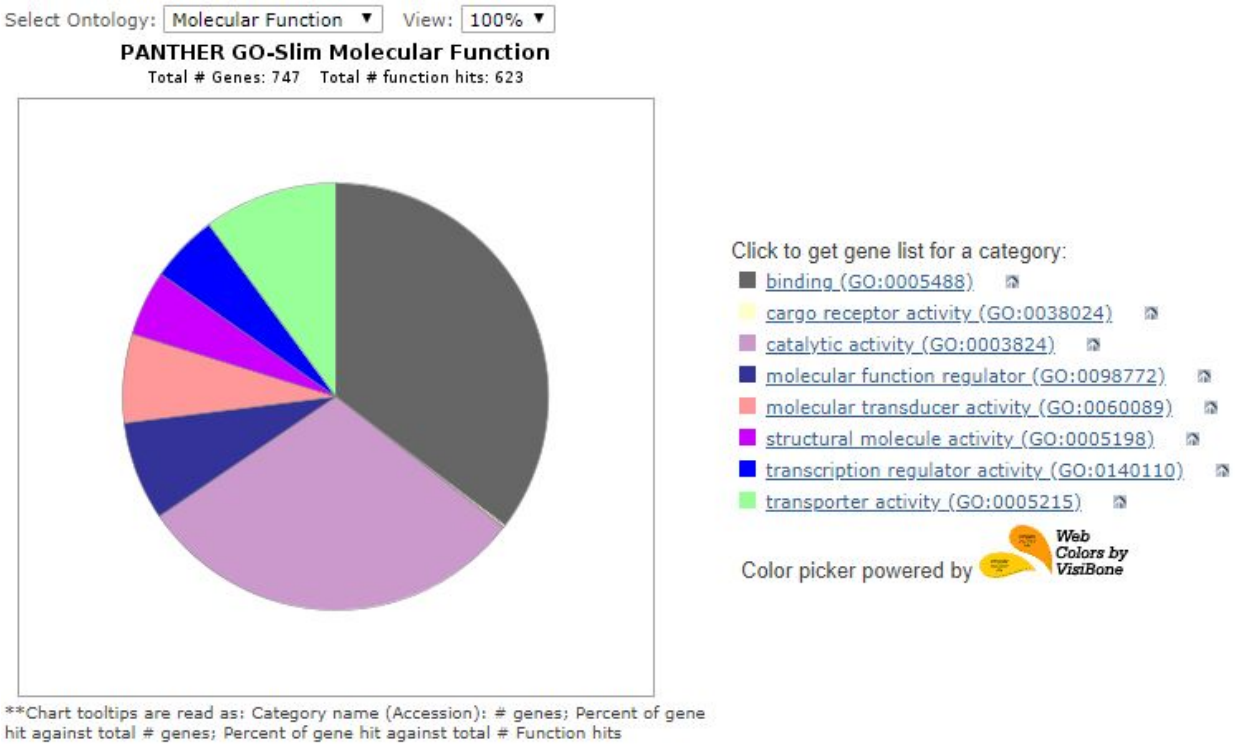

Figure 8: Distribution of molecular function of genes coming from gene pairs across all chromosomes with Pearson correlation >0.2. Genes were filtered before being correlated, according to their frequency of expression across cells. Only genes that were expressed in at least 20% and at most 80% of the cells were allowed. A gene is expressed in a cell if the RNA-seq detects at least one transcript of the gene.

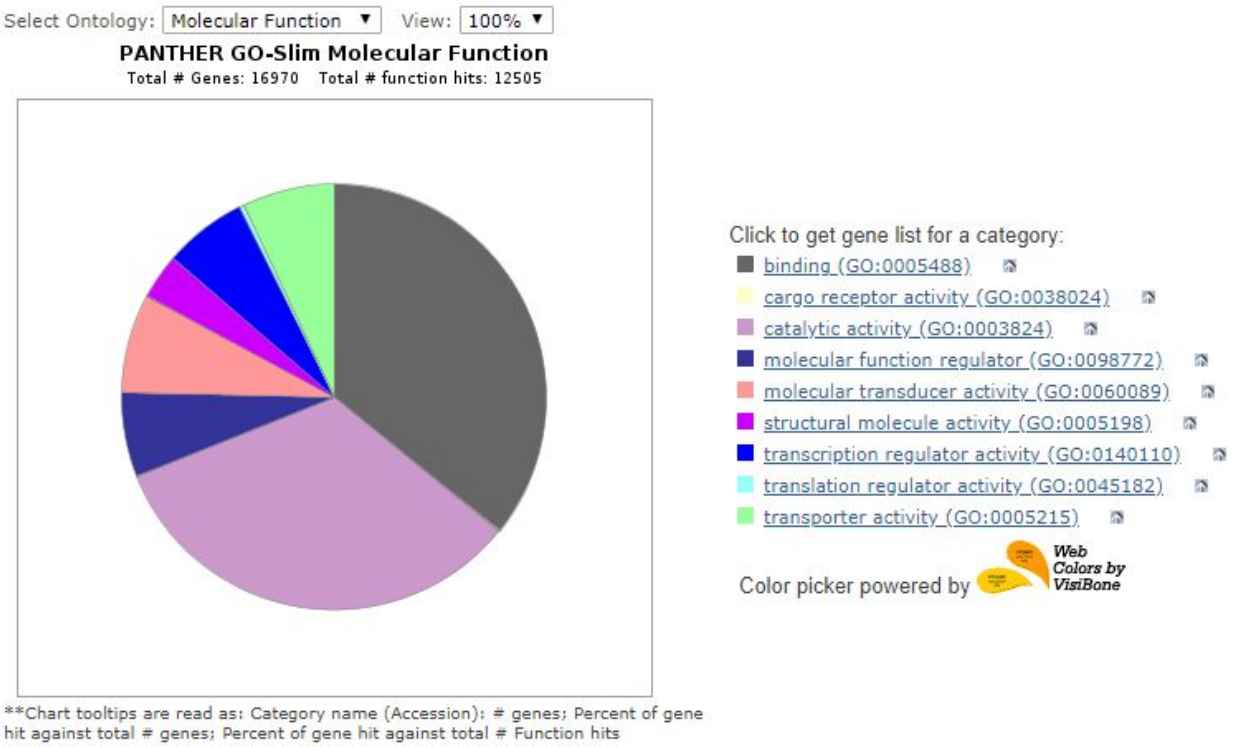

Figure 9: Distribution of molecular function of all genes from the RNA-seq database of He et. al.

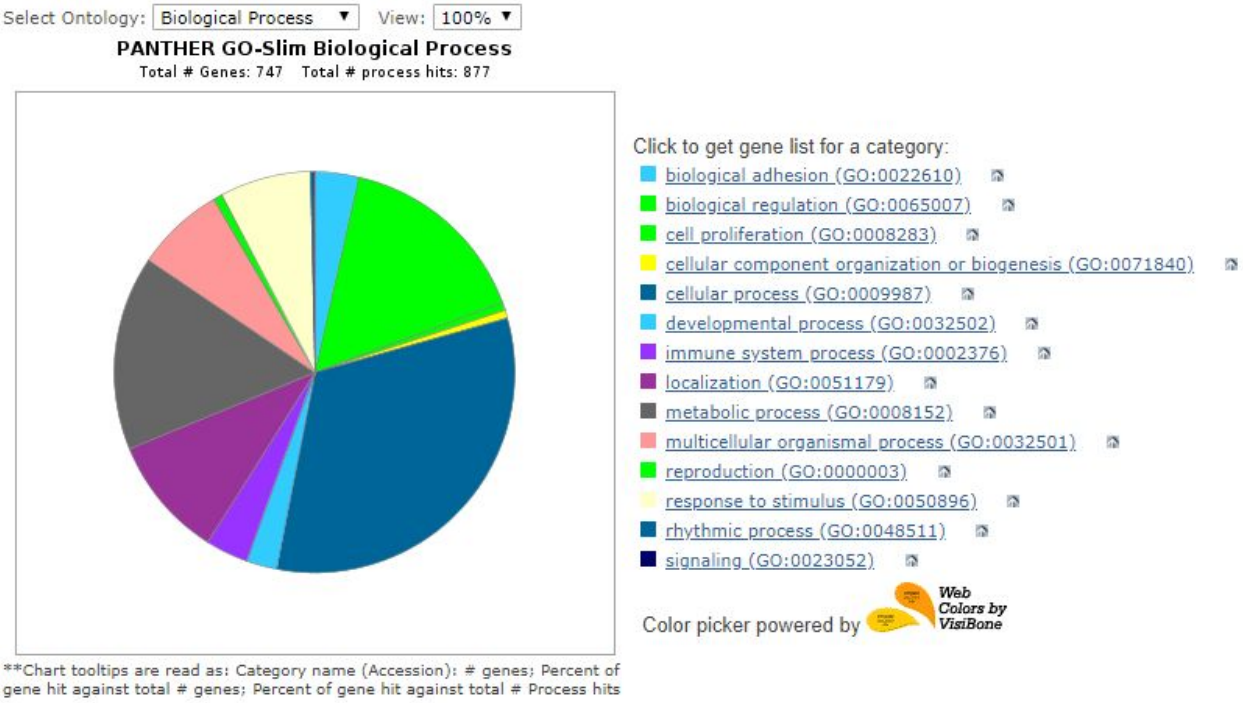

Figure 10: Distribution of biological processes of genes coming

from gene pairs across all chromosomes with Pearson correlation >0.2. Genes were filtered before being correlated, according to their frequency of expression across cells. Only genes that were expressed in at least 20% and at most 80% of the cells were allowed. A gene is expressed in a cell if the RNA-seq detects at least one transcript of the gene.

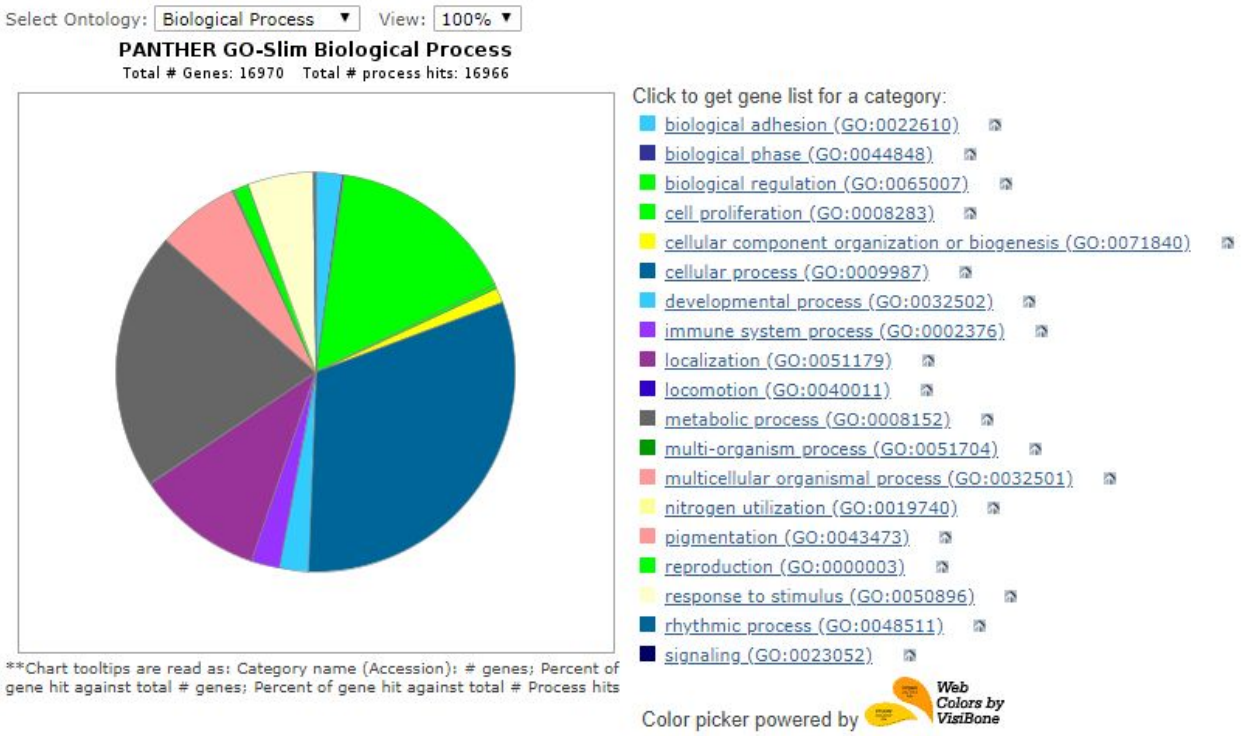

Figure 11: Distribution of biological processes of all genes from the RNA-seq database of He et. al.

Shared Zeros and Correlation:

Outside of the biological function of genes, we found some interesting patterns in the RNA-seq data. When we randomly shuffled all entries of the RNA-seq matrix, we found that the resulting distribution of gene pair correlations had a much less negative minimum than that of the original data. Specifically, for correlations between genes from chromosome 1, the lowest correlation coefficient from the shuffled data was about -0.039 (Fig 12), while the lowest coefficient from the real data was about -0.448 (Fig 13).

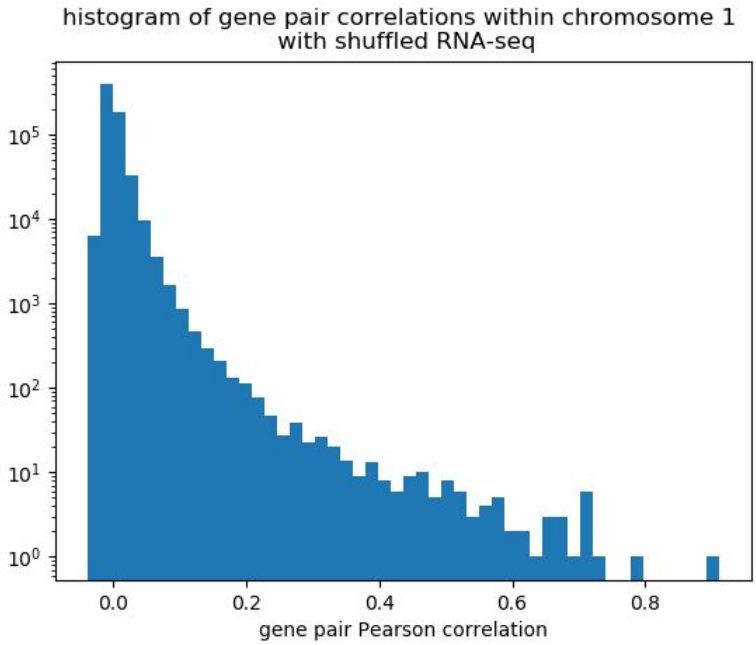

Figure 12: Histogram of gene pair correlations in chromosome 1 when the RNA-seq matrix cells are randomly shuffled. The lowest Pearson correlation in this graph is -0.039

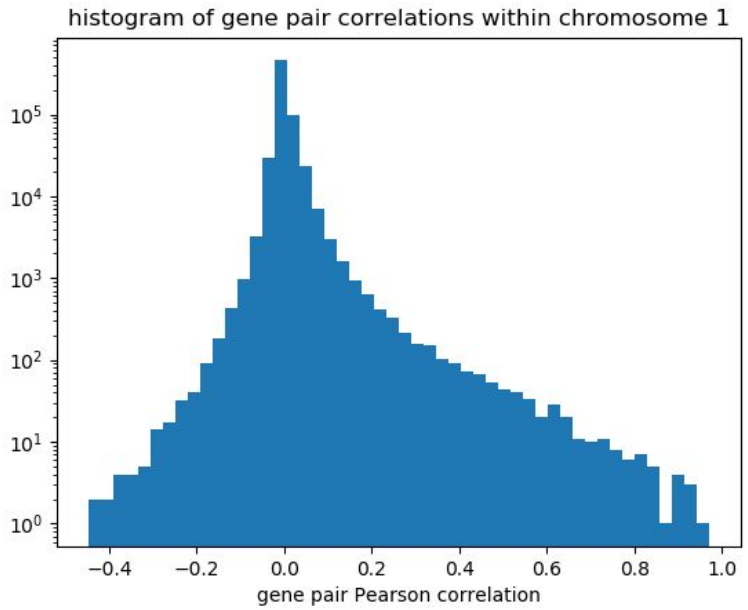

Figure 13: Histogram of gene pair correlations in chromosome 1 when using the original RNA-seq data. The lowest Pearson correlation in this graph is -0.448.

We decided to look at the number of shared zeros of each gene pair and see if they were affecting correlations, since RNA-seq tables tend to have the number 0 appear frequently, due to dropout, and having many duplicate numbers can have a significant effect on the resulting correlation coefficient.

To examine the effect of shared zeros on gene pair correlation, we focused on pairs of genes from the real data with correlation less than the minimum correlation in the shuffled data. For each pair of genes, we plotted the RNA-seq values of one gene across all cells onto the x-axis and the RNA-seq values of the other gene onto the y-axis, and we looked at the number of times the point (0,0) was plotted. For gene pairs with correlation less than the minimum correlation on the shuffled data, the origin was plotted on average 1377.15 times. This is significantly lower than the average number for all pairs of genes from the real data, which was 2441.87, and the average number for all pairs of genes from the shuffled data, which was 2415.99. In the real data, gene pairs with highly negative correlation tend to have a small number of (0,0) points, and gene pairs with highly positive correlation tend to have a high number of (0,0) points, as indicated by the two tails in the scatterplot of Fig. 14. In Figure 15, the range of the number of (0,0) points for each pair of genes is much smaller, and there is a less clear relationship between correlation and number of shared zeros. This suggests that the number of shared zeros affects how negative gene pair correlation can be. Although Figure 14 and 15 only apply to chromosome 1, the same trend where gene pairs with few shared zeros are associated with highly negative correlation occurs for genes in all other chromosomes.

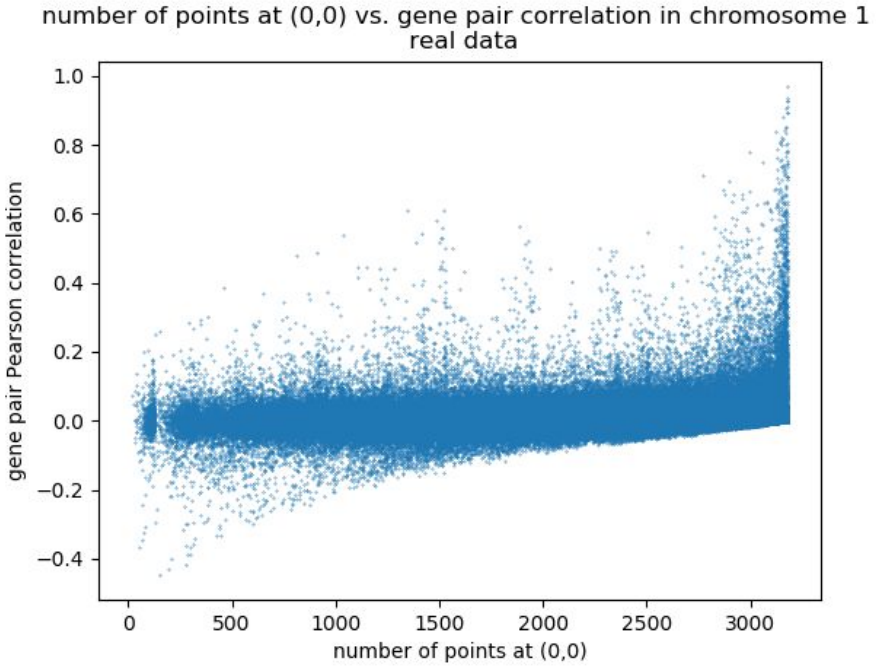

Figure 14: Plot of number of times the point (0,0) appears in the 2D-plot of gene expressions for each gene pair in the original data. The range of the number of times (0,0) appears is huge. Additionally, gene pairs with highly negative correlation tend to have a small number of (0,0) points, and gene pairs with highly positive correlation tend to have a high number of (0,0) points, as indicated by the two tails in the scatterplot.

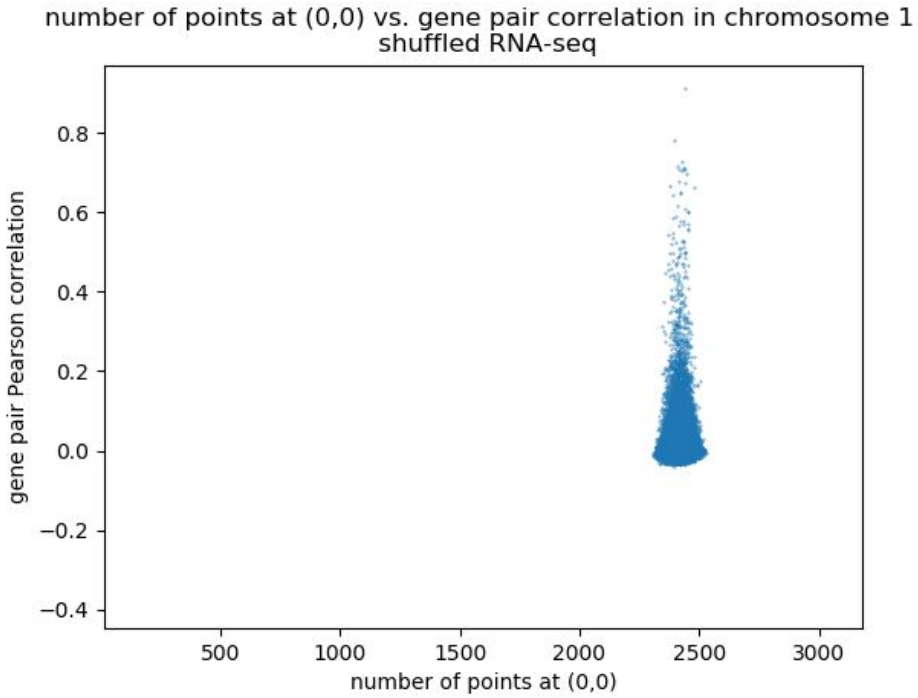

Figure 15: Plot of number of times the point (0,0) appears in the 2D-plot of gene expressions for each gene pair in the shuffled data. The range of the number of times (0,0) appears is very small.

Conclusion:

Because there is a relationship between gene function, structure, and coexpression, in organisms such as *E. coli* and *S. cerevisiae*, we looked to see if the same relationship is present in *Mus musculus* cells. However, we did not find a relationship between gene pair correlation of single-cell RNA-seq gene expression and gene pair distance. Additionally, there was no change in this result when considering gene pairs within a single TAD versus those across two TADs. Even when examining the few highly-correlated genes, we found that correlation did not have much relation with the biological function of the genes. However, after comparing the scatterplots of shared zeros and gene pair correlations for the shuffled and un-shuffled data, we did see that the shared zeros in the original data allows some genes to have highly negative correlation. In the future, we will redo this experiment with other gene sequencing data, such as single-cell ATAC-seq.

Bibliography

Luecken, Malte D, and Fabian J Theis. “Current Best Practices in Single‐Cell RNA‐Seq Analysis: a Tutorial.” Molecular Systems Biology, John Wiley & Sons, Ltd, 19 June 2019, https://www.embopress.org/doi/10.15252/msb.20188746.

Kruglyak, Semyon, and Haixu Tang. "Regulation of adjacent yeast genes." Trends in Genetics. Cell, www.cell.com/trends/genetics/comments/S0168-9525(99)01941-1. Accessed 6 Sept. 2019.

Szabo, Quentin, et al. “Principles of Genome Folding into Topologically Associating Domains.” Science Advances, American Association for the Advancement of Science, 1 Apr. 2019, https://advances.sciencemag.org/content/5/4/eaaw1668.

van Berkum, Nynke L, et al. “Hi-C: a Method to Study the Three-Dimensional Architecture of Genomes.” Journal of Visualized Experiments : JoVE, MyJove Corporation, 6 May 2010, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3149993/.

Works Cited

Falk, Martin, et al. “Heterochromatin Drives Compartmentalization of Inverted and Conventional Nuclei.” Nature, U.S. National Library of Medicine, 5 June 2019, https://www.ncbi.nlm.nih.gov/pubmed/31168090.

He, Liqun, et al. “Single-Cell RNA Sequencing of Mouse Brain and Lung Vascular and Vessel-Associated Cell Types.” Nature News, Nature Publishing Group, 21 Aug. 2018, https://www.nature.com/articles/sdata2018160.

Pannier, Lucia, et al. “Effect of Genomic Distance on Coexpression of Coregulated Genes in E. Coli.” PloS One, Public Library of Science, 18 Apr. 2017, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5395161/.

Acknowledgments:

We would like to thank Martin Falk and Sameer Abraham for their mentorship, along with PRIMES Computational Biology and MIT Mirny Lab for their support in making this research possible.