**Distribution Agreement**

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Hongyue Chen Date

# Exploring the Genetics and Biological Pathways of Obesity through Computational Biology and Statistical Approaches

By

Hongyue Chen

Master of Public Health

Department of Biostatistics and Bioinformatics

---

Steve Qin, PhD

(Thesis Advisor)

---

Yijuan Hu, PhD

(Reader)

# Exploring the Genetics and Biological Pathways of Obesity through Computational Biology and Statistical Approaches

By

Hongyue Chen

B.S., Rensselaer Polytechnic Institute, 2021

Thesis Committee Chair: Steve Qin, PhD

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of Master of Public Health

in Department of Biostatistics and Bioinformatics

2023

**Abstract**

Exploring the Genetics and Biological Pathways of Obesity through Computational Biology and Statistical Approaches

By Hongyue Chen

**Background:** Obesity is a chronic and complex disease that has become one of the most serious public health concerns of our time. One significant factor contributing to the development of obesity is genetics, with the fat mass and obesity-associated (FTO) gene considered to carry the highest risk of developing the obesity phenotype. Specifically, the rs1421085 variant of the FTO gene has been shown to have the strongest association with obesity.

**Objectives:** In this study, we aim to investigate the rs1421085 associated genotype within 49 specific tissues provided by the GTEx project to identify potential obesity-associated genes, tissues, and biological mechanisms. This research will provide valuable insights into the complex genetic and molecular mechanisms underlying obesity and may lead to new approaches for preventing and treating this growing public health issue.

**Methods:** We utilized various data resources in this study, including 49 raw tissue-specific datasets from the open-access GTEx Analysis version 8, as well as curated gene sets from BioCarta and KEGG subsets of canonical pathways. We performed differential gene expression analysis on the normalized TPM gene data to discover quantitative changes in expression levels between groups with rs1421085. Additionally, we employed Gene Set Enrichment Analysis (GSEA) to determine whether a previously defined set of genes showed statistically significant, concordant differences between phenotypes. To carry out these analyses, we utilized the PLINK 1.07 and RStudio Version 4.1.2 software tools.

**Results:** The "C" allele of the rs1421085 gene variant is a risk allele for obesity, according to GTEx statistics. The small intestine was found to be the most rs1421085-associated tissue, with increased transit time potentially due to effective nutrient absorption and decreased satiety signals. The valid genes associated with rs1421085 were TBC1D3E, CCL3L3, CSF3, CXCL3, and IL6, with evidence from various studies. Pathway analysis revealed cytokine-cytokine receptor interaction and IL-17 signaling pathway as associated pathways, potentially linking chronic inflammation with obesity.

**Conclusions:** In the studies, we explored candidate genes, biological pathways, and tissues associated with obesity through computational biology and statistical methods. The research given the ideas of obesity is inflammation associated, which give researchers an insight in the field of future obesity study.

# Exploring the Genetics and Biological Pathways of Obesity through Computational Biology and Statistical Approaches

By

Hongyue Chen

B.S., Rensselaer Polytechnic Institute, 2021

Thesis Committee Chair: Steve Qin, PhD

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of Master of Public Health
in Department of Biostatistics and Bioinformatics
2023

# 1.Introduction:

Obesity is a chronic and complex disease that results from various factors with the presence of abnormal or excessive accumulation of body fat [1], [2]. According to the standards from the World Health Organization (WHO), an individual who is 20 years of age or older with the body mass index (BMI) greater than 25 (kg/m^2) is classified as an obese patient [2]. Nowadays, obesity has reached pandemic levels, with the number of global cases nearly tripling from 1975 to 2016 [3]. In 2022, there were approximately 650 million adults, 340 million adolescents, and 39 million children who were considered obese. The worldwide prevalence of obesity continues to increase, despite efforts to combat this health issue [4]. As the number of people affected by obesity continues to increase, so does the associated financial burden on healthcare systems. For instance, in the United States, where obesity affects around 100.1 million adults (41.9%) and 14.7 million children (19.7%), the annual healthcare costs are estimated to be $147 billion by 2023 [5].

Obesity is regarded as one of the serious public health concerns not only for the enormous health problems and healthcare costs it generates but also for numerous comorbidities associated with itself. Individuals with obesity have a higher risk of type 2 diabetes, pre-diabetes, hypertension, and cardiovascular disease compared to those who with lower BMI [6]. Besides these common comorbidities, obesity is also associated with coronavirus disease 2019 (COVID-19), a disease caused by SARS-CoV-2. Compared with normal weight patients, obese group had higher risk of COVID-19 related hospitalizations and death in the previous systematic review and meta-analysis [7]. In addition, there seems to be a correlation between severe COVID-19 outcomes and excessive visceral adiposity [8]. Given the numerous health problems associated with obesity, scientists have conducted extensive researches to understand causation and develop prevention methods to the disease. Various studies on obesity included lifestyle factors, behavior factors, educational factors, environmental factors [9], and genetic factors [10] have made progress on prevention treatments. Structured lifestyle,

realistic weight-loss goal, food monitoring, healthy home food environment and higher education have been proved to be plausible ways to reduce the risk of developing obesity [9]. Focusing on genetic perspective, obesity is considered as genetic associated traits. Around 250 specific genetic variants associated with obesity are identified from previous genome-wide association studies (GWAS) after the Human Genome project. The fat mass and obesity associated (FTO) gene located on chromosome 16 is considered the most significant gene associated with obesity, carrying the highest risk of developing the obesity phenotype [10]. The FTO genetic variant single nucleotide polymorphisms (SNP) rs1421085 is identified to be most associated SNPs with obesity [11]. SNPs are common genetic variants that occur in the DNA sequence between genes. These variants can serve as biological markers for locating genes that are associated with particular diseases, especially useful in complex genetic studies [12]. By examining rs1421085, more potential obesity associated genes and biological pathways can be identified.

Gene expression is an indicator of the extent to which a gene is functioning, reflecting the level of genetic information being expressed. With the development of statistical models and computational biology techniques conducting on gene expression data, genomic traits and biological mechanisms underneath are determined [13]. The objective of this paper is to investigate rs1421085 associated genes expression within 49 specific tissues provided by GTEx project, unitizing Differential Gene Expression Analysis, and Gene Set Enrichment Analysis techniques in order determine potential tissues and biological pathways related to obesity.

# 2.Method:

## 2.1 Sample resources

The Genotype-tissue Expression (GTEx) project is a comprehensive public resource to that enabled the study of tissue-specific gene expression and regulation [14]. The GTEx donor tissues are collected without consideration for any disease state and are a representation of the

overall US population. In the GTEx project, an extensive amount of molecular data is collected from various tissues, providing a glimpse into the genetic and genomic diversity of individuals and tissues in a non-diseased, healthy condition. [15] The 49 raw tissue-specific data utilized in this paper was derived from the open-access data of GTEx Analysis version 8, including tissue-specific annotation file, phenotype annotation file, RNA Sequencing (RNA-Seq) data, and gene count by tissues. The 49 biospeciemens were: Adipose-Subcutaneous, Adipose-Visceral (Omentum), Adrenal Gland, Artery-Aorta, Artery-Coronary, Artery-Tibial, Brain-Amygdala, Brain-Anterior cingulate cortex (BA24), Brain-Caudate (basal ganglia), Brain-Cerebellar Hemisphere, Brain-Cerebellum, Brain-Cortex, Brain-Frontal Cortex (BA9), Brain-Hippocampus, Brain-Hypothalamus, Brain-Nucleus accumbent (basal ganglia), Brain-Putamen (basal ganglia), Brain-Spinal cord (cervical c-1), Brain-Substantia nigra, Breast-Mammary Tissue, Cells-Cultured fibroblasts, Cells-EBV-transformed lymphocytes, Colon-Sigmoid, Colon-Transverse, Esophagus-Gastroesophageal Junction, Esophagus-Mucosa, Esophagus-Muscularis, Heart-Atrial Appendage, Heart-Left Ventricle, Kidney-Cortex, Liver, Lung, Minor Salivary Gland, Muscle-Skeletal, Nerve-Tibial, Ovary, Pancreas, Pituitary, Prostate, Skin-Not Sun Exposed (Suprapubic), Skin-Sun Exposed (Lower leg), Small Intestine-Terminal Ileum, Spleen, Stomach, Testis, Thyroid, Uterus, Vagina, and Whole Blood.

## 2.2 Differential Gene Expression Analysis (DE analysis)

By exploring the distinctions of gene expression between disease and disease-free states, it might be possible to identify candidate genes associated with disease. Differential gene expression analysis toke the normalized read count data and performed statistical analysis to discover quantitative changes in expression levels between experimental groups [16].

In the DE analysis, the parametric Wald test was conducted through DESeq2. For the Wald test analysis between genetic groups, the gene count data was used, which was formatted into two matrices. The first matrix included gene expression data from gene count, while the

other contained allele genotype information. The step includesutilizes shrinkage estimation for dispersions and fold changes since it was difficult to accurately estimate within-group variance with a small number of replicates. To estimate the dispersion value for each gene, DESeq2 employs a model fit procedure, which required biological replicates for each experimental condition to produce reliable results. In cases duplicates were removed, DESeq2 would estimate the dispersion value using the tissue samples from different conditions as replicates [17]. The significance threshold was set for 0.05, and the genes significantly associated with rs1421085 were then selected based on their p-values.

## 2.3 Gene Set Enrichment Analysis (GSEA):

Besides differential expression analysis on gene sets, gene set enrichment analysis (GSEA) was able to interpret the results to the underlying biological processes. It could detect subtle changes in gene expression through analysis gene sets as a group compared with traditional single gene exploration [18]. The reference gene sets were established based on previous biological studies, categorizing in aspects such as biological pathways, cellular component, chromosome locations, and molecular functions [18]. The reference group of gene sets used in this paper was BioCarta subset (292 gene sets) and KEGG subset (186 gene sets) under the Canonical Pathways.

Before the genetic statistical analysis, determine the rank order of all members of the gene set within the ranked dataset followed by ranking significant gene expression data based on fold changes. Calculate the enrichment score (ES) to quantify the degrees to which the given canonical pathway associated gene set was overrepresented at the extreme of the ranked list of genes, assuming the rank was randomly distributed [18]. Then estimate the significance level of ES through permutation test, and adjust for multiple hypothesis testing based on false discovery rate (FDR) [19]. The significance level of the ES played a crucial role in determining the degree to which the heat map reflects the significance of the pathway and the tissues that were strongly

associated with obesity. After conducting pathway analysis, pathways of interest were visualized. Pathway visualization helped to gain a more comprehensive understanding of the interactions and signaling pathways involved in the biological mechanism under investigation. In this paper, the permutation were set as 1000, and pathways with p-values lower than 0.05 were considered as to be significant.

## Analysis Tools:

The GWAS part was performed through PLINK 1.07 [20] on Linux environment, and the gene expression data were analyzed and visualized using RStudio [21]. The Packages used through the analyses were: biomaRT, data.table, DESeq2, dplyr, ggplot2, readr, org.Hs.eg.db, stringr, and tidyr.

# 3. Results:

## 3.1 GTEx Statistics

The GTEx portal provided comprehensive statistics on genes, SNPs, RNA-Seq, quantitative trait locus (QTL), and tissue histology, serving as the valuable resources for investigation on the complex relationship between genetics and phenotype traits [14]. As not all tissues had available RNA-Seq data, all subsequent analysis were conducted on those tissue samples with valid RNA-Seq information (Table 3.1.1).

According to the eQTL violin plot available on the GTEx portal, the risk allele associated with the rs1421085 appeared to be the "C" allele, indicating individuals with C alleles on rs1421085 were more likely to be affected by obesity (Figure 3.1.1). The statistical results obtained were consistent with the information presented on the SNPedia, which suggested that the CC genotype is associated with a 1.7- fold increase in obesity risk, while the CT genotype is linked with a 1.3-fold increase in risk. In contrast, individuals with the TT genotype were reported to have a normal obesity risk [22].

The GTEx portal also provided the multi-tissue eQTL Comparison statistical expression results with rs1421085 on FTO gene. Cumulative results from multiple tissues to identify eQTLs had been shown to improve accuracy from enhancing statistical power and reducing the risk of type I and type II error compared to examining tissues individually. This approach was similar to the concept of meta-analysis, involving combination of results from multiple GWAS to improve the statistical power to detect associations [23]. In the eQTL analysis, differentially expressed

Table 3.1.1 Bulk Tissue RNA-Seq Sample Info

| Tissue | #Genotyped RNASeq samples |
|---|---|
| Adipose - Subcutaneous | 581 |
| Adipose - Visceral (Omentum) | 469 |
| Adrenal Gland | 233 |
| Artery - Aorta | 387 |
| Artery - Coronary | 213 |
| Artery - Tibial | 584 |
| Brain - Amygdala | 129 |
| Brain - Anterior cingulate cortex (BA24) | 147 |
| Brain - Caudate (basal ganglia) | 194 |
| Brain - Cerebellar Hemisphere | 175 |
| Brain - Cerebellum | 209 |
| Brain - Cortex | 205 |
| Brain - Frontal Cortex (BA9) | 175 |
| Brain - Hippocampus | 165 |
| Brain - Hypothalamus | 170 |
| Brain - Nucleus accumbens (basal ganglia) | 202 |
| Brain - Putamen (basal ganglia) | 170 |
| Brain - Spinal cord (cervical c-1) | 126 |
| Brain - Substantia nigra | 114 |
| Breast - Mammary Tissue | 396 |
| Cells - Cultured fibroblasts | 483 |
| Cells - EBV-transformed lymphocytes | 147 |
| Colon - Sigmoid | 318 |
| Colon - Transverse | 368 |
| Esophagus - Gastroesophageal Junction | 330 |
| Esophagus - Mucosa | 497 |
| Esophagus - Muscularis | 465 |
| Heart - Atrial Appendage | 372 |
| Heart - Left Ventricle | 386 |
| Kidney - Cortex | 73 |
| Liver | 208 |
| Lung | 515 |
| Minor Salivary Gland | 144 |
| Muscle - Skeletal | 706 |
| Nerve - Tibial | 532 |
| Ovary | 167 |
| Pancreas | 305 |
| Pituitary | 237 |
| Prostate | 221 |
| Skin - Not Sun Exposed (Suprapubic) | 517 |
| Skin - Sun Exposed (Lower leg) | 605 |
| Small Intestine - Terminal Ileum | 174 |
| Spleen | 227 |
| Stomach | 324 |
| Testis | 322 |
| Thyroid | 574 |
| Uterus | 129 |
| Vagina | 141 |
| Whole Blood | 670 |

**Figure 3.1.1 rs1421085 eQTL on FTO (muscle-skeletal tissue)**

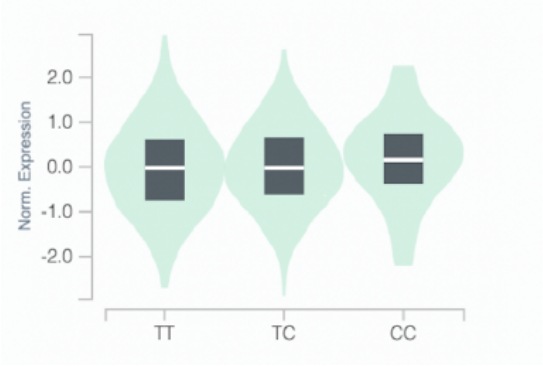
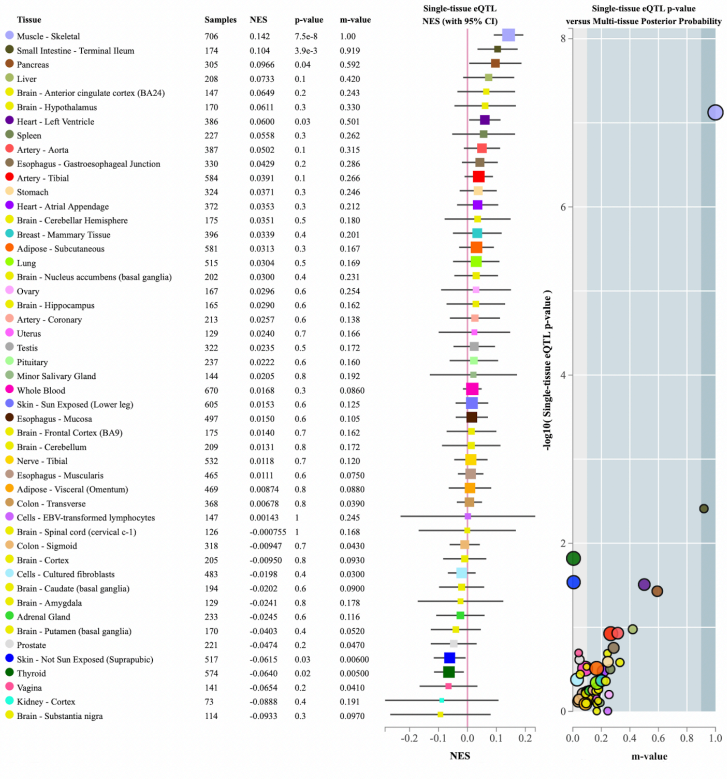



**Figure 3.1.2 Multi-tissue eQTL Comparison: rs1421085 on FTO**



genes (DEGs) were divided into up-regulated and down-regulated, followed by an investigation into whether these gene sets are preferentially enriched at either the top or bottom of each other's ranked transcriptome datasets. The GTEx statistical results provided normalized enrichment score (NES) in order to indicate the which regulated categories DEGs in the tissues belonged. A positive NES value in eQTL indicates enrichment of a gene set at the top of the ranked transcriptome data, while a negative NES value indicates enrichment at the bottom [24]. In GTEx statistics (Figure 3.1.2), the top 5 tissues with significantly expressed gene sets were muscle-skeletal, small intestine-terminal ileum, pancreas, liver and brain-anterior cingulate cortex (BA24).

## 3.2 Differential Gene Expression Analysis Statistics

Table 3.2.1 rs1421085 Associated Genes count

| Tissue | #Significant Expressed Genes |
|---|---|
| Adipose - Subcutaneous | 3116 |
| Adipose - Visceral (Omentum) | 4863 |
| Adrenal Gland | 5653 |
| Artery - Aorta | 2697 |
| Artery - Coronary | 1096 |
| Artery - Tibial | 3211 |
| Brain - Amygdala | 951 |
| Brain - Anterior cingulate cortex (BA24) | 1182 |
| Brain - Caudate (basal ganglia) | 1315 |
| Brain - Cerebellar Hemisphere | 1807 |
| Brain - Cerebellum | 1373 |
| Brain - Cortex | 1901 |
| Brain - Frontal Cortex (BA9) | 1796 |
| Brain - Hippocampus | 1106 |
| Brain - Hypothalamus | 3795 |
| Brain - Nucleus accumbens (basal ganglia) | 1578 |
| Brain - Putamen (basal ganglia) | 1763 |
| Brain - Spinal cord (cervical c-1) | 1911 |
| Brain - Substantia nigra | 2181 |
| Breast - Mammary Tissue | 2573 |
| Cells - Cultured fibroblasts | 1006 |
| Cells - EBV-transformed lymphocytes | 2166 |
| Colon - Sigmoid | 2739 |
| Colon - Transverse | 1418 |
| Esophagus - Gastroesophageal Junction | 2124 |
| Esophagus - Mucosa | 4532 |
| Esophagus - Muscularis | 3035 |
| Heart - Atrial Appendage | 5694 |
| Heart - Left Ventricle | 4011 |
| Kidney - Cortex | 980 |
| Liver | 3820 |
| Lung | 5214 |
| Minor Salivary Gland | 1490 |
| Muscle - Skeletal | 2484 |
| Nerve - Tibial | 2391 |
| Ovary | 1928 |
| Pancreas | 1505 |
| Pituitary | 4218 |
| Prostate | 1471 |
| Skin - Not Sun Exposed (Suprapubic) | 3641 |
| Skin - Sun Exposed (Lower leg) | 2815 |
| Small Intestine - Terminal Ileum | 3672 |
| Spleen | 2030 |
| Stomach | 3018 |
| Testis | 2907 |
| Thyroid | 4437 |
| Uterus | 2925 |
| Vagina | 3954 |
| Whole Blood | 7037 |

The single-tissue DE analysis on RNA-Seq data were performed from 49 tissue-specific samples focusing on only genes on autosomal chromosomes. The numbers of rs1421085 associated genes identified as differentially expressed in each tissues, with a p-value less than 0.05, were summarized in Table 3.2.1. According to the table, the top 5 tissues with most significant expressed genes were whole blood, heart-atrial appendage, adrenal gland, lung and adipose-visceral (omentum). Among 49 tissues, the top 10 rs1421085 associated genes were MTND1P23, MTCO1P12, RPL10P6, RPL10P9, SORD2P, TBC1D3E, CCL3L3, CSF3, CXCL3, IL6, and MTATP8P2 (Table 3.2.2). The genes which were not highlighted were confirmed as pseudogene in GeneCards, the Human Gene Database [25]. Pseudogenes were gene copies that had accumulated mutations over evolutionary time, rendering them unable to encode messenger RNA or produce functional proteins due to alterations in essential regions of the genetic code [26]. With accumulated

mutation, the pseudogenes usually showed higher expression level than normal genes.

When the highly expressed gene sets were identified from DE analysis, it was important to investigate its biological function and determined the pathways that they were involved in and interacted with. The visualization could be achieved through EnrichR, which provided information on biological pathways that were related to given gene sets and showed the interaction within each pathways [27,28,29]. The top 5 rs1421085 associated gene biological pathways were Cytokine-cytokine receptor interaction, Rheumatoid arthritis, IL-17 signaling pathway, Viral protein interaction with cytokine and cytokine receptor, and Lipid and atherosclerosis from KEGG pathway reference (Figure 3.2.1).

**Figure 3.2.1 Top candidate rs1421085 associated gene pathway network**

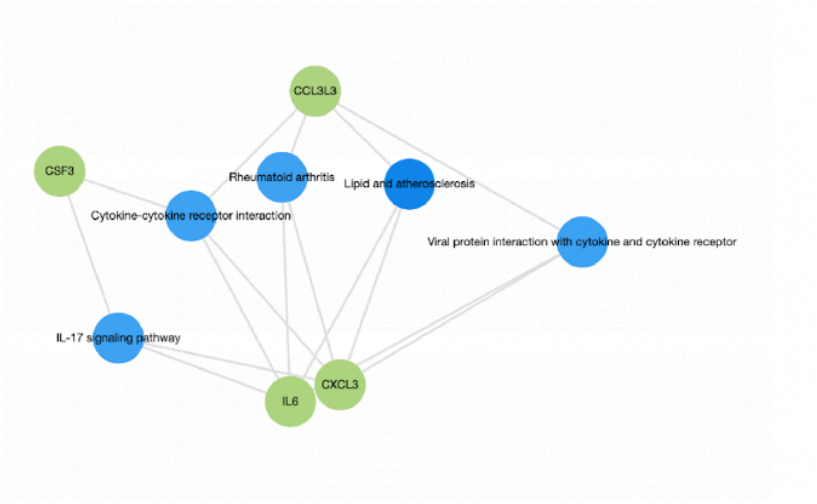



Table 3.2.2 rs1421085 Associated Genes

| Genes | #Significant Expressed Genes |
|---|---|
| MTND1P23 | 47 |
| MTCO1P12 | 26 |
| RPL10P6 | 10 |
| RPL10P9 | 10 |
| SORD2P | 5 |
| **TBC1D3E** | **4** |
| **CCL3L3** | **3** |
| **CSF3** | **3** |
| **CXCL3** | **3** |
| **IL6** | **3** |
| MTATP8P2 | 3 |

## 3.3 Gene Set Enrichment Analysis Statistics

The GSEA was performed with gene sets derived from the significant differentially expressed genes identified in the previous DE analysis among 49 tissue samples. The reference gene sets used in the analysis were BioCarta subset and KEGG subset under the Canonical Pathways. In the analysis, the heat map was generated to visualize the both significant rs1421085 associated pathways with tissues identified in the GSEA analysis. The heat maps for two different reference gene sets displayed the p-values in negative logarithm

transformation of all pathways and tissues. The heat map also indicated the degree of enrichment in each pathway and tissue with a color gradient. The most enriched pathways and tissues were presented by a shade of red, while less enriched pathways and tissues were presented by lighter shades of orange and yellow. There were four heat maps in total: one groups for all BioCarta pathway subsets and only significant BioCarta pathway subsets, while the other groups for all KEGG pathway subsets and only significant KEGG pathway subsets (Figure 3.3.1, Figure 3.3.2).

**Figure 3.3.1 rs1421085 associated pathways (BioCarta)**

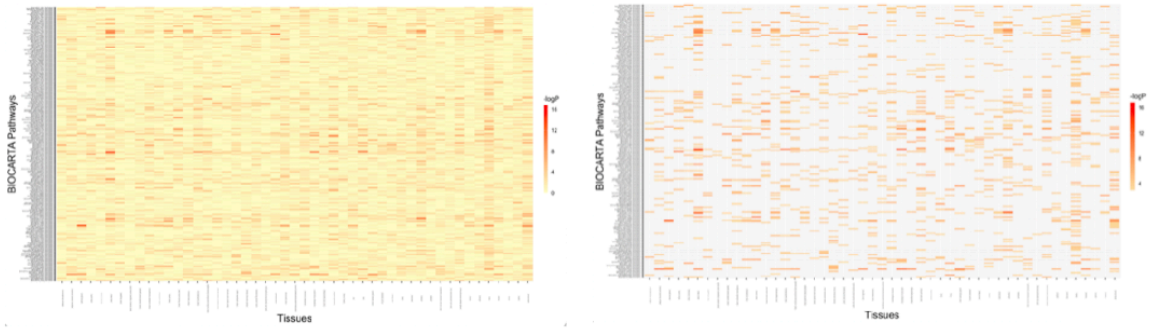
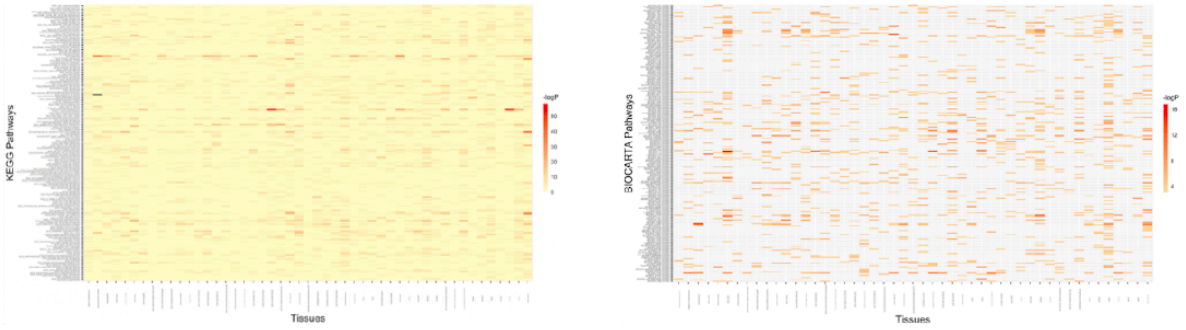



Table 3.3.5 rs1421085 Associated Pathway with tissue (KEGG)

| Pathway | Log P-value | Tissue |
|---|---|---|
| KEGG_OLFACTORY_TRANSDUCTION | 58.25117 | cells cultured fibroblasts |
| KEGG_OLFACTORY_TRANSDUCTION | 53.95712 | uterus |
| KEGG_DRUG_METABOLISM_CYTOCHROME_P450 | 42.40114 | whole blood |
| KEGG_METABOLISM_OF_XENOBIOTICS_BY_CYTOCHROME_P450 | 41.55602 | whole blood |
| KEGG_RIBOSOME | 34.88142 | adipose visceral omentum |
| KEGG_RETINOL_METABOLISM | 33.50327 | whole blood |
| KEGG_RIBOSOME | 33.35330 | nerve tibial |
| KEGG_OLFACTORY_TRANSDUCTION | 27.41691 | cells ebv transformed lymphocytes |
| KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION | 26.80357 | artery tibial |
| KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION | 26.65939 | cells ebv transformed lymphocytes |

**Figure 3.3.2 rs1421085 associated pathways (KEGG)**




According to the GSEA analysis, the tissues and pathways with most rs1421085 associated genes expressed were determined. The tissues with high expression levels were colon transverse, heart left ventricle, pituitary, small intestine-terminal ileum, testis, and whole blood from the summary in BioCarta gene sets and KEGG gene sets (Table 3.3.1 and Table 3.3.2).

Table 3.3.1 rs1421085 Associated Tissues (BIOCARTA)

| Tissue | Frequency |
|---|---|
| testis | 72 |
| artery tibial | 50 |
| pituitary | 46 |
| whole blood | 45 |
| heart left ventricle | 41 |
| esophagus mucosa | 35 |
| small intestine terminal ileum | 34 |
| brain hypothalamus | 32 |
| brain cerebellar hemisphere | 30 |
| colon transverse | 30 |

Table 3.3.2 rs1421085 Associated Tissues (KEGG)

| Tissue | Frequency |
|---|---|
| heart left ventricle | 60 |
| pituitary | 57 |
| vagina | 55 |
| whole blood | 51 |
| testis | 48 |
| colon transverse | 44 |
| cells cultured fibroblasts | 41 |
| small intestine terminal ileum | 41 |
| colon sigmoid | 37 |
| adrenal gland | 35 |

The following lists showed the significantly expressed biological pathways and tissue combinations associated with rs1421085, as identified using BioCarta and KEGG gene sets (Table 3.3.3, Table 3.3.4, Table 3.3.5, and Table 3.3.6).

Table 3.3.3 rs1421085 Associated Pathway (BIOCARTA)

| Pathway | Frequency |
|---|---|
| BIOCARTA_AHSP_PATHWAY | 20 |
| BIOCARTA_NEUTROPHIL_PATHWAY | 19 |
| BIOCARTA_CTLA4_PATHWAY | 18 |
| BIOCARTA_FIBRINOLYSIS_PATHWAY | 17 |
| BIOCARTA_IL17_PATHWAY | 17 |
| BIOCARTA_INFLAM_PATHWAY | 17 |
| BIOCARTA_LAIR_PATHWAY | 17 |
| BIOCARTA_LYMPHOCYTE_PATHWAY | 17 |
| BIOCARTA_MONOCYTE_PATHWAY | 16 |
| BIOCARTA_STEM_PATHWAY | 16 |

Table 3.3.4 rs1421085 Associated Pathway (KEGG)

| Pathway | Frequency |
|---|---|
| KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION | 36 |
| KEGG_RIBOSOME | 30 |
| KEGG_OLFACTORY_TRANSDUCTION | 29 |
| KEGG_DRUG_METABOLISM_CYTOCHROME_P450 | 26 |
| KEGG_HEMATOPOIETIC_CELL_LINEAGE | 26 |
| KEGG_METABOLISM_OF_XENOBIOTICS_BY_CYTOCHROME_P450 | 24 |
| KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION | 24 |
| KEGG_NOD_LIKE_RECEPTOR_SIGNALING_PATHWAY | 24 |
| KEGG_PENTOSE_AND_GLUCURONATE_INTERCONVERSIONS | 24 |
| KEGG_LEISHMANIA_INFECTION | 23 |

Table 3.3.5 rs1421085 Associated Pathway with tissue (BIOCARTA)

| Pathway | Log P-value | Tissue |
|---|---|---|
| BIOCARTA_IL17_PATHWAY | 16.78579 | artery tibial |
| BIOCARTA_IL17_PATHWAY | 13.69025 | esophagus muscularis |
| BIOCARTA_COMP_PATHWAY | 13.37424 | adrenal gland |
| BIOCARTA_TALL1_PATHWAY | 12.98777 | colon sigmoid |
| BIOCARTA_CTLA4_PATHWAY | 12.44362 | pituitary |
| BIOCARTA_CTL_PATHWAY | 12.09623 | artery tibial |
| BIOCARTA_AHSP_PATHWAY | 11.89515 | esophagus muscularis |
| BIOCARTA_IL17_PATHWAY | 11.71420 | brain hypothalamus |
| BIOCARTA_AHSP_PATHWAY | 11.31371 | adipose visceral omentum |
| BIOCARTA_TCAPOPTOSIS_PATHWAY | 11.05490 | artery tibial |

The analysis of gene sets from BioCarta and KEGG pathways revealed significant pathways from individual and combination analysis. In BioCarta gene sets, the pathways that showed significance were IL17 pathway (IL-17 signaling pathway), AHSP pathway (Hemoglobin's Chaperone), and CTLA4 pathway (The Co-Stimulatory Signal During T-cell Activation) [18]. Similarly, in KEGG gene sets, significant pathways were observed in Olfactory transduction, Drug metabolism cytochrome p450, Metabolism of xenobiotics by cytochrome p450, Ribosome, Cytokine Cytokine receptor interaction, and neuroactive ligand receptor interaction. Additionally, the cytokine cytokine receptor interaction and IL-17 signaling pathway

were also found to be significantly associated with FTO candidate genes, as determined through previous EnrichR analysis.

# 4. Discussion

In the study, the aim was to uncover potential obesity-associated genes and biological mechanisms with statistical methods and computational biology techniques among 49 tissue samples from GTEx portal. From the results, the study revealed some important insights into the obesity associated tissues, genes and pathways.

According to the GTEx statistics, the "C" allele of the rs1421085 gene variant was considered a risk allele, indicating that individuals who carried this allele were more susceptible to obesity. The small intestine was considered as most rs1421085 associated tissued from the results in eQTL and GSEA analysis. In previous obesity studies from Wisen et al., the obese patients had a significantly higher rate of absorption in the proximal intestine but a shorter intestinal transit time compared to individuals with normal weight. The rapid absorption occurring in the small intestine may lead to a decrease in satiety signals, ultimately impacting the motility of the small intestine [30]. In obesity, the transit time in the proximal small intestine appears to be heightened, which may be attributed to effective nutrient absorption and subsequent absence of satiety signals triggered by nutrients in the small intestine [31].

The rs1421085 associated valid genes were TBC1D3E, CCL3L3, CSF3, CXCL3, and IL6, and all of them had scientific supports from different studies. The associated pathways were cytokine cytokine receptor interaction and IL-17 signaling pathway from results in BioCarta, KEGG, and Enrichr analysis. Some studies indicated the relationship between cytokines and obesity. Obesity was associated with consistently elevated levels of various cytokines, such as TNFα, IL6, IL10, and CRP, primarily produced by adipose tissues, that contribute to chronic inflammatory conditions [32]. The presence of chronic low-grade inflammation symptoms in obese individuals implies that adipose tissue may be affected by IL-17A from IL-17 signaling

pathway [33]. The analysis could give the future study directions on inflammation conditions, which was associated with obesity.

Despite the promising results, the study also had several limitations. One notable limitation was the inconsistency in the results observed for highly expressed tissues among the eQTL analysis, DE analysis, and GSEA analysis. This discrepancy may be due to the difference in methods between DE analysis and eQTL analysis with covariates. Specifically, the DE analysis did not include all of the phenotypes in the statistical model, while the eQTL analysis included all phenotypes. Furthermore, the limited phenotype information available from the GTEx portal resulted in different statistical formulas being used for DESeq and eQTL analyses. Additionally, the DE analysis only accounted for two genotypes (CC and TT), whereas the eQTL analysis accounted for all three genotypes (CC, CT, and TT). These limitations suggest that the observed results may not fully reflect the true association between gene expression and the phenotype of interest. Another limitation of the study was the sample size used for the analysis, which may not have been large enough to detect subtle differences in gene expression between the groups of interest. To address this limitation, future studies could increase the sample size from the GTEx portal and conduct a more comprehensive analysis of the data to better understand the underlying mechanisms of gene expression in relation to the phenotype of interest.

# 5.References:


(1) Hruby, A., & Hu, F. B. (2015). The Epidemiology of Obesity: A Big Picture. *PharmacoEconomics*, *33*(7), 673–689. https://doi.org/10.1007/s40273-014-0243-x

(2) World Health Organization. (2023, March 27). Obesity. Retrieved from: https://www.who.int/health-topics/obesity

(3) World Health Organization. (2023, March 27). Obesity and overweight. Retrieved from: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

(4) World Health Organization. (2023, March 27).World Obesity Day 2022 - Accelerating action to stop obesity. Retrieved from: https://www.who.int/news/item/04-03-2022-world-obesity-day-2022-accelerating-action-to-stop-obesity

(5) CDC – Division of Nutrition Physical Activity, and Obesity – HOP 2023. https://www.cdc.gov/nccdphp/dnpao/state-local-programs/fundingopp/2023/hop.html 2023, March 27

(6) Pantalone, K. M., Hobbs, T. M., Chagin, K. M., Kong, S. X., Wells, B. J., Kattan, M. W., Bouchard, J., Sakurada, B., Milinovich, A., Weng, W., Bauman, J., Misra-Hebert, A. D., Zimmerman, R. S., & Burguera, B. (2017). Prevalence and recognition of obesity and its associated comorbidities: cross-sectional analysis of electronic health record data from a large US integrated health system. *BMJ open*, *7*(11), e017583. https://doi.org/10.1136/bmjopen-2017-017583

(7) Sawadogo W, Tsegaye M, Gizaw A, et al Overweight and obesity as risk factors for COVID-19-associated hospitalisations and death: systematic review and meta-analysis BMJ Nutrition, Prevention & Health 2022;e000375. doi: 10.1136/bmjnph-2021-000375

(8) Huang, Y., Lu, Y., Huang, Y. M., Wang, M., Ling, W., Sui, Y., & Zhao, H. L. (2020). Obesity in patients with COVID-19: a systematic review and meta-analysis. *Metabolism: clinical and experimental*, *113*, 154378. https://doi.org/10.1016/j.metabol.2020.154378

(9) Fruh S. M. (2017). Obesity: Risk factors, complications, and strategies for sustainable long-term weight management. *Journal of the American Association of Nurse Practitioners*, *29*(S1), S3–S14. https://doi.org/10.1002/2327-6924.12510

(10) Tirthani E, Said MS, Rehman A. Genetics and Obesity. [Updated 2023 Feb 1]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK573068/

(11) Dina, C., Meyre, D., Gallina, S., Durand, E., Körner, A., Jacobson, P., Carlsson, L. M., Kiess, W., Vatin, V., Lecoeur, C., Delplanque, J., Vaillant, E., Pattou, F., Ruiz, J., Weill, J., Levy-Marchal, C., Horber, F., Potoczna, N., Hercberg, S., Le Stunff, C., … Froguel, P. (2007). Variation in FTO contributes to childhood obesity and severe adult obesity. *Nature genetics*, *39*(6), 724–726. https://doi.org/10.1038/ng2048

(12)Kumar, S., Banks, T. W., & Cloutier, S. (2012). SNP Discovery through Next-Generation Sequencing and Its Applications. *International journal of plant genomics*, *2012*, 831460. https://doi.org/10.1155/2012/831460

(13)Park, J., Xu, K., Park, T., & Yi, S. V. (2012). What are the determinants of gene expression levels and breadths in the human genome?. *Human molecular genetics*, *21*(1), 46–56. https://doi.org/10.1093/hmg/ddr436

(14)The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from: [insert, where appropriate] the GTEx Portal on 03/29/2023

(15)eGTEx Project. Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease. *Nat Genet* **49**, 1664–1670 (2017). https://doi.org/10.1038/ng.3969

(16) Adam McDermaid, Brandon Monier, Jing Zhao, Bingqiang Liu, Qin Ma, Interpretation of differential gene expression results of RNA-seq data: review and integration, *Briefings in Bioinformatics*, Volume 20, Issue 6, November 2019, Pages 2044–2054, https://doi.org/10.1093/bib/bby067

(17)Love, M.I., Huber, W., Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**:550. 10.1186/s13059-014-0550-8

(18)Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., & Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics (Oxford, England)*, *27*(12), 1739–1740. https://doi.org/10.1093/bioinformatics/btr260

(19)Reiner, A., Yekutieli, D., & Benjamini, Y. (2003). Identifying differentially expressed genes usingfalse discovery rate controlling procedures. *Bioinformatics*, *19*(3), 368–375. https://doi.org/10.1093/bioinformatics/btf877

(20)Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, de Bakker PIW, Daly MJ & Sham PC (in press) PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics.

(21)RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL *http://www.rstudio.com/*.

(22)SNPedia: a wiki supporting personal genome annotation, interpretation and analysis Michael Cariaso; Greg Lennon Nucleic Acids Research 2011; doi: 10.1093/nar/gkr798

(23)Sul JH, Han B, Ye C, Choi T, Eskin E (2013) Effectively Identifying eQTLs from Multiple Tissues by Combining Mixed Model and Meta-analytic Approaches. PLOS Genetics 9(6): e1003491. https://doi.org/10.1371/journal.pgen.1003491

(24)Kim, K., Jeong, J., Kim, J., Lee, N., Kim, M. E., Lee, S., Chang Kim, S., & Choi, G. (2016). PIF1 Regulates Plastid Development by Repressing Photosynthetic Genes in the Endodermis. *Molecular plant*, *9*(10), 1415–1427. https://doi.org/10.1016/j.molp.2016.08.007

(25)Stelzer G, Rosen R, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Iny Stein T, Nudel R, Lieder I, Mazor Y, Kaplan S, Dahary, D, Warshawsky D, Guan - Golan Y, Kohn A, Rappaport N, Safran M, and Lancet D Current Protocols in Bioinformatics(2016), 54:1.30.1 - 1.30.33.doi: 10.1002 / cpbi.5

(26) *Elise Feingold retires after spearheading encode project*. National Human Genome Research Institute Home | NHGRI. (n.d.). Retrieved April 6, 2023, from https://www.genome.gov/

(27)Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics.2013; 128(14).

(28)Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW, Ma'ayan A. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Research. 2016; gkw377 .

(29)Xie Z, Bailey A, Kuleshov MV, Clarke DJB., Evangelista JE, Jenkins SL, Lachmann A, Wojciechowicz ML, Kropiwnicki E, Jagodnik KM, Jeon M, & Ma'ayan A. Gene set knowledge discovery with Enrichr. Current Protocols, 1, e90. 2021. doi: 10.1002/cpz1.90

(30)Wisén O, Johansson C. Gastrointestinal function in obesity: motility, secretion, and absorption following a liquid test meal. Metabolism 1992;41:390-5.

(31)Mushref, M., & Srinivasan, S. (2012). Effect of high fat-diet and obesity on gastrointestinal motility. *Annals Of Translational Medicine, 1*(2), 14. doi:10.3978/

(32)j.issn.2305-5839.2012.11.01Hosgood, H. D., Gunter, M. J., Murphy, N., Rohan, T. E., & Strickler, H. D. (2018). The Relation of Obesity-Related Hormonal and Cytokine Levels With Multiple Myeloma and Non-Hodgkin Lymphoma. *Frontiers in oncology*, *8*, 103. https://doi.org/10.3389/fonc.2018.00103

(33)Ahmed, M., & Gaffen, S. L. (2010). IL-17 in obesity and adipogenesis. *Cytokine & growth factor reviews*, *21*(6), 449–453. https://doi.org/10.1016/j.cytogfr.2010.10.005