Multi-Omics Analysis of Adipogenesis

Vidal Marcel Arroyo
University College
University of Oxford

A thesis submitted for the degree of
Master of Statistics by Research

Trinity 2021

For my Ita, who showed me how to fight disease with resilience, tenacity, and grace.

Acknowledgements

It takes a village to raise a lot of things, including a master's thesis. In the case of ***my*** master's thesis, it took even more than a village - an entity that I will now refer to as a super-village. It will not be possible to thank everyone in this super-village with a finite number of pages, so two will have to suffice. Hence, I want to apologize up front to those I have forgotten - it is perhaps the case that our friendship is so natural that I cannot parse it out as an external factor that has partially sustained this work.

I first want to start with those who are closest to this thesis: the Lindgren Group. When I first came into the Lindgren Group, I was soggy, frustrated, and in need of a 20-minute nap - primarily due to the underwhelming Oxford rain on my bike ride up to the Big Data Institute. Beyond these acute qualities, I was also at my scientific low, having spent 4 months in Oxford with no work towards my thesis completed. It was Cecilia, my supervisor, who took in this muppet of a man and turned him into an independent, rigorous, and fearless scientist. I have nothing but gratitude for her love, encouragement, and support throughout the course of my master's. I also want to give special thanks to Chris, Helen, Samvida, and Saskia, who have given a tremendous amount of feedback throughout the course of this project - especially during the writing phase. Additionally, I want to give particular thanks to Richard, who will never know how much I looked forward to our Thursday lunches at the Wellcome Centre. Penultimately, I want to thank our collaborators in the Claussnitzer Lab who have been an absolute joy to work with. Finally, I want to thank the Lindgren Group at large for creating a wonderful atmosphere to do science, albeit virtually. I consider you all cherished teachers, colleagues, and friends.

Now, I want to thank those who lie between my academic and family life - a category of people who I will now refer to as 'friends'. These include those who I have met through The Rhodes Trust, The OCCA, Calvary Chapel Oxford, and elsewhere. I have made too many new friendships in Oxford to mention them all, so here I will just highlight a few that fall in the 'elsewhere' category. This thesis was made across three different homes, so I want to thank those who I have had the pleasure of sharing my home(s) with: for Kristina, who showed me how to appreciate the simple things in life (namely a slice of toast with unsalted butter); for Leah, who inspired me to channel my inner cyanobacteria and spend more time enjoying the sun; for Shruti, who was always willing to share her Labneh with me; for Jaz, who imbued an all-consuming love for cheese within the very crevices of my soul; for Max,

who gave me his skateboard; and for Bella, for just being herself. I also want to thank Blake, who became a big brother to me. If best friends were defined by the number of hours spent with someone outside of unavoidable contexts, then Blake was my best friend here in Oxford. Thank you for being there when I needed it most, dude.

As far as physical proximity goes, I have never been farther from my family then while in Oxford. But, as far as emotional proximity goes, I have never been closer. I can go on and on (and on and on) about how much my family means to me, but again here I will keep it brief. For this work, I want to give particular thanks to the members of my immediate family who I was blessed to talk with weekly (and, during a good week, daily): for my pops, who showed me how to lead a life of character; for my momma, who never failed to infect me with her endless positivity; for Lucci, who sent me a fantastic YouTube video on how to properly trim a beard; for Joaquo, who always sent the perfect memes; and for Paloma, for chasing her dreams. *Te amo.*

Finally, I want to thank the patients for whom this work was done for and by whom this work was done by. This year was full of many unexpected surprises for myself, including the fact that I ended up turning down a career as a physician because I knew that caring directly for patients was not in my calling. That being said, as a humanitarian-scientist, I will always continue to strive towards the highest bar in academic medicine: to make a real difference for a real patient in real life. Thank you in particular for the patients whose data I had the pleasure of analyzing in this work. Albeit I have never and probably will never meet you face-to-face, I will never take for granted that every row in my data frame and every point on my plot represents the story of another human who has suffered in some way from disease. Thank you for giving me the opportunity to have a small role in turning those sorrows into insights. Hopefully, one day someone else can turn these insights into joys.

SOLI DEO GLORIA

Contents

Abstract

Understanding adipocyte development, also known as adipogenesis, is critical for improving metabolic health. Additionally, adipose tissue depot, sex, age, body-mass index, Type 2 diabetes status, and cell size are known to play important roles in adipogenesis. Here, two classes of unsupervised learning methods were used to investigate which of these phenotypes were most strongly associated with differences in omics data between human adipocytes. Data consisted of two omics data types, including RNA-Sequencing (RNA-Seq) and Adipocyte Profiling (AP), as well as nine patient phenotypes: day, depot, sex, age, body-mass index, Type 2 diabetes status, mean cell area, cell area variance, and % small cells. Data was clustered with a systematic pipeline using either single-view learning or multi-view learning methods, where single-view learning was performed using Principal Component Analysis (PCA) and multi-view learning was performed using either Multi-Omics Factor Analysis (MOFA+) or Multiple Canonical Correlation Analysis (MCCA). Chi-square permutation tests were used to test for association between K-Means clusters and phenotype labels and an extra layer of FDR-Correction was performed to account for multiple testing across all clustering analyses. Day (RNA-Seq PCA P-Value = 0.0092; AP PCA P-Value = 0.4224; MOFA+ P-Value = 0.0078; MCCA P-Value = 0.0092) and depot (RNA-Seq PCA P-Value = 0.0352; AP PCA P-Value = 0.0736; MOFA+ P-Value = 0.0450; MCCA P-Value = 0.0092) were identified as the primary drivers of clustering. However, none of the other phenotypes drove clustering, potentially due to unbalanced clusters or reduced statistical power from stratification. This multi-omics analysis of adipogenesis demonstrates the power of unsupervised clustering methods for the generation of robust biological insights that are consistent with the current literature surrounding day and depot. Nonetheless, since neither single-view nor multi-view learning methods detected associations for any of the other phenotypes, the inclusion of more samples and more omics is preferred for future studies.

Abbreviations

AP = Adipocyte Profiling
BMI = Body-Mass Index
D# = Day #
FDR = False-Discovery Rate
MCCA = Multiple Canonical Correlation Analysis
MOBB = Munich Obesity BioBank
MOFA+ = Multi-Omics Factor Analysis
mRNA = Messenger RNA
NA = Not Available
PAC = Pre-Adipocyte
PC = Principal Component
PCA = Principal Components Analysis
RNA = Ribonucleic Acid
RNA-Seq = RNA-Sequencing
s = Seconds
scWAT = Subcutaneous White Adipose Tissue
T2D = Type 2 Diabetes
UMAP = Uniform Manifold Approximation and Projection
vcWAT = Visceral White Adipose Tissue

Chapter 1

Introduction

1.1 Background

The application of big data to biology and medicine has greatly benefited from the rise of omics, which are high-dimensional biological data types that are typically generated using high-throughput experimental biology techniques. The generation and analysis of omics data has helped tease apart complex biology, like in the case of a 2016 study that used two types of omics to identify a genetic program underlying metastatic lung cancer cells (Denny et al. 2016). The power of omics is due to its ability to capture information systematically - thus enabling the generation of plausible functional hypotheses without prior knowledge. For instance, genomic technologies like genotyping capture genome-wide human genetic variation. More recently, functional genomic technologies like RNA-Sequencing (RNA-Seq), which captures global gene expression (Wang et al. 2009), and ATAC-Seq, which captures chromatin accessibility (Minnoye et al. 2021), have enabled the investigation of gene functions and interactions across the genome. Some omic technologies have also broadened beyond the genome to the cell as the system of study. An example of such a method is Cell Painting, which uses multiplexed dyes to profile cellular morphological features (Bray et al. 2016). Ultimately, integrating both genomic and cellular phenotyping data in multi-omics analyses may lead to new insights for biological processes like human metabolism.

Human metabolism is a complex physiological process that involves the interplay between multiple cell, tissue, and organ types. In particular, organs such as the liver, brain, gut, and skeletal muscle serve as both providers and recipients of system wide physiological signals (Priest and Tontonoz 2019). A critical center piece to this system-wide communication is adipose tissue (Figure 1.1), and there is a substantial body of work that suggests that dysregulation of adipocyte development, also known

as adipogenesis, plays a functional role in development of metabolic diseases like Type 2 diabetes (T2D; Ghaben and Scherer 2019). Multiple studies have also demonstrated that obesity, which is defined by having a body-mass index (BMI) above 30.0, affects adipogenesis (Nadler et al. 2000; Dubois et al. 2006). Abnormal adipocyte function caused by obesity can also lead to downstream insulin resistance and potentially T2D (Guilherme et al. 2008). Given that nearly 650 million people today are obese (*Obesity and overweight* n.d.), it is paramount that research continues to develop a deeper understanding of adipogenesis and its role in human metabolism.

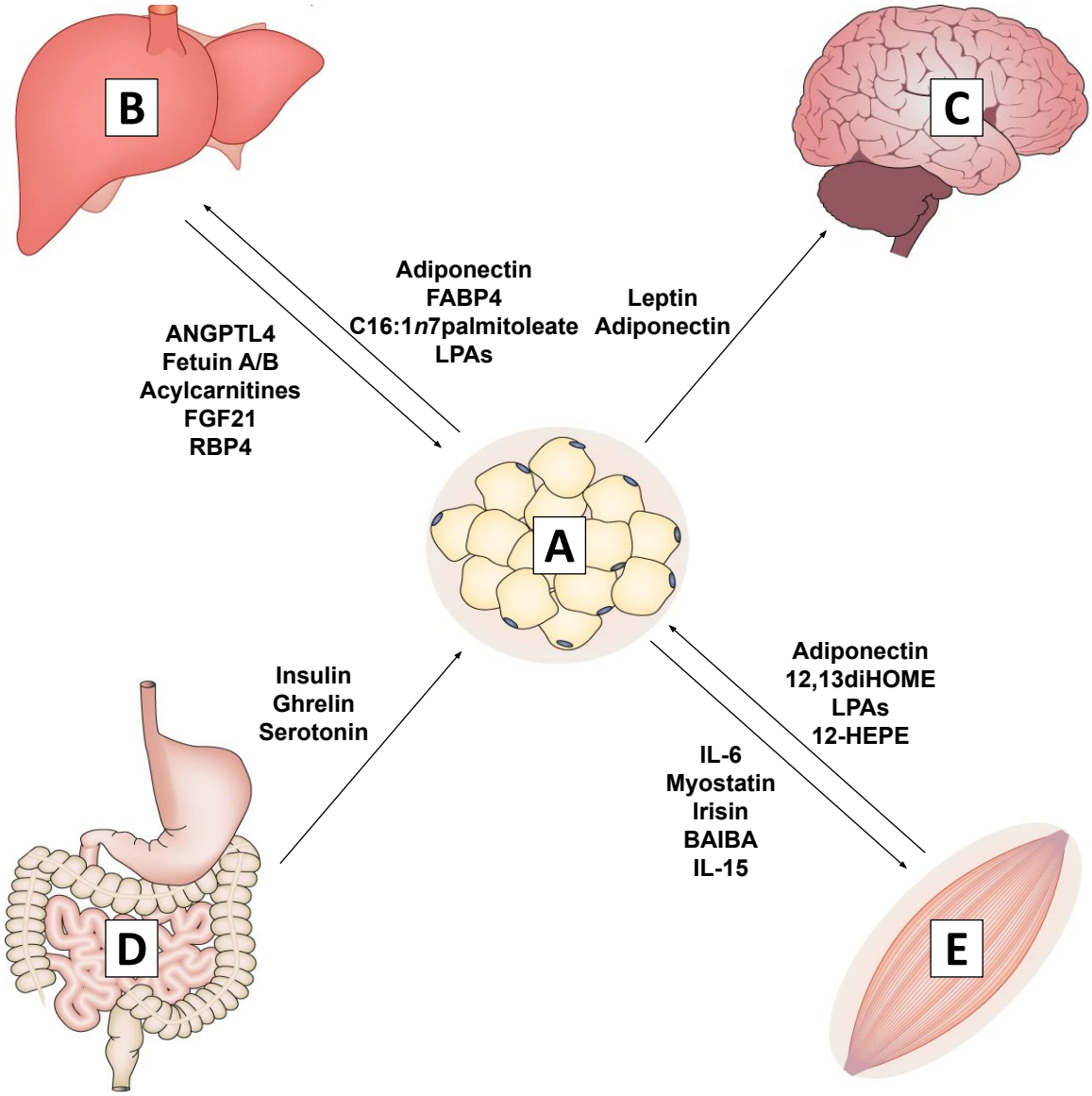

Figure 1.1: Inter-organ communication between (A) adipose tissue, (B) the liver, (C) the brain, (D) the gut, and (E) skeletal muscle during the maintenance of metabolic homeostasis. Adapted with permission from Priest and Tontonoz 2019.

1.2 Adipose Tissue and Adipogenesis

Adipose tissue in the human body plays a critical role in energy balance, hormonal regulation, and inflammation (Chait and Hartigh 2020). Multiple cell types make up adipose tissue, including stem cells, preadipocytes, adipocytes, endothelial cells, and immune cells (Chait and Hartigh 2020). Adipose tissue is also patterned across various depots throughout the body (Bjørndal et al. 2011). For instance, white adipose tissue is either located in subcutaneous depots, which are just under the skin, or visceral depots, which surround internal organs (Mittal 2019). Brown adipose tissue, on the other hand, is located along the vasculature and around the neck (Sacks and Symonds 2013).

Different types of adipose tissue harbor different types of adipocytes. White adipose tissue is primarily composed of white adipocytes, which are used for fat storage (Inagaki et al. 2016). Brown adipose tissue is made up of brown adipocytes, which are known to be involved in thermogenesis (Inagaki et al. 2016). There is also a third type of adipocyte known as beige or 'brite' (brown-white) adipocytes, which primarily reside in subcutaneous white adipose tissue and are plastic enough to switch between white and brown cellular phenotypes (Ikeda et al. 2018).

Adipocytes, just like any cell type, are derived from a common progenitor stem cell. In the case of white adipocytes, the development from mesenchymal stem cell to mature adipocyte occurs through a process known as adipogenesis (Figure 1.2; Tang and Lane 2012). Adipogenesis is primarily regulated by a transcription factor known as PPAR-G, which regulates the expression of transcriptional programs that lead to adipocyte development (Lefterova et al. 2014). As a result of this process, lipids accumulate in adipocytes and lead to the formation of lipid droplets, which increase cell size (Yu and Li 2016). Hence, increased cell size and/or the formation of lipid droplets form morphological evidence that adipogenesis has taken place.

Several studies have characterized differences in adipogenesis between white adipocytes as a consequence of underlying biological variation. For instance, one study demonstrated that adipocytes have different gene expression patterns across adipogenesis depending on whether they come from subcutaneous or visceral adipose tissue depots (Perrini et al. 2007). Differences in gene expression between these two depots have also been demonstrated outside of the context of adipogenesis (Atzmon et al. 2002; Linder et al. 2004). Additionally, sex-specific differences have been observed for adipogenesis genes (Anderson et al. 2020). Like in the case for depot, differences in gene expression within static adipose tissue have also been observed between the

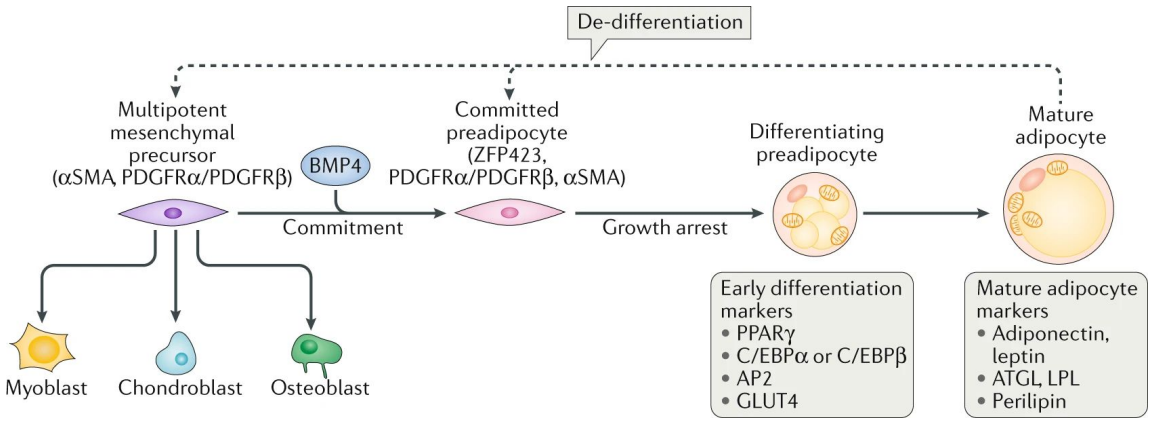

Figure 1.2: Overview of mammalian adipogenesis. Adapted with permission from Ghaben and Scherer 2019.

sexes (White and Tchoukalova 2014). Beyond the influence of depot and sex, aging is known to have a prominent effect on fat mass and distribution, which can affect adipocyte development (Kirkland et al. 2002, Cartwright et al. 2007, Mancuso and Bouchard 2019). Based on previous work, it is apparent that day, depot, sex, age, BMI, T2D status, and cell size are all associated with differences across adipogenesis. However, little work has been done to determine which of these phenotypes are most strongly associated with variation in adipogenesis when compared to one another.

1.3 Overview of Thesis

Multi-view clustering methods have yet to be used to study adipogenesis. So, this work will use both unsupervised single-view and multi-view clustering methods with the aim to investigate whether phenotypes including day, depot, sex, age, BMI, T2D status, and histology-derived cell sizes are associated with large-scale transcriptional and/or morphological differences in multi-omics data. This is split into two sub-aims: (1) To compare and contrast the utility of single-view and multi view learning in the study of adipogenesis (statistical) and (2) To identify which phenotypes drive biological diversity among human adipocytes using both single-view and multi-view learning (biological). Through these analyses, this work aims to identify which phenotypes are most strongly associated with variation between differentiating adipocytes.

Chapter 2

Materials and Methods

2.1 Overview of Data

With the goal of understanding how transcriptional and/or morphological differences are related to adipogenic phenotypes, data for this study were generated by the Claussnitzer and Lindgren Labs at the Broad Institute, USA, from patient preadipocytes (PACs) from 30 individuals in the Munich Obesity BioBank (MOBB), which was described in prior work (Glastonbury et al. 2020). PACs were taken from patient subcutaneous and/or visceral adipose tissue samples and then placed in cell culture, where cells were cultured over 14 days using an adipocyte-specific differentiation medium in order to promote differentiation into mature adipocytes. Over the course of adipogenesis, cells were collected at 3 (Day 0, Day 3, and Day 14) or 4 (Day 0, Day 3, Day 8, and Day 14) time points to represent different stages of adipogenesis (Niemelä 2008). Functional multi-omics data was produced using these collected cells, generating a large, longitudinal, multi-omics dataset of human adipogenesis in vitro (Figure 2.1). There were 2 types of omics data: (1) genome-wide transcriptional data generated using RNA-Sequencing (RNA Seq), and 2) cell-wide morphological data generated using Adipocyte Profiling. In addition, for each patient and at each stage of adipogenesis, phenotype data was available on day, depot of origin, sex, age, BMI, T2D status, and histology derived information on cellular size for 18/30 individuals. Even though there were only 30 individuals in this study, nearly 180 paired omics samples were generated across time on a per-patient and per-depot basis, making this one of the largest multi-omics studies of its kind.

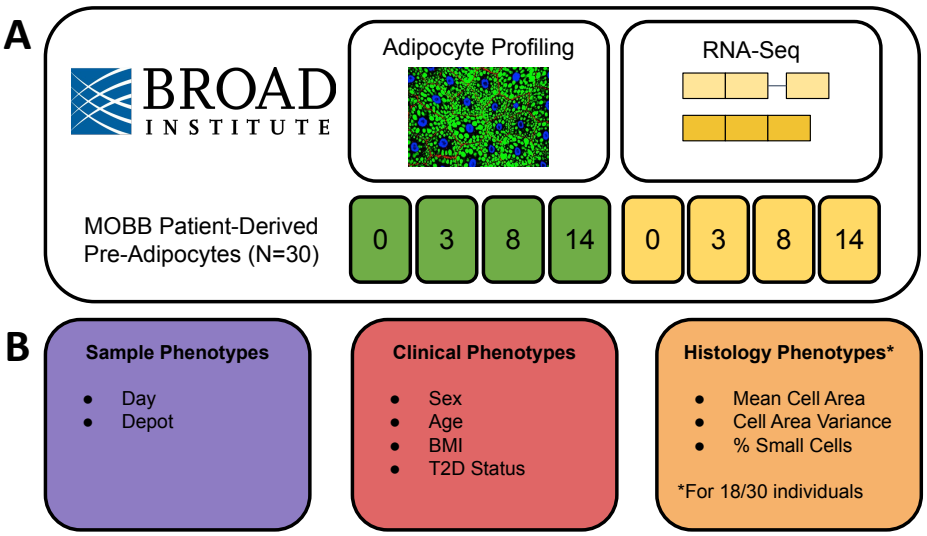

Figure 2.1: Overview of data. (A) Multi-omics data were generated across four time points (Day 0, Day 3, Day 8, and Day 14) for 30 Patient-Derived Pre-Adipocytes from the Munich Obesity BioBank (MOBB) cohort. (B) Phenotype data for this study, which were split into (1) cell line-specific sample phenotypes (day, depot), (2) patient-specific clinical phenotypes (sex, age, BMI, and T2D status), and (3) histology-based cell size phenotypes that were generated in prior work (Glastonbury et al. 2020). Broad Institute logo adapted with permission. Adipocyte profiling image adapted with permission from Zhang et al. 2017.

2.2 Phenotype Data

In order to study which phenotypes were most strongly associated with variation among differentiating adipocytes, nine phenotypes were included in these analyses: two sample phenotypes that were specific to cell line samples (day and depot), four patient phenotypes that were specific to individuals (sex, age, BMI, and T2D status), and three histology-based cell size phenotypes (mean cell area, cell area variance, and % small cells) (Figure 2.1B). Patient-derived PACs were collected from two depots: (1) subcutaneous white adipose tissue (scWAT) or (2) visceral white adipose tissue (vcWAT). The PACs from these separate depots were analyzed across time, where day was represented as four broad timepoint categories: Day 0, Day 2/3, Day 6/7/8, and Day 13/14/24. These time points were grouped according to four broad stages of adipocytes during adipogenesis: determined pre-adipocyte, committed pre-adipocyte, early-stage mature adipocyte, and late-stage mature adipocyte (Niemelä 2008).

In order to facilitate effective downstream clustering, all phenotype variables had to be represented categorically. Thus, continuous clinical phenotypes (BMI and age) were transformed using inverse-rank normal transformation. Following inverse-rank

normal transformation, variables were split into two categorical groups at 0. The same procedure was also applied to the three histology-based phenotypes (mean cell area, cell area variance, and % small cells), which were generated by a colleague at Oxford using a deep-learning model to automatically extract these cell size features (Glastonbury et al. 2020). This process resulted in two categorical groups which were stratified at the median of each continuous variable, which was 46 years for age, 51 for BMI, 3300 μm^2 for mean cell area, 2900 μm^2 for cell area variance, and 48% for percentage small cells (Figures 2.2 - 2.3). Overall, the categorical transformation of these continuous variables made them viable for downstream clustering analyses but at the possible expense of separating them at a median value that is not representative of the population from which other studies have generated insights.

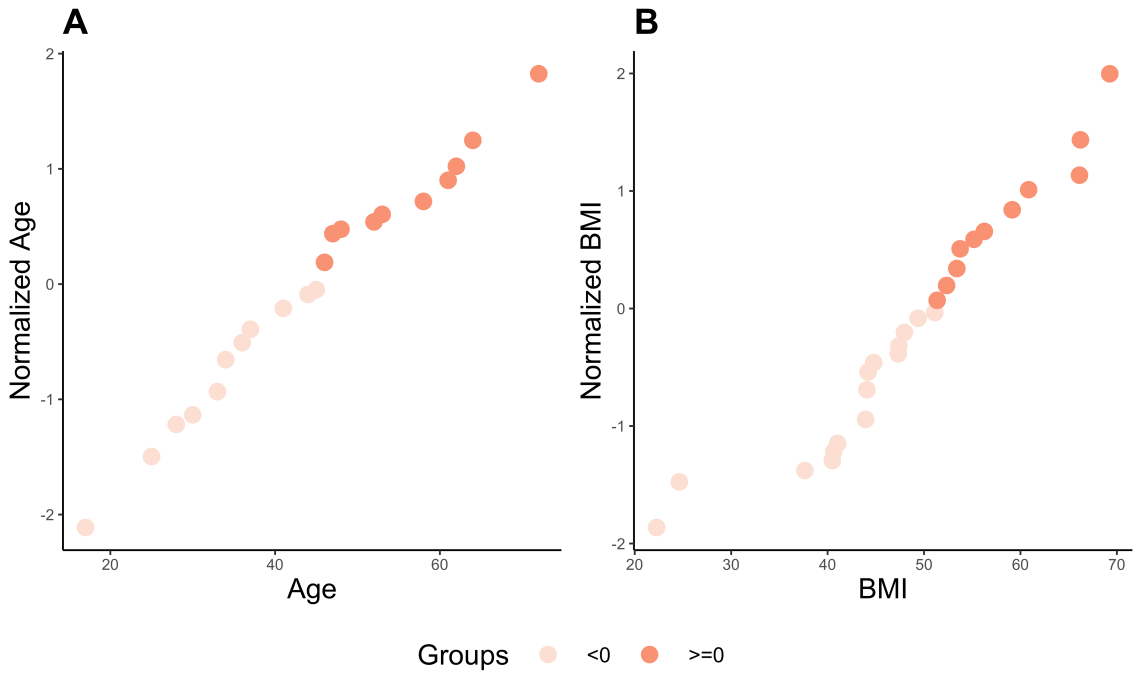

Figure 2.2: Continuous to categorical transformations of continuous clinical phenotypes. Transformation was performed using inverse rank normalization followed by splitting the transformed data at the median of the normalized distribution. (A) After transformation, age was split into two groups: less than 46 years of age and greater than or equal to 46 years of age. (B) After transformation, BMI was split into two groups: less than 51 and greater than or equal to 51.

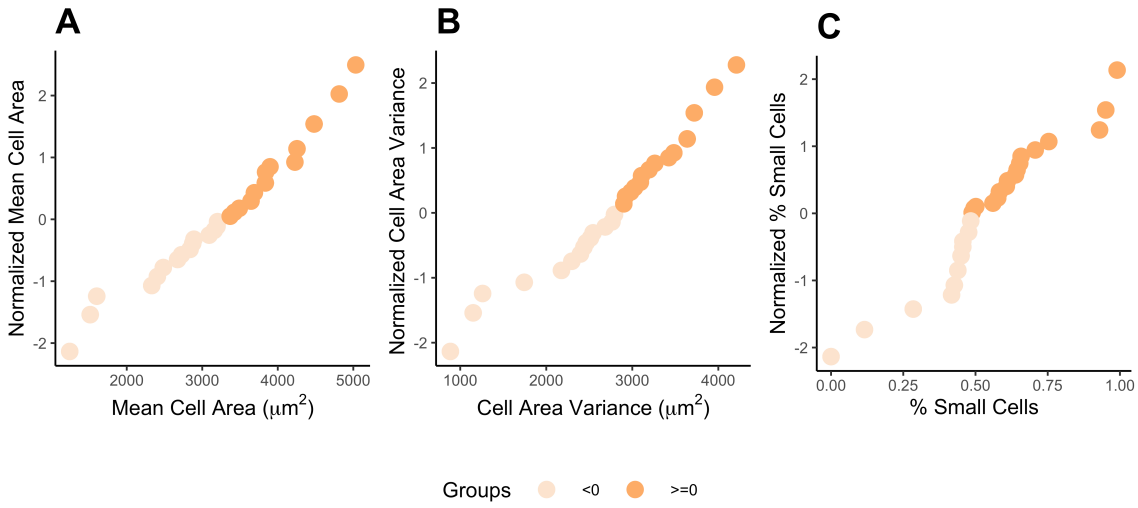

Figure 2.3: Continuous to categorical transformations of histology phenotypes. Transformation was performed using inverse rank normalization followed by splitting the transformed data at the median of the normalized distribution. (A) After transformation, mean cell area was split into two groups: less than 3300 µm^2 and greater than or equal to 3300 µm^2. (B) After transformation, cell area variance was split into two groups: less than 2900 µm^2 and greater than or equal to 2900 µm^2. (C) After transformation, % small cells was split into two groups: less than 48% and greater than or equal to 48%.

2.3 Cellular Phenotyping Data

2.3.1 Data Generation

It is not clear which adipogenic phenotypes are most strongly associated with morphological differences in adipocytes. So, high-throughput cellular phenotyping data was generated for this study with a method derivative of Cell Painter, which is a cellular morphology profiling assay based on six fluorescent dyes which are multiplexed and imaged in five channels (Bray et al. 2016). Through this methodology, Cell Painter measures quantitative phenotypes on eight broadly relevant cellular components or organelles. In this study, high-dimensional morphological features were created using Adipocyte Profiling, where BODIPY had been added to the other stains in the Cell Painter protocol to stain for lipid droplets in adipocytes - thus making it adipocyte-specific. After staining adipocytes in a multiplexed fashion, morphological features were automatically extracted using analytical software housed in CellProfiler 3.1.9 (Jones et al. 2008). 25 fields per well were imaged and 3005 features were extracted per field. Extracted features were classified along different axes: based

on cellular compartment (nuclei, cell, or cytoplasm), based on stain (mitochondria, BODIPY, DNA, or actin), or based on feature (granularity, texture, or intensity). Before pre-processing and quality control, flat field correction was performed using average per-plate intensity. Although Adipocyte Profiling is a new method that has not yet been used extensively, it generated morphological data that was both broad and specific to adipocytes - making it the preferred cellular phenotyping method for this study.

2.3.2 Pre-Processing and Quality Control

In order to prepare Adipocyte Profiling data for downstream clustering analyses, pre-processing and quality control of the Adipocyte Profiling data were performed by the Claussnitzer Lab (Figure 2.4). Any fields with a cell count less than 50 were removed in order to ensure that features were quantified from fields with enough cells. Fields that were corrupted by experimentally induced technical artifacts were also removed by using a manually-defined quality control mask. Data was normalized on a per-plate and per-patient basis using a robust scaling approach that subtracts the median from each feature and divides it by the inter-quartile range (Pedregosa et al. 2011). Individual wells were further aggregated for downstream analysis by depot and day. Before analysis, samples with missing features were removed. Additionally, nuclear block-listed features that were known to be noisy and generally unreliable were also removed (Way 2020). Additionally, features that were identical across all samples were removed. Inverse rank normalization was performed per morphological feature in order to ensure comparability between features during clustering. Despite the fact that Adipocyte Profiling data is a new data type with no previously established pipeline, the pipeline described in this work was robust and consistent with the outside literature when possible, like in the case of data normalization (Pedregosa et al. 2011) and handling of block-listed features (Way 2020).

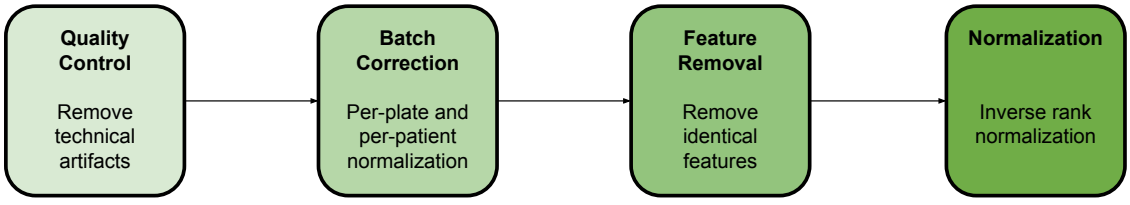

Figure 2.4: Pre-processing and quality control pipeline for cellular phenotyping data, which consisted of quality control, batch correction, feature removal, and normalization.

2.4 mRNA Expression Data

2.4.1 Data Generation

Even though gene expression during adipogenesis has been extensively studied (Ambele et al. 2016; Ehrlund et al. 2016; Côté et al. 2017; Ramirez et al. 2020), it has never been tied together with cellular phenotyping data in an association analysis with phenotypes that are known to be relevant for adipogenesis. Thus, mRNA expression data was generated for this study from each sample across either 3 (N=18 individuals) or 4 (N=12 individuals) stages of adipocyte differentiation via paired-end Illumina bulk RNA-Sequencing (RNA-Seq). Briefly, total RNA was extracted from cells collected at each time point using a spin column kit. The integrity and purity of total RNA were assessed using an Agilent 4200 Bioanalyzer (Agilent Technologies) before proceeding to preparation of RNA-Seq libraries. The Illumina TruSeq Stranded mRNA protocol was used for the preparation of RNA-seq libraries and sequenced on a NovaSeq6000 machine (Illumina). The use of bulk RNA-Sequencing in this study trades the higher resolution of single-cell methods for more robust signals that can be leveraged during clustering analyses.

2.4.2 Pre-Processing and Quality Control

RNA-seq data were prepared for downstream clustering analyses with the following pipeline (Figure 2.5). Sequencing quality control was performed using FastQC 0.11.9 (Andrews et al. 2019) and adaptor trimming of paired-end Illumina adaptors was done with TrimGalore! 0.6.2 (Krueger et al. 2019). Sequence quality was assessed with phred score plots, where all reads exceeded a phred score of 28 which is indicative of high quality (Figure 2.6). Sequences were aligned to the ENSEMBL hg38 genome using STAR 2.7.1 (Dobin et al. 2012), where no more than 1 base-pair mismatch was allowed during alignment. All samples had at least 15M aligned reads (Figure 2.7). RNAseq counts tables were generated by RSEM 1.3.2 (Li and Dewey 2011), where each read was counted and assigned to genes by specifying the hg38 reference genome and paired-end sequencing. 15M reads for 9/12 samples were uniquely assigned to genes (Figure 2.8). Transcript counts were normalised using variance-stabilising transformations implemented in DESeq2 (Anders and Huber 2010; Love et al. 2014). This transformation was effective in reducing homoscedasticity i.e. the variance-mean dependence (Figure 2.9). All quality control plots were generated with

MultiQC (Ewels et al. 2016). This pipeline did not quantify mRNA counts at an isoform level, but it did do so at a per-gene basis using robust tools that have been used for thousands of other studies.

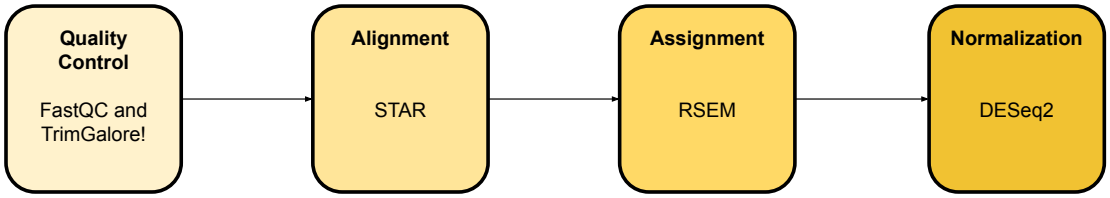

Figure 2.5: Pre-processing and quality control pipeline for mRNA expression data, which consisted of quality control, alignment, assignment, and normalization.

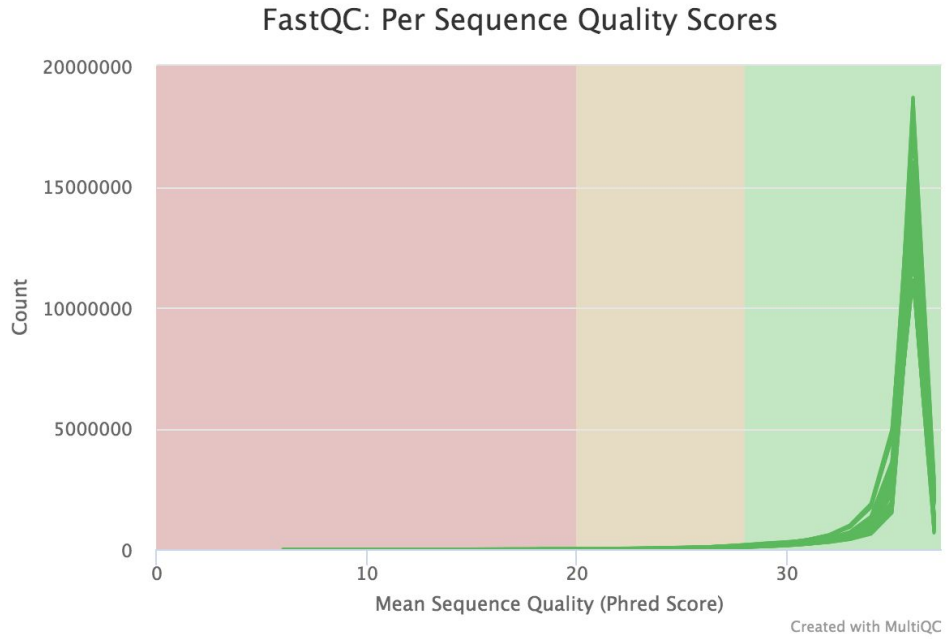

Figure 2.6: FastQC per sequence quality scores for for 12 samples, where all samples had a mean sequence quality score (Phred Score) above the cut-off for high quality sequences (28). Made using MultiQC (Ewels et al. 2016).

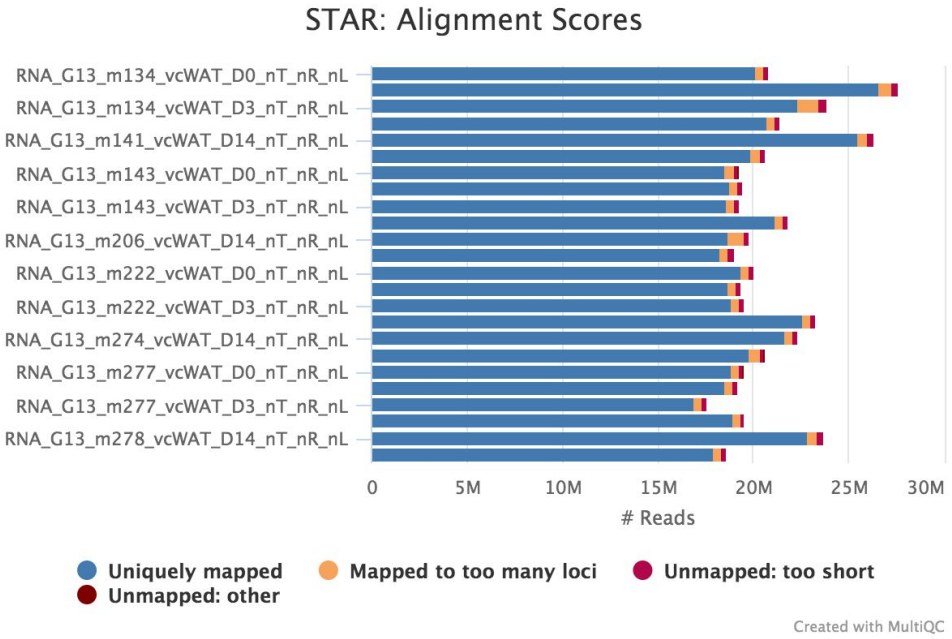

Figure 2.7: STAR alignment scores for 12 samples, where all samples had at least 20M aligned reads. Made using MultiQC (Ewels et al. 2016).

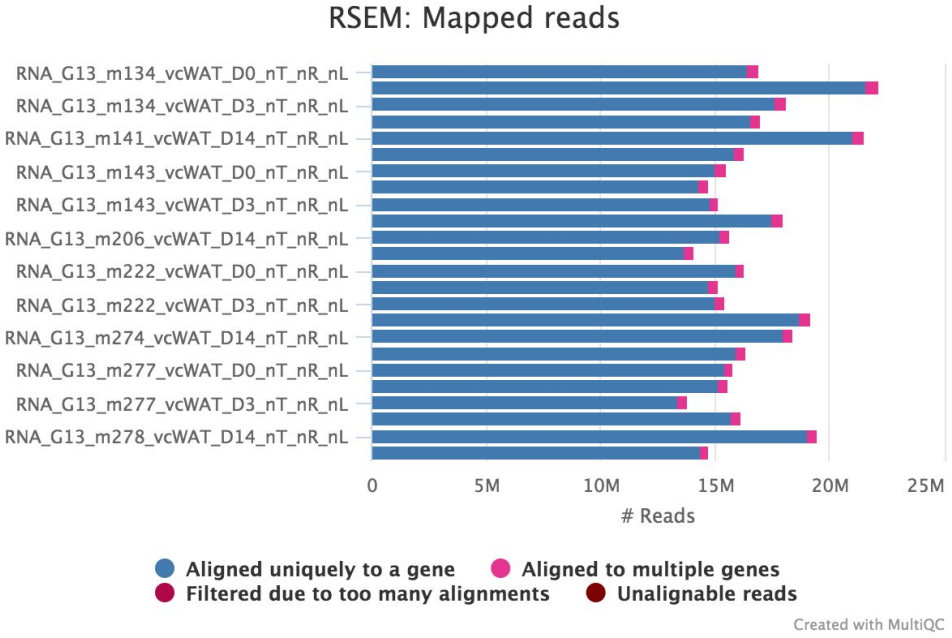

Figure 2.8: RSEM mapped reads for 12 samples, where 9/12 samples had at least 15M assigned reads. Made using MultiQC (Ewels et al. 2016).

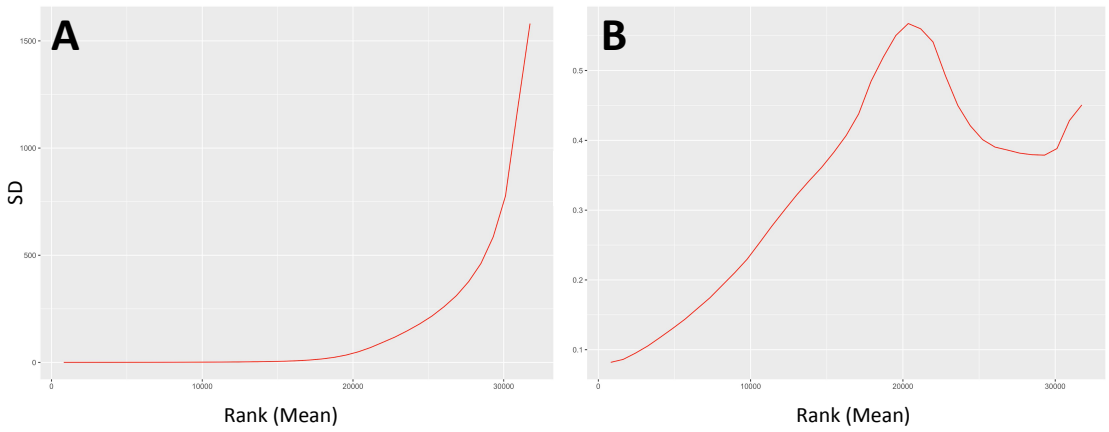

Figure 2.9: Variance-mean dependence of transcriptomic count data (A) before and (B) after variance stabilizing transformation using DESeq2 (Anders and Huber 2010; Love et al. 2014).

2.5 Selection of Clustering Methods

2.5.1 Single-View Learning using PCA

In order to study adipogenesis using multi-omics data, both single-view and multi view learning methods were used for clustering analysis to compare and contrast the utility of integrating multiple layers of data into views at once. For single-view learning, principal component analysis (PCA) was selected, which is a dimensionality-reduction technique that is commonly used for exploratory analysis of large-scale biological data (Lever et al. 2017). From a mathematical perspective, PCA is a change of basis method used to identify major sources of variability in a single view. In the first step of the PCA algorithm, eigenvalues are computed that maximize the variation explained in a given view. These eigenvalues are linear weights that are multiplied against features to produce a principal component for each sample (Lever et al. 2017). In this study, the term feature is used to describe a variable shared across observations, which can be either a gene for RNA-Seq data or a morphological feature for Adipocyte Profiling data. For each successive step of PCA, another set of eigenvalues are computed that are orthogonal to the first - thus giving independence between the set of computed principal components. In this work, PCA was run on each view separately after all pre-processing, quality control, and normalization had been performed. The use of PCA in this work allowed the discovery of independent, uncorrelated components for each view at the expense of only providing information for one view at a time.

2.5.2 Systematic Selection of Multi-View Learning Methods

Since several multi-view learning methods could have been used for this study, a systematic pipeline was developed to select methods for downstream clustering analyses (Figure 2.10). With this pipeline, the suitability of 9 multi-view learning methods with well-implemented R packages as reviewed in Nguyen and Wang 2020 were evaluated for their application for multi-view clustering. The first criteria ruled out any methods that had a run-time above 2000 seconds in a previous benchmark, which eliminated one method (Rappoport and Shamir 2018). Methods were also only kept if they could be used for sample classification or clustering, which eliminated two methods. Finally, methods that could be used for the analysis of more than two omics were retained since these would be methods that would be more broadly applicable to the field. After applying all of these criteria, only two methods remained: Multi Omics Factor Anaysis (MOFA+; Argelaguet et al. 2020) and Multiple Canonical

Correlation Analysis (MCCA; Witten and Tibshirani 2009). Many selection pipelines have their limitations, but the one used in this work did enable the selection of two methods that have been used for several studies and happen to integrate multi-omics data with different assumptions.

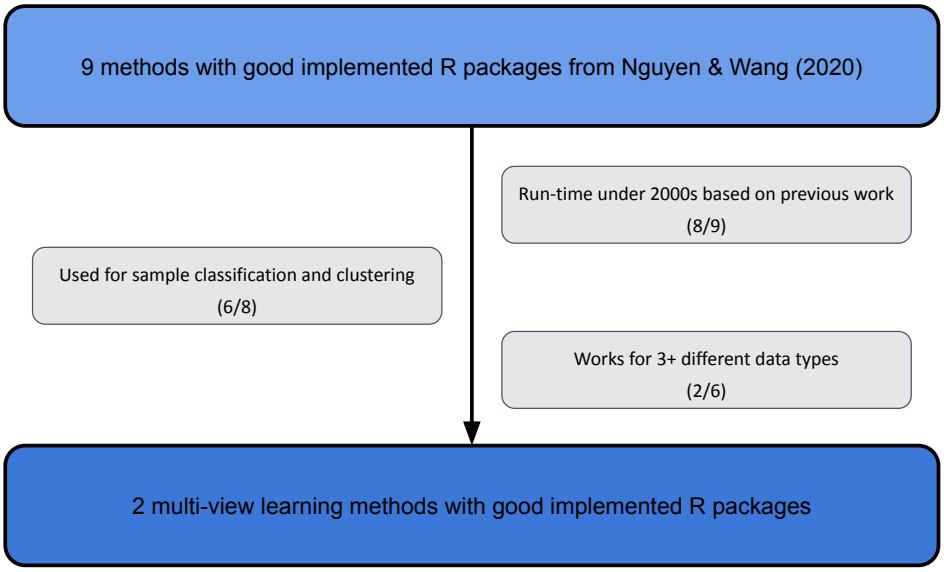

Figure 2.10: Systematic selection pipeline for multi-view learning methods. Methods from Nguyen and Wang 2020 were chosen on the basis of (1) run time less than 2000 seconds (s) from the analysis in Rappoport and Shamir 2018, (2) ability to work with more than 2 data types for broader applicability, and (3) applicability for clustering analysis.

2.5.3 Multi-View Learning using MOFA+

One of the multi-view learning methods that was selected for downstream clustering analyses was MOFA+, which is a Bayesian machine-learning method used to identify major sources of variability called factors that can be shared across and between views (Figure 2.11). MOFA+ can be thought of as a multi-dimensional PCA because it seeks to identify major sources of variability across two or more omics. MOFA+ identifies these sources of variability by using a non-negative matrix factorization framework to learn a shared set of latent factors between multiple views (Argelaguet et al. 2018). MOFA+ was run with default parameters recommended by the authors as implemented in the package. The contribution of each view to each factor was visualized across all samples in order to explore the extent to which MOFA+ explained

variation jointly. The primary strength of MOFA+ is that it enables the identification of independent factors that can be separately or jointly explained by multiple omics. However, by enabling separate explanation of variation, MOFA+ does risk reducing to single-view insights if there is no relationship between the omics.

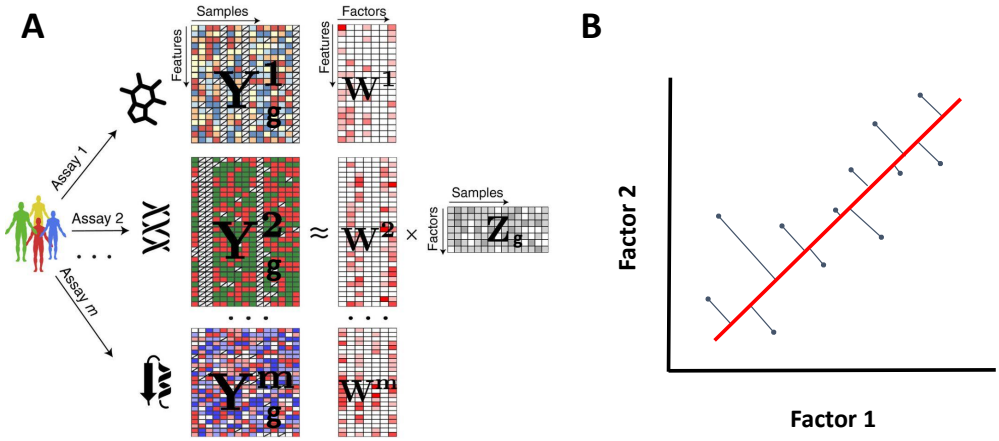

Figure 2.11: Overview of MOFA+. (A) MOFA+ learns a set of shared latent factors shared between multiple omics. Adapted with permission from Argelaguet et al. 2018. (B) Simplistic representation of MOFA+, where factors are orthogonal of one another.

2.5.4 Multi-View Learning using MCCA

The second method selected for downstream multi-view clustering analyses was MCCA, which is a correlation-based method used to identify linear combinations of features per view called variates that maximize correlation between views (Figure 2.12). At each step of the MCCA algorithm, a set of variates is computed where the length of the variate set is equal to the number of data layers processed by MCCA. MCCA generalizes canonical correlation analysis to more than two types of data by maximizing correlation between all views simultaneously instead of doing so pair-wise (Witten and Tibshirani 2009). MCCA was implemented with parameters recommended by the authors. In addition, since MCCA cannot run if features are identical across samples, features with variance 0 were removed. In contrast to MOFA+, MCCA assumes a direct relationship between all omics used in the framework - thus enabling equal

contribution of each omic to the variates. However, this may force a relationship between omics that is not actually substantiated by the data.

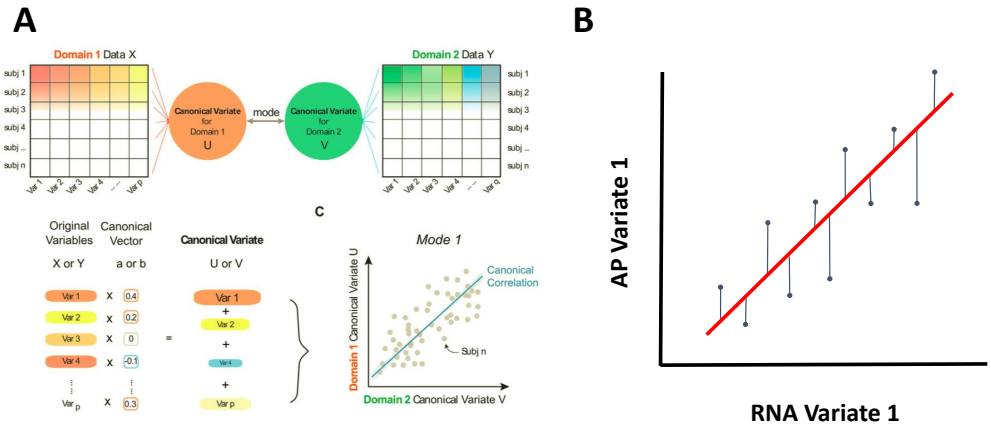

Figure 2.12: Overview of MCCA. (A) Canonical correlation methods compute canonical variates that maximize the correlation between two different types of data. Adapted with permission from Wang et al. 2020. (B) Simplistic representation of MCCA, where one variate for RNA-Seq (RNA) is regressed against its paired variates for Adipocyte Profiling (AP).

2.6 Clustering Pipeline

A clustering pipeline was created in order to systematically identify which phenotypes drive clustering among omics data (Figure 2.13). Features, which are variables shared across observations (i.e. a gene in RNA-Seq), were first reduced to a smaller number of factors, which are combinations of features. The top factors (25 if there were at least 25 samples, 15 otherwise) were then set aside. Each of the different methods called these 'factors' by a different name: in PCA, they were principal components; in MOFA+, they were factors; in MCCA, they were variate pairs - one for RNA-Seq and one for Adiopcyte Profiling. Since none of these methods assign cluster labels to samples, K-Means clustering was then used to cluster samples by these top factors, where K was chosen to match each of the phenotypes that it would later be associated with (Lloyd 1982). To test for association, FDR-Corrected permutation tests were

performed for each phenotype. These tests resulted in a set of P-Values - one for each phenotype. If there was at least one significant hit ($P < 0.05$), the omics data was stratified by the most significant hit among the phenotypes - which was defined by the highest test statistic. These steps were repeated iteratively until there were no more significant hits. After, FDR-correction was performed once again to account for multiple testing across all strata. After, the corrected P-Values were assessed in order to which phenotypic associations were significant after correction, which would then be visualized across the top factors using UMAP. For instance, if the data had been stratified by day and there was a significant association with sex at D0, then sex would be visualized across all day strata (D0, D3, D8, and D14). This entire pipeline was run to completion for (1) RNA-Seq data using PCA, (2) Adipoctye Profiling data using PCA, (3) both omics using MOFA+, and (4) both omics using MCCA. This enabled the systematic testing and visualization of which phenotypes were most strongly associated with large-scale variation in RNA-Seq and/or Adipocyte Profiling data across all samples. The use of unsupervised learning methods here is a particular strength because phenotypic associations are able to be tested without any prior hypotheses, but multiple rounds of stratification may result in a significant loss of power and a smaller number of associations being detected.

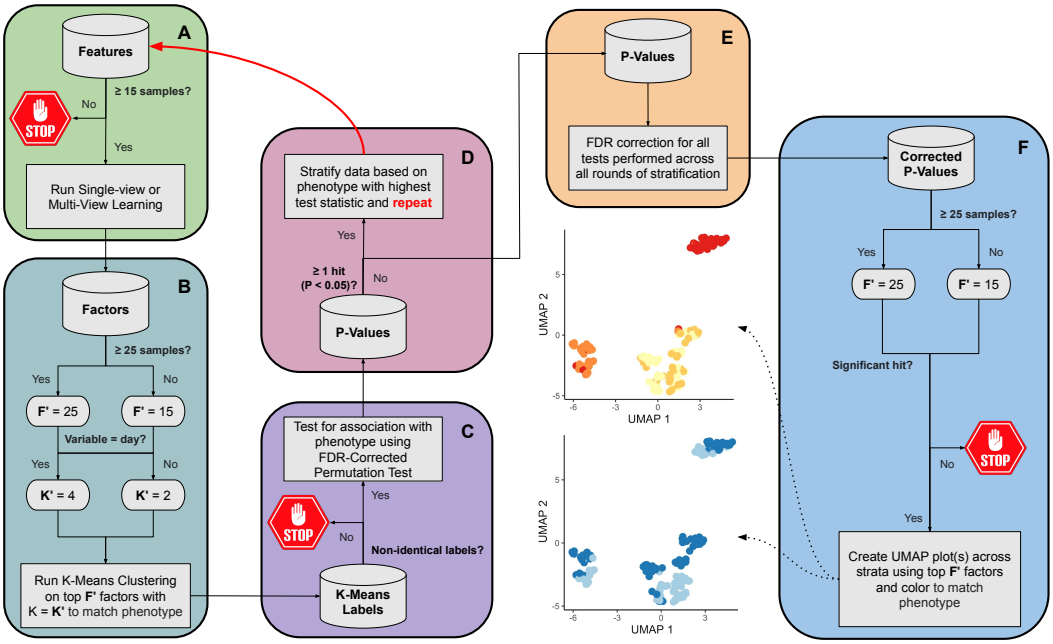

Figure 2.13: Clustering pipeline used to systematically identify phenotypic drivers of clustering. (A) Features for one (singe-view) or both (multi-view) omics are used for single-view or multi-view learning if at least 15 samples are present. (B) Factors produced by single-view or multi-view learning are used for K-Means clustering for each phenotype. If at least 25 samples are present, then the top 25 factors are used. If not, the top 15 factors are used. If the phenotype variable is day, then the number of categories for K-means clustering is 4; otherwise, it is 2. (C) K-Means clustering assigns clustering labels for each sample, which are then tested for association with the associated phenotype if samples were assigned to different clusters. 1000 permutations are performed for all FDR-corrected permutation tests. (D) Permutation tests across all phenotypes produce a set of P-values, where a significant hit is defined at $P < 0.05$. (E) FDR-correction is performed over all the p-values generated during stratified clustering. (F) Corrected P-Values are evaluated in order to determine which results to visualize per phenotype. If there is a significant hit, then it is visualized across all strata at a given round of stratification.

Chapter 3

Results

3.1 Descriptive Statistics of Phenotype Data

Before clustering analyses were performed, descriptive statistics were produced in order to check if there were any unbalanced categorical variables before clustering analyses. To do this, counts and proportions of each phenotype were enumerated across all samples in the data set, and then for all variables except day a two-sided binomial test at P=0.5 was performed to test how well balanced each phenotype was (Table 3.1). Day was not tested because it had four categories and because it was expected to be unbalanced since 18 individuals did not have a D6/D7/D8 sample. Nonetheless, it is clear from assessing these statistics that D0, D2/D3, and D13/D14/D24 do not have the same number of samples, which would have been expected given the study design. This is because over 50 samples were removed from the Adipocyte Profiling data due to missing morphological features. None of the remaining phenotypes had a significant difference in cluster sizes except for sex, where the ratio of female to male samples was about 5 to 2 (127 vs. 53, P <0.0001).

Phenotype Category	Count	Percentage	P-Value
Day			
D0	42	23%	
D2/D3	43	24%	NA
D6/D7/D8	36	20%	
D13/D14/D24	59	33%	
Depot			
PAC scWAT	81	45%	0.2050
PAC vcWAT	99	55%	
Sex			
Male	53	29%	3.432e-08
Female	127	71%	
Age			
<46	88	49%	0.8231
≥ 46	92	51%	
BMI			
<51	89	49%	0.9406
≥ 51	91	51%	
T2D Status			
Yes	90	50%	1.0000
No	90	50%	
Mean Cell Area*			
$<3300\,\mu m^2$	50	51%	1.0000
$\geq 3300\,\mu m^2$	49	49%	
Cell Area Variance*			
$<2900\,\mu m^2$	52	53%	0.6879
$\geq 2900\,\mu m^2$	47	47%	
% Small Cells*			
$<48\%$	48	48%	0.8408
$\geq 48\%$	51	52%	

Table 3.1: Descriptive statistics of phenotype data. P-Values computed using two-sided binomial test, where P=0.5. Test not performed for day due to more than 2 categories. * = data available for 18/30 individuals.

3.2 Single-View Clustering of Adipocyte Profiling Data

In order to determine how large-scale variation in Adipocyte Profiling data was associated with phenotypic differences, the clustering pipeline was run for Adipocyte Profiling principal components (Figure 2.13). Both day and depot were both iden-

tified as significant drivers of clustering before further analyses were performed (P <0.05; Table 3.2). Accordingly, since day had the highest chi-square statistic among these two significant hits (23.7449 vs. 12.4346), it was used to stratify samples into four groups for further clustering analyses as outlined in the clustering pipeline (Figure 2.13). However, after FDR-correcting across all tests performed across all rounds, neither day nor depot remained significant so UMAP plots were not visualized (Figure 2.13).

Phenotype	**Chi-Square Test Statistic**	**P-Value**	**FDR-Corrected P-Value**
Day	23.7449	0.0048	0.4224
Depot	12.4346	0.0008	0.0736
Sex	2.7124	0.1060	1.0000
Age	1.7621	0.2260	1.0000
BMI	0.0942	0.7640	1.0000
T2D Status	0.2006	0.7670	1.0000
Mean Cell Area	0.0121	0.9990	1.0000
Cell Area Variance	0.0216	0.9990	1.0000
% Small Cells	0.7765	0.4220	1.0000

Table 3.2: Adipocyte Profiling PCA clustering results (no stratification). FDR Corrected P-Values were produced after all Adipocyte Profiling PCA analyses were finished.

The Adipocyte Profiling data was stratified by day and then analyzed again using the clustering pipeline in order to determine if there were any day-specific clustering effects for any of the remaining phenotypes. Through these analyses, depot was found to be significantly associated with clusters among D2/D3 (FDR Corrected P-Value = 0.0182) and D13/D14/D24 strata (FDR-Corrected P-Value = 0.0092; Tables 3.4, 3.6 and Figure 3.1). Depot was not significantly associated with clustering differences among D0 (FDR-Corrected P-Value = 1.0000) nor D6/D7/D8 (FDR-Corrected P-Value = 1.0000) strata (Tables 3.3, 3.5 and Figure 3.1). None of the other phenotypes were significantly associated with clustering differences at this step or within depot-stratified D2/D3 or D13/D14/D24 groups after FDR-Correction (Supplementary Table 1).

Phenotype	**Chi-Square Test Statistic**	**P-Value**	**FDR-Corrected P-Value**
Depot	0.6841	0.5380	1.0000
Sex	1.2302	0.3020	1.0000
Age	0.4229	0.5520	1.0000
BMI	0.4229	0.5600	1.0000
T2D Status	0.9873	0.3540	1.0000
Mean Cell Area	NA	NA	NA
Cell Area Variance	NA	NA	NA
% Small Cells	NA	NA	NA

Table 3.3: Adipocyte Profiling PCA clustering results (stratified by day = D0). FDR Corrected P-Values were produced after all Adipocyte Profiling PCA analyses were finished. Histology-based phenotypes were not analyzed due to identical K-Means labels between samples.

Phenotype	**Chi-Square Test Statistic**	**P-Value**	**FDR-Corrected P-Value**
Depot	14.9300	0.0002	0.0182
Sex	3.1189	0.0918	1.0000
Age	0.4309	0.5360	1.0000
BMI	6.5081	0.0231	1.0000
T2D Status	2.2160	0.2050	1.0000
Mean Cell Area	0.3500	0.6630	1.0000
Cell Area Variance	3.1988	0.1620	1.0000
% Small Cells	0.3500	0.6730	1.0000

Table 3.4: Adipocyte Profiling PCA clustering results (stratified by day = D2/D3). FDR-Corrected P-Values were produced after all Adipocyte Profiling PCA analyses were finished.

Phenotype	**Chi-Square Test Statistic**	**P-Value**	**FDR-Corrected P-Value**
Depot	12.0000	0.0008	0.0720
Sex	1.4026	0.4380	1.0000
Age	2.4000	0.2470	1.0000
BMI	0.6000	0.7050	1.0000
T2D Status	0.6000	0.6980	1.0000
Mean Cell Area	1.5867	0.4790	1.0000
Cell Area Variance	2.5500	0.2070	1.0000
% Small Cells	1.2364	0.5120	1.0000

Table 3.5: Adipocyte Profiling PCA clustering results (stratified by day = D6/D7/D8). FDR-Corrected P-Values were produced after all Adipocyte Profiling PCA analyses were finished.

Phenotype	**Chi-Square Test Statistic**	**P-Value**	**FDR-Corrected P-Value**
Depot	20.6865	0.0001	0.0092
Sex	2.3127	0.1530	1.0000
Age	0.1582	0.7960	1.0000
BMI	2.0047	0.2000	1.0000
T2D Status	1.3613	0.3030	1.0000
Mean Cell Area	1.0588	0.4920	1.0000
Cell Area Variance	2.9412	0.1670	1.0000
% Small Cells	1.8889	0.3090	1.0000

Table 3.6: Adipocyte Profiling PCA clustering results (stratified by day = D13/D14/D24). FDR-Corrected P-Values were produced after all Adipocyte Pro filing PCA analyses were finished.

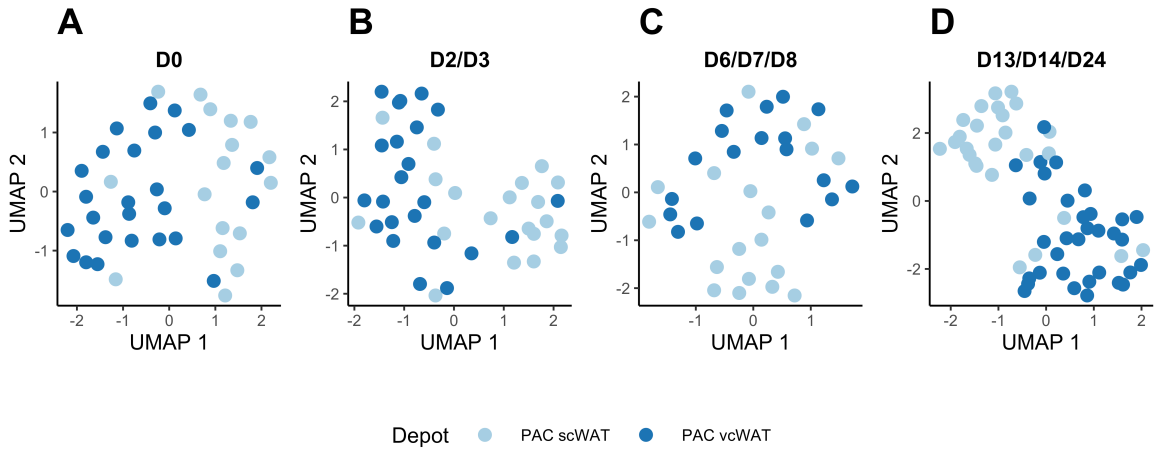

Figure 3.1: Visualization of (A) D0, (B) D2/D3, (C) D6/D7/D8, and (D) D13/D14/D24 Adipocyte Profiling PCA components using UMAP, colored by depot.

3.3 Single-View Clustering of RNA-Seq Data

The same clustering pipeline used to analyze Adipocyte Profiling principal components was also applied to RNA-Seq principal components in order to determine which phenotypes were most strongly associated with variation in mRNA expression. Both day (FDR-Corrected P-Value = 0.0092) and depot (FDR-Corrected P-Value = 0.0352) were significantly associated with clustering differences among RNA-Seq samples after FDR-Correction (Table 3.7 and Figure 3.2). Day was the most significant phenotype with a chi-square test statistic of 273.6797, so it was used to stratify the data for further clustering analysis. As consistent with the Adipocyte Profiling results, no other variables were significantly associated with clustering before further stratification.

Phenotype	**Chi-Square Test Statistic**	**P-Value**	**FDR-Corrected P-Value**
Day	273.6797	0.0001	0.0092
Depot	13.6403	0.0004	0.0352
Sex	0.0028	0.9990	1.0000
Age	1.3363	0.2710	1.0000
BMI	0.5382	0.5370	1.0000
T2D Status	0.3879	0.6430	1.0000
Mean Cell Area	0.0810	0.8310	1.0000
Cell Area Variance	2.3177	0.1440	1.0000
% Small Cells	0.1664	0.8330	1.0000

Table 3.7: RNA-Seq PCA clustering results (no stratification). FDR-Corrected P Values were produced after all RNA-Seq PCA analyses were finished.

After stratifying by day, the RNA-Seq data was once again analyzed using the clustering pipeline to determine if any phenotypes had day-specific clustering effects. Analogous to what was found when analyzing the Adipocyte Profiling data, significant clustering differences between depot groups were found within D2/D3 (FDR Corrected P-Value = 0.0092) and D13/D14/D24 (FDR-Corrected P-Value = 0.0267) strata (Tables 3.9, 3.11 and Figure 3.3). In addition, there was also a significant difference between depots within the D6/D7/D8 strata (FDR-Corrected P-Value = 0.0092; Table 3.10 and Figure 3.3). However, there was not a statistically significant depot difference between D0 samples (FDR-Corrected P-Value = 1.0000; Table 3.8

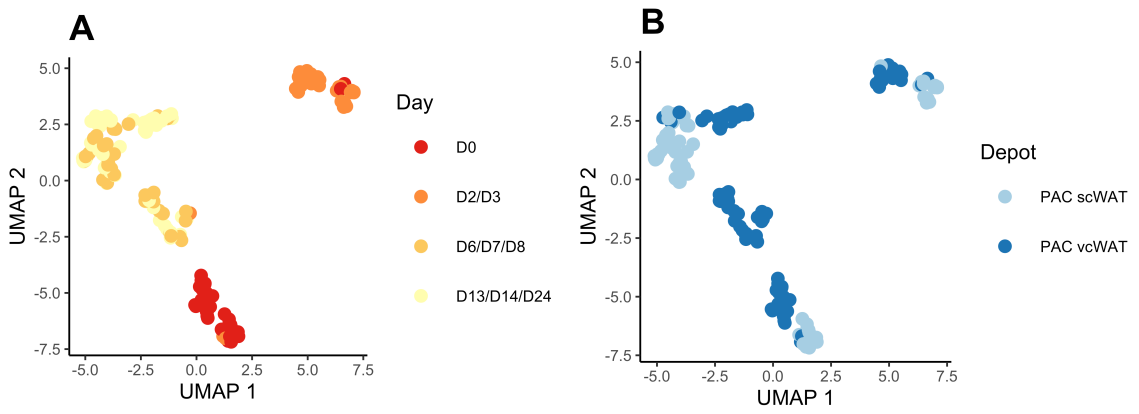

Figure 3.2: Visualization of top 25 RNA-Seq PCA components using UMAP, colored by (A) day and (B) depot.

and Figure 3.3). There were no further associations with any other phenotypes that were found to be significant after FDR-Correction (Supplementary Table 2).

Phenotype	**Chi-Square Test Statistic**	**P-Value**	**FDR-Corrected P-Value**
Depot	0.9198	0.5560	1.0000
Sex	1.0096	0.5720	1.0000
Age	3.5538	0.0964	1.0000
BMI	3.5538	0.0986	1.0000
T2D Status	0.5988	0.5760	1.0000
Mean Cell Area	NA	NA	NA
Cell Area Variance	NA	NA	NA
% Small Cells	NA	NA	NA

Table 3.8: RNA-Seq PCA clustering results (stratified by day = D0). FDR-Corrected P-Values were produced after all RNA-Seq PCA analyses were finished. Histology-based phenotypes were not analyzed due to identical K-Means labels between samples.

Phenotype	**Chi-Square Test Statistic**	**P-Value**	**FDR-Corrected P-Value**
Depot	28.5921	0.0001	0.0092
Sex	0.1411	0.7480	1.0000
Age	0.1514	0.7700	1.0000
BMI	0.7231	0.5360	1.0000
T2D Status	0.2391	0.7560	1.0000
Mean Cell Area	2.1034	0.2070	1.0000
Cell Area Variance	7.0783	0.0159	1.0000
% Small Cells	9.9913	0.0024	0.2088

Table 3.9: RNA-Seq PCA clustering results (stratified by day = D2/D3). FDR Corrected P-Values were produced after all RNA-Seq PCA analyses were finished.

Phenotype	**Chi-Square Test Statistic**	**P-Value**	**FDR-Corrected P-Value**
Depot	21.7778	0.0001	0.0092
Sex	0.0000	0.9990	1.0000
Age	0.4500	0.7360	1.0000
BMI	0.0000	0.9990	1.0000
T2D Status	0.4500	0.7360	1.0000
Mean Cell Area	0.0139	0.9990	1.0000
Cell Area Variance	0.4857	0.6280	1.0000
% Small Cells	0.2355	0.9990	1.0000

Table 3.10: RNA-Seq PCA clustering results (stratified by day = D6/D7/D8). FDR Corrected P-Values were produced after all RNA-Seq PCA analyses were finished.

Phenotype	Chi-Square Test Statistic	P-Value	FDR-Corrected P-Value
Depot	14.7197	0.0003	0.0267
Sex	0.2097	0.7640	1.0000
Age	0.4856	0.5990	1.0000
BMI	0.2394	0.7860	1.0000
T2D Status	0.1377	0.7960	1.0000
Mean Cell Area	0.0000	0.9990	1.0000
Cell Area Variance	1.9429	0.2890	1.0000
% Small Cells	0.0826	0.9990	1.0000

Table 3.11: RNA-Seq PCA clustering results (stratified by day = D13/D14/D24). FDR-Corrected P-Values were produced after all RNA-Seq PCA analyses were finished.

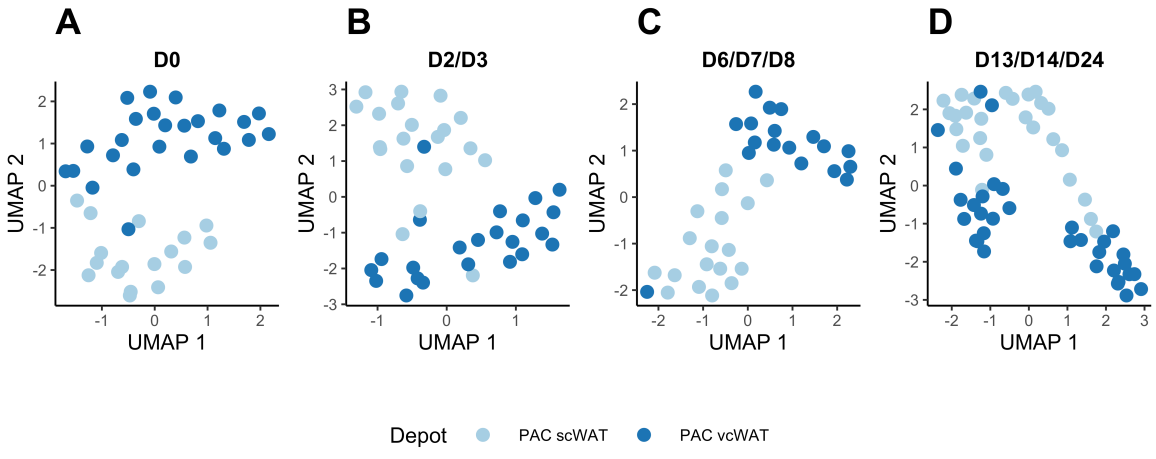

Figure 3.3: Visualiztion of (A) D0, (B) D2/D3, (C) D6/D7/D8, and (D) D13/D14/D24 RNA-Seq PCA components using UMAP, colored by depot.

3.4 Multi-View Clustering using MOFA+

The clustering pipeline was applied to the analysis of MOFA+ factors in order to determine which phenotypes were associated with differences across both RNA-Seq and Adipocyte Profiling data. Similar to what was found in the single-view RNA-Seq PCA clustering analysis, both day (FDR Corrected P-Value = 0.0078) and depot (FDR-Corrected P-Value = 0.0450) were significantly associated with clustering differences among samples after FDR-Correction (Table 3.12 and Figure 3.4). Again, day was the most significant hit with a chi-square test statistic of 307.1041, so it was used to stratify samples into day-specific groups for further analysis. There were no further significant associations with any other phenotype before stratification. In addition, MOFA+ latent factors that explained more variability were more likely to do so for each view separately, while factors that explained less variability were able to jointly model variation better (Figure 3.5).

Phenotype	**Chi-Square Test Statistic**	**P-Value**	**FDR-Corrected P-Value**
Day	307.1041	0.0001	0.0078
Depot	11.8107	0.0006	0.0450
Sex	0.0656	0.8580	1.0000
Age	0.7119	0.4310	1.0000
BMI	1.1009	0.3450	1.0000
T2D Status	0.8857	0.4350	1.0000
Mean Cell Area	0.0202	0.9990	1.0000
Cell Area Variance	1.2962	0.2960	1.0000
% Small Cells	0.7279	0.5240	1.0000

Table 3.12: MOFA+ clustering results (no stratification). FDR-Corrected P-Values were produced after all MOFA+ analyses.

Samples were stratified by day and then carried forth for another round of clustering analysis in order to determine if there were any day-specific clustering effects shared between both omics. There was a significant association with depot among D6/D7/D8 (FDR-Corrected P-Value = 0.0078) and D13/D14/D24 (FDR-Corrected P-Value = 0.0078) strata, which was consistent with what was found in the RNA-Seq PCA clustering analysis (Tables 3.15, 3.16 and Figure 3.6). There was no significant association with depot among D0 samples, which was also consistent with what was

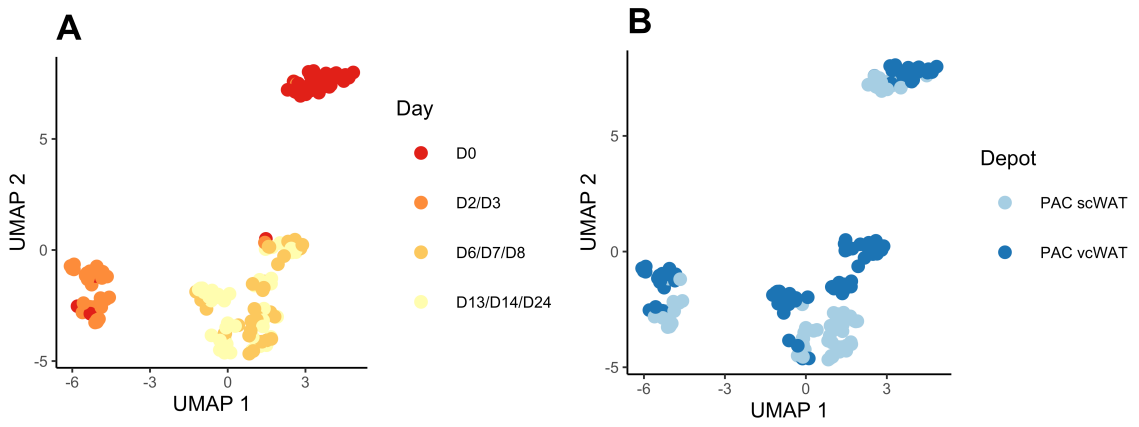

Figure 3.4: Visualization of MOFA+ factors using UMAP, colored by (A) day and (B) depot.

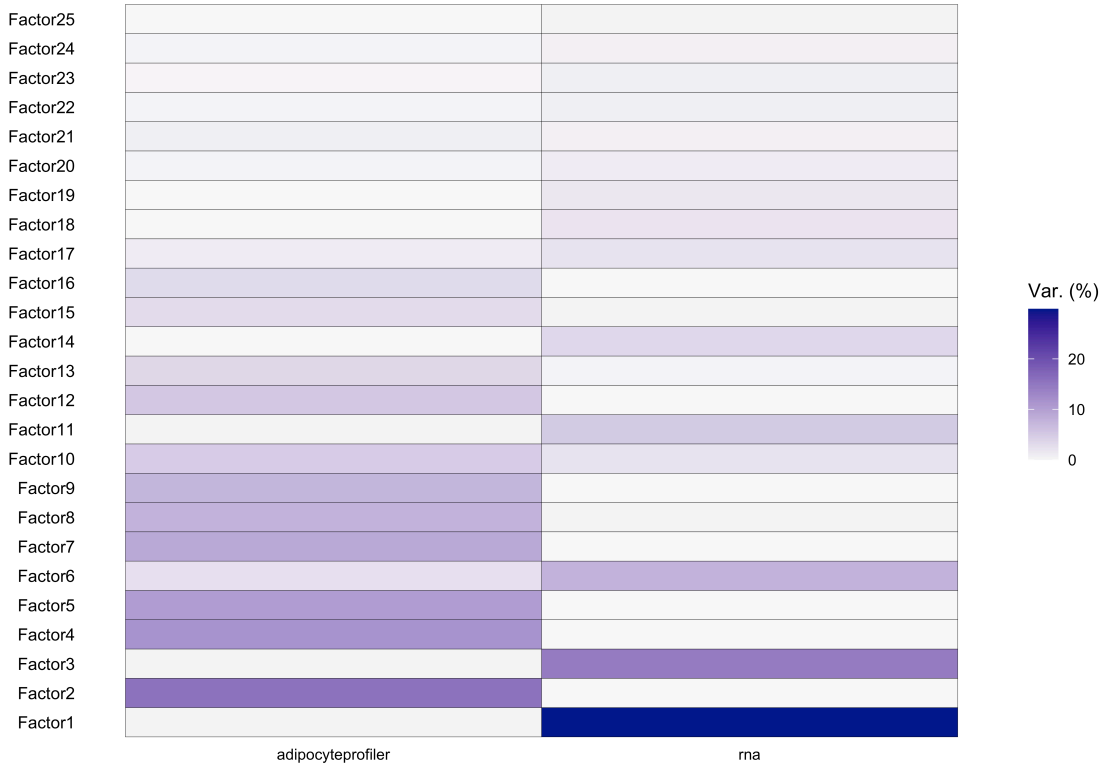

Figure 3.5: Variance contribution per view of top 25 MOFA+ factors. Generated using Argelaguet et al. 2020.

previously found (Table 3.13 and Figure 3.6). However, there was also no significant association found with depot among the D2/D3 samples (Table 3.14 and Figure 3.6). After FDR-correction, there were no further significant phenotype associations

(Supplementary Table 3).

Phenotype	**Chi-Square Test Statistic**	**P-Value**	**FDR-Corrected P-Value**
Depot	0.9198	0.5660	1.0000
Sex	1.0096	0.5710	1.0000
Age	3.5538	0.1020	1.0000
BMI	3.5538	0.1040	1.0000
T2D Status	0.5988	0.5720	1.0000
Mean Cell Area	NA	NA	NA
Cell Area Variance	NA	NA	NA
% Small Cells	NA	NA	NA

Table 3.13: MOFA+ clustering results (stratified by day = D0). FDR-Corrected P-Values were produced after all RNA-Seq PCA analyses were finished. Histology-based phenotypes were not analyzed due to identical K-Means labels between samples.

Phenotype	**Chi-Square Test Statistic**	**P-Value**	**FDR-Corrected P-Value**
Depot	0.5266	0.5910	1.0000
Sex	1.3975	0.5370	1.0000
Age	2.8043	0.2370	1.0000
BMI	2.8043	0.2350	1.0000
T2D Status	0.3102	0.9990	1.0000
Mean Cell Area	0.0041	0.9990	1.0000
Cell Area Variance	0.0379	0.9990	1.0000
% Small Cells	2.0079	0.4790	1.0000

Table 3.14: MOFA+ clustering results (stratified by day = D2/D3). FDR-Corrected P-Values were produced after all RNA-Seq PCA analyses were finished.

Phenotype	**Chi-Square Test Statistic**	**P-Value**	**FDR-Corrected P-Value**
Depot	22.0500	0.0001	0.0078
Sex	0.2864	0.7320	1.0000
Age	0.0056	0.9990	1.0000
BMI	0.0056	0.9990	1.0000
T2D Status	0.3600	0.7400	1.0000
Mean Cell Area	0.2355	0.9990	1.0000
Cell Area Variance	1.4310	0.3310	1.0000
% Small Cells	0.0156	0.9990	1.0000

Table 3.15: MOFA+ clustering results (stratified by day = D6/D7/D8). FDR Corrected P-Values were produced after all RNA-Seq PCA analyses were finished.

Phenotype	**Chi-Square Test Statistic**	**P-Value**	**FDR-Corrected P-Value**
Depot	16.1853	0.0001	0.0078
Sex	0.8730	0.3870	1.0000
Age	0.0153	0.9990	1.0000
BMI	0.1381	0.8000	1.0000
T2D Status	0.0153	0.9990	1.0000
Mean Cell Area	0.1245	0.9990	1.0000
Cell Area Variance	3.1136	0.1560	1.0000
% Small Cells	0.0069	0.9990	1.0000

Table 3.16: MOFA+ clustering results (stratified by day = D13/D14/D24). FDR Corrected P-Values were produced after all RNA-Seq PCA analyses were finished.

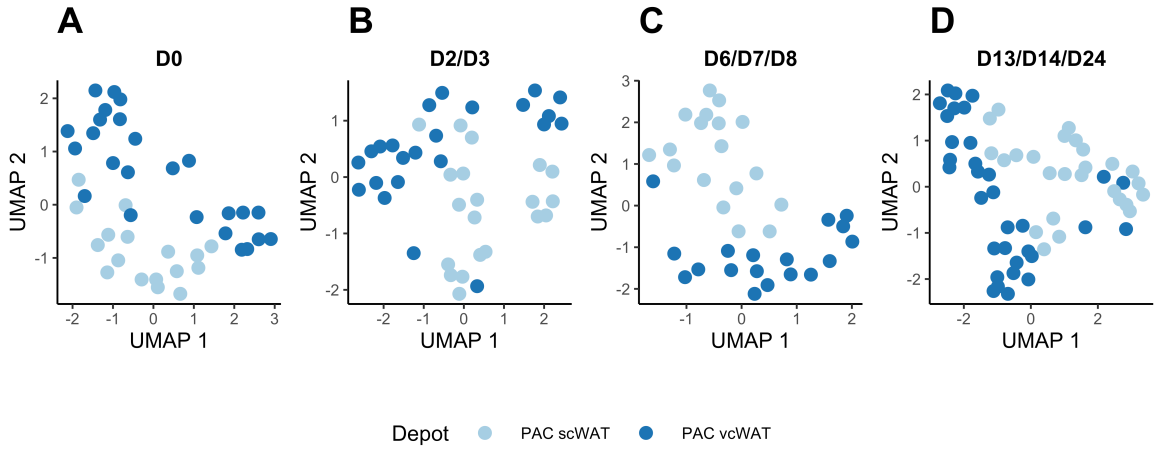

Figure 3.6: Visualization of (A) D0, (B) D2/D3, (C) D6/D7/D8, and (D) D13/D14/D24 MOFA+ factors using UMAP, colored by depot.

3.5 Multi-View Clustering using MCCA

The clustering pipeline was applied to 25 pairs of MCCA variates in order to determine which phenotypes were associated with large-scale variation shared between RNA-Seq and Adipocyte Profiling data. Similar to what was found in the both RNA-Seq PCA clustering and MOFA+ clustering analyses, day (FDR-Corrected P-Value = 0.0092) and depot (FDR-Corrected P-Value = 0.0092) were again significantly associated with clustering differences among samples after FDR-Correction (Table 3.17 and Figure 3.7). Day had the highest chi-square test statistic (153.4811), so it was used to stratify the samples for further clustering. MCCA did not detect any further associations with any other phenotypes before stratification, which is consistent with all previous single-view and multi-view analyses.

Phenotype	**Chi-Square Test Statistic**	**P-Value**	**FDR-Corrected P-Value**
Day	153.4811	0.0001	0.0092
Depot	35.3571	0.0001	0.0092
Sex	0.1724	0.7380	1.0000
Age	0.3339	0.6540	1.0000
BMI	1.0731	0.3730	1.0000
T2D Status	0.3571	0.6620	1.0000
Mean Cell Area	0.0881	0.8390	1.0000
Cell Area Variance	1.7831	0.2330	1.0000
% Small Cells	1.2605	0.3180	1.0000

Table 3.17: MCCA clustering results (no stratification). FDR-Corrected P-Values were produced after all MCCA analyses.

The MCCA data was grouped by day and analyzed once again with the clustering pipeline to determine which remaining phenotypes had day-specific clustering effects among MCCA variates. After stratifying by day, there were significant differences among samples from different depots within the D6/D7/D8 (FDR-Corrected P-Value = 0.0092) and D13/D14/D24 (FDR-Corrected P-Value = 0.0092) strata (Tables 3.20,3.21 and Figure 3.8). Like the RNA-Seq PCA clustering analyses, there were no significant depot differences among D0 samples (FDR-Corrected P-Value = 1.0000; Table 3.18 and Figure 3.8) but there were among D2/D3 samples (FDR-Corrected P-Value = 0.0092; Table 3.19 and Figure 3.8). No further significant associations with any other phenotype were detected (Supplementary Table 4).

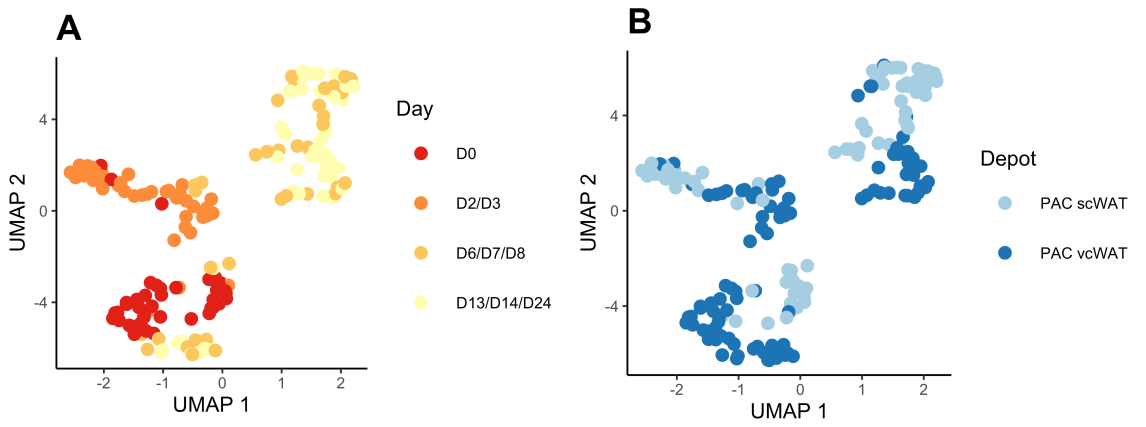

Figure 3.7: Visualization of top 25 MCCA variate pairs using UMAP, colored by (A) day and (B) depot.

Phenotype	**Chi-Square Test Statistic**	**P-Value**	**FDR-Corrected P-Value**
Depot	0.9198	0.5660	1.0000
Sex	1.0096	0.5710	1.0000
Age	3.5538	0.1020	1.0000
BMI	3.5538	0.1040	1.0000
T2D Status	0.5988	0.5720	1.0000
Mean Cell Area	NA	NA	NA
Cell Area Variance	NA	NA	NA
% Small Cells	NA	NA	NA

Table 3.18: MCCA clustering results (stratified by day = D0. FDR-Corrected P Values were produced after all MCCA analyses were finished. Histology-based phenotypes were not analyzed due to identical K-Means labels between samples.

Phenotype	**Chi-Square Test Statistic**	**P-Value**	**FDR-Corrected P-Value**
Depot	22.3484	0.0001	0.0092
Sex	0.9418	0.5040	1.0000
Age	0.1514	0.7700	1.0000
BMI	0.7231	0.5360	1.0000
T2D Status	0.0168	0.9990	1.0000
Mean Cell Area	0.0232	0.9990	1.0000
Cell Area Variance	1.8061	0.2250	1.0000
% Small Cells	3.6301	0.0861	1.0000

Table 3.19: MCCA clustering results (stratified by day = D2/D3. FDR-Corrected P-Values were produced after all MCCA analyses were finished.

Phenotype	**Chi-Square Test Statistic**	**P-Value**	**FDR-Corrected P-Value**
Depot	28.8000	0.0001	0.0092
Sex	0.2864	0.7340	1.0000
Age	0.0056	0.9990	1.0000
BMI	0.3600	0.7300	1.0000
T2D Status	0.0056	0.9990	1.0000
Mean Cell Area	0.4857	0.6340	1.0000
Cell Area Variance	2.9514	0.1540	1.0000
% Small Cells	1.4310	0.3350	1.0000

Table 3.20: MCCA clustering results (stratified by day = D6/D7/D8. FDR-Corrected P-Values were produced after all MCCA analyses were finished.

Phenotype	Chi-Square Test Statistic	P-Value	FDR-Corrected P-Value
Depot	23.3907	0.0001	0.0092
Sex	0.2564	0.7690	1.0000
Age	0.1509	0.7960	1.0000
BMI	0.0153	0.9990	1.0000
T2D Status	0.1509	0.7960	1.0000
Mean Cell Area	0.1193	0.9990	1.0000
Cell Area Variance	2.9825	0.1640	1.0000
% Small Cells	0.4242	0.7430	1.0000

Table 3.21: MCCA clustering results (stratified by day = D13/D14/D24. FDR Corrected P-Values were produced after all MCCA analyses were finished.

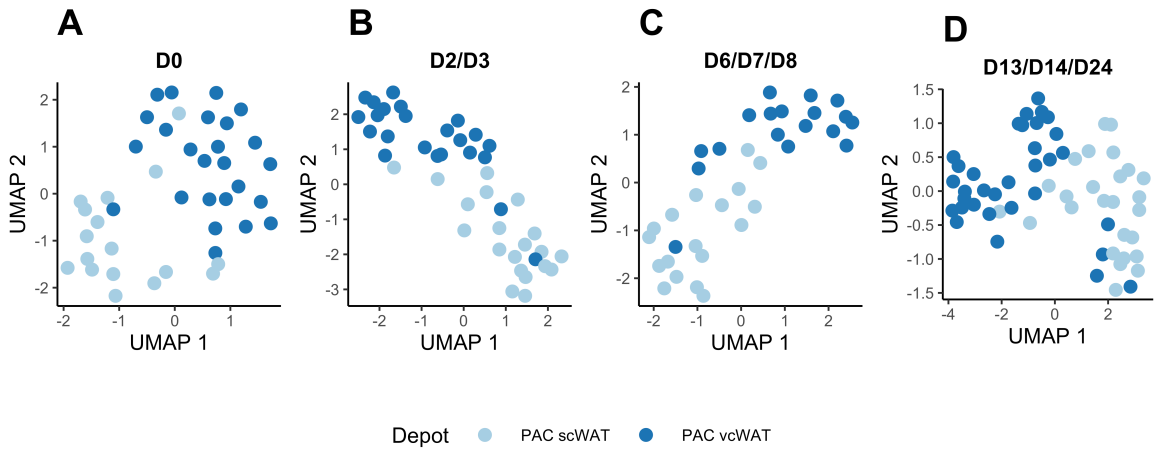

Figure 3.8: Visualization of top 25 (A) D0, (B) D2/D3, (C) D6/D7/D8, and (D) D13/D14/D24 MCCA variate pairs using UMAP, colored by depot.

Chapter 4

Discussion

4.1 Summary of Results

This work applied both single-view and multi-view learning clustering methods to study adipogenesis. One goal was to compare and contrast the utility of these methods for identifying phenotypes that drive biological diversity between human adipocytes. For single-view learning, PCA was used to cluster Adipocyte Profiling and RNA-Seq data separately. In the Adipocyte Profiling data there was no significant association with day or depot before stratification, but there were differences in depot within D2/D3 and D13/D14/D24 strata after stratifying by day (Table 4.1). Day and depot were both significantly associated with clustering differences in RNA-Seq data before stratification, and after stratifying by day there were day-specific differences in depot within D2/D3, D6/D7/D8, and D13/D14/D24 strata (Table 4.1). For multi-view learning, both MOFA+ and MCCA were used to cluster both omics jointly. Clustering of MOFA+ factors identified significant differences with regards to day and depot before stratification, and after stratifying by day there were depot differences within D6/D7/D8 and D13/D14/D24 groups (Table 4.1). MCCA replicated all of the same significant hits seen through the single-view RNA-Seq analysis (Table 4.1).

Phenotype	Stratification	AP PCA	RNA PCA	MOFA+	MCCA
Day	None	0.4224	0.0092	0.0078	0.0092
Depot	None	0.0736	0.0352	0.0450	0.0092
Depot	D0	1.0000	1.0000	1.0000	1.0000
Depot	D2/D3	0.0182	0.0092	1.0000	0.0092
Depot	D6/D7/D8	0.0720	0.0092	0.0078	0.0092
Depot	D13/D14/D24	0.0092	0.0267	0.0078	0.0092

Table 4.1: Overview of single-view and multi-view clustering results, where FDR-Corrected P-Values are displayed for each data-method pair. AP = adipocyte profiling; MCCA = multiple canonical correlation analysis; MOFA+ = multi-omics factor analysis; PCA = principal components analysis.

4.2 Interpretation of Results

Single-view PCA, multi-view MOFA+, and multi-view MCCA all led to the same broad conclusion: in this study, day and depot are major drivers of clustering differences among human adipocytes. However, they disagreed to the extent that these differences existed. This was the case for single-view analyses, where day and depot were significantly associated with differences in RNA-Seq data before stratification but not with Adipocyte Profiling data (Table 4.1). A lack of concurrence also existed more broadly between single-view and multi-view methods with regards day-specific depot differences (Table 4.1). For single-view analyses, this may be partly due to the different omics used, which represent different underlying biological signals. But, there are also different assumptions between these methods: single-view PCA assumes that variation is driven separately, multi-view MCCA assumes that variation is driven jointly, and MOFA+ allows for both separate and joint explanation of variation. These differing assumptions may lead to different strengths for each method and hence different findings. For instance, an advantage of multi-view learning over single-view learning is that it is able to integrate signals from multiple views at once, but this may risk losing a signal that is present in one view if the opposite effect is seen in the other. This may be the case for MOFA+, which did not retain the D2/D3 depot signal seen in the RNA-Seq PCA clustering analyses. Between multi-view learning methods, the possibility of joint explanation of variation may compensate for this, but at the possible expense of making an assumption that may be too strong. MCCA may have suffered from this limitation given that it kept day, depot, and a

day-specific depot association at D6/D7/D8 even though these were not significantly associated with clustering differences for the Adipocyte Profiling data. These findings suggest that every tool has its own unique strengths and weaknesses, so utilizing multiple tools to 'triangulate' biological discovery may be the preferred approach for generating robust insights (Munafò and Smith 2018).

The finding that day and depot were major drivers of clustering differences was expected since both are known to be primary sources of heterogeneity among differentiating adipocytes. Day was expected to be significant because adipogenesis is a developmental process involving the activation of time-specific genetic programs (Tang and Lane 2012). Also, there have been longitudinal gene expression studies within human adipocytes that have observed notable differences in gene expression over the course of adipogenesis (Ambele et al. 2016; Ehrlund et al. 2016; Côté et al. 2017; Ramirez et al. 2020). Differences in gene expression have also been observed for adipocytes coming from different adipose tissue depots, suggesting that adipocyte function depends on spatial location (Atzmon et al. 2002; Linder et al. 2004). A recent study has also demonstrated that differences in gene expression between depots across time may be related to obesity, drawing a connection between day, depot, and metabolic disease (Breitfeld et al. 2020). Through the systematic, unbiased clustering approach used in this study, day and depot were identified as the primary sources of heterogeneity in these data, corroborating findings in the broader literature. These findings also provide further confidence in systematic pipelines like the one used in this study for their ability to generate robust biological insights in an unbiased manner.

Despite the finding that day and depot were primary drivers of clustering among human adipocytes, there were several phenotypes that were expected to drive clustering differences but were not found to do so in this study. For instance, some studies have demonstrated that the expression of adipogenic genes are reduced in obese individuals, defined as having a BMI above 30.0 (Nadler et al. 2000; Dubois et al. 2006). The current literature also suggests that there are sex-specific differences in adipose tissue development (Anderson et al. 2020) and function (White and Tchoukalova 2014; Chang et al. 2018). These studies, in addition to the broader literature, support the notion that both sex and BMI drive differences in adipocyte biology, so the lack of replication in this study was unexpected. One possibility for this null result is that these phenotypes were too unbalanced or skewed in these data. This is supported by the descriptive statistics for sex, where there were about 5 female samples for every 2 male samples (Table 3.1). Most of the individuals in this study also had a BMI above 30 since they were recruited from an obesity cohort, so this skew could

have affected downstream analyses since the full spectrum of BMI was not well represented. More balance within these phenotypes would have helped detect possible clustering differences. Additionally, the loss of power from successive stratification could have also contributed. Since stratification reduces power via reduced sample size, it is possible that even a very strong signal for either of these phenotypes could have been lost after stratifying by potentially stronger signals like day and depot. For other phenotypes outside of sex and BMI, this issue of power may have had a role since age (Kirkland et al. 2002, Cartwright et al. 2007, Mancuso and Bouchard 2019), T2D status (Guilherme et al. 2008), and cell size (Stenkula and Erlanson-Albertsson 2018) have all been associated with differences in adipogenesis.

4.3 Future Directions and Broader Applicability

This work described a multi-omics analysis of adipogenesis, using single-view and multi-view clustering methods to systematically identify major drivers of variation among differentiating human adipocytes. Systematic clustering analyses identified both day and depot as drivers of clustering, which is consistent with previous literature. This consistent finding suggests that systematic clustering pipelines like the one used in this work may be effective for the identification of robust differences between biological entities. Nonetheless, multi-view clustering methods did not garner any new biological insights nor identify associations with other expected phenotypes like sex, age, BMI, T2D status, or cell size. The discrepancy between this work and the outer literature may exist because it is not only the methods but also the underlying data that powers multi-omics discovery.

One way to potentially advance this work is to build upon the underlying data. For one, the inclusion of more samples could minimize the risk of under-powered statistical analyses from stratification. Ensuring that phenotypes like sex and BMI are more balanced like sex and BMI could also help increase statistical power for biological discovery. In addition to both of these suggestions, the inclusion of more omics layers is likely to also improve power for multi-omics clustering analyses. To date, genotyping methods have already been used to explore how common genetic variation is associated with differential risk for metabolic diseases (Speliotes et al. 2010; Wang et al. 2011; Locke et al. 2015; Rask-Andersen et al. 2019). In addition, ATAC-Sequencing, which can be used to measure genome-wide accessibility, has already been used to assess which accessible genomics regions in adipocytes harbor this genetic variation (Cannon et al. 2019). Given these findings, the inclusion of genotype and

ATAC-Seq data may be of particular interest for this study. Both of these omics, paired with the RNA-Seq data already generated for this study, may enable the investigation of how common human genetic variation and chromatin architecture is related to differential gene regulation during adipogenesis in both healthy and diseased states. Further inclusion of Adipocyte Profiling Data could also enable a high-level investigation of how differences in gene regulation among human adipocytes are associated with morphological differences across time.

In addition to what omics are currently available, multi-view learning may find applicability for new kinds of omics in the near future. One emerging area is that of spatial omics, where new technologies like Slide-Seq (Rodriques et al. 2019) and seqFISH+ (Eng et al. 2019) are able to generate high-dimensional biological data within their spatial context. For instance, spatially-resolved transcriptomics, which is a set of techniques that generate gene expression data in intact tissues, was recently recognized as Nature Method of the Year (Marx 2021; Larsson et al. 2021). Along with the development of spatial transcriptomic methods, spatial genomic methods have also been developed that can enable 3-dimensional genome analysis at single-cell resolution (Zhuang 2021). As the technology underlying these methods progresses further, multi-view learning may be used for spatial multi-omics analyses - generating multi-omic biological insights within their spatial context. Outside of spatial multi-omics, multi-view learning may also find applicability within the nascent field of temporal omics. To date, recent improvements in live cell tracking methods using deep learning have enabled improved analysis of real-time cellular phenotyping data (Moen et al. 2019). In addition, early work has demonstrated that live cell tracking combined with single-cell RNA-Seq data can help tease apart morphodynamic states between microglia (Wu et al. 2020). There is still a lot of room for technological innovation in this space, and if researchers are able to develop real-time molecular omic technologies then it is possible to envision a future where multi-view signal-processing methods may be used for the analysis of real-time omics. Given that this work found that both space (depot) and time (day) were major drivers of biological heterogeneity among human adipocytes, it is fitting that multi-view learning may one day be paired with new technologies to measure at higher resolution how space and time influence other complex biological processes.

The development and application of new spatial and temporal technologies is likely to yield new insights where much is still unknown, including adipocyte biology. The analyses here suggest that both day and depot are major drivers of differences among human adipocytes and perhaps even more important than patient-specific clinical

characteristics. Nonetheless, there is still much to be learned with regards to how temporal and spatial variation intersect with patient-specific clinical differences in the context of human metabolism. With the inclusion of new types of data and more of it, scientists will be able to generate new basic insights about adipogenesis and address these fundamental questions. Then, one day, the field may be able to use this knowledge to develop new therapeutic strategies for metabolic disease and hopefully alleviate human suffering in a meaningful way.

Bibliography

Ambele et al. (May 2016). "Genome-wide analysis of gene expression during adipogenesis in human adipose-derived stromal cells reveals novel patterns of gene expression during adipocyte differentiation". In: *Stem Cell Research* 16.3, pp. 725–734. DOI: 10.1016/j.scr.2016.04.011. URL: https://doi.org/10.1016/j.scr.2016.04.011.

Anders and Huber (Oct. 2010). "Differential expression analysis for sequence count data". In: *Genome Biology* 11.10. DOI: 10.1186/gb-2010-11-10-r106. URL: https://doi.org/10.1186/gb-2010-11-10-r106.

Anderson et al. (Sept. 2020). "Sex differences in human adipose tissue gene expression and genetic regulation involve adipogenesis". In: *Genome Research* 30.10, pp. 1379–1392. DOI: 10.1101/gr.264614.120. URL: https://doi.org/10.1101/gr.264614.120.

Andrews et al. (Jan. 2019). *FastQC*. Babraham Institute. Babraham, UK.

Argelaguet et al. (May 2020). "MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data". In: *Genome Biology* 21.111. DOI: 10.1186/s13059-020-02015-1.

— (June 2018). "Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets". In: *Molecular Systems Biology* 14.6. DOI: 10.15252/msb.20178124. URL: https://doi.org/10.15252/msb.20178124.

Atzmon et al. (Nov. 2002). "Differential Gene Expression Between Visceral and Subcutaneous Fat Depots". In: *Hormone and Metabolic Research* 34.11/12, pp. 622–628. DOI: 10.1055/s-2002-38250. URL: https://doi.org/10.1055/s-2002-38250.

Bjørndal et al. (2011). "Different Adipose Depots: Their Role in the Development of Metabolic Syndrome and Mitochondrial Response to Hypolipidemic Agents". In: *Journal of Obesity* 2011, pp. 1–15. DOI: 10.1155/2011/490650. URL: https://doi.org/10.1155/2011/490650.

Bray et al. (Aug. 2016). "Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes". In: *Nature Protocols* 11.9, pp. 1757–1774. DOI: 10.1038/nprot.2016.105. URL: https://doi.org/10.1038/nprot.2016.105.

Breitfeld et al. (Mar. 2020). "Developmentally Driven Changes in Adipogenesis in Different Fat Depots Are Related to Obesity". In: *Frontiers in Endocrinology* 11. DOI: 10.3389/fendo.2020.00138. URL: https://doi.org/10.3389/fendo.2020.00138.

Cannon et al. (Aug. 2019). "Open Chromatin Profiling in Adipose Tissue Marks Genomic Regions with Functional Roles in Cardiometabolic Traits". In: *G3 Genes|Genomes|Geneti* 9.8, pp. 2521–2533. DOI: 10.1534/g3.119.400294. URL: https://doi.org/10.1534/g3.119.400294.

Cartwright et al. (June 2007). "Aging in adipocytes: Potential impact of inherent, depot-specific mechanisms". In: *Experimental Gerontology* 42.6, pp. 463–471. DOI: 10.1016/j.exger.2007.03.003. URL: https://doi.org/10.1016/j.exger.2007.03.003.

Chait and Hartigh (Feb. 2020). "Adipose Tissue Distribution, Inflammation and Its Metabolic Consequences, Including Diabetes and Cardiovascular Disease". In: *Frontiers in Cardiovascular Medicine* 7. DOI: 10.3389/fcvm.2020.00022. URL: https://doi.org/10.3389/fcvm.2020.00022.

Chang et al. (July 2018). "Gender and Sex Differences in Adipose Tissue". In: *Current Diabetes Reports* 18.9. DOI: 10.1007/s11892-018-1031-3. URL: https://doi.org/10.1007/s11892-018-1031-3.

Côté et al. (2017). "Temporal Changes in Gene Expression Profile during Mature Adipocyte Dedifferentiation". In: *International Journal of Genomics* 2017, pp. 1–11. DOI: 10.1155/2017/5149362. URL: https://doi.org/10.1155/2017/5149362.

Denny et al. (July 2016). "Nfib Promotes Metastasis through a Widespread Increase in Chromatin Accessibility". In: *Cell* 166.2, pp. 328–342. DOI: 10.1016/j.cell.2016.05.052. URL: https://doi.org/10.1016/j.cell.2016.05.052.

Dobin et al. (Oct. 2012). "STAR: ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29.1, pp. 15–21. DOI: 10.1093/bioinformatics/bts635. URL: https://doi.org/10.1093/bioinformatics/bts635.

Dubois et al. (Sept. 2006). "Decreased Expression of Adipogenic Genes in Obese Subjects with Type 2 Diabetes∗". In: *Obesity* 14.9, pp. 1543–1552. DOI: 10.1038/oby.2006.178. URL: https://doi.org/10.1038/oby.2006.178.

Ehrlund et al. (Nov. 2016). "Transcriptional Dynamics During Human Adipogenesis and Its Link to Adipose Morphology and Distribution". In: *Diabetes* 66.1, pp. 218–230. DOI: 10.2337/db16-0631. URL: https://doi.org/10.2337/db16-0631.

Eng et al. (Mar. 2019). "Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH". In: *Nature* 568.7751, pp. 235–239. DOI: 10.1038/s41586-019-1049-y. URL: https://doi.org/10.1038/s41586-019-1049-y.

Ewels et al. (June 2016). "MultiQC: summarize analysis results for multiple tools and samples in a single report". In: *Bioinformatics* 32.19, pp. 3047–3048. DOI: 10.1093/bioinformatics/btw354. URL: https://doi.org/10.1093/bioinformatics/btw354.

Ghaben and Scherer (Jan. 2019). "Adipogenesis and metabolic health". In: *Nature Reviews Molecular Cell Biology* 20.4, pp. 242–258. DOI: 10.1038/s41580-018-0093-z. URL: https://doi.org/10.1038/s41580-018-0093-z.

Glastonbury et al. (Aug. 2020). "Machine Learning based histology phenotyping to investigate the epidemiologic and genetic basis of adipocyte morphology and cardiometabolic traits". In: *PLOS Computational Biology* 16.8. Ed. by Lilia M. Iak-

oucheva, e1008044. DOI: 10.1371/journal.pcbi.1008044. URL: https://doi.org/10.1371/journal.pcbi.1008044.

Guilherme et al. (May 2008). "Adipocyte dysfunctions linking obesity to insulin resistance and type 2 diabetes". In: *Nature Reviews Molecular Cell Biology* 9.5, pp. 367–377. DOI: 10.1038/nrm2391. URL: https://doi.org/10.1038/nrm2391.

Ikeda et al. (Mar. 2018). "The Common and Distinct Features of Brown and Beige Adipocytes". In: *Trends in Endocrinology & Metabolism* 29.3, pp. 191–200. DOI: 10.1016/j.tem.2018.01.001. URL: https://doi.org/10.1016/j.tem.2018.01.001.

Inagaki et al. (June 2016). "Transcriptional and epigenetic control of brown and beige adipose cell fate and function". In: *Nature Reviews Molecular Cell Biology* 17.8, pp. 480–495. DOI: 10.1038/nrm.2016.62. URL: https://doi.org/10.1038/nrm.2016.62.

Jones et al. (2008). "CellProfiler Analyst: data exploration and analysis software for complex image-based screens". In: *BMC Bioinformatics* 9.1, p. 482. DOI: 10.1186/1471-2105-9-482. URL: https://doi.org/10.1186/1471-2105-9-482.

Kirkland et al. (June 2002). "Adipogenesis and aging: does aging make fat go MAD?" In: *Experimental Gerontology* 37.6, pp. 757–767. DOI: 10.1016/s0531-5565(02)00014-1. URL: https://doi.org/10.1016/s0531-5565(02)00014-1.

Krueger et al. (Nov. 2019). *Trim Galore!* Babraham Institute. Babraham, UK.

Larsson et al. (Jan. 2021). "Spatially resolved transcriptomics adds a new dimension to genomics". In: *Nature Methods* 18.1, pp. 15–18. DOI: 10.1038/s41592-020-01038-7. URL: https://doi.org/10.1038/s41592-020-01038-7.

Lefterova et al. (June 2014). "PPAR and the global map of adipogenesis and beyond". In: *Trends in Endocrinology & Metabolism* 25.6, pp. 293–302. DOI: 10.1016/j.tem.2014.04.001. URL: https://doi.org/10.1016/j.tem.2014.04.001.

Lever et al. (July 2017). "Principal component analysis". In: *Nature Methods* 14.7, pp. 641–642. DOI: 10.1038/nmeth.4346. URL: https://doi.org/10.1038/nmeth.4346.

Li and Dewey (Aug. 2011). "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome". In: *BMC Bioinformatics* 12.1. DOI: 10.1186/1471-2105-12-323. URL: https://doi.org/10.1186/1471-2105-12-323.

Linder et al. (Jan. 2004). "Differentially expressed genes in visceral or subcutaneous adipose tissue of obese men and women". In: *Journal of Lipid Research* 45.1, pp. 148–154. DOI: 10.1194/jlr.m300256-jlr200. URL: https://doi.org/10.1194/jlr.m300256-jlr200.

Lloyd (Mar. 1982). "Least squares quantization in PCM". In: *IEEE Transactions on Information Theory* 28.2, pp. 129–137. DOI: 10.1109/tit.1982.1056489. URL: https://doi.org/10.1109/tit.1982.1056489.

Locke, Adam E. et al. (Feb. 2015). "Genetic studies of body mass index yield new insights for obesity biology". In: *Nature* 518.7538, pp. 197–206. DOI: 10.1038/nature14177. URL: https://doi.org/10.1038/nature14177.

Love et al. (Dec. 2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15.12. DOI: 10.1186/s13059-014-0550-8. URL: https://doi.org/10.1186/s13059-014-0550-8.

Mancuso and Bouchard (Mar. 2019). "The Impact of Aging on Adipose Function and Adipokine Synthesis". In: *Frontiers in Endocrinology* 10. DOI: 10.3389/fendo.2019.00137. URL: https://doi.org/10.3389/fendo.2019.00137.

Marx (Jan. 2021). "Method of the Year: spatially resolved transcriptomics". In: *Nature Methods* 18.1, pp. 9–14. DOI: 10.1038/s41592-020-01033-y. URL: https://doi.org/10.1038/s41592-020-01033-y.

Minnoye et al. (Jan. 2021). "Chromatin accessibility profiling methods". In: *Nature Reviews Methods Primers* 1.1. DOI: 10.1038/s43586-020-00008-9. URL: https://doi.org/10.1038/s43586-020-00008-9.

Mittal (2019). "Subcutaneous adipose tissue & visceral adipose tissue". In: *Indian Journal of Medical Research* 149.5, p. 571. DOI: 10.4103/ijmr.ijmr_1910_18. URL: https://doi.org/10.4103/ijmr.ijmr_1910_18.

Moen et al. (Oct. 2019). "Accurate cell tracking and lineage construction in live-cell imaging experiments with deep learning". In: DOI: 10.1101/803205. URL: https://doi.org/10.1101/803205.

Munafò and Smith (Jan. 2018). "Robust research needs many lines of evidence". In: *Nature* 553.7689, pp. 399–401. DOI: 10.1038/d41586-018-01023-3. URL: https://doi.org/10.1038/d41586-018-01023-3.

Nadler et al. (Oct. 2000). "The expression of adipogenic genes is decreased in obesity and diabetes mellitus". In: *Proceedings of the National Academy of Sciences* 97.21, pp. 11371–11376. DOI: 10.1073/pnas.97.21.11371. URL: https://doi.org/10.1073/pnas.97.21.11371.

Nguyen and Wang (Apr. 2020). "Multiview learning for understanding functional multiomics". In: *PLoS Computational Biology* 16.4. DOI: 10.1371/journal.pcbi.1007677.

Niemelä (2008). "Adipose Tissue and Adipocyte Differentiation: Molecular and Cellular Aspects and Tissue Engineering Applications". In: *Topics in Tissue Engineering*. Ed. by Ashammakhi, Reis, and Chiellini. Mary Ann Liebert, Inc. Chap. 4, pp. 1–26.

Obesity and overweight (n.d.). https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight. Accessed: 2021-05-06.

Pedregosa Fabian, Varoquaux et al. (Nov. 2011). "Scikit-Learn: Machine Learning in Python". In: *J. Mach. Learn. Res.* 12.null, pp. 2825–2830. ISSN: 1532-4435.

Perrini et al. (Oct. 2007). "Fat depot-related differences in gene expression, adiponectin secretion, and insulin action and signalling in human adipocytes differentiated in vitro from precursor stromal cells". In: *Diabetologia* 51.1, pp. 155–164. DOI: 10.1007/s00125-007-0841-7. URL: https://doi.org/10.1007/s00125-007-0841-7.

Priest and Tontonoz (Dec. 2019). "Inter-organ cross-talk in metabolic syndrome". In: *Nature Metabolism* 1.12, pp. 1177–1188. DOI: 10.1038/s42255-019-0145-5. URL: https://doi.org/10.1038/s42255-019-0145-5.

Ramirez et al. (Apr. 2020). "Single-cell transcriptional networks in differentiating preadipocytes suggest drivers associated with tissue heterogeneity". In: *Nature Communications* 11.1. DOI: 10.1038/s41467-020-16019-9. URL: https://doi.org/10.1038/s41467-020-16019-9.

Rappoport and Shamir (Oct. 2018). "Multi-omic and multi-view clustering algorithms: review and cancer benchmark". In: *Nucleic Acids Research* 46.20, pp. 10546–10562. DOI: 10.1093/nar/gky889. URL: https://doi.org/10.1093/nar/gky889.

Rask-Andersen et al. (Jan. 2019). "Genome-wide association study of body fat distribution identifies adiposity loci and sex-specific genetic effects". In: *Nature Communications* 10.1. DOI: 10.1038/s41467-018-08000-4. URL: https://doi.org/10.1038/s41467-018-08000-4.

Rodriques et al. (Mar. 2019). "Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution". In: *Science* 363.6434, pp. 1463–1467. DOI: 10.1126/science.aaw1219. URL: https://doi.org/10.1126/science.aaw1219.

Sacks and Symonds (May 2013). "Anatomical Locations of Human Brown Adipose Tissue: Functional Relevance and Implications in Obesity and Type 2 Diabetes". In: *Diabetes* 62.6, pp. 1783–1790. DOI: 10.2337/db12-1430. URL: https://doi.org/10.2337/db12-1430.

Speliotes et al. (Oct. 2010). "Association analyses of 249, 796 individuals reveal 18 new loci associated with body mass index". In: *Nature Genetics* 42.11, pp. 937–948. DOI: 10.1038/ng.686. URL: https://doi.org/10.1038/ng.686.

Stenkula and Erlanson-Albertsson (Aug. 2018). "Adipose cell size: importance in health and disease". In: *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* 315.2, R284–R295. DOI: 10.1152/ajpregu.00257.2017. URL: https://doi.org/10.1152/ajpregu.00257.2017.

Tang and Lane (July 2012). "Adipogenesis: From Stem Cell to Adipocyte". In: *Annual Review of Biochemistry* 81.1, pp. 715–736. DOI: 10.1146/annurev-biochem-052110-115718. URL: https://doi.org/10.1146/annurev-biochem-052110-115718.

Wang et al. (Apr. 2011). "A Genome-Wide Association Study on Obesity and Obesity-Related Traits". In: *PLoS ONE* 6.4. Ed. by Zhongming Zhao, e18939. DOI: 10.1371/journal.pone.0018939. URL: https://doi.org/10.1371/journal.pone.0018939.

— (Aug. 2020). "Finding the needle in a high-dimensional haystack: Canonical correlation analysis for neuroscientists". In: *NeuroImage* 216, p. 116745. DOI: 10.1016/j.neuroimage.2020.116745. URL: https://doi.org/10.1016/j.neuroimage.2020.116745.

— (Jan. 2009). "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nature Reviews Genetics* 10.1, pp. 57–63. DOI: 10.1038/nrg2484. URL: https://doi.org/10.1038/nrg2484.

Way (2020). *Blocklist Features - Cell Profiler*. DOI: 10.6084/M9.FIGSHARE.10255811.V3. URL: https://figshare.com/articles/Blacklist_Features_-_Cell_Profiler/10255811/3.

White and Tchoukalova (Mar. 2014). "Sex dimorphism and depot differences in adipose tissue function". In: *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 1842.3, pp. 377–392. DOI: 10.1016/j.bbadis.2013.05.006. URL: https://doi.org/10.1016/j.bbadis.2013.05.006.

Witten and Tibshirani (Jan. 2009). "Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data". In: *Statistical Applications in Genetics and Molecular Biology* 8.1, pp. 1–27. DOI: 10.2202/1544-6115.1470. URL: https://doi.org/10.2202/1544-6115.1470.

Wu et al. (July 2020). "DynaMorph: learning morphodynamic states of human cells with live imaging and sc-RNAseq". In: DOI: 10.1101/2020.07.20.213074. URL: https://doi.org/10.1101/2020.07.20.213074.

Yu and Li (Dec. 2016). "The size matters: regulation of lipid storage by lipid droplet dynamics". In: *Science China Life Sciences* 60.1, pp. 46–56. DOI: 10.1007/s11427-016-0322-x. URL: https://doi.org/10.1007/s11427-016-0322-x.

Zhang et al. (Aug. 2017). "Reversal of hyperactive Wnt signaling-dependent adipocyte defects by peptide boronic acids". In: *Proceedings of the National Academy of Sciences* 114.36, E7469–E7478. DOI: 10.1073/pnas.1621048114. URL: https://doi.org/10.1073/pnas.1621048114.

Zhuang (Jan. 2021). "Spatially resolved single-cell genomics and transcriptomics by imaging". In: *Nature Methods* 18.1, pp. 18–22. DOI: 10.1038/s41592-020-01037-8. URL: https://doi.org/10.1038/s41592-020-01037-8.