Gene expression in the Brain

October 12, 2022

1 Team Members: Ben Agyare, Yumeng Wang, Jake Trauger, Yash Patel

We present herein the findings from our explorations on differential gene expression in the brain. In this report, we present some initial genes that could be of interest for further investigation. Such suggestions are further qualified with our corresponding level of confidence, largely based on replication across the statistical analyses we present. The report is, therefore, broken into the following sections:

- Data exploration
- Normalization
- T-test analysis
- GLM analysis
- Permutation test analysis
- Rank-sum analysis
- Conclusion

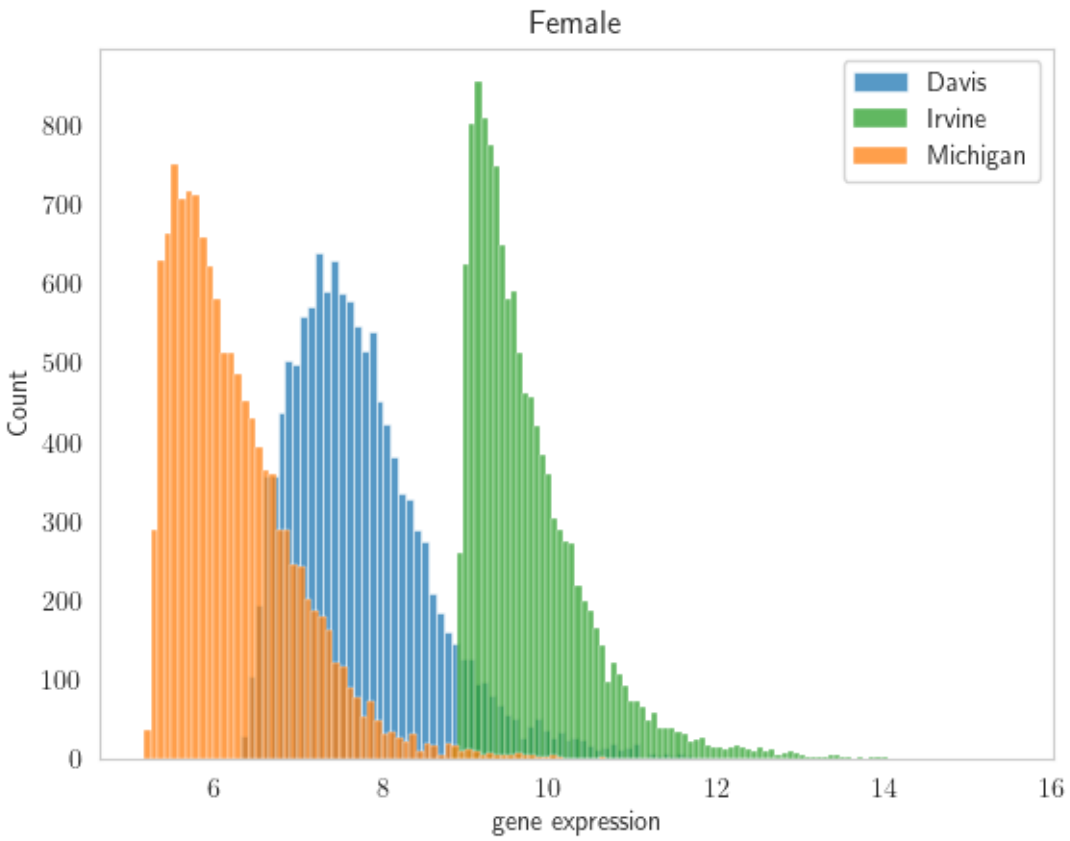
Each section is annotated with the corresponding code and figures in the sections below along with associated exposition.

2 Exploratory analysis

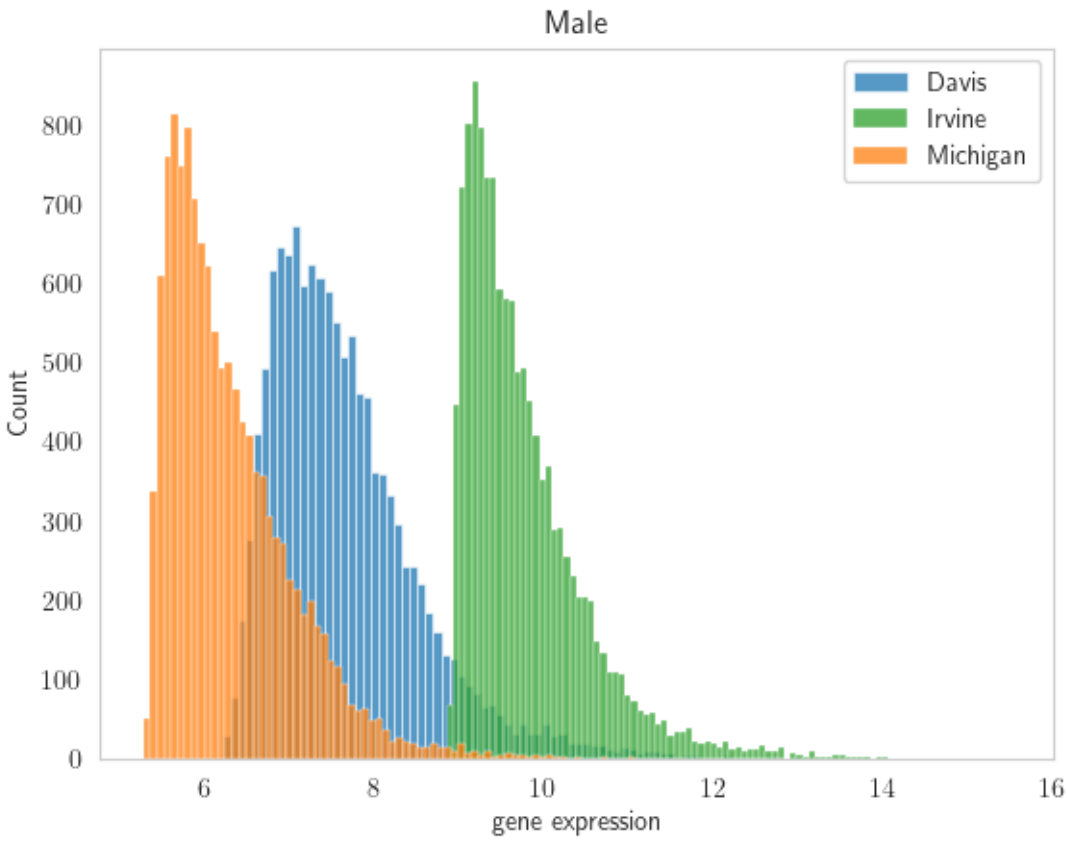
For a good experimental design for gene differential analysis, measurements have to be accurate. Due to uncertainty in measurements, there is the need for replication. The most common types of replications are biological and technical replications. Biological replication involves taking measurements mostly in the same lab and with same technology for several samples of the same cell, while technical replication deals with repeating identical labs and protocols for sequencing on a single sample. While replication fosters confidence in our analysis as it enables us to quantify uncertainty, it posses yet a potential problem in gene differential analysis. For example, technical replication, if exists, could confound the true gene expression differentiation or can cause expression for reasons unrelated to the levels of the expression. Thus, it is highly imperative to investigate the presence of variations arising from replications before conducting further statistical analysis.

The dataset has technical replicates. Measurements were taken from three labs Michigan, Irvine and Davis. We first make examine the distribution of genes by Lab for both sexes (first two figures below) and the look at same for the ACC & DLPFC brain regions by lab (the two figures thereafter).

```
<matplotlib.legend.Legend at 0x179c67040>
```

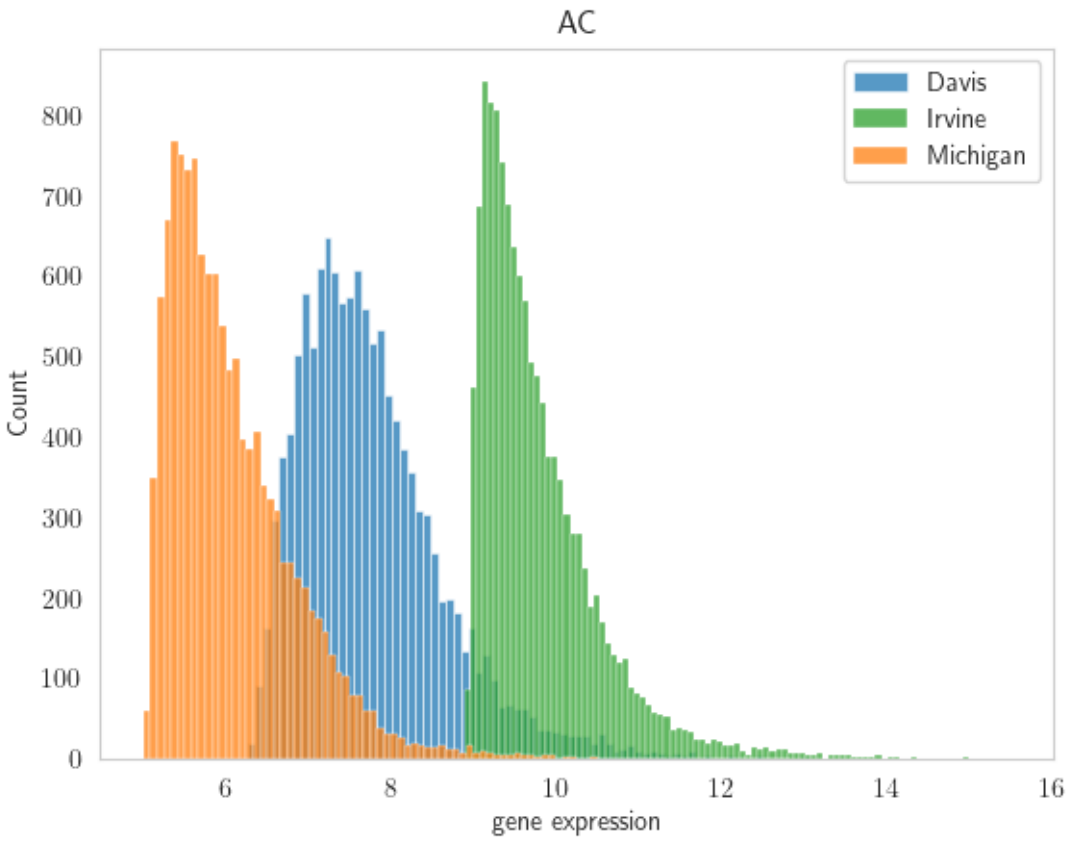



```
<matplotlib.legend.Legend at 0x17884aa60>
```

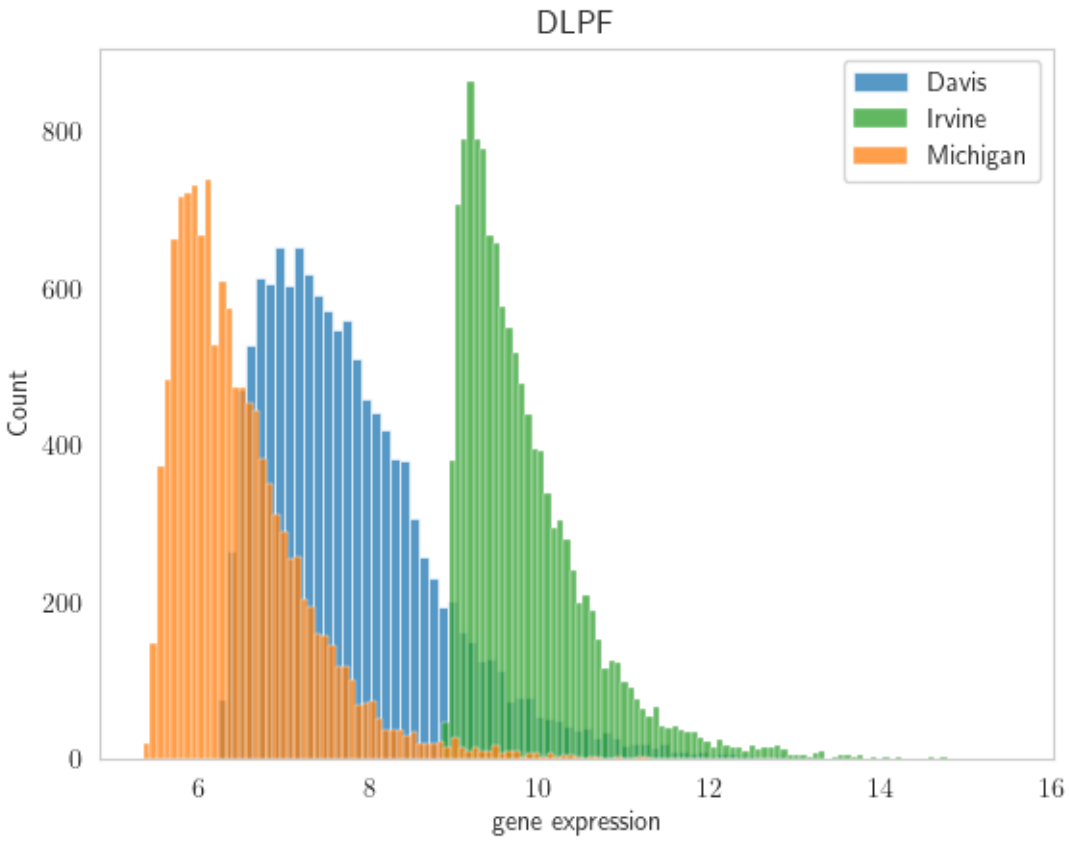


Repeating the above for AC and DLPF:

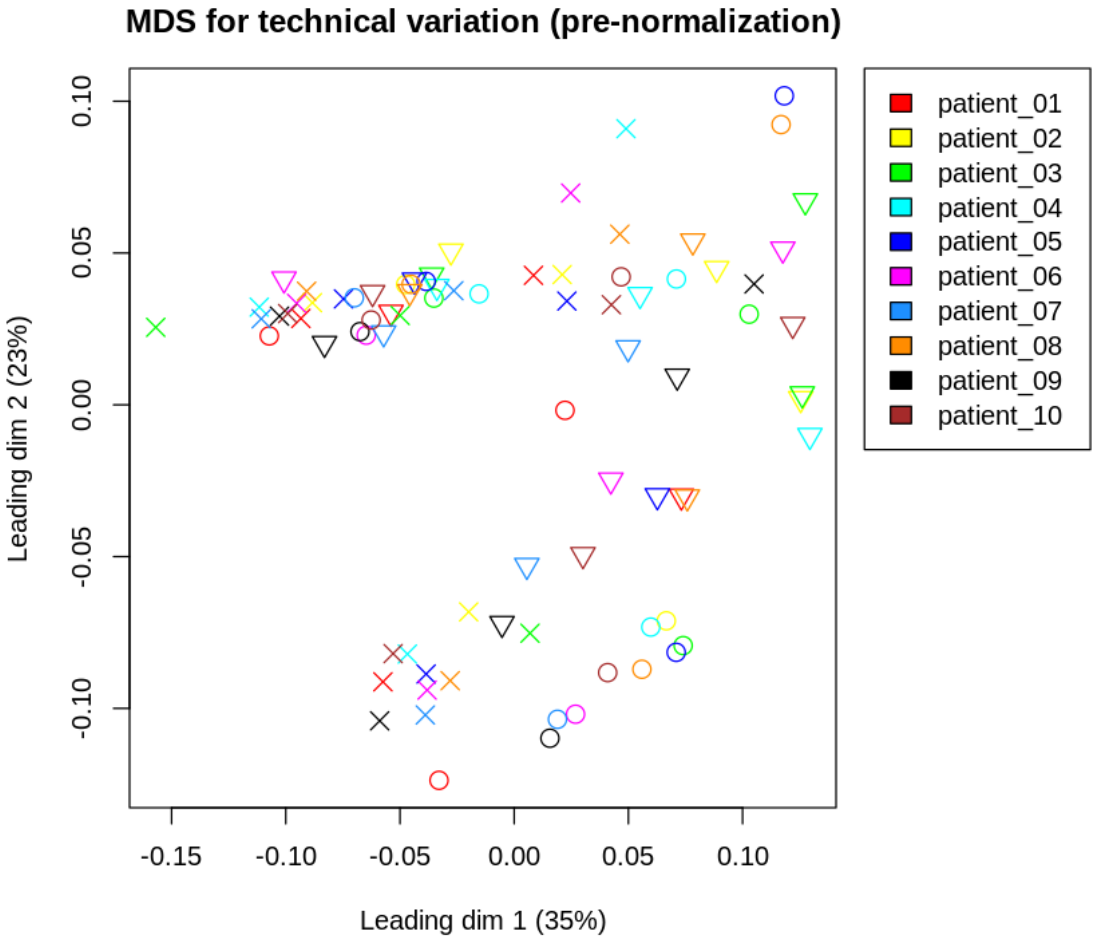
```
<matplotlib.legend.Legend at 0x17d81bb50>
```




```
<matplotlib.legend.Legend at 0x17dc25fd0>
```

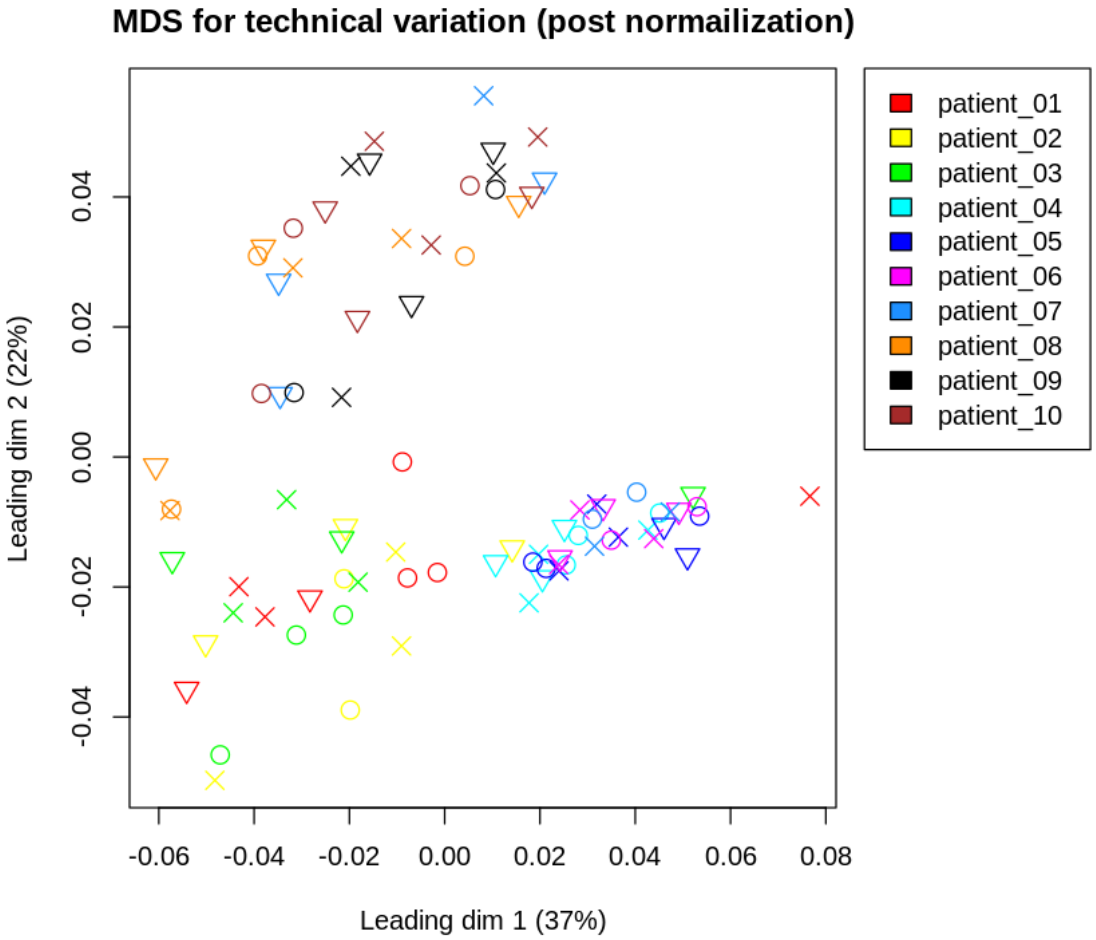


We observe in Fig 1 & 2 that the distributions of the gene expressions vary by lab for both sexes. This is a clear indication of variability due to technical replication. Same story can be told about the distributions of the genes for the ACC & DLPFC brain regions. Further, we make a Multi-Dimensional Scale (MDS) plot to see how the various replicates cluster among each other for the 10 patients.



This above figure show the first two dimensions of the MDS. The first dimension explains 35% while the second explains 23% of the total variation explained by the MDS. Clearly, there is significant sparsity for the replicates of each patient as the various measurements across the three labs are farther apart.

The issue of variation due to technical replications is one that had left researches with serious decisions to make in the past years. Should the scientist throw the data away and start all over? Recent developments have provided a good solution to handle this problem with Normalization. There have been several normalization techniques that scientists have employed in reducing technical variations including the use of housekeeping (control) genes as baseline or the expression level from a particular quantile of the distribution of gene expression values of each sample as well as using variance stabilizing transform from statistical modeling. We employ one of the Normalization techniques which would be discussed in the subsequent section and visualize the MDS plot post normalization in the figure below.



We observe that, normalizating the data has resulted in significant reduction in the variation due to technical replication (comparing Fig 5 and 6). This gives us some confidence in assessing the true gene differentiation as we perform the appropriate statistical analyses in the rest of the sections of this report.

2.1 Exploratory Analysis Takeaways

From these initial findings, we find that it is necessary to perform normalization of the data to allow for comparisons between labs. Doing normalization, however, can take a number of forms, which is the primary difference in the statistical analyses presented next. In the first, we perform explicit normalization using a combination of spike-in normalization and latent "gene expression" extraction followed by t-tests. For GLMs, the nuisance parameters are explicitly added as variables to the model, thereby allowing us to look at the isolated effect of the genes. In the last, we similarly perform "implicit normalization" by just analyzing the ranks of the data.

3 Normalization

We need to do some normalization with the bacterial spike-ins. According to the following two sources:

- https://support.bioconductor.org/p/49150/
- https://bioinformatics.mdanderson.org/MicroarrayCourse/Lectures/ma07b.pdf

Two common normalization techniques are spike-in normalizatio and quantile normalization. The former doesn't really involve much assumption beyond the experimental consistency of spike-in amounts. Quantile normalization assumes that most genes are *not* differentially expressed (that only a handful are). We use the former herein.

3.1 Spike-In Normalization

Given the consistency of the spike in *per lab*, however, it seems plausible to actually go ahead with this spike-in normalization. To do so, we count gene expressed in a "normalized space" based on the total amount of control genes expressed:

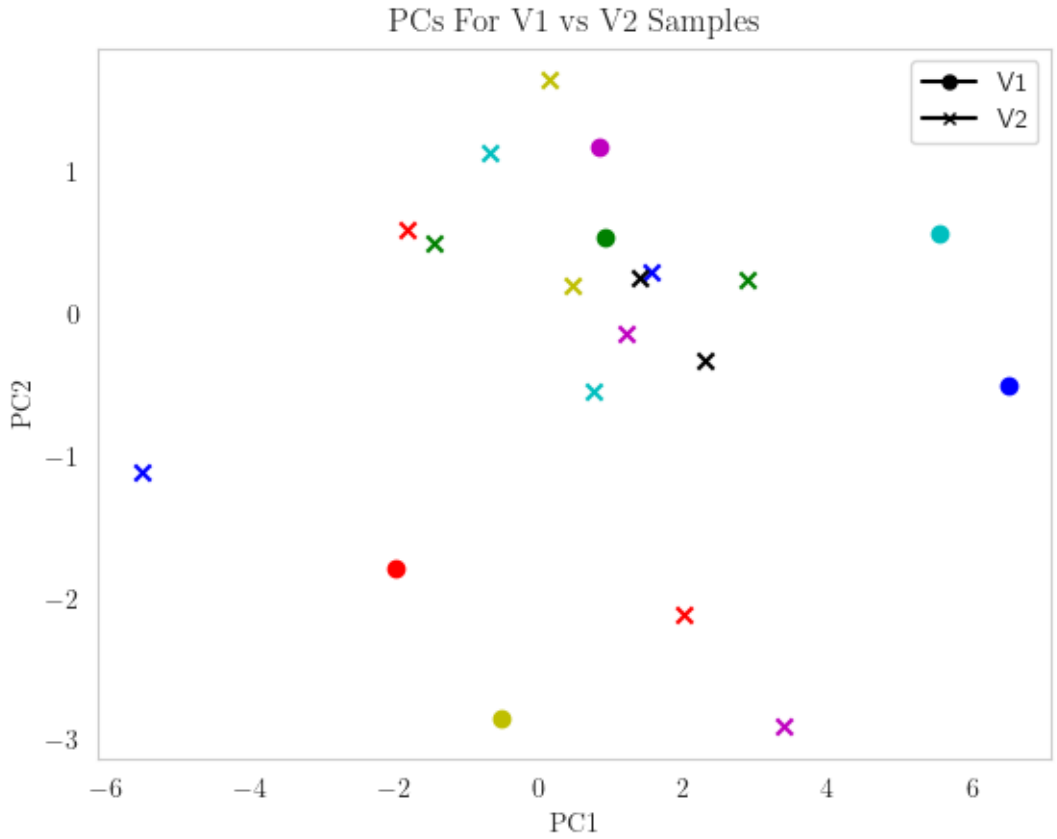
$$\hat{g} = \frac{g}{\frac{1}{|\{c_g\}|}\sum c_g}$$

Where g is some gene of interest and c_g the set of controls.

4 Investigating Combining Versions

We can perform this analysis in two ways: take the labs separately and treat them as separate experiments for the analysis. Or combine them and do the analysis on this combined dataset. The rationale behind the latter is to view the measurements of each labs as noisy representations of the "true" gene expressions. Before doing so, one thing that differed between otherwise seeming "replicates" of samples between labs was the version of the chip that was used. We first wish to investigate if there is any systematic difference between the versions by visualizing paired samples (i.e. samples that are otherwise replicates) after projecting them to a lower dimensional space via PCA.

```
Text(0.5, 1.0, '$\\mathrm{PCs \\ For \\ V1 \\ vs \\ V2 \\ Samples}$')
```



This plot suggests that there are no systematic differences between the two version, which suggests that combining the samples is reasonable.

5 T-test Analysis

We currently have normalized data from which we can analyze the difference in *probe expression* between males and females (and similarly for the brain regions), but recall that the true object of interest for comparison is the *gene expression*. We can currently only do this analysis in an obfuscated light, where the probe sets highlight differences that we then map back to the corresponding genes. It, however, makes sense to eliminate chip-level effects to uncover the latent "gene expression" and only then compare them across the subjects of interest.

Specifically, recalling that each gene has several corresponding probes, for an array i, probe j, and gene k, we can model the measured expression as being:

$$Y_{ijk} = \mu_{ik} + \alpha_{jk} + \epsilon_{ijk}$$

- μ_{ik}: "True" gene expression
- α_{jk}: Probe affinity effect
- ϵ_{ijk}: Noise

We wish then to estimate μ_{ik}. Let's fix a gene k for this discussion, from which we now wish to estimate μ_i. We estimate this using Tukey's Median Polish algorithm, which breaks the above relation into:

$$Y_{ij} = \mu + \delta_i + \alpha_j + \epsilon_{ij}$$

Where:

- $\mu + \delta_i$: "True" gene expression
- α_j: Probe affinity effect
- ϵ_{ij}: Noise

So, after running a Median Polish, we take our estimate to be: $\widehat{\mu_i} = \hat{\mu} + \widehat{\delta_i}$

5.1 T-test [Sex]

We investigate first the differential expression of sex from the gene data extracted above. Such gene expressions are analyzed *per lab*, which we then analyze for replication. That is, from the individual analyses, we see which genes are found repeatedly across the different labs, which gives us greater confidence in such findings. This analysis is performed with a simply independent t-test using a Benjamini-Hochberg procedure to control the FDR to 0.11. Note that we are using a relatively high FDR value because we wish to find all plausible candidates for further investigation.

To provide further validation, we additionally investigate these results if the data are further split by the brain region. That is, we then compare ACC males vs. ACC females, DLPF males vs. DLPF females, and so on. Given the reduced sample size in doing this splitting, the power of this procedure is reduced, but it serves as a means of finding a more conservative set of genes for further exploration.

```
Y -> ['USP9Y', 'KDM5D', 'DDX3Y', 'RPS4Y1', 'UTY']
X -> ['XIST']
KDM5D -> 3
DDX3Y -> 3
RPS4Y1 -> 3
USP9Y -> 2
UTY -> 1
XIST -> 1
============================================

Y -> ['RPS4Y1']
RPS4Y1 -> 1
============================================
Y -> ['RPS4Y1', 'DDX3Y']
RPS4Y1 -> 2
DDX3Y -> 1
============================================
Y -> ['RPS4Y1', 'DDX3Y']
RPS4Y1 -> 1
DDX3Y -> 1
============================================
```

5.2 T-test [Brain Region]

We repeat the procedure above, now instead using a dependent t-test, since each ACC and DLPF sample pair comes from the same patient.

```
FLOT2 -> 1
UBR4 -> 1
GALT -> 1
```

The brain region findings seem to be highly spurious given the tremendous differential expression found between the ACC in the DLPF in the Michigan data but complete lack thereof in other labs' datasets. This is further compounded by the fact we had to reduce the FDR control to 0.0001 to restrict the total findings to a reasonable number. Such a disparity leads us to believe these results are spurious and should *not* be trusted. We instead opt to suggest the results found in the latter two analyses due to their replication.

5.3 T-test Takeaways

Therefore, from the t-test analysis, we find the following genes, along with the corresponding replications. Italicized results also significant in combined lab analysis:

Sex

Gene	Replications
KDM5D	3
DDX3Y	3
RPS4Y1	3
USP9Y	2
XIST	2
UTY	1

Region

Gene	Replications
FLOT2	1
UBR4	1
GALT	1

6 GLM Analysis

Another test we ran was done by using generalized linear models (GLMs). When doing prelimary research on the topic of differential expressions we found an R package called DESeq2. This package has over 27000 citations and it uses GLMs on the raw count data, along with more technical processes to get differentially expressed genes. Therefore, since it seems that GLMs are a well precedented method in the bioinformatics community we decided to try our own GLM model to test the research questions.

As for the specifics, we ran a GLM on the original dataset with a gaussian family and canonical

link. For the formula we use gene_expression ~ [variable we are looking at] + lab + chip.version + constant. The lab and chip version were added to control for any confounding they might have since we are still using the original dataset. Therefore, this method can also be used as a sanity check for making sure our normalization works as desired. If the results from the GLM and t-test analyses concur, then we have more confidence in both our results and techniques.

6.1 GLM [Sex]

For the males and females split, we ran a GLM on each gene and found the p-value for the sex variable. We then ran the Bonferroni family-wise error rate correction with $\alpha = .05$ and got 2 genes that are significant: DDX3Y and RPS4Y1. We choose to use a Bonferroni correction due to its conservative nature. Since GLMs are well precedented and the Bonferroni procedure is a very conserative correction, we can be very confident that the findings we get for our GLMs are worth looking into.

```
DDX3Y
RPS4Y1
```

6.2 GLM [Brain Region]

For the region split, we ran a GLM on each gene and found the p-value for the region variable. We then ran the Bonferroni family-wise error rate correction with $\alpha = .05$ and got 4 genes that are significant: CABP1, CARTPT, SCN1B, COX7A1.

```
CABP1
CARTPT
SCN1B
COX7A1
```

7 Permutation Test Analysis

We also ran a Monte-Carlo permutation test on the difference of means on our research questions using the normalized data. Given a gene, if under the null hypothesis males and females (or A.C.C. vs D.L.P.F.C.) have no differences, then if we aggregate the male and female data points and go through all possible permutations of male/female (A.C.C./D.L.P.F.C.) assignment, then we can see how unlikely our found difference in means was. However, since we have many genes and many samples for each gene, going through all possible permutations is computationally expensive. Thus, we only run 1000 random permutations and get our p-value from these permutations. It is known that the standard deviation of this is $\sqrt{\frac{\hat{p}(1-\hat{p})}{1000}} \leq \sqrt{\frac{.5 \cdot .5}{1000}} \approx .0158$, which we deemed acceptable due to the leniency we have been given to find differentably expressed genes.

7.1 Permutation Test [Sex]

When running our Monte-Carlo permutation tests for male vs. female, using the Benjamini-Hochberg false discovery rate correction with $\alpha = .1$, we got the following 5 differentiably expressed genes: UTY, USP9Y, KDM5D, DDX3Y, RPS4Y1. Note that executing the corresponding cell below may yield slightly different results, since this permutation test is a Monte Carlo algorithm.

7.2 Permutation Test [Brain Region]

When running our Monte-Carlo permutation tests for A.C. cortex vs. D.L.P.F. cortex, using the Benjamini-Hochberg false discovery rate correction with $\alpha = .1$, we got the following 10 differentiably expressed genes: ZNF609, DAPK3, CARTPT, AP1S1, ZYX, NDUFS8, CAMTA1, CCDC106, SLC9A3R2, RHBDD3. While these only share 1 common gene with the GLM model, we will note that the others declared significant in the GLM model were still among the genes with the smallest $\hat{p}$ with the largest of them having a $\hat{p}$ of .015.

Note once again that executing the corresponding cell below may yield slightly different results, since this permutation test is a Monte Carlo algorithm.

8 Wilcoxon Rank Sum Test

We consider a nonparametric statistical test method, the Wilcoxon rank sum test, which compares the population median of two samples. The null hypothesis is that the population distributions of female and male are the same. Next, we run the test in three different labs separately, and also run the test in combined data.

8.1 Rank-Sum Test [Sex]

We start with running the analysis separately for the labs, from which we find the following genes: DDX3Y and RPS4Y1

Repeating the above in the combined data gives the same findings, namely DDX3Y and RPS4Y1

```
DDX3Y
RPS4Y1
```

If we use the Bonferroni procedure with $\alpha = 0.1$, no gene stands out. The reason might be that the smallest p-value is not that significant (~0.001) over 12600 comparisons.

8.2 Rank-Sum Test Takeaways [Sex]

We find that the genes with less than 0.05 p-value in three labs simultaneously are the same as that in the combined data. These genes selected by the Wilcoxon rank sum test are consistent with that of the GLM method and t-test.

8.3 Rank-Sum Test [Brain Region]

Repeating the above for the brain region, we once again perform separate and combined analyses. The separate analysis finds no genes of significance.

Repeating the above in the combined data gives the findings that are consistent with the GLM findings, namely CABP1, CARTPT, SCN1B, and COX7A1. This, however, is prior to Bonferroni correction.

```
CABP1
CARTPT
SCN1B
COX7A1
```

If we use the Bonferroni procedure with $\alpha = 0.1$ similar to what we did for female and male data, no gene stands out.

8.4 Rank-Sum Test Takeaways [Brain Region]

There is no common gene that stands out in three different labs at the same time. For the test in combined data, we observe four genes with p-value lower than 0.1. These four genes are consistent with what we find in the GLM method.

9 Conclusion

From the above analysis, we have the following findings for genes we recommend you explore for understanding differential expression by sex followed by those for brain regions. Along with these suggestions, we give our confidences, which arise from the replication across the previously presented analyses. In the presented table, we abbreviate tests as follows:

- **T**: t-test
- **G**: GLM
- **P**: permutation test
- **R**: rank-sum test

For all but the rank-sum test, we list the analysis in the below table if the corresponding gene was found to be significant (post-FDR control) with that test. Since the rank-sum test did not have any significant findings, we simply took the top 5 most significant findings.

Sex

Gene	Confidence	Tests
KDM5D	High	TP
DDX3Y	High	TGRP
RPS4Y1	High	TGRP
USP9Y	Medium	TP
UTY	Medium	TP
XIST	Medium	T

Region We repeat this for brain regions, but again emphasize that our confidence in the t-test analysis is far less in this instance given the great disparity observed between labs.

Gene	Confidence	Tests
CABP1	Medium	GR
CARTPT	Medium	GR
SCN1B	Medium	GR
COX7A1	Medium	GR