

Project Title:

**Differences in the Gene Expression Profile of Schizophrenia Patients
Who Experience Antipsychotic-induced Weight Gain Versus Patients
Who Don't Gain Weight.**

Name: Shaista Madad (sm8847@nyu.edu)

Course: Applied Genomics

Table of Contents

Abstract	3
Introduction	5
Experimental Design	6
Method	8
Results	11
Results 1.1: DGE Between Weight Gain and No Weight Gain Group	11
Result 1.2: DGE within Weight Gain Group (Presumably Between Two Time Points)	12
Results 1.3: DGE Between Two Time Periods in No Weight Gain Group.....	16
Result 1.4 GSEA in Weight Gain Versus No Weight Gain Group Using All GeneSets from MSig Database	19
Result 1.5: GSEA Using Immunologic Signature Gene Set Only	20
Discussion	22
1.1Single-Gene Expression Analysis	23
1.2 Pathway Analysis.....	25
Limitations and Further Analysis.....	29
Appendix 1: Screenshot of the Wget Command Use to Download the Fastq Files	30
Appendix 2: Fastqc and Index Bash Scripts	31
Appendix 3: Fastqc Summary Reports	32
Appendix 4: .cls file format for GSEA	33
Appendix 5: Alignment and Bam Conversion Bash Scripts	33
Appendix 6: RSubread Code	34
Appendix 7: Function in R for Normalization for GSEA	35
Appendix 8: 62 DE Genes between Weight Gain and No Weight Gain Group	36
Appendix 9: 14 Gene Sets Upregulated in Weight Gain Group	38
Bibliography	39

Abstract

Antipsychotic induced weight gain (AIWG) is an important contributor to the high cardiovascular mortality seen in patients in schizophrenia. However, not all schizophrenia patients who take the medication gain weight. This suggests a potential genetic underpinning for the differential effect of antipsychotic treatment on the BMI of schizophrenia patients. In this project, I investigated the differences in the gene expression profile of two groups (n=36 in each group) of schizophrenia patients based on a study by Crespo-Facorro and colleagues (2019). One group, WG, gained weight following three months of antipsychotic treatment, and the other, NWG, did not. I found 62 differentially expressed (DE) genes between the WG group and the NWG group. Geneset Enrichment Analysis (GSEA) using the genesets from MSig database revealed 12 Gene Sets to be significantly enriched in the WG group at FDR q-value < .05. Five of these gene sets are associated with the cardiovascular system, five with metabolism and four with regulation of the nervous system. These gene sets have the potential to be molecular signatures to identify schizophrenia patients at risk of AIWG. Thus, knowledge of differences in gene patterns between subgroups of patients can help develop more precise and customized treatment plans. However, there were limitations to my data analysis due to incomplete sample information. The original study had tested the two groups (n=18) before and after three months of treatment. The differences in gene expression profile of the two groups were then tested independently (paired t-tests) and compared. However, the information on samples from the same individual before and after treatment was missing in the metadata. MDS plot based clustering helped divide the WG and NWG groups into sub-groups to do a differential gene analysis. However, the usefulness of the results is questionable since the clustering may have been driven by factors such as sex rather than treatment time period. My project highlights the importance of adhering to the good scholarly practice of providing complete metadata when publishing research. The lack of adequate sample information is an important reason for the

reproducibility crisis in biomedical research. Future research to find short nucleotide polymorphisms between the WG and NWG group using SNP analysis will help find genetic markers for differential response to antipsychotic treatment.

Introduction

Schizophrenia is a psychotic disorder which affects approximately 1% of the population worldwide (Leucht et al., 2007) with onset typically between the end of puberty and the end of the third decade (Khandaker, 2015).

Schizophrenia patients have been reported to have a 15-20-year shorter life span (Khandaker, 2015).

Cardiovascular disease related mortality is the leading natural cause of death in the schizophrenia population (Kim et al., 2001).

Identifying the exact cause of the high incidence of cardiovascular mortality is tricky due to the interplay of a multitude of risk factors, which all contribute to the development of cardiovascular disease. One cause may be lifestyle factors, such as high rates of smoking, lack of exercise and a tendency not to seek medical help, which predispose to cardiovascular complications in the schizophrenic population (Hennekens et al., 2005).

However, antipsychotic medication, particularly the atypical antipsychotics (AAPs), have been associated with a number of side effects including weight gain and metabolic syndrome, which are significant risk factors for cardiovascular mortality. Metabolic syndrome is a combination of symptoms which include impaired glucose clearance, insulin resistance, hyperlipidaemia, weight gain and hypertension (Leung et al., 2012). The pharmacological actions underlying these side effects are not fully characterized, although weight gain has been attributed to 5HT_{2C} antagonism and activation of hypothalamic AMP-Kinase via Histamine (H₁) receptors which increases food intake (Kim et al., 2001).

Interestingly, not all patients who take antipsychotic medication gain weight. Crespo-Facorro et al., (2019) conducted RNA-Seq analysis on blood samples from two groups of first-episode schizophrenia patients. One group, which had gained weight, hereafter referred to as WG group, after three months of antipsychotic treatment. The other group (NWG) did not gain weight. I chose this dataset to understand why a subset of schizophrenia patients gain weight but others don't. What is the difference in the gene expression patterns between the WG and NWG group? Is there an overlap? Which pathways are these genes involved in? Which

physiological pathways are dysregulated by antipsychotic treatment in the WG group? Are there signature gene sets enriched only in the WG group which can become predictors of whether a patient is susceptible to weight gain or not? These questions are important from a translational perspective as antipsychotic induced weight gain (AIWG) is an important contributor to the high incidence of cardiovascular mortality in patients of schizophrenia.

Detecting genetic factors that cause antipsychotic induced weight gain (AIWG) will help generate prediction tests to filter patients at risk of AIWG in the future, once studies with larger data sets confirm the preliminary findings. The findings from this project have the potential to contribute towards the development of targeted therapeutic interventions for schizophrenia patients: e.g, avoid prescription of antipsychotics with greater adverse effects on metabolic profile to patients at risk of AIWG.

Experimental Design

The experimental design (Table 1) is a simple 2 sample t-test with two groups: Weight Gain group (WG) and the No Weight Gain group (NWG). Each group has 36 samples. Two samples in each group came from one patient, one before and one after the three-month antipsychotic treatment. Unfortunately, the metadata provided by the authors of the original study did not provide the identity of samples belonging to the same patient. Had the information been available, a two-factor paired design (Table 2) with the necessary adjustments for a time course experiment would have been the suitable approach (Limma User Guide, 2020).

Table 1: Experimental Design	
Phenotype	Name
WG	C23ELACXX_8_19.bam
WG	C2LRRACXX_1_18.bam
WG	C2LRRACXX_1_22.bam
WG	HHW7MBBXX_4_15.bam
NWG	C23ELACXX_6_16.bam
NWG	C23ELACXX_7_18.bam
NWG	C23K9ACXX_7_4.bam
NWG	C24A4ACXX_1_23.bam
NWG	C2LRRACXX_4_21.bam

Table 2: Suitable Experimental Design		
Phenotype	Time	Name
WG	Before	C23ELACXX_8_19.bam
WG	Before	C2LRRACXX_1_18.bam
WG	Before	C2LRRACXX_1_22.bam
WG	After	C6KALANXX_1_27.bam
WG	After	C6KALANXX_7_14.bam
WG	After	C6KALANXX_7_15.bam
NWG	Before	C9F62ANXX_1_20.bam
NWG	Before	C9F62ANXX_3_15.bam
NWG	After	C9F6AANXX_6_16.bam
NWG	After	C9FT8ANXX_1_4.bam
NWG	After	D259AACXX_1_7.bam
NWG	Before	D259AACXX_2_6.bam

Method

The sample data consisted of 144 fastq files. These 144 files represent 72 paired end RNASeq samples. Using the `wget` command (Appendix 1), I downloaded the 144 fastq file to my scratch directory in NYU HPC from the SRA database (SRA Run Selector, 2020). The SRA Study ID of my data is ERP114104.

The first step was conducting quality checks on the fastq files to see if they needed modifications such as adaptor trimming or marking of duplicates at later stages. Using a bash script (Appendix 2) on the command line of NYU HPC cluster, I ran `fastqc` on all 144 files to generate 144 summary reports. From the analysis of the html files, I did not see the need for running `trimmomatic`. Adaptor content was zero, which suggested that the files were already trimmed (An example is Figures 1 and 2). Analysis of summary reports of the `fastqc` showed that most files had 0 sequences flagged as poor quality, and all fastq files passed the Adaptor Content metric (Appendix 3 shows summary statistics for a few fastq files).

The reference genome fasta file and annotation gtf file were downloaded from the UCSC website (UCSC, 2020). Gene Set Enrichment Analysis (GSEA) was carried out using the GSEA software developed by Broad Institute and UC San Diego (GSEA, 2020). The process consisted of downloading GSEA version 4.0.3 to my local computer, preparing the input files based on the format specified in the GSEA user guide (GSEA User Guide, 2020). Three files were needed for the GSEA (Table 3). The counts data was required to be normalized. I used a normalization function (provided in Appendix 7) provided on GSEA documentation published on Github (hxin/ograph, 2020).

Table 3: Files Required for GSEA Using the GSEA Software

File Name	Requirements and Format
Expression Dataset	Normalized Data in .gct format
Phenotype Labels	72 samples in total in .cls format (Appendix 3)
Genesets	Genesets from Molecular Signatures Database (Liberzon et al., 2011) prepared to be used with the GSEA software (GSEA, 2020). The 8 major gene set collections, consisting of 25724 gene sets in total were used. A gmt file called = “msigdb.v7.1.symbols.gmt” (GSEA Downloads, 2020) with gene sets described using gene symbols was used as the annotation and reference genome files in this analysis also used HGCN gene symbols.

Step 1: Quality Checks (Fastqc)

Step 2: Collection of MetaData

Step 3: Development of Index files (hisat2)
(reference genome and annotation files from UCSC website)

Step 4: Alignment rate for all 72 fasta pair files (144) >90-95 percent. (hisat2) (bash script in Appendix 5)

Step 5: Conversion from sam to bam files (72 in total, bash script in Appendix 5)

Step 6: Get raw reads(Rsubread; featureCounts (>75% reads successfully assigned for all 72 bam files)

Step 7: Differential Gene Analysis (limma and edgeR, Ritchie et al., 2015; Conesa et al., 2016) R Code Provided in Supplementary Information

Step 8: Gene Set Enrichment Analysis

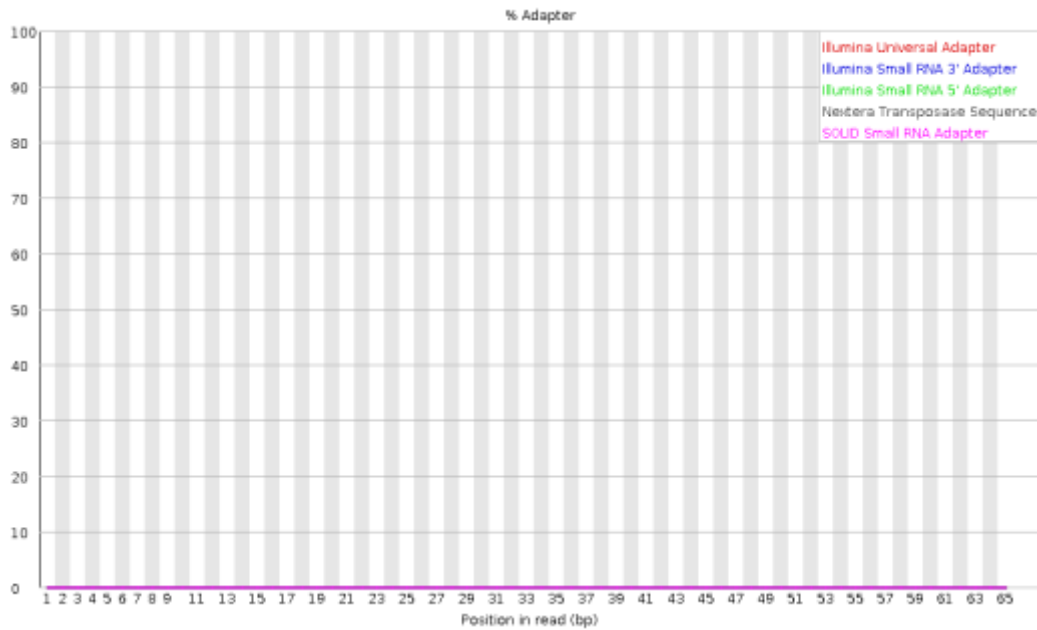


FIGURE 1 ADAPTOR CONTENT GRAPH FOR ALL 144 FILES LOOKED SIMILAR TO THE GRAPH FOR SAMPLE CBVKEANXX_7_20_1.FASTQ. THIS SUGGESTS THAT THE FASTQ FILES WERE ALREADY TRIMMED

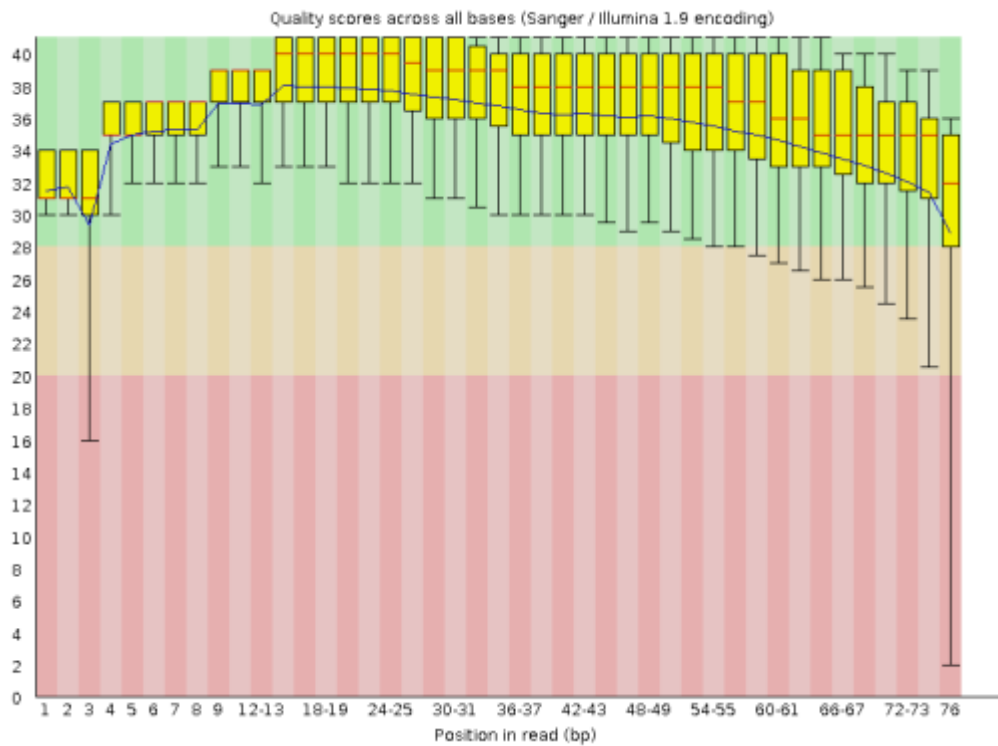


FIGURE 2 PER BASE QUALITY SCORE FOR SAMPLE C8VKEANXX_7_20_1.FASTQ ACROSS THE ENTIRE LENGTH (76 BPS) THE QUALITY SCORE REMAINS ABOVE 30. THIS IS A GOOD SCORE. MOST OF THE SAMPLES HAD SIMILAR GRAPHS

Results

Results 1.1: DGE Between Weight Gain and No Weight Gain Group

Differential gene expression analysis between the WG and NWG group revealed 62 genes to be differentially expressed between the two groups at $FDR < 5\%$ (Appendix 8 shows all genes with details). The top 20 genes (based on descending adjusted p values) are mostly downregulated with a negative log-fold change between 0 and -1. The clustering of the 72 samples using MDS plot (Ritchie et al., 2015) revealed an interesting trend (Figure 3). The 36 samples on the left side of the plot constituted of 18 weight gain samples and 18 no weight gain samples. It is possible that this clustering is based on gene expression profiles before and after the three-month-antipsychotic treatment. Therefore, using this clustering, I further divided the weight gain and no weight gain groups into two groups (each group consisting of 18 samples, presumably from before and after treatment, however, this is speculation).

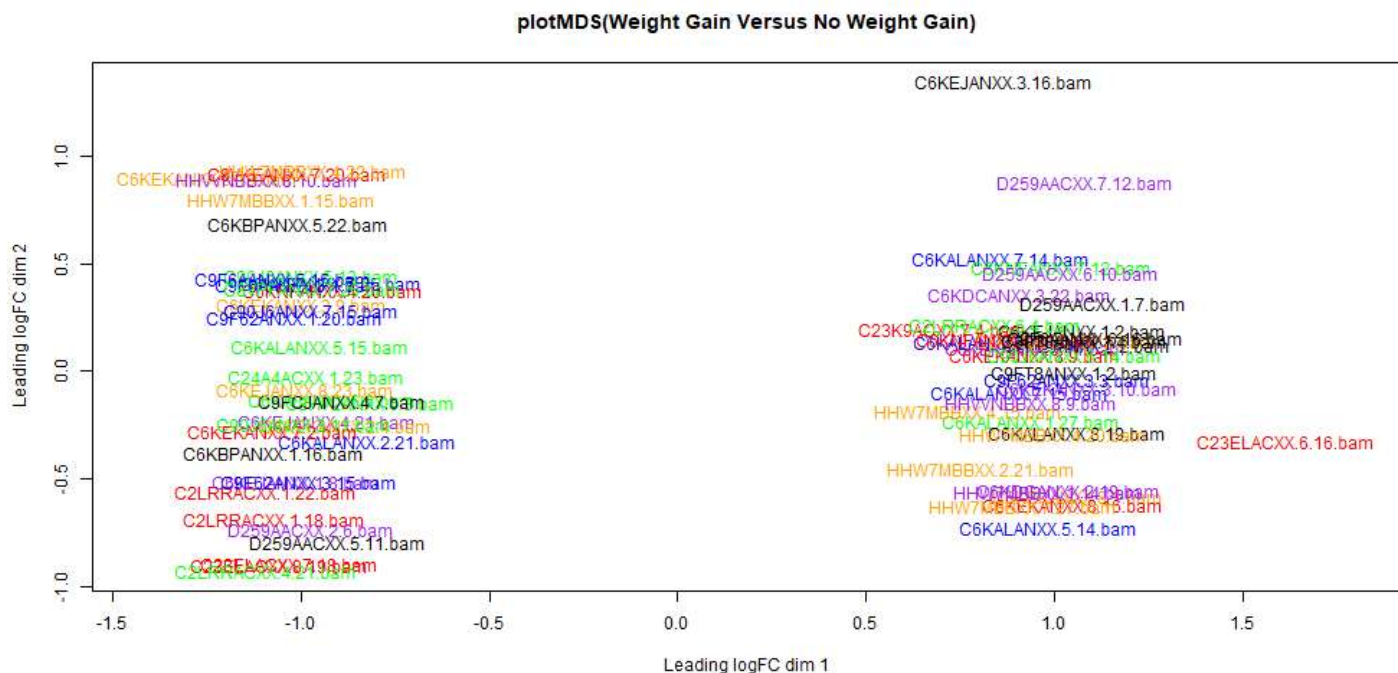


FIGURE 3: MDS PLOT OF THE WEIGHT GAIN VERSUS NO WEIGHT GAIN GROUP. 36 SAMPLES CLUSTERED TO THE RIGHT SIDE OF THE PLOT AND 36 SAMPLES CLUSTERED TO THE LEFT SIDE. INTERESTINGLY, 18 SAMPLES ON THE RIGHT SIDE WERE THE WEIGHT GAIN SAMPLES, AND 18 WERE NO WEIGHT GAIN SAMPLES. THERE WAS NO SUB-CLUSTERING BETWEEN THE 18 WEIGHT GAIN AND 18 NO WEIGHT GAIN SAMPLES ON RIGHT AND LEFT SIDES. THEY WERE RANDOMLY SPREAD. IT IS POSSIBLE THAT THIS CLUSTERING RESULTED FROM GENE EXPRESSION CHANGES DUE TO THE THREE MONTH-ANTIPSYCHOTIC TREATMENT. HENCE, ALL THE SAMPLES FROM THE LEFT SIDE MAY BE REPRESENTING THE GENE EXPRESSION OF BOTH GROUPS BEFORE OR AFTER THE ANTIPSYCHOTIC TREATMENT.

Result 1.2: DGE within Weight Gain Group (Presumably Between Two Time Points)

The experimental design (Table 3) was again a simple 2-sample design (two levels: WGB and WGA representing presumably two different time points; ideally a paired t-test should have been used here if sample information was complete). Eighteen samples belonged to time A (WGA) and eighteen samples to time B (WGB). DGE analysis using limma and edgeR revealed 158 genes to be differentially expressed between the two time periods at FDR< 5%. Table 4 shows the top 20 differentially expressed genes. MDS plot (Figure 4) clustering of the weight gain only analysis corresponded to the clustering in Figure 1, i.e., the 18 samples on the left side in Figure 3 clustered together in Figure 4.

Table 3: Experimental Design Weight Gain only	
stressor	name
WGB	C23ELACXX_8_19.bam
WGB	C2LRRACXX_1_18.bam
WGB	C2LRRACXX_1_22.bam
WGA	C6KALANXX_1_27.bam
WGB	C6KALANXX_5_15.bam
WGB	C6KALANXX_6_6.bam
WGA	C6KALANXX_7_14.bam
WGA	C6KALANXX_7_15.bam
WGB	C6KBPANXX_7_25.bam
WGA	C6KEJANXX_1_2.bam
WGA	C6KEJANXX_2_2.bam
WGA	C6KEJANXX_3_16.bam
WGB	C6KEJANXX_4_21.bam

Table 4: Top 20 Differentially Expressed Genes Between Two Time Periods in WG Group

Gene	logFC	AveExpr	t	P.Value	adj.P.Val	B
PRKY	-5.91661	2.763412	-46.4577	3.01E-36	1.47E-32	69.82349
DDX3Y	-12.3524	0.689887	-60.4466	9.99E-41	1.75E-36	65.68665
UTY	-11.1414	-0.12562	-59.8863	1.44E-40	1.75E-36	65.32989
KDM5D	-12.1499	0.50157	-54.3832	6.34E-39	5.15E-35	63.91451
TXLNGY	-11.0252	-0.14736	-46.5209	2.86E-36	1.47E-32	61.01087
RPS4Y1	-12.5725	1.101853	-44.5372	1.56E-35	6.35E-32	60.69297
USP9Y	-11.0325	0.169541	-38.324	5.36E-33	1.45E-29	57.12998
TTY15	-7.96528	-1.71019	-39.3615	1.90E-33	6.62E-30	57.1115
ZFY	-9.55398	-0.83228	-38.5766	4.15E-33	1.26E-29	56.96953
EIF1AY	-10.4018	-0.25575	-31.8296	6.82E-30	1.51E-26	52.38783
LINC00278	-5.9055	-2.83928	-32.1855	4.45E-30	1.09E-26	51.62589
TMSB4Y	-6.72398	-2.3391	-31.1342	1.59E-29	3.22E-26	51.14184
LOC107987346	-5.52032	-3.01728	-30.0207	6.37E-29	1.19E-25	49.71962
XIST	12.19701	1.032335	28.47155	1.53E-27	2.67E-24	48.60167
LOC107987347	-5.49512	-3.04832	-27.2379	2.55E-27	4.14E-24	47.14966
BCORP1	-8.74839	-1.40816	-26.1821	1.13E-26	1.72E-23	46.78717
LOC105377223	-4.77442	-3.40031	-24.9539	6.86E-26	9.83E-23	44.47466
PRY	-5.33376	-3.12575	-24.2391	2.03E-25	2.75E-22	43.89018
TTY14	-5.0024	-3.29193	-22.4313	3.60E-24	4.61E-21	41.57696
LOC105377224	-4.29668	-3.64546	-21.8935	8.79E-24	1.07E-20	40.61321

Results 1.3: DGE Between Two Time Periods in No Weight Gain Group

Using a similar experimental design (Table 5), DGE analysis revealed 132 genes to be differentially expressed in the NWG group between the two time periods at FDR < 5%. Table 6 shows the top 20 differentially expressed genes. Figure 5 shows clustering. Again, the clustered 18 samples to the left of figure 5 were the same 18 samples clustered together in Figure 3.

Table 5: Experimental Design NWG Group	
Stressor	Name
NWGA	C23ELACXX_6_16.bam
NWGB	C23ELACXX_7_18.bam
NWGA	C23K9ACXX_7_4.bam
NWGB	C24A4ACXX_1_23.bam
NWGB	C2LRRACXX_4_21.bam
NWGA	C2LRRACXX_6_4.bam
NWGB	C6KALANXX_2_21.bam
NWGA	C6KALANXX_5_14.bam
NWGA	C6KALANXX_5_16.bam

Table 6: Top 20 Differentially Expressed Genes Between Two Time Periods in the NWG Group

Gene	logFC	AveExpr	t	P.Value	adj.P.Val	B
PRKY	-5.61738	2.86912	-50.695	4.68E-37	1.16E-32	71.934
ZFY	-9.65348	-0.75559	-44.6511	5.93E-35	7.35E-31	59.57072
DDX3Y	-11.8556	0.944893	-40.9275	1.62E-33	8.91E-30	58.5141
USP9Y	-11.1132	0.261217	-40.922	1.63E-33	8.91E-30	58.08914
TXLNGY	-10.9863	-0.0795	-40.8152	1.80E-33	8.91E-30	57.75026
KDM5D	-11.8785	0.735663	-39.5187	6.11E-33	2.16E-29	57.42217
RPS4Y1	-12.0994	1.321742	-37.8971	2.98E-32	9.23E-29	57.00095
TTY15	-8.38277	-1.75892	-40.1564	3.33E-33	1.38E-29	56.86299
UTY	-10.8069	0.075868	-36.5974	1.11E-31	2.95E-28	55.43817
EIF1AY	-10.4523	-0.37281	-36.5323	1.19E-31	2.95E-28	55.14836
BCORP1	-8.99659	-1.3024	-32.5164	9.46E-30	2.13E-26	52.04657
TMSB4Y	-7.08425	-2.47136	-31.6034	2.74E-29	5.67E-26	50.86261
LOC107987346	-5.64686	-3.21847	-29.2061	5.18E-28	9.88E-25	48.32766
LOC105377223	-5.31605	-3.3698	-29.1167	5.81E-28	1.03E-24	48.10989
XIST	12.38592	0.86669	26.52833	1.81E-26	3.00E-23	46.90736
LINC00278	-6.06902	-2.98413	-26.3366	2.37E-26	3.67E-23	45.77587
LOC107987347	-5.51616	-3.16901	-26.1096	3.25E-26	4.74E-23	45.40508
TTY10	-4.56202	-3.73579	-24.8779	1.91E-25	2.63E-22	43.64278
PRY	-5.94363	-3.08952	-23.9945	7.15E-25	8.85E-22	43.14225
TTY14	-5.14234	-3.44032	-24.1525	5.63E-25	7.34E-22	43.09036

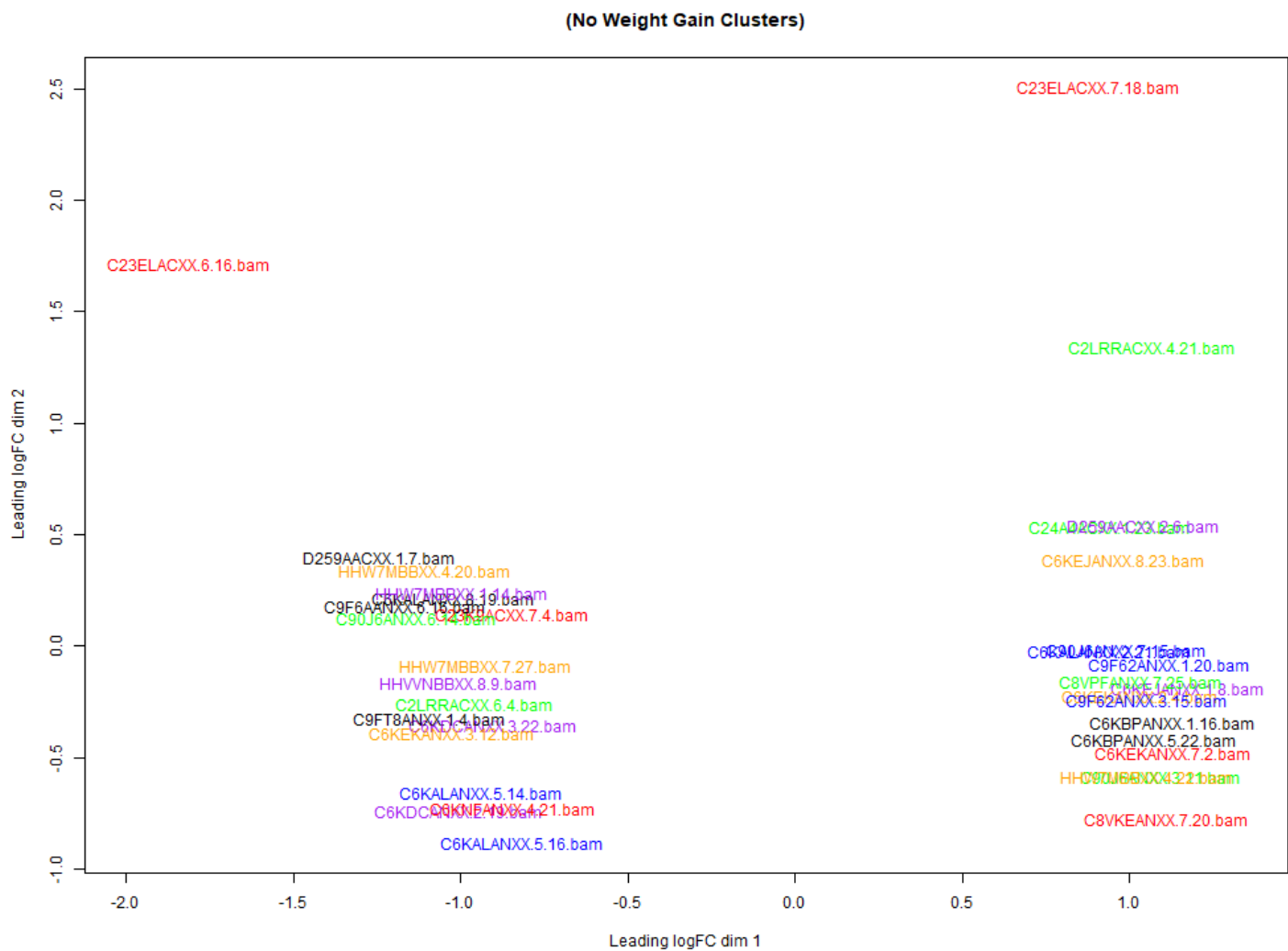


FIGURE 5 18 SAMPLES CLUSTERED TO THE RIGHT AND 18 TO THE LEFT. THE CLUSTERING GROUPS ARE THE SAME AS THAT IN FIGURE 3

Result 1.4 GSEA in Weight Gain Versus No Weight Gain Group Using All GeneSets from MSig Database

GSEA revealed 14 gene sets to be significantly enriched at nominal pvalue < 1% in the Weight Gain Group versus the No Weight Gain Group (Appendix 8) However, the nominal p value does not account for multiple testing. The FDR-q value is a better metric as it corrects for the effect of multiple testing on p-value. 12 gene sets were significantly enriched at FDR q-value< 0.05 in the WG group. A positive ES indicates gene set enrichment at the top of the ranked list (Figure 6); a negative ES indicates gene set enrichment at the bottom of the ranked list, i.e., downregulation (GSEA Guide, 2020). In the analysis results, the enrichment plot provides a graphical view of the enrichment score for a gene set (GSEA Guide, 2020).

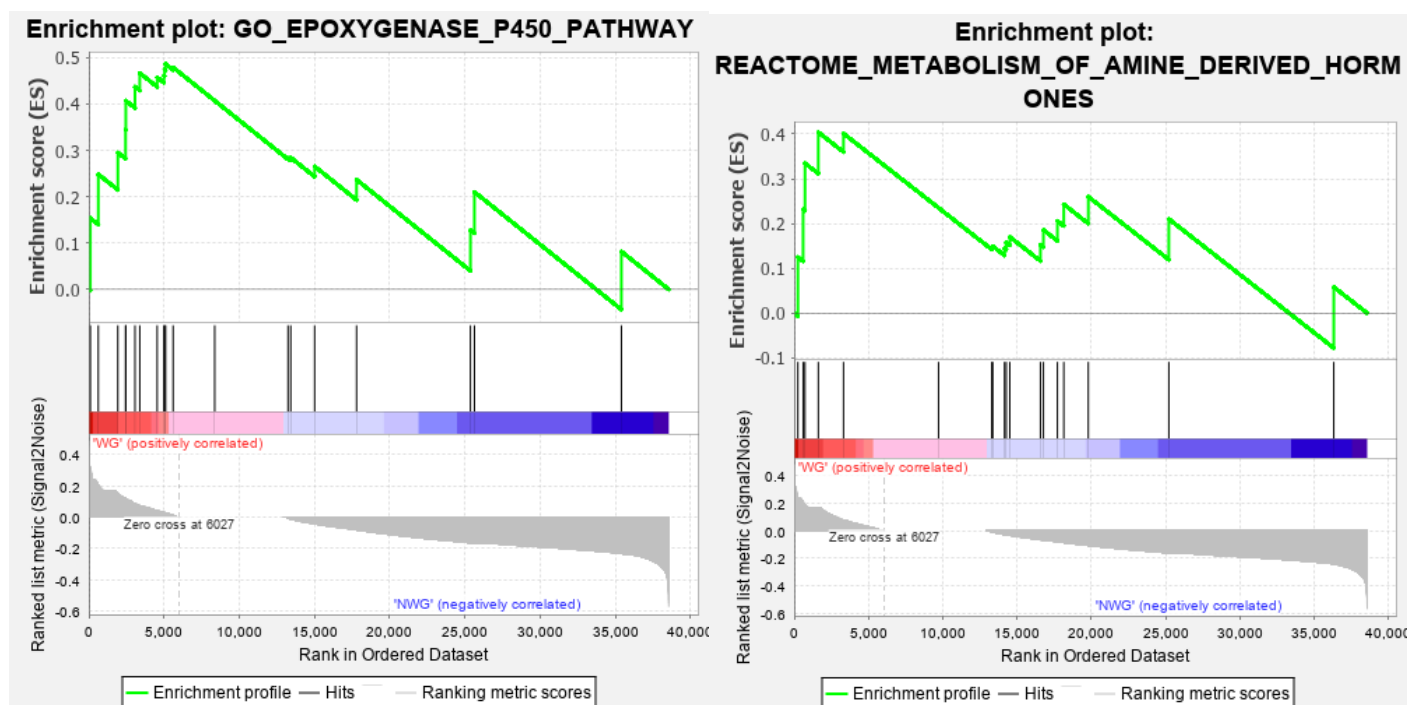
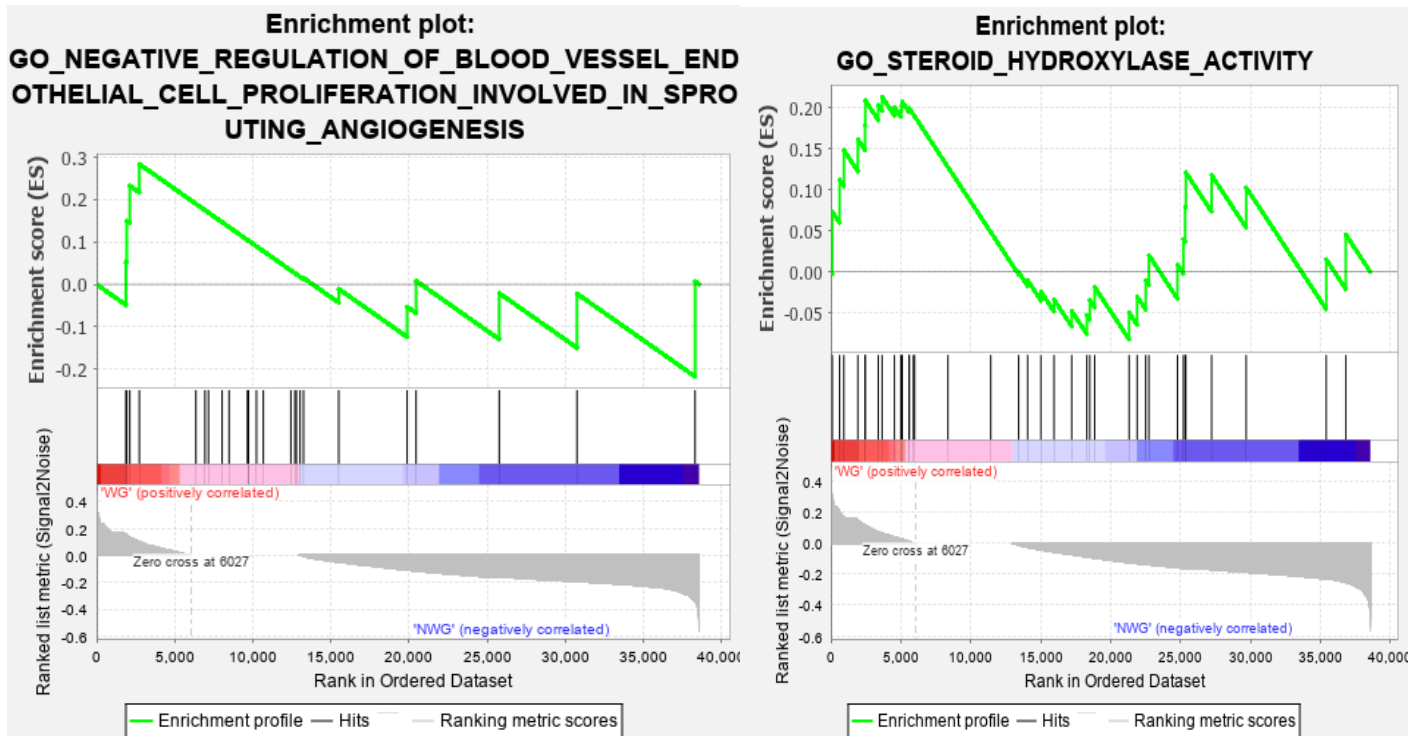


FIGURE 6 ENRICHMENT PLOTS 4 OF THE 12 GENE SETS SIGNIFICANT GENE SETS



Result 1.5: GSEA Using Immunologic Signature Gene Set Only

A second GSEA using the C7: immunologic signatures gene sets (GSEA Downloads, 2020) revealed 82 gene sets were significantly enriched at nominal pvalue < 1% in the No Weight Gain Group (NWG) . i.e., the 82 gene sets were downregulated in the NWG group. Figures 7 and 8 show the enrichment plot and heat map respectively for one of the gene sets. However, the nominal p-value does not correct for multiple testing. The FDR-q value is the better metric to test their significance. None of the gene sets were significantly enriched when the FDR q-value cut off of 0.05 was used (Figure 9).

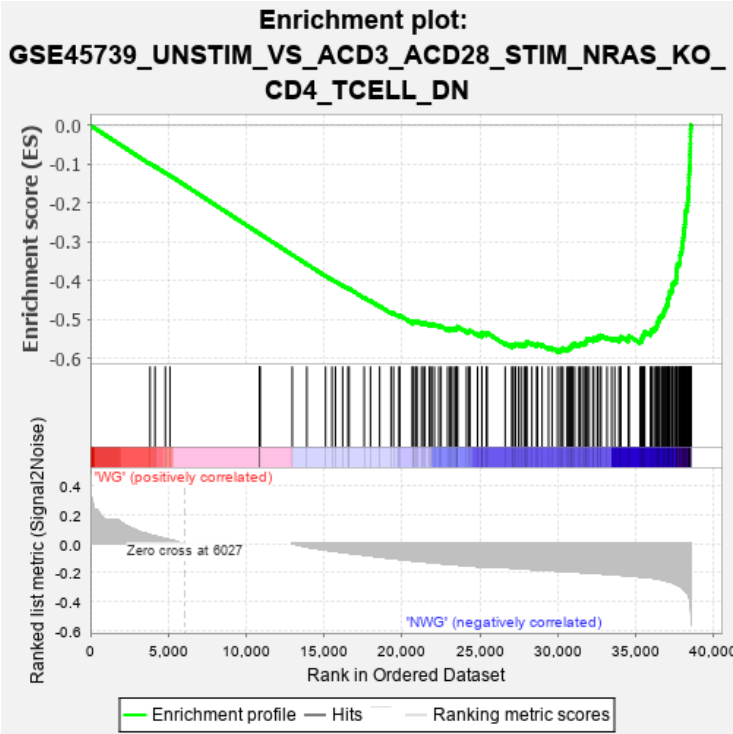


FIGURE 7 AN EXAMLE OF ONE OF THE TOP 10 GENESETS ENRICHED

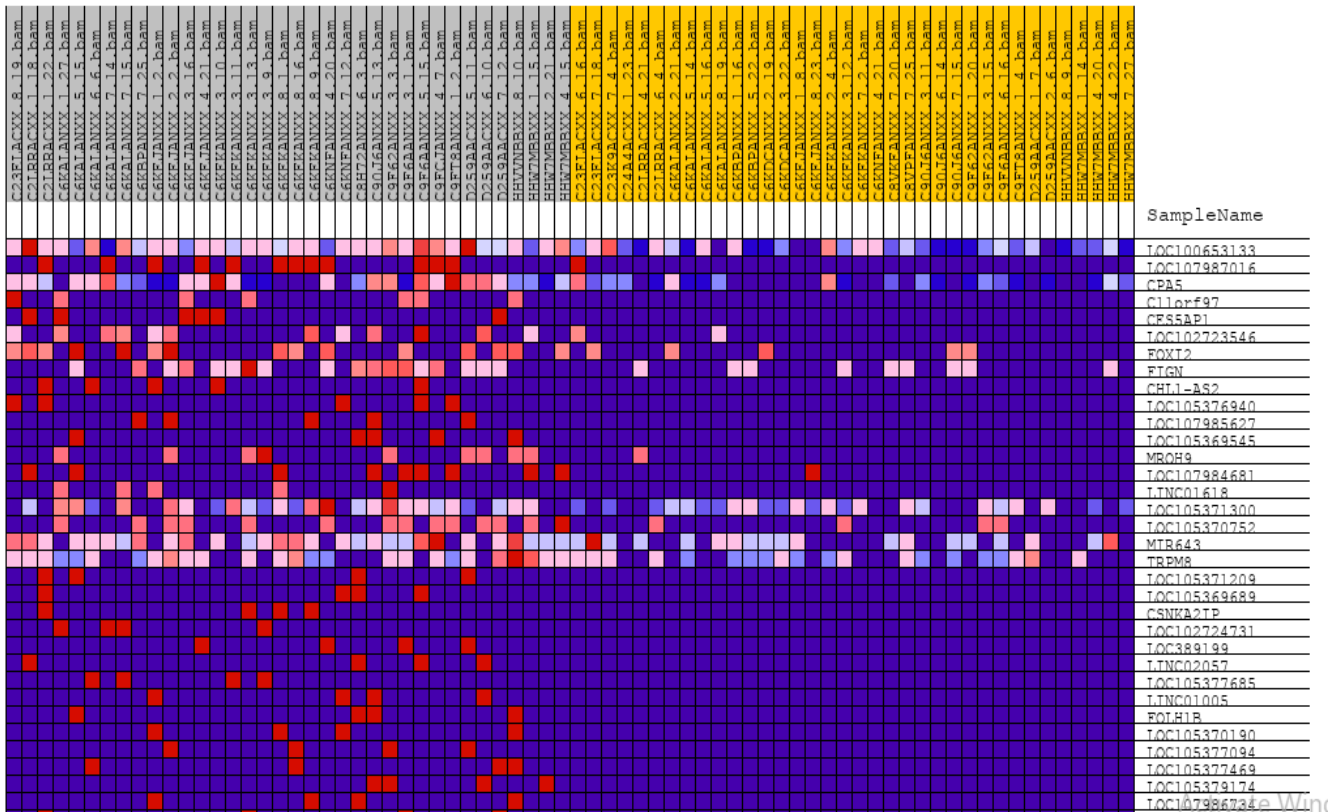


FIGURE 8 HEAT MAP FOR GENES IN THE GENE SET IN ENRICHMENT PLOT IN FIGURE 4. THE GREY CELL SAMPLES ARE THE WEIGHT GAIN GROUP. THE YELLOW CELL SAMPLES ARE THE NO WEIGHT GAIN GROUP. RED COLOUR SHOWS HIGH EXPRESSION

NAME	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val
GSE45739_UNSTIM_VS_ACD3_ACD28_STIM_NR AS_KO_CD4_TCELL_DN	198	-0.58347	-1.569	0.005988	1	0.197
GSE4748_CYANOBACTERIUM_LPSLIKE_VS_LPS_A ND_CYANOBACTERIUM_LPSLIKE_STIM_DC_3H_ UP	183	-0.63418	-1.56431	0.005769	0.863299	0.203
GSE10325_BCELL_VS_MYELOID_UP	188	-0.65436	-1.53533	0.003846	0.97986	0.25
GSE34205_HEALTHY_VS_RSV_INF_INFANT_PBM C_UP	192	-0.59652	-1.51794	0	0.975205	0.283
GSE29614_CTRL_VS_DAY7_TIV_FLU_VACCINE_P BMC_DN	185	-0.61411	-1.51673	0.031621	0.80107	0.284
GSE29614_DAY3_VS_DAY7_TIV_FLU_VACCINE_P BMC_DN	179	-0.60744	-1.50426	0.021526	0.827933	0.301
GSE3982_DC_VS_EFF_MEMORY_CD4_TCELL_DN	195	-0.51219	-1.49258	0.001901	0.824918	0.317
GSE24634_TREG_VS_TCONV_POST_DAY7_IL4_C ONVERSION_UP	197	-0.5608	-1.48581	0.009843	0.801354	0.329
GSE3982_EOSINOPHIL_VS_NKCELL_DN	196	-0.57467	-1.48549	0.005814	0.715225	0.329
GSE3982_EOSINOPHIL_VS_EFF_MEMORY_CD4_T CELL_DN	196	-0.58347	-1.46974	0.015414	0.794603	0.363

FIGURE 9 TOP 10 GENE SETS BASED ON NORMALIZED ENRICHMENT SCORE FOR GSEA FOR THE WG VERSUS NWG GROUPS USING IMMUNOLOGIC GENE SET. ALTHOUGH THE NEGATIVE ENRICHMENT SCORES SUGGEST A HIGHER EXPRESSION LEVEL FOR THESE IMMUNE SYSTEM RELATED GENE SETS IN THE WEIGHT GAIN GROUP, NOTE THE FDR-Q VALUE. THE VERY HIGH Q VALUE TELLS US THE ENRICHMENT IS NOT SIGNIFICANT AFTER CORRECTING FOR MULTIPLE TESTING



















Discussion

The original paper had analyzed the transcriptome of the two groups independently to detect differentially expressed genes before and after medication using the Deseq package (Crespo-Fecorro et al., 2019). Table 8 gives a summary of the numbers of DE genes in the study. Furthermore, the authors used the Gene Reference into function (GeneRIF) database to characterize the function of the genes. The GeneRif database (GeneRif,2020) is maintained by NCBI and provides functional annotations for genes based on all biomedical

research submitted to NCBI. The authors defined “bmi”, “obesity”, “cholesterol” related genes as those genes which included these terms in their description in the database. (Crespo-Fecorro et al., 2019).

1.1 Single-Gene Expression Analysis

Result 1.1 shows 62 differentially expressed genes between the weight gain and the no weight gain group. Table 7 shows the functional annotations from GeneRIF database for the top 18 of these 62 genes. Two of these genes OSBPL10, KLF11, are associated with cholesterol regulation. They are downregulated in the WG group. These genes are important in regulation of metabolic processes to maintain appropriate levels of lipids and cholesterol. Hence, it will be interesting to explore if these two genes were downregulated pre-treatment as well. If they were downregulated pre-treatment, it will point towards inherent dysregulation in the expression of these genes which might be a causal factor for AIWG susceptibility in this group. If the genes were downregulated post-treatment, it is possible that antipsychotic treatment disrupts the functions of these two genes only in the WG group. However, due to lack of information about samples from before and after treatment, further analysis is not possible. One gene, ZBTB32, which is involved in immune cell development, is downregulated in the WG group. It is difficult to compare these results with the original paper as the authors had compared the gene expression profile between the two groups before medication (155 DE genes) and after medication (300 DE genes) separately. Only one gene, KLF11, from the 62 of my analysis, was also present in the 300 genes DE in the WG versus NWG group after medication in the original study.

Table 7: Top 18 DE genes in the WG versus NWG Group				
Gene Name	logFC	Up/Down	adj.P.Val	Function
LOC100653133	1.139959		0.00024	Protein coding; not clear (GeneRif,2020)
LOC105375130	-2.95529		0.000778	uncharacterized(GeneRif,2020)
LOC105370969	0.707861		0.005912	uncharacterized(GeneRif,2020)
LOC105378184	-0.74866		0.009825	uncharacterized(GeneRif,2020)
EZR	-0.20731		0.009825	Cell surface structure adhesion, migration and organization (GeneRif,2020)
JAKMIP1	-0.53665		0.010568	GO annotations related to this gene are RNA Binding and Kinase binding (Database,2020)
BSPRY	-0.83319		0.008332	Epithelial calcium transport regulation (Database,2020)
OSBPL10	-0.5049		0.010853	Go Annotation Related to this is Cholesterol Binding (Database,2020)
FBXO10	-0.28786		0.013038	F-Box family protein; act as protein-ubiquitin ligases(GeneRif,2020)
LOC105376505	1.064324		0.013038	Uncharacterized (GeneRif,2020)
LOC105376504	0.879177		0.013038	Uncharacterized (GeneRif,2020)
LOC105378187	-0.66653		0.013038	Uncharacterized (GeneRif,2020)
PAXBP1-AS1	0.383807		0.019935	Ubiquitous expression in brain and testis (GeneRif,2020)
LOC105376412	-1.19754		0.010853	Uncharacterized; biased expression in lymph node (GeneRif,2020)
KLF11	-0.31516		0.019935	Defects associated with early onset diabetes (GeneRif,2020)
GDF7	-0.77542		0.020263	Defects associated with esophageal adenocarcinoma (GeneRif,2020)
LOC105372793	0.512987		0.019935	Uncharacterized; biased expression in kidney (GeneRif,2020)
ZBTB32	-0.50138		0.022048	Involved in immune cell development (Beaulieu et al., 2020)

In my analysis of the sub-groups of WG only group, 158 genes were differentially expressed, whereas 132 genes were differentially expressed in the NWG group (Supplementary Information 1 gives lists of all genes). The top 20 genes (based on descending adjusted p-value) in WG only sub-group analysis were also differentially expressed in the NWG-only sub-group analysis and vice versa, although as expected, the adjusted p-values differed. In comparison, in the original study 115 genes were DE in the WG group, and 156 in the NWG group; 33 genes were common in these two DE analyses. The remaining 82 genes, related to obesity, cholesterol and immune system were upregulated in the WG group only. Lack of sample information did not make it possible for me to make similar single gene analyses for the NWG and WG groups. For one, the

clustering seen in Figures 3, 4 and 5 may not necessarily represent the clustering based on gene expression patterns before and after medication. It is possible that the clustering is driven by another factor such as sex. The male to female ratio in both groups is 1:1 (9 males and 9 females). Therefore, no confident deductions are possible from sub-group analysis of the WG and NWG groups.

1.2 Pathway Analysis

Subramanian and colleagues note that focusing on single-gene analysis may miss differences in expression patterns of sets of genes involved in various cellular pathways. A 20 percent increase in all genes of a metabolic pathway may give more important information than a 200 percent increase in expression of a single gene (Subramanian et al., 2005). Therefore, GSEA is a more powerful analytical tool as it allows to detect “unifying biological themes” (Subramanian et al., 2005) within long list of statistically significant genes to make sense of our data.

Table 8: Summary of Gene Expression Changes in the Original Paper			
	Weight Gain (Within Group)	No Weight Gain (Within Group)	Weight Gain Versus No Weight Gain Groups
	115 genes were DE; significantly enriched for “bmi”, “cholesterol” or “obesity” related genes based on gene annotations from the GeneRif database; 33 genes also DE in NWG group post treatment; 46 genes related to immune system related genes were significantly enriched post-treatment	156 genes were differentially expressed following three months of treatment; not enriched for “BMI”, “cholesterol,” or “triglyceride”, and had a weak enrichment for “obesity-” related genes according to the GeneRIF database.	
Before Medication			155 genes were differentially expressed; not significantly enriched for “bmi”, “cholesterol” or “obesity” related genes; enrichment in immune related pathways
After Medication			300 genes were differentially expressed’ 75 of these genes were DE before medication as well DE genes were enriched most significantly in the four pathways: neutrophil degranulation (Reactome) ; the immune system pathway (Reactome); graft-vs.-host disease (KEGG); immunoregulatory interactions between a lymphoid and a nonlymphoid cell (Reactome)

Table : Distribution of the 14 Gene Sets Significantly Enriched in WG Group		
Gene Sets relevant to Cardiovascular System	Gene Sets Relevant to Metabolism	Gene Sets Relevant to Central Nervous System
GO_EPOXYGENASE_P450_PATHWAY	REACTOME_METABOLISM_OF_AMINE_DERIVED_HORMONES	GO_SEROTONIN_RECEPTOR_ACTIVITY
GO_NEGATIVE_REGULATION_OF_BLOOD_VESSEL_ENDOTHELIAL_CELL_PROLIFERATION_INVOLVED_IN_SPROUTING_ANGIOGENESIS	GO_STEROID_HYDROXYLASE_ACTIVITY	GO_NEUROPEPTIDE_BINDING
GO_NEGATIVE_REGULATION_OF_CELLULAR_RESPONSE_TO_VASCULAR_ENDOTHELIAL_GROWTH_FACTOR_STIMULUS	GO_METALLOCARBOXYPEPTIDASE_ACTIVITY	GO_DETECTION_OF_LIGHT_STIMULUS_INVOLVED_IN_SENSORIAL_PERCEPTION
GO_POSITIVE_REGULATION_OF_BLOOD_VESSEL_ENDOTHELIAL_CELL_PROLIFERATION_INVOLVED_IN_SPROUTING_ANGIOGENESIS	GO_CELLULAR_GLUCURONIDATION	
GO_GAP_JUNCTION_CHANNEL_ACTIVITY		

GSEA analysis revealed 14 gene sets which were upregulated in the WG group compared to NWG group at a nominal p-value <1%. However, two of these gene sets had an FDR q-value > 5% and hence will not be considered in this analysis. The gene sets fall into three categories. The most interesting one is gene sets implicated in the regulation of the cardiovascular system. The GO_EPOXYGENASE_P450_PATHWAY is important in metabolizing arachidonic acid into eicosanoids which play important role in physiological processes such as inflammation and maintaining the integrity of the vasculature by preventing atherosclerosis and defects in endothelial lining (Theken and Lee, 2007). Furthermore, genetic polymorphisms in this pathway have been associated with differing risks to cardiovascular risk (Theken and Lee, 2007). However, the exact mechanism of this pathway is not clear.

The second gene set, “Go negative regulation of blood vessel endothelial cell proliferation involved in sprouting angiogenesis”, has been associated with the DLL4-Notch pathway, which modulates angiogenesis (creation of blood vessels) and plays an important role in modulation of vasoconstriction and blood flow (You et al., 2013). The gene set GO_GAP_JUNCTION_CHANNEL_ACTIVITY, represents genes involved in the modulation of gap junctions between cells which enable transport of electrical signals, solutes etc. Genes include those involved in electrical conduction in the cardiac cells (Consortium, 2020).

These genesets are enriched in the WG group even after accounting for the confounding variable of anti-psychotic treatment affecting gene expression profile of the same individual before and after treatment. Therefore, these five cardiovascular system associated gene sets have the potential to be signature gene pathways which are upregulated in a subset of schizophrenia patients who gain weight on antipsychotic treatment. Hence, they can help develop more targeted medical treatment plans: patients with this genetic profile who represent a risk group for cardiovascular disease should be prescribed antipsychotic medication which cause less weight gain. Antipsychotic medicines (APs) have unique receptor binding and functional profiles and some of them may have more adverse effects on metabolic profile than others (Kim et al., 2001). APs with higher antagonism at cholinergic receptors, e.g., clozapine, have a more negative effect on cardio-autonomic regulation which contributes to weight gain, whereas those with minimal effects on cholinergic receptors, e.g., amisulpride may even improve cardio-autonomic regulation (Birkhofer et al., 2013).

The pathway analysis in the original paper shows a dominance in immune system related pathways (Table 8). 42 of the 115 DE in WG group before and after medication were immune related. Furthermore, in the WG only group and the WG versus NWG group after medication, four most significantly enriched pathways were immune system related in the WG groups. The synergistic effects of immune responses, such as inflammation, and metabolic abnormalities has received attention in recent years, i.e., they act like positive feedback loops, upregulation of one causes worsening of the other (Sajadieh et al., 2004; de Heredia et al., 2012). If AIWG occurs via a disruption in the immune system pathways, this has implications for treatment plans. Or, it is possible that upregulation in the immune system pathways causes the AIWG. GSEA using the gene sets belonging to immunologic markers (C7) from the SigDB database did not show any significant enrichment. Although, it is important to note that the upregulation in immune related pathways in WG group was present; it did not reach statistical significance. Of course, the confounding variable of lack of proper sample identity may have caused this lack of results.

Limitations and Further Analysis

The major limitation in my analysis was the inadequate sample information. Since the paired samples were not distinguishable, I was unable to compare gene expression profiles within the WG and NGW groups before and after the medication. The authors of the paper did not respond to my email request for information. The clustering in my MDS plots dividing the WG and NWG samples into 2 evenly divided clusters may have been driven by sex, rather than time (before and after medication). Furthermore, as the authors have noted, the WG sample's median age is 4.7 years lower than the NWG group. This could be a confounding variable.

Different antipsychotic drugs have different metabolic profiles. The authors of this study did not fully take this into account. Finally, transcriptomic analysis was carried out on RNA from blood samples, whereas the genes of interest were related to pathways regulating metabolic system and immune system. Samples from tissue fluid from adipose tissue and lymph nodes (regions involved in innate immune response) may give a more accurate picture.

Further analysis using larger samples can be carried out to see if the cardiovascular and metabolism related pathways identified in this report are consistently upregulated in schizophrenia patients who experience AIWG. Finally, snp analysis can be carried out to detect polymorphisms and genetic markers distinguishing patients who are more resilient to AIWG.

Appendix 1: Screenshot of the Wget Command Use to Download the Fastq Files

```
Continuing in background, pid 171216.  
Output will be written to 'wget-log'.  
[sm8847@log-1 sm8847]$ cd P  
-bash: cd: P: No such file or directory  
[sm8847@log-1 sm8847]$ cd FinalProject/  
[sm8847@log-1 FinalProject]$ wget -b http://ftp.sra.ebi.ac.uk/vol1/run/ERR319/ERR3199736/C2LRRACXX\_4\_21\_1.fastq.gz  
Continuing in background, pid 172719.  
Output will be written to 'wget-log.11'.  
[sm8847@log-1 FinalProject]$ wget -b http://ftp.sra.ebi.ac.uk/vol1/run/ERR319/ERR3199736/C2LRRACXX\_4\_21\_2.fastq.gz  
Continuing in background, pid 172957.  
Output will be written to 'wget-log.12'.  
[sm8847@log-1 FinalProject]$ wget -b http://ftp.sra.ebi.ac.uk/vol1/run/ERR319/ERR3199749/C2LRRACXX\_6\_4\_1.fastq.gz  
Continuing in background, pid 176035.  
Output will be written to 'wget-log.13'.  
[sm8847@log-1 FinalProject]$ wget -b http://ftp.sra.ebi.ac.uk/vol1/run/ERR319/ERR3199749/C2LRRACXX\_6\_4\_2.fastq.gz  
Continuing in background, pid 176205.  
Output will be written to 'wget-log.14'.  
[sm8847@log-1 FinalProject]$ wget -b http://ftp.sra.ebi.ac.uk/vol1/run/ERR319/ERR3199729/C6KALANXX\_1\_27\_1.fastq.gz  
Continuing in background, pid 176334.  
Output will be written to 'wget-log.15'.  
[sm8847@log-1 FinalProject]$ wget -b http://ftp.sra.ebi.ac.uk/vol1/run/ERR319/ERR3199729/C6KALANXX\_1\_27\_2.fastq.gz  
Continuing in background, pid 176504.  
Output will be written to 'wget-log.16'.  
[sm8847@log-1 FinalProject]$ wget -b http://ftp.sra.ebi.ac.uk/vol1/run/ERR319/ERR3199703/C6KALANXX\_2\_21\_1.fastq.gz  
Continuing in background, pid 176903.  
Output will be written to 'wget-log.17'.  
[sm8847@log-1 FinalProject]$ wget -b http://ftp.sra.ebi.ac.uk/vol1/run/ERR319/ERR3199703/C6KALANXX\_2\_21\_2.fastq.gz  
Continuing in background, pid 177395.  
Output will be written to 'wget-log.18'.  
[sm8847@log-1 FinalProject]$ wget -b http://ftp.sra.ebi.ac.uk/vol1/run/ERR319/ERR3199746/C6KALANXX\_5\_14\_1.fastq.gz  
Continuing in background, pid 177592.  
Output will be written to 'wget-log.19'.  
[sm8847@log-1 FinalProject]$ wget -b http://ftp.sra.ebi.ac.uk/vol1/run/ERR319/ERR3199746/C6KALANXX\_5\_14\_2.fastq.gz  
Continuing in background, pid 177781.  
Output will be written to 'wget-log.20'.  
[sm8847@log-1 FinalProject]$ █
```

Activate Windows
Go to Settings to activate Windows.

Appendix 2: Fastqc and Index Bash Scripts

```
#!/bin/bash
#SBATCH --job-name=fastqc
#SBATCH --nodes=6
#SBATCH --cpus-per-task=1
#SBATCH --mem 32GB
#SBATCH --time=06:30:00
```

```
module purge
module load fastqc/0.11.8
fastqc *.fastq
echo "job complete"
```

```
#!/bin/bash
#SBATCH --job-name=hisat2_build_index
#SBATCH --nodes=1
#SBATCH --cpus-per-task=16
#SBATCH --mem 32GB
#SBATCH --time=02:30:00
module purge
module load hisat2/intel/2.0.5
hisat2-build -p 16 GRCh38.primary_assembly.genome.fa hg38
echo "job complete"
bamfile=$2

java -jar $PICARD_JAR SortSam \
INPUT=$samfile \
OUTPUT=$bamfile \
SORT_ORDER=coordinate

echo "job complete"
```

Appendix 3: Fastqc Summary Reports

PQS= Number of Poor Quality Sequences; ORS= overrepresented sequences; PTQS= Per Tile Quality Score; PBSQ= Per Base Quality

Score; PSQS= Per Sequence Quality Score

Name	Basic Stats.	PBSQ	PTS Q	PSQS	Sequence Duplication Levels	ORS	Adaptor Content	PQS	comments
C2LRRACXX_1_18_1.fastq	pass	pass	warn	pass	fail	warn	pass	0	no hit for adaptor in over-represented sequences
C2LRRACXX_1_18_2.fastq	pass	pass	pass	pass	warn	warn	pass	0	no hit for adaptor in over-represented sequences
C2LRRACXX_1_22_1.fastq	pass	pass	warn	pass	fail	warn	pass	0	adaptor overrepresented
C2LRRACXX_1_22_2.fastq	pass	pass	pass	pass	fail	warn	pass	0	no hit for adaptor in over-represented sequences
C2LRRACXX_4_21_1.fastq	pass	pass	warn	pass	fail	warn	pass	0	no hit for adaptor in over-represented sequences
C2LRRACXX_4_21_2.fastq	pass	pass	warn	pass	fail	warn	pass	0	no hit for adaptor in over-represented sequences
C2LRRACXX_6_4_1.fastq	pass	pass	warn	pass	fail	warn	pass	0	no hit for adaptor in over-represented sequences
C2LRRACXX_6_4_2.fastq	pass	pass	warn	pass	warn	warn	pass	0	no hit for adaptor in over-represented sequences
C6KALANXX_1_27_1.fastq	pass	pass	pass	pass	warn	warn	pass	0	no hit for adaptor in over-represented sequences
C6KALANXX_1_27_2.fastq	pass	pass	pass	pass	warn	warn	pass	0	no hit for adaptor in over-represented sequences
C6KALANXX_2_21_1.fastq	pass	pass	pass	pass	warn	warn	pass	0	no hit for adaptor in over-represented sequences

Appendix 4: .cls file format for GSEA

```
72 2 1
#WG NWG
WG      WG      WG      WG      WG      WG      WG      WG
```

72 labels in total. The order of the labels should correspond to the order of the sample columns in the normalized data file

Appendix 5: Alignment and Bam Conversion Bash Scripts

```
#!/bin/bash
#SBATCH --job-name=hisat2
#SBATCH --nodes=1
#SBATCH --cpus-per-task=16
#SBATCH --mem 32GB
#SBATCH --time=02:30:00
module purge
module load hisat2/intel/2.0.5
FASTQFILE1=$1
FASTQFILE2=$2
SAMFILE=$3
hisat2 -p 16 -x hg38 -1 $FASTQFILE1 -2 $FASTQFILE2 -S $SAMFILE

echo "job complete"
```

```
#!/bin/bash
#SBATCH --job-name=sort_picard
#SBATCH --nodes=1
#SBATCH --cpus-per-task=16
#SBATCH --mem 32GB
#SBATCH --time=02:30:00

module purge

module load picard/2.17.11

samfile=$1
bamfile=$2

java -jar $PICARD_JAR SortSam \
INPUT=$samfile \
OUTPUT=$bamfile \
SORT_ORDER=coordinate

echo "job complete"
```

Appendix 6: RSubread Code

Starting R

The following command was given on HPC command line in my scratch directory containing the 72 bam files representing the 72 samples: 36 No Weight Gain, 36 Weight Gain
module load r/intel/3.6.0

R

#####

The R console showed up.

I installed the Rsubread package which is part of Bioconductor package.

```
BiocManager::install("Rsubread")
```

```
library(Rsubread)
```

Summarize single-end reads using a user-provided

GTF annotation file:

```
read_counts= featureCounts(files=dir(pattern="bam"),  
annot.ext="hg38.ncbiRefSeq.gtf",  
  isGTFAnnotationFile=TRUE,  
  GTF.featureType="exon",isPairedEnd=TRUE,  
  GTF.attrType="gene_id")
```

#actual data

```
read_counts_data = read_counts$counts
```

```
head(read_counts_data)
```

```
save(read_counts_data_NWG, file = "TotalCountsNCBI.RData")
```

```
annotation= read_counts $annotation
```

```
save(annotation, file = "TotalCountsAnnotationNCBI.RData")
```

Appendix 7: Function in R for Normalization for GSEA

```
```{r}
GSEA.NormalizeRows <-
function(V) {
 # Takes as input a matrix or dataframe with raw read counts. Each row represents reads for
 # a single gene. Each column is reads from a single sample for all genes
 #The function assumed that each row of count data represents reads for a single gene for all
 # samples. The mean and standard deviation for each row was calculated.
 row.mean <- apply(V, MARGIN = 1, FUN = mean) # mean for each row calculated. This is the
 mean counts for a given gene
 row.sd <- apply(V, MARGIN = 1, FUN = sd) # the standard deviation for the read count for
 each gene is calculated
 row.n <- length(V[, 1]) # this is the number of rows in the counts dataframe
 for (i in 1:row.n) {
 if (row.sd[i] == 0) { # if standard deviation is zero, then all read counts should be equated
 to zero
 V[i,] <- 0
 } else { # is sd is not zeor, then for each read in a given row, substract the row mean from it
 and then divide the answer by the standard deviation of that row.
 V[i,] <- (V[i,] - row.mean[i])/row.sd[i] # replaced the raw read count by this normalised
 score
 }
 }
 return(V)
}
```

```{r}
Normalised_Data= GSEA.NormalizeRows(TotalCountsNCBI)
```

```{r}
write.csv(Normalised_Data, file= "Normalised_Data.csv")
```
```

Appendix 8: 62 DE Genes between Weight Gain and No Weight Gain Group

| Gene Name | logFC | AveExpr | t | P.Value | adj.P.Val | B |
|--------------|----------|----------|----------|----------|-----------|----------|
| LOC100653133 | 1.139959 | 0.889136 | 6.424932 | 1.13E-08 | 0.00024 | 9.219582 |
| LOC105375130 | -2.95529 | 0.155804 | -5.99901 | 7.36E-08 | 0.000778 | 7.035418 |
| LOC105370969 | 0.707861 | -0.71036 | 5.374112 | 8.39E-07 | 0.005912 | 4.872404 |
| LOC105378184 | -0.74866 | 3.986051 | -5.12238 | 2.41E-06 | 0.009825 | 4.555812 |
| EZR | -0.20731 | 7.316402 | -5.0705 | 2.79E-06 | 0.009825 | 4.35201 |
| JAKMIP1 | -0.53665 | 2.588948 | -5.01432 | 3.50E-06 | 0.010568 | 4.265436 |
| BSPRY | -0.83319 | -1.08068 | -5.21556 | 1.58E-06 | 0.008332 | 4.152236 |
| OSBPL10 | -0.5049 | 3.04712 | -4.94356 | 4.62E-06 | 0.010853 | 3.992116 |
| FBXO10 | -0.28786 | 2.234394 | -4.83508 | 6.93E-06 | 0.013038 | 3.633936 |
| LOC105376505 | 1.064324 | 3.607926 | 4.83684 | 7.33E-06 | 0.013038 | 3.530843 |
| LOC105376504 | 0.879177 | 1.305126 | 4.812625 | 7.73E-06 | 0.013038 | 3.498164 |
| LOC105378187 | -0.66653 | 3.28388 | -4.80587 | 8.02E-06 | 0.013038 | 3.464987 |
| PAXBP1-AS1 | 0.383807 | 0.991203 | 4.657979 | 1.36E-05 | 0.019935 | 2.967944 |
| LOC105376412 | -1.19754 | -2.08581 | -4.96833 | 4.15E-06 | 0.010853 | 2.814263 |
| DMWD | -0.25705 | 3.550737 | -4.6137 | 1.60E-05 | 0.019935 | 2.7865 |
| KLF11 | -0.31516 | 4.348417 | -4.63016 | 1.51E-05 | 0.019935 | 2.776748 |
| GDF7 | -0.77542 | 1.233915 | -4.5977 | 1.72E-05 | 0.020263 | 2.771819 |
| LOC105372793 | 0.512987 | 0.282663 | 4.621268 | 1.56E-05 | 0.019935 | 2.75058 |
| ZBTB32 | -0.50138 | 1.022701 | -4.55692 | 1.98E-05 | 0.022048 | 2.633407 |
| MIR8071-2 | -0.75116 | 1.354815 | -4.52182 | 2.29E-05 | 0.024185 | 2.5272 |
| TBXA2R | 0.420634 | 1.498556 | 4.478289 | 2.65E-05 | 0.02468 | 2.40063 |
| SIGLEC6 | -0.43233 | 2.138853 | -4.4579 | 2.86E-05 | 0.02468 | 2.333727 |
| ZNF860 | -0.45238 | 1.813868 | -4.42576 | 3.22E-05 | 0.02468 | 2.229838 |
| EML6 | -0.46796 | 1.813528 | -4.42382 | 3.24E-05 | 0.02468 | 2.223401 |
| COCH | -0.73713 | 0.792275 | -4.429 | 3.20E-05 | 0.02468 | 2.189102 |
| SNHG16 | 0.254621 | 5.169603 | 4.473482 | 2.70E-05 | 0.02468 | 2.178978 |
| TRIB3 | -0.24701 | 3.047875 | -4.40682 | 3.45E-05 | 0.02514 | 2.111258 |
| RAPGEF1 | -0.17775 | 7.659714 | -4.4377 | 3.08E-05 | 0.02468 | 2.073237 |
| LOC101927432 | 0.734172 | 0.027186 | 4.388326 | 3.69E-05 | 0.02565 | 1.967799 |
| SYNPO | -0.46576 | 2.481729 | -4.33446 | 4.51E-05 | 0.028625 | 1.903104 |
| COL4A3 | -0.50042 | 1.703778 | -4.3127 | 4.86E-05 | 0.028666 | 1.856845 |
| TNFRSF13B | -0.60332 | 1.532072 | -4.29379 | 5.23E-05 | 0.02989 | 1.790012 |
| KNSTRN | 0.293356 | 2.444839 | 4.266192 | 5.74E-05 | 0.031968 | 1.679537 |
| FAN1 | 0.254186 | 4.89315 | 4.311371 | 4.88E-05 | 0.028666 | 1.629299 |
| LOC105379504 | 0.481891 | -0.06025 | 4.253471 | 6.01E-05 | 0.032607 | 1.541184 |

| | | | | | | |
|--------------|----------|----------|----------|----------|----------|----------|
| OSBPL1A | 0.331453 | 3.085822 | 4.196103 | 7.39E-05 | 0.037085 | 1.397874 |
| FA2H | -0.77014 | -1.78401 | -4.38308 | 3.76E-05 | 0.02565 | 1.389897 |
| SYT11 | -0.26657 | 4.895657 | -4.21985 | 6.79E-05 | 0.035837 | 1.317483 |
| LOC102724971 | -0.63496 | -0.12491 | -4.18381 | 7.72E-05 | 0.037085 | 1.315738 |
| BHLHE41 | -0.50497 | 1.135219 | -4.13256 | 9.26E-05 | 0.041648 | 1.272895 |
| LARGE2 | -0.33285 | 2.746368 | -4.14457 | 8.87E-05 | 0.041423 | 1.259714 |
| LOC105374209 | 0.606629 | -2.07407 | 4.36368 | 4.04E-05 | 0.02667 | 1.22437 |
| CPA5 | 1.81665 | -1.78299 | 4.333904 | 4.60E-05 | 0.028625 | 1.215348 |
| EBF1 | -0.45369 | 1.843499 | -4.10909 | 0.000101 | 0.044308 | 1.196558 |
| GRIK1 | 0.45477 | -0.20855 | 4.140165 | 9.01E-05 | 0.041423 | 1.177739 |
| ILDR1 | -0.69322 | -1.24521 | -4.21318 | 6.95E-05 | 0.035837 | 1.130256 |
| C11orf80 | -0.22926 | 2.282732 | -4.07493 | 0.000113 | 0.047031 | 1.066717 |
| WWOX | -0.29384 | 1.030262 | -4.05101 | 0.000123 | 0.047428 | 1.016674 |
| GRIK4 | -1.72647 | -1.30388 | -4.1912 | 7.70E-05 | 0.037085 | 0.978399 |
| LOC105372321 | 0.69769 | -0.50591 | 4.064302 | 0.000118 | 0.047428 | 0.902299 |
| SREBF1 | 0.255464 | 5.792802 | 4.096305 | 0.000105 | 0.045456 | 0.879108 |
| COL4A4 | -0.4854 | 1.624654 | -4.00637 | 0.000144 | 0.049625 | 0.877947 |
| LAMA5 | -0.46899 | 2.282652 | -4.01065 | 0.000143 | 0.049625 | 0.858931 |
| LOC105371453 | -1.08281 | -3.01236 | -4.42144 | 3.27E-05 | 0.02468 | 0.857188 |
| PYHIN1 | -0.29662 | 4.958068 | -4.06332 | 0.000119 | 0.047428 | 0.787496 |
| NIPA1 | -0.19363 | 4.047231 | -4.02653 | 0.000134 | 0.048993 | 0.741789 |
| SIX5 | -0.48217 | -0.95292 | -4.04204 | 0.000127 | 0.047902 | 0.741 |
| SH2D2A | -0.38244 | 4.623593 | -4.04241 | 0.000129 | 0.047902 | 0.732688 |
| NODAL | 0.555341 | -1.39472 | 4.051958 | 0.000123 | 0.047428 | 0.652478 |
| TRPV6 | 0.613667 | -1.72209 | 4.088065 | 0.000108 | 0.045807 | 0.640821 |
| SEL1L3 | -0.25396 | 6.289872 | -4.0088 | 0.000143 | 0.049625 | 0.593369 |
| APOBEC3G | -0.25024 | 5.767228 | -4.00401 | 0.000145 | 0.049625 | 0.574897 |

Appendix 9: 14 Gene Sets Upregulated in Weight Gain Group

| NAME | SIZE | ES | NES | nominal p-val | FDR q-val | FWER p-val |
|---|------|-------|-------|---------------|-----------|------------|
| GO_EPOXYGENASE_P450_PATHWAY | 20 | 0.486 | 2.376 | 0 | 0.0000 | 0 |
| REACTOME_METABOLISM_OF_AMINE_DERIVED_HORMONES | 18 | 0.404 | 2.321 | 0 | 0.0000 | 0 |
| GO_NEGATIVE_REGULATION_OF_BLOOD_VESSEL_ENDOTHELIAL_CELL_PROLIFERATION_INVOLVED_IN_SPROUTING_ANGIOGENESIS | 25 | 0.284 | 1.897 | 0 | 0.0000 | 0 |
| GO_SEROTONIN_RECEPTOR_ACTIVITY | 32 | 0.318 | 1.505 | 0 | 0.0033 | 0.2 |
| GO_CELLULAR_GLUCURONIDATION | 18 | 0.377 | 1.500 | 0 | 0.0026 | 0.2 |
| GO_NEUROPEPTIDE_BINDING | 24 | 0.303 | 1.442 | 0 | 0.0022 | 0.2 |
| GO_METALLOCARBOXYPEPTIDASE_ACTIVITY | 29 | 0.285 | 1.400 | 0 | 0.0019 | 0.2 |
| GO_NEGATIVE_REGULATION_OF_CELLULAR_RESPONSE_TO_VASCULAR_ENDOTHELIAL_GROWTH_FACTOR_STIMULUS | 15 | 0.332 | 1.342 | 0 | 0.0072 | 0.4 |
| GO_POSITIVE_REGULATION_OF_BLOOD_VESSEL_ENDOTHELIAL_CELL_PROLIFERATION_INVOLVED_IN_SPROUTING_ANGIOGENESIS | 22 | 0.278 | 1.247 | 0 | 0.0218 | 0.8 |
| GO_DETECTION_OF_LIGHT_STIMULUS_INVOLVED_IN_SENSORY_PERCEPTION | 18 | 0.260 | 1.202 | 0 | 0.0267 | 1 |
| GO_GAP_JUNCTION_CHANNEL_ACTIVITY
http://www.informatics.jax.org/go/term/GO:0005243 | 16 | 0.235 | 1.132 | 0 | 0.0578 | 1 |
| GO_STEROID_HYDROXYLASE_ACTIVITY | 38 | 0.213 | 1.125 | 0 | 0.0555 | 1 |
| GO_NEGATIVE_REGULATION_OF_VOLTAGE_GATED_POTASSIUM_CHANNEL_ACTIVITY | 21 | 0.243 | 1.076 | 0 | 0.0830 | 1 |
| GO_CELL_PROLIFERATION_INVOLVED_IN_HEART_MORPHOGENESIS | 18 | 0.313 | 1.035 | 0 | 0.1100 | 1 |

Bibliography

1. Birkhofer, A., Geissendoerfer, J., Alger, P., Mueller, A., Rentrop, M., Strubel, T., Leucht, S., Förstl, H., Bär, K.J., and Schmidt, G. (2013). The deceleration capacity - a new measure of heart rate variability evaluated in patients with schizophrenia and antipsychotic treatment. *Eur. Psychiatry* 28, 81–86.
2. Beaulieu AM., Madera, S., Sun, J.C., (2015). "Molecular Programming of Immunological Memory in Natural Killer Cells". *Advances in Experimental Medicine and Biology*. 850: 81–91. doi:10.1007/978-3-319-15774-0_7. PMID 26324348
3. Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szczesniak, M., Gaffney, D., Elo, L., Zhang, X. and Mortazavi, A., 2016. A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1).
4. Consortium, G., 2020. Amigo 2: Term Details For "Gap Junction Channel Activity" (GO:0005243). [online] Amigo.geneontology.org. Available at: <<http://amigo.geneontology.org/amigo/term/GO:0005243>> [Accessed 14 May 2020].
5. Ccb.jhu.edu. 2020. *HISAT2*. [online] Available at: <<https://ccb.jhu.edu/software/hisat2/manual.shtml>> [Accessed 06 May 2020].
6. Crespo-Facorro, B., Prieto, C. and Sainz, J., 2019. Altered gene expression in antipsychotic-induced weight gain. *npj Schizophrenia*, 5(1).
7. Database, G., 2020. *JAKMIP1 Gene - Genecards | JKIP1 Protein | JKIP1 Antibody*. [online] Genecards.org. Available at: <<https://www.genecards.org/cgi-bin/carddisp.pl?gene=JAKMIP1>> [Accessed 14 May 2020].
8. de Heredia, F. P., Gomez-Martinez, S. & Marcos, A. Obesity, inflammation and the

immune system. *Proc. Nutr. Soc.* 71, 332–338 (2012).

9. Gsea-msigdb.org. 2020. *GSEA / Downloads*. [online] Available at: <<https://www.gsea-msigdb.org/gsea/downloads.jsp>> [Accessed 13 May 2020].
10. Gsea-msigdb.org. 2020. *GSEA / Msigdb*. [online] Available at: <<https://www.gsea-msigdb.org/gsea/msigdb/index.jsp>> [Accessed 06 May 2020].
11. Gsea-msigdb.org. 2020. *GSEA User Guide*. [online] Available at: <<https://www.gsea-msigdb.org/gsea/doc/GSEAUUserGuideFrame.html>> [Accessed 13 May 2020].
12. GitHub. 2020. Hxin/Ograph. [online] Available at: <<https://github.com/hxin/ograph/blob/master/R/GSEA.1.0.R>> [Accessed 13 May 2020].
13. Hennekens, C.H., Hennekens, A.R., Hollar, D., and Casey, D.E. (2005). Schizophrenia and increased risks of cardiovascular disease. *Am. Heart J.* 150, 1115–1121.
14. Hgdownload.soe.ucsc.edu. 2020. UCSC. [online] Available at: <<http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/genes/>> [Accessed 13 May 2020].
15. Khandaker, G.M. (2015). Europe PMC Funders Group Inflammation and immunity in schizophrenia : implications for pathophysiology and treatment. 2, 258–270.
16. Kim, D.J., Kim, W., Yoon, S.J., Go, H.J., Choi, B.M., Jun, T.Y., and Kim, Y.K. (2001). Effect of risperidone on serum cytokines. *Int. J. Neurosci.* 111, 11–19.
17. Leucht, S., Burkard, T., Henderson, J., Maj, M., and Sartorius, N. (2007). Physical illness and schizophrenia: A review of the literature. *Acta Psychiatr. Scand.* 116, 317–333.
18. Leung, J.Y.T., Barr, A.M., Procyshyn, R.M., Honer, W.G., and Pang, C.C.Y. (2012). Cardiovascular side-effects of antipsychotic drugs: The role of the autonomic nervous system. *Pharmacol. Ther.* 135, 113–122.
19. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P. and Mesirov, J., 2011. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12), pp.1739-1740.

20. Mootha, V., Lindgren, C., Eriksson, K., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M., Patterson, N., Mesirov, J., Golub, T., Tamayo, P., Spiegelman, B., Lander, E., Hirschhorn, J., Altshuler, D. and Groop, L., 2003. PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34(3), pp.267-273.
21. Ncbi.nlm.nih.gov. 2020. Generif. [online] Available at: <<https://www.ncbi.nlm.nih.gov/gene/about-generif>> [Accessed 14 May 2020].
22. Ritchie, M., Phipson, B., Wu, D., Hu, Y., Law, C., Shi, W. and Smyth, G., 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), pp.e47-e47.
23. Robinson, M. and Oshlack, A., 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3), p.R25.
24. Sajadieh, A., Nielsen, O.W., Rasmussen, V., Hein, H.O., Abedini, S., and Hansen, J.F. (2004). Increased heart rate and reduced heart-rate variability are associated with subclinical inflammation in middle-aged and elderly subjects with no apparent heart disease. *Eur. Heart J.* 25, 363–370.
25. Smyth, G., Ritchie, M., Thorne, N. and Wettenhall, J., 2020. *Limma User Guide*. [online] Bioconductor.org. Available at: <<https://www.bioconductor.org/packages/release/bioc/vignettes/limma/inst/doc/usersguide.pdf>> [Accessed 13 May 2020].
26. Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E. and Mesirov, J., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43), pp.15545-15550.

27. Trace.ncbi.nlm.nih.gov. 2020. SRA Run Selector. [online] Available at:
<https://trace.ncbi.nlm.nih.gov/Traces/study/?acc=ERP114104&o=acc_s%3Aa> [Accessed 14 May 2020].
28. Theken, K. and Lee, C., 2007. Genetic variation in the cytochrome P450 epoxigenase pathway and cardiovascular disease risk. *Pharmacogenomics*, 8(10), pp.1369-1383.
29. You, C., Erol Sandalcioğlu, I., Dammann, P., Felbor, U., Sure, U. and Zhu, Y., 2013. Loss of CCM3 impairs DLL4-Notch signalling: implication in endothelial angiogenesis and in inherited cerebral cavernous malformations. *Journal of Cellular and Molecular Medicine*, 17(3), pp.407-418.