The Data Open

Final Report - Team 19

Niccolo Dalmasso, Kwangho Kim, Alan Mishler, Chirag Nagpal

October 20, 2017

1 Introduction and Topic Questions

Large-scale genomic data has the potential to revolutionize medicine. An understanding of how genes and gene expression levels interact with various diseases can pave the way for the development of treatments that target specific genes or gene expression pathways, including treatments that are customized to individuals' genetic makeup.

The human genome is vast, consisting of roughly 3 billion base pairs distributed across 23 chromosomes. These base pairs together comprise some 20,000 genes, each of which has the potential to be associated with a given disease. Finding genuine (non-spurious) relationships between genes and diseases is therefore a non-trivial task.

Here, we investigate gene expression data to investigate two substantive questions:

1. Are there groups of genes whose expression levels differ between cancerous and non-cancerous tissues?
2. Can we use these same groups of genes to identify expression level patterns in other diseases besides cancer?

2 Datasets

Our data comes from consists of the following:

1. `genes`
   This dataset includes descriptive information of the $5,000$ that are included in the Genotype-Tissue Expression (GTEx) experiment.

2. **GTEx datasets**
   We have available a number of datasets related to the Genotype-Tissue Expression experiment, which was carried out on 570 healthy adult and post-mortem donors, with a total of $8,555$ samples across 53 tissues and 2 cell line lines. On top of information about the modeling of such genes, information about such tissues and the integrity of such samples are available to us. Finally, the full gene expressions are given using the "RPKM" (Reads Per Kilobase Million) measurement, for a total of ~ 43 million rows.

3. **TCGA datasets**
   We have data from the Cancer Genome Atlas (TCGA) experiment, in which genes expressions are recorded considering patients who were cancer-positive and across 8 different tissues. The expression level is measured in "FPKM" (Fragments Per Kilobase Million).

4. **Chemical and Disease Dataset**
   We have data relating the disease which appear to be more associated with a given gene, expressed in terms of the "inference score", which roughly indicates the degree of association between a gene and a disease. We also have data relative to the presence of a given chemical component in a specific genes.

Overall, while we will be working more with the two gene expressions datasets for GTEx and TCGA, information given by the other datasets are going to be useful to provide guidance on which genes and tissues could lead to the most insight.

3 Exploratory Data Analysis and Data Manipulation

The following issues are to be considered in this case:

1. As measures, "RPKM" (Reads Per Kilobase Million) and "FPKM" (Fragments Per Kilobase Million) are not directly comparable, as they do not measure the same biological quantities.[1] It can actually happen that it some cases the two quantities are equal, but since we do not have any information in order to address this issue we are going to consider them as different measurement. Hence, they are not directly comparable, which implies that a direct comparison between the gene expression for the two experiments, GTEx and TCGA, might not be possible;

2. Not all the $5,000$ genes included in GTEx experiment are included in TCGA. We therefore we reduce the number of genes we take into consideration to the number of genes included in both studies, which is $4,961$;

3. The number of tissues for the two experiments is quite different and not all of the 8 tissues for TCGA do actually match tissues in GTEx. in order to account for this, we have only included samples which come from tissues that are available in both cases. We then reduced to six tissues - namely *Breast*, *Lung*, *Utherus*, *Prostate*, *Thyroid* and *Brain*. This category are also generally coarser then the ones included only for GTEx;

4. When considering GTEx and TCGA together it is necessary to take into consideration that the first experiment was carried out on healthy adults, while the second one obtained sample from cancer-positive individuals. This is indeed a big difference which is thought to interfere with gene expressions as well.

Given the points above, we have also considered taking into consideration the RNA integrity number and the autolysis score. For the first one usually something above 7 might be considered acceptable, with a value above 9 seen the best rule of thumb, while the latter indicates the level of self-destruction of the sample in consideration. Figure 3 shows two boxplots, first considering all the samples available to us and then considering the removal of sample with RNA integrity number lower than 7. It is interesting to see a negative correlation in the first

[1] "StatQuest: A gentle introduction to RNA-seq", `https://statquest.org/`

case (left figure), in which the higher the autolysis score the lower the RNA integrity number becomes. This is natural as they both measure in some degree the integrity of the sample available. When considering samples with only RNA integrity number above 7 we do not observe the same negative correlation, which is assuring from an integrity of samples point of view.

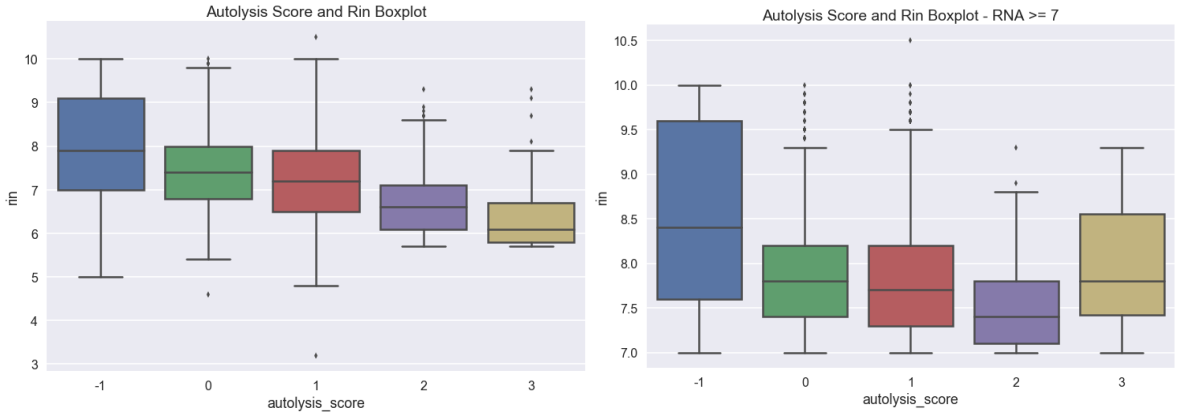

Figure 1: Boxplot for RNA Integrity number against autolysis score when considering all the samples (on the left) and just the samples with RNA bigger or equal to 7 in the second case. Negative correlation is not observable when removing samples with poor integrity numbers.

Hence, considering removing the genes and tissues that are not common and by only having samples with integrity number above 7 we are left with around $6,5$ million gene expression dataset for $GTEx$ experiment. Figure 2 shows the number of samples we have from each of the tissue after performing such process.

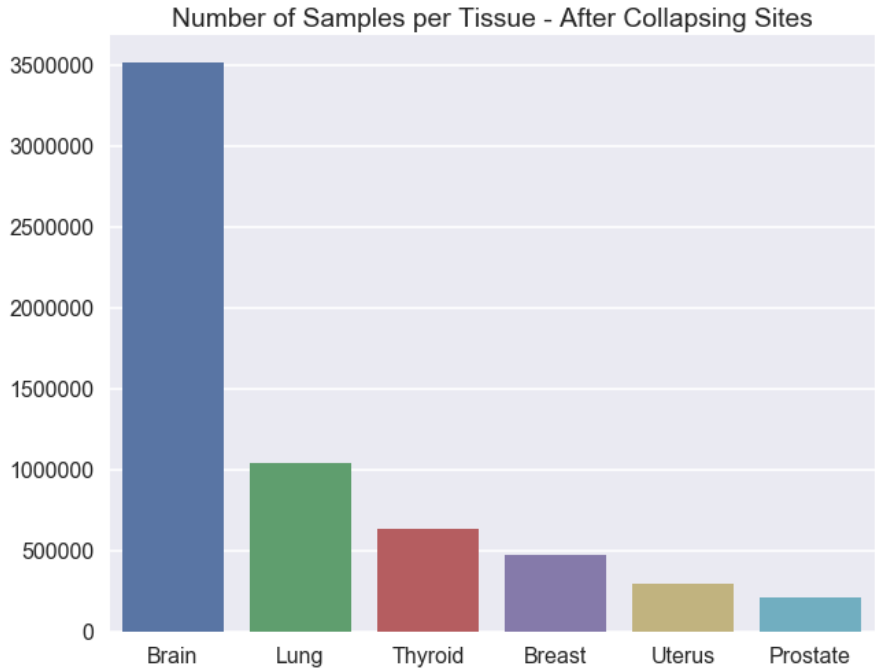

Figure 2: Number of samples in GTEx from each of the tissue after we just consider the tissues that are in common between GTEx and TCGA.

In order to assess wether gene expressions follow certain structure, we proceed to perform qualitative analysis on the gene expressions from GTEx and TCGA in various tissues. Since the expression units in GTEx and TCGA is in RPKM and FPKM respectively they are not

directly comparable across the two datasets. We first reduce each gene in the space of its expression in different tissue cites, for which we use the genes Median, Mean and Standard Deviation of expression. In order to perform qualitative evaluation, to find if the genes follow similar patterns of activity, we perform a t-SNE to embed the featurized gene expression profiles to a lower dimensional manifold, in order to visually estimate inter-gene similarity between healthy individuals, and individuals suffering from cancer.

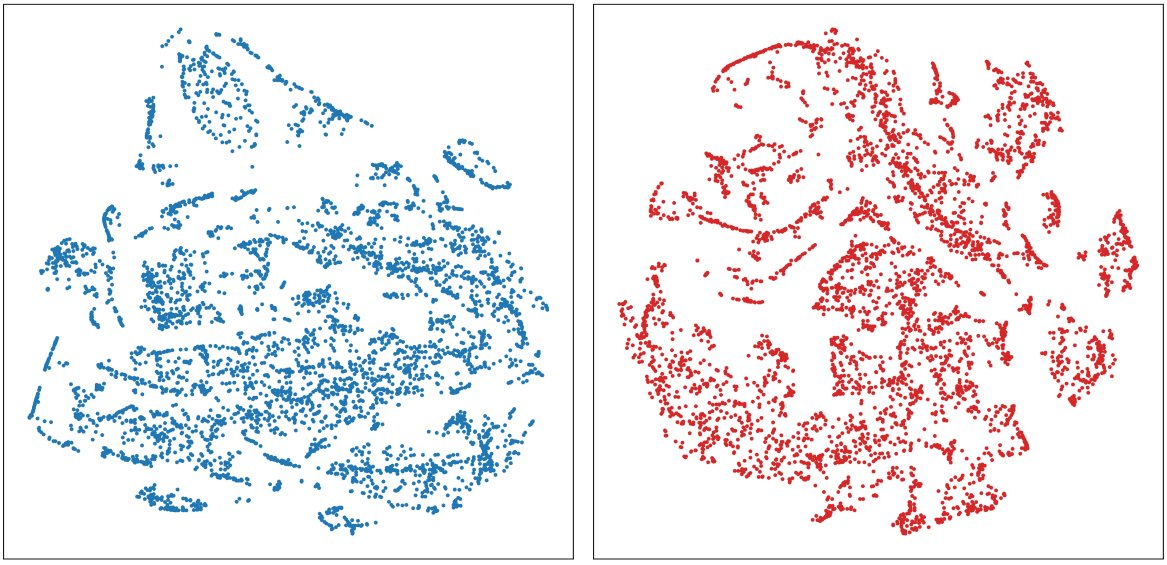

Figure 3: t-SNE Gene Profile for TCGA

4 Modeling

4.1 Clustering and ranking

4.1.1 Clustering

Our data contains expression levels for 4,691 genes, far too many to easily visualize or concisely analyze. In order to reduce this complexity, and in order to look for low-dimensional structure in this high-dimensional space, we clustered the genes using k-means clustering.
For each gene, we calculated the mean, median, and standard deviation of the expression levels, considering samples from each tissue separately. We repeated this calculation for the TCGA data, the GTEx data, and the combined TCGA and GTEx data. In other words, each gene was represented as a vector of 54 features: 3 measures x 6 tissues x 3 combinations of the data (TCGA alone, GTEx alone, and the combination of the two).
Since K-means is sensitive to scaling, we scaled each feature to have mean 0 and variance 1. We then extracted 5, 8, 10, 20, 50, 100, and 1,000 clusters. We selected 8 clusters to use for downstream analysis, both for relative ease of visualization and because other numbers of clusters tended to produce highly lopsided distributions across the clusters.

4.1.2 Ranking

Since the TCGA and GTEx datasets contain distinct, non-translatable expression level measures (fpkm vs. rpkm), they cannot be compared directly. To render these two datasets comparable, we calculated the expression level *rank* for each gene within each sample. We included only the 4,691 genes that the two datasets have in common, so that the ranks would

be comparable. Ties were recorded as the mean rank within a tie group. For example, a vector of expression levels (2, 5, 5, 23) would be translated to ranks (1, 2.5, 2.5, 4).
The cluster membership labels and ranking representations of genes were used downstream in 3 ways: (1) to investigate differences in expression levels between the cancerous (TCGA) and non-cancerous (GTEx) samples, (2) to investigate whether diseases other than cancer show different patterns of association with different clusters of genes, and (3) to build a classifier to predict whether a particular sample comes from cancerous or non-cancerous tissue.

4.2 Gene expression levels in cancerous vs. non-cancerous samples

Using the cluster and rank representations for each gene described above, we computed kernel density estimates of the ranks by tissue type and cluster number (Figures 4-7). (Note that the figures are divided by cluster number and tissue type, for convenience.) We observe striking differences across tissue types and clusters, and between the two datasets. In general, the relative gene expression levels are much more stable across samples for the TCGA data as compared to the GTEx data. The ranks for the GTEx data tend to vary much more across tissues and clusters. Clusters where the ranks for the TCGA samples are much higher than for the GTEx samples, such as for the Prostate samples in cluster 3, may indicate genes that have higher expression levels with cancer.

4.3 Cancer Prediction from Gene Activation Profile

We use the ranked gene expressions from all the samples from GTEx and TGCA and proceed to build a model to predict if a sample is drawn from TGCA or GTEx, or more generally, if a sample is drawn from the healthy population or the individuals suffering from cancer. The use of ranked feature representations allow projecting the activations from GTEx and TGCA to a common space for the application of supervised classification models. As opposed to black-box ensemble methods we leverage models with simpler learnt hypotheses space, since they offer a level of interpretability, to directly identify which gene correspond the the activity of interest. We proceed to train a Logisitc Regression model with an $\ell 2$ penalty on the weights vector in a 5 fold cross validation fashion. Surprisingly, the model worked extremely well, although we are not sure if the performance is reflective of the ease of the classification problem, or due to some idiosyncrasies in the way the data was collected in the two experiments which make the problem easier. We also proceed to use the weights vector coefficient to identify which genes actually contribute to cancer.

Rank	Gene
1	ENSG00000273407
2	ENSG00000273455
3	ENSG00000273473
4	ENSG00000273493
5	ENSGR0000182162
6	ENSGR0000197976
7	ENSGR0000205755
8	ENSGR0000228572
9	ENSGR0000229232

4.4 New Framework to Associate Collection of Genes With Disease

How can we talk about if a collection of genes is associated with certain disease, not about a single gene? For example, each person has slightly different collection of genes. Traditionally, we can use inference score. However, 1) the inference score is not available for every person, and 2) probably the collection of sampled genes from medical examination will vary from time to time. Utilizing the cluster information for each gene we previously defined, for each disease we can identify a fingerprint for each disease in terms of the distribution of inference score in each cluster label, particularly for diseases whose inference score is in general very high across genes. We estimate the distribution of inference score in each cluster via kernel density estimation. Below is the result for top 10 diseases. With this information, if we generate the same distribution fingerprint for someone we can use the above information to make an inference which disease is particularly highly associated with him. From the results, we can guess donor GTEX-N7MS is potentially likely associated with Necrosis.

Shot 2017-09-16 at 3.02.43 PM.png Shot 2017-09-16 at 3.02.43 PM.png
Figure 4: Top 10 diseases with highest average inference score

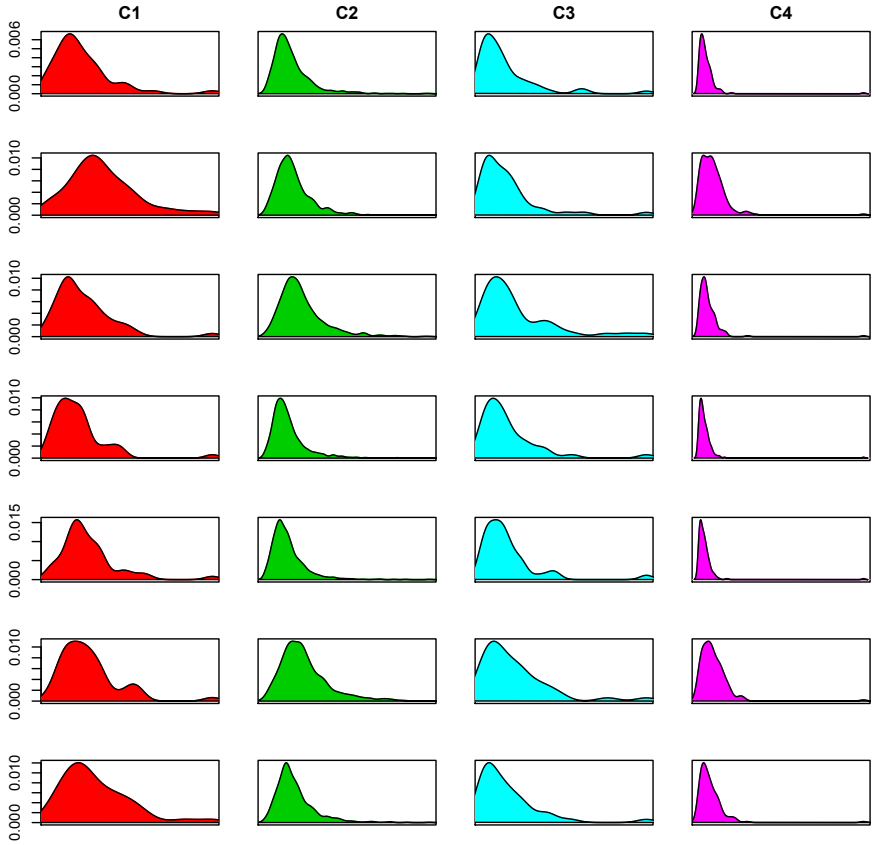

Figure 5: The distribution of inference score in each cluster via kernel density estimation for top 7 diseases with highest average inference score.

distribution of donor GTEX-N7MS.pdf distribution of donor GTEX-N7MS.pdf

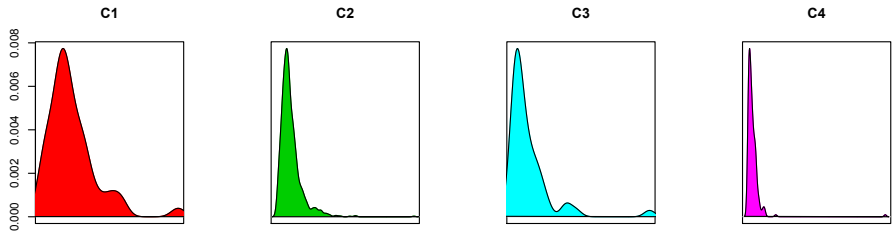

Figure 6: Cluster-distribution fingerprint for donor GTEX-N7MS.

5 Discussion and Further Work

Given the challenge presented by the use of different expression level measures across the two data sets, our clustering and ranking approach was successful in identifying differences in gene expression levels for the TCGA vs. the GTEx data. These differences in general are striking, with the distributions in Figures 7-11 showing very different shapes and locations.

Additionally, we were able to identify a fingerprint for each disease in terms of the distribution of inference scores within each cluster label.

Finally, we find that we are able to use the gene expression ranks to predict whether a sample is cancerous or not, with perfect accuracy. (We checked our representations many times to make sure we hadn't introduced we bug; we believe this classification problem happens to be fairly easy.)

In light of the above, we believe the following steps are necessary to take this work forward:

1. Exploration on binary classification in using gene expression to detect samples coming from a cancer-positive individual. The perfect accuracy achieved in multiclass logistic regression reflects how using different classification methods and classifiers might be needed to further explore the problem;

2. Pattern stability across different number of clusters will need to be assessed, in order to be able to take further conclusions into the association of the change in ranking of certain genes in presence of cancer-positive samples. On top of that, a further investigation on genes with very high ranking in cancer-positive samples might be beneficial to explore their association;

3. Further investigation on different ways of associating fingerprints to diseases from inference score and their respective stability across different cluster numbers.

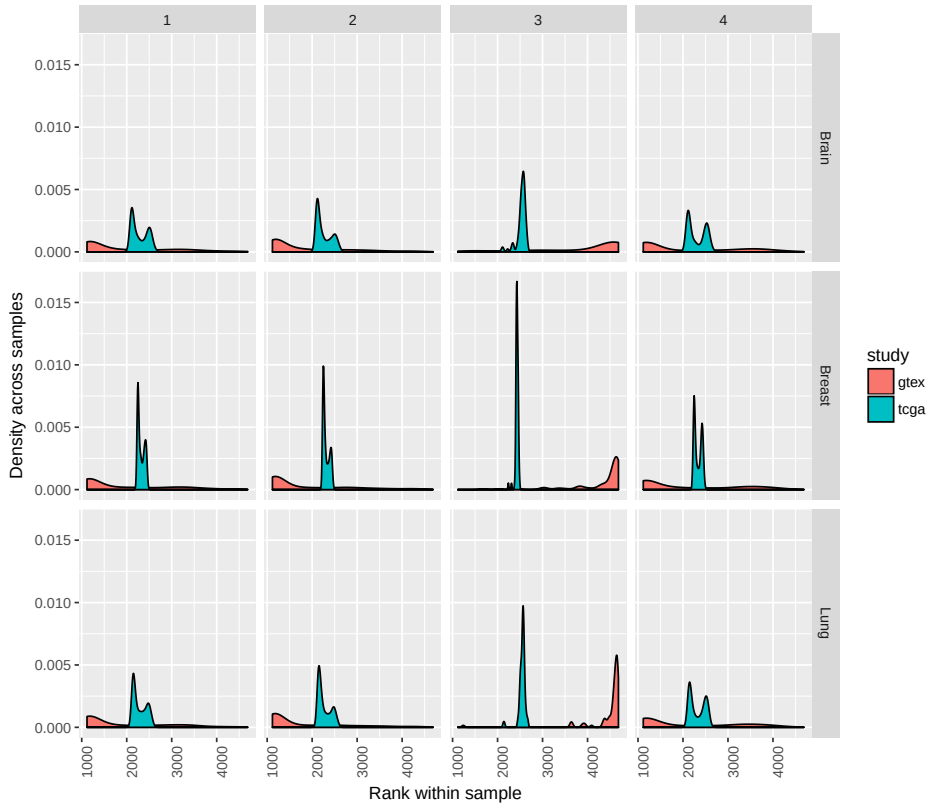

Figure 7: Density of ranks between GTEx and TCGA samples, from clusters 1 to 4 and for Brain, Breast and Lung Tissues.

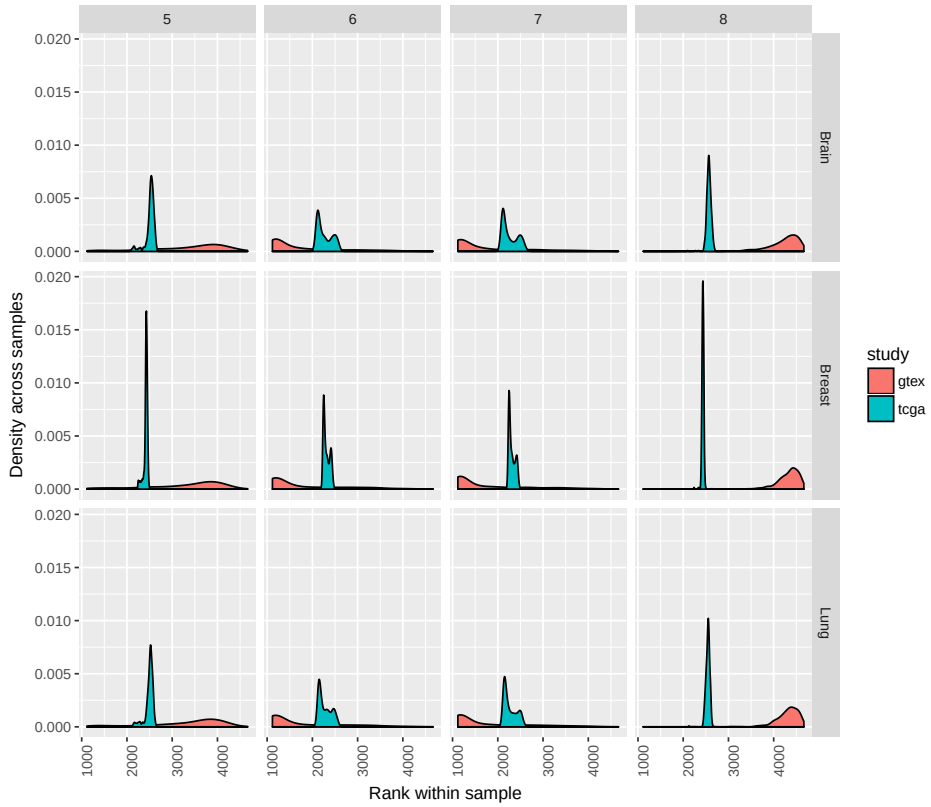

Figure 8: Density of ranks between GTEx and TCGA samples, from clusters 5 to 8 and for Brain, Breast and Lung Tissues.

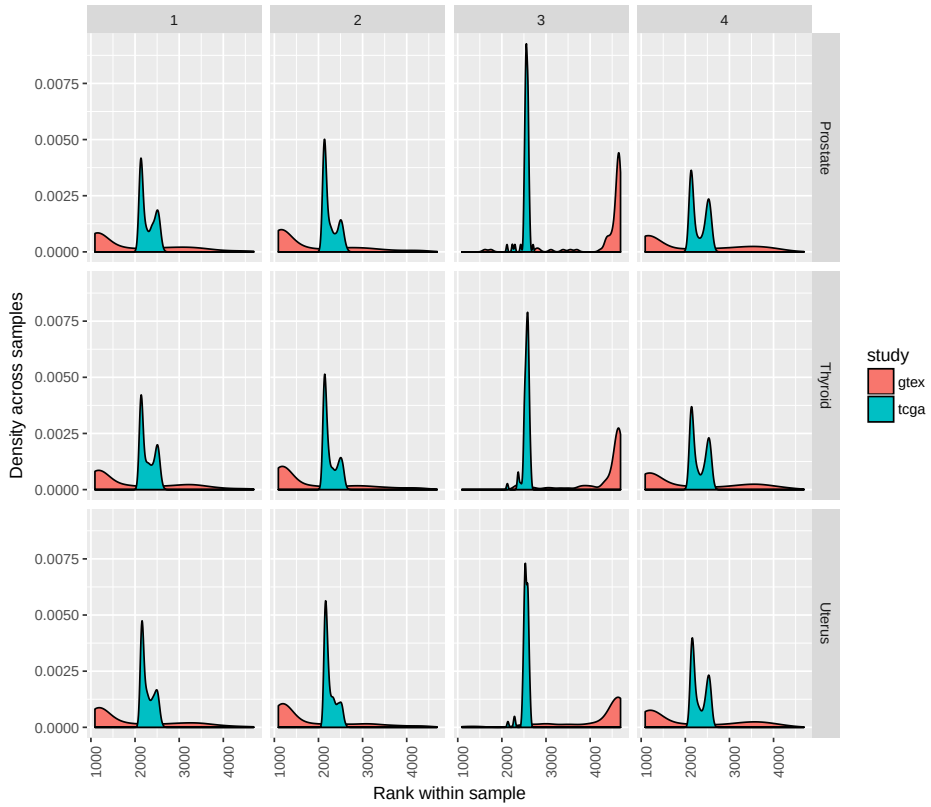

Figure 9: Density of ranks between GTEx and TCGA samples, from clusters 1 to 4 and for Prostate, Thyroid and Uterus.

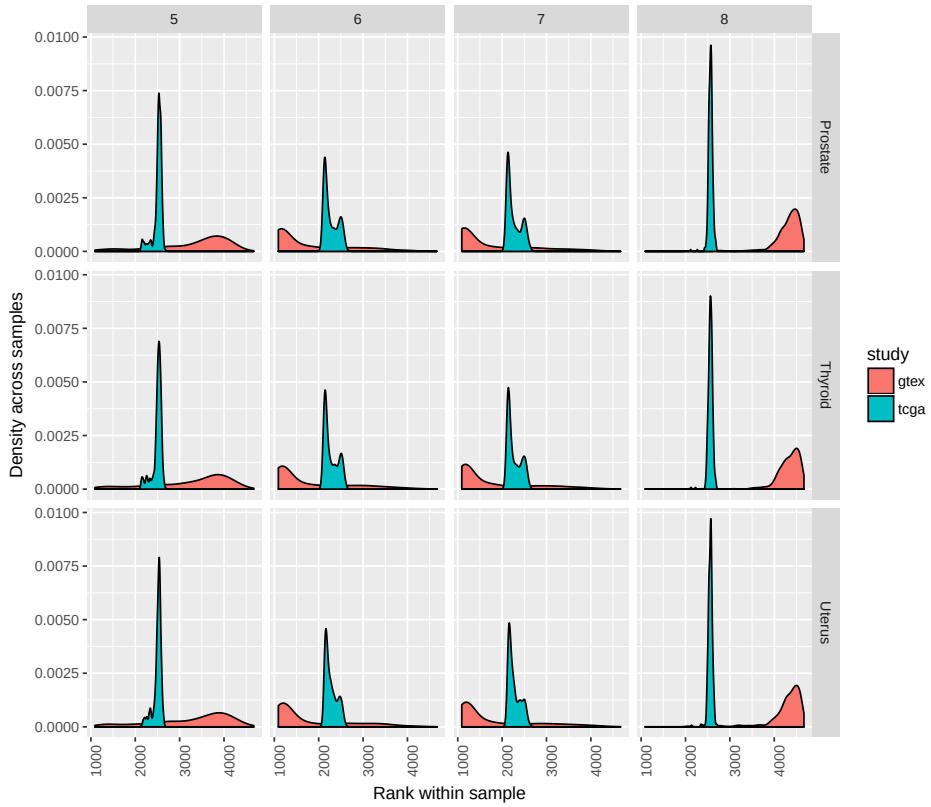

Figure 10: Density of ranks between GTEx and TCGA samples, from clusters 4 to 8 and for Prostate, Thyroid and Uterus.