RNA-Seq Data Analysis of Bipolar Disorder

Dohyoung Ko
Akshay Nataraja

CMPSC 497 Genomic Data Science Final Presentation

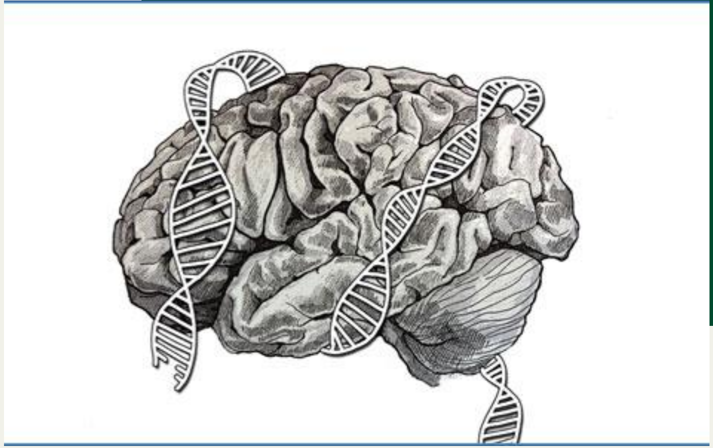
Contents

01	Project Motivation
02	Background
03	Data Retrieval
04	Data Preprocessing
05	Analysis Methods
06	Results
07	Problems
08	Conclusion

Project Motivation

This project aims to analyze RNA-Sequence data by preprocessing samples and developing a CLI pipeline to identify genes that are significantly overexpressed or underexpressed among bipolar disorder patient cohorts.

Background



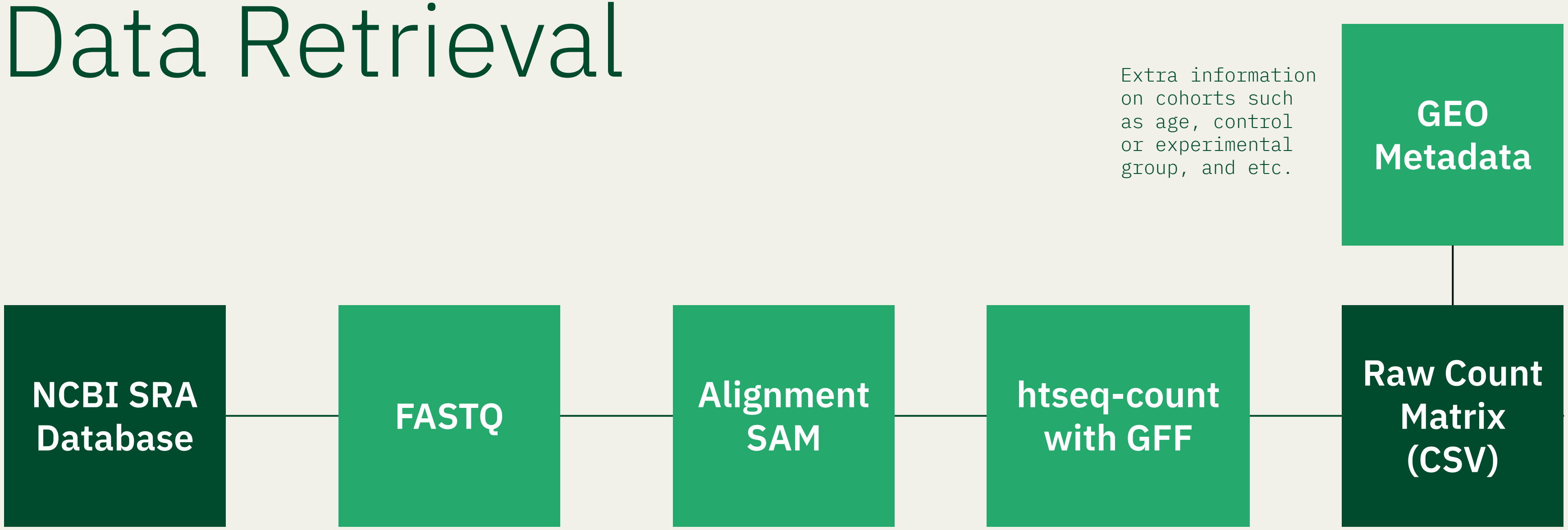
Bipolar Disorder

Overview: Formerly known as manic-depressive illness, it is a chronic mental health condition characterized by extreme mood swings that include emotional highs and lows such as mania and depression. These mood swings can affect a person's sleep, energy, activity, judgment, behavior, and many other daily functions. In severe cases, visual or auditory hallucinations could occur.

Why did we chose to focus on it?

- Genetic Complexity
- Updated Datasets
- Familiarity with Topic

Data Retrieval



Extra information on cohorts such as age, control or experimental group, and etc.

GEO Metadata

NCBI SRA Database

FASTQ

Alignment SAM

htseq-count with GFF

Raw Count Matrix (CSV)

Access the dataset of cohorts from the existing study on SRA via EntreZ

FASTQ files of raw Illumina HiSeq-2000 high throughput RNA-Seq read data

SAM files storing alignments to reference genome after running HISAT2

Download GFF gene mappings from NCBI Homo Sapiens database

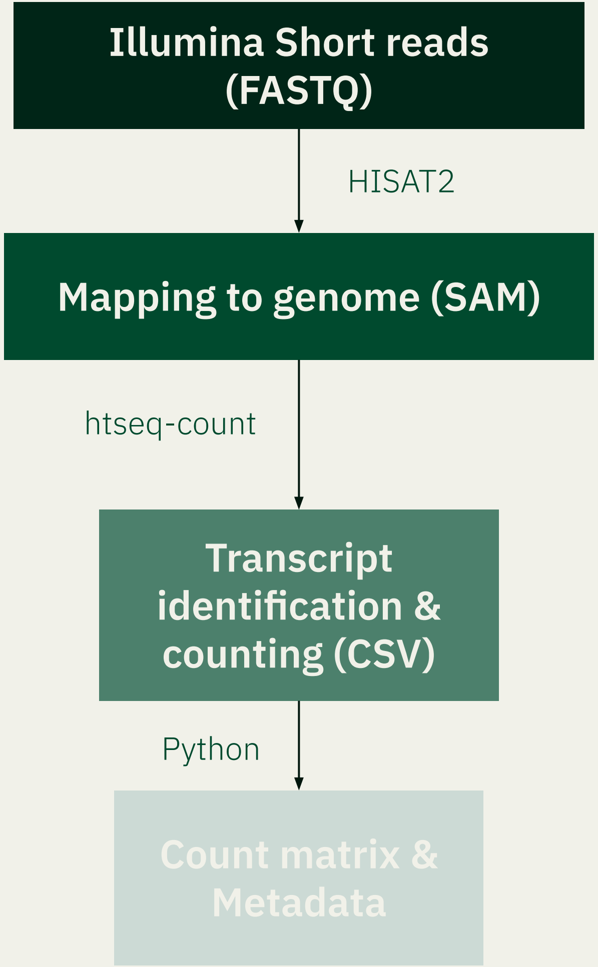
Matrix stored in CSV storing raw counts for each cohort in each row

https://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE80336, https://www.ncbi.nlm.nih.gov/sra/?term=Transcriptome+profiling+of+the+human+dorsal+striatum+in+bipolar+disorder

Data Preprocessing

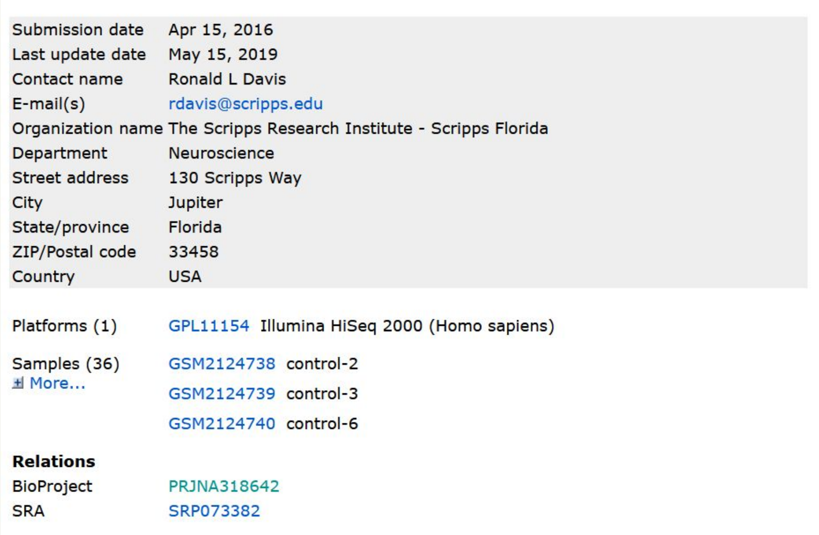
HISAT2: A sequence alignment program that maps our RNA sequences to a reference genome. Indexes the reference genome by creating a hash table of short and non-overlapping substrings of the reference genome that are used to find potential alignment locations. The tool scans the short reads we inputted and matches it to the substrings previously identified.

Htseq-count: Takes in the SAM file along with the GFF file that holds the gene annotations. It then counts the number of reads that overlap with each feature matched with the gene annotations.


Illumina Short reads (FASTQ)
HISAT2
Mapping to genome (SAM)
htseq-count
Transcript identification & counting (CSV)
Python
Count matrix & Metadata

DGE Analysis Pipeline

- Command line application streamlining the whole RNA-Seq DGE analysis process
- Provides one-click solution under High Performance Computing Environment
  - BioProject ID on NCBI and GEO
- Written in Python, R, and shell
- https://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE80336



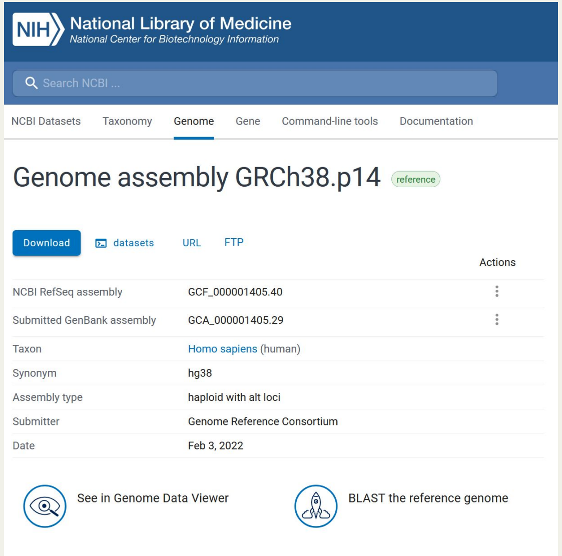
Submission date	Apr 15, 2016
Last update date	May 15, 2019
Contact name	Ronald L Davis
E-mail(s)	rdavis@scripps.edu
Organization name	The Scripps Research Institute - Scripps Florida
Department	Neuroscience
Street address	130 Scripps Way
City	Jupiter
State/province	Florida
ZIP/Postal code	33458
Country	USA
Platforms (1)	GPL11154 Illumina HiSeq 2000 (Homo sapiens)
Samples (36) ⊞ More...	GSM2124738 control-2 GSM2124739 control-3 GSM2124740 control-6
Relations	
BioProject	PRJNA318642
SRA	SRP073382

Digital Research Alliance of Canada | Alliance de recherche numérique du Canada

NCBI Reference Genome

GCF_000001405.40_GRCh38.p14_genomic.gtf.gz

GCF_000001405.40_GRCh38.p14_genomic.fna.gz

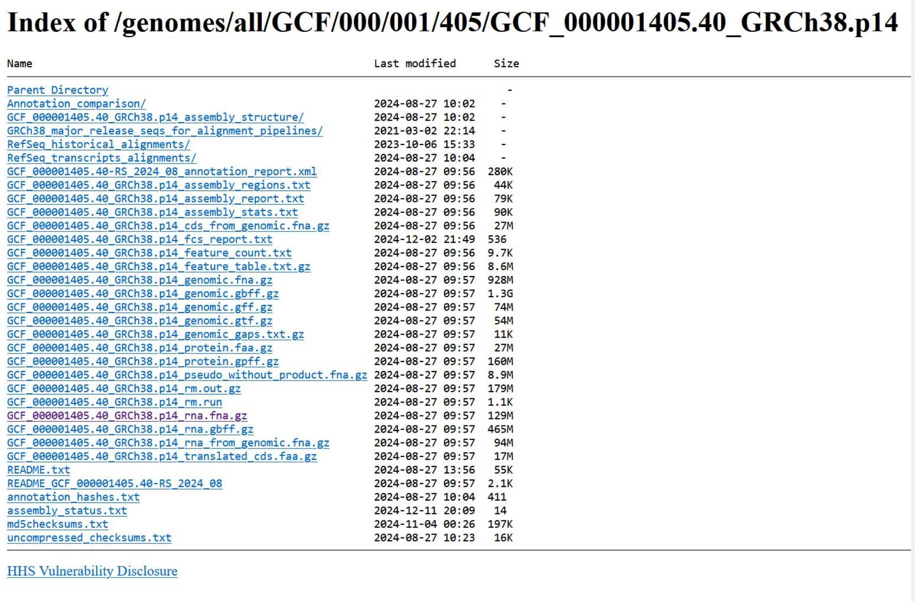

NIH
National Library of Medicine
National Center for Biotechnology Information
Search NCBI ...
NCBI Datasets
Taxonomy
Genome
Gene
Command-line tools
Documentation

Genome assembly GRCh38.p14

reference
Download
datasets
URL
FTP

		Actions
NCBI RefSeq assembly	GCF_000001405.40	⋮
Submitted GenBank assembly	GCA_000001405.29	⋮
Taxon	Homo sapiens (human)	
Synonym	hg38	
Assembly type	haploid with alt loci	
Submitter	Genome Reference Consortium	
Date	Feb 3, 2022	

See in Genome Data Viewer
BLAST the reference genome

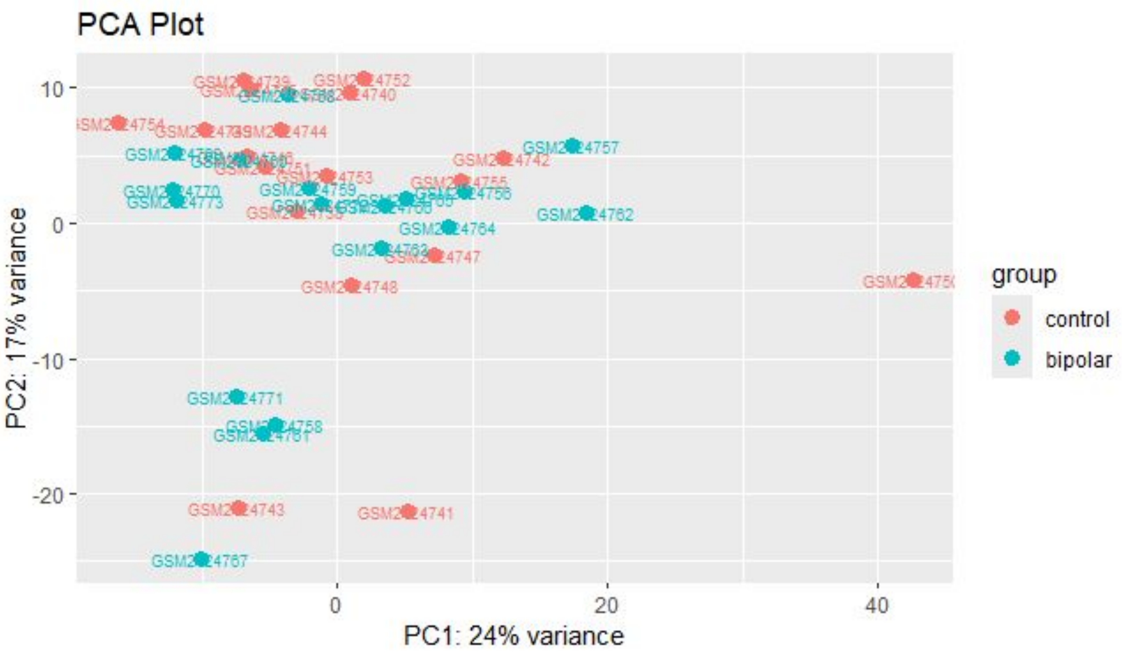


Index of /genomes/all/GCF/000/001/405/GCF_000001405.40_GRCh38.p14

```
Name                                                      Last modified      Size

Parent Directory                                                               -
Annotation_comparison/                                    2024-08-27 10:02    -
GCF_000001405.40_GRCh38.p14_assembly_structure/           2024-08-27 10:02    -
GRCh38_major_release_seqs_for_alignment_pipelines/        2021-03-02 22:14    -
RefSeq_historical_alignments/                             2023-10-06 15:33    -
RefSeq_transcripts_alignments/                            2024-08-27 10:04    -
GCF_000001405.40-RS_2024_08_annotation_report.xml         2024-08-27 09:56  280K
GCF_000001405.40_GRCh38.p14_assembly_regions.txt          2024-08-27 09:56   44K
GCF_000001405.40_GRCh38.p14_assembly_report.txt           2024-08-27 09:56   79K
GCF_000001405.40_GRCh38.p14_assembly_stats.txt            2024-08-27 09:56   90K
GCF_000001405.40_GRCh38.p14_cds_from_genomic.fna.gz       2024-08-27 09:56   27M
GCF_000001405.40_GRCh38.p14_fcs_report.txt                2024-12-02 21:49  536
GCF_000001405.40_GRCh38.p14_feature_count.txt             2024-08-27 09:56  9.7K
GCF_000001405.40_GRCh38.p14_feature_table.txt.gz          2024-08-27 09:56  8.6M
GCF_000001405.40_GRCh38.p14_genomic.fna.gz                2024-08-27 09:57  928M
GCF_000001405.40_GRCh38.p14_genomic.gbff.gz               2024-08-27 09:57  1.3G
GCF_000001405.40_GRCh38.p14_genomic.gff.gz                2024-08-27 09:57   74M
GCF_000001405.40_GRCh38.p14_genomic.gtf.gz                2024-08-27 09:57   54M
GCF_000001405.40_GRCh38.p14_genomic_gaps.txt.gz           2024-08-27 09:57   11K
GCF_000001405.40_GRCh38.p14_protein.faa.gz                2024-08-27 09:57   27M
GCF_000001405.40_GRCh38.p14_protein.gpff.gz               2024-08-27 09:57  160M
GCF_000001405.40_GRCh38.p14_pseudo_without_product.fna.gz 2024-08-27 09:57  8.9M
GCF_000001405.40_GRCh38.p14_rm.out.gz                     2024-08-27 09:57  179M
GCF_000001405.40_GRCh38.p14_rm.run                        2024-08-27 09:57  1.1K
GCF_000001405.40_GRCh38.p14_rna.fna.gz                    2024-08-27 09:57  129M
GCF_000001405.40_GRCh38.p14_rna.gbff.gz                   2024-08-27 09:57  465M
GCF_000001405.40_GRCh38.p14_rna_from_genomic.fna.gz       2024-08-27 09:57   94M
GCF_000001405.40_GRCh38.p14_translated_cds.faa.gz         2024-08-27 09:57   17M
README.txt                                                2024-08-27 13:56   55K
README_GCF_000001405.40-RS_2024_08                        2024-08-27 09:57  2.1K
annotation_hashes.txt                                     2024-08-27 10:04  411
assembly_status.txt                                       2024-12-11 20:09   14
md5checksums.txt                                          2024-11-04 00:26  197K
uncompressed_checksums.txt                                2024-08-27 10:23   16K
```

HHS Vulnerability Disclosure



PCA’s goal is to reduce the dimensionality between the two groups and visualize the main sources of variability within our data. Our data is split between control and the bipolar cohorts.

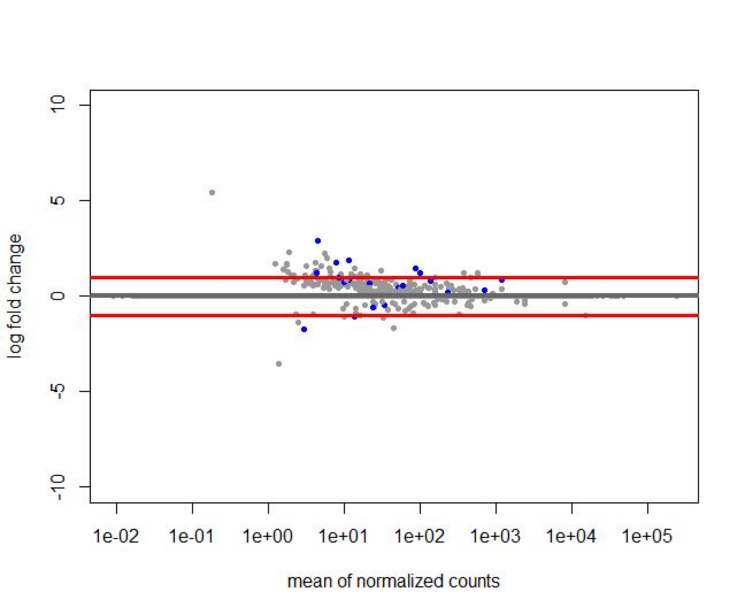
Observations:

Partial Overlap between the control and bipolar group.

But they key thing to note is that the total variance captured by the plot is only 41% (24% from PC1 and (17% from PC2). Ideally, our total variance should be at least 70% captured.

An explanation could be high dimensionality of the RNA-Seq, or notable localized changes in the gene expression.

Results

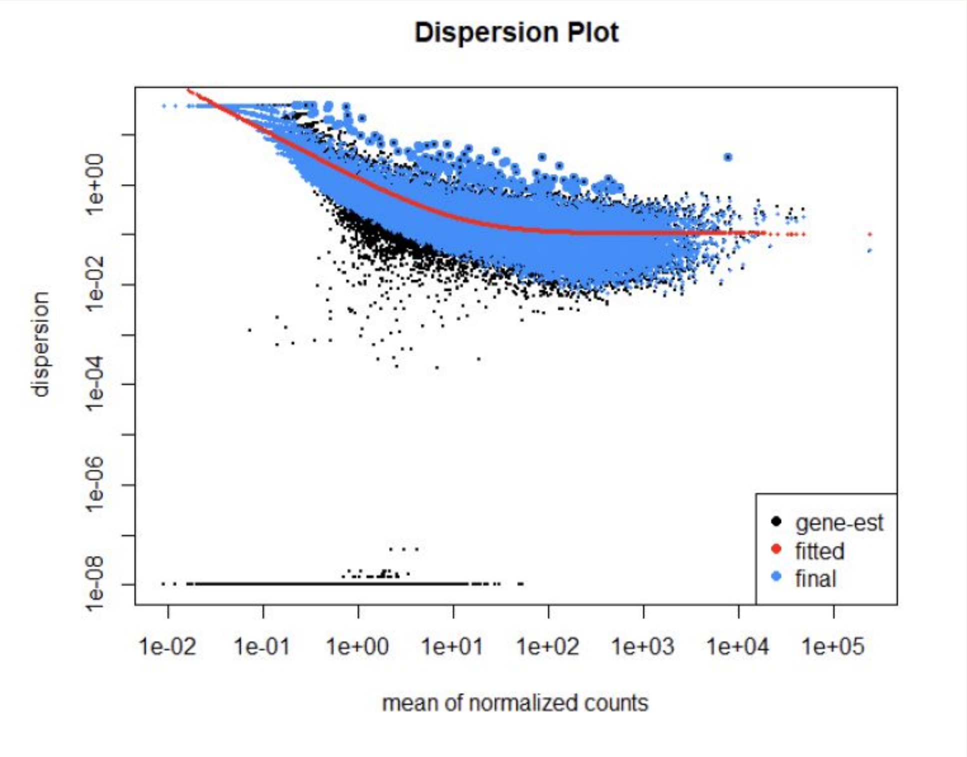


This MA plot illustrates gene expression changes between two cohorts, "control" and "bipolar," in this case study.

The plot indicates that genes with lower mean expression values exhibit highly variable log fold changes. The blue dots in the figure represent differentially expressed genes (DEGs) that have adjusted p-values below the threshold of 0.05. Genes that meet the filtering criteria of a p-value less than 0.05 and a fold change of either greater than 2 (i.e., $\ell > 2$) or less than -2 (i.e.,$\ell < -2$) are considered statistically significant.

The red vertical lines in the graph mark these thresholds. A fold change greater than 2 (i.e., log $\ell > 0$) indicates up-regulated genes, while a fold change less than 1 (i.e., log $\ell < 0$) indicates down-regulated genes. Among the DEGs identified are five up-regulated genes and eight down-regulated genes. This analysis is further illustrated in the following plot.

Results



DESeq2 offers a valuable criterion known as dispersion to assess whether its negative binomial model is appropriate for the given data.

Goal of the dispersion plot is to see if our dataset is well-fitted to the DESeq2 Model.

We can observe that the fitting function is monotonically decreasing. This graph indicates a continuous decrease in dispersion as the mean expression increases. The dispersion estimates generally cluster around the curve, suggesting that it it is well-fitted. However, there is some shrinking as we only have one replicate for each sample in the group.

Results

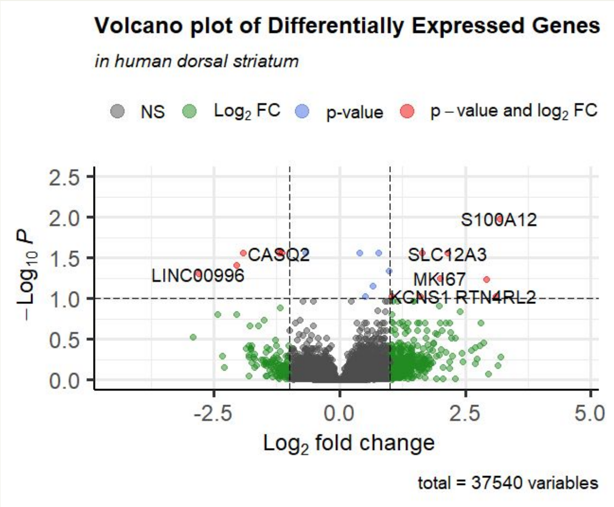
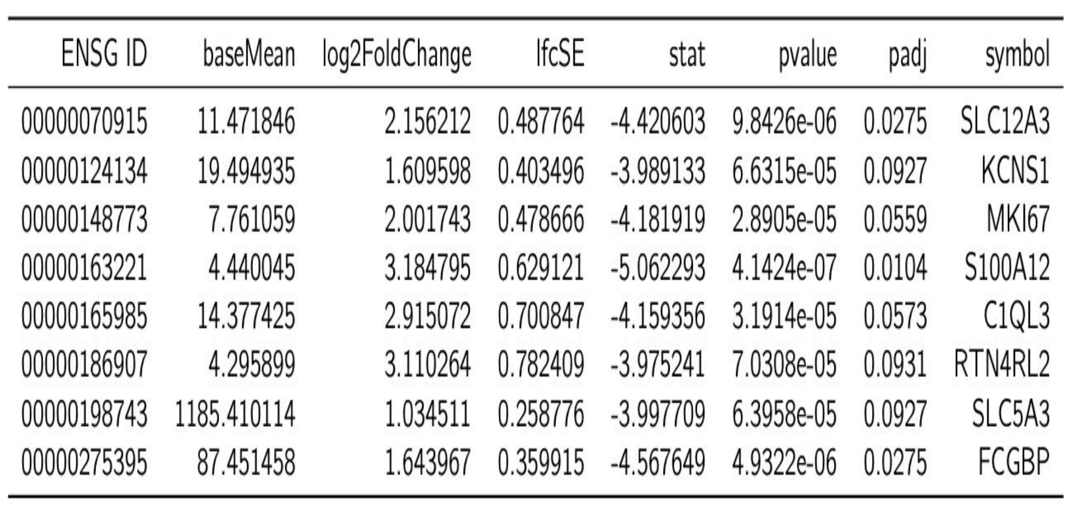


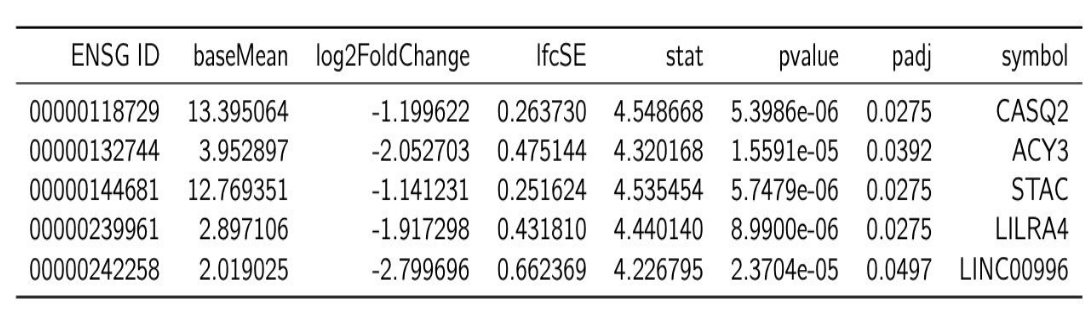
Figure 1: Volcano plot of DEGs in human dorsal striatum

A fold change greater than 2 (i.e., log ℓ > 0) indicates up-regulated genes, while a fold change less than 1 (i.e., log ℓ < 0) indicates down-regulated genes.



ENSG ID	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	symbol
00000070915	11.471846	2.156212	0.487764	-4.420603	9.8426e-06	0.0275	SLC12A3
00000124134	19.494935	1.609598	0.403496	-3.989133	6.6315e-05	0.0927	KCNS1
00000148773	7.761059	2.001743	0.478666	-4.181919	2.8905e-05	0.0559	MKI67
00000163221	4.440045	3.184795	0.629121	-5.062293	4.1424e-07	0.0104	S100A12
00000165985	14.377425	2.915072	0.700847	-4.159356	3.1914e-05	0.0573	C1QL3
00000186907	4.295899	3.110264	0.782409	-3.975241	7.0308e-05	0.0931	RTN4RL2
00000198743	1185.410114	1.034511	0.258776	-3.997709	6.3958e-05	0.0927	SLC5A3
00000275395	87.451458	1.643967	0.359915	-4.567649	4.9322e-06	0.0275	FCGBP

Table 1: Complete list of up-regulated genes



ENSG ID	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	symbol
00000118729	13.395064	-1.199622	0.263730	4.548668	5.3986e-06	0.0275	CASQ2
00000132744	3.952897	-2.052703	0.475144	4.320168	1.5591e-05	0.0392	ACY3
00000144681	12.769351	-1.141231	0.251624	4.535454	5.7479e-06	0.0275	STAC
00000239961	2.897106	-1.917298	0.431810	4.440140	8.9900e-06	0.0275	LILRA4
00000242258	2.019025	-2.799696	0.662369	4.226795	2.3704e-05	0.0497	LINC00996

Table 2: Complete list of down-regulated genes

Results

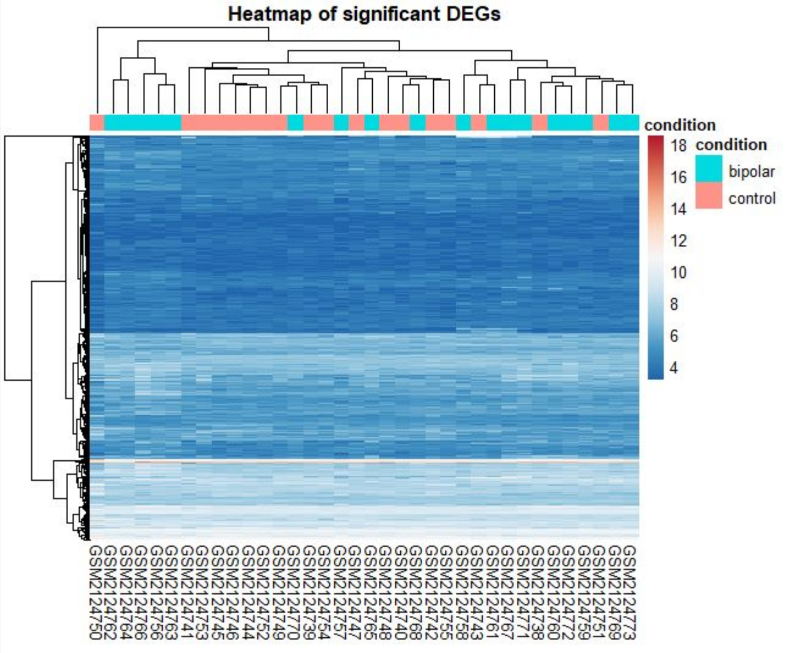


Figure 2: Heatmap of significant DEGs

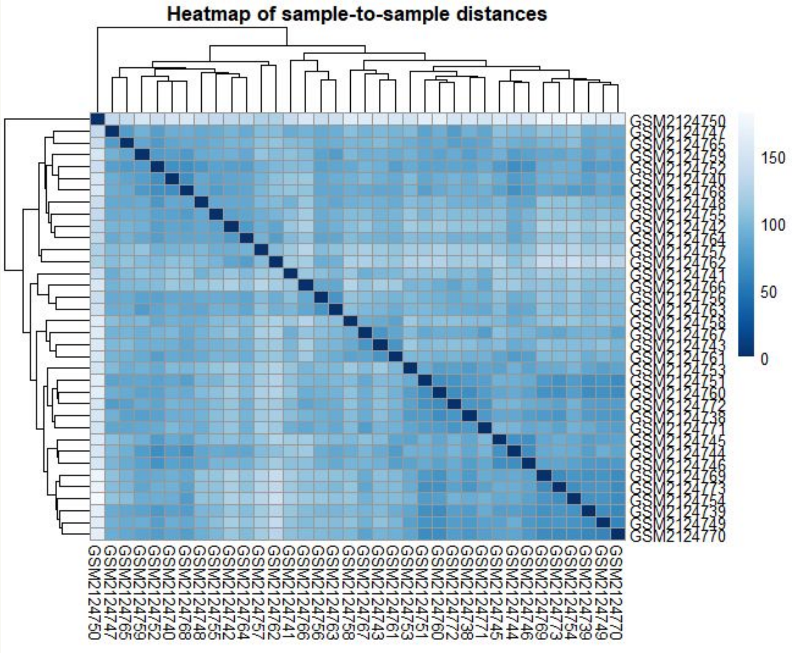


Figure 2: Heatmap of sample-to-sample distances

Results on Normalized Dataset

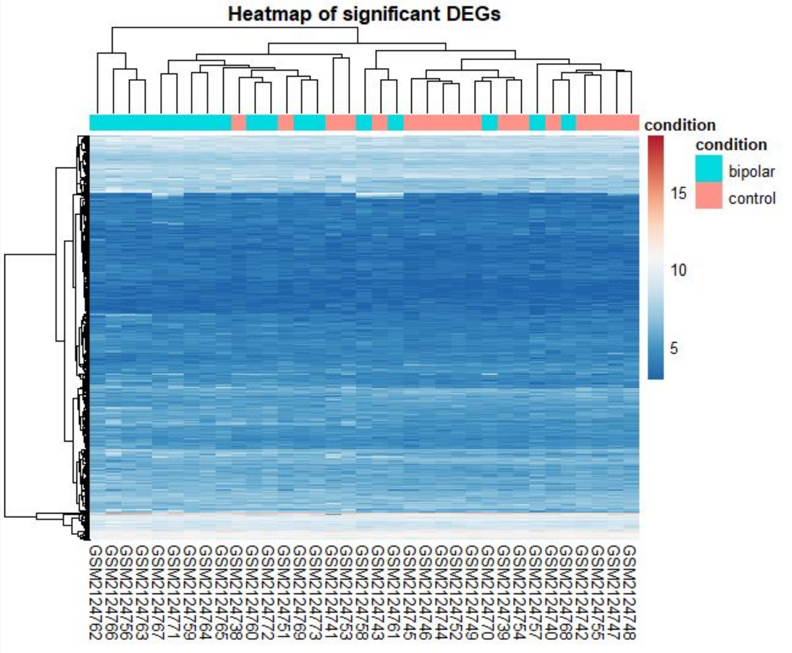


Figure 2: Heatmap of significant DEGs

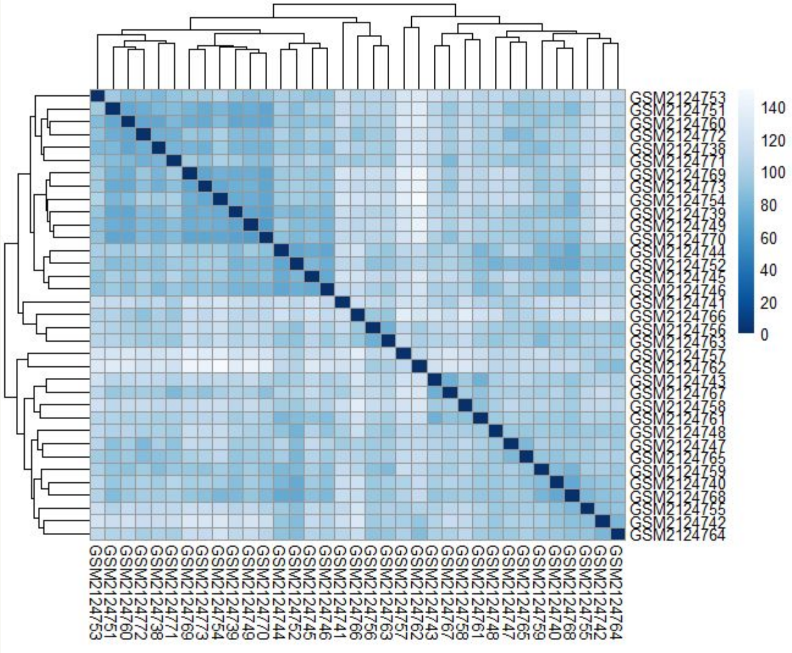


Figure 2: Heatmap of sample-to-sample distances

Under-Expressed

LILRA4- Its expression enhances immune regulation, contributing to the innate and adaptive immune response, including during viral infections and inflammatory diseases.

ACY3- This gene plays a significant role in maintaining renal function and detoxification processes.

LINC00996- A long intergenic non-coding RNA, has been shown to suppress lung adenocarcinoma cells, indicating its role in tumor suppression and cellular homeostasis.

CASQ2- It is pivotal in regulating calcium homeostasis, influencing muscle contraction and relaxation cycles.

STAC- STAC enhances the activity of calcium channels and slows the inactivation of CACNA1C, which is critical for maintaining cellular calcium signaling.

Over-Expressed

SLC12A3- Down-regulation of SLC12A3 is predicted to disrupt processes in regulating ion homeostasis, cell volume, and neuronal excitability.

KCNS1- Plays a part in regulating neuronal excitability and synaptic function. Its down-regulation has been associated with altered synaptic transmission.

MKI67- Down-regulation of MKI67 has been linked to reduced cell division, often serving as a marker for low proliferative activity in cells.

S100A12- Plays a part in regulating calcium-dependent processes, including muscle contraction and neuronal signaling, down-regulation may impair immune responses and antibacterial properties.

C1QL3- Involved in synaptic regulation, glucose homeostasis, and neuronal projections.

RTN4RL2- Is critical for neuronal development and plasticity; its down-regulation correlates strongly with bipolar disorder.

SLC5A3- Is essential for glucose uptake and metabolism.

FCGBP- Functions as a tumor suppressor gene, particularly in head and neck squamous cell carcinoma.

Problems

Data Intensive Processes

- RAW FASTQ files were data intensive.
- Processing could not be done locally.
- Required a HPC to handle requests.

Preprocessing Complexity

- Preprocessing the data required multiple frameworks.
- Data was sourced from different locations.
- No streamlined pipeline to preprocess raw RNA-Seq data.

Thank You

Any

Questions

:(:

ACTG