Summer 2023

Meta-Analysis of Mesenchymal Stem Cell Gene Expression Data from Obese and Non-Obese Patients

Dakota William Shields
Missouri University of Science and Technology

Department:

META-ANALYSIS OF MESENCHYMAL STEM CELL GENE EXPRESSION MICROARRAY DATA FROM OBESE AND NON-OBESE PATIENTS

by

DAKOTA WILLIAM SHIELDS

A THESIS

Presented to the Graduate Faculty of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

in

APPLIED MATHEMATICS: STATISTICS

2023

Approved by:

Dr. Gayla R. Olbricht, Advisor
Dr. Akim Adekpedjou
Dr. Julie Semon

ABSTRACT

The prevalence of gene expression microarray datasets in public repositories gives opportunity to analyze biologically interesting datasets without running the laboratory aspect in house. Such experimentation is expensive in terms of finances, time, and expertise, which often results in low numbers of replicates. Meta-analysis techniques attempt to overcome issues due to few biological or technical replicates by combining separate experiments together to increase statistical power. Proper statistical considerations help to offset issues like simultaneous testing of thousands of genes, unintended hybridization, and other noises.

Microarrays contain light intensities from tens of thousands of hybridized probes giving a measure of gene expression for much of the human genome. This work focuses on identifying differentially expressed genes between obese and non-obese patients using microarray data from two studies collected from mesenchymal stem cell samples. Obesity is associated with poorer quality stem cells that are less readily available to differentiate and it is of interest to identify genes associated with this condition. Meta-analysis performed to increase statistical power from low replicate microarray experiments is an attempt to gain a better idea of the gene expression differences between obese and non-obese individuals compared to results from an individual study. Increased statistical power translates to improved ability to discover genes or sets of genes associated with this observed decrease in differentiation efficacy. Furthermore, pathway analysis could be completed to identify pathways of interest from this differential expression analysis.

ACKNOWLEDGMENTS

Thank you, Dr. Gayla, for your dedication to my research and education. Our continued efforts in multiple research groups allowed me to climb to new heights in ways I did not realize I would. The expertise and kindness of such a wonderful person and professor as yourself cannot be overstated.

Thank you, Dr. Akim and Dr. Semon, for supporting me throughout my journey at Missouri S&T as well as in our research together. Your encouragement and counsel have had a significant impact on myself and my work.

Thank you, Hailey, for inviting me to join you in our research. I've enjoyed dearly the great efforts and fruitions of our work together.

Thank you, Missouri S&T, for the opportunity to grow in both sides of the classroom. Giving back to my university through the education of others has been one of my greatest joys here in Rolla.

Thank you, Mom and Dad, for raising me to be steadfastly curious and strong willed. I am truly grateful to love you and to be loved by you.

And, of course, thank you Juniper, Clementine, Milkshake, Boba and Mungbean for making our house a home.

TABLE OF CONTENTS

Page

LIST OF ILLUSTRATIONS

LIST OF TABLES

1. INTRODUCTION

1.1. MOTIVATION AND BIOLOGICAL BACKGROUND

Mesenchymal stem cells (MSCs) are a type of multipotent stromal cell with self-renewal and cell differentiation properties first isolated by A. J. Freidenstein and colleagues in 1970 from bone marrow. Since then, MSCs from several other tissues have been isolated including blood, umbilical cord, and adipose tissue (Bianco et al., 2008). MSCs are derived from adult or young adult (fetal/perinatal tissues) and utilize tissue that is either renewable or unwanted. MSCs are different than embryonic stem cells which pose ethical issues for research. The two most common sources for MSCs are from bone tissue, which is renewable, or adipose tissue, which is often unwanted. Furthermore, MSCs derived from the placenta or Wharton's Jelly (umbilical cord tissue) are often otherwise disposed at birth. These sources provide a basis for collecting and analyzing MSCs for research purposes from adult (or young adult) tissues (Pittenger et al., 2019). One reason MSCs are interesting is their ability to differentiate into multiple cell lineages including osteoblasts (bone cells), chondrocytes (cartilage cells), and myocytes (muscle cells) in order to replace damaged or diseased tissues (Pittenger et al., 2019). Furthermore, MSCs have the ability to regulate the immune system by signaling immune cells and by secreting cytokines and growth factors which assist in cell repair, metabolism, and inflammation (Han et al., 2022; Gao et al., 2021). These properties as well as the MSCs' low immune response to a foreign body make them good candidates as a therapeutic for autoimmune diseases.

MSCs have been used to treat a variety of autoimmune diseases including Crohn's Disease and lupus (Gao et al., 2021). MSCs were shown to increase circulating T_{Reg} cells and balance cytokines associated with systematic lupus erythematosus (Pistoia & Raffaghello, 2017). Fistula are abnormal connections in the body that join two body parts that are not typically connected (e.g., such as the colon and the surface of body) (Anal Fistula, 2022). In a study on patients with fistulas caused by Crohn's disease, autologous (patient derived) MSC treatment either improved or completely closed the fistula of all 12 patients (Gao et al., 2021). In another study, patients suffering from graft-versus-host disease, an immune disease caused from donor T-cells attacking the host' cells, had a significantly higher 1-year survival rate when using MSCs derived from the placenta compared to historical data using other therapies (Baygan et al., 2017). These studies represent a broad and active area of research involving the investigation of the clinical utility of MSCs in a variety of settings. In 2019, over 950 clinical trials involving MSCs were registered with the Food and Drug Administration, which illustrates their therapeutic potential (Pittenger et al., 2019). Much work is still needed to understand different aspects of MSCs and their capacity to treat diseases.

Of particular interest to the research in this thesis is the comparative efficacy of MSCs derived from obese and non-obese patients. Oñate et. al. studied subcutaneous white adipose-derived stem cells from obese and non-obese patients (Oñate et al., 2013). They found a downregulation of genes associated with differentiation and an upregulation of genes associated with inflammation in obese patients. That is, stem cells from obese patients were dedicated to differentiation into adipocytes (fat cells) and were not effectively differentiating into other lineages as well as stem cells from non-obese

patients. Pestel et. al. recently reported adipose-derived MSCs (ASCs) exhibit pro-inflammatory or anti-inflammatory properties dependent on the microenvironment encompassing the ASCs (Pestel et al., 2023). Thus, a cycle continues in which an obese person's inflammatory environment induces pro-inflammatory properties of the ASCs, which further contributes to inflammation. This inflammatory environment then contributes to the pathologies of cancer and autoimmune disease (Pestel et al., 2023). These studies motivate the work in this thesis to further study MSCs through analysis of the gene expression profile from multiple sources of stem cells in an attempt to detect significantly upregulated or downregulated genes between obese and non-obese patients. This is accomplished by implementing a statistical framework that employs meta-analysis methodology to obtain the most information from several complex experiments and addresses the multiple testing issue to reduce the number of false positives across the thousands of genes tested. After a thorough review of potential studies available in the National Center for Biotechnology (NCBI) Gene Expression Omnibus (GEO), two studies were selected and each utilized unique sources of MSCs and technologies for measuring gene expression.

In this section, genetic concepts related to gene expression and how it is measured via microarray technology are introduced. Details are provided for the types of microarrays (Affymetrix GeneChip and Illumina BeadChip) utilized in the two studies selected for analysis. A general review of statistical methods for differential expression in individual microarray experiments is then provided, with a focus on the methods used in this work. Finally, a review of meta-analysis methods used to pool information from

multiple gene expression studies is given, with a focus on the rank product method that is employed in this work.

1.2. THE CENTRAL DOGMA AND DIFFERENTIAL GENE EXPRESSION

The "Central Dogma of Molecular Biology" is the concept that the transfer of genetic information is unidirectional, that is, deoxyribonucleic acid (DNA) transcribes to make ribonucleic acid (RNA) and RNA translates to make protein. More specifically, DNA transcribes to make messenger RNA (mRNA) and then mRNA is translated in the cell's ribosomes to make protein (Sookdeo, 2022). Proteins facilitate the behavior and structure of the cell. Genes are the regions of DNA that have the ability to encode proteins and/or produce a functional RNA. Gene expression is defined by this process in which phenotypes are derived from DNA. For example, the same set of instructions (DNA) are present for muscle cells and neurons, yet these two cell types possess starkly different behaviors and shapes (phenotypes). This difference indicates a different set of genes are being expressed in muscle cells versus neurons. Genes in diseased tissue (e.g., cancer) versus healthy tissue or immune-compromised cells versus healthy cells also display differential expression. The purpose of studying differential gene expression is to see which genes in a treatment or disease group are overexpressed, underexpressed, or equally expressed compared to the same genes from a control or healthy group. In this work, if the disease group has higher expression of a gene compared to the healthy group, then this gene is said to be upregulated. If the disease group has lower expression compared to the healthy group, then this gene is said to be downregulated. This indicates which subset of instructions a cell is working with and provides a deeper insight into the

mechanisms behind treatment/disease groups and their phenotypes. A visualization of differential expression is given in Figures 1.1 and 1.2 below.

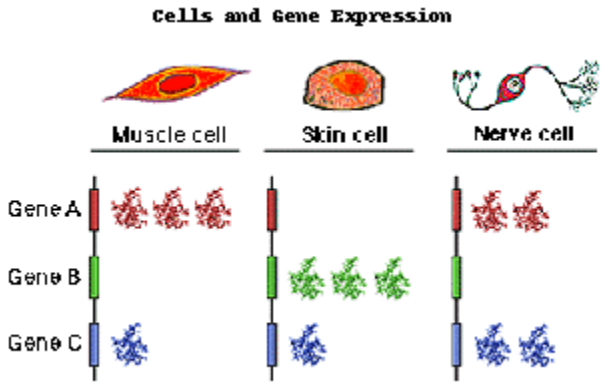

Figure 1.1 Differential Gene Expression Between Cell Types (muscle, skin, and nerve cells). Figure from the U.S. National Library of Medicine (.n.d.) MLA CE Course Manuel: Molecular Biology Information Resources.

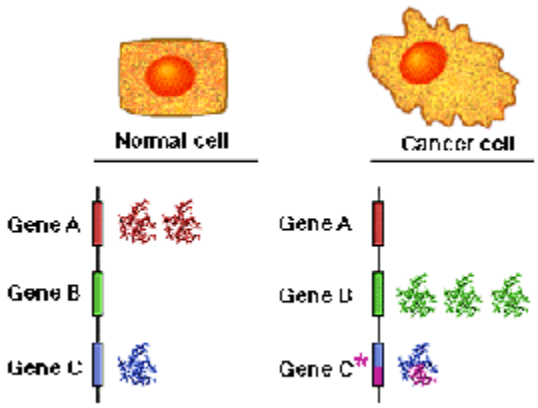

Figure 1.2 Differential Gene Expression Example Between Disease States (normal and cancer cells). Figure from the U.S. National Library of Medicine (n.d.), MLA CE Course Manuel: Molecular Biology Information Resources.

In this section, concepts related to the Central Dogma are described in more detail to provide a more thorough understanding of the gene expression process. DNA consists of a double-stranded chain of deoxyribose sugar molecules, phosphorous backbones, and

nitrogen bases called nucleotides (Sookdeo, 2022). These nucleotides include adenine (A), guanine (G), cytosine (C), and thymine (T). The double-stranded nature of DNA involves the anti-parallel pairing of complementary DNA strands and the exclusive pairings of adenine with thymine and cytosine with guanine. Furthermore, adenine and thymine are bonded by two hydrogen bonds while guanine and cytosine are bonded by three hydrogen bonds. These nitrogen bases are attached to a sugar phosphate backbone via covalent bonds (Sookdeo, 2022). The general structure of DNA is given in Figures 1.3 and 1.4 below.

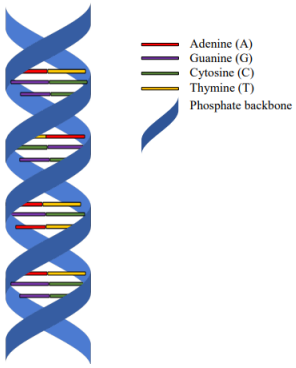

Figure 1.3 The Double Helix Structure of DNA. Figure from Sookdeo (2022).

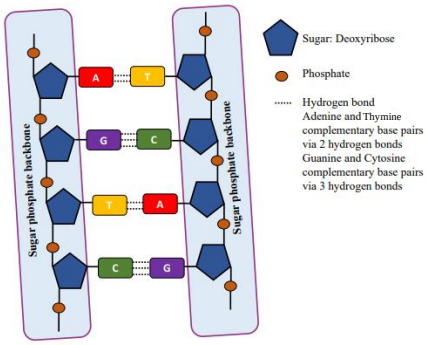

Figure 1.4 DNA Structure with Sugar Phosphate Backbone and Hydrogen Bonds between Complementary Base Pairs. Figures from Sookdeo (2022).

RNA is a molecule similar to DNA, in that it also consists of a phosphate group, nitrogen base, and sugar molecule. However, the sugar molecule in RNA is ribose, which has an additional hydroxyl group attached to the second carbon. Additionally, RNA is single stranded and consists of uracil (U) instead of thymine (Sookdeo, 2022). Thus, the nitrogen bases of RNA are then adenine, uracil, cytosine, and guanine. A visualization of RNA is given in Figure 1.5 below.

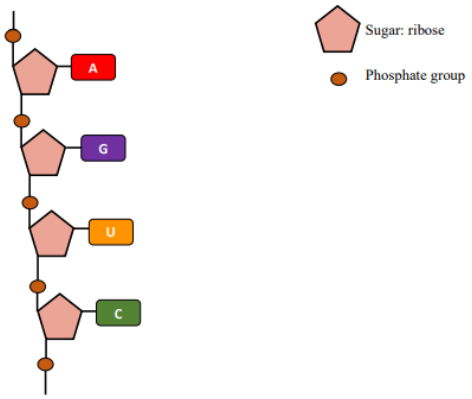

Figure 1.5 The Single Stranded Structure of RNA. Figure from Sookdeo (2022).

There are several kinds of RNA necessary for gene expression, namely, messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA). The mRNA facilitates the transfer of genetic information and is made from transcription of the DNA strand. This single strand of RNA is complementary to the DNA from which it came. Transcription is the process of making complementary mRNA from a template DNA strand (Sookdeo, 2022). Transcription begins when RNA polymerase attaches to sequences of DNA called promoter sequences and separates the template strand and non-template strand of DNA in a process called initialization (Figure 1.6). RNA polymerase

moves along the 3’ to 5’ direction of the template strand of DNA while attaching complementary nucleotides to the 3’ end of the newly forming RNA (i.e., RNA is synthesized in the 5’ to 3’ direction) in a process called elongation (Figure 1.7). This process stops when RNA polymerase reaches sequences of DNA called terminators, which signal to RNA polymerase to detach from the DNA strand (Figure 1.8).

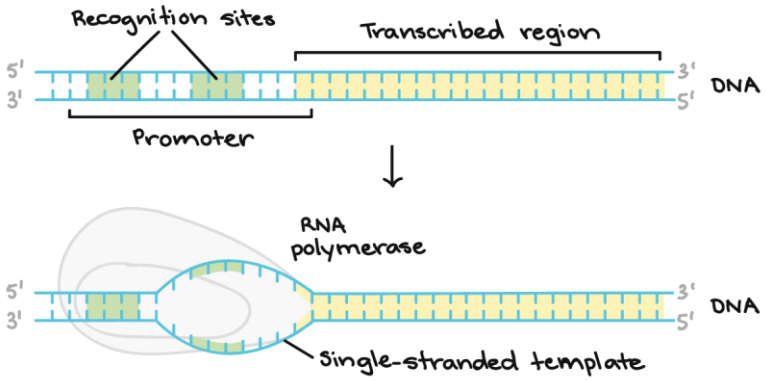

Figure 1.6 Initiation Stage of Transcription. Figure from Khan Academy (n.d.).

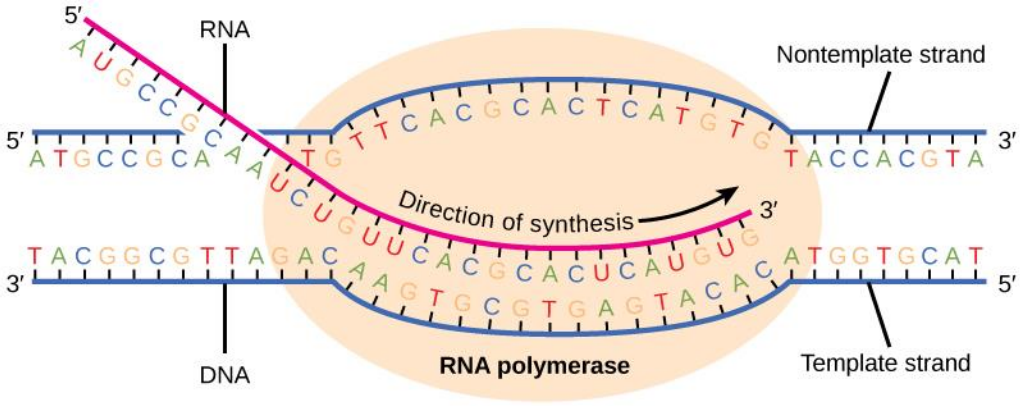

Figure 1.7 Elongation Stage of Transcription. Figure from OpenStax College (n.d.).

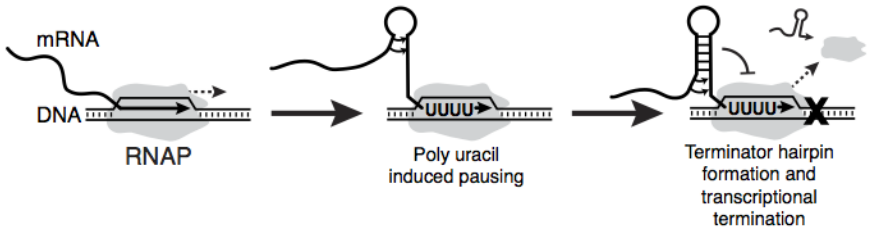

Figure 1.8 Termination Stage of Transcription. Figure from Chappell and Lucks (n.d.).

After transcription is completed, a modified guanine cap is attached to the 5’ end of the RNA strand, which helps to protect it from degradation, and a sequence of adenines called a polyadenylation (poly-A) tail is attached to the 3’ end (Sookdeo, 2022). The mRNA is then ready to undergo the translation process. Each set of three nucleotides in the mRNA forms a codon, which are associated with amino acids. There are 64 possible arrangements of three nucleotides, but only 20 amino acids meaning there is redundancy between codons and associated amino acids. In the translation process, these amino acids are joined together via a polypeptide chain to form a protein, which performs cellular functions.

tRNA facilitates the gathering and connection of the amino acids forming the polypeptide chain (protein) coded for by the mRNA (Sookdeo, 2022). On one end of the tRNA is the amino acid receptor, which holds the amino acid associated with the anticodon loop on the other end of the tRNA. The anticodon loop consists of the three complementary nucleotides on the tRNA that are associated with the codon on the mRNA which is being added to the protein being translated. A visualization of a tRNA is given in Figure 1.9 below.

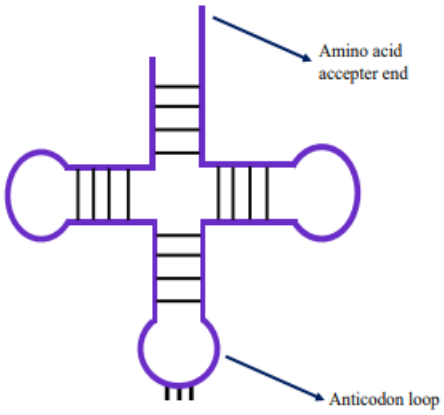

Figure 1.9 Visualization of the tRNA Molecule. Figure from Sookdeo (2022).

The process of translation is carried out in a cell's ribosomes from the 5' to 3' ends of the mRNA. The rRNA and ribosomal proteins are the principal components of a cell's ribosomes, which consists of a large ribosomal subunit and a small ribosomal subunit (Sookdeo, 2022). A visualization of a ribosome is given in Figure 1.10 below.

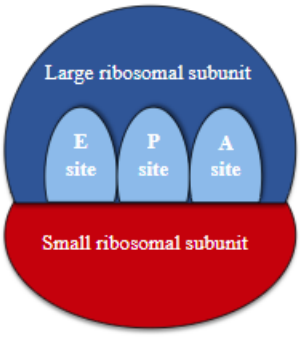

Figure 1.10 Structure of a Ribosome with E, P, and A Sites. Figure from Sookdeo (2022).

There are 3 sites on the large ribosomal subunit used in protein synthesis. The A (aminoacyl) site is the first site the mRNA interacts with in a process called initiation

(Figure 1.11). A start codon on the mRNA strand, consisting of nucleotides AUG and coding for methionine, signals the beginning of translation when a complementary tRNA carrying the amino acid methionine attaches to the A site and the mRNA (Sookdeo, 2022). Then the next codon is read in the A site while the methionine and its tRNA are transferred to the P (peptidyl) site. When the next tRNA and amino acid are introduced to the ribosome, the tRNAs shift from the A to P site and then from the P to E (exit) site while the amino acid from the A site attaches to the growing chain in the P site. This process of attaching amino acids to the incomplete polypeptide chain is referred to as elongation (Figure 1.12-1.14). The polypeptide chain remains in the P site until the protein has been fully translated. The tRNA transferred to the E site now no longer has an amino acid and detaches from the ribosome into the cytoplasm. Three of the 64 codons are stop codons that cannot be translated and are indicators to stop protein synthesis, signaling that the protein is completely translated (Sookdeo, 2022). The completed protein and mRNA are then released and the ribosomal subunits detach to find another mRNA to translate in a process called termination (Figure 1.15).

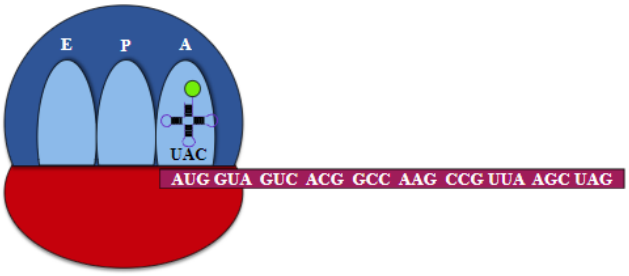

Figure 1.11 Translation Initiation. Figure from Sookdeo (2022).

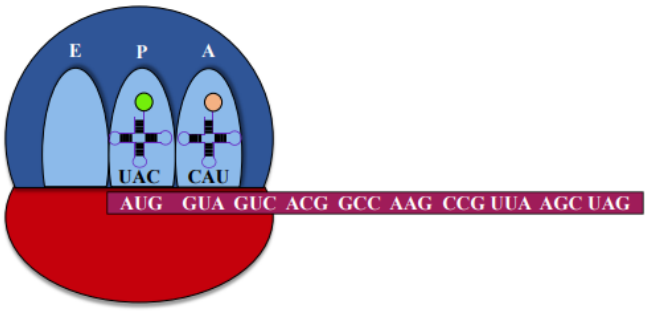

Figure 1.12 Translation Elongation Part 1. Figure from Sookdeo (2022).

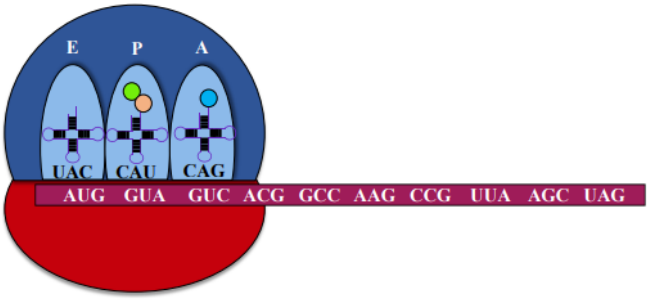

Figure 1.13 Translation Elongation Part 2. Figure from Sookdeo (2022).

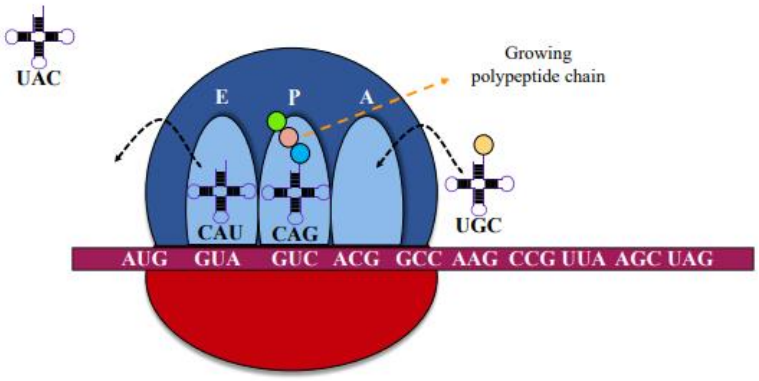

Figure 1.14 Translation Elongation Part 3. Figure from Sookdeo (2022).

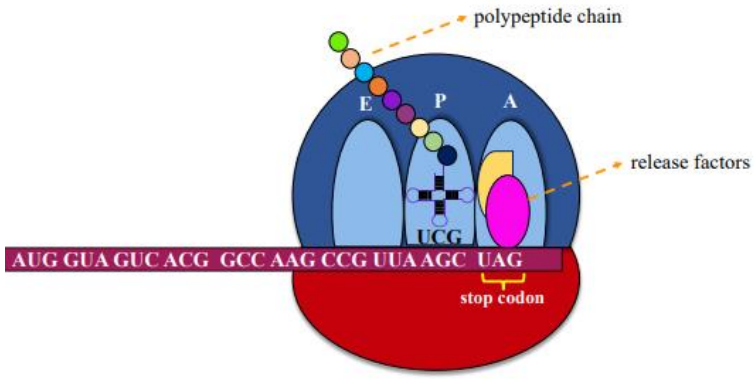

Figure 1.15 Translation Termination. Figure from Sookdeo (2022).

This Central Dogma and process of gene expression are important in understanding how genes are connected to phenotypes. Data can be collected at each stage to better understand these associations. At the DNA level, genotype information can be collected to investigate how differences in the DNA sequence between individuals are associated with phenotypic differences. Other studies investigate the proteins created from the process. In this work, the focus is on data obtained from the intermediate mRNA step. Measuring mRNA levels (often called expression levels) provide more information about which genes are being transcribed in a given sample. Comparing the mRNA expression levels between conditions can provide insights into how differences in the transcriptional activity is related to phenotypes.

1.3. MICROARRAY TECHNOLOGY

Microarrays have revolutionized scientific studies of gene expression since their introduction in 1995, by allowing the expression levels thousands of genes to be

measured simultaneously (Clough and Barrett, 2016). This involves attaching synthetic DNA probes to a glass slide (or chip) and hybridizing a sample obtained from mRNA in an individual person to these probes with the goal of studying expression levels of genes of interest in the human genome. By attaching a fluorescent molecule to these probes, lasers allow for the quantification of gene expression in the form of light intensity (Jaksik et al., 2015).

In 2000, the National Center for Biotechnology Information (NCBI) introduced the Gene Expression Omnibus (GEO) database for researchers to deposit their gene expression microarray data (Clough and Barrett, 2016). This database provides a publicly available, data-rich resource for high-throughput genomic data that has expanded from gene expression microarray experiments to a variety of other genomic applications (e.g., DNA methylation) and technologies (e.g., sequencing data). In 2002, many major journals began requiring authors to make their microarray data publicly available through databases such as GEO (Clough and Barrett, 2016). To ensure the available data meet a set of standard criteria that other researchers can interpret and verify, the Minimum Information About a Microarray Experiment (MIAME) guidelines (Brazma et al., 2001) are required for data submitted to GEO. As of May 2023, the GEO database contains data on over 199,000 studies consisting of over 5.7 million samples (*Geo summary - geo – NCBI*, n.d.). This culture of data sharing has made it possible for researchers to further advance genomic research by independently reproducing analyses, exploring alternative analysis methods, and combining data from multiple studies through meta-analysis techniques.

In this work, the NCBI GEO database was searched to identify microarray studies measuring gene expression levels from an MSC source in obese and non-obese patients (Clough and Barrett, 2016). Studies were included if: 1) there was a comparison between MSCs that were isolated from healthy individuals and MSCs isolated from obese individuals; 2) studies were published and accessible in English; 3) studies were peer-reviewed. Studies were excluded if: 1) MSCs were derived from non-human species; 2) MSCs were treated with any pharmaceutical agent or biomaterial for the duration of the study; 3) the article was a review, conference proceeding, or retracted study. Two studies were identified for inclusion, with each study using a different type of microarray technology (Affymetrix GeneChip and Illumina BeadChip) and obtaining an MSC sample from a different source (adipose tissue and Wharton's Jelly). The goal of this work is to combine information from both of these past experiments to identify a common set of differentially expressed genes derived from MSCs in obese and non-obese patients. This section provides an overview of how microarrays are used to measure genome-wide expression. Details are provided for the two types of microarray technologies used in this research, the Affymetrix GeneChip (which is currently produced by Thermo Fisher Scientific) and the Illumina BeadChip.

1.3.1. Affymetrix GeneChip Technology. Affymetrix GeneChip microarrays were the first commercially available gene expression microarray. These microarrays give a measure of a gene's expression level by representing each gene with multiple probes, referred to as a probeset, that hybridize with the target mRNA from an individual sample. Probes are synthesized by creating short (25 bases, also called 25-mer) single stranded copies of DNA (called oligonucleotides) that represent segments of genes (Luo,

2007). Each probeset consists of 11 to 20 probes derived from a gene of interest. This redundancy gives a more reliable measure than a single probe, which is especially necessary since probes can map to multiple genes. The probes are covalently bonded to a glass slide known as a chip. Hundreds of thousands of copies of each probe are attached to a small section of this glass slide known as a feature. A visualization of how the probes from a probeset are represented on the microarray is given in Figure 1.16 (Jaksik et al., 2015).

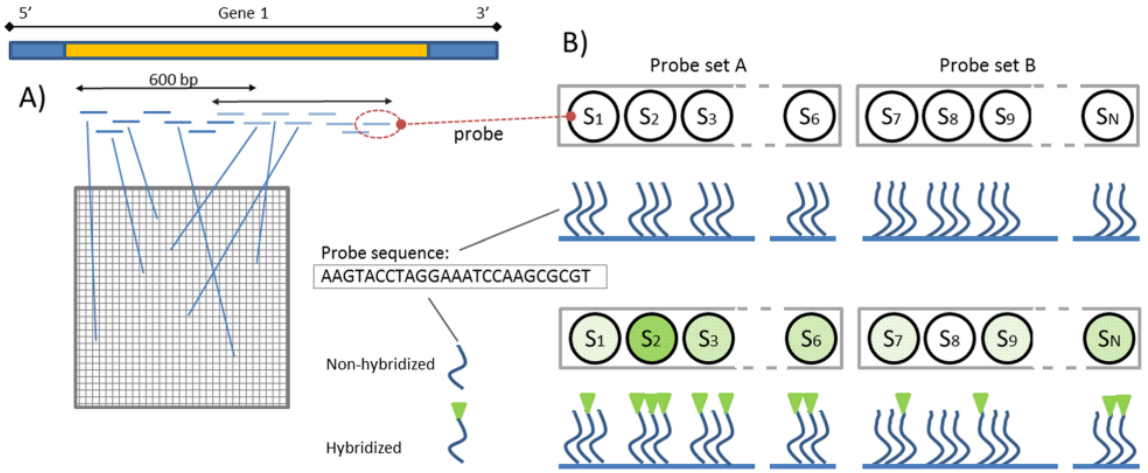

Figure 1.16 Visualization of Probesets Associated with a Gene of Interest. Figure from Jaksik et al. (2015).

The probes hybridize through complementary base paring to biotin labeled complementary RNA (cRNA) segments of interest derived from the biological target sample (Luo, 2007). Complementary RNA is the antisense copy of RNA made from double stranded complementary DNA (cDNA). Complementary DNA is made from reverse transcription of the RNA segment of interest. The mRNA is transcribed from the original DNA segment of interest and the cDNA is a complementary copy of that DNA segment made from reverse transcription of the mRNA segment. This is necessary

because mRNA degrades faster than DNA. Biotin allows for a fluorescent molecule to be attached to the cRNA (the antisense strand of the cDNA) (Luo, 2007). When the chip is put under an optical scanner, a laser light is shined on the chip and the fluorescent molecule on the biotin gives a measure of gene expression for that probe in the form of a light intensity. Thus, probes associated with a gene not expressed in the sample are dark since the complementary segments were not present in the sample and probes associated with a gene highly expressed in the sample are bright since the sample contained the complementary biotin-labeled segments that hybridized to the probes. These light intensities provide quantitative readings that represent gene expression levels. The probe-level intensities are then summarized into a single number for each probeset during the Robust Multi-Array Average (RMA) preprocessing method using Tukey's Median Polish approach. The details of this method are provided in Section 1.4.1. Note that for Affymetrix arrays, one biological sample is hybridized to each array.

1.3.2. Illumina BeadChip Technology. In contrast to Affymetrix GeneChip technology, Illumina later developed what's known as BeadChip technology, which includes randomly selected silica beads put into wells on a silicon chip (Illumina, n.d.). These silicon beads are known as bead types and each bead type has hundreds of thousands of copies of a 50-mer oligonucleotide probe specific to the bead type attached to the bead (Luo, 2007). The increase in the probe length compared to the 25-mer Affymetrix probes allow for more specific binding and improved performance, according to Illumina (Kuhn et al., 2004). Also, in contrast to the Affymetrix array, the vast majority of genes on the array are represented by only one of the 50-mer probes. However, an average of 30 replicates of each bead type are placed randomly on the

Illumina BeadChip, providing for technical replicates of each probe. These replicates are summarized into probe-level data using a trimmed mean method utilized in Illumina's default settings of BeadStudio (Luo, 2007). Microarray datasets found on the NCBI GEO repository often reflect probe-level data, which has already been summarized from the bead-level data. The randomness of the bead type placement on the array controls the effects of spatial artifacts while the redundancy in the number of each bead type allows for higher precision (Kuhn et al., 2004). Since the placement of bead types is random, chips are put through a decoding process by Illumina to determine which beads correspond to which gene. Similar to Affymetrix, complimentary strands of cRNA from the target biological sample are labeled with a fluorescent molecule give a quantitative measure of gene expression in the form of light intensity when hybridized to the probes. Note that Illumina BeadChips contain multiple BeadArrays per chip, with one biological sample being hybridized to each BeadArray. Thus, multiple samples are measured in parallel on each BeadChip, reducing the experimental differences that may occur if the samples are processed at different times.

1.3.3. Other Microarray Considerations. There are some common issues to consider for both Affymetrix and Illumina microarray technology that are important for obtaining meaningful data. These include laboratory batch effects, non-specific hybridization, reliance on existing knowledge of the human genome, non-unique probe(set) to gene mapping, and the need to address multiple testing. Critical stages of microarray preparation that can affect the quality of the gene expression estimates include the amplification of biotin labeled cRNA from the cDNA and the hybridization of cRNA to the probes on the microarray. Amplification provides the microarray with enough

cRNA that a light intensity can be read. The hybridization process is time-consuming and sensitive to reaction conditions and individual cRNA molecule structures that contribute to variation in light intensity readings (Jaksik et al., 2015). Thus, it is important to reduce variation from these types of technical artifacts that can arise between samples. Normalization is an important pre-processing step for microarray data that removes this type of technical variation between samples that would otherwise be confounded with true biological differences (Jaksik et al., 2015). A detailed description of the normalization methods used in this work is given in Section 1.4.1.

Furthermore, non-specific hybridization of probes reduces the statistical power of the experiment by causing background intensity to mask true signal intensity. Non-specific hybridization (also called cross-hybridization) occurs when the target cRNA sequences bind to probes that are not strictly complementary to their sequence (Jaksik et al., 2015). The Affymetrix array attempts to address this by include both a perfect match (PM) and mismatch (MM) sequence for each probe. The PM probe represents the exact sequence where target cRNA sequence should hybridize and the MM probe differs from the PM probe at the 13^{th} base. Although the MM probe was designed to detect non-specific hybridization, research has shown that it is also detects true signal and thus does not provide an accurate measurement of the background noise (Shi et al., 2010). Thus, typically only the PM probes are utilized. The Illumina array contains a set of negative control probes that are not expected to hybridize with the target cRNA sample since they do not correspond to an expressed sequence in the genome. These control probes can provide a way to measure the non-specific binding and background noise (Xie et al., 2009). For both arrays, background correction is an additional pre-processing step that is

important for extracting signal from background noise by reducing the impact of this non-specific binding. A detailed description of the background correction methods used in this work is given in Section 1.4.1.

Microarray technology from companies like Affymetrix and Illumina is possible due to existing knowledge of the human genome. That is, microarrays are made from predetermined probes and genes from the existing knowledge of the human genome at the time of their production. Thus, they are limited to the genes humanity was aware of when the array were made and was able to reproduce through the creation of probes or probesets. A newer technology called RNA-seq that is utilized to measure gene expression does not require the use of pre-determined probes and thus can detect novel transcripts that were not represented on microarrays. As the body of knowledge of the human genome increases, methods for measuring gene expression will continue to improve.

The methodology of the meta-analysis method used in this research (rank product) requires a dataset with one sample per column and one gene per row. Furthermore, a gene cannot be represented more than once. For this reason, a unique gene to probe(set) mapping must be established. Note, however, that a probe(set) can be associated with multiple genes and a gene can be associated with multiple probe(set)s. A description of the establishment of this unique mapping is found in Sections 2.1.1 and 2.1.2.

Finally, it is important to point out that microarray experiments are a double-edged sword. On one hand, there is advantage to studying 1000’s of genes at the same time since the expression of a cell’s gene could change as time changes. On the other

hand, testing for significant expression differences between groups (e.g., disease vs. healthy) in a large number of genes with few samples can lead to a large number of false positives if care is not taken. That is, genes which appear to be differentially expressed but are not truly differentially expressed are more likely to be discovered when there is a small sample size and a large number of tests. Thus, a multiple testing correction is necessary to reduce the number of false positives. The Benjamini-Hochberg multiple testing correction, which controls the false discovery rate is used in this work (Benjamini & Hochberg, 1995). This method is described in further detail in Section 1.4.3.

1.4. STATISTICAL METHODS FOR DIFFERENTIAL EXPRESSION TESTING FOR INDIVIDUAL MICROARRAY EXPERIMENTS

In this section, a review of the statistical methods needed for analyzing gene expression microarray data from an individual experiment are given. The preprocessing and differential expression methods that are used in this work are described. Although this section focuses on analyzing data from an individual experiment, the methods for the Affymetrix and Illumina arrays are chosen with the goal of utilizing similar methods that are comparable between the technologies.

1.4.1. Preprocessing: RMA and NEQC. As described in Section 1.3.3, preprocessing is an important part of microarray data analysis that includes background correction and normalization steps. The choice of preprocessing method greatly affects downstream analysis, thus normalizing in a consistent way across technologies and experiments is of great interest. With this consideration, the Robust Multi-Array Average (RMA) and the Normal Exponential Convolution model followed by Quantile Normalization (NEQC) were selected for the Affymetrix and Illumina technologies,

respectively. Both of these methods use a normal-exponential convolution model, which assumes the true probe signal follows an exponential distribution and background noise follows a normal distribution. Although these distributional assumptions are approximate, these methods are widely accepted and utilized in preprocessing microarray data (Silver et al., 2008).

RMA preprocesses Affymetrix microarrays through background correction, quantile normalization, $\log_2$ transformation, and probe intensity summarization using only the perfect match (PM) probes. For Illumina BeadChip microarrays, probe-level data are preprocessed using the NEQC procedure, which performs background correction using negative control probes, adds a positive offset to probe intensities, and then performs quantile normalization using negative control probes, positive control probes, and regular probes. A $\log_2$ transformation is then applied after normalization. The offset is only applied in the NEQC method, which balances bias and noise by providing an overall shift of the intensities away from zero (Shi et al., 2010). Both of these preprocessing methods are readily available in the form of R functions, namely, "rma" and "neqc" (RMA: Robust Multi-Array Average Expression Measure, n.d.; R: Normexp Background Correction and Normalization Using Control Probes, n.d.).

The purpose of background correction is to account for technical artifacts and ambient noise. This is accomplished through fitting a normal-exponential convolution model to estimate signal and noise through maximum likelihood estimation (R: Fit Normal+Exp Convolution Model to Observed Intensities, n.d.). This ambient noise is due to non-specific hybridization of the probes and technical artifacts such as optical noise from the scanning of the microarray (Ritchie et al., 2007). For Affymetrix arrays, let P_{ijk}

be the total PM intensity measurement, s_{ijk} be the true signal intensity, and b_{ijk} be the background intensity for array i, probe j, and probeset k. The observed PM intensity is modeled as follows:

$$P_{ijk} = s_{ijk} + b_{ijk} \quad (1)$$

$$s_{ijk} \sim Exp(\lambda_i) \quad (2)$$

$$b_{ijk} \sim N(\mu_i, \sigma_i) \quad (3)$$

The observed intensity (P_{ijk}) is the sum of the true and background signal intensities. The true signal (s_{ijk}) is assumed to follow an exponential distribution with mean λ_i and the background signal is assumed to follow a normal distribution with mean μ_i and variance σ_i^2. The background corrected values are estimated by finding $E(s_{ijk}|P_{ijk})$, which is strictly positive, as follows:

$$E(s_{ijk}|P_{ijk}) = a + b\frac{\varphi\left(\frac{a}{b}\right) - \varphi\left(\frac{P_{ijk}-a}{b}\right)}{\Phi\left(\frac{a}{b}\right) + \Phi\left(\frac{P_{ijk}-a}{b}\right) - 1} \quad (4)$$

where $a = P_{ijk} - \mu_i - \sigma_i^2\lambda_i$, $b = \sigma_i$, $\varphi(*)$ is the probability density function of the standard normal distribution, and $\Phi(*)$ is the cumulative distribution function of the standard normal distribution (Xie et al., 2009). The parameters λ_i, μ_i, and σ_i are estimated by either saddle-point approximation or maximum likelihood estimation using the saddle-point estimates as starting values when the saddle-point approximation struggles with numerical programming issues (Silver et al., 2008).

For Illumina BeadChip arrays, a similar approach is taken with the true signal assumed to follow an exponential distribution and the background assumed to follow a normal distribution. However, for Illumina arrays, negative control probes are used by

assuming they are the background signal which follow a normal distribution (Xie et al., 2009). Similar to Affymetrix arrays, the background corrected expression value is E(signal | observed). The parameters for the normal-exponential convolution model are estimated through maximum likelihood estimation utilizing all probes (negative control, positive control, regular probes). An offset (by default 16) is added to the background corrected expression values to shift the values away from 0, which would not be represented in the $\log_2$ scale (Shi et al., 2010).

Quantile normalization is useful for removing technical/experimental variation between arrays and allows comparisons across arrays by ensuring probes in each array have the same distribution of intensities. This method is necessary because an assumption of the statistical tests are constant variance of probe measurements between arrays. Quantile normalization involves taking the probe of minimum intensity from each array, calculating the mean of these lowest intensity probes, and setting all of these lowest probe intensities to be the calculated mean. This process continues with the next smallest intensity probes until all probes have been normalized (Bolstad et al., 2003). Note although the overall distribution of probe intensities are the same across all arrays, the ordering of the probes are different due to variation of a probe's intensity between arrays. Both the Affymetrix and Illumina arrays undergo quantile normalization. For Affymetrix arrays, only the PM probes are normalized. For Illumina arrays, the control probes and the regular probes are quantile normalized together. After quantile normalization the background corrected, quantile normalized probes undergo a $\log_2$ transformation.

For Affymetrix arrays, one final preprocessing step is needed since multiple probes map to a single gene. It is necessary to summarize across the multiple probes to

get an accurate measure of a gene's expression. This involves combining probe intensities into a single measurement of a gene's expression through Tukey's median polish which is robust to outlier probes (Irizarry et al., 2003). Tukey's median polish operates by utilizing an additive model of the form:

$$Y_{ijk} = \mu_{ik} + \alpha_{jk} + \epsilon_{ijk} \quad (5)$$

where Y_{ijk} is the background corrected, quantile normalized, $\log_2$ intensity of PM probes. μ_{ik} is the expression level for probeset k on array i, α_{jk} is the probe affinity effect for probe j in probeset k, and ϵ_{ijk} is the random error that is independent and identically distributed with mean zero. The parameters are estimated using Tukey's median polish and the estimate for μ_{ik} gives the expression level for probeset k on the $\log_2$ scale (Irizarry et al., 2003). Note that Illumina arrays do not require the summary step since the technical replicates of the probe are already summarized prior to the preprocessing through a trimmed mean procedure used by Illumina's BeadStudio software (Luo, 2007).

1.4.2. Differential Expression: LIMMA. A common goal of microarray experiments is to identify genes that are differentially expressed between conditions (e.g., disease vs healthy). Although many statistical methods have been developed to accomplish this goal, a popular approach is the LIMMA method, which is available in R/Bioconductor. LIMMA (linear models for microarray data) is a flexible modeling approach for differential expression analysis that utilizes gene-wise linear models that can be applied to data generated from many different types of genomic technologies (Smyth, 2004). By fitting a linear model to each gene, information across samples is conglomerated and the simplicity of the linear model allows for flexibility in incorporating different experimental design elements, conducting hypotheses tests, and

testing specific contrasts (Ritchie et al., 2015). For instance, using this framework a researcher has the ability to test for batch effects or interaction effects. This is specified by the design and contrast matrices created from R functions model.matrix() and makeContrasts(), respectively. The design matrix informs LIMMA of which samples are from which treatment group and the model parameters to be estimated while the contrast matrix informs LIMMA of which groups to compare.

A key feature of LIMMA is that it uses an empirical Bayes method to borrow information across genes to give a more precise estimate of gene-wise variability (Law et al., 2016). Shrinking the sample variances of the genes towards a pooled estimate allows for increased statistical power, which is especially important for the low sample sizes frequently encountered in microarray studies (Smyth, 2004). When comparing average expression levels between two groups, a p-value is derived from a moderated *t*-statistic with higher degrees of freedom than the classic *t*-statistic due to the degrees of freedom gained from prior information. Details of LIMMA's moderated *t*-test for the goal of comparing two groups are given below. The gene-wise hypotheses for LIMMA's moderated *t*-test are as follows:

H_o: There is no difference in population mean expression levels between groups i and i', namely $\mu_i - \mu_{i'} = 0$ [Gene is not differentially expressed]

H_a: There is a difference in population mean expression levels between groups i and i', namely $\mu_i - \mu_{i'} \neq 0$ [Gene is differentially expressed]

For each such contrast, a moderated *t*-statistic is calculated as follows:

$$t_{gi} = \frac{\bar{y}_i - \bar{y}_{i'}}{s_g\sqrt{v_{gi}}} \quad (6)$$

where i and i' designate the groups of interest and g represents the gene of interest being tested. $\bar{y}_i$ and $\bar{y}_{i'}$ represent the sample average expression values in groups i and i', s_g is the standard deviation of the gene of interest (estimated via empirical Bayes), and v_{gi} is equal to $\frac{2}{n}$ when the sample size is equal in the two groups, which is true in this study. Under the null hypothesis t_{gi} follows a t distribution with $d_o + d_g$ degrees of freedom, where d_o is the degrees of freedom from the prior information using all genes and d_g is the degrees of freedom from the gene of interest. This moderated *t*-statistic is significant (the null hypothesis is rejected) if it has a magnitude greater than the *t* critical value given by $|t_{gi}| > t_{\frac{\alpha}{2}, d_0 + d_g}$, where α is the significance level.

The empirical Bayes method for obtaining s_g works by assuming a prior distribution on the population variance (σ_g^2) of each gene, $\frac{1}{\sigma_g^2} \sim \frac{1}{d_o s_o^2} \chi_{d_o}^2$ where χ^2 represents the chi-squared distribution. The $d_o$and s_0^2 terms represent hyperparameters (degrees of freedom and variance, respectively) that are estimated from expression levels of all genes, see (Smyth, 2004) for estimation details. It is also assumed that $\hat{\sigma}_g^2 | \, \sigma_g^2 \sim \frac{\sigma_g^2}{d_g} \chi_{d_g}^2$, where d_g and $\hat{\sigma}_g^2$ are the residual degrees of freedom and sample variance for an individual gene. Using this information, Bayes' rule is then applied to yield the posterior mean for σ_g^2, namely $s_g^2 = E\left[\sigma_g^2 \middle| \hat{\sigma}_g^2\right] = \frac{d_o s_0^2 + d_g \, \hat{\sigma}_g^2}{d_0 + d_g}$. The square root of this estimate, s_g, is used in the denominator of the moderated t-statistic (Smyth, 2004). It reflects a combination of the gene's own sample variance and the prior variance obtained from all the genes, thereby shrinking the observed variances towards the prior values and increasing the degrees of freedom. Note that for the two-group comparison, it is assumed

that the samples are independent between and within groups, but the model and test can be easily modified to address technical replicates or paired samples. The LIMMA package in R allows for such flexibility (Ritchie et al., 2015). In addition to the distributional assumptions on the variance, the expression values within each group are assumed to follow a normal distribution with equal variance between groups.

After implementing LIMMA in R/Bioconductor, the results include the moderated *t*-statistic, the log fold change, a p-value, a q-value referred to as an adjusted p-value, and a B-statistic for each gene. The B-statistic is the log posterior odds ratio of differential expression. P-values give the probability of obtaining a test statistic at least as extreme as the one obtained if the null were true. Thus, small p-values (less than α) result in rejecting the null hypothesis and concluding the gene is differentially expressed. P-values are calculated individually for each gene using the $t_{d_0+d_g}$ distribution described previously. The q-value makes an adjustment to the p-values that provides a way to control the number of false positives among the genes where the null is rejected. The need for this adjustment is due to the multiple tests conducted (one for each gene) and the method used to handle this issue is described in the next section.

1.4.3. Multiple Testing Corrections. Using the LIMMA approach for differential expression testing, a hypothesis test is conducted for each gene, resulting in thousands of tests for a single experiment. As discussed in Section 1.3.3, this can result in a higher probability of finding false positive results across the set of tests compared to conducting a single test. Multiple testing procedures provide a way to combat this issue and provide a stricter control on the number of false positives across the set of tests. Although there are many possible multiple testing methods available, the Benjamini-

Hochberg approach to controlling the false discovery rate (FDR) is widely used in differential expression testing (Pawitan et al., 2005).

The FDR is the expected proportion of significant tests (rejected null hypotheses) that are false. The Benjamini-Hochberg approach to controlling the false discovery rate at level L is as follows. Let $P_{(i)}$ be the i^{th} ordered p-value from m hypotheses tests such that $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(m)}$. Find the largest rank, r, which satisfies $P_{(r)} \leq \frac{r}{m} L$. All $P_{(i)}$ for $i = 1, 2, \ldots, r$ are considered significant (null is rejected). Furthermore, the q-value, which is the adjusted p-value, is given by $P_{(i)}^{(adj)} = \min_{i \leq r}\{\frac{m}{r} P_{(r)}\}$ (Benjamini & Hochberg, 1995). This provides a way to compare the adjusted p-value to the desired FDR level L. Any gene with $P_{(i)}^{(adj)} < L$ is declared to be differentially expressed.

1.5. META-ANALYSIS TECHNIQUES FOR GENE EXPRESSION DATA

With the wealth of gene expression data publicly available in databases such as NCBI GEO, it is possible for researchers to integrate data from multiple studies to address new research questions or test the robustness of previous findings. Meta-analysis methods provide one way to combine data from multiple independent studies to identify a set of common differentially expressed genes. Since many microarray studies often have relatively small sample sizes, combining data from multiple studies that investigate differences between the same set of conditions can improve the statistical power of detecting differentially expressed genes. The results can also yield more robust findings by identifying common biomarkers that emerge from an analysis that incorporates

multiple independent studies that may utilize different technologies (Toro-Domínguez et al., 2021).

The choice of the meta-analysis method is critical in appropriately analyzing multiple experiments together to address a specific research question. Experiments performed with different conditions or platforms are better analyzed with a different method than experiments performed with the same microarray platform or the same condition of interest. For example, some methods are sensitive to false negatives while others are sensitive to false positives and therefore lead to different results. Interpretation of significance is another consideration as some methods take significance of a gene to mean a gene is statistically significant in all studies (HSA), while other methods consider significance when a gene is statistically significant in at least one study (HSB). Certain methods are also better when the number of available studies is small. A flow chart (Figure 1.17) provided by Toro-Dominguez et. al. (2021) gives guidance on the appropriate meta-analysis method to use in different situations. In this research, the microarray data available are from different platforms (Affymetrix and Illumina) and there are only two studies. Thus, the rank product method is the most appropriate meta-analysis method based on this approach for selecting an analysis (Toro-Dominguez et al., 2021). In this section, each of the meta-analysis methods listed in the flowchart is briefly described following the Toro-Dominguez et al. (2021) review paper. A more thorough review is given for the rank product method utilized in this work.

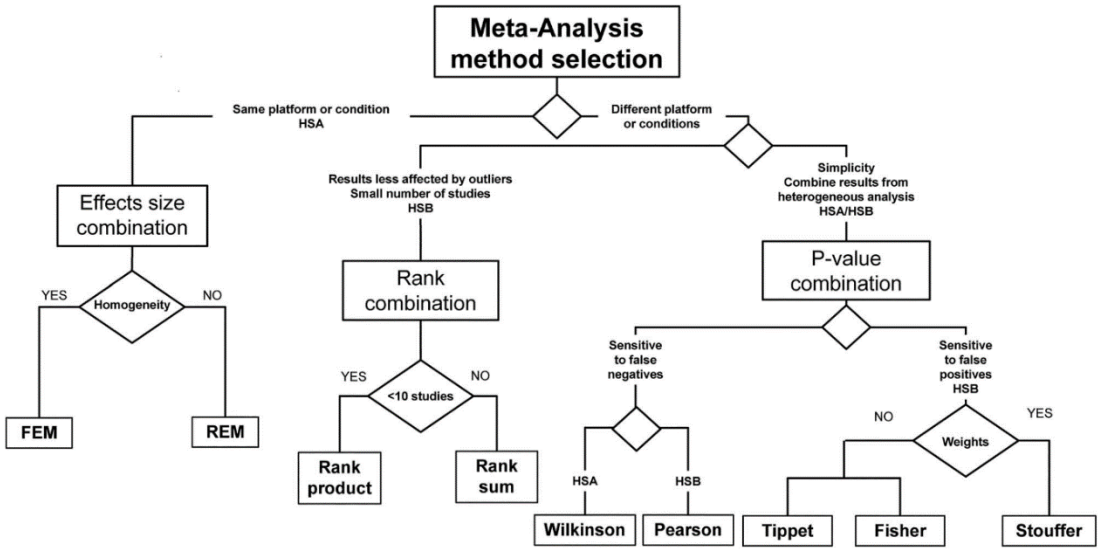

Figure 1.17 Meta-Analysis Method Selection Flowchart. Figure from Toro-Domínguez et al. (2021).

1.5.1. P-Value Based Methods. Using a p-value based method involves obtaining p-values from each study individually and then combining those p-values using an appropriately chosen method. P-value based methods are particularly useful if there are different platforms or conditions between experiments since only the p-values are considered and the individual measurements are not. Further choice of method depends on interpretation of significance and sensitivity to either false positives or false negatives. For instance, if it is important to consider the sensitivity to false negatives, the HSA condition is best handled with the Wilkinson method and the HSB condition works best with the Pearson method. A study that prioritizes sensitivity to false positives with the HSB condition is best handled by the Tipper, Fisher, or Stouffer methods (Toro-Domínguez et al., 2021). A disadvantage of these methods is the loss of fold change directionality, which is preserved in the effect size based methods. The p-value does not

maintain any information about the gene expression fold change so separate analyses of upregulated and downregulated genes are useful. Furthermore, outliers can also contribute to false positives. Despite these disadvantages, the p-value based methods are attractive due to their ability to combine heterogeneous datasets as well as the simplicity of their application (Toro-Domínguez et al., 2021).

1.5.1.1. Fisher's method. The null hypothesis of Fisher's method is that the gene being tested is not differentially expressed across studies. If a gene is significant using Fisher's method, then it is significant in one or more studies. The statistic of the Fisher's method is given by the following where p_i is the p-value of a gene for study i.

$$-2 * \sum_{i=1}^{k} \ln (p_i) \tag{7}$$

This statistic follows a Chi-Squared distribution under the null hypothesis with $2 * k$ degrees of freedom where k is the number of studies. Furthermore, this method is sensitive to small p-values such that a significant combined p-value can be obtained from a single small p-value (Toro-Domínguez et al., 2021).

1.5.1.2. Pearson's method. Pearson's method is quite similar to the Fisher's method though sensitive to large p-values instead of small ones. Because of this, more false negatives are obtained compared to Fisher's method which is more sensitive to false positives. The Pearson's method statistic is given by:

$$-2 * \sum_{i=1}^{k} \ln (1 - p_i) \tag{8}$$

where p_i is the p-value of a gene for study i. Similar to Fisher's method, this statistic follows a Chi-Squared distribution under the null hypothesis with $2 * k$ degrees of freedom where k is the number of studies (Toro-Domínguez et al., 2021).

1.5.1.3. Stouffer's method. The test statistic of Stouffer's method is given by:

$$\frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}} \tag{9}$$

where Z_i is equal to $\Phi^{-1}(1 - p_i)$. Φ represents the standard normal cumulative distribution function, p_i is the p-value of a gene for study i, and k is the number of studies. Under the null hypothesis that the gene being tested is not differentially expressed between studies, this statistic follows a standard normal distribution. A weighted version of the statistic is also available if weights can be calculated from the inverse variance of the statistics on which the p-values were found. In this scenario, the weighted Stouffer's method gives more reliable results than Fisher's method. Let ω_i be the calculated weight from the inverse variance of the statistic used to obtain a study's p-value, then the weighted statistic is given by:

$$\frac{\sum_{i=1}^{k} \omega_i Z_i}{\sqrt{\sum_{i=1}^{k} \omega_i^2}} \tag{10}$$

If the inverse variance of the statistics which made the p-values for each study cannot be calculated, the square roots of the sample sizes for each study may be used instead (Toro-Domínguez et al., 2021).

1.5.1.4. Tippet's method. Tippet's method is simple in that statistic of interest is the minimum p-value for a gene in a set of k studies. This statistic follows a Beta(1,k) distribution under the null hypothesis that the gene is not differentially expressed between studies. Since only the smallest p-value for a gene is considered, this method finds a gene significant if statistically significant in any of the studies (Toro-Domínguez et al., 2021).

1.5.1.5. Wilkinson's method. On the other end of Tippet's method is Wilkinson's method whose statistic of interest is the maximum p-value for a gene across k studies. This is the only p-value based method discussed in the Toro-Dominguez et al. (2021) paper in which a significant gene is one that is statistically significant across all of the studies. This statistic follows a Beta(k,1) distribution under the null hypothesis that the gene is not differentially expressed (Toro-Domínguez et al., 2021).

1.5.2. Effect Size Based Methods. Effect size based meta-analysis methods combine the effect or evidence of differential expression across studies. Effect size is most commonly calculated using Hedges' g estimator (Toro-Domínguez et al., 2021). Given the null hypothesis that there is no differential expression between treatment groups for a given gene, the effect size denoted ES_i for data set i is given below (Toro-Domínguez et al., 2021):

$$ES_i = c(m)\frac{\bar{y}_M - \bar{y}_W}{s} \tag{11}$$

where $\bar{y}_M$ is the average expression values from n_M "mutant-type" or case patients and $\bar{y}_W$ is the average expression value from n_W "wild-type" or control patients. The degrees of freedom is given by $m = n_M + n_W - 2$ and $c(m)$ is a constant that corrects positive bias of this statistic, which is equal to $1 - \frac{3}{4m-1}$. Note $S = \sqrt{\frac{(n_M-1)S_M^2+(n_W-1)S_W^2}{n_M+n_W-2}}$, namely, the pooled standard deviation between studies where S_M^2 and S_W^2 are the sample variances of the mutant and wild type patients, respectively. Furthermore, the variance of ES_i is given by:

$$V(ES_i) = \frac{n_M+n_W}{n_M*n_W} + \frac{(ES_i)^2}{2(n_M+n_W)} \tag{12}$$

Note that these equations specify the effect size of a single gene between groups for a single study. There are two ways to combine the individual gene effect sizes into a combined effect size across studies, namely, either the fixed effects model (FEM) or the random effects model (REM) (Toro-Domínguez et al., 2021).

1.5.2.1. Fixed effects model. In the fixed effects model, different studies are assumed to share a common effect size and that differences in effect size are due to sampling error (Borenstein et al., 2009). Because of this, the FEM is only appropriate when all samples are coming from the same population or when there is homogeneity between the samples (Toro-Domínguez et al., 2021). In the FEM, the combined effect size is:

$$\overline{ES} = \frac{\sum \omega_i ES_i}{\sum \omega_i} \tag{13}$$

where ω_i are the weights assigned to each experiment using the within-study variance, $V(ES_i)$, and are defined as:

$$\omega_i = \frac{1}{V(ES_i)} \tag{14}$$

The variance of $\overline{ES}$ is given by:

$$V(\overline{ES}) = \frac{1}{\sum \omega_i} \tag{15}$$

and the combined effect value is given by:

$$Z = \frac{\overline{ES}}{\sqrt{V(\overline{ES})}} \tag{16}$$

The combined effect Z follows a standard normal distribution, which yields a two-tailed p-value given by:

$$P = 2[1 - \Phi(|Z|)] \tag{17}$$

where $\Phi(*)$ is the standard normal cumulative distribution function.

1.5.2.2. Random effects model. The random effects model (REM) assumes that the true effect differs between studies, unlike the FEM which assumes a shared true effect between studies. The combined effect is then the expected value of the population of true effect sizes (Toro-Domínguez et al., 2021). The combined effect size ($\overline{ES}^*$) is the same as the FEM (Equation 13) but the weights of the REM (ω_i^*) are calculated differently than the weights of the FEM (ω_i) because there are now two sources of variance: within-study and between-study variance. Let Q represent the total variance of k studies, ES_i be the observed effect of a study, ω_i be the calculated FEM weight, and $\overline{ES}$ be the FEM combined effect (Toro-Domínguez et al., 2021). Then Q is given by:

$$Q = \sum_{i=1}^{k} \omega_i (ES_i - \overline{ES}) \tag{18}$$

Let τ^2 represent the between-study variance that was not present in the FEM and df represent the degrees of freedom equal to $k - 1$. Then τ^2 can be derived from Q using the following equations (Toro-Domínguez et al., 2021).

$$\tau^2 = \begin{cases} \frac{Q-df}{C} \; if \; Q > df \\ 0, Otherwise \end{cases} \tag{19}$$

where

$$C = \sum_{i=1}^{k} \omega_i - \frac{\sum_{i=1}^{k} \omega_i^2}{\sum_{i=1}^{k} \omega_i} \tag{20}$$

Let $v_i^* = V(ES_i) + \tau^2$ then the weights assigned to each study are given by:

$$\omega_i^* = \frac{1}{v_i^*} \tag{21}$$

Then similar to the FEM, the variance of the combined effect, test statistic of the combined effect (Z), and associated p-value (P) are as follows:

$$V(\overline{ES}^*) = \frac{1}{\sum_{i=1}^{k} \omega_i^*} \tag{22}$$

$$Z = \frac{\overline{ES^*}}{\sqrt{V(\overline{ES^*})}} \quad (23)$$

$$P = 2[1 - \Phi(|Z|)] \quad (24)$$

The assumption of varying true effect between studies of the REM is more biologically reasonable than the FEM and allows for heterogeneity between studies (Toro-Domínguez et al., 2021). For these reasons, the REM is more often utilized than the FEM.

1.5.3. Nonparametric Based Methods. The rank-based methods are an attractive choice for meta-analysis due to their few, weak assumptions about the data, and their non-parametric nature allowing for analysis of heterogeneous datasets. For instance, rank-based methods are readily available to conglomerate several technologies into one analysis (Toro-Domínguez et al., 2021). The technologies may vary in what they are measuring (e.g., DNA methylation, RNA sequencing, and gene expression) or in the platform for given type of data (e.g. Affymetrix vs. Illumina gene expression microarrays). Rank-based methods allow for an increase in power by pulling information from these numerous technologies when, typically, there are very few samples from any one technology or experiment. This is the case for the datasets available in this work, which were collected from two different microarray technologies (Affymetrix and Illumina) and have a limited number of samples in each experiment (3 and 7 biological replicates per group, respectively). This motivates the use of a rank-based method for this work.

The assumptions about the datasets for the rank-based methods include: a small number of genes are differentially expressed, independence of measurements between arrays, the differential expression of most genes are independent of one another, and the

measurement variance is the same for all genes (Breitling et al., 2004). The underlying justification for the rank-based methods is simple and biologically motivated: whether upregulated or downregulated, genes with low rank have the strongest evidence of differential expression for that array and genes with low rank across several arrays are the most likely to be differentially expressed.

Ranks for each gene are determined by the $\log_2$ fold change (Hong et al., 2023). This log fold change is determined by the disease (D) and control (C) expression levels (Y_{ij}^{D} and Y_{ij}^{C}, respectively, for study i and array/sample j) and then calculating their $\log_2$ fold changes, $\log_2(\frac{Y_{ij}^{D}}{Y_{ij}^{C}})$, for all pairs of samples within a study to yield $k = 1, \dots, K$ total pairs of fold changes across studies (Hong & Breitling, 2008). Note that $K = \sum_i n_i^D * n_i^C$, where n_i^D and n_i^C are the sample sizes for the disease and control groups in study i. The ranks (r_{gk}) for each gene g are obtained by ranking the fold changes within each comparison/pair k. The ranking is performed two different ways, depending on whether the test is for upregulation or downregulation. For upregulation, the largest fold changes is assigned a rank of 1 and for downregulation the smallest fold change is given a rank of 1 (Heskes et al., 2014). Next the ranks are either summed in the rank sum method or multiplied together in the rank product method depending on the computational demand of the meta-analysis of interest.

1.5.3.1. Rank product. Combining the calculated ranks across arrays through a geometric mean is the essence of the rank product (RP) method (Eisinga et al., 2013). Let p_g^C denote the true expression level for gene g in the control group and let p_g^D denote the true expression level for gene g in the disease group. Then the null hypothesis of the rank product method is that no genes are differentially expressed, namely $\log_2\left(\frac{p_g^D}{p_g c}\right) = 0$ for all g.

Let RP_g denote the rank product statistic of gene g considering $k = 1, \ldots, K$ pairs of fold changes as described in Section 1.5.3. The rank product statistic is given by:

$$RP_g = \left(\prod_{k=1}^{K} r_{gk}\right)^{1/K} \tag{25}$$

Note that as an alternative to taking all pairs, a subset of random pairings can be selected equal to the smaller of the sample sizes in the disease or control group for each study. Multiple datasets with these random pairings will be generated and the rank product statistic calculated. The median of the rank product statistics across the datasets is taken as the test statistic. This approach was developed to help reduce the false discovery rate (Del Carratore et al., 2017; Bioconductor Rankprod Package Vignette, n.d.).

The rank product method originally utilized a permutation testing method for calculating p-values for each gene. Essentially, permutations of ranks randomly assigned to genes were formed and then the p-value was determined by calculating the proportion of these random permutations' ranks products that were smaller than the calculated rank product statistic for a gene.

Due to the time-consuming nature of the permutation testing for large sample sizes and accurate p-values, Koziol sought to find a method for estimating the p-values without permutations (Eisinga et al., 2013). This was done by relating the RP distribution to the Gamma distribution with the null hypothesis of interest being no genes are differentially expressed. Using this null hypothesis, Koziol proposes an approximate uniform distribution for $\frac{r_{gk}}{G+1}$ on the interval [0,1] for G genes, thus $-\log\left(\frac{r_{gk}}{G+1}\right)$ approximately follows an exponential distribution with scale parameter equal to 1. Furthermore, the sum of independent exponential random variables with scale parameter, 1, follows a gamma distribution with the same scale parameter and a shape parameter, K, equal to the number of independent, identically distributed exponential random variables summed together. In context, let K be the pairs/comparisons, as described previously, then the log-transformed random variable, $-\log\left(RP_g\right) + K * log(G+1)$, approximately follows a gamma distribution with shape parameter, K, and scale parameter, 1 under the null hypothesis (Koziol, 2010).This distribution can then be used to calculate p-values. While this is overall an accurate and computationally efficient approximation compared to the permutation re-sampling method, this method falls short in the tails of the distribution, which contains the most interesting insights into differential expression (Heskes et al., 2014). An exact distribution was of great interest to remedy this approximation error.

Tom Heskes, Rob Eisinga, and Rainer Breitling introduced the exact probability distribution of the rank product (Eisinga et al., 2013). Shortly afterwards, this influential group published a method for quickly and accurately approximating p-values by using the geometric mean of a lower bound and a slightly conservative upper bound of the exact p-

value (Heskes et al., 2014). Deriving p-values from the exact probability distribution is quick for the most interesting genes, i.e. genes with small RP values, but requires large computation times for the exact p-value when the RP statistic is large. This motivated the use of the geometric mean of bounds of the exact p-value as a computationally feasible and accurate approximate p-value calculation. This method is implemented using the R function RP.advance() (Bioconductor Rankprod Package Vignette, n.d.). Note that since the RP test is applied to all G genes, it is important to control the false discovery rate, as described in Section 1.4.3.

1.5.3.2. Rank sum. The statistic for the rank sum is quite similar to the rank product found by summing the ranks instead of finding the product of the ranks. Let RS_g denote the rank sum statistic of gene g considering K comparisons. The rank sum statistic is given by:

$$RS_g = \sum_{k=1}^{K} r_{gk} \tag{26}$$

Although less robust than the rank product, the rank sum requires less time to compute. Thus, the rank sum should be used if there are an infeasibly large number of arrays being analyzed (Toro-Domínguez et al., 2021).

2. METHODS

2.1. DATA

In this work, two historical datasets were identified from the NCBI GEO database that met criteria (described in Section 1.3) of including gene expression (transcription) level measurements derived from some type of MSC in obese and non-obese patients. In one of the datasets (GEO accession: GSE48964), gene expression levels from adipose stem cells were measured on an Affymetrix GeneChip array (Oñate et al., 2013). In the second dataset (GEO accession: GSE107214), stem cells were collected from Wharton's Jelly and gene expression was measured with an Illumina Expression BeadChip (Badraiq et al., 2017). The goal of this work is to combine information from both of these past experiments to identify a common set of differentially expressed genes derived from MSCs in obese and non-obese patients. This approach has the potential to reveal new insights through the use of meta-analysis to improve statistical power and detecting genes that show a robust signal across the array technology and type of MSC. Each of the two datasets is described in detail below.

2.1.1. Affymetrix Dataset. The experiment referred to herein as the "Affymetrix Dataset" collected subcutaneous abdominal white adipose tissue stem cells from 6 patients, 3 obese and 3 non-obese, with no technical replicates (Oñate et al., 2013). These adipose-derived stem cells (ASC) are a form of mesenchymal stem cell. The threshold for non-obesity was BMI less than 25 $\frac{kg}{m^2}$ while the threshold for obesity was BMI greater than 40 $\frac{kg}{m^2}$. These gene expression values measuring transcriptional activity were derived using the Affymetrix's Human Gene 1.0 ST Array GeneChip technology.

Beginning with 33,297 probesets, there are a number of filters necessary to adequately prepare the dataset. The R statistical software (Version 4.2.2) was utilized for this filtering task. Ultimately, it is important to identify a common set of genes with high-quality measurements that are represented on both array technologies. For each array, this involves obtaining a dataset with one gene per row and one sample or array per column. First, probesets that did not have a known gene symbol association were filtered out using the mapIDs() function in R and the hugene10sttranscriptcluster.db annotation file. To combat the many-to-many relationship between probesets and genes, only probesets mapping to exactly 1 gene are considered. This was done using the mapIDs() function with the "filter" option in R (AnnotationDBI: Introduction to Bioconductor Annotation Packages, n.d.). After removing probesets that did not map to any gene or that mapped to more than one gene, 19,922 probesets remained. Furthermore, 1 probeset can map to many genes, but 1 gene can also map to many probesets. The unique() function in R completes the one-to-one relationship between probesets and genes. The matter of selecting which gene-to-probeset relationship to keep for the meta-analysis was determined by the p-value from LIMMA analysis. The gene-to-probeset mapping with the smallest LIMMA p-value, and thus the most statistically significant, was selected for further analysis. After removing probesets with no affiliated gene, probesets that map to multiple genes, and determining unique probesets for each gene, 18,876 genes remain for analysis.

2.1.2. Illumina Dataset. The experiment referred to herein as the "Illumina Dataset" collected Wharton's Jelly mesenchymal stem cells (WJ-MSC) from the umbilical cord of 14 patients, 7 obese and 7 non-obese, with 3 technical replicates each

for a total of 42 samples (Badraiq et al., 2017). All subjects were women at the gestational age of 37 weeks who had Caesarean section deliveries without complications. The thresholds for non-obesity was BMI less than or equal to 25 kg/m^2 while the threshold for obesity was BMI of greater than or equal to 30 kg/m^2. These expression values were derived using Illumina's Human HT-12 v4.0 Expression BeadChip technology.

Beginning with 47,323 probes, the same filters used in the Affymetrix Dataset were applied to the Illumina Dataset. Annotation file, "HumanHT-12_V4_0_R2_15002873_B.bgx", provided the gene symbols for most probes, though 44,053 probes were left after filtering out those with no gene associations. Note, although less prevalent than in the Affymetrix dataset, there were probes mapping to multiple genes. Using LIMMA to determine the most statistically significant, unique probe-to-gene relationship, 31,426 genes remain for analysis. Furthermore, the intersection of these 31,426 genes from the Illumina Dataset and the 18,876 genes from the Affymetrix Dataset revealed 15,767 genes shared between the two. The expression levels associated with these 15,767 genes were utilized in the rank product analysis. Note, although the LIMMA differential expression analysis was run separately for Affymetrix and Illumina datasets, this intersection of 15,767 genes was used to make the Venn diagrams in Section 3.1 representing the number of shared statistically significant genes between the two technologies.

2.2. PREPROCESSING: RMA AND NEQC

2.2.1. RMA and the Affymetrix Dataset. The Affymetrix Dataset consisted of CEL and CHP file types used by Affymetrix, where CEL files hold the probe-level intensities and CHP files contain the gene-level information. Once the files have been extracted, they are called with function, read.celfiles(), which make an object of class "gene feature set". This gene feature set contains data about intensities, assays, phenotype, experiment, and protocol. The gene feature set object is the input to the rma() function which performs the RMA steps described in Section 1.4.1 and creates an object of class "expression set". This expression set object is part of the input for the lmFit() function used in differential expression analysis with LIMMA.

2.2.2. NEQC and the Illumina Dataset. The Illumina Dataset consisted of the "idat" file type used by Illumina to store the probe-level BeadArray data. Once extracted, these idat files are read into R with the function read.idat() which creates an object of class "e list raw". This "e list raw" object contains the expression values on a raw scale as opposed to the log2 scale utilized later in the analysis and serves as the input for the neqc() function. Note gene annotation for each probe is automatic in this step by specifying the "HumanHT-12_V4_0_R2_15002873_B.bgx" annotation file in the neqc() function. Utilizing neqc() with default setting and an offset of 16 preprocesses the intensities as described in Section 1.4.1 and creates an object of class "E List" which serves as the input for lmFit() for differential expression analysis in LIMMA.

2.3. DIFFERENTIAL EXPRESSION: LIMMA

An individual differential expression analysis is performed on each of the two microarray datasets using the LIMMA method described in Section 1.4.2. The goals of this analysis are three-fold. First, differentially expressed genes can be compared between the two datasets (for genes represented on both arrays) to determine if there are any common significant genes. Second, for genes that map to more than one probe/probeset, the LIMMA p-value is utilized to select which probeset to represent the gene. The probeset with the smallest LIMMA p-value is selected. Finally, the genes identified as differentially expressed in LIMMA and the rank product method will be compared to identify a robust set of genes that are identified in multiple methods. Details about the implementation of LIMMA in R/Bioconductor are described below.

The first step in implementing LIMMA is crafting the design matrix that specifies which samples are from obese or non-obese patients. In this study, a "L" arbitrarily designates the non-obese patients and an "O" designates the obese patients. Let "OorL" denote this list of L's and O's, which is an input to the function, model.matrix() that creates the design matrix. Note the means model was utilized using the notation, model.matrix(~0+OorL).

Although run separately, most of the implementation of the LIMMA method is the same between the Affymetrix Dataset and the Illumina Dataset. A major difference is the need to aggregate the technical replicates from the Illumina Dataset using the function, duplicateCorrelation(). This function assumes between-replicate correlation of the technical replicates is constant across genes in order to improve the precision of the estimation of gene-wise variance more so than a simple average (Smyth et al., 2005).

Next, linear models are fit to each gene using lmFit(). For the Illumina Dataset, the "block" option was specified as a vector designating the biological replicates and the "cor" option was specified as the "consensus" output from duplicateCorrelation().

After fitting the linear models, the contrast matrix is made to designate which comparisons are of interest. Namely, in each dataset the contrast between obese and non-obese patients is of interest. The contrast matrix was made using the function, makeContrasts() specifying the relevant contrasts of interest, namely "O-L", and the "levels" option to be the design matrix. Next the function, contrasts.fit(), estimates the coefficients and standard errors of the linear model for this contrast (Contrasts.fit: Compute Contrasts from Linear Model Fit, n.d.)

This contrast fit is the input for the empirical Bayes method which smooths the variances for individual genes towards a shared value. This is implemented using the function, eBayes(), with the "robust" option set to "TRUE". The output of this method is a list of differentially expressed genes between non-obese and obese patients. The most statistically significant genes can be retrieved and filtered with the topTable() function as well as exported to a csv file using the function, write.csv(). Note the adjusted p-value in this output is the gene's Benjamini-Hochberg multiple testing corrected p-values (as described in Section 1.4.3).

The genes from the Illumina Dataset have been annotated by the use of the neqc() function, but the Affymetrix results were not yet annotated with gene symbols. Instead, the probeset ID has been used until this point. Annotating the Affymetrix Dataset probeset IDs with gene symbols is done using the hugene10sttranscriptcluster.db file and the mapIds() function with key type equal to "PROBEID" and multivals option equal to

"filter". This filter option removes any probes that map to multiple genes, which helps with the probe-to-gene direction of the necessary one-to-one relationship between probes and genes to run the rank product method. Furthermore, some probes do not have associated genes and will give "NA" as the gene's associated symbol. These probes with no associated genes are removed using the complete.cases() function. For both the Affymetrix and Illumina Datasets, a unique relationship between probes-to-genes was established by removing all gene-to-probe mappings other than the most statistically significant gene-to-probe mapping as determined by the adjusted p-value from the LIMMA output using the duplicated() function. Now the LIMMA results contain only the single most significant probe-to-gene relationship with no missing values.

2.4. META-ANALYSIS: RANK PRODUCT

Note the lists of genes between the two technologies are different due to differences in probes used. In order to run the rank product method, only genes common to both technologies can be considered. The inner_join() function completes this necessary set up by making a list of only genes shared between the two technologies, namely 15,767 genes. Note every column is a microarray sample and every row is the log2 expression values of a gene with exactly 1 row referring to 1 gene.

The function, RP.advance(), runs the rank product method and requires some other inputs including class, origin, and gene name. The class refers to the whether the microarray sample is from an obese or non-obese patient while the origin refers to the experiment or laboratory that produced the microarray sample. Note option "calculateProduct" is set to "TRUE" in order to run the rank product method. If

“calculateProduct” is set to “FALSE” then the rank sum method is run. Now that the rank product method has been applied and a list of genes, p-values, and q-values have been created, the results are separated by upregulated and downregulated genes and exported to csv files.

Of interest are the ranked gene lists of the rank product themselves, but also the intersection between the LIMMA and rank product results. Certainly, there is a higher degree of confidence for differential expression between genes determined significant by both parametric and non-parametric methods. The adjusted p-value and q-value (both estimates of the FDR) from the LIMMA results and rank product results, respectively, were used to determine statistically significant gene intersections with the inner_join() function.

3. RESULTS

3.1. LIMMA

The Affymetrix Dataset found very few significant genes in the LIMMA analysis, 1 for FDR (adj. P-Value) ≤ 0.05 and 3 for FDR ≤ 0.10. On the other hand, the LIMMA analysis of the Illumina Dataset provided many significant results, 468 for FDR ≤ 0.05 and 809 for FDR ≤ 0.10. There were no differentially expressed genes in common between the two datasets. The Venn diagrams showing the intersection of significant genes (at 0.05 and 0.10 FDR levels) between the Affymetrix and Illumina Datasets are given below in Figures 3.1 and 3.2. Tables of the top 20 significant genes using the adjusted p-values are given in Tables 3.1 and 3.2. For each gene in the LIMMA output, "Symbol" is its associated gene symbol, "logFC" is its log fold-change between disease and healthy groups, "AveExpr" is its overall average log-expression values, "t" is its *t*-statistic, "P.value" is its associated p-value of that *t*-statistic, "adj.P-Value" is its Benjamini-Hochberg multiple testing corrected p-value, and "B" is its log-posterior odds ratio. It should be noted that only genes represented on both arrays are included in this comparison.

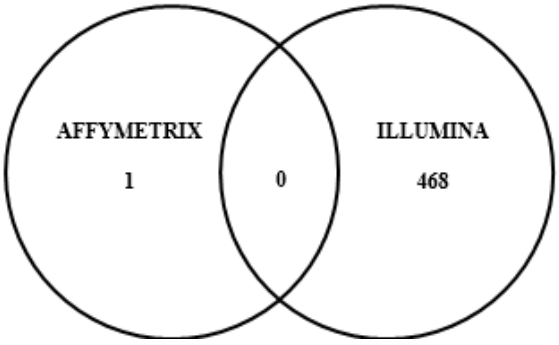

Figure 3.1 Venn Diagram of Significant Genes found by LIMMA in the Affymetrix and Illumina Datasets at FDR ≤ 0.05.

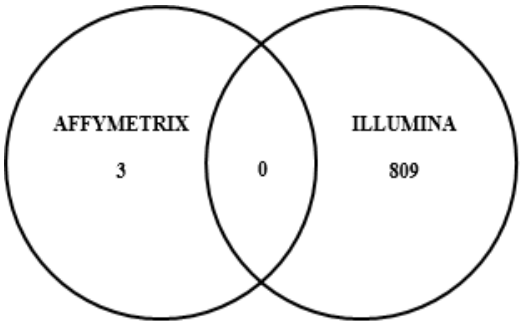

Figure 3.2 Venn Diagram of Significant Genes found by LIMMA in the Affymetrix and Illumina Datasets at FDR ≤ 0.10.

Table 3.1 Top 20 Significant Genes from LIMMA Analysis of Affymetrix Dataset

Symbol	logFC	AveExpr	t	P.Value	adj. P-Value	B
FOS	-4.18448	10.70464	-16.8476	2.40E-06	0.039923	3.590005
FOSB	-4.10931	9.668905	-14.1487	6.80E-06	0.075426	3.165728
SNORD14E	-1.62045	7.320688	-12.5277	1.20E-05	0.099974	2.940161
H2BC21	-1.61267	9.922421	-10.9779	2.67E-05	0.126902	2.492363
NR4A1	-3.01399	9.925258	-10.8341	3.28E-05	0.127766	2.340112
ACTA2	-0.80466	12.336	-9.31922	6.97E-05	0.128977	1.881272
VN1R1	0.639423	6.030341	9.707352	5.48E-05	0.128977	2.04184
NR4A2	-3.66304	9.325393	-9.51166	6.97E-05	0.128977	1.862473
CXCL2	-3.97925	8.463512	-9.51272	6.97E-05	0.128977	1.862901
EGR1	-1.49552	12.41596	-9.41508	6.81E-05	0.128977	1.891909
CLK4	0.799922	7.75098	9.585365	5.91E-05	0.128977	1.992536
TNFAIP3	-1.68934	9.633427	-10.3242	4.01E-05	0.128977	2.234926
UQCRB	0.675286	6.593843	9.971584	4.68E-05	0.128977	2.145146
SPON1	1.304484	8.054788	9.226348	7.41E-05	0.129855	1.838994
FRY	-0.7781	8.234832	-8.29757	0.000137	0.168744	1.402293
TRIM52	0.56355	7.556802	8.37612	0.00013	0.168744	1.442287
MTRES1	0.600237	7.584736	8.372787	0.00013	0.168744	1.440601
SUGCT	-0.92008	7.070217	-8.34295	0.000133	0.168744	1.425468
MEST	-0.69408	11.87815	-8.20555	0.000146	0.168744	1.354717
MCMDC2	0.56354	6.721801	8.116737	0.000155	0.172515	1.308042

Table 3.2 Top 20 Significant Genes from LIMMA Analysis of Illumina Dataset

Symbol	logFC	AveExpr	t	P-Value	adj. P-Value	B
TIPARP	1.100755	10.04988	10.24939	5.10E-13	2.41E-08	18.29808
PFAAP5	-0.70257	7.303626	-8.71604	5.55E-11	1.31E-06	14.23974
KCNK6	0.574748	9.009732	8.246865	2.46E-10	3.88E-06	12.92758
NR4A2	1.30919	5.673715	8.053705	4.57E-10	5.41E-06	12.37869
CYP1B1	2.085074	9.257347	7.784656	1.09E-09	1.03E-05	11.6063
SLC22A4	0.604954	7.632294	7.564351	2.23E-09	1.76E-05	10.96747
ZFP36	0.977747	9.345253	7.279023	5.68E-09	3.84E-05	10.13233
CTPS	0.583268	10.13531	7.131637	9.22E-09	5.45E-05	9.697828
FDFT1	-0.38828	12.33779	-7.07697	1.08E-08	5.66E-05	9.558904
ACSS2	-0.51595	9.903562	-6.74619	3.29E-08	0.000156	8.55303
UBA7	-0.59202	8.303163	-6.69745	3.87E-08	0.00016	8.407539
FKBP14	0.391501	10.52148	6.679269	4.06E-08	0.00016	8.36424
LOC440895	1.039098	7.643119	6.594955	5.43E-08	0.000185	8.101094
ACACA	-0.43072	9.801296	-6.59257	5.48E-08	0.000185	8.093948
GOLIM4	0.738393	7.637447	6.527131	6.80E-08	0.000202	7.898001
MGST1	0.525969	10.56022	6.525604	6.84E-08	0.000202	7.893428
HIST2H2BE	0.559808	7.597725	6.401505	1.03E-07	0.000271	7.521265
LIMS3	0.846456	5.361331	6.357816	1.19E-07	0.000297	7.390097
TOMM5	0.366382	9.990882	6.204759	1.97E-07	0.000465	6.938225
ATF3	1.028282	8.70235	6.176535	2.18E-07	0.000485	6.84521

3.2. RANK PRODUCT

The rank product analysis found 1093 upregulated genes for PFP ≤ 0.05 and 1406 upregulated genes for PFP ≤ 0.10 as well as 1091 downregulated genes for PFP ≤ 0.05 and 1316 downregulated genes for PFP ≤ 0.10. For each gene in the rank product output, "Symbol" is its associated gene symbol, "RP Statistic" is its test statistic as described in Section 1.5.3.1, "Rank" is its rank according to the RP Statistic, "PFP" is its estimated percentage of false predictions (an estimate of the FDR), "P-value" is its associated p-value, and "AveFC" is the log fold-change of average expression levels. In the following results, the terms upregulated and downregulated are consistent with the definitions used in the LIMMA analysis. That is, a gene is said to be upregulated if expressed higher in the disease (obese) group than the healthy (non-obese) group and said to be downregulated if expressed lower in the disease (obese) group than the healthy (non-obese) group. Tables of the top 20 upregulated (Table 3.3) and downregulated (Table 3.4) significant genes as determined by the rank product method are displayed below.

3.3. INTERSECTION OF LIMMA AND RANK PRODUCT RESULTS

A natural next step is to determine if there are any significant genes shared between the Illumina LIMMA results and the rank product results. The Affymetrix results are not as meaningful since at most 3 genes were identified as significant in the LIMMA analysis. The Venn diagrams showing the intersection of significant genes (at 0.05 and 0.10 FDR level) between the Illumina LIMMA and the rank product results are given in Figures 3.3 and 3.4 for upregulated genes and Figures 3.5 and 3.6 for downregulated genes. Note the results of LIMMA analysis from the Illumina Dataset are

also separated by upregulated or downregulated genes in Figures 3.3 through 3.6. Additionally, the top 20 upregulated and downregulated genes in this intersection are given in Tables 3.5 and 3.6, where the genes are sorted by their rank product estimated percentage of false prediction (PFP) which is an estimate of the FDR. The adj. p-value is the p-value from LIMMA after Benjamini-Hochberg multiple testing correction. These genes may be a priority for further investigation.

In general, there are a similar number of upregulated and downregulated genes found in both analyses. The rank product method identified more unique differentially expressed genes than the Illumina LIMMA analysis. At the 0.05 FDR level, there were 151 differentially expressed upregulated genes found in both methods and 125 downregulated genes. At the 0.10 FDR level, there were 261 upregulated differentially expressed genes found by both methods and 207 downregulated genes.

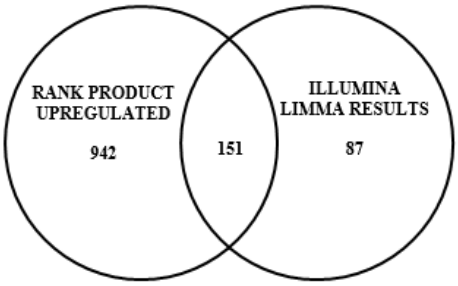

Figure 3.3 Venn Diagram of Significant Upregulated Genes found by Rank Product and Illumina LIMMA Analysis at FDR ≤ 0.05.

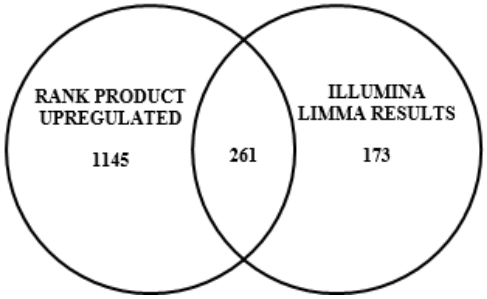

Figure 3.4 Venn Diagram of Significant Upregulated Genes found by Rank Product and Illumina LIMMA Analysis at FDR ≤ 0.10.

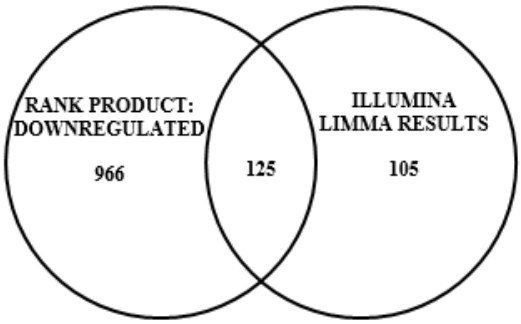

Figure 3.5 Venn Diagram of Significant Downregulated Genes found by Rank Product and Illumina LIMMA Analysis at FDR ≤ 0.05.

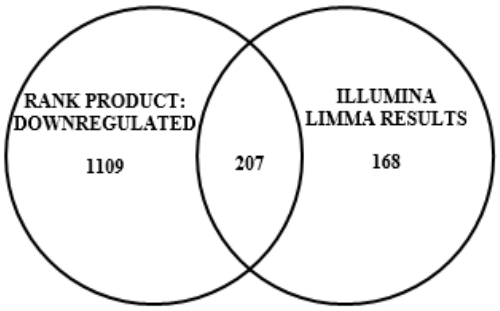

Figure 3.6 Venn Diagram of Significant Downregulated Genes found by Rank Product and Illumina LIMMA Analysis at FDR ≤ 0.10.

Table 3.3 Top 20 Upregulated Genes from Rank Product Analysis

Symbol	RP Statistic	Rank	PFP	P-Value	AveFC
CCND2	164.6232	1	1.32E-21	8.39E-26	1.163465
GSTT1	245.1409	2	4.53E-18	5.74E-22	1.226781
LAMC2	257.0491	3	8.38E-18	1.60E-21	0.266875
XRN2	297.7699	4	1.43E-16	3.64E-20	0.828796
GAS7	318.2059	5	4.60E-16	1.46E-19	0.65691
PARM1	337.0383	6	1.27E-15	4.82E-19	0.46402
FMOD	410.7209	7	6.09E-14	2.70E-17	0.615994
TMEM119	442.8592	8	2.40E-13	1.22E-16	0.599924
OLFML2A	478.0035	9	9.59E-13	5.48E-16	0.571239
SPON1	482.2179	10	1.03E-12	6.50E-16	0.899825
CCDC58	494.7691	11	1.54E-12	1.07E-15	0.59451
SH3GL3	515.4157	12	3.12E-12	2.37E-15	0.467107
SCRG1	534.1756	13	5.74E-12	4.73E-15	0.164335
SLC7A2	572.3062	14	1.99E-11	1.77E-14	0.456646
INHBE	634.361	15	1.29E-10	1.23E-13	0.716559
SULF2	636.9323	16	1.30E-10	1.32E-13	0.22193
ISLR	645.0158	17	1.55E-10	1.67E-13	0.390864
RUNX1T1	645.0747	18	1.47E-10	1.67E-13	0.678072
DDIT3	651.5379	19	1.67E-10	2.01E-13	0.640007
RPL14	668.1458	20	2.53E-10	3.21E-13	0.486816

Table 3.4 Top 20 Downregulated Genes from Rank Product Analysis

Symbol	RP Statistic	Rank	PFP	P-Value	AveFC
CYP1B1	110.3616	1	1.19E-25	7.55E-30	-0.8824
HAPLN1	117.0993	2	2.43E-25	3.09E-29	-2.22241
NR4A2	137.2349	3	6.72E-24	1.28E-27	-2.48611
ANKRD1	170.0921	4	6.94E-22	1.76E-25	-2.06558
FOXQ1	183.9045	5	3.23E-21	1.02E-24	-0.89715
XYLT1	191.706	6	6.81E-21	2.59E-24	-1.05992
MMP1	198.342	7	1.25E-20	5.53E-24	-1.42278
TNFRSF11B	271.5984	9	9.08E-18	5.18E-21	-1.18235
CXCL2	270.4295	8	9.31E-18	4.73E-21	-2.48239
TIPARP	312.3667	10	1.56E-16	9.92E-20	-0.96647
IL1A	316.1843	11	1.83E-16	1.28E-19	-1.16975
NTN4	317.7663	12	1.86E-16	1.42E-19	-1.07926
ATF3	328.4732	13	3.43E-16	2.83E-19	-1.3067
TMEFF2	336.7948	14	5.34E-16	4.75E-19	-0.66666
ANGPTL4	345.8524	15	8.62E-16	8.21E-19	-0.52269
HES1	351.2207	16	1.11E-15	1.13E-18	-1.88204
ZFP36	422.5749	17	4.43E-14	4.78E-17	-0.89493
GSTM1	448.0674	18	1.34E-13	1.53E-16	-0.62266
IL1B	459.861	19	2.12E-13	2.56E-16	-1.46796
CALB2	483.0702	20	5.31E-13	6.73E-16	-0.86066

Table 3.5 Top 20 Upregulated Genes from Rank Product Analysis Intersected with Illumina LIMMA Results. PFP = percentage of false predictions, logFC = log fold change, adj.P-value = Benjamini-Hochberg adjusted p-value.

Symbol	PFP	logFC	adj. P-Value
LAMC2	8.38E-18	1.3981	0.020061
XRN2	1.43E-16	1.44513	0.003184
OLFML2A	9.59E-13	0.89803	0.04916
SH3GL3	3.12E-12	0.99424	0.014791
DDIT3	1.67E-10	0.79942	0.003036
RPL14	2.53E-10	0.87949	0.013981
CYGB	3.77E-10	0.72141	0.001069
SYTL2	3.16E-09	0.8286	0.000488
LIFR	9.40E-09	0.75172	0.017624
NUDT7	1.46E-08	0.64625	0.004487
EFNB3	2.84E-08	0.75755	0.010106
SALL2	6.21E-08	0.72537	0.004559
CLDN23	7.88E-08	0.69327	0.036707
PNPLA7	8.31E-08	0.68979	0.002417
GEM	1.05E-07	0.75501	0.004956
PKDCC	1.79E-07	0.62881	0.038131
DHX58	2.10E-07	0.66968	0.001418
DDIT4	2.32E-07	0.56729	0.041999
HMGCS1	5.14E-07	0.67208	0.001656
CAND2	5.45E-07	0.69421	0.019045

Table 3.6 Top 20 Downregulated Genes from Rank Product Analysis Intersected with Illumina LIMMA Results. PFP = percentage of false predictions, logFC = log fold change, adj.P-value = Benjamini-Hochberg adjusted p-value.

Symbol	PFP	logFC	adj. P-Value
CYP1B1	1.19E-25	2.085074	1.03E-05
HAPLN1	2.43E-25	1.583964	0.002359
NR4A2	6.72E-24	1.30919	5.41E-06
ANKRD1	6.94E-22	1.37031	0.022189
FOXQ1	3.23E-21	1.663924	0.02109
XYLT1	6.81E-21	1.948587	0.014774
MMP1	1.25E-20	1.668332	0.037908
CXCL2	9.31E-18	0.98554	0.006014
TIPARP	1.56E-16	1.100755	2.41E-08
IL1A	1.83E-16	1.132063	0.001418
NTN4	1.86E-16	1.161478	0.014803
ATF3	3.43E-16	1.028282	0.000485
ANGPTL4	8.62E-16	1.406806	0.010789
HES1	1.11E-15	0.897127	0.000554
ZFP36	4.43E-14	0.977747	3.84E-05
ALDH1A3	8.09E-12	0.954061	0.001966
DYNC1I1	1.96E-11	0.932067	0.007891
GREM1	8.86E-11	0.955859	0.003036
GBP3	9.67E-11	0.794511	0.001656
SPP1	1.17E-10	0.838993	0.025464

4. DISCUSSION

4.1. CONCLUSIONS

It is not surprising that the Affymetrix Dataset provided very few significant genes considering there were only 3 patients in each treatment group with no additional technical replicates. The Illumina Dataset provided 7 patients in each treatment group with 3 technical replicates for each patient, yielding hundreds of statistically significant genes. The rank product results provided the largest number of significant genes, which is in good agreement with the concept of conglomerating information across multiple studies to increase the statistical power and heighten the ability to detect differentially expressed genes. Furthermore, of interest are the significant genes identified by both the parametric LIMMA procedure as well as the non-parametric rank product procedure. Certainly, there is greater trust in the genes designated significant by these two types of statistical analyses.

4.2. LIMITATIONS

The Illumina Dataset contained a plethora of information, namely, 7 patients in each treatment group with 3 technical replicates for each patient. On the contrary, the Affymetrix Dataset provided only 3 patients in each treatment group with no additional technical replicates. It would be beneficial to identify more microarray experiments with obese/non-obese patient MSC datasets in order to increase the power of the statistical methods. Furthermore, it would increase statistical power to include experiments that are not microarray experiments since the rank product method allows for heterogeneity.

It should be noted that there are several differences between the two datasets used in this work. The Affymetrix Dataset had a stricter definition of obesity (BMI > 40 $\frac{kg}{m^2}$) than the Illumina Dataset (BMI ≥ 30 $\frac{kg}{m^2}$). The study population also differs between the two studies. The Illumina Dataset consisted of only pregnant women who had a Caesarean section birth whereas the Affymetrix Dataset included both men and women. Thus, it may be difficult to make conclusions about a broader population without further investigations. Both datasets are observational studies and thus there is the potential for confounding variables that are driving the differences between the obese and non-obese groups. However, both studies lack available data on characteristics of the subjects (e.g., demographics) that would allow investigating differences between the obese and non-obese groups on potential confounding variables. Finally, by only utilizing the genes that are in common between the two microarrays, many genes are filtered out. It would be beneficial to explore a method that would enable the inclusion of more genes.

4.3. FUTURE WORK

With a significant gene list identified, the next step is to use gene pathway analysis to identify significant pathways of interest. This brings a broader biological significance to the results beyond individual genes of interest. Several programs allow for pathway analysis including gProfiler, Cytoscape, and Enrichment Map. Furthermore, the rank product method is readily available to conglomerate information from heterogenous datasets giving potential for combining data from different types of technologies to give unique insights. For instance, RNA sequencing or DNA methylation experiments could be incorporated into the meta-analysis to increase the power to detect robust biomarkers.

BIBLIOGRAPHY

Anal fistula. Anal Fistula | Johns Hopkins Medicine. (2022, April 19). Retrieved April 26, 2023, from https://www.hopkinsmedicine.org/health/conditions-and-diseases/anal-fistula

Annotationdbi: Introduction to Bioconductor Annotation Packages. (n.d.). Retrieved March 5, 2023, from https://www.bioconductor.org/packages/devel/bioc/vignettes/AnnotationDbi/inst/doc/IntroToAnnotationPackages.pdf

Badraiq, H., Cvoro, A., Galleu, A., Simon, M., Miere, C., Hobbs, C., Schulz, R., Siow, R., Dazzi, F., & Ilic, D. (2017). Effects of maternal obesity on Wharton's Jelly mesenchymal stromal cells. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-18034-1

Baygan, A., Aronsson-Kurttila, W., Moretti, G., Tibert, B., Dahllöf, G., Klingspor, L., Gustafsson, B., Khoein, B., Moll, G., Hausmann, C., Svahn, B. M., Westgren, M., Remberger, M., Sadeghi, B., & Ringden, O. (2017). Safety and side effects of using placenta-derived decidual stromal cells for graft-versus-host disease and hemorrhagic cystitis. *Frontiers in Immunology*, *8*(JUL). https://doi.org/10.3389/fimmu.2017.00795

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bianco, P., Robey, P. G., & Simmons, P. J. (2008). Mesenchymal Stem Cells: Revisiting History, Concepts, and Assays. In *Cell Stem Cell* (Vol. 2, Issue 4, pp. 313–319). Elsevier Inc. https://doi.org/10.1016/j.stem.2008.03.002

Bioconductor Rankprod Package Vignette. (n.d.). Retrieved March 5, 2023, from https://www.bioconductor.org/packages/devel/bioc/vignettes/RankProd/inst/doc/RankProd.pdf

Bolstad, B. M., Irizarry, R. A., Astrand, M. ˚, & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. In *BIOINFORMATICS* (Vol. 19, Issue 2). http://www.bioconductor.org.

Brazma, A., Hingamp, P., Quackenbush, J. *et al.* Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. (2001). *Nat Genet* **29**, 365–371. https://doi.org/10.1038/ng1201-365

Breitling, R., Armengaud, P., Amtmann, A., & Herzyk, P. (2004). Rank products: A simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters*, *573*(1–3), 83–92. https://doi.org/10.1016/j.febslet.2004.07.055

Chappell, J., & Lucks, J. Transcribing with the STARS. (n.d.). Retrieved April 22, 2023, from https://benchling.com/pub/star.

Clough, E., & Barrett, T. (2016). The Gene Expression Omnibus Database. *Methods in molecular biology (Clifton, N.J.), 1418*, 93–110. https://doi.org/10.1007/978-1-4939-3578-9_5

Contrasts.fit: Compute Contrasts From Linear Model Fit. RDocumentation. (n.d.). Retrieved March 5, 2023, from https://www.rdocumentation.org/packages/limma/versions/3.28.14/topics/contrasts.fit

Del Carratore, F., Jankevics, A., Eisinga, R., Heskes, T., Hong, F., & Breitling, R. (2017). Rankprod 2.0: A refactored bioconductor package for detecting differentially expressed features in molecular profiling datasets. *Bioinformatics*, *33*(17), 2774–2775. https://doi.org/10.1093/bioinformatics/btx292

Eisinga, R., Breitling, R., & Heskes, T. (2013). The exact probability distribution of the rank product statistics for replicated experiments. *FEBS Letters*, *587*(6), 677–682. https://doi.org/10.1016/j.febslet.2013.01.037

Gao, G., Fan, C., Li, W., Liang, R., Wei, C., Chen, X., Yang, Y., Zhong, Y., Shao, Y., Kong, Y., Li, Z., & Zhu, X. (2021). Mesenchymal stem cells: ideal seeds for treating diseases. In *Human Cell* (Vol. 34, Issue 6, pp. 1585–1600). Springer Japan. https://doi.org/10.1007/s13577-021-00578-0

Han, Y., Yang, J., Fang, J., Zhou, Y., Candi, E., Wang, J., Hua, D., Shao, C., & Shi, Y. (2022). The secretion profile of mesenchymal stem cells and potential applications in treating human diseases. In *Signal Transduction and Targeted Therapy* (Vol. 7, Issue 1). Springer Nature. https://doi.org/10.1038/s41392-022-00932-0

Hong, F., & Breitling R. (2008). A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*, *24*(3), 374–382. https://doi.org/10.1093/bioinformatics/btm620

Hong, F., Del Carratore, F., Wittner, B., Breitling, R., & Janckevics, A. (2023, April 25). *Bioconductor Rankprod Package Vignette*. Bioconductor RankProd Package Vignette. https://www.bioconductor.org/packages/devel/bioc/vignettes/RankProd/inst/doc/RankProd.pdf

Heskes, T., Eisinga, R., & Breitling, R. (2014). A fast algorithm for determining bounds and accurate approximate -values of the rank product statistic for replicate experiments. *BMC Bioinformatics*, *15*(1). https://doi.org/10.1186/s12859-014-0367-1

Illumina microarray technology. Illumina. (n.d.). Retrieved April 22, 2023, from https://emea.illumina.com/science/technology/microarray.html

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. (2003). *Biostatistics*. *4*(2):249-64. doi: 10.1093/biostatistics/4.2.249. PMID: 12925520.

Jaksik, R., Iwanaszko, M., Rzeszowska-Wolny, J., & Kimmel, M. (2015). Microarray experiments and factors which affect their reliability. *Biology Direct*, *10*(1). https://doi.org/10.1186/s13062-015-0077-2

Khan Academy. (n.d.). *Stages of Transcription: Initiation, Elongation & Termination (article)*. Khan Academy. Retrieved April 22, 2023, from https://www.khanacademy.org/science/biology/gene-expression-central-dogma/transcription-of-dna-into-rna/a/stages-of-transcription

Koziol, J. A. (2010). Comments on the rank product method for analyzing replicated experiments. *FEBS Letters*, *584*(5), 941–944. https://doi.org/10.1016/j.febslet.2010.01.031

Kuhn, K., Baker, S. C., Chudin, E., Lieu, M. H., Oeser, S., Bennett, H., Rigault, P., Barker, D., McDaniel, T. K., & Chee, M. S. (2004). A novel, high-performance random array platform for quantitative gene expression profiling. *Genome Research*, *14*(11), 2347–2356. https://doi.org/10.1101/gr.2739104

Law, C. W., Alhamdoosh, M., Su, S., Smyth, G. K., & Ritchie, M. E. (2016). RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. *F1000Research*, *5*, 1408. https://doi.org/10.12688/f1000research.9005.1

Luo, Y. (2007, June). Comparison between Affymetrix and Illumina gene expression microarray platforms. Retrieved from http://hdl.handle.net/11375/21286

Oñate, B., Vilahur, G., Camino-López, S., Díez-Caballero, A., Ballesta-López, C., Ybarra, J., Moscatiello, F., Herrero, J., & Badimon, L. (2013). Stem cells isolated from adipose tissue of obese patients show changes in their transcriptomic profile that indicate loss in stemcellness and increased commitment to an adipocyte-like phenotype. *BMC Genomics*, *14*(1). https://doi.org/10.1186/1471-2164-14-625

OpenStax College, Concepts of Biology. (n.d.). Retrieved April 22, 2023, from https://philschatz.com/biology-concepts-book/contents/m45476.html.

Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A., & Ploner, A. (2005). False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*, *21*(13), 3017–3024. https://doi.org/10.1093/bioinformatics/bti448

Pestel, J., Blangero, F., & Eljaafari, A. (2023, January 17). Pathogenic role of adipose tissue-derived mesenchymal stem cells in obesity and obesity-related inflammatory diseases. *Cells*. https://ncbi.nlm.nih.gov/pmc/articles/PMC9913687/

Pistoia, V., & Raffaghello, L. (2017). Mesenchymal stromal cells and autoimmunity. In *International Immunology* (Vol. 29, Issue 2, pp. 49–58). Oxford University Press. https://doi.org/10.1093/intimm/dxx008

Pittenger, M. F., Discher, D. E., Péault, B. M., Phinney, D. G., Hare, J. M., & Caplan, A. I. (2019). Mesenchymal stem cell perspective: cell biology to clinical progress. In *npj Regenerative Medicine* (Vol. 4, Issue 1). Nature Research. https://doi.org/10.1038/s41536-019-0083-6

R: Fit Normal+Exp Convolution Model to Observed Intensities. (n.d.). Retrieved February 25, 2023, from http://web.mit.edu/~r/current/arch/i386_linux26/lib/R/library/limma/html/normexpfit.html

R: Normexp Background Correction and Normalization Using Control Probes. (n.d.). Retrieved February 25, 2023, from http://web.mit.edu/~r/current/arch/i386_linux26/lib/R/library/limma/html/nec.html

RMA: Robust Multi-Array Average Expression Measure. RDocumentation. (n.d.). Retrieved February 25, 2023, from https://www.rdocumentation.org/packages/affy/versions/1.50.0/topics/rma

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

Ritchie, M. E., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A., & Smyth, G. K. (2007). A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, *23*(20), 2700–2707. https://doi.org/10.1093/bioinformatics/btm412

Shi, W., Oshlack, A., & Smyth, G. K. (2010). Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips. *Nucleic Acids Research*, *38*(22). https://doi.org/10.1093/nar/gkq871

Silver, J. D., Ritchie, M. E., & Smyth, G. K. (2008). Microarray background correction: Maximum likelihood estimation for the normal-exponential convolution. *Biostatistics*, *10*(2), 352–363. https://doi.org/10.1093/biostatistics/kxn042

Smyth, G. K., Michaud, J., & Scott, H. S. (2005). Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, *21*(9), 2067–2075. https://doi.org/10.1093/bioinformatics/bti270

Smyth, Gordon K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. (2004). *Statistical applications in genetics and molecular biology* 3.1.

Sookdeo, A. (2022). *The Central Dogma: Gene Expression*. https://academicworks.cuny.edu

Toro-Domínguez, D., Villatoro-Garciá, J. A., Martorell-Marugán, J., Román-Montoya, Y., Alarcón-Riquelme, M. E., & Carmona-Saéz, P. (2021). A survey of gene expression meta-analysis: Methods and applications. In *Briefings in Bioinformatics* (Vol. 22, Issue 2, pp. 1694–1705). Oxford University Press. https://doi.org/10.1093/bib/bbaa019

U.S. National Library of Medicine. (n.d.). *Geo summary - geo - NCBI*. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/geo/summary/?type=tax

U.S. National Library of Medicine. (n.d.). *MLA CE Course Manual: Molecular Biology Information Resources (Genetics Review)*. National Center for Biotechnology Information. Retrieved April 22, 2023, from https://www.ncbi.nlm.nih.gov/Class/MLACourse/Original8Hour/Genetics/gene_expression.html

Xie, Y., Wang, X., & Story, M. (2009). Statistical methods of background correction for Illumina BeadArray data. *Bioinformatics*, *25*(6), 751–757. https://doi.org/10.1093/bioinformatics/btp040

VITA

Dakota William Shields attended elementary, middle, and high school in Maryville, Missouri. In Spring 2020, Dakota graduated from Northwest Missouri State University with a degree of Bachelor of Science in Nanoscale Science with an emphasis in Physics and minors in Business and Calculus. Dakota contributed to several published works during their time developing models of endofullerenes with Dr. Chakraborty. After graduating, they furthered their computational skill by joining the Master's program at Missouri S&T in Applied Mathematics with an emphasis in Statistics in Fall 2020. At S&T, Dakota enjoyed contributing to the Fly Sleep Modeling research group with Dr. Olbricht, Dr. Thimgan, Dr. Sam, and Dr. Wu over several semesters developing new models of predicting lifespans. Dakota also enjoyed collaborating with Dr. Semon and Hailey Swain analyzing stem cell experiments. They received their M.S. degree in Applied Mathematics with an emphasis in Statistics in July 2023.