DSC 180B

Differential Gene Expression Analysis of Human Opioid Abusers

Dennis Wu, Zhaoyi Guo, Cathleen Peña

HYDROCODONE
TAKE 1-2 TABLETS BY MOUTH EVERY 4-6 HOURS AS NEEDED FOR PAIN

Source: Photo by ©iStock.com/smartstock

Introduction

Opioids are now one of the most common causes of accidental death in the US. According to the CDC, two out of three drug overdose deaths in 2018 involved an opioid [11], so opioid abuse can not only affect people physically and mentally but can also deprive their lives [1]. Opioid addiction has a unique background in that a large reason for why people become addicted is that patients in hospitals are often prescribed opioids to treat pain, however these patients wind up misusing their prescriptions and become addicted. [12]

The growing severity of opioid abuse impacted us as we learned more about it, and because of this, our group wants to address the burden of drug abuse and provide new therapeutic targets for human substance abuse. We'd like to explore the effects of opioids (including codeine, fentanyl, heroin, hydrocodone, methadone, morphine, and oxycodone) and see how gene expression differs in those who abuse opioids. We do this by identifying specific gene expression differences between the control group and opioid abusers on postmortem ventral midbrain. To do so, we first obtained our data from the NCBI Sequence Read Archive [9] under the accession number: PRJNA492904. The dataset contains 29 of opioid abusers and 20 of control (total of 49 individuals). After obtaining the dataset, we converted the format of the data from an SRA dump to a FASTQ format, which is an accessible raw sequencing file format [3]. With this FASTQ format, we processed the data using FastQC [4] for data quality control . Next, we applied cutadapt [5] to trim the adapter sequences from the reads . After the RNA sequences were trimmed, we used kallisto [6] to align our sequences and get gene expression counts. With the gene expression count, we filtered the data for quality reads, combined the gene counts together, and used that data as input to perform DESeq2 [7] analysis. In addition to DESeq2, we utilized other methods such as the weighted correlation network analysis (WGCNA) [8], which can be used to build a co-expression network in an attempt to discover modules of highly correlated genes related with opioid abuse [1].

Ideally, we want to present genome-wide changes in midbrain gene expression associated with human opioids abuse. This way, we could find the midbrain gene expression difference between opioid users and control groups. Also, we want to identify drug-responsive modules associated with responses with opioids abuse. Hopefully, the opioid-regulated genes identified in this project might provide new therapeutic targets and implicate important biomarkers for human substance abuse [1]. The project output will be a report/paper with the findings and figures we've made from our analysis.

Methods

Quality Control and Data Cleaning

Fast-qc

To start off, we ran ten of our raw data files through Fastqc[4]. The raw data is made of base pair nucleotides (ATCGU). The inputs for Fastqc are a sample of the data that we'd like to get measurements of and an output path to direct where to put the files generated by running Fastqc. The html file visualizes statistics which give us information on the quality of our reads. Those statistics include information such as the total number of sequences processed and the length of the shortest and longest sequence in the set. Fastqc also reports information on the per base sequence quality, per sequence quality scores, per base sequence content, per base GC content, per sequence GC content, per base N content, sequence length distribution, sequence duplication levels, and overrepresented sequences. We used the generated "summary.txt" file to give us an indication of the overall quality of the sample and whether or not we should keep it to incorporate into analysis. The "summary.txt" file gives a general 'pass' or 'fail' for each file which lets us know if it is normal (meaning 'random and diverse').

Cutadapt

The main reason for running Fastqc on our samples is to determine whether or not we would have to perform adapter trimming on our data. If there is adapter content, then we would remove the adapter sequences contained in the data with a library such as cutadapt[5]. Adapter sequences tell us where the gene is about to start and points to where it would start transcription of that gene. Adapter sequences are only needed within the cell to point to where to start transcription for genes. Therefore, it can be removed since it does not tell us any genetic information and is only a flag for where the sequencing read from the sample starts.

Kallisto

Next we used Kallisto[6] to perform pseudo-alignment on our samples. Kallisto aligns the reads against a reference transcriptome and counts how many times those reads appear, giving us those counts for each gene. The input for Kallisto includes a reference index, the number of bootstraps we'd like to perform and the sample. We first had to create the kallisto index, using the reference transcriptome GRCh38 [10]. Once the index is built, it is used to align reads (similar to using a reference genome to compare reads against) and provide gene counts for all our samples. We decided to set the number of bootstraps to the default one hundred as is standard and lastly, we used the --pseudobam flag to get bam files output from Kallisto as well as the following. When running all the data through, we use both read files (the .1 and .2 files since it is paired-end reads) as input and for each sample we get an output directory that gives us three files: run_info.json, abundance.tsv, and abundance.h5. The run_info.json file gives us details such as the quantification, number of bootstraps, and program version. The abundance.tsv gives the results of the quantification (gene counts). However, the main output file of importance is the pseudoalignment.bam file which gives us the alignment sequence information. The h5 files give the quantifications with the bootstraps. Once we finished running Kallisto on all the data and received all the output files, we performed some data cleaning and filtering.

Samtools

We first used Samtools[15] to drop PCR duplicates from the bam files output from kallisto, so that amplified gene counts wouldn't be overrepresented and skew our results. Next, we sorted the bam files based on gene name so that the bam files would be ready to be used in the HT-seq step of our analysis. Lastly, we removed multi-mapped reads with value q=10.

GTF File Filtering and HT-Seq

HTSeq[14] is used to assemble the read counts of the genes. HTSeq requires a gene annotation file and the aligned bam file. We used the Ensembl gene-level annotations from Gencode release 24 (GRCh38.p5)[13] as the annotation file. Before continuing with HTSeq, we filtered the annotation GTF file for only "gene" type, and also removed rows that were from chromosome M (mitochondrial DNA). Lastly, we removed genes that had less than 1 reads. We then proceeded with acquiring the read counts per gene for each bam file using HTSeq and merged all the files together at the end to create a holistic gene count csv to be used in the next differential gene expression analysis.

Differential Gene Expression Analysis

DESeq2

DESeq2[7] requires a gene counts matrix of all the samples, the SRA run table which links the samples to other values such as the patients age, sex, race, brain pH, RIN, cause of death, whether or not they used cocaine, and what drugs were in their system when experimental group overdosed, and lastly a design parameter which specifies to return the fold change of the results from whether or not the subject used opioids. We made a DESeqDataSet object, using brainpH, RIN, and age as covariates and then filter to keep only the counts that are greater than 10. We then use that output to create a volcano plot. The volcano plot shows us how many genes are up-regulated and down-regulated in comparison to the control which can give us an idea of to what extent the expression of genes differ if you use opioids. We also conducted principal component analysis (PCA) in order to get an idea of how much the experimental group varies in comparison to the control as well. We use the same DESeqDataSet object to this as well.

WGCNA

Taking advantage of the R WGCNA [8] package, we are able to divide the genes based on the topological overlap matrix dissimilarity. Topological overlapping matrix of dissimilarity is calculated based on the adjacency matrix between the genes. It’s an n by n matrix signifying the weighted correlation between the genes. Basically this matrix measures how similar different genes are after the principal component analysis conducted in DESeq2. Therefore, in each color module, similar genes are grouped together based on their quantification and are represented by different colors in the bottom of the graph. Using this clustering, we are able to show the correlation between the color modules and the up- or down-regulation with opioid abuse.

Results

Exploratory Data Analysis

SRA Run Table

First and foremost, we did an exploratory data analysis on our SRA run table that is associated with the samples, so that we can get to know the general trends and information that our data contains. First we looked at the differences in means for the quantitative values in our data for the experimental group versus the control. In doing this, we compared the values for age, brain pH, and RIN by using a box plot. We also looked at the difference in means for just the experimental group to see if cocaine use was a confounding variable. Below in Figure 1, you can see some of our findings, and supplementary figures in the Appendix.

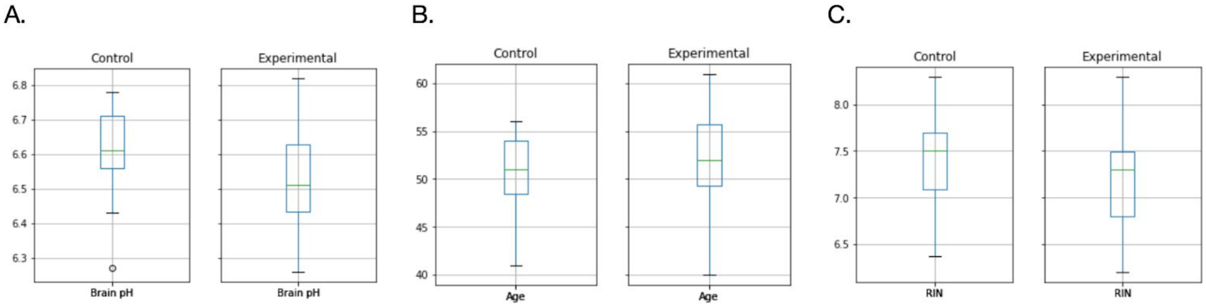


Figure 1. A) Boxplot of Brain pH values among experimental group and control group. B) Boxplot of Age values among experimental group and control group. C) Boxplot of RIN values among experimental group and control group. Exploratory Data Analysis on SRA table data, shows that there is not a significant difference in values for age, brain pH, or RIN between the control group and experimental group.

Raw Data

Next, after running a subset of our data through Fast-qc, and evaluating the pass/fail values we received, we found that there was no trace of adapter content, example shown in Figure 2 below. This means that we don't have to use cutadapt to trim the adapter sequences in the data in order to be ready for the alignment process with kallisto.

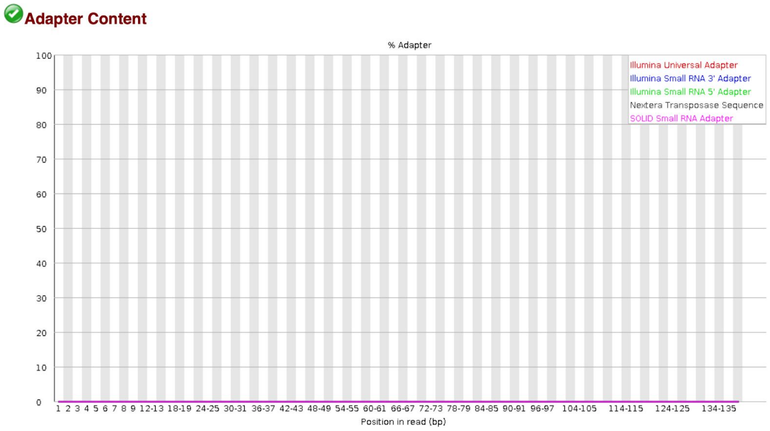


Figure 2. Fastqc results in adapter content of one of the files that were run in order to assess the quality of our reads. This visualization shows that the sample being analyzed did not contain any adapter content. This means that it would be unnecessary to perform adapter trimming on this sample in order to prepare the data for alignment.

Kallisto Counts

Once alignment was finished and we merged the counts that kallisto output for each sample, into one csv, we performed a simple eda on the values in order to check the validity of our data. We performed a t-test on the data for the experimental group versus the control group, then calculated the log-fold change. We also wanted to see the relative abundance of specific types of genes after running alignment on our samples. We did this by taking the first two letters of the genes name representation and finding out what they stood for. We then made the pie chart out of the counts of those values. (Figure 3)

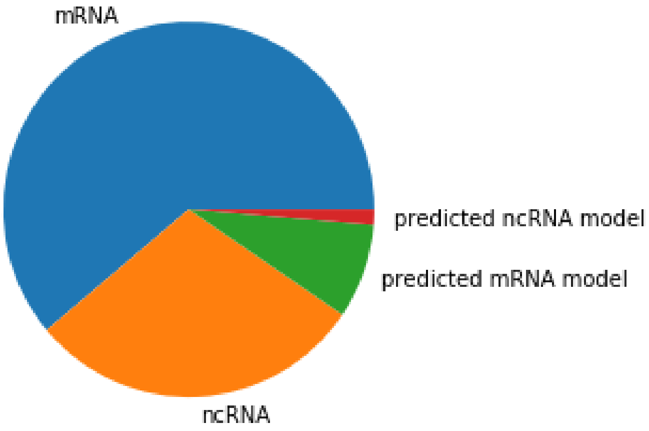


Figure 3. Different types of genes observed overall after aligning samples with kallisto.

Differential Gene Expression Analysis

DESeq-2

We created a volcano plot, using Enhanced Volcano [16], to help visualize the data from DESeq-2. The volcano plot can show us relatively how many genes are either upregulated or downregulated, with the y-axis telling us how significant those differences are (negative log 10 p-value). We have 16 up-regulated genes and 28 down regulated genes. (Figure 4)

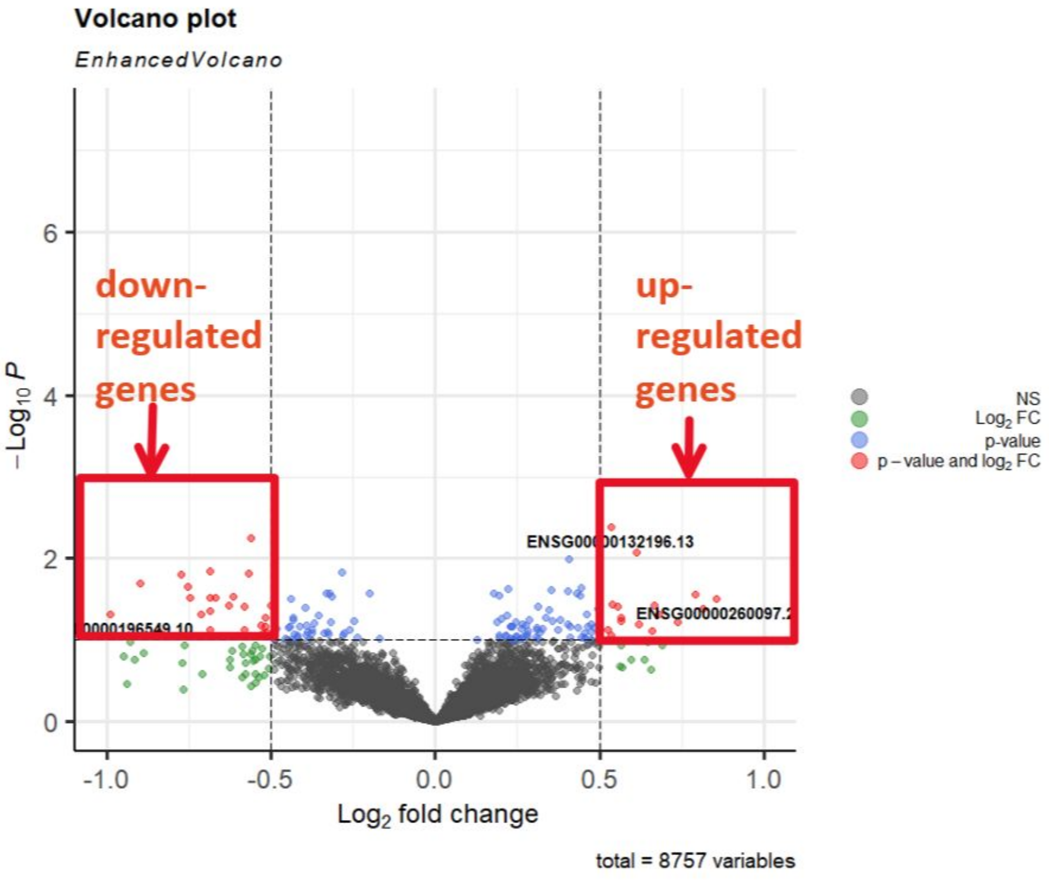


Figure 4. Volcano plot of genes reveals upregulation and downregulation of genes.

Next, we performed PCA in order to see the heterogeneity of subjects between the control and experimental groups. The first principal component explained 36% of the variation and the second explained 20% of the variation. From this, we can see that patients who abused drugs have a larger variance compared to those who did not (control patients). (Figure 5)

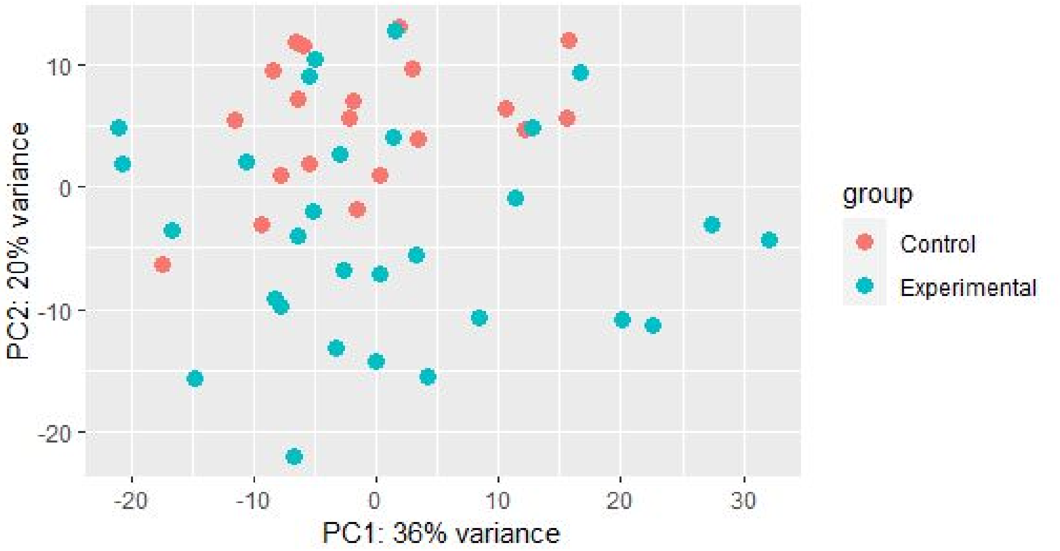


Figure 5. Results of principal component analysis (PCA) done on patients. Principal component 1 on the x-axis and principal component 2 on the y-axis.

WGCNA

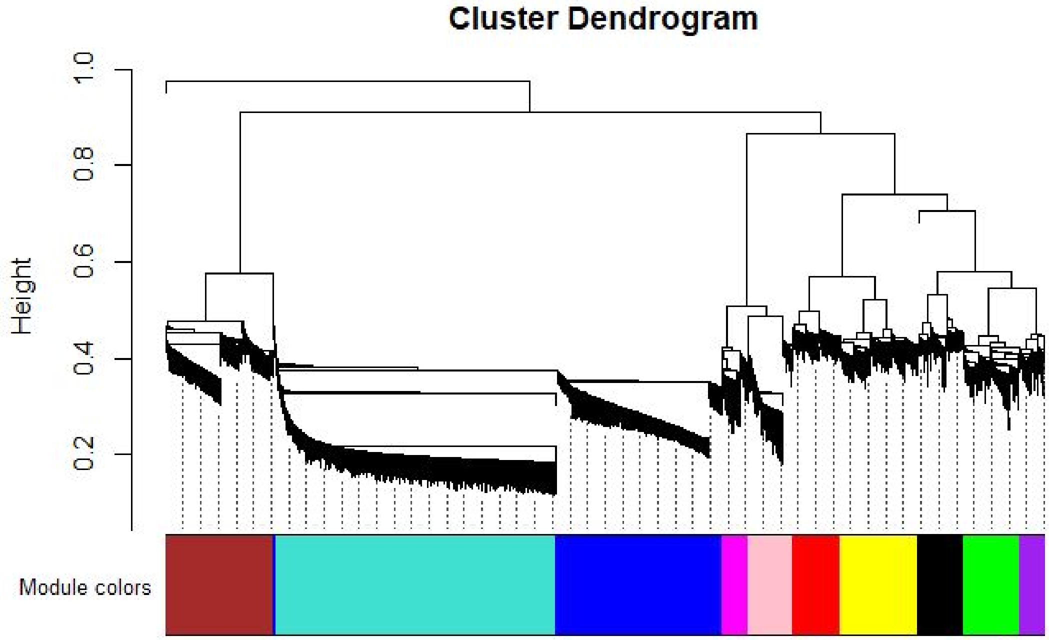


Figure 6. Results of Weighted Gene Correlation Analysis (WGCNA) cluster dendrogram done on expression of genes in the experimental group versus expression of genes in the control group.

Discussion

We believe that there may have been some preprocessing steps that did not run as we expected. Our lack of knowledge in this particular field left us a bit unsure of what else we could possibly do to better clean our data and we feel that this may have skewed the results of our analysis. We had some trouble downsizing our data after getting the counts of HT-Seq, even after filtering for only “gene” type and excluding those that were on chromosome M.

When it comes to the dendrogram result of WGCNA we believe that it was also affected by the variance of genes we decided to use since we restricted it to only protein coding genes. Without that variability, we believe that is why we were only able to see 10 groups of genes output by WGCNA.

The data used in this experiment was very limited. It focused on all males and had no female representation. Oftentimes results can be much different for females and gene expression differences may appear to be different. Therefore, we should consider that this analysis should not be seen as widely applicable to all opioid users. Furthermore, the patients that the samples were collected from were predominantly black, making up 72% of the population. Here arises another issue in that different races and regions around the world can respond much differently than others, therefore, perhaps doing race-specific among multiple races might be a better idea. Another future problem we could look into is seeing whether or not age makes a difference in gene expression. By using a wider range of age in patients, we could analyze whether or not age could possibly be a confounding variable in gene expression.

Conclusion

In our study, we found 28 significantly down-regulated genes and 16 significantly up-regulated genes. We also identified specific gene networks which grouped genes similar in expression together. This gave us insight into what kinds of genes are most affected by opioid drug abuse. We also learned that people who use opioids have greater gene expression variability than those who do not use them. Finding specific genes such as MME and SPDYE6 that were most downregulated and upregulated, can help us think about new genes to use as therapeutic targets to fight opioid addiction.

References

1. Saad, Manal H. *Differentially Expressed Gene Networks, Biomarkers, Long Noncoding RNAs, and Shared Responses with Cocaine Identified in the Midbrains of Human Opioid Abusers*, 2019, www.nature.com/articles/s41598-018-38209-8.pdf?proof=t.

2. "Products - Data Briefs - Number 384 - October 2020." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 7 Oct. 2020, www.cdc.gov/nchs/products/databriefs/db384.htm.

3. Staff, Sequence Read Archive Submissions. "Using the SRA Toolkit to Convert .Sra Files into Other Formats." *SRA Knowledge Base [Internet].*, U.S. National Library of Medicine, 1 Jan. 1970, www.ncbi.nlm.nih.gov/books/NBK158900/.

4. *Babraham Bioinformatics - FastQC A Quality Control Tool for High Throughput Sequence Data*, www.bioinformatics.babraham.ac.uk/projects/fastqc/.

5. Martin, Marcel. "Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads." *EMBnet.journal*.

6. Lab, Pachter. "About." *Sitewide ATOM*, pachterlab.github.io/kallisto/about.

7. Lönnstedt, T. Speed, et al. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology*, BioMed Central, 1 Jan. 1970, genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8.

8. Fisher, RA., et al. "WGCNA: an R Package for Weighted Correlation Network Analysis." *BMC Bioinformatics*, BioMed Central, 1 Jan. 1970, bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-559.

9. "Home - SRA - NCBI." *National Center for Biotechnology Information*, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/sra.

10. "GRCh38.p13 - Genome - Assembly - NCBI." *National Center for Biotechnology Information*, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39.

11. Centers for Disease Control and Prevention. *Data Overview*. 7 Dec. 2020, www.cdc.gov/drugoverdose/data/index.html.

12. Centers for Disease Control and Prevention. "Prescription Opioid Data." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 12 Mar. 2020, www.cdc.gov/drugoverdose/data/prescribing.html.

13. Human Release 24." GENCODE, www.gencodegenes.org/human/release_24.html.

14. Anders, S., Pyl, P. T. & Huber, W. HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics 31(2), 166–9 (2015).

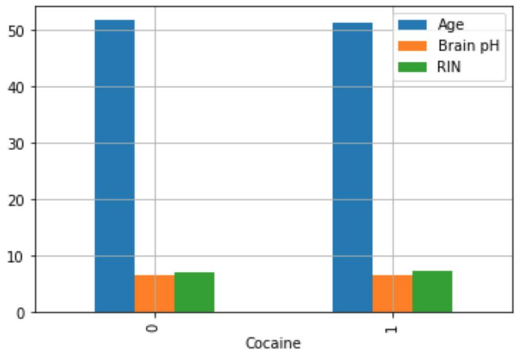
15. Li, H. et al. The sequence alignment/map format and SAMtools. Bioinformatics 25(16), 2078–9 (2009).

16. Blighe K, Rana S, Lewis M (2020). EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling. R package version 1.8.0, https://github.com/kevinblighe/EnhancedVolcano.

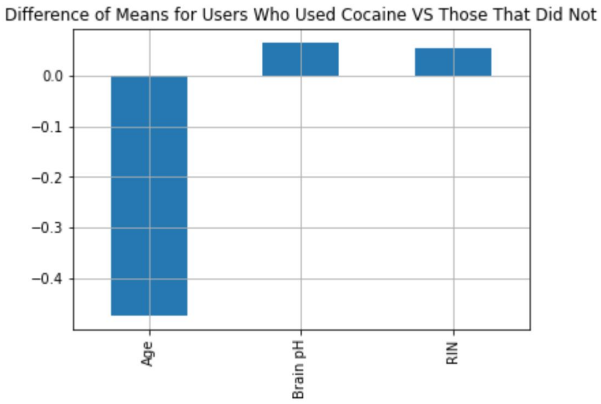
Appendix

SRA table exploratory data analysis visualizations

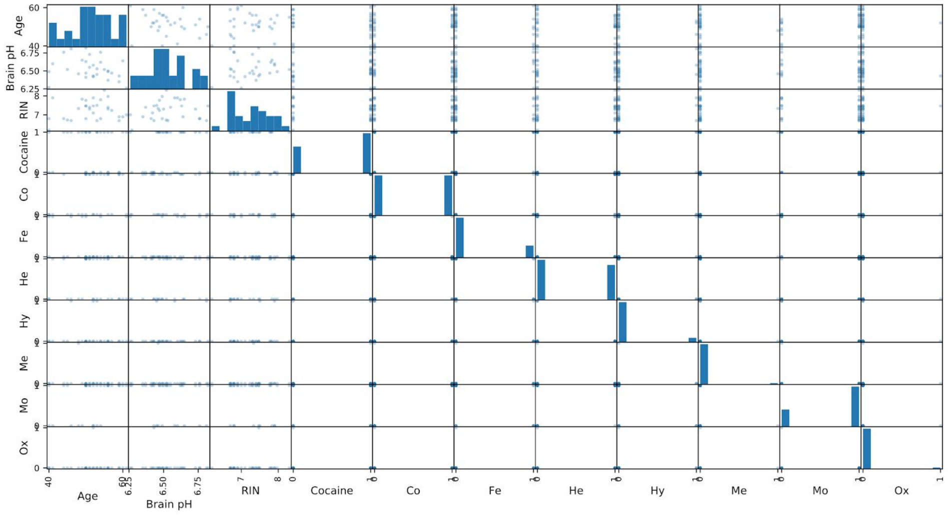
1) Mean values for age, brain pH, and RIN for two subgroups in the experimental group- opioid users who used cocaine and those that did not. 2) Difference of means between opioid users who used cocaine and opioid users who did not use cocaine. 3) Scatterplot matrix of experimental group data to see any correlation within quantitative values. 4) Pie chart of the percentage of each drug found in our experimental group as a whole. 5) Pie chart of the percentage of representation we have in all of our data (control and experimental). 6) Small subset of data after being transformed by one hot encoding "Opioids" column to get counts of specific drugs and find correlations.



1.

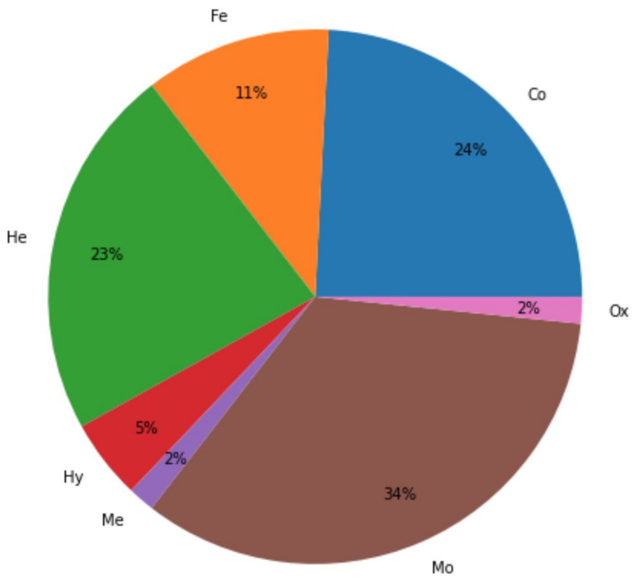

Difference of Means for Users Who Used Cocaine VS Those That Did Not

2.



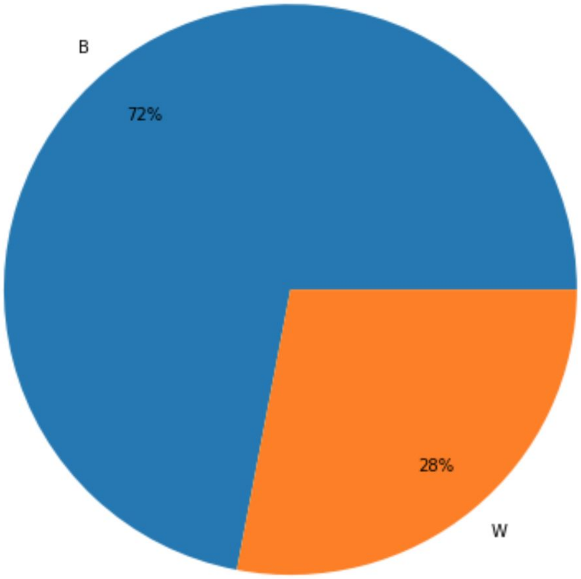
3.

4. Distribution of the kinds of drugs found in all people in the experimental group



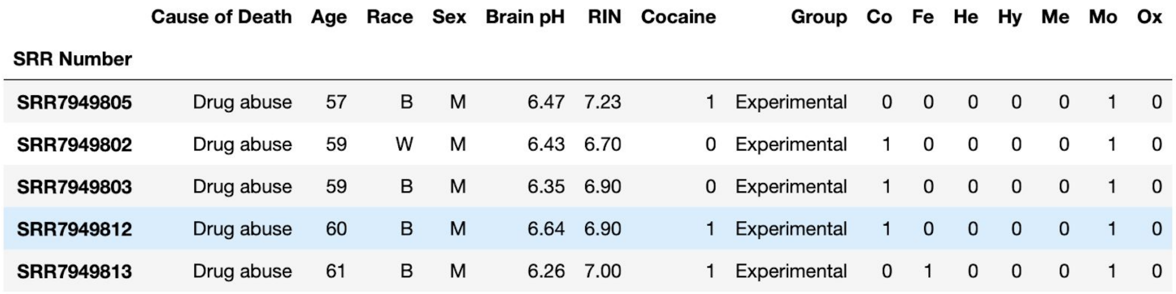
**Co - codeine, Fe - fentanyl, He - heroin, Hy - hydrocodone, Me - methadone, Mo - morphine, Ox - oxycodone

5. Percentage of Race Representation in the control group and experimental group.



** B - Black, W - White

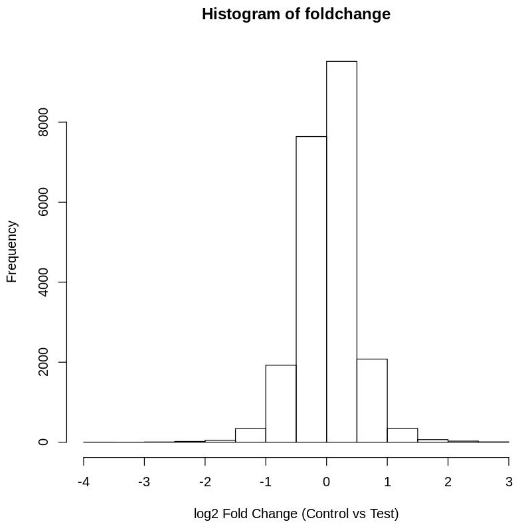
6.



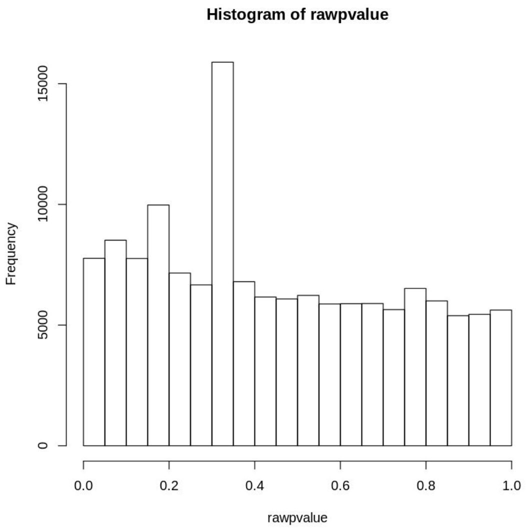
SRR Number	Cause of Death	Age	Race	Sex	Brain pH	RIN	Cocaine	Group	Co	Fe	He	Hy	Me	Mo	Ox
SRR7949805	Drug abuse	57	B	M	6.47	7.23	1	Experimental	0	0	0	0	0	1	0
SRR7949802	Drug abuse	59	W	M	6.43	6.70	0	Experimental	1	0	0	0	0	1	0
SRR7949803	Drug abuse	59	B	M	6.35	6.90	0	Experimental	1	0	0	0	0	1	0
SRR7949812	Drug abuse	60	B	M	6.64	6.90	1	Experimental	1	0	0	0	0	1	0
SRR7949813	Drug abuse	61	B	M	6.26	7.00	1	Experimental	0	1	0	0	0	1	0

Kallisto Output EDA

1) Is a histogram of the frequency of the log2 fold change values that came from the t-test performed between the control and experimental data. 2) Tells us the frequency of p-value from the t-test.



1.



2.