Contextualizing Causal Genes from Bone Mineral Density GWAS Using Single-Cell Transcriptomics in Diversity Outbred Mice

Luke J. Dillard

Bachelor of Science (B.S.), Microbiology, Virginia Tech, 2017

A Dissertation presented to the Graduate Faculty of the University of Virginia in Candidacy for the Degree of Doctor of Philosophy

Department of Biology

University of Virginia

April 2024

Dr. Charles R. Farber (Advisor)
Dr. John Campbell (First Reader)
Dr. Ani Manichaikul
Dr. Martin Wu
Dr. Clint Miller

Abstract

Genome-wide association studies (GWASs) have advanced our understanding of the genetics of human bone mineral density (BMD), a clinical predictor of fracture risk and osteoporosis. Aside from the identification of causal genes, other difficult challenges to leveraging GWAS include characterizing the roles of predicted causal genes in disease and providing additional functional context, such as the cell type predictions or biological pathways in which causal genes operate. Using single cell transcriptomics (scRNA-seq) can augment the utility of BMD GWAS by linking variants and genes to a cell type context in which these causal genes may drive disease; however, few population-level single-cell transcriptomics data sets have been generated on bone. To address this challenge, we demonstrate the utility of bone marrow–derived stromal cells cultured under osteogenic conditions (BMSC-OBs) in large populations of Diversity Outbred (DO) mice. The BMSC-OB model can be used to generate cell type–specific transcriptomic profiles of mesenchymal lineage cells to inform human genomic studies. By enriching for mesenchymal lineage cells *in vitro*, coupled with pooling of multiple samples for scRNA-seq preparation and downstream genotype deconvolution, we demonstrate the scalability of the BMSC-OB model for population-level studies. Furthermore, we show that BMSC-OBs are diverse and consist of bone-relevant cells, such as osteoblasts, osteocyte-like cells, marrow adipogenic lineage precursors (MALPs), and cells with characteristics of mesenchymal progenitors. Through the use of scRNA-seq analytical tools, we confirm the biological identities of BMSC-OBs and show that their transcriptomic profiles are similar to cells isolated *in vivo*. To contextualize BMD GWAS-implicated causal genes and prioritize targets for subsequent investigations,

we generated cell type-specific gene co-expression networks (GCNs). Using pseudotime trajectories inferred from the BMSC-OB scRNA-seq data, we identify networks enriched with genes that exhibit the most dynamic changes in expression across trajectories. We discover 21 network driver genes, which are also causal genes that have human BMD GWAS associations that colocalize with expression/splicing quantitative trait loci (eQTL/sQTL). These driver genes, including *Fgfrl1* and *Tpx2* (along with their associated networks), are predicted to be novel regulators of BMD via their roles in the differentiation of mesenchymal lineage cells. In this work, we showcase the power of single-cell transcriptomics from mouse bone-relevant cells and human BMD GWAS to prioritize genetic targets with potential causal roles in the development of osteoporosis.

Table of Contents

Abstract I

Table of Contents III

List of Figures VI

Chapter 1 - Introduction 1

1.1. An overview of the genetics of osteoporosis and bone mineral density 2

1.2. Characterizing the genetics of BMD via GWAS 3

1.3. Systems genetics approaches utilizing transcriptomics data 4

1.3.1. Expression Quantitative Trait Loci (eQTL) Analysis 6

1.3.2. Splicing Quantitative Trait Loci (sQTL) Analysis 9

1.3.3. Network Analyses using transcriptomics data 12

1.4. Single-cell transcriptomics 19

1.4.1. Summary of bulk transcriptomics (RNA-seq) 20

1.4.2. Single-cell and RNA capture 21

1.4.3. Dimensionality reduction and clustering 23

1.4.4. Pseudotime Trajectory Analysis 27

1.4.5. Limitations 30

1.5. Using the Diversity Outbred mouse population as a model to study bone 31

1.6. BMSCs and in vitro osteogenic differentiation 34

1.7. Summary 37

1.8. Chapter 1 - Main Figures 39

Chapter 2 - Single-Cell Transcriptomics of Bone Marrow Stromal Cells in Diversity Outbred Mice: A Model for Population-Level scRNA-Seq Studies 45

2.1. Abstract 46

2.2. Introduction 47

2.3. Materials and Methods 49

2.3.1. Sample preparation and in vitro cell culture of BMSCs 49

2.3.2. Single-cell isolation procedure 50

2.3.3. Single-cell analysis pipeline 51

2.3.4. Bulk RNA-seq analysis 52

2.3.5. Integration of data sets via canonical correlation analysis (CCA) 52

2.3.6. Souporcell 53

2.3.7. Scenic 53

2.3.8. CELLECT 54

2.4. Results 55

2.4.1. BMSC cultures grown under osteogenic differentiation conditions are heterogenous 55
2.4.2. Cell clustering is robust to the effects of cell isolation 56
2.4.3. Cell types isolated from BMSC-OBs are similar to their in vivo counterparts 58
2.4.4. Transcriptomic profiles from scRNA-seq for individual cell types are robust 59
2.4.5. Frequency of osteogenic cell types are highly variable across DO mice 60
2.4.6. BMSC-OBs show expected gene regulatory networks 61
2.4.7. MALPs and osteogenic cells capture BMD heritability identified by GWAS 62
2.5. Discussion 63
2.6. Acknowledgements 68
2.7. Author Contributions 69
2.8. Disclosures 69
2.9. Data Availability Statement 69
2.10. Chapter 2 - Main Figures 70
2.11. Chapter 2 - Tables 78
Chapter 3 - Cell Type-Specific Network Analysis in Diversity Outbred Mice Identifies Genes Potentially Responsible for Human Bone Mineral Density GWAS Associations. 80
3.1. Abstract 81
3.2. Introduction 82
3.3. Results 84
3.3.1. BMSC-OBs derived from DO mice yield diverse cell types that are enriched for mesenchymal lineage cells 84
3.3.2. Mesenchymal lineage cells are enriched in BMD heritability 87
3.3.3. Generating mesenchymal cell type specific Bayesian networks to inform BMD GWAS 87
3.3.4. Identifying putative drivers of mesenchymal cell differentiation 88
3.3.5. Identification of differentiation driver genes (DDG): 91
3.3.6. Network analysis predict Fgfrl1 and Tpx2 as novel regulators of BMD: 93
3.4. Discussion 94
3.5. Methods 99
3.5.1. Sample preparation and scRNA-seq 99
3.5.2. scRNA-seq analysis pipeline 99
3.5.3. Trajectory and tradeSeq Analysis 101
3.5.4. CELLECT Analysis 102
3.5.5. WGCNA 103
3.5.6. Bayesian network learning 103
3.5.7. DO eQTL mapping 104
3.5.8. Cell type proportion analysis 105

3.6. Acknowledgements **105**
3.7. Author Contributions **106**
3.8. Disclosures **106**
3.9. Chapter 3 - Main Figures **107**
Chapter 4 - Concluding Remarks and Future Directions **116**
4.1. Summary and Conclusions **117**
4.1.1. Assessing the utility of the BMSC-OB model 118
4.1.2. Leveraging scRNA-seq data from the BMSC-OB model to inform GWAS 119
4.2. Future Directions **119**
4.2.1. In vitro investigation of prioritized targets 119
4.2.2. RNAi-mediated knock-down 120
4.2.3. Prime-editing 122
4.3. Long-read, single-cell transcriptomics **124**
Appendix A - Supplementary Figures **127**
A. Chapter 2 - Supplementary Figures **128**
B. Chapter 3 - Supplementary Figures **133**
Appendix B - Supplementary Tables **136**
A. Chapter 2 - Supplementary Tables **137**
B. Chapter 3 - Supplementary Tables **137**
References **138**

List of Main Figures

Chapter 1

Figure 1. Expression Quantitative Trait Loci (eQTL) can affect the expression of genes in a cell type-specific fashion. 39

Figure 2. Splicing Quantitative Trait Loci (sQTL) can affect isoform-specific gene expression patterns. 40

Figure 3. Summary of preperation strategy for single cells using droplet-based scRNA-seq. 41

Figure 4. Summary of Unique Molecular Identifer (UMI) de-duplication. 42

Figure 5. Single cells captured via scRNA-seq are projected in a multi-dimentional, Principle Component (PC) space to highlight variability in gene expression. . 43

Figure 6. Overview of the BMSC-OB model. Bone marrow derived stromal cells (BMSCs) are extracted from the femurs of Diversity Outbred (DO) mice. 44

Chapter 2

Figure 1. ScRNA-seq of BMSC-OBs identifies multiple cell-types. 70

Figure 2. Liberation of single cells from a heavily mineralized matrix in vitro has minimal impact on transcriptomic signatures of BMSC-OBs. 71

Figure 3. ScRNA-seq of BMSC-OB and scRNA-seq data derived from cells harvested in vivo cluster together and have few transcriptomic differences. 72

Figure 4. Transcriptomic profiles of individual cell-types from scRNA-seq of BMSC-OBs are robust and representative of bulk RNA-seq data. 74

Figure 5. Cell-type frequencies captured by scRNA-seq are highly variable across individual DO mice. 75

Figure 6. SCENIC gene regulatory network (GRN) analysis reveals expected transcriptomic activity and validates the identities of cell-types in BMSC-OBs. 77

Chapter 3

Figure 1. Analysis of single cell RNA-seq (scRNA-seq) data for BMSC-OBs derived from 80 Diversity Outbred (DO) 108

Figure 2. Overview of the network analysis pipeline 109

Figure 3. Pseudotime Trajectory Inference analysis and establishment of cell type boundaries for tradeSeq analysis 111

Figure 4. TradeSeq-identified genes associated with BMSC-OB differentiation exhibit eQTL effects. 113

Figure 5. Fgfrl1 and Tpx2 are prioritized DDGs and putative drivers of mesenchymal differentiation. 115

Chapter 1

Introduction

1.1. An overview of the genetics of osteoporosis and bone mineral density (BMD)

Osteoporosis is often regarded as a multifactorial disease and commonly characterized by low bone mineral density (BMD), a measure of the amount of bone mineral in bone tissue, and an increased risk of fracture[1,2]. Millions of patients in the US have been clinically diagnosed with osteoporosis and billions are spent annually on fractures attributed to osteoporosis[3].

BMD remains one of the most significant clinical predictors of fracture and can be affected by a myriad of factors, such as age, sex, nutrition, and environmental exposures[1,4]; however, BMD is also a highly heritable trait and genetics is suggested to play a large role[5]. In terms of narrow sense heritability (h^2), which is a measurement used to quantify the variance of a phenotypic trait that may be due to genetic variation[6], the heritability of BMD is estimated to be $h^2 = 0.5 - 0.8$, further supporting the role of genetics in this complex trait[5]. Characterizing the genetic factors underlying BMD is a critical prerequisite to advancing the clinical treatment of osteoporosis.

Decades of work have cultivated a foundational understanding of some genetic determinants of osteoporosis, such as associations with polymorphisms in canonical genes (e.g., vitamin D receptor[7] or type I collagen[8]). We are now equipped with modern techniques, such as genome-wide association studies (GWASs), that can enable the resolution of millions of single nucleotide polymorphisms (SNPs) associated with disease. These genetic variants which are scattered across the genome can be associated with specific genes and variation in BMD, or other processes related to bone.

1.2. Characterizing the genetics of BMD via GWAS

GWAS remains one of the most powerful approaches to identify variants associated with a disease. Typically the output of these investigations is thousands of SNPs that exhibit statistically significant associations with the trait or disease of interest[9]. In the context of osteoporosis research, the largest GWAS to date analyzed estimated bone mineral density (eBMD) from the heel in approximately 420,000 individuals[10]. While BMD is typically measured clinically by dual-energy X-ray absorptiometry (DXA) from the femoral neck or lumbar spine, quantitative ultrasound is and often used alternative and measures eBMD on the calcaneus[11], as performed in the aforenoted GWAS study. While a number of other GWASs have been performed with the same goal of identifying genetic associations for BMD, the BMD GWAS from Morris and colleagues[10] remains the most comprehensive. From this study, a total of 1,103 independent associations were identified across 518 loci, of which 301 were novel, and together explain 20.3% of total eBMD variance.

While results from large GWAS studies have revolutionized our understandings of the genetics of a variety of human diseases, they are not without their shortcomings. A noteworthy challenge of GWAS studies is that many associations fall in close proximity to one another along the genome and tend to be in extensive linkage disequilibrium (LD) with one another. Thus, the identification of causal SNPs is convoluted by LD; numerous statistically significant associations with a disease-state or phenotype may not be causal, but act as surrogates for causal variants[12]. Additionally, functional characterization of these SNPs is complicated by the fact that the vast majority (>80%) of the statistically-significant associations reside in non-coding regions of the human genome[13]. For

example, causal variants may be located in DNA regulatory regions (e.g., enhancers or promotors) and affect the regulation of transcription and subsequent expression of a disease-relevant gene. Despite the mentioned challenges, GWASs have been successful at resolving hundreds of putative causal loci impacting BMD. Nevertheless, the specific genes affected by many of the identified associations remain unknown. Leveraging systems-based approaches and the integration of a variety of "-omics" level data have proven invaluable to facilitating the functional characterization of GWAS associations.

1.3. Systems genetics approaches utilizing transcriptomics data

In the past decade alone, the advancement of Next Generation Sequencing (NGS) technology, which often functions to sequence DNA at high-throughput in a short period of time[14], has enabled the generation of massive volumes of molecular data. The quantification of diverse biological features and deconvolution of cellular processes can be achieved with NGS technology, which has fueled the emergence of the field of systems biology. Ultimately, systems biology aims to make a molecular map or network, sometimes in order to resolve the functional role of perturbations to any layer of biology in a certain state, like a disease. In this field, data can take on many forms; broadly, it is referred to as "-omics" data, but it can be further classified based on what layer of biology is captured in the data. For example, data derived from genomic studies aims to characterize the relationships, functions, or interactions between genes that comprise a genome[15]; likewise, proteomics data aims to characterize protein isoforms, structures, or modifications in the cellular proteome[16].

Given the reliance on population-scale genomics data for GWAS analyses, the application of systems genetics approaches have been instrumental to predicting the molecular function of GWAS variants. The field of systems genetics, a derivative of systems biology, is concerned with characterizing sources of genetic variation that may perturb other molecular features (e.g., gene expression, protein, metabolite levels), especially in relation to a complex trait or disease, like BMD and osteoporosis[17]. The interdisciplinary nature of systems genetics requires the generation of various -omic data modalities, such as transcriptomics, one of the most powerful sources of -omic data.

Single-cell transcriptomics data aims to characterize RNA transcripts within a cell, typically to quantify changes in expression patterns under various states, such as a developmental stage, experimental, or physiological condition[18]. When used in unison, transcriptomics and GWAS data can be used to predict the consequences of genetic variation at certain loci; we can substantially improve our ability to resolve the connections between disease-associated variants and the genes they may regulate or functionally impact.

As shown in our recent work, bulk transcriptomics (RNA-seq) data and BMD GWAS data can be used in conjunction to discover multiple genes predicted to impact BMD[19]. Al-Barghouthi and colleagues generated RNA-seq data from mouse cortical bone samples and applied a network-based approach to highlight genes (and their associated networks) with strong evidence of impacting BMD[19]. To prioritize the identification of causal genes, they made use of human BMD GWAS from Morris and colleagues[10]; they required a gene to be implicated by GWAS and potentially serve as an eQTL (discussed below). Ultimately, they identified 66 genes (19 of which were novel or not reported

previously to be implicated in bone processes) likely to be causal for human BMD GWAS associations. While many studies have utilized transcriptomics data and GWAS to highlight causal genes, this work functioned to inform bone biology, a field currently lacking in molecular or "-omic" data, and thus remarkably contributed to the ongoing effort to generate and robustly analyze bone transcriptomics data. The remainder of this section will focus on transcriptomic analyses commonly used in systems genetics to predict the molecular effect of GWAS associations.

1.3.1. Expression Quantitative Trait Loci (eQTL) Analysis

Genetic variation has a significant influence on transcriptional regulation and many GWAS variants are presumably located in regulatory regions of the genome. Regulatory regions, or regulatory elements, are stretches of DNA that have roles in orchestrating many cellular processes, such as transcriptional regulation and chromatin organization[20]. Some of the best-known examples of regulatory regions of the genome are enhancers and promotors, and regulatory proteins are often recruited to specific motifs within these regions to enhance (or suppress) the transcription of target genes. Many GWAS-identified variants do not reside in the protein-coding regions of genes, but are located in regulatory regions of the genome and function to perturb expression of causal genes.

Genetic variation that is associated with the expression of genes is commonly referred to as expression quantitative trait loci (eQTL) while the associated gene is sometimes called an "eGene"[21]. In order to identify eQTL, transcriptomics data must be acquired, ideally, from a tissue or cell type relevant to the trait or disease in question.

Further, genotype information from these sample must also be obtained, either from a genotype array capturing a panel of SNPs or from whole-genome sequencing (WGS). Statistically significant SNPs associating with the expression of a gene are identified, which can exert their effects or operate in a cell type-specific manner (**Figure 1**).

There are two categories of eQTL: *cis*-eQTL, which operate in close genomic proximity (approximately 1 Mbp) to the associated eGene (e.g., in the promotor) (**Figure 1**), and distal or *trans*-eQTL, which operate much farther away, sometimes on completely different chromosomes, from the affected eGene (e.g., in a distal enhancer)[22]. In terms of eQTL discovery, historically, most studies have only been equipped to identify locally-acting *cis*-eQTLs due to statistical power difficulties in detecting *trans*-eQTLs[23]; however, much genetic variation is harbored in regions that induce *trans*-effects on target genes[24] and *trans*-eQTL are expected to regulate the expression of a wider range of genes[25].

To identify disease-relevant eQTL, a common approach is to perform a colocalization of disease-associated SNPs identified from GWAS. In this case, a colocalization analysis would function to identify eQTL and GWAS signal that originate from a shared genomic locus, which harbors potentially many causal variants. Therefore, not only can colocalizing sources of genetic variation begin to predict the function of uncharacterized GWAS variants, but also describe a potential causal mechanism of disease. For example, strongly colocalizing eQTL can be further fine-mapped to predict the precise SNP(s) that govern the expression of essential genes driving the manifestation of disease or a complex trait, like BMD.

An excellent resource that has been frequently leveraged in eQTL studies is the Genotype-Tissue Expression (GTEx) Consortium. The GTEx project was first launched in 2010 and is an ongoing effort to study the genetic effects on tissue-specific gene expression across a collection of human tissues[26]. To date, the project has provided open public access to data captured from over 54 non-diseased tissue sites from over 1000 post-mortem human donors; this data is derived from WGS and bulk transcriptomics (RNA-seq) assays performed on most samples[27]. However, one pitfall of this resource is that essential bone cell types (e.g., osteoblasts and osteocytes) are not represented in this data. While these cell types are most canonically associated with bone diseases and traits, such as osteoporosis and BMD, the GTEx project and others demonstrate that many eQTL are shared across tissues and various cell types[28]. Therefore, the utility of this resource can extend to research investigating bone. For example, our recently published work leveraged data provided by GTEx; it was used in unison with transcriptome-wide association study (TWAS) and eQTL colocalization in order to predict causal genes underlying BMD GWAS associations[29].

Leveraging these data, Al-Barghouthi and colleagues identified 512 putatively causal genes associated with BMD[29]. Notably, their approach functioned to make use of non-bone gene expression data (from GTEx) and currently available BMD GWAS data to prioritize candidates. First, a TWAS was performed using the BMD GWAS data from Morris and colleagues[10], as well as the gene expression reference available on GTEx; a TWAS functions to prioritize genes associated with a trait or disease, which are further associated with significant SNPs identified from a GWAS[30]. Next, an eQTL analysis was performed using 49 GTEx tissues; using the same BMD GWAS data, local eQTL were

identified for genes located within a GWAS locus via colocalization to determine if signals highlighted via both analyses are due to the same sets of genetic variants. The TWAS and eQTL analyses identified 2156 and 1182 genes, respectively. A total of 512 genes that were significant in both analyses were subsequently deemed prioritized causal genes with putative roles as regulators of BMD or bone-relevant biological processes (e.g., osteoblast differentiation, ossification, etc.)

1.3.2. Splicing Quantitative Trait Loci (sQTL) Analysis

Another mechanism by which genetic variation can impact gene regulation is via alternative splicing. The process of splicing is most canonically associated with the generation of multiple different messenger RNA (mRNA) isoforms from a transcript of an expressed gene, thus contributing to a wealth of diversity in terms of protein structure and function[31]. However, splicing can also operate on non-coding RNA species as well[32], further expanding the possible functional impacts tied to splicing patterns in a cell. Splicing is mediated by the spliceosome, which functions to recognize splice sites located at the 5' (donor site) or 3' (acceptor site) ends of introns located within a pre-processed RNA transcript[32,33]. The spliceosome subsequently removes (or retains) various combinations of introns to yield alternatively spliced transcripts. Aberrant splicing events can result in proteins with altered structural domains, functional features, or regulatory sites that can ultimately impact a cellular phenotype (**Figure 2**); alternatively, variation in splicing patterns can affect the abundance (or ratio) of specific protein isoforms in a cell (**Figure 2**). Importantly, the intricate process of alternative splicing can be altered by genetic variation.

Genetic variants associated with splicing events are referred to as splicing quantitative trait loci (sQTL) (**Figure 2**); they can impact splice site selection or the assembly of the spliceosome to result in dramatic changes in alternative splicing patterns[34]. Such variants can act to yield functional effects in *cis* or *trans*; *cis*-acting variants are located within the transcript undergoing splicing while *trans*-acting variants may result in the formation of a mutant core splicing protein or associated regulatory protein that functions to impact the process of splicing[34]. Therefore, genetic variation can affect the abundance, diversity, and ratios of various spliced RNA transcripts, which can lead to the manifestation of different molecular phenotypes and disease.

In order to identify sQTL, similar to eQTL discovery, transcriptomic and genotype data must be acquired on relevant samples. However, some challenges are incurred with sQTL studies (that are less of a hinderance in eQTL studies). For example, while commonly used and sufficient to quantify the expression of a gene, short-read RNA-seq data often does not enable a characterization of the complete repertoire of transcript isoforms that a gene can be spliced to yield. Transcriptomics data that is derived from short-read RNA-seq is comprised of short segments of sequenced RNA, which often do not contain the full-length sequence of RNA transcripts. Due to the significant degree of similarity that is often observed between alternatively spliced transcripts, mapping short-read data to a reference and associating a read with a specific isoform is statistically challenging[35,36]. Ultimately, this challenge inhibits the ability to adequately identify alternatively spliced RNA transcripts. However, recent advances in long-read RNA-seq technology now permit the capture and subsequent sequencing of

full-length RNA transcripts, therefore increasing our abilities to resolve novel transcript isoforms and better understand the nature of alternative splicing patterns.

As mentioned previously, colocalization approaches to combine GWAS and eQTL results have been successful in identifying disease-associated SNPs, thereby predicting the functional mechanism through which they can cause disease (i.e., modulating the expression of a gene). A colocalization approach can be employed to combine sQTL and GWAS to highlight a different mechanism through which genetic variation can cause disease (i.e., impacting alternative splicing of transcripts). Strongly colocalizing sQTL/GWAS signals are fine-mapped to further predict the impact that the genetic variation has on splicing, such as observed differences in causal protein isoform ratio, function, stability, and related protein-protein interactions. In our recent work, we leveraged long-read RNA-seq and identified colocalizing BMD GWAS and sQTL associations from GTEx-related tissues in order to identify effectors of BMD[37].

Abood and colleagues identified a total of 732 protein-coding genes with sQTL and associated with BMD[37]. First, a colocalization was performed using the BMD GWAS data from Morris and colleagues[10], as well as the sQTLs captured in GTEx, to yield the 732 "sGenes." Of these, over half (367 total) were sGenes with sQTL shared across more than one GTEx tissue. The novelty of this study was the use of long-read transcriptomics in bone-relevant cells types to interrogate predictions made from colocalization of sQTL and BMD GWAS associations. From an immortalized osteoblast cell line (hFOBs), long-read RNA-seq was performed to subsequently generated 22 million full-length RNA transcripts; 74% (50,588) of the transcripts were known (annotated previously on GENCODE) while the remaining 25% (17,375) were novel.

Using this data, they connected sQTL to 441 genes expressed in osteoblasts, further identifying putative genes impacted by splicing and associated with BMD.

1.3.3. Network Analyses using transcriptomics data

Polygenic diseases and complex traits, like osteoporosis and BMD, are a product of the cumulative effects of multiple loci that ultimately result in a phenotype[38]. In other words, we often expect to observe the consequences of genetic variation on more than one causal gene. Aside from identifying discrete sources of genetic variation that are predicted to drive disease, a comprehensive and systems-level understanding of disease can be achieved by characterizing how the interactions between genes (and related cellular processes) are affected by genetic variation. In order to infer interactions between genes, a network approach is often employed to resolve putative relationships, which is an essential strategy leveraged in the field of systems genetics. Many various types of networks can be resolved depending on the molecular data available. For example, network-level analyses, such as gene co-expression networks (GCNs), commonly take transcriptomics data as input to predict interactions between genes based on shared expression profiles. Many studies, including much of our recently published work, have demonstrated the utility of GCNs and a network-based strategy to predict the interactions of causal genes implicated by GWAS.

1.3.3.1 Gene co-expression network (GCN) analyses

The purpose of a GCN analysis is to group genes based on correlation in their expression, thus describing pairwise relationships between genes. As mentioned above,

transcriptomics data from multiple samples is used as input, providing a quantitative metric for each gene's expression. The classic output of a GCN analysis are often visualized as web-like networks where each gene is referred to as a "node" and the connections between each node are referred to as "edges." An edge represents a correlative connection between nodes, but in GCNs, edges are not assigned a directionality to further describe the interaction between connecting nodes, thus rendering GCNs undirected[39].

One approach commonly leveraged to generate such GCNs is weighted gene co-expression network analysis (WGCNA)[40]. The goal of WGCNA is to group genes based on co-expression and then highlight clusters of densely connected genes, referred to as modules. Biologically insightful information can be gleaned from investigating the specific genes contained in various modules. As a general trend, co-expressed genes that comprise a module can be functionally related or play a role in similar cellular pathways or processes[41]. Further, WGCNA analysis can be extended to discover correlations between specific co-expression modules and a particular trait, such as a quantitative metric taken on a sample[42]. GCNs can also be used to contextualize causal genes implicated by GWAS associations; we frequently generate GCNs from bulk transcriptomics (RNA-seq) data from bone cells gathered from large populations of mice to inform BMD GWAS[19,43,44].

For example, Calabrese and colleagues generated bulk transcriptomics data from cortical bone from a large panel of mice (n = 96) and subsequently generated GCNs[43]. First, they defined a list of 167 genes implicated by 64 lead SNP associations from a large BMD GWAS. Subsequent GCN analysis identified two modules of co-expressed

genes that contained many GWAS-implicated genes. A high degree of correlation was observed between the eigengenes for the modules (r = 0.63); functionally, an eigengene serves as a summary of the gene expression for a module, or the first principal component for the gene expression profile. A gene ontology (GO) enrichment analysis was performed and indicated that the modules were significantly enriched for genes associated with processes relevant to osteoblasts, such as ossification and osteoblast differentiation, thus portraying the shared functional roles of genes in GCNs.

Additionally, between the modules, many genes have previously reported evidence of impacting human BMD, such as sclerostin (*Sost*), osterix (*Sp7*; transcription factor Sp7), and osteoprotegerin (*Tnfrsf11b*; tumor necrosis factor receptor superfamily member 11b), all of which have canonical regulatory roles in osteogenic cells. Overall, they were able to associate genes to 30 of the 64 BMD GWAS loci, identifying two genes to investigate further: Spectrin, beta, nonerythrocytic 1 (*Sptbn1*) and MAP/microtubule affinity-regulating kinase 3 (*Mark3*), both of which had strong BMD GWAS associations. Interestingly, the International Mouse Phenotype Consortium (IMPC)[45], which is an organization that functions to systematically study the phenotypic effects of knocking-out various genes in the mouse genome, reported that heterozygous mice for the gene-trap *Sptbn1* allele exhibited increased whole-body BMD, while the opposite was observed for female mice, which had a decrease in whole-body BMD. In regards to *Mark3*, when knocked-down via siRNA in calvarial osteoblasts, cells exhibited an increase in mineralization *in vitro*. Further, heterozygous mice for the gene-trap *Mark3* allele exhibited increased femoral BMD. These data suggest a putative role of *Sptbn1* and *Mark3* in BMD.

In another one of our more recent works, Sabik and colleagues generated GCNs on calvarial osteoblasts using bulk RNA-seq data from a large panel of mice (n = 42)[44]. They also identified a module that contained many genes tied to osteoblast differentiation and mineralization, as well as genes implicated by BMD GWAS associations identified from Morris and colleagues[10]. Importantly, the eigengene of this module was the only one (of 13 total modules) significantly correlated (r = 0.49) with *in vitro* mineralization, a quantitative trait measured by the accrual of alizarin red-stained calcified nodules of cultured osteoblasts from the mice.

Furthermore, they found that genes within this module fall into two distinct categories based on their gene expression patterns observed over the course of *in vitro* osteoblast differentiation: genes that exhibit high expression *early* during the process of osteoblast differentiation (EDS; n = 175 genes) and genes that exhibit higher expression *later* during differentiation (LDS; n = 323 genes). LDS genes were enriched for GWAS-implicated genes; additionally, LDS genes had a significantly higher module membership (which is the correlation between a gene's expression and the eigengene for the module) as compared to EDS genes, thus indicating a more central role of LDS genes in the module. Of the LDS genes, 48 overlapped with a BMD GWAS association, of which 12 genes had a BMD GWAS association that also colocalized with an eQTL in a GTEx tissue. Of these 12 genes, 4 were measured via the IMPC and had significant ($P_{adj} < 0.05$) whole-body alteration to BMD: cell adhesion molecule 1 (*Cadm1)*, beta-1,4-N-acetyl-galactosaminyl transferase 3 (*B4galnt3)*, dedicator of cytokinesis 9 (*Dock9)*, and adhesion G protein-coupled receptor D1 (*Adgrd1*, or *Gpr133)*. These data suggest a potential casual role for these genes as drivers of BMD in humans.

Both of the aforementioned studies showcase the utility of GCN analysis to aid in the prioritization of candidate causal genes implicated by GWAS. However, because GCNs are undirected networks and generated solely based on correlations in the expression patterns amongst genes, a limitation to GCNs is that they only reveal which genes are active simultaneously and provide minimal information about causal interactions between genes[46]. Gene regulatory networks (GRNs), on the other hand, function to portray a network with interactions describing a regulatory relationship between nodes, ideally to describe a transcriptional program[47]. The generation of GRNs also requires additional data that can describe specific aspects of the regulatory nature of the network (e.g., transcription factor binding interactions); further, the application of a detailed mathematical function is needed to model regulatory relationships[48]. Nevertheless, gene co-expression networks (GCN) analysis remains a valid network-approach to organizing genes and aids in forming initial predictions of disease-relevant genes. To glean more biologically insightful information from GCNs, additional methods can be applied, such as Bayesian network reconstructions, to model directed interactions from these correlation-based networks.

1.3.3.2 Bayesian network analyses

The overarching goal of performing network analyses is to learn the structure of a network and subsequently infer relationships between the constituents of the network. As described above, GCNs are constructed based on genes' co-expression, but are limited in terms of further describing the interactions between genes in the networks. To ascribe

more meaning to networks, Bayesian network learning methods can be applied to predict causal interactions, for example, amongst co-expressed genes of a GCN.

Bayesian networks (BN) graphically depict real dependencies between measured variables[49], or the expression of genes in our case. The connections of a BN form to represent interactions between nodes (genes) and indicate that a gene's expression depends on the expression of the gene(s) upstream of it in the network, which are referred to as parent nodes[50,51]. In a BN, the edges between the nodes are directed, often represented as arrows. The directed edges of the network indicate the conditionally dependent relationships between the nodes; furthermore, for each node, a probability is calculated that functions to describe the relationship it has with another node of the network[50,51]. To summarize all relationships between nodes within the network, joint probabilities are determined by taking the geometric sum of all probabilities for individual nodes[50]. To generate a network topology that best describes the gene expression data, methods such as the Max-Min Hill-Climbing (MMHC) algorithm[52] are applied to maximize a score function and learn the optimal structure of a BN[49].

BNs are acyclic, requiring that no cyclic paths or "loops" are formed in the global structure of the network[49]. Thus, BNs are often classified as a type of directed acyclic graph (DAG)[49]. Dynamic Bayesian networks (DBNs), on the other hand, can be applied to understand more "cyclic" biological phenomena, such as feedback loops. DBNs take into account temporal or time-dependent relationships that are not modeled in acyclic BNs described above[53]. Given the inclusion of data taken on a variable at multiple time points, perhaps gene expression measurements taken over a time course experiment, DBNs can model temporal interactions in a network that change over time[53]. In practice,

DBNs may resolve relationships in which the expression of a gene at a certain time point may be dependent on the expression of various genes in a prior timepoint[53]. For example, DBNs can be employed to map changes in the expression patterns of genes and their associated partner genes to resolve feedback loops underlying potential transcriptional landscapes driving a phenotype[54]. Nevertheless, BNs are frequently used in systems genetics and are suffice in resolving meaningful, static interactions and infer causal relationships amongst genes. As we showcase in our recent work[19], we leveraged BNs generated from WGCNA modules to highlight networks and genes potentially responsible for BMD GWAS associations.

Al-Barghouthi and colleagues generated BNs from bulk transcriptomics data (RNA-seq) from cortical bone samples derived from a large cohort of Diversity Outbred (DO) mice (n = 192)[19]. A total of 142 WGCNA modules of co-expressed genes were used as input and subsequently reconstructed to generate BNs, which was necessary to resolve directionality between genes in the networks and model causal interactions. The structures of the BNs were learned using the MMHC algorithm. Networks were then prioritized based on having more neighbors (genes) in the BN than average as well as having a significant enrichment of genes with previously established roles in various processes related to bone; this yielded a total of 1370 “key driver” networks. Key drivers (KD) are nodes (genes) that are central to a network topology and enable the formation of a large neighborhood of genes that constitute a BN. Of the 1370 KDs, 1174 had a corresponding homolog in humans. Additionally, 688 of the KDs were genes that were within 1 Mbp of a lead SNP identified from BMD GWAS. From their Bayesian network analysis, they highlight two putative novel regulators of BMD: SERTA domain-

containing protein 4 (*Sertad4*) and Glycosyltransferase 8 domain containing 2 (*Glt8d2)*, both of which were identified in their network analysis as KDs of BNs containing many canonical genes involved with bone cell processes (e.g., *Postn*, *Wnt16,* and *Pappa2*). By leveraging BNs, they were able to model causal interactions between co-expressed, and potentially co-regulated, genes to ultimately identify novel key drivers of the regulation of BMD.

1.4. Single-cell transcriptomics

Pioneering breakthroughs in single-cell transcriptomics (e.g., scRNA-seq) have revolutionized our abilities to characterize gene expression. In the past decade alone, scRNA-seq technology has exploded in popularity and its applications to biomedical research are seemingly endless; we now can delineate the dynamic patterns of cell type and tissue-specific gene expression across a variety of conditions in order to generate experimental hypotheses or discover novel transcriptomic markers for cells.

Currently, there is a lack of -omics data at single-cell resolution in the field of bone and osteoporosis research. To date, much of the transcriptomics data generated on bone has been microarray-based or from bulk RNA-seq. In fact, all of our work has leveraged RNA-seq data in various analytical frameworks to showcase its utility to informing BMD GWAS; however, the cell type context in which many of these GWAS associations are operational is unclear. In regards to informing GWAS, the integration of scRNA-seq data can provide cell type predictions for which observed genetic variation may have a functional impact. Given that numerous cell types are vital to bone function (e.g., osteoblasts, osteocytes, MSCs, adipocytes, etc.), capturing their transcriptomic

profiles at single-cell resolution, particularly at population-scale, is essential to better characterizing their roles and understanding how alterations to their function may be associated with a phenotype or disease.

Single-cell transcriptomics (scRNA-seq) now enables the elucidation of complex biology at much higher resolutions; however, bulk RNA-seq has historically served as a staple to investigating the transcriptome of cells associated with many disease states. While both serve very specific purposes, their differences are also apparent.

1.4.1. Summary of bulk transcriptomics (RNA-seq)

Bulk RNA-seq studies capture and subsequently sequence RNA transcripts. RNA-seq effectively provides a global "average" summary of gene expression for biological samples[55]; however, an often harped on limitation of using bulk data is that it does not enable the association of transcriptomic signal to individual cells for a sample. While methodological interventions can be performed prior to bulk sequencing in order to isolate a homogenous population of a discrete cell type (e.g., flow cytometry)[56], this is not a feasible approach (from a labor and fiscal perspective) for high-throughput sample processing required for population-scale studies.

Additionally, when RNA-seq is performed on samples from a heterogenous tissue comprised of many different cell types (e.g., peripheral blood), the resulting gene expression signatures are often challenging to associate with a specific cell type. However, deconvolution of bulk transcriptomics data can be achieved bioinformatically to estimate cell type abundance and infer expression profiles using computational tools

(e.g., CIBERSORTx)[57], but a heavy reliance on reference data (derived from scRNA-seq or bulk-sorted data) as well as larger sample sizes can impact the accuracy of results[58].

Despite these drawbacks, performing bulk transcriptomics enables the capture of not only polyadenylated mRNA species, but also other diverse RNA species as well (depending on the selection/depletion method used), such as non-coding RNA[59,60]. Estimates suggest that the majority of the genome can be transcribed, but less than 2% encodes proteins, leaving the remaining to presumably encode non-coding RNA[61]; thousands of long non-coding RNAs (lncRNA) have been documented[62,63]. The implications of non-coding RNA in disease are now at the forefront of research in the pharmaceutical industry[64]; therefore, bulk RNA-seq technologies will likely remain the most feasible approach to characterizing the roles of diverse RNA species. Further, bulk RNA-seq experiments are becoming more economical to perform given the appreciable decrease in sequencing costs in recent years[65]. However, leveraging the most current transcriptomics technology can enable a more granular understanding of the biological processes impacted by disease at cell type-specific resolution. The remainder of this section will focus on some important concepts underlying single-cell transcriptomics, including analysis and current strategies used to investigate cellular heterogeneity captured in scRNA-seq data.

1.4.2. Single-cell and RNA capture

The appeal to performing single cell transcriptomics (scRNA-seq) is the ability to associate gene expression signatures to individual cells. Upon the disassociation of cells from tissue, droplet-based strategies (e.g. 10X Genomics), which are among the more

popular options for sample preparation[66], partition individual cells into nanoliter-scale, oil-based droplets[67]. These droplets, called GEMs (Gel Bead-in-Emulsion), are generated using microfluidic technologies that encapsulate a single cell, along with the necessary enzyme and reagents required for a reverse transcription reaction, and importantly, a gel bead carrying thousands of oligonucleotides which are essential to the functionality of scRNA-seq (**Figure 3**)[68].

While slight variations exist in the sequence of the oligonucleotide of the gel bead, for the commonly used 3' 10X Genomics protocols, the oligonucleotide structure is comprised of a primer sequence, a barcode sequence, a unique molecular identifier (UMI) sequence, and an oligo-dT sequence (**Figure 3**)[68]. The oligo-dT sequence serves the essential purpose of annealing to the 3' polyadenylated tails of RNA transcripts to prime reverse transcription. The UMI and barcode sequences serve to index the captured RNA transcripts and associate them with a specific single cell, respectively, while the primer sequence enables subsequent polymerase-chain reaction (PCR) amplification for library construction.

The overall structure of all oligonucleotides is the same for all gel beads in the employed protocol; however, each gel bead has a distinct barcode sequence. During the preparation of GEM droplets, every individual cell is captured with a single gel bead, and thus a distinct barcode sequence is associated with the cell: a critical feature for downstream analysis that enables the discrimination between the sequencing products derived from individual cells.

While barcode sequences function to index individual cells, the UMI sequences function to index the captured RNA transcripts from the cell. Unlike the barcode

sequences, the nucleotide sequence of a UMI is variable[69]. In other words, for all oligonucleotides of a gel bead, the barcode sequences are identical, but the UMI sequences are different. The UMI sequence is valuable for improving transcript quantification, which can be skewed by nonlinear amplification during library construction PCR[70]; cDNA (generated from RNA transcripts) are copied via PCR and result in the accumulation of amplified progeny from original transcripts, which are referred to as PCR duplicates[69]. Exponential amplification of PCR duplicates can sometimes lead to their over representation and result in skewed gene expression quantification[70]. However, UMIs remedy this issue by enabling the "de-duplication," or collapsing, of PCR duplicates, thus improving the ability to distinguish between the transcripts and their duplicates (**Figure 4**)[69,70]. Subsequent transcript quantification occurs by counting the collapsed UMIs for a gene, ultimately yielding a raw count to be used as expression measurements for all originally captured transcripts for any given gene (**Figure 4**).

1.4.3. Dimensionality reduction and clustering

An attribute of scRNA-seq data is its high dimensionality as a result of capturing the expression of thousands of genes for thousands of individual cells (**Figure 5**). As the complexity of the data increases, measures must be taken to reduce this dimensionality to make data analysis practical, but also ensure as much meaningful information is retained as reasonably possible. Typically, more than 20,000 genes, or "features," can be captured in scRNA-seq data; however, not all of the genes will be biologically informative to highlighting meaningful differences between cells or cell clusters[66,71]. Identifying the

most highly variable genes (HVG), or those genes exhibiting higher expression in some cells and lower expression in other cells[72], functions to prioritize these genes as informative sources of heterogeneity to investigate[73]. HVGs are selected by identifying those with the largest standardized variance, typically as determined via approaches such as variance-stabilizing transformation (VST)[74]. By focusing on a subset of HVG (usually between 1000 - 5000 genes), the dimensionality of the data is reduced; nevertheless, even after feature selection, it remains quite high and additional measures must be taken for biological insight to be extrapolated[73].

While dozens of techniques can be implemented to lower the dimensionality of scRNA-seq data[75], principal component analysis (PCA) is often employed to aid in summarizing the data[76]. In practice, PCA is a orthogonal linear dimensionality technique used to project gene expression for all single cells in a multidimensional space[77]. Multiple principal components (PCs) are calculated to summarize the variability observed in the gene expression data, typically on the HVGs. Genes whose expression exhibit the largest variation between cells will have the most impact on the resulting PCs. As a general trend, the first and second PCs capture the most variation observed in the data. Additionally, feature (gene) loadings are calculated, which are numeric values assigned to each gene that describes the influence or contribution its expression has in any given PC. These feature loadings are subsequently used to acquire cell embeddings, which represent the placement of each single cell in a lower dimensional PC space (**Figure 5**). The number of PCs to include during data analysis corresponds to how segregated the cells will be in PC space; however, as the vast majority of the variability is retained in the first 15-30 PCs, the inclusion of dozens of PCs beyond this range does not necessarily

enhance the capture of relevant information and would likely capture technical noise rather than meaningful biological heterogeneity.

One of the main goals of performing scRNA-seq is to capture cellular heterogeneity with an endpoint of generating "clusters" comprised of multiple individual cells that have a similar gene expression profile. Collectively, clusters can represent a distinct cell type or cell state (e.g., an intermediary cell state between precursor cells and differentiated cells). In the lower dimensional, PC space, single cells group together (or separate apart) based on their gene expression profiles (**Figure 5**). To define these groups, a kNN (k-nearest neighbors) algorithm is applied to generate a network in which each cell is connected to a specified number of "nearest neighbors" (other cells) in a specified number of dimensions (e.g., k = 20 neighbors in 15 PCs). The algorithm first calculates the Euclidean distances between cells in a PC space, then for any given cell, the shortest (nearest) distances to other cells (neighbors) are identified.

Using the kNN network, a shared nearest neighbors (SNN) network is subsequently generated. In the SNN, for every cell, the connections to its neighbors can be assigned a weight - the Jaccard similarity coefficient. Its calculation considers the number of connections a cell, and one of its neighbors, has in common[72,78]. Next, community detection algorithms, such as the Louvain algorithm, are applied to identify communities of single cells using the SNN[78,79]. In an iterative process, the Louvain algorithm functions to maximize the modularity of the identified communities of cells; it optimizes the best possible connections between communities and aggregates communities that are closely connected to form "clusters"[78].

After the identification of cell clusters, visualizing the clusters in a meaningful way requires the use of another dimensionality reduction technique. PCA is classically defined as being a linear dimensionality reduction technique; it serves its purpose of lowering the dimensionality and summarizing the variance of gene expression across cells captured in the scRNA-seq data. However, PCA can falls short in effectively reducing the dimensionality into as few as two components, which is a prerequisite for endpoint interpretation and visualization of the scRNA-seq data[71]. To generate a visual representation of the identified cell clusters, accurate portrayal of the distinct spatial distributions between them must be achieved after reducing the dimensionality to such low levels (e.g., two dimensions)[71]. In order to adequately reproduce the global structure of scRNA-seq data, nonlinear dimensionality reduction methods are commonly used, such as t-SNE (t-distributed stochastic neighbor embedding)[80] and UMAP (Uniform Manifold Approximation and Projection)[81]. Both methods function to project the single cells in the lowest dimensional space, but differ in how they mathematically represent the distances between cell clusters[82]. A contentious debate still remains in the single-cell field regarding which method performs better. While the use of either is accepted, some suggest that UMAP preserves the global structure of the data better than tSNE[83], while others have refuted this notion and suggest that they both perform (and underperform) equally[84]. Nevertheless, nonlinear dimensionality reduction approaches ultimately yield essential output of scRNA-seq analysis: typically, 2D graphical representations of the processed scRNA-seq that captures heterogeneity, and ideally, distinct clusters of single cells. Subsequent differential gene expression analysis of the resulting clusters can assist in assigning cell type annotations.

Marker gene identification for the clusters is often achieved by quantifying the expression of each gene in all cells of a cluster and subsequently comparing expression across all clusters in a "global" analysis of gene expression. Comparison is typically performed using the log fold change of the average expression of genes for a cluster. A "pseudobulk" approach can used to aggregate (sum) the expression of each gene across all cells of a cluster for one sample. Pseudobulk approaches are employed in an attempt to more accurately represent transcriptomic variance observed between individual biological samples[85,86]. Sample-specific, pseudobulk profiles can subsequently be used for differential gene expression (DEG) analysis between cell types across samples[86]. Nevertheless, the identification of various cell types or cell states is achievable by comparing marker genes to current literature or other scRNA-seq annotation databases.

1.4.4. Pseudotime Trajectory Analysis

The identification of marker gene expression for clusters is easily achievable and is often suffice as an initial pass of characterizing the heterogeneity of cells captured in scRNA-seq data. However, typical approaches function to identify marker genes in what can be classified as a "static" analysis of gene expression and do not adequately capture the dynamic changes in expression underlying biological processes.

Gene expression can be viewed as a continuum and temporal changes in gene expression can be quantified, for example, across a path of cellular differentiation in which an undifferentiated cell (e.g., stem cells or progenitor cells) transitions from one cellular state to ultimately become a fully differentiated cell type. Paths of differentiation can be mapped in scRNA-seq data, depending on the degree of cellular heterogeneity of

the captured cells (i.e., not a homogenous population of identical cells). Numerous scRNA-seq tools function to infer cellular trajectories, or "lineages," which can be used to characterize paths of differentiation. The most popular tools operate in a reduced dimensionality space containing single cells (e.g., PCA) and then infer trajectories (e.g., Slingshot[87], Monocle[88]); however, other tools operate in a gene-space (e.g., CSHMM[89]) or analyzing RNA transcript splicing patterns (e.g., RNA velocity via scVelo[90])[91].

In the context of trajectory inference in a reduced dimensionality space, trajectories are most commonly constructed using a minimal spanning tree (MST) approach[91]. MST is an algorithm that learns the most efficient network (tree) of contiguous connections of all nodes such that the sum of all distance-weighted edges (e.g., Euclidean distance) is minimized[92]. For example, Slingshot begins to infer trajectories in a multi-dimensional space by treating cell clusters as nodes, or "centroids," which are subsequently used as input to generate a MST[87,91]. Next, simultaneous principal curves are fit to summarize the MST and individual cells are orthogonally projected to them; importantly, a curve can bifurcate to indicate cells diverging from a lineage[87].

From these curves, "pseudotime" values are subsequently obtained for each cell, which are singular values assigned to each cell and describes both its transcriptional progression and placement along its associated trajectory[87]. When multiple lineages are inferred, cells are assigned to a particular lineage based on weights, which are derived from their projected distance to any given trajectory[87]. The output of the trajectory analysis is a pseudo-temporal ordering of single cells along a trajectory. Additionally, trajectories can be graphically represented as smooth curves which subsequently can be

overlayed onto the single cells in a lower dimensional space (e.g., UMAP) in order to visualize their progression to various terminal cell fates.

Trajectory inference enables the analysis of temporal gene expression along a resulting trajectory. Differential gene expression can be assessed across the entire breadth of a single trajectory, or between specific points, which can be defined by pseudotime. Additionally, more than one trajectory can be compared to identify differing transcriptional patterns underlying the cells traversing along them. TradeSeq[93], for example, is a tool that relies on the pseudo-temporal ordering of cells on a lineage to interpret the continuous nature of gene expression.

Ideally, a linear relationship would be observed between pseudotime and the expression of a gene; however, this is not always true and nonlinearity must be accounted for[93–95]. To overcome this challenge, generalized additive models (GAM) are often implemented. GAMs estimate nonlinear relationships between predictor (independent) variables and the response (dependent) variable via smoothing functions that attempt to summarize the data using curves[96]. For example, tradeSeq employs a GAM to model gene expression as a function of non-linear pseudotime. Smooth functions are inferred for gene expression along pseudotime for a trajectory. These smoothers provide the foundation for which temporal gene expression analysis can be performed. For example, pseudotime boundaries can be established to define a particular region of interest along an inferred smoother (for a specific trajectory), then gene expression from cells mapping to the start and end of the pseudotime boundary can be assessed for dynamic changes in expression. Thus, pseudotime trajectory analysis can effectively facilitate temporal gene expression investigations.

1.4.5. Limitations

While scRNA-seq may seem like the all-cure for high-throughput and high-resolution transcriptomic studies, it will not single-handedly replace bulk transcriptomics at this point in time and is not without its limitations. A pitfall of scRNA-seq is the "zero-inflated" nature of the data that yields a sparser matrix with zero values for the expression of some genes in a proportion of cells, which are colloquially known as "drop-outs"[97]. The reason for the occurrence of a drop-out in the resulting data is commonly attributed to technical reasons, such as low sequencing depth, inefficient capture, or poor amplification of a transcript[98,99]. Alternatively, some drop-outs may indicate the true absence of the transcripts, thus capturing valid biological signal[98,99]. As a result, zero-inflated scRNA-seq can impact downstream analysis of the data, such as normalization.[73,97] Numerous imputation methods have been developed to correct drop-outs by estimating their expression values in an attempt to improve subsequent data analysis; however, the accuracy of imputation methods and whether or not the resulting improvement in transcriptomic signal recovery is enough to warrant such intervention are under scrutiny[99,100].

Additionally, the most commonly used scRNA-seq methods are often confined to capturing only polyadenylated RNA species, namely protein-coding mRNA and long noncoding RNAs (lncRNA) that are polyadenylated[101]. Therefore, much of the remaining RNA diversity of a cell (e.g., microRNA (miRNA), small-interfering RNA (siRNA)) is not captured via conventional scRNA-seq data. Another limitation is that short-read sequencers (e.g., Illumina-based sequencers) are commonly used to generate data for

scRNA-seq studies. As such, cDNA fragmentation is a procedural requirement for samples to be eligible for sequencing, thus resulting in a library composed of small fragments of RNA transcripts for sequencing (300-600 bp in length)[102,103] rather than full length RNA transcripts. Furthermore, only the 3' (or 5', depending on the kit used) of the fragments are amplified and subsequently sequenced[102], thus limiting our abilities to identify all spliced isoforms of a given RNA transcript. Together, these procedural constraints and limitations of current scRNA-seq technology impede a comprehensive analysis of the entire cellular transcriptome in one pass; however, great strides are being made to overcome these hurdles, such as long-read scRNA-seq. Nevertheless, the ability to resolve even a fraction of the transcriptome of a single cell (in a throughput fashion) has revolutionized biomedical research.

1.5. Using the Diversity Outbred mouse population as a model to study bone

One of the motivating goals of biomedical research is to have the work be applicable to humans and ideally, translational such that it may lead to clinical interventions that improve human health (e.g., vaccines or novel therapeutics)[104]. However, performing studies exclusively in human subjects is often not feasible; therefore, the establishment of mice as an animal model for human biomedical research began decades ago to provide researchers with the means to experimentally investigate biological phenomena related to human diseases[105,106]. In the context of bone biology research, mice have been used extensively as a model organism to study various aspects of skeletal development, bone cell function, remodeling, fracture healing, and the manifestation of osteoporosis[107]. Moreover, substantial similarities exist between human

and murine skeleton[108], making it a logical choice of model organism for research investigating bone-related phenotypes and diseases.

Genetics studies are popular in a murine model due to the feasibility of genetic manipulations required to generate engineered mouse strains, along with other practical elements that enable these studies, such as the shorter life cycle, more control over breeding and subsequent genetic composition of progeny, and lower cost of animal maintenance[109–111]. For some genetic studies, inbred mouse strains are employed, namely for the purposes of generating a reproducible and nearly identical population of mice[112], which is pertinent for certain experiments where genetic variation must be controlled[113]. Conversely, outbred populations are generated for the very purpose of retaining genetic variation across individual mice and may be more applicable to modeling genetic diversity representative of a human population[113]. The Diversity Outbred (DO) mouse population, for example, was developed as a genetically diverse stock of mice to enable studies to characterize the genetic basis of complex human diseases[114,115].

DO mice were first generated in 2009 from randomized breeding of mice from eight founder mice strains: A/J, CAST/EiJ, C57BL/6J, NOD/ShiLtJ, NZO/HILtJ, PWK/PhJ, WSB/EiJ, and 129S1/SvImJ[112]. Over the course of dozens of generations, individual DO mice have genetic backgrounds that exhibit a diverse assortment of allelic combinations derived from the founders[115]. Quantitative trait measurements and other phenotypes in the DO can be associated to specific regions of genetic variation across individual mice, which are typically the output of QTL analyses. A prerequisite to such studies is genotyping of the mice, which is feasibly achieved with high-density SNP arrays, such as the Mouse Universal Genotyping Array (MUGA)[116]; the GigaMUGA panel employs a

total of 143,259 markers spanning all mouse chromosomes to sufficiently discriminate between the DO founder strains[116]. These SNP arrays are subsequently used to reconstruct the haplotype of each mouse; for any given interrogated genomic locus, there are 36 genotype possibilities[117]. QTL analyses in the DO have been successful in identifying loci associated with complex and disease-related traits,[112] including those associated with bone. In our recent work[19], we utilized the DO mouse population to highlight specific haplotype backgrounds that are associated with quantifiable bone strength-related traits, as well as the expression of specific genes.

In a large cohort of DO mice (N = 619), Al-Barghouthi and colleagues, captured 55 skeletal phenotypes commonly used to evaluate bone, ranging from histomorphometry to microarchitecture and mechanics assessments[19]. Additionally, bulk transcriptomics (RNA-seq) data was collected from the femoral diaphyseal bone for a fraction of the cohort (N = 192). Using these data, they identified 28 significant QTLs for 20 of the traits, some of which reside in loci that were associated with more than one trait. In particular, one locus on chromosome 1 was associated with traits such as cross-sectional size (e.g., medial–lateral femoral width), cortical tissue mineral density, and cortical porosity. At this locus, QTL effects for one DO founder background, WSB/EiJ, was identified and potentially responsible for influencing the aforementioned traits. Upon performing an eQTL analysis, they identified 18 genes with eQTL that colocalized with six of the loci from the trait-QTL analysis; the locus on chromosome 1 was one of these high-priority loci. Candidate genes mapping to this locus included Immediate Early Response 5 (*Ier5*) and Quiescin sulfhydryl oxidase 1 (*Qsox1*), both of which indicated strong negative eQTL effects in mice with a WSB/EiJ background at this locus. Results

from this study portray the utility of the DO as a mouse population in the discovery of the genetic underpinnings of complex traits, particularly those implicated in bone.

1.6. BMSCs and *in vitro* osteogenic differentiation

One of the essential functions of most skeletal bone is to house and protect the bone marrow cavity, which contains the microenvironments where resident stem cells give rise to a diverse array of cell types critical to human function. In the marrow, two populations of stem cells exist: hematopoietic stem cells (HSCs) and mesenchymal stem cells (MSCs), both of which have the capacity for self-renewal and multipotency[118–121], or the ability to subsequently differentiate into lineage-specific cell types. Because the use of the phrase "MSC" can be considered contentious in the field, as the characterization and precise definition of these cells is currently under scrutiny[118], BMSCs (bone marrow-derived stromal cells) will be the nomenclature used henceforth to refer those non-hematopoietic multipotent cells derived exclusively from the bone marrow[122]. HSCs can differentiate into most cells canonically associated with the immune system, like myeloid (e.g., monocytes, granulocytes) and lymphoid (e.g., B-lymphocytes, T-lymphocytes) lineage cells. Conversely, BMSCs can differentiate into cells belonging to a variety of lineages that are responsible for the formation of bone, fat, and cartilage tissues. Hematopoietic lineage cells have important roles in bone biology and disease, such as osteoclasts (derived from the myeloid lineage) that are responsible for the reabsorption of bone[123,124], while the mesenchymal lineage cells give rise to multiple other bone-relevant cell types, such as osteoblasts, which are responsible for synthesizing new bone matrix[123];

osteocytes, that embedded in bone and orchestrate bone maintenance[123]; chondrocytes, which generate cartilage[125]; and adipocytes, that store lipids[126].

Adipocytes account for a large proportion of the total bone marrow volume in adult humans (up to 70% depending on age)[127,128], which is similarly observed in mice[128]; however, multipotent cells are far less abundant. For example, BMSCs are roughly estimated to comprise 0.01-0.1% of cells in bone marrow in humans[120]. Despite their low abundance in the marrow, a tremendous amount of research has been done to investigate the localization of BMSCs in the various compartments or niches of the bone marrow cavity[121,129]. In fact, they are an essential component of the perisinusoidal niches necessary for HSC maintenance in the marrow[121], thus portraying another critical function of MSCs.

As mentioned previously, BMSCs are capable of differentiating into both osteogenic (e.g., osteoblasts, osteocytes) and adipogenic (e.g. adipocytes) cells, and the commitment of BMSCs down either lineage has been implicated in osteoporosis and alterations in bone remodeling[130]. Further, the relationship and homeostatic balance of osteogenic and adipogenic cells have a clear connection to disease. For example, marrow adipose tissue (MAT) content has a well-defined inverse correlation with BMD[131]. While this relationship is observed in healthy patients (independent of age, sex, and bodily fat), in osteoporotic individuals, marrow adiposity is noticeably higher[132]. Thus, the aforenoted cell types, among the others derived from BMSCs, are essential to understand as they relate to the manifestation of various bone-related diseases, like osteoporosis.

In the context of clinical and biomedical research, to overcome the challenge of the scarcity of BMSCs in raw marrow, several methods have been established to enrich

for sufficient quantities of them[133]. Among such methods include the strategy of exploiting the unique adherent property of BMSCs as the basis for their selection *in vitro* (**Figure 6**). Adherent selection and subsequent expansion during culturing typically yields homogenous populations of BMSCs from marrow. This method is regarded as feasible and timely[134], lending to its popularity as a method for BMSC enrichment. Further, BMSCs can be induced to subsequently differentiate into a wide array of cell types upon the application of various culture mediums[135]. A well-established approach to accumulating osteogenic cells is by applying an osteogenic culture medium (containing components such as, Ascorbic acid, B-glycerophosphate, and Dexamethasome) to the BMSCs *in vitro*, thus yielding mature, mineralizing osteoblasts and potentially osteocyte-like cells embedding in mineralized nodules (**Figure 6**).

Accumulating adequate quantities of BMSCs and osteogenic cells is an essential prerequisite for studies aiming to characterize them, particularly via "-omics" technology; however, one obvious limitation is that by taking an *in vitro* approach, we do not replicate a perfect *in vivo* environment, in terms of biochemical composition or physical properties of bone tissue[135]. Differences in the expression of discrete genes or alteration of cellular signaling pathways could be affected by culturing cells *in vitro*[135]. Notwithstanding, as we and others show, *in vitro* usage of BMSCs and osteogenic cells remains a valid approach to characterizing many aspects of bone cell biology and findings can often be applicable or connected to *in vivo* phenomena[136–139].

1.7. Summary

Osteoporosis is a complex disease characterized by low bone mineral density (BMD), which can contribute to skeletal bone fragility and a drastic increase in the risk of bone fracture. While genome-wide association studies (GWAS) for BMD have discovered over 1100 associations, understanding how causal genes drive disease is convoluted by a lack of other molecular data (e.g., transcriptomics) for bone cell phenotypes. These "-omic" level datasets are essential to enhancing the utility of BMD GWAS. To overcome this challenge, we perform single-cell RNA-seq (scRNA-seq) on a large population of bone-relevant cell types (BMSC-OBs) from genetically diverse mice (DO mice) in order to characterize the transcriptomic landscape of these cells, elucidate gene co-expression networks (GCNs), and contextualize BMD GWAS-implicated genes to provide high priority targets for future investigations.

1) In **Chapter 2**, we demonstrate that bone marrow-derived stromal cells cultured under osteogenic conditions (BMSC-OBs) from the Diversity Outbred (DO) mouse population can be used as an *in vitro* model to generate single-cell transcriptomics data (scRNA-seq) for mesenchymal lineage cells in large numbers of mice (and potentially humans). We assessed the impact of the single-cell isolation procedure (used to liberate cells from a heavily mineralized matrix) and also compared the cell clusters generated in our *in vitro* model to cell types isolated directly from bone *in vivo*. Further, we made use of multiple scRNA-seq analytical tools to rigorously characterize the BMSC-OBs (e.g., SCENIC and CELLECT).

2) In **Chapter 3**, we showcase the scalability of our model and perform subsequent single-cell analyses from a larger sample pool. We perform scRNA-seq on BMSC-OBs from 80 DO mice in the same fashion as we previously described. We perform a pseudotime trajectory analysis to infer paths of differentiation across the cell clusters. Additionally, we perform a temporal gene expression analysis and identify genes with predicted roles in BMSC-OB differentiation. Further, in a cell type-specific expression quantitative trait locus (eQTL) analysis, we identify two eGenes (*Pkm*, *S100a1*) that can also be associated with significant differences in cell type proportion in mice with the genetic background driving the eQTL effect at each eGene locus. To inform BMD GWAS, we perform a cell type-specific network analysis. We aimed to contextualize genes with significant BMD GWAS associations and predict *Fgfrl1* and *Tpx2* as novel regulators of BMD.
3) In **Chapter 4**, I provide suggested future experiments to validate the roles of predicted targets in **Chapter 3**. Additionally, I provide my final thoughts, considerations, and future directions for the project.

The work described in this dissertation aims to showcase examples of "systems-level" approaches to investigate the genetics of osteoporosis and highlight targets with putative roles in human BMD by 1) leveraging our *in vitro* approach to generating osteogenic cells (BMSC-OBs) from the DO mouse population, 2) generating large-scale "-omics" data (scRNA-seq), and 3) performing a diverse array of single-cell analyses.

1.8. Chapter 1 - Main Figures

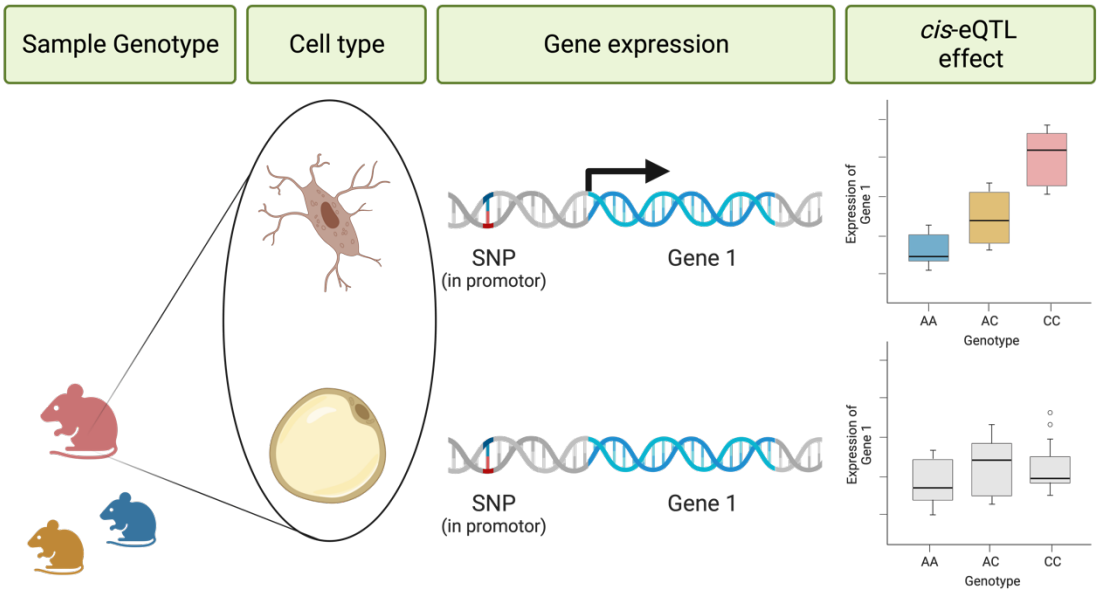


Figure 1. Expression Quantitative Trait Loci (eQTL) can affect the expression of genes in a cell type-specific fashion.
Based upon the genotype of a sample, transcriptomics data can be acquired from biological tissues of interest to resolve the effects of genetic variation on the expression of specific genes. Single-cell transcriptomics can resolve cell type-specific eQTLs to further characterize the effects of individual samples' genotype on the expression of genes. These genetic effects can be observed in specific cell types, but not others. For example, SNPs (associated with a specific genotype) found in a promotor region for a specific gene in osteocytes (brown cell) may only exert an observable effect on the expression of the gene in osteocytes, as opposed to adipocytes (yellow). Made with Biorender.

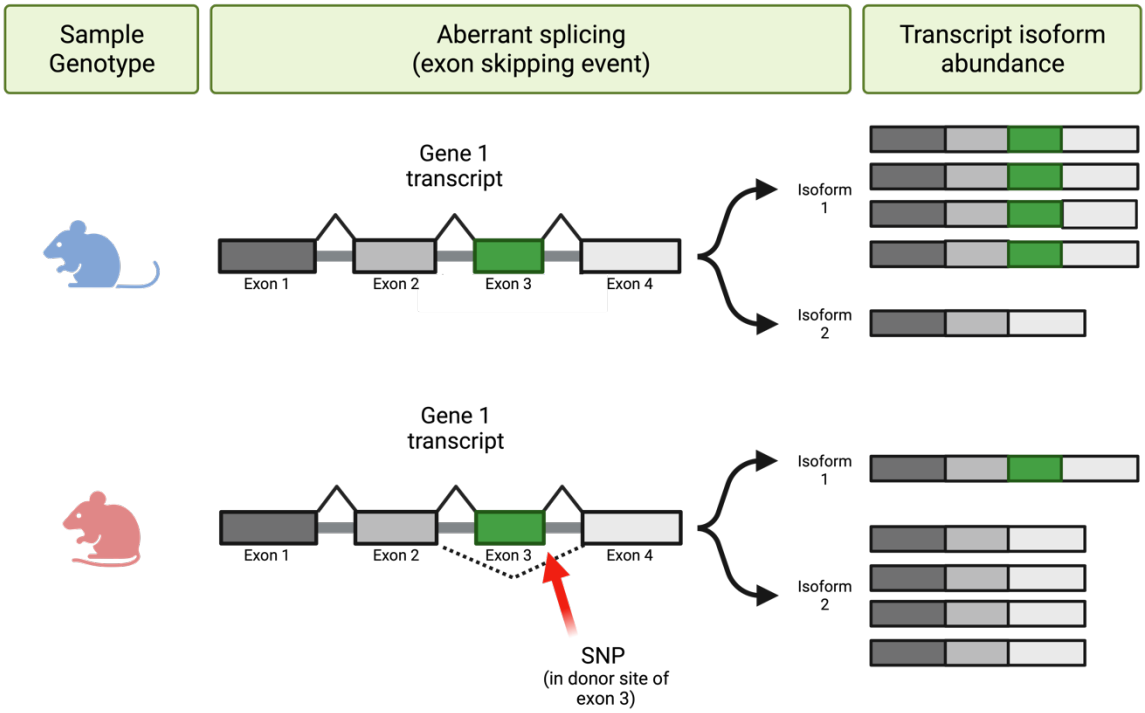


Figure 2. Splicing Quantitative Trait Loci (sQTL) can affect isoform-specific gene expression patterns.
Upon acquiring either bulk or long-read RNA sequencing data, isoform-specific gene expression can be resolved from samples of interest. Genetic variation associated with the samples' genotype can induce their effects on the biological process of splicing, resulting in a number of various events that can affect transcript expression, diveristy, abundance, etc. For exmaple, an aberrant splicing event can occur as a result of SNPs in donor sites of specific exons of expressed transcripts for a gene. In samples with genotypes confering such genetic variation, the isoform-specific gene expression patterns will be altered; the abundance and ratio of specific transcript isoforms can be skewed. Made with Biorender.

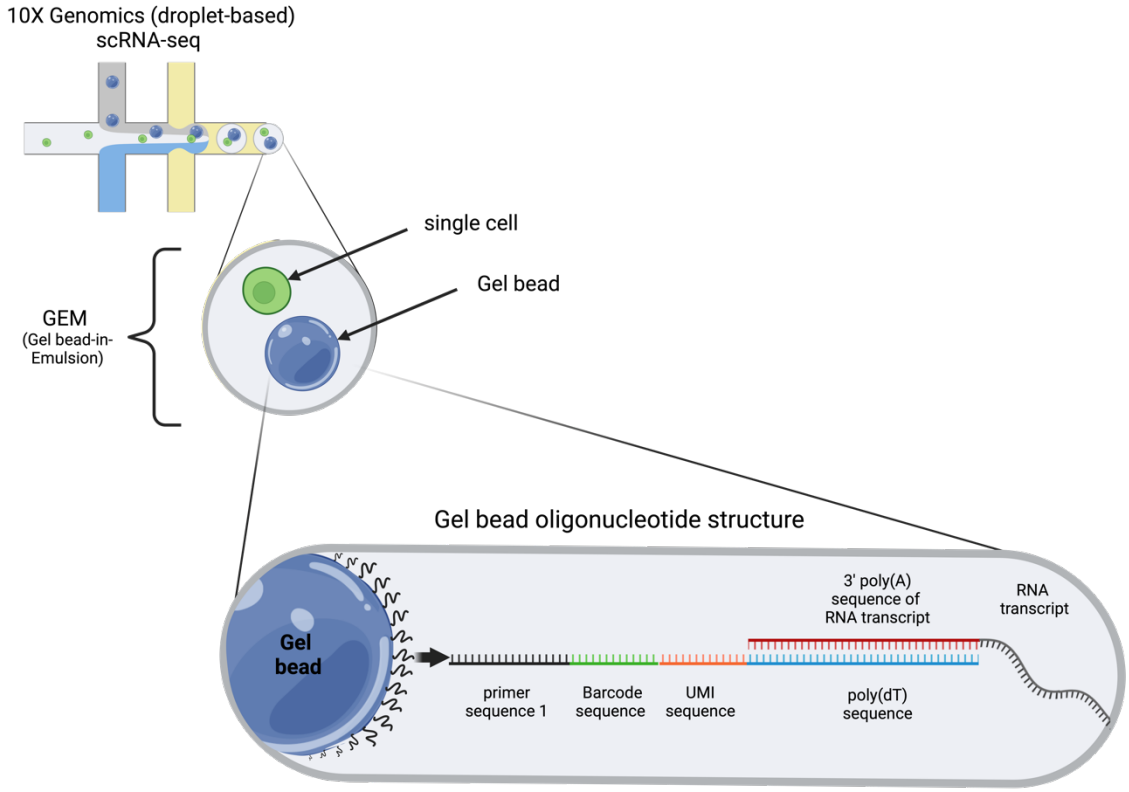


Figure 3. Summary of preperation strategy for single cells using droplet-based scRNA-seq.
Droplet-based strategies for scRNA-seq (e.g., 10X Genomics) employ microfluidics and oil-based chemistry to encapsulate indivudal cells in a GEM (Gel bead-in-Emulsion). The GEM bead contains a captured single cell along with a Gel bead (3' protocol depicted here). The Gel bead is composed of oligonucleotides which function to hybridize to polyadenylated RNA species (e.g., mRNA) released by the lysed single cell in the GEM. Additionally, the oligonucleotide contains specific sequences that are essential to downstream scRNA-seq analysis, such as the Barcode, Unique Molecular Identifier (UMI), and associated primer sequences. Made with Biorender.

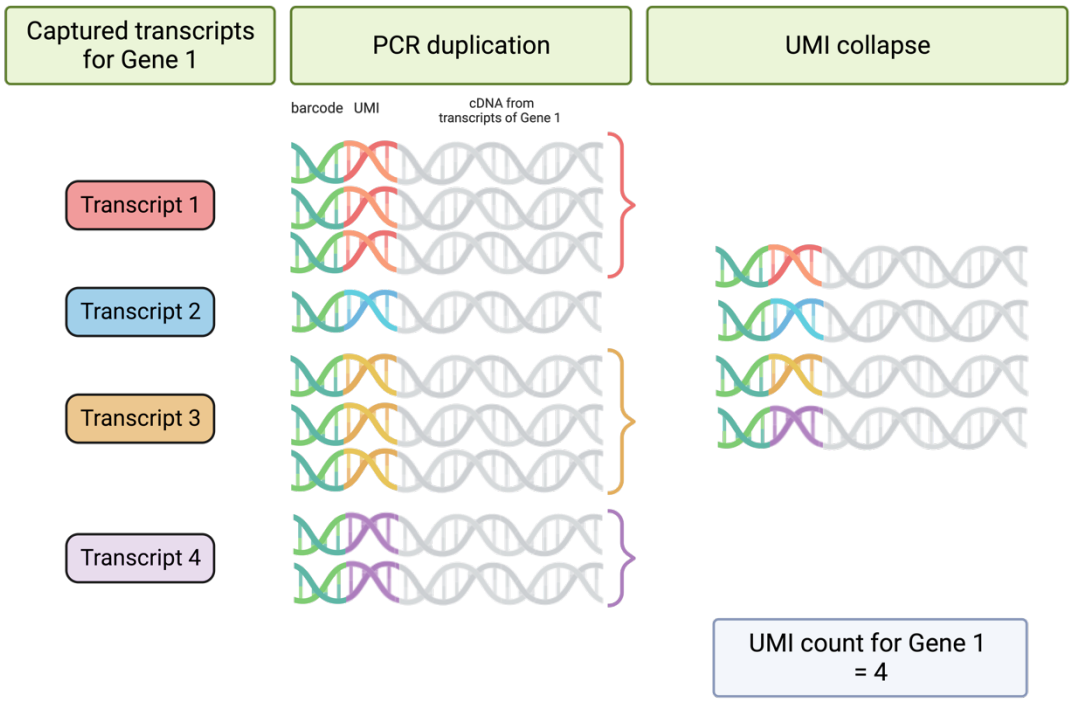


Figure 4. Summary of Unique Molecular Identifer (UMI) de-duplication.
UMI sequences index the RNA transcripts captured during scRNA-seq. They function to improve transcript quantification which can be skewed by PCR amplification during library construction. After sequencing, downstream analyses the UMIs enable de-duplication of exponentially amplified PCR duplicates. For example, Gene 1 is expressed in a cell to yield four transcripts, which are captured during scRNA-seq. During the PCR amplification process, some of the captured transcripts are duplicated in various amounts. During downstream bioinformatic analysis of the resulting scRNA-seq data, UMIs are collapsed to de-duplicate the sequenced transcripts. Therefore, the de-duplicated UMI counts yield a more accurate quantification of gene expression for Gene 1. Made with Biorender.

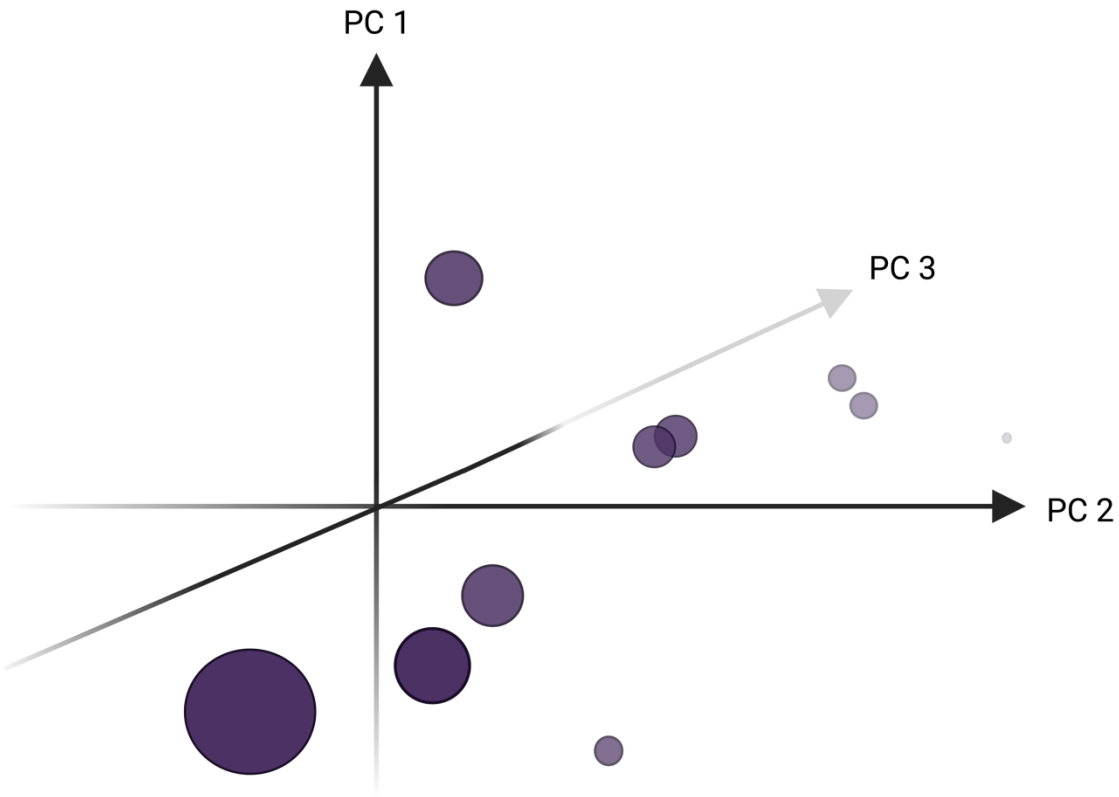


Figure 5. Single cells captured via scRNA-seq are projected in a multi-dimentional, Principle Component (PC) space to highlight variability in gene expression.
During downstream bioinformatic analysis of scRNA-seq data, individual single cells are typically projected into a highly-dimentional PC space (three PCs shown here for visualization purposes). Cells cluster together based upon their transcriptomic profiles; cells sharing similar gene expression signatures cluster together while variability in gene expression drives the seperation of groups of cells. Subsequent dimentionality reduction techniques are applied to summarize this multi-dimentional space for visualization purposes (e.g., Uniform Manifold Approximation and Projection, UMAP). Made with Biorender.

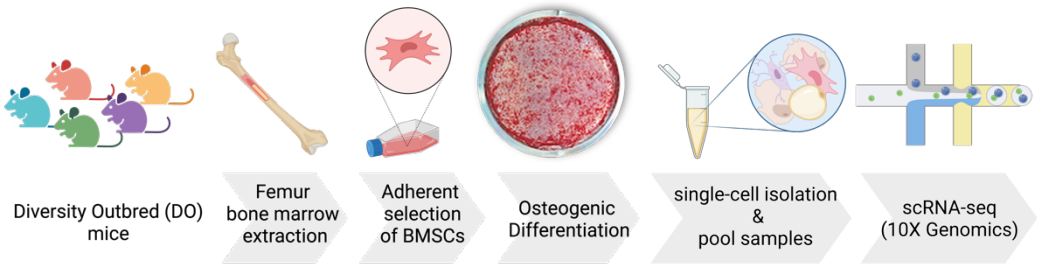


Figure 6. Overview of the BMSC-OB model. Bone marrow derived stromal cells (BMSCs) are extracted from the femurs of Diversity Outbred (DO) mice.
These mutli-potent mesenchmyal stems cells are subseqently selected for and expanded in vitro (~3 days). Osteogenic differentiation culture medium is applied over the course of 10-12 days to facilitate the differnetiation of the BMSCs towards bone cell fates (e.g., osteoblasts (OB), osteocyte-like cells). BMSC-OBs are then released from their mineralized matrix using a single cell isolation procedure and prepared for droplet-based (e.g., 10X Genomics) scRNA-seq. Made with Biorender.

Chapter 2

Single-Cell Transcriptomics of Bone Marrow Stromal Cells in Diversity Outbred Mice: A Model for Population-Level scRNA-Seq Studies

Published in:

Dillard LJ, Rosenow WT, Calabrese GM, et al. Single-Cell Transcriptomics of Bone Marrow Stromal Cells in Diversity Outbred Mice: A Model for Population-Level scRNA-Seq Studies. J Bone Miner Res. 2023;38(9):1350-1363. doi:10.1002/jbmr.4882

2.1. Abstract

Genome-wide association studies (GWASs) have advanced our understanding of the genetics of osteoporosis; however, the challenge has been converting associations to causal genes. Studies have utilized transcriptomics data to link disease-associated variants to genes, but few population transcriptomics data sets have been generated on bone at the single-cell level. To address this challenge, we profiled the transcriptomes of bone marrow–derived stromal cells (BMSCs) cultured under osteogenic conditions from five diversity outbred (DO) mice using single-cell RNA-seq (scRNA-seq). The goal of the study was to determine if BMSCs could serve as a model to generate cell type–specific transcriptomic profiles of mesenchymal lineage cells from large populations of mice to inform genetic studies. By enriching for mesenchymal lineage cells in vitro, coupled with pooling of multiple samples and downstream genotype deconvolution, we demonstrate the scalability of this model for population-level studies. We demonstrate that dissociation of BMSCs from a heavily mineralized matrix had little effect on viability or their transcriptomic signatures. Furthermore, we show that BMSCs cultured under osteogenic conditions are diverse and consist of cells with characteristics of mesenchymal progenitors, marrow adipogenic lineage precursors (MALPs), osteoblasts, osteocyte-like cells, and immune cells. Importantly, all cells were similar from a transcriptomic perspective to cells isolated in vivo. We employed scRNA-seq analytical tools to confirm the biological identity of profiled cell types. SCENIC was used to reconstruct gene regulatory networks (GRNs), and we observed that cell types show GRNs expected of osteogenic and pre-adipogenic lineage cells. Further, CELLECT analysis showed that osteoblasts, osteocyte-like cells, and MALPs captured a significant

component of bone mineral density (BMD) heritability. Together, these data suggest that BMSCs cultured under osteogenic conditions coupled with scRNA-seq can be used as a scalable and biologically informative model to generate cell type–specific transcriptomic profiles of mesenchymal lineage cells in large populations.

2.2. Introduction

Osteoporosis is a disease characterized by low bone mineral density (BMD) and an increased risk of fracture[140]. Osteoporosis-related quantitative traits, such as BMD, are highly heritable[141], and genome-wide association studies (GWASs) for BMD have identified more than 1100 independent associations[10]. The goal of BMD GWAS is to identify responsible causal genes[142,143]. However, this is often difficult because of challenges such as linkage disequilibrium between potentially causal variants[143] and the observation that most associations implicate non-coding variation[144]. The generation of transcriptomics data and use of systems genetics approaches to interpret GWAS can address these limitations by assisting in prioritizing putatively causal genes for further investigation[145,146]

The utility of transcriptomic data to inform BMD GWAS has been demonstrated through studies using approaches such as expression quantitative trait locus (eQTL) mapping and colocalization[147–149], transcriptome-wide association studies (TWASs)[29,150], and reconstruction of transcriptomic networks (e.g., gene-regulatory and co-expression networks)[19,43,44]. These studies have utilized bone, non-bone (e.g., the Gene Tissue Expression [GTEx] project)[151], and mouse bone transcriptomic data. However, all of the transcriptomic data used to inform BMD GWAS to date has been

generated using bulk RNA-seq. These samples are a mixture of data derived from all cells associated with a particular microenvironment, and downstream data analysis is often constrained by the inability to definitively attribute transcriptomic signatures to a single cell type[55]. Further, signals from potentially rare cell types can be masked by the presence of more abundant cell populations[152]. As a result, there is currently a need to generate population-scale (i.e., hundreds of samples) cell type–specific expression data on cells directly relevant to bone to aid in the identification of causal BMD GWAS genes.

In recent years, single-cell RNA-seq (scRNA-seq) has enabled the efficient generation of high-quality transcriptomes from individual cells[153]. ScRNA-seq can remedy the aforementioned challenges posed with bulk RNA-seq by enabling the generation of single-cell transcriptomic profiles from heterogeneous tissues or primary cell cultures. ScRNA-seq has provided significant insight into the landscape of bone cell types[154–158]. However, we still lack cost-effective approaches capable of generating scRNA-seq data at scale for key bone cell types.

Here, we explored the use of bone marrow–derived stromal cells (BMSCs) cultured under osteogenic conditions (BMSC-OBs), a popular in vitro model of osteoblast differentiation, to address the above limitations by generating scRNA-seq data on cells of the mesenchymal lineage. We sought to explore technical challenges, cellular heterogeneity, and compare cultured cells to the same cells isolated directly from bone. We show that this approach not only enriches for osteogenic cells but also is a scalable approach capable of generating biologically informative cell type–specific transcriptomic profiles relevant to BMD GWAS. Our results suggest that scRNA-seq of BMSC-OBs has the potential to enable the large-scale generation of cell type–specific transcriptomic data

on mesenchymal lineage cells that can be used to inform genetic studies in mice and potentially humans.

2.3. Materials and Methods

2.3.1. Sample preparation and in vitro cell culture of BMSCs

From a large cohort of diversity outbred (DO) mice characterized in Al-Barghouthi and colleagues[19], 5 mice (12, 45, 48, 50, 84) were selected randomly for in vitro culture of BMSCs and subsequent scRNA-seq. Bone marrow extraction and subsequent cell culture was performed as described in Al-Barghouthi and colleagues[19]. In brief, femurs were isolated and marrow was exuded by centrifuging at 2000g for 30 seconds and suspended in 35 µL of fetal bovine serum (FBS, Atlantic Biologicals, Miami, FL, USA). After the addition of 150 µL of cold media (90% FBS, 10% dimethyl sulfoxide [DMSO; Thermo Fisher Scientific, Waltham, MA, USA]), marrow was triturated six times, placed into a Mr. Frosty Freezing Container (Nalgene, Rochester, NY, USA), and stored in liquid nitrogen for storage. In preparation for cell culture, samples were thawed at 37°C and resuspended in 5 mL of bone marrow growth media (MEM alpha [Gibco, Thermo Fisher Scientific], 10% FBS, 1% penicillin streptomycin [pen/strep, Gibco], and 1% glutamax [Gibco]). Samples were subjected to red blood cell lysis by resuspending in 5 mL of 0.2% NaCl for 20 seconds, then thorough mixing of 1.6% NaCl. Cells were pelleted, resuspended in 1 mL bone marrow growth media, and cells from each mouse were cultured in separate wells of a 48-well tissue culture plate. Samples were incubated in 37°C, 5% CO2, 100% humidity incubator for 3 days. Thereafter, media was aspirated and replaced daily and adherent cells were allowed to

grow to confluence. After 6 days, cells were washed and underwent in vitro osteoblast differentiation for 10 days by replacing bone marrow growth media with 300 µL of osteogenic differentiation media (alpha MEM, 10% FBS, 1% pen/strep, 1% glutamax, 50 µg/µL ascorbic acid [Sigma, St. Louis, MO, USA], 10 nM B-glycerophosphate [Sigma], 10 nM dexamethasome [Sigma]).

2.3.2. Single-cell isolation procedure

The isolation procedure outlined below was inspired by Hanna and colleagues[159]. Mineralizing cultures were washed twice with Dulbecco's phosphate-buffered saline (DPBS, Gibco). A total of 0.5 mL of 60 mM ethylenediaminetetraacetic acid pH 7.4 (EDTA [Thermo Fisher Scientific], made in DPBS) was added to cultures and incubated at room temperature (RT) for 15 minutes. EDTA solution was aspirated, replaced, and cultures were incubated again at RT for 15 minutes. Cultures were washed with 0.5 mL Hank's balanced salt solution (HBSS, Gibco) and subsequently incubated with 0.5 mL 8 mg/mL collagenase (Gibco) in HBSS/4 mM CaCl2 (Fisher) for 10 minutes at 37°C with shaking. Cultures were then triturated 10 times and incubated again for 20 minutes. Samples were then transferred to a 1.5 mL Eppendorf tube, centrifuged at 500g for 5 minutes at RT, resuspended in 0.5 mL 0.25% trypsin–EDTA (Gibco), and incubated for 15 minutes at 37°C. After trituration, samples were incubated for an additional 15 minutes, after which 0.5 mL of media was added, incubated once more for 15 minutes, and centrifuged at 500g for 5 minutes at RT. Cultures were resuspended in 0.5 mL osteogenic differentiation media and cells were counted. After single-cell isolation, cells

from each of the five individual culture wells were pooled and concentrated to 800 cells/μL in PBS supplemented with 0.1% BSA (bovine serum albumin).

2.3.3. Single-cell analysis pipeline

Pooled single cells were prepared for sequencing using the 10× Chromium Controller (10× Genomics, Pleasanton, CA, USA), as described in Al-Barghouthi and colleagues[19]. After the library was sequenced on the NextSeq500 (Illumina, San Diego, CA, USA), data were processed using 10× Genomics Cell Ranger toolkit (version 5.0.0) and reads were mapped to the GRCm38 reference genome[160]. Overall, 8990 cells were captured and data are available on the Gene Expression Omnibus (GEO) at accession code GSE152806.

Seurat[161] (version 4.1.1) was used for analysis of the scRNA-seq data. A Seurat object was generated with the inclusion of features detected in at least three cells and cells with at least 200 features detected. Souporcell[162] (described below) was used to remove doublet cells. Additionally, cells with more than 5800 reads and less than 800 reads were removed, as well as those cells with more than 10% mitochondrial reads. After filtering, 7357 cells remained for further analysis. The resulting object underwent standard normalization, scaling, and the top 3000 features were modeled from a variance stabilizing transformation (VST) using the Seurat "FindVariableFeatures" function. Cell-cycle markers based on Tirosh and colleagues[163] were regressed out using the "CellCycleScoring" and scaling functions. For subsequent dimensionality reduction, 14 principal components (PCs) were summarized, which was the last PC in which the percent change in variation between the consecutive PCs was quantified to be more than

0.1%, as described in Piper and colleagues[164]. A kNN (k = 20) graph was created and the Louvain algorithm was used to cluster cells at a resolution of 0.22. Annotation of cell-type clusters was performed manually based on differential gene expression analysis using the Seurat "FindAllMarkers" function (**Supplemental Table S1**).

2.3.4. Bulk RNA-seq analysis

Total RNA was extracted using a RNeasy Micro Kit (QIAGEN, Valencia, CA, USA) and poly-A selected RNA was sequenced via GENEWIZ (South Plainfield, NJ, USA). RNA-seq analysis was performed using a custom bioinformatics pipeline. Briefly, FastqQC[165] and RSeQC[166] were used to assess the quality of raw reads. Adapter trimming was completed using Trimmomatic[167]. Sequences were aligned to the GRCm38 reference genome[160] using the single-nucleotide polymorphism (SNP) and splice aware aligner HISAT2[168]. Genome assembly and abundances in counts per million (CPM) were quantified using StringTie[169]. Differential expression analysis was performed using the DESeq2[170] package in R.

2.3.5. Integration of data sets via canonical correlation analysis (CCA)

CCA[171] in Seurat was used to integrate *in vivo* scRNA-seq data derived from Zhong and colleagues[156] (1-, 1.5-, and 3-month time points) with the BMSC-OB in vitro data. The Zhong and colleagues[156] data were first pre-processed in the same fashion as the BMSC-OBs scRNA-seq data set and clustered at a final resolution of 0.675 (**Supplemental Fig. S2**). Cell types not present in the BMSC-OBs data set were removed from the Zhong and colleagues[156] data in order to portray only osteogenic and adipogenic

lineage cells. After integration, the combined data set was analyzed as described in the single-cell analysis pipeline (above) and clustered at a final resolution of 0.22 (**Supplemental Fig. S5**).

2.3.6. Souporcell

Upon performing Souporcell[162] (version 2.0.0), barcoded cells identified as doublets were removed from the scRNA-seq count matrix during pre-processing of the data. Additionally, Souporcell was used to perform genotype deconvolution using the GRCm38 reference genome[160]. Five genotypically distinct clusters (genotypes) were inferred based on variants in the sequenced reads. Genotype clusters were assigned their corresponding DO mouse ID by comparing allele calls made by the shared variants captured between Souporcell and GigaMUGA arrays previously performed on all mice in the cohort. DO mouse IDs were assigned by making a pairwise comparison between each Souporcell genotype cluster and GigaMUGA array. The comparison yielding the highest percentage of matching allele calls indicated the identity/genotype of each mouse (**Supplemental Table S7**).

2.3.7. Scenic

pySCENIC[172] (Single-Cell rEgulatory Network Inference and Clustering) (version 0.11.2) was used to infer gene regulatory networks. A fully processed Seurat object containing cell-type annotations was transformed into a loom file by using SeuratDisk[173] (version 0.0.0.9019). The loom file was subsequently used as input to the SCENIC workflow[172]. In brief, gene regulatory networks (GRNs) were built using GRNBoost[174] to

identify potential gene targets for each transcription factor (TF) based on co-expression. CisTarget[175] was then used to select potential direct target genes of the governing TF of the co-expression modules (**Supplemental Table S11**). The activities of the final regulons were calculated using AUCell[176] (**Supplemental Tables S12 and S13**). Regulon specificity score (RSS) is based on Jensen-Shannon divergence measurements, as described in Suo and colleagues[177] (**Supplemental Table S14**). The most active and specific regulons as well as associated target genes were resolved for each cell-type cluster.

2.3.8. CELLECT

CELLECT[178] (CELL-type Expression-specific integration for Complex Traits) (version 1.1.0) was used to identify likely etiologic cell types underlying complex traits of both the BMSC-OBs and Zhong and colleagues[156] data sets. CELLECT quantifies the association between the GWAS signal and cell-type expression specificity using the S-LDSC genetic prioritization model[179]. Summary statistics from the UK Biobank eBMD and Fracture GWAS[10] (Data Release 2018) and cell-type annotations from each scRNA-seq data set were used as input. Cell-type expression specificities were estimated using CELLEX(45) (CELL-type EXpression-specificity) (version 1.2.1). The CELLECT output prioritizes likely etiologic cell types for BMD (**Table 1**).

2.4. Results

2.4.1. BMSC cultures grown under osteogenic differentiation conditions are heterogenous

We isolated BMSCs from 5 DO mice (n = 1 male and n = 4 females). The DO is a genetically diverse outbred population derived from eight inbred laboratory strains[114]. We have previously used the DO to perform GWAS for bone strength traits[19]. BMSCs were cultured under osteogenic conditions for 10 days and cells generated mineralized nodules as previously shown in Al-Barghouthi and colleagues[19] (**Supplemental Fig. S1**). After differentiation, cells were liberated from mineralized cultures and profiled using scRNA-seq. After stringent pre-processing and quality control of the data (Materials and Methods), 17,457 genes were identified in 7357 cells across all 5 mice. Unsupervised clustering identified eight cell clusters ranging in size from 46 to 2367 cells (**Fig. 1A**).

We manually annotated the cell-type identity of each cluster using the "FindAllMarkers" function in Seurat[161] to highlight differentially expressed genes (DEGs) for each cluster relative to all other clusters (**Supplemental Table S1**). As a framework, we used the nomenclature of Zhong and colleagues[156], who labeled, FACS-selected, and sequenced single cells from bone marrow using Col2-Cre Rosa26 < lsl-tdTomato > reporter mice[156]. In these mice, tdTomato (Td) labels cells spanning the mesenchymal lineage. From Td+ selected cells, three types of mesenchymal progenitors were identified: early (EMPs), intermediate (IMPs), and late (LMPs). None of the BMSC-OB clusters reported here had signatures of EMPs or IMPs (**Fig. 1A**); however, cluster 0 (32.2% of the cells) had high expression of marker genes associated with LMPs, such as Aspn, Timp3, Thbs2, and Itm2a (**Fig. 1A, B**). Clusters 1, 2, and 4 (49.7% of the

cells) all had signatures of cells in the osteoblast lineage. Mature osteoblasts (OB) in cluster 1 exhibited expression of Bglap and Mepe, whereas cluster 4 had a transcriptomic signature of osteocyte-like cells (Ocy) with high expression of Bambi and Sost (**Fig. 1A, B**). Cells in cluster 2 resembled an osteoblast progenitor (OBP) population differentiating into mature osteoblasts and expressed genes such as Sgms2, Ifitm5, and P4ha1 (**Fig. 1A, B**). Relative to the Zhong and colleagues[156] data, we observed an enrichment in the proportion of mature bone cell types. In the BMSC-OB scRNA-seq data set, OBs and Ocy-like cells accounted for 29.1% and 5.6% of all sequenced cells, respectively, a notable increase from 8.0% (OB) and 0.9% (Ocy) in the Zhong and colleagues[156] data set (**Supplemental Fig. S2**). Marrow adipogenic lineage precursors (MALPs), identified as a novel component of bone marrow in Zhong and colleagues[156], were represented in cluster 3 (accounting for 9.7% of the cells) and expressed known MALP markers (*Cxcl12, Adipoq, H19, Hp, Lpl*) (**Fig. 1A, B**). Clusters 5, 6, and 7 (8.3% of the cells) were cells not associated with the mesenchymal cell lineage and have transcriptomic signatures of immune cells derived from the hematopoietic cell lineage (**Fig. 1A, B**). The expression of select marker genes representative of all cell types were consistent with cell-type annotations (**Fig. 1C**).

2.4.2. Cell clustering is robust to the effects of cell isolation

The isolation of cells from their heavily mineralized matrix took ~2 hours, raising the possibility that the procedure itself could have an effect on gene expression, transcriptomic signatures, and downstream clustering of cells. To directly assess the effects of the single-cell isolation procedure, we performed a separate experiment in

which we generated two identical cultures of BMSC-OBs (10 days post differentiation as in the scRNA-seq experiment) from C57BL/6J mice ($n = 7$) (**Fig. 2A**). From one culture (bulk), we extracted RNA from the entire culture and performed RNA-seq. From the other culture (pooled single-cell bulk, psc-bulk), cells were harvested via the single-cell isolation procedure, pooled into one sample, and profiled using RNA-seq. Overall gene expression between the bulk and psc-bulk samples was highly correlated ($r = 0.99$, $p < 2.2 \times 10^{-16}$) (**Fig. 2B**). However, a total of 776 genes were differentially expressed ($p_{adj} < 0.05$) with a fold-change (FC) less than 0.5 and greater than 2.0 in the psc-bulk versus bulk samples (**Supplemental Table S2**). A PANTHER[180] Gene Ontology (GO) enrichment analysis revealed that DEGs consisted of "acute inflammatory response" (GO:0002675, $n = 11$, $p = 2.43 \times 10^{-8}$) and "response to stress" (GO:0080134, $n = 111$, $p = 4.96 \times 10^{-19}$) signatures (**Supplemental Table S3**). Of the 776 DEGs identified in the psc-bulk versus bulk samples, 684 (88%) were captured in the entire BMSC-OB scRNA-seq count matrix and 107 (14%) were quantified as a DEG ($p_{adj} < 0.05$, FC greater than 2.0 or less than 0.5) in any given cell cluster of the scRNA-seq data (**Supplemental Table S4**). The majority of the expression of these DEGs are localized to hematopoietic immune cell clusters in the scRNA-seq data and are mediators of inflammation (eg, chemokines) or involved with immune cell activation (eg, cell surface or immunoglobulin receptors) (**Supplemental Fig. S3, Supplemental Table S5**). Of the 107 DEGs, 79 (74%) of them were expressed exclusively in the immune cell clusters (**Supplemental Table S6**). To evaluate the impact of the single-cell isolation procedures on cell clustering of the scRNA-seq data set, we removed all 776 DEGs from the scRNA-seq count matrix. Upon removal, a negligible effect was observed on the cells

clustering in Uniform Manifold Approximation and Projection **(**UMAP) space and six distinct cell clusters (five mesenchymal lineage cell clusters) were annotated, similar to the original UMAP (**Fig. 2C**). Only 8.1% of cells shifted from their original cell cluster assignments upon removal of the DEGs (**Fig. 2D**). Most of the cells with shifted assignments were located on the boundaries of cell clusters (**Fig. 2D**). These data indicate that gene expression is altered in a predictable manner by the cell isolation procedure but has little meaningful impact on cell clustering.

2.4.3. Cell types isolated from BMSC-OBs are similar to their in vivo counterparts

We next wanted to determine if mesenchymal cells generated in vitro were similar, in terms of global gene expression, to cell types isolated directly from bone. Zhong and colleagues[156] performed scRNA-seq on Td+ bone marrow cells from mice at 1, 1.5, and 3 months of age. We jointly processed the data from both experiments and integrated the data sets using canonical correlation analysis (CCA)[171]. Overall, the cells from both experiments displayed significant overlap (**Fig. 3A**). This was even more apparent when clusters were annotated and cell types (LMPs, MALPs, OBs, and Ocy-like cells) overlapped in UMAP space between the data sets (**Fig. 3B**). A notable difference between cell types was the absence of EMPs and IMPs in the cultured BMSC-OBs. However, an appreciable enrichment of osteoblast lineage cells, particularly in the OB population, was observed in the BMSC-OB data compared with the cells isolated directly from bone (**Fig. 3B, C**). Importantly, the overlap of cells from the two studies suggests few transcriptional differences as a consequence of cell culture and in vitro differentiation.

2.4.4. Transcriptomic profiles from scRNA-seq for individual cell types are robust

One of our goals for future experiments will be to generate expression profiles for multiple mesenchymal cell types in large populations of mice (or humans) for use in downstream applications such as eQTL analysis or the generation of networks to inform human GWAS. To evaluate how well cell type–specific expression profiles from scRNA-seq align with profiles generated via traditional bulk RNA-seq, we performed a correlation analysis between the expression profiles derived from each of the six defined cell types (five mesenchymal + one grouped immune cell cluster) examined in this study to bulk RNA-seq data (derived from psc-bulk data described above). We generated a "pseudobulk" profile (PB) from the entire scRNA-seq data by aggregating unique molecular identifier (UMI) counts for each gene across all cells to simulate a data set representative of one derived from bulk sequencing methods. Additionally, cell type–specific PB profiles were generated by aggregating UMI counts for cells belonging to a specific cell type. A high correlation was observed between both the bulk/psc-bulk profiles and the PB profile generated for the entire scRNA-seq data set ($r = 0.84$ and $r = 0.85$, respectively; $p < 2.2 \times 10^{-16}$) (**Fig. 4A**). We generally observed high pairwise correlations ($r > 0.9$) among osteogenic cell cluster PB profiles (LMP, OBP, OB, and Ocy) and between osteogenic clusters and MALPs (**Fig. 4B**). The immune cell cluster PB profile had a relatively lower correlation ($r < 0.9$) to both the osteogenic clusters and MALPs (**Fig. 4B**). Cell type–specific PB profiles were compared individually to the psc-bulk profile as well (**Fig. 4B**). As expected, the correlations were

similar to the correlation observed between the psc-bulk profile and the PB profile for the entire scRNA-seq data set ($r = 0.85$, **Fig. 4A**).

Additionally, we estimated the minimum number of cells per cluster required to generate robust cell-type expression profiles by randomly selecting from 2 to 400 cells from each cluster, generating a PB profile (as described above), and subsequently calculating the correlation between each cell-type PB profile to the psc-bulk sample. Calculated correlations plateaued for all cell types at ~100 cells (**Fig. 4C**). These data indicate that aggregated data across at least 100 cells from a given cell type approximates data generated from bulk RNA-seq.

2.4.5. Frequency of osteogenic cell types are highly variable across DO mice

Because the BMSC-OB scRNA-seq data set consisted of multiple samples pooled into one, we used Souporcell[162] for genotype deconvolution to assign a mouse-of-origin for each cell. Five genotypically distinct clusters (genotypes) were inferred by Souporcell from the scRNA-seq data based on SNPs captured in the sequenced cDNA. Genotype clusters were assigned to their corresponding DO mouse ID by comparing allele calls made for the variants captured between Souporcell and genotypes previously generated on all five DO mice using GigaMUGA genotyping arrays[19,116]. Of the 67,056 total variants identified by Souporcell, 0.87% (581) were also captured by the GigaMUGA arrays (143,259 total). DO mouse IDs were assigned based on the highest percentage (all ~90%) of matching allele calls made upon pairwise comparison between Souporcell clusters and GigaMUGA arrays (**Supplemental Table S7**). Upon quantifying percent *Xist* expression in all single cells, we confirmed accurate Souporcell genotype

clustering of the male mouse in our cohort (**Supplemental Table S8, Supplemental Fig. S4**). After assigning a mouse-of-origin to all cells in the scRNA-seq data, we quantified differences in the frequencies of various cell types contributed by each mouse (**Fig. 5A**). For example, mouse 50 had a higher frequency of LMPs and MALPs and fewer osteoblasts and osteocytes compared with the other four mice (**Fig. 5A, B**). Additionally, we recorded a variety of phenotypic trait measurements on the five mice (**Supplemental Tables S9 and S10**); however, inferring correlation between cell-type proportions and trait metrics will require a larger sample size. Pooling samples for scRNA-seq, coupled with downstream genotype deconvolution, is an approach that is scalable for multi-sample input, which is necessary to reduce costs associated with performing population-level investigations.

2.4.6. BMSC-OBs show expected gene regulatory networks

Cell-type identification is largely based on the association of canonical and highly expressed genes within certain cell types; however, underlying gene regulatory networks (GRNs) provide insight into how expression is coordinated[181]. Moreover, GRN inference can be used to establish gene expression profiles for cell types of interest by elucidating which distinct transcription factors (TFs) are responsible for modulating the expression of downstream target genes[181]. We used SCENIC[176] to better understand the GRNs that characterize the cell states in BMSC-OBs. The SCENIC analysis pipeline first generates regulatory modules inferred from co-expression patterns, which are used to form "regulons" consisting of a core TF that governs the expression of predicted target genes. Next, direct target genes are selected based on enrichment of the TF cis-regulatory motifs

located upstream or downstream of target genes in the regulon (**Supplemental Table S11**). Across all individual cells, the activity of each regulon is quantified (**Supplemental Table S12**).

We applied the SCENIC analysis pipeline to the BMSC-OBs and resolved distinct regulons associated with each cell cluster in the BMSC-OB data set (**Fig. 6**). Regulons were robust in activity (**Fig. 6A, B**; **Supplemental Table S13**) and specific for each cell type (**Fig. 6C, D**; **Supplemental Table S14**). For example, *Sp7* (Osterix), a key TF known to be involved in osteoblast differentiation, was found to be more specifically associated and highly active in the OBP cell cluster (**Fig. 6C, D**). Similarly, we show *Pparg* is a highly active regulon and more exclusively associated with MALPs (**Fig. 6C, D**), consistent with its role as a master regulator of adipogenesis. This analysis suggests that not only do BMSC-OB cell types show similar transcriptomic signatures to the same cells isolated directly from bone, but also cell circuits (ie, GRNs) are similar.

2.4.7. MALPs and osteogenic cells capture BMD heritability identified by GWAS

We next used CELLECT[178] to evaluate the relevance of identified cell types with regard to mediating the effects of GWAS. CELLECT is designed to use GWAS and scRNA-seq data to identify cell populations that are enriched for BMD GWAS heritability[178]. We applied CELLECT to the cell types identified in BMSC-OBs and those identified by Zhong and colleagues[156]. Gene expression specificity (ESμ), which is quantified as the marginal likelihood of a gene being specifically expressed in a given cell type[178], was determined for each gene in each cell type across both scRNA-seq data sets (**Supplemental Tables S15 and S16**). We observed that genes with selective

expression in MALPs, OBs, and Ocy-like cells from both data sets were significantly ($p < 0.05$) enriched for BMD heritability (**Table 1**). In addition, IMPs and LMPs in the Zhong and colleagues[156] data set were also significant. Non-mesenchymal lineage cells, which are mostly immune cells in both data sets, were not significant (**Table 1**). Interestingly, osteoclasts captured in the Zhong and colleagues[156] data set were not identified as significant in the CELLECT analysis (**Table 1**).

2.5. Discussion

A considerable challenge faced upon analyzing GWAS is identifying the causal genes impacted by significant associations. Integrating transcriptomics data has proven invaluable for accomplishing this goal. Colocalizing genetic variation impacting gene expression with GWAS associations can identify putative causal genes influencing disease. Moreover, integrating single-cell transcriptomics data can provide the cellular context in which causal genes are most likely to be impactful. In the context of osteoporosis research, the generation of population-scale transcriptomics data at single-cell resolution would aid in gene discovery. Here, we demonstrate the use of BMSCs cultured under osteogenic conditions (BMSC-OBs) from the DO mouse population coupled with scRNA-seq can serve as a model to generate single-cell transcriptomics data of mesenchymal cell types relevant to bone. Further, we demonstrate the utility of the BMSC-OB model for feasibly generating population-scale scRNA-seq data in a cost-efficient manner.

A number of approaches have been used to profile individual bone cells[182]. These include scRNA-seq on whole bone marrow, using fluorescence-activated cell sorting

(FACS) on marrow to enrich for mesenchymal lineage cells, the digestion of bone combined with FACS, and FACS on lineage-specific reporter mice. These studies have provided important insights into the cellular landscape of bone and the identity of skeletal stem cells. However, none of these approaches were developed with the goal of investigating bone cells at the population scale in mice or humans. These approaches isolate a wide range of cells, many of which provide little insight in the context of informing BMD GWAS. Profiling non-relevant cells significantly increases cost and makes population screening less feasible. As an alternative, BMSC-OBs have several attractive attributes. First, it is simple, marrow is relatively easy to isolate from populations of mice, or even humans, and isolating BMSCs based on plastic adherence is cost-effective and straightforward. Second, we show that osteoblasts and osteocyte-like cells are some of the most relevant to BMD GWAS, and we were able to enrich for these cells by culturing under osteogenic conditions. Third, we do not need to use FACS or specific reporter mice, making it possible to perform this approach in any population of mice and potentially humans.

We show that after subsequent culturing of BMSCs under osteogenic differentiation, there was an enrichment in the relative frequencies of osteoblasts and osteocyte-like cells compared with cells isolated in vivo using a mesenchymal lineage reporter in Zhong and colleagues[156]. Additionally, the model yields adipogenic progenitor cells and their transcriptomic signature is similar to the MALPs identified in Zhong and colleagues[156]. These cells are classified as a stable intermediary cell type along the adipogenic differentiation route after mesenchymal progenitors and before more mature, lipid-laden adipocytes (LiLAs)[156]. Although more cell heterogeneity was observed than

initially expected after adherent selection of BMSCs and in vitro osteoblast differentiation, the sequenced BMSC-OBs yielded by our model were enriched for mature osteogenic cells (OBs and Ocy-like cells), which we demonstrate are likely to be the most relevant for informing BMD GWAS.

We addressed the technical challenges posed with our approach, such as the single-cell isolation procedure used to liberate BMSC-OBs from a highly mineralized matrix in vitro. This procedure consists of an ~2-hour process involving incubations with proteases and EDTA, raising the concern of technical effects impacting the integrity and quality of the isolated cells for scRNA-seq. In the bulk versus psc-bulk experiment, we sought to characterize the impact of the single-cell isolation procedure on gene expression. Despite the induction of inflammation/stress-related genes in the psc-bulk sample, the overall gene expression profiles between bulk and psc-bulk samples were highly correlated and any observed change in gene expression had a negligible impact on global transcriptomic signatures or downstream annotation of mesenchymal cell types. However, care should be taken when interpreting the expression of individual genes, especially those identified to be responsive to the isolation procedure.

We also assessed the biological informativeness of BMSC-OBs by comparing them to the same cells isolated directly from bone. Upon comparison of both scRNA-seq data sets, we found that the transcriptomic signatures of BMSC-OB cell types did not substantially differ from the cells isolated by Zhong and colleagues[156]. Nevertheless, differences between the two data sets were observed, namely the absence of early/intermediate mesenchymal progenitor (EMP, IMP) populations in the BMSC-OB data set, which is likely due to the maturation of LMPs beyond EMP/IMP cell stages

during the in vitro osteoblast differentiation. Importantly, these results indicate that individual cell types in BMSC-OBs are similar, in the context of transcriptomic signatures, to their counterparts in bone.

A valid concern of population-based scRNA-seq studies is cost associated with increasing scalability and sample throughput. Using BMSC-OBs, we address this challenge by pooling cells derived from multiple mice into a single sample for scRNA-seq. Because each DO mouse is genetically unique, we were able to assign each cell to a mouse based on coding variants. Our approach to genotype deconvolution employed Souporcell to cluster cells using genetic variants detected in scRNA-seq reads to effectively associate a "mouse-of-origin" to each single cell. Our ability to pool cells from multiple mice before library preparation and sequencing, then deconvolute the data bioinformatically, significantly reduced costs associated with generating scRNA-seq data (via 10× Genomics technology) by approximately a factor of five, which will make population-based experiments in hundreds of samples more cost efficient.

Highlighting significant variation in cell-type abundances is often valuable and provides insight into many biological contexts, such as differences between experimental conditions, patient samples, or tissues. Additionally, in future experiments with larger sample sizes, we will be able to correlate variation in cell-type proportions to bone phenotypes (eg, bone mass, bone strength, etc.) with effect sizes necessary to make informative conclusions. Although the sample size of our mouse cohort in this study was small ($n = 5$), we observed notable differences in cell-type frequencies between our mice. These differences likely reflect variation in cell-type composition of the starting BMSCs from each mouse and differences in the rate/efficiency of osteoblast differentiation

arising as a function of mouse-specific genotype and environmental effects. With this study serving as a proof-of-concept, BMSC-OBs feasibly permit scalability and increased sample throughput, which can enable population-level connections between cell-type abundance and many quantifiable phenotypes.

A limitation to our approach is that in vitro culture of BMSC-OBs may have an impact on the expression of certain genes; however, we show that the transcriptomic landscape of our BMSC-OBs are not vastly different from in vivo–derived BMSCs, suggesting that plastic adherence does not have a large "global" effect on the transcriptome. Another potential limitation is that marrow samples were frozen immediately after extraction for storage; however, in vitro culturing and osteogenic differentiation of the cells were successful thereafter, indicating the single freeze–thaw had minimal impact on transcriptome integrity and cell viability. Additionally, Zhong and colleagues[156] scraped and digested the endosteal surface for the extraction of cells, whereas our approach begins with a bone marrow flush and subsequent enrichment for BMSCs in vitro. Therefore, a limitation of the BMSC-OB model is that it does not capture all cell types relevant to bone, such as osteoclasts. However, it is important to note that in our CELLECT analysis, BMD heritability was not enriched in genes whose expression was more specific to osteoclasts from the Zhong and colleagues[156] data set. It is unclear why osteoclasts were not significant and may be due to cross-sectional measures of BMD being more so a product of peak bone mass and osteoblast-mediated bone accrual than bone loss, a process driven by osteoclasts, or the fact that these were likely immature osteoclasts as mature cells would be too large to be captured for sequencing. Interestingly, one of the strongest signals from the CELLECT analysis was

enrichment of BMD heritability in MALPs. MALPs express high levels of *Cxcl12*, a marker of *Cxcl12*-abundant reticular (CAR) cells[183]. Although data suggest the existence of two subsets of CAR cells, each with either osteogenic or adipogenic potential, Zhong and colleagues suggest that most CAR cells with adipogenic potential are MALPs[184]. Future studies will be needed to clarify the precise nature of these cell types, but the CELLECT analysis suggests that a subset of BMD associations impacts genes influencing BMD through their expression in MALPs.

Here, we described how the osteogenic differentiation of BMSCs can facilitate the generation of large-scale scRNA-seq data for mesenchymal lineage cells derived from the DO mouse population. Based on findings gathered here, the transcriptomic profiles generated from BMSC-OBs will serve as a valuable biological input for future genetic analyses. For example, cell type–specific, co-expression networks can be used as input to perform directed Bayesian network reconstruction and key driver analysis (KDA), as previously described in Al-Barghouthi and colleagues[19]. These subsequent analyses can aid in informing GWAS and highlighting putatively novel genes driving disease. We have demonstrated that the BMSC-OB model has the potential to facilitate more holistic genotype-to-phenotype investigations, which will aid in our understanding of the genetics of bone mass and lead to the identification of novel therapeutic targets to treat and prevent osteoporosis.

2.6. Acknowledgements

Research reported in this publication was supported in part by the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of

Health under award number R01AR68345 to CRF, MAH, and CJR and R01AR077992 to CRF.

2.7. Author Contributions

Luke J Dillard: Writing - review & editing; Visualization; Conceptualization; Methodology; Formal analysis; Data curation; Writing - original draft. Will T Rosenow: Conceptualization; Methodology. Gina M Calabrese: Methodology. Larry D Mesner: Methodology. Basel M Al-Barghouthi: Data curation. Abdullah Abood: Data curation. Emily A Farber: Methodology. Suna Onengut-Gumuscu: Conceptualization. Steven M Tommasini: Conceptualization. Mark A Horowitz: Conceptualization. Clifford J Rosen: Conceptualization. Lutian Yao: Data curation; Conceptualization. Ling Qin: Conceptualization. Charles R Farber: Writing - review & editing; Funding acquisition; Supervision; Conceptualization.

2.8. Disclosures

The authors declare that they have no conflicts of interest with the contents of this article.

2.9. Data Availability Statement

The data that support the findings of this study are openly available in GEO at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE152806, reference number GSE152806.

2.10. Chapter 2 - Main Figures

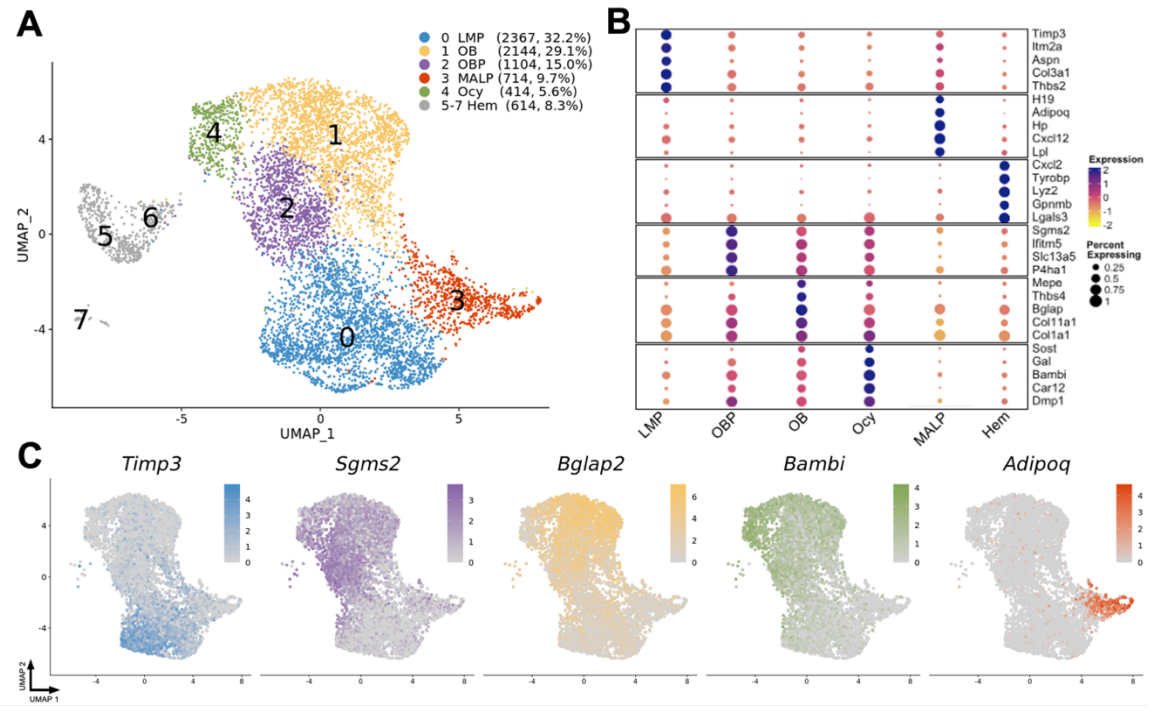


Figure 1. ScRNA-seq of BMSC-OBs identifies multiple cell-types.
(A) Uniform Manifold Approximation and Projection (UMAP) cell clusters of 7357 single BMSC-OBs isolated from five Diversity Outbred (DO) mice. Cell numbers and corresponding percentages are listed in parenthesis to the right of the annotated cluster name. LMP: = late mesenchymal progenitor cells; MALP: = marrow adipogenic lineage precursors; OBP: = osteoblast progenitor cells; OB: = osteoblasts; Ocy: = osteocyte-like cells; Hem: = Hematopoietic lineage cells. **(B)** Dot plot[185] of some of the most highly expressed genes for all annotated cell clusters. The size of the dots are proportional to the percentage of cells of a given cluster that express a given gene while the color of the dot corresponds to the scaled average gene expression. **(C)** Feature plots portraying the normalized expression of select marker genes associated with each cell cluster.

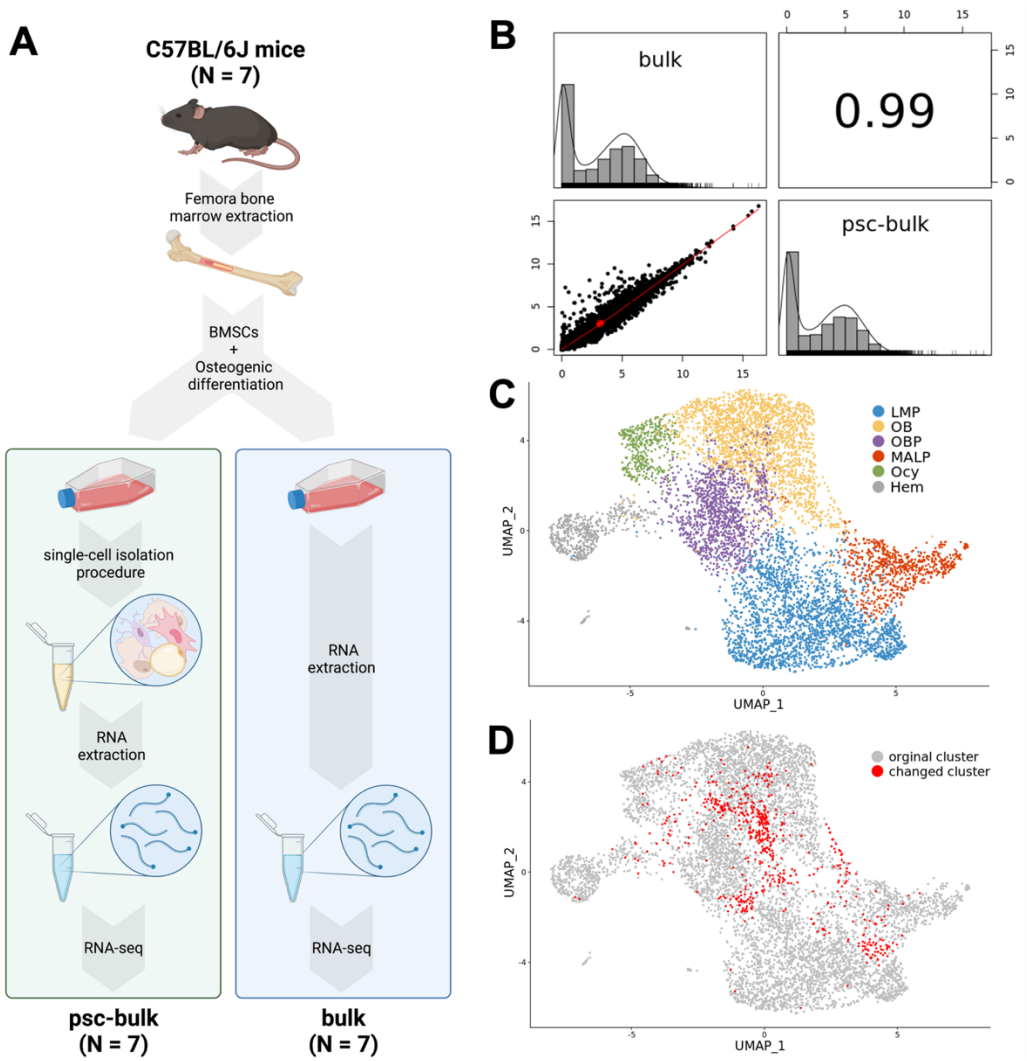


Figure 2. Liberation of single cells from a heavily mineralized matrix in vitro has minimal impact on transcriptomic signatures of BMSC-OBs.
(A) Flow chart diagram portraying the design of the bulk versus. pooled single cell-bulk (psc-bulk) experiment in C57BL/6J mice (N = 7). Cultured BMSC-OBs were harvested and underwent either immediate RNA extraction (bulk) or the single-cell isolation procedure, pooled, and then subsequent RNA extraction (psc-bulk). Extracted RNA from both conditions was sequenced via traditional RNA-seq. Created with BioRender.com**. (B)** RNA read counts for the bulk and psc-bulk gene expression profiles were converted to counts per million (CPM) values, log2-transformed, and the average for each gene was calculated across all samples within each group (bulk and psc-bulk). Correlation (R = 0.99, $p < 2.2\ 2 \times \times 10{-}16$) was performed using the subset of genes shared between the two profiles (N = 17,924). **(C)** ScRNA-seq UMAP clusters of BMSC-OBs derived from the five DO mice after removal of differentially expressed genes (identified from the psc-bulk vs. bulk experiment, 684 total genes) from the scRNA-seq count matrix. **(D)** Cells highlighted in red represent those that changed from their original cell cluster annotation as a result of removal of DEGs (8.1% of cells).

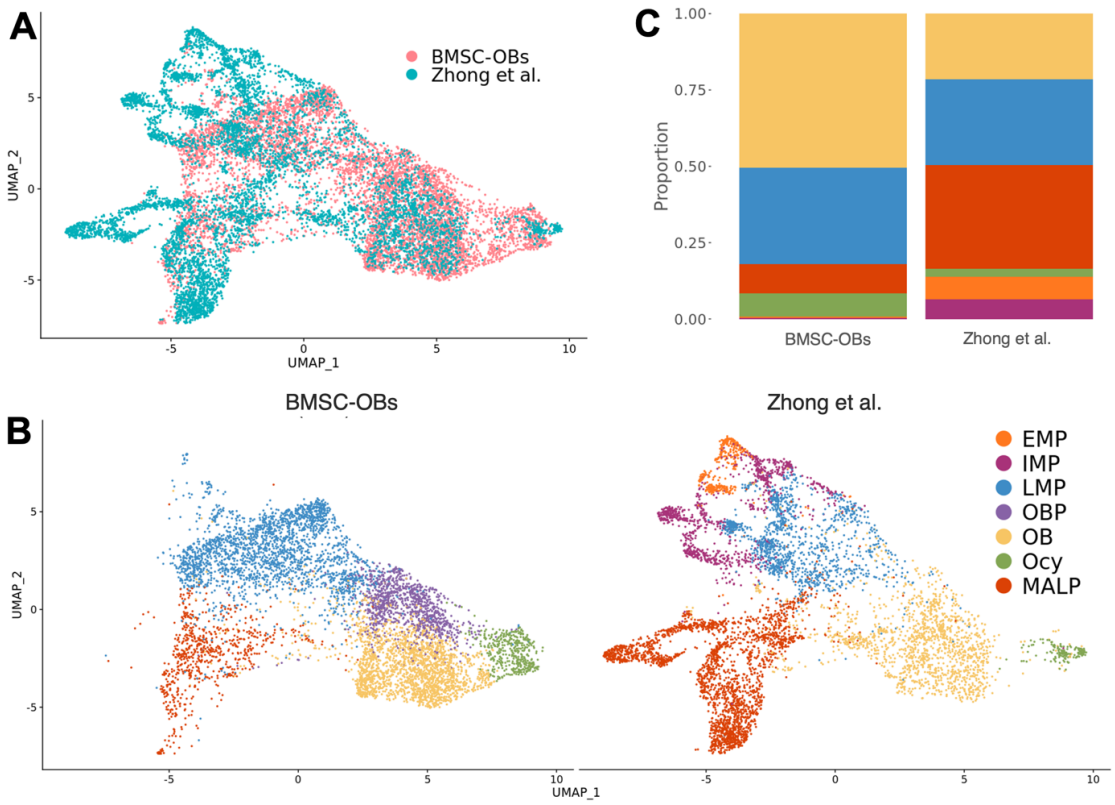


Figure 3. ScRNA-seq of BMSC-OB and scRNA-seq data derived from cells harvested in vivo cluster together and have few transcriptomic differences.
(A) Overlap of 13,310 single cells in UMAP space after integration of both the BMSC-OBs and Zhong and colleagues[156] scRNA-seq datasets. Integration was performed using Canonical Correlation Analysis (CCA) and using only the osteogenic and adipogenic lineage cells from each dataset as input. The integrated data was processed in the same fashion as the BMSC-OBs scRNA-seq data (Methods) and clustered at a resolution of 0.22 (Supplemental Fig. S5). **(B)** UMAPs of the integrated data and split based on dataset origin (BMSC-OBs or Zhong and colleagues[156]). Cells are labeled with their original cell annotations from either BMSC-OBs or Zhong and colleagues[156] datasets. **(C)** Bar chart representing the proportion of each annotated cell cluster in the integrated data based on dataset origin (BMSC-OBs or Zhong and colleagues[156]).

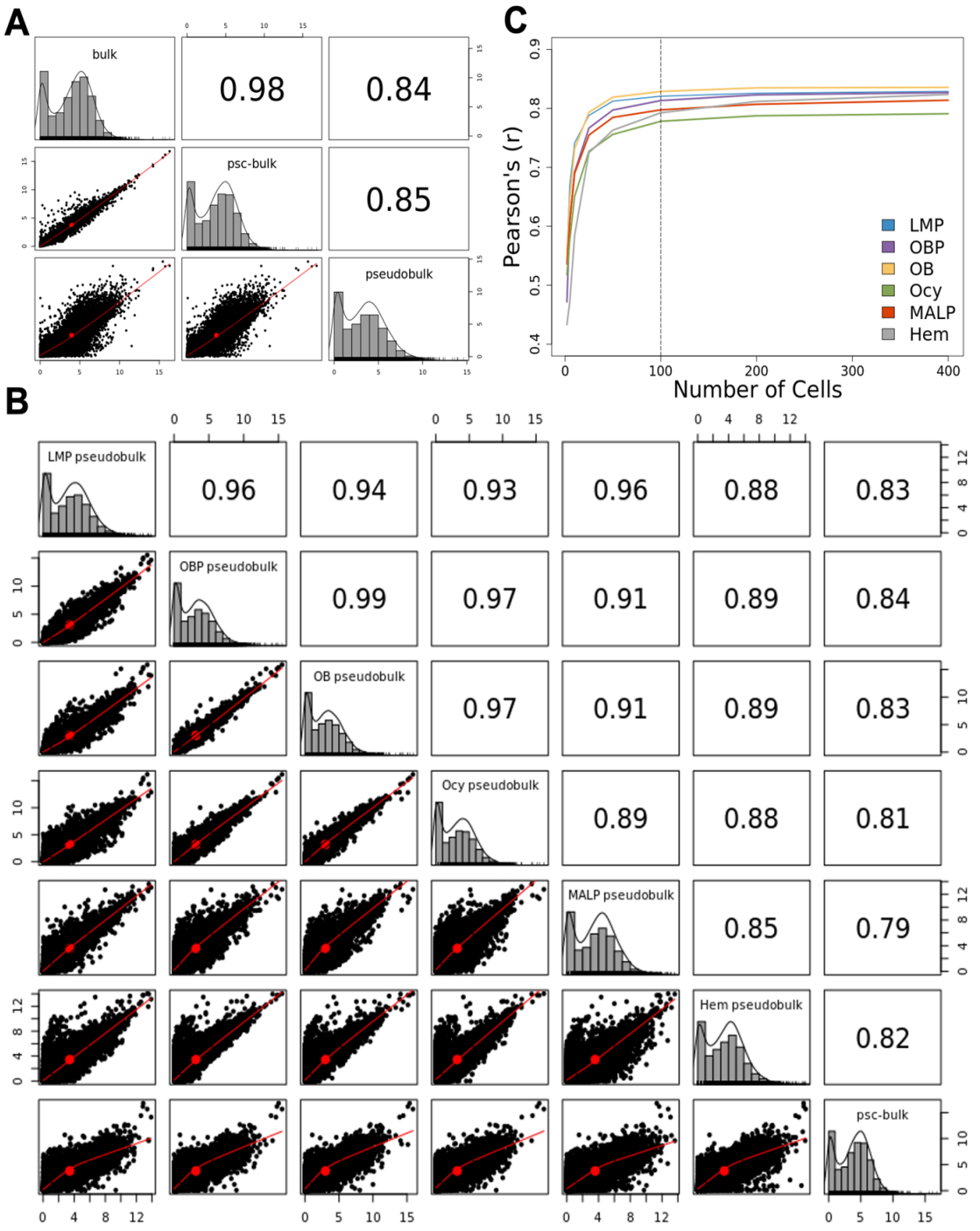

A
bulk
0.98
0.84
psc-bulk
0.85
pseudobulk
C
Pearson's (r)
Number of Cells
LMP
OBP
OB
Ocy
MALP
Hem
B
LMP pseudobulk
0.96
0.94
0.93
0.96
0.88
0.83
OBP pseudobulk
0.99
0.97
0.91
0.89
0.84
OB pseudobulk
0.97
0.91
0.89
0.83
Ocy pseudobulk
0.89
0.88
0.81
MALP pseudobulk
0.85
0.79
Hem pseudobulk
0.82
psc-bulk

Figure 4. Transcriptomic profiles of individual cell-types from scRNA-seq of BMSC-OBs are robust and representative of bulk RNA-seq data.
(A) Correlations between the bulk, pooled single cell-bulk (psc-bulk), and pseudobulk gene expression profiles. A pseudobulk (PB) profile was generated from the entire BMSC-OB scRNA-seq dataset by aggregating Unique Molecular Identifier (UMI) counts, converting to counts per million (CPM), and $\log_2$-transforming the counts. Counts for the PB, bulk, psc-bulk gene expression profiles were performed using the subset of genes shared between all three profiles ($N = 13{,}920$). **(B)** Correlations between the psc-bulk and cell-type specific PB gene expression profiles. Cell-type specific PB profiles were generated for individual cell clusters in the same fashion described above. **(C)** Correlations between psc-bulk and cell-type specific PB profiles generated using different numbers of sampled cells. Cell-type specific PB profiles were generated from random sampling of cells from the cell-type cluster. Eight samples were taken, ranging in size from 2 to 400 cells, and profiles were correlated to the psc-bulk profile.

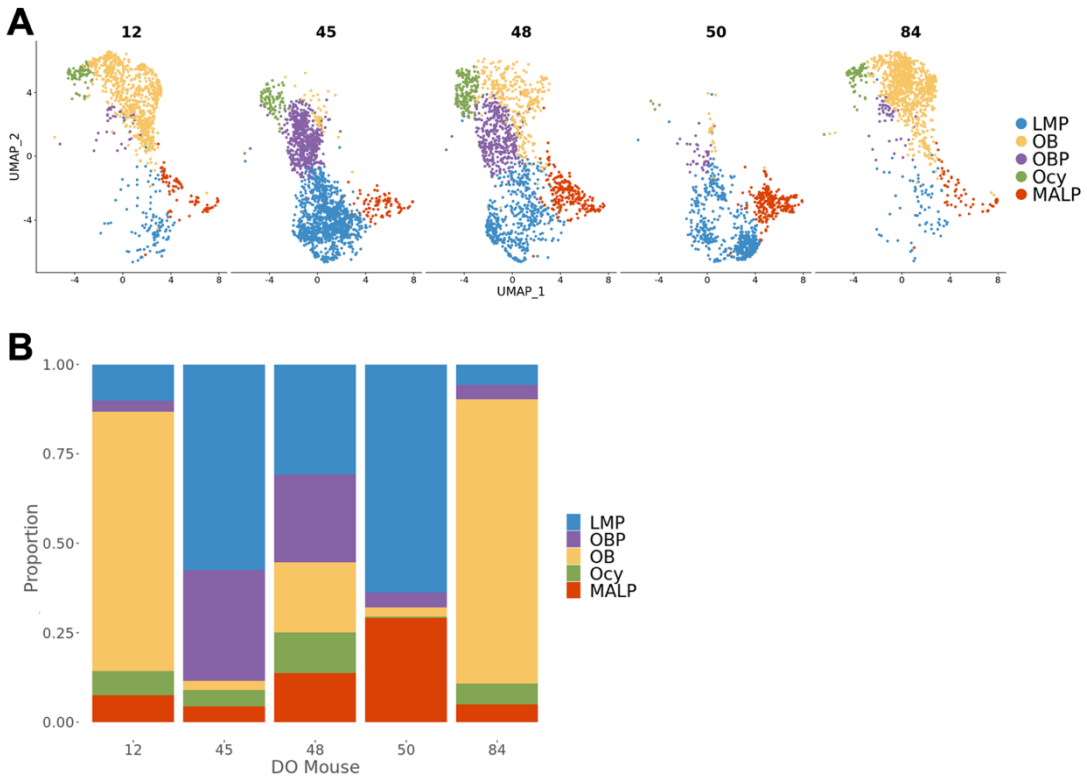


Figure 5. Cell-type frequencies captured by scRNA-seq are highly variable across individual DO mice.
(A) Uniform Manifold Approximation and Projection (UMAP) of cell clusters of the BMSC-OB scRNA-seq dataset split based on the five Diversity Outbred (DO) mice (12, 45, 48, 50, 84). **(B)** Stacked bar chart representing the proportion of each cell-type derived from each mouse.

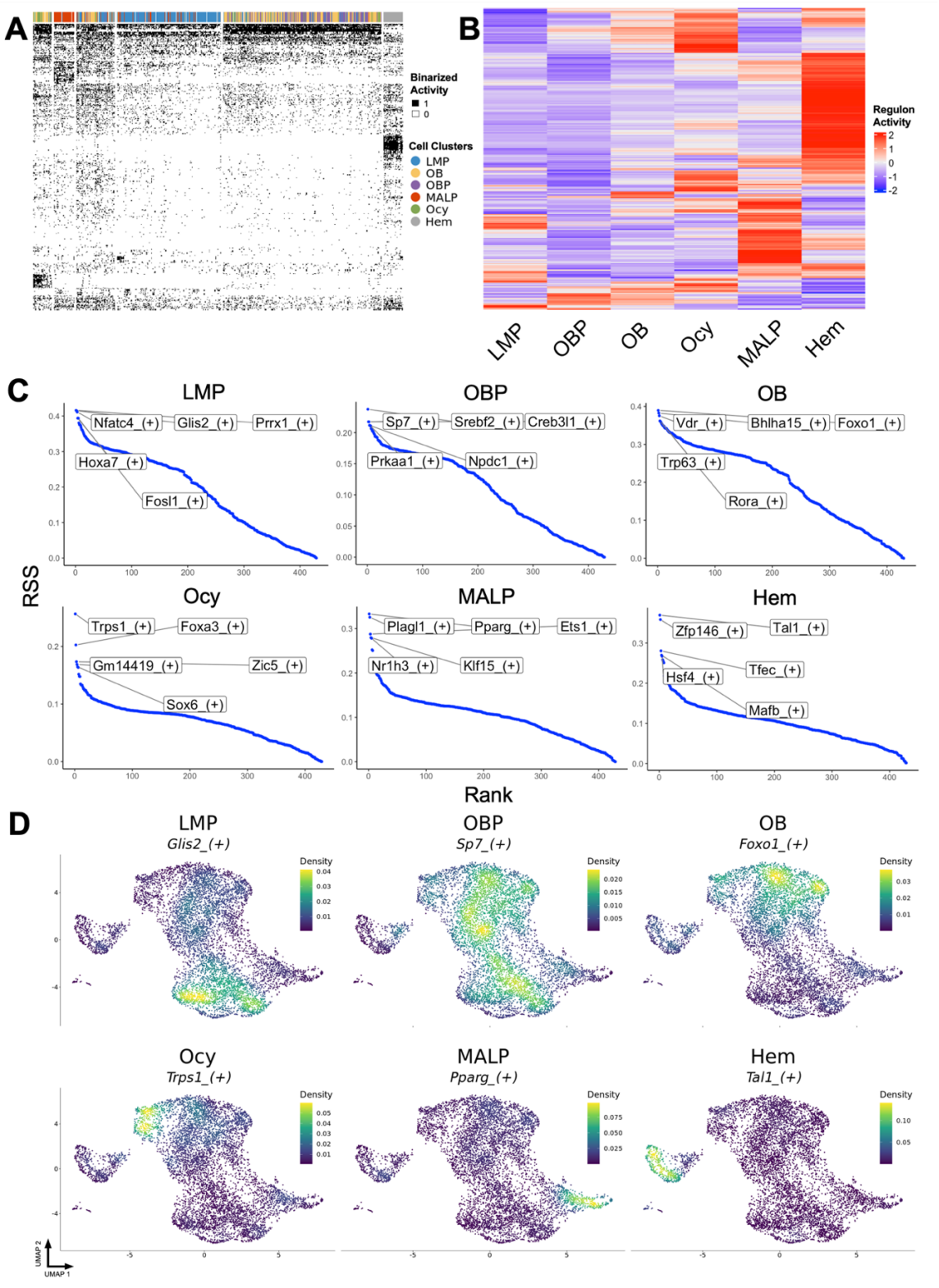

A
Binarized Activity
1
0
Cell Clusters
LMP
OB
OBP
MALP
Ocy
Hem
B
Regulon Activity
2
1
0
-1
-2
LMP
OBP
OB
Ocy
MALP
Hem
C
LMP
Nfatc4_(+)
Glis2_(+)
Prrx1_(+)
Hoxa7_(+)
Fosl1_(+)
OBP
Sp7_(+)
Srebf2_(+)
Creb3l1_(+)
Prkaa1_(+)
Npdc1_(+)
OB
Vdr_(+)
Bhlha15_(+)
Foxo1_(+)
Trp63_(+)
Rora_(+)
Ocy
Trps1_(+)
Foxa3_(+)
Gm14419_(+)
Zic5_(+)
Sox6_(+)
MALP
Plagl1_(+)
Pparg_(+)
Ets1_(+)
Nr1h3_(+)
Klf15_(+)
Hem
Zfp146_(+)
Tal1_(+)
Hsf4_(+)
Tfec_(+)
Mafb_(+)
RSS
Rank
D
LMP
Glis2_(+)
OBP
Sp7_(+)
OB
Foxo1_(+)
Ocy
Trps1_(+)
MALP
Pparg_(+)
Hem
Tal1_(+)
Density
UMAP 2
UMAP 1

Figure 6. SCENIC gene regulatory network (GRN) analysis reveals expected transcriptomic activity and validates the identities of cell-types in BMSC-OBs.
(A) Binarized heatmap SCENIC regulon activity results, where "1" indicates active regulons; "0" indicates inactive regulons. **(B)** Heatmap of SCENIC results portraying the scaled average for regulon activity in each annotated cell cluster, where the color key from blue to red indicates activity levels from low to high, respectively. (**C)** Plots of the top five regulons with the highest specificity score (RSS) for each cell cluster. RSS is quantified from 0 to 1, where "1" indicates the activity of a regulon is exclusively specific to one cell-type, while "0" indicates the lowest level of exclusivity. **(D)** Cell density plots portraying the regulon-weighted kernel density of select regulons for each cell cluster. Cell density is weighted by the activity of a given regulon in the single cells. Plots represent regulon activity by leveraging signal from cells that are more likely to have a given regulon active in their neighboring cells[186].

2.11. Chapter 2 - Tables

Table 1. CELLECT cell-type prioritization on all cell-types annotated in the BMSC-OBs and Zhong et al. (2020) scRNA-seq datasets.

scRNA-seq dataset	Cell-type	Beta	Beta SE	P-value
Zhong et al.	MALP	5.88×10^{-8}	1.84×10^{-8}	**6.92×10^{-4}**
Zhong et al.	OB	4.80×10^{-8}	1.56×10^{-8}	**1.05×10^{-3}**
Zhong et al.	Ocy	5.91×10^{-8}	2.15×10^{-8}	**3.03×10^{-3}**
BMSC-OBs	Ocy	5.70×10^{-8}	2.16×10^{-8}	**4.18×10^{-3}**
BMSC-OBs	MALP	4.86×10^{-8}	1.86×10^{-8}	**4.57×10^{-3}**
Zhong et al.	IMP	3.61×10^{-8}	1.68×10^{-8}	**1.57×10^{-2}**
Zhong et al.	LMP	3.09×10^{-8}	1.71×10^{-8}	**3.55×10^{-2}**
BMSC-OBs	OB	6.24×10^{-8}	3.56×10^{-8}	**3.98×10^{-2}**
Zhong et al.	EMP	2.86×10^{-8}	1.79×10^{-8}	5.52×10^{-2}
Zhong et al.	CH	1.96×10^{-8}	1.38×10^{-8}	7.81×10^{-2}
Zhong et al.	Mural	9.12×10^{-9}	1.66×10^{-8}	2.91×10^{-2}
BMSC-OBs	LMP	-4.57×10^{-9}	2.04×10^{-8}	5.89×10^{-2}
BMSC-OBs	OBP	-5.85×10^{-9}	1.77×10^{-8}	6.30×10^{-1}
Zhong et al.	Erythrocyte	-8.07×10^{-9}	1.73×10^{-8}	6.79×10^{-1}
Zhong et al.	Mono	-3.03×10^{-8}	1.60×10^{-8}	9.71×10^{-1}
Zhong et al.	MF	-2.98×10^{-8}	1.52×10^{-8}	9.75×10^{-1}
Zhong et al.	EC	-2.20×10^{-8}	1.10×10^{-8}	9.77×10^{-1}
Zhong et al.	B-cell	-3.47×10^{-8}	1.63×10^{-8}	9.83×10^{-1}
Zhong et al.	OC	-4.66×10^{-8}	1.55×10^{-8}	1
Zhong et al.	Granulo	-3.56×10^{-8}	9.95×10^{-9}	1
BMSC-OBs	Hem	-5.90×10^{-8}	1.36×10^{-8}	1
Zhong et al.	T-cell	-6.45×10^{-8}	1.35×10^{-8}	1
Zhong et al.	HSC	-6.28×10^{-8}	1.28×10^{-8}	1
Zhong et al.	GP	-5.44×10^{-8}	1.11×10^{-8}	1

Note: Beta is regression effect size estimate for given annotation. Beta SE is the standard error for the regression coefficient. The p value is the one-sided test (beta >0) association between bone mineral density genomewide association study signal heritability and each annotated cell type. Any p values <0.05 are in bold. B cell = B lymphocyte; CH = chondrocyte; EC = endothelial cell; EMP = early mesenchymal progenitor; GP = granulocyte progenitor; Granulo = granulocyte; Hem = hematopoietic lineage cells; HSC = hematopoietic stem cell; IMP = intermediate mesenchymal progenitor; LMP = late mesenchymal progenitor; MALP = marrow adipogenic lineage precursors; MF = macrophage; Mono = monocyte; Mural = mural cells; OB = osteoblast; OBP = osteoblast progenitor; Ocy = osteocyte; OC = osteoclast; T cell = T lymphocyte.

Chapter 3

Cell Type-Specific Network Analysis in Diversity Outbred Mice Identifies Genes Potentially Responsible for Human Bone Mineral Density GWAS Associations.

(*In preparation*)

Dillard LJ, Calabrese GM, Mesner LD, Farber CR

3.1. Abstract

Genome-wide association studies (GWASs) have identified many sources of genetic variation associated with bone mineral density (BMD), a clinical predictor of fracture risk and osteoporosis. Aside from the identification of causal genes, other difficult challenges to leveraging findings from GWAS include characterizing the roles of predicted causal genes in disease and providing additional functional context, such as the cell type predictions or biological pathways in which causal genes operate. Leveraging single cell transcriptomics (scRNA-seq) can assist in informing BMD GWAS by linking disease-associated variants to genes and providing a cell type context in which these causal genes drive disease. Here, we use large-scale scRNA-seq data from bone marrow–derived stromal cells cultured under osteogenic conditions (BMSC-OBs) from Diversity Outbred (DO) mice to generate cell type-specific networks and contextualize BMD GWAS-implicated genes. Using trajectories inferred from the scRNA-seq data, we identify networks enriched with genes that exhibit the most dynamic changes in expression across trajectories. We discover 21 network driver genes, which are likely to be causal for human BMD GWAS associations that colocalize with expression/splicing quantitative trait loci (eQTL/sQTL). These driver genes, including *Fgfrl1* and *Tpx2* (along with their associated networks), are predicted to be novel regulators of BMD via their roles in the differentiation of mesenchymal lineage cells. In this work, we showcase the use of single-cell transcriptomics from mouse bone-relevant cells to inform human BMD GWAS and prioritize genetic targets with potential causal roles in the development of osteoporosis.

3.2. Introduction

Osteoporosis is a complex disease characterized by low bone mineral density (BMD), bone fragility, and an increased risk of fracture[140]. BMD is a highly heritable trait and a significant clinical predictor of osteoporotic fracture[4,141]. Increasing our understanding of the genetic basis of osteoporosis and BMD is critical for the development of approaches for disease treatment and prevention. Genome-wide association studies (GWAS) have identified thousands of genetic variants putatively influencing BMD. The largest BMD GWAS to date discovered over 1,100 associations[10]. However, the challenge lies in pinpointing causal genes, which has impeded the translation of genetic findings into novel therapies.

A number of approaches exist to identify genes responsible for GWAS associations[29,147,149,150]. Most rely on population-based "-omics" data[145], which are scarce for human bone, to connect associations to causal genes. However, most approaches do not provide information on how causal genes impact "systems-level" function[146]. To address these limitations, we recently used co-expression networks generated from mouse bone transcriptomic datasets to inform BMD GWAS. The idea is simple – genes that play a central role in the regulation of a complex trait are often functionally-related and functionally-related genes are often co-expressed[41]. For example, we generated gene co-expression networks using RNA-seq data from mouse cortical bone to identify potential causal genes, such as *MARK3* and *SPTBN1*[43]. In a similar study, we generated networks on mouse calvarial osteoblasts using RNA-seq data and identified four novel GWAS-implicated genes associated with osteoblast differentiation and predicted regulators of BMD (e.g., *CADM1*, *B4GALNT3*, *DOCK9*, and *GPR133*)[44]. Recently, we extended our use

of co-expression networks and used a Bayesian approach to generate directed networks using cortical bone RNA-seq data from 192 Diversity Outbred (DO) mice. We discovered 19 novel genes (and their associated networks), such as *SERTAD4* and *GLT8D2*, which are likely causal for human BMD GWAS associations[19].

To date, our analyses have been solely reliant on informing GWAS using networks generated from heterogeneous bulk transcriptomics (RNA-seq) datasets from bone. Leveraging single-cell transcriptomics (scRNA-seq) data, however, offers the added benefit of resolving the transcriptomic profiles of discrete cell type populations. Using scRNA-seq data can provide additional context to inform GWAS by predicting the specific cell types in which causal genes and their associated networks operate; however, generating scRNA-seq data on bone-relevant cell types at the population-scale (i.e., hundreds of samples) is an essential prerequisite. In recent work, we characterized the BMSC-OB model (bone marrow-derived stromal cells cultured under osteogenic conditions) from a small cohort of mice and assessed its utility for the generation of population-scale scRNA-seq data[136]. The BMSC-OB model effectively enriches for mesenchymal lineage cells (e.g., mesenchymal progenitors, osteoblasts, osteocyte-like cells) at single-cell resolution. These cell type-specific transcriptomic profiles can be leveraged in network analyses to prioritize and infer the function of putatively causal GWAS genes, particularly in the context of mesenchymal cell differentiation. Here, we showcase the scalability of this model and generated scRNA-seq data of BMSC-OBs from 80 Diversity Outbred (DO) mice to generate cell type specific networks in order to inform BMD GWAS.

In this work, our goal was to prioritize and contextualize genes implicated by BMD GWAS using large-scale, scRNA-seq on bone-relevant cell types. We accomplished this

by using our previously established strategy of generating Bayesians networks[19]; however, here we do so in a cell type-specific fashion using scRNA-seq data. We subsequently prioritized networks based on their enrichment for genes exhibiting the most dynamic changes in expression across trajectories inferred from the scRNA-seq data, thus highlighting networks likely associated with the differentiation of BMSC-OBs. We then use these networks to prioritize genes with expression/splicing quantitative trait loci (eQTL/sQTL) which colocalize with BMD GWAS associations[29,37]. In doing so, this analysis provides additional support for a role of these genes in the regulation of BMD and highlights their potential roles in differentiation of cell types essential to bone tissue.

3.3. Results

3.3.1. BMSC-OBs derived from DO mice yield diverse cell types that are enriched for mesenchymal lineage cells

We isolated BMSCs from Diversity Outbred (DO) mice (N=75 from the current study and N=5 from Al-Barghouthi and colleagues[19] for a total of N=80) (N = 49 male and N = 31 females). The DO is a genetically diverse outbred mouse population[114,115]. We cultured BMSCs under osteogenic conditions and subsequently performed scRNA-seq, as described by Dillard and colleagues[136]. After stringent processing and quality control (Materials and Methods), the dataset consisted of 21,831 genes quantified across 139,392 total cells. We manually annotated 15 clusters ranging in size from 270 to 27,291 cells and identified cell types of the mesenchymal lineage as well as various other cell types (**Fig. 1A**, **Supplementary Table S1**).

Based on our prior BMSC-OB scRNA-seq study[136], we expected to identify a large proportion of mesenchymal cells and a smaller fraction of non-mesenchymal cell types. Consistent with this hypothesis, clusters associated with mesenchymal lineages accounted for 74.14% of all cells (**Fig. 1A**). These included mesenchymal progenitor cells (MPCs), late mesenchymal progenitors (LMPs), osteoblast progenitors (OBPs), two mature osteoblast populations (OB1 and OB2), osteocyte-like cells (Ocy), and marrow adipogenic lineage progenitors (MALPs). The non-mesenchymal cell-types observed included macrophages, monocytes, granulocytes, T-cells, B-cells, endothelial cells, and osteoclast-like cells (**Fig. 1A**). With regards to the mesenchymal cell types, the only differences in cell clusters relative to our previous report[136] were the presence of MPCs and two mature osteoblast clusters. Interestingly, MPCs did not have transcriptomic profiles similar to other mesenchymal progenitor cells previously identified by our group and Zhong and colleagues[156]. All other mesenchymal cell types demonstrated specific expression of canonical marker genes (**Fig. 1A, B**).

Upon comparing the two distinct osteoblast cell clusters, OB1 and OB2 (**Fig. 1A**), both clusters had ubiquitous expression of genes associated with mature osteoblasts (e.g., *Col1a1*, *Bglap, Sparc*, and *Ibsp*) (**Supplementary Table S1**). Interestingly, many of the "canonical" osteoblast markers were more highly expressed in OB1 compared to OB2 (**Supplementary Table S2**). A PANTHER[180] Gene Ontology (GO) analysis indicated that genes more highly expressed ($|\text{average } \log_2 FC| > 0.25$, N = 467) in OB2 relative to OB1 were enriched with genes associated with cellular response to hypoxia (GO:0071456, N = 20, $P = 2.30 \times 10^{-13}$) (**Supplementary Table S2, Supplementary Table S3)**. Additionally, we used CELLEX[178] to calculate gene expression specificity (ESμ), which are metrics

assigned to each gene to quantitatively assess the specificity of its expression in a given cell type (**Supplementary Table S4)**. We compared the top ESμ values for genes between OB1 and OB2 (**Supplemental Fig. 1**, **Supplementary Table S5**). Genes with high specificity in OB2 (ESμ > 0.8, N = 215) were also enriched for genes associated with cellular response to hypoxia (GO:0071456, N = 5, P = 2.27 x 10^{-3}) (**Supplementary Table S6).** Among the hypoxia-related genes identified in both GO queries, which included *Egln3*, *Ak4*, *Fndc1*, *Tbl2*, and *Mgarp,* they exhibited more specific expression in OB2 relative to OB1 (**Supplemental Fig. 1, Supplementary Table S7**).

We next assessed the variability in cell type frequencies across DO mice by quantifying the proportions of each annotated cell type of the mesenchymal lineage. All other clusters, which mainly consisted of immune cells of hematopoietic origin, were aggregated into one group (Hem) for each mouse. We observed high variability in the raw proportional abundances of cell types derived from each mouse (**Fig. 1C, Supplementary Table S8**). For example, the proportions of osteoblasts (OB1 and OB2) varied significantly among individual DO mice (**Fig. 1D**). All mice used in the current experiment had been extensively phenotyped for a wide range of bone traits (microCT, histomorphometry, biomechanical bone properties, etc.) as part of a previous genetic analysis of bone strength[19]. We correlated cell type frequencies with bone traits, however, none of the cell type proportions were strongly correlated with any bone trait (**Supplementary Table S9-10**).

3.3.2. Mesenchymal lineage cells are enriched in BMD heritability

The primary goal of this work was to prioritize and contextualize genes implicated by BMD GWAS. As a first step towards this goal, we sought to determine the individual cell types identified in this study that are the most relevant to the genetics of BMD. Using the BMD GWAS and the BMSC-OB scRNA-seq data, we performed a CELLECT[178] analysis to identify cell clusters enriched for BMD heritability and observed that all mesenchymal cells were significantly ($P < 0.05$) enriched for BMD heritability (**Fig. 1E**, **Supplementary Tables S11**). None of the non-mesenchymal cells identified were significant ($P > 0.05$) (**Fig. 1E**). As a result, our downstream efforts using these data focused solely on mesenchymal cell types to inform GWAS.

3.3.3. Generating mesenchymal cell type specific Bayesian networks to inform BMD GWAS

We have previously shown that network-based approaches using bulk RNA-seq are powerful tools for the identification of putative causal genes identified via BMD GWAS[19,43,44]. Here, our goal was to apply these same approaches using the BMSC-OB scRNA-seq data to prioritize and contextualize genes we previously identified as being putative regulators of BMD[29,37], such as those genes with human BMD GWAS associations that also colocalize with expression quantitative trait locus (eQTL; N=512) or splicing QTL (sQTL; N=732) in a tissue from Genotype-Tissue Expression (GTEx) project[26]. Genes identified in each study (or both) yield a list of high priority target genes (N = 1037). While GTEx does not currently contain data for bone tissue, eQTL can be shared across many tissues and may exert their effects in cell types resident to bone. Therefore, utilizing our

previous work, we hypothesized that generating cell type-specific networks would yield more biologically relevant representations of processes occurring within particular mesenchymal cell types. Additionally, by leveraging pseudo-temporal gene expression along inferred cell trajectories, our network analysis (**Fig. 2)** aims to identify driver genes of networks influencing BMD via their roles in mesenchymal cell differentiation.

Our network analysis begins by partitioning genes into groups based on co-expression by applying iterative weighted gene co-expression network analysis (iterativeWGCNA)[187] to each mesenchymal cell type (Step 1, **Fig. 2**). In total, 535 modules were identified from the BMSC-OB scRNA-seq data, and the number of modules identified for each mesenchymal cell cluster ranged from 26 to 153 with an average of 76 modules per cluster (**Supplementary Table S12, S13**). We sought to infer causal relationships between genes in each cell type-specific co-expression module and subsequently identify networks involved in processes relevant to BMSC-OB differentiation. To this end, we generated Bayesian networks for each co-expression module, thus enabling us to model directed interactions between co-expressed genes based on conditional independence[19] (Step 2, **Fig. 2**).

3.3.4. Identifying putative drivers of mesenchymal cell differentiation

We hypothesized that many genes impacting BMD do so by influencing osteogenic differentiation or possibly bone marrow adipogenic differentiation, as suggested by the CELLECT analysis above. Therefore, the next step of our network analysis was to identify cell type-specific Bayesian networks enriched for genes potentially driving mesenchymal differentiation (Step 3, **Fig. 2**). To accomplish this, we first performed a pseudotime

trajectory analysis to infer paths of differentiation in the mesenchymal lineage cells. We resolved three pseudotime trajectories (two osteogenic, one adipogenic) originating from the MPC cell cluster and ending in either Ocy, OB2, or MALP cell fates (**Fig. 3A**). It is important to note that given the identification of multiple skeletal stem cells[154,188–190], we do not view these trajectories as lineages, but rather "differentiation paths" (progenitor to mature and/or terminally differentiated cells) that are likely traversed by different subsets of skeletal stem cells.

To identify genes likely impacting BMSC-OB differentiation, we used tradeSeq to identify genes that exhibit dynamic changes in expression along pseudotime[93]. Prior to performing the tradeSeq analysis, we parsed the pseudotime trajectories into regions that encompass cells associated with each cell type along their respective trajectories (**Fig. 3B**). We defined multiple cell type boundaries (nine in total) using pseudotime values, which represent points along a trajectory. The tradeSeq analysis was performed for each boundary (**Supplementary Table S14**). For example, trajectories bifurcate in the LMP cell cluster (**Fig. 3A**); therefore, cells belonging to the LMP cluster can map to adipogenic and/or osteogenic trajectories depending on their placement along pseudotime. Between a cell type boundary, cells spanning a specific cluster (e.g., LMP) and mapping to a specific lineage (e.g., osteogenic trajectory) are used as input to analyze gene expression between the start and end points of the cell type boundary (e.g., LMP_to_OBP). We analyzed gene expression within the established cell type boundaries for all trajectories and identified genes that exhibit the most significant differences in expression between the start and end points of the cell type boundaries. The total number of significant trajectory-specific tradeSeq genes ($P_{adj} \leq 0.05$) ranged from 87 to 1697 across the 9 cell type boundaries

(**Supplementary Table S14, S16-18**). The expression of representative marker genes for all cell types as a function of pseudotime were consistent with boundaries defined for each cell type (**Fig. 3C**).

To provide further support that tradeSeq-identified genes are enriched for genes involved in differentiation, we performed a cell type-specific expression quantitative trait locus (eQTL) analysis for each mesenchymal cell type. We identified 563 genes (eGenes) regulated by a significant *cis*-eQTL in specific cell types of the BMSC-OB scRNA-seq data (**Supplementary Table S19**). In total, 73 eGenes were also tradeSeq-identified genes in one or more cell type boundaries along their respective lineages (**Supplementary Table S14**).

We hypothesized that if tradeSeq genes were responsible for driving mesenchymal differentiation, then the eQTLs that perturb their expression would also impact the proportion of cells at different stages along the cell trajectories. Despite being significantly underpowered for this analysis due to our relatively small sample size (N = 80), we identified two cell type-specific eGenes where the genotype responsible for the *cis*-eQTL effect was also associated with cell type proportions. The first of these genes was Pyruvate Kinase, muscle (*Pkm*), which was identified as a significant global tradeSeq gene (P_{adj} = 8.35 x 10^{-8}; **Supplementary Table S14, S15**) associated with the transition from LMPs to OBPs along an osteogenic trajectory (**Fig. 4A**). Moreover, *Pkm* serves as an eGene in the LMP cell cluster (LOD = 9.72; **Fig. 4B, Supplementary Table S19**). Mice inheriting at least one *Pkm* PWK allele at this locus (N = 15) demonstrated lower *Pkm* expression (**Fig. 4C**) and a notable reduction in mature osteoblasts (OB1) and osteocyte-like cells (Ocy)

proportions (P = 0.030 and P = 0.026, respectively), while LMP proportions were unaffected (**Fig. 4D, Supplementary Table 20**).

Similarly, S100 calcium binding protein A1 (*S100a1*) was an OBP to OB1 transition tradeSeq gene (P_{adj} = 0.023; **Fig. 4A, Supplementary Table S14, S15**) and an eGene in the OBP cell cluster (LOD = 10.12; **Fig. 4B, Supplementary Table S19**). Mice inheriting at least one 129 allele at this locus (N = 30) had higher *S100a1* expression, while the opposite was observed for mice inheriting NZO alleles (N = 14) (**Fig. 4C**). Additionally, 129 mice showed a significant decrease in LMP proportion and increase in OB1 proportion (P = 0.008 and P = 0.016, respectively) (**Fig. 4D, Supplementary Table S20**), while no significant differences were observed in cell type proportions among NZO mice (**Supplementary Fig. 3, Supplementary Table S20**). These data support the role of tradeSeq-identified genes in the differentiation of mesenchymal cell types.

3.3.5. Identification of differentiation driver genes (DDG):

We hypothesized that tradeSeq-identified genes involved in BMSC-OB differentiation would be highly connected and play central roles in various cell type-specific Bayesian networks. In order to test this hypothesis and discover BMSC-OB differentiation genes potentially responsible for BMD GWAS associations, the next step of our network analysis leveraged the trajectory-specific tradeSeq genes identified for each cell type boundary (**Supplementary Table S16-18**) to identify differentiation driver genes (DDGs) (Step 3, **Fig. 2**). We identified DDGs by extracting subnetworks (i.e., large 3-step neighborhood; see Methods) for each gene in each cell type-specific Bayesian network and identifying those subnetworks enriched ($P_{adj} < 0.05$) for lineage-specific tradeSeq genes for

the cell type boundary. The analysis identified 408 significant DDGs (**Supplementary Table S21, S22-24**). We performed a PANTHER[180] GO analysis for the cell type boundaries yielding a sufficient number of DDGs and found that DDGs for the osteogenic cell type boundaries (LMP_to_OBP, OBP_to_OB1, OBP_to_OB2) were enriched for genes associated with the cell cycle (GO:0007049; N = 23, 18, 23; P = 1.12 x 10^{-6}, 1.29 x 10^{-13}, 1.0 x 10^{-14}, respectively) (**Supplementary Table S25-27**). The DDGs for the adipogenic cell type boundary (LMP_to_MALP, MALP_to_end) were enriched for genes associated with extracellular matrix organization (GO:0030198; N = 10; P = 1.62 x 10^{-7}) and lipid metabolic processes (GO:0006629; N = 25; P = 1.83 x 10^{-11}), respectively (**Supplementary Table S28-29**). Across all 408 DDGs, 49 were identified in one or more cell type boundaries as genes with a significant alteration ($P < 0.05$) of whole-body BMD when knocked-out/down in mice, as reported by the International Mouse Knockout Consortium (IMPC)[45] (**Supplementary Table S22-24**).

We used our previously generated list of potentially causal BMD GWAS genes (N=1037) to subsequently prioritize the DDGs (Step 4, **Fig. 2**). Of the 408 DDGs, 21 DDGs in one or more cell type boundaries were genes that have BMD GWAS associations that colocalize with sQTL/eQTL. The majority of these DDGs were identified in LMPs along both the osteogenic (LMP_to_OBP) and adipogenic (LMP_to_MALP) trajectories (N = 10 and 6, respectively; **Supplementary Table S21, S30**). The remaining DDGs were identified in OBPs along both osteoblast trajectories (OBP_to_OB1, OBP_to_OB2; N = 1 and 3, respectively) and MALPs (MALP_to_end; N = 6). Additionally, 3 of the 21 DDGs are IMPC genes that exhibit a significant alteration of BMD (**Supplementary Table S21, S30**).

3.3.6. Network analysis predict Fgfrl1 and Tpx2 as novel regulators of BMD:

Of the 21 prioritized DDGs and their associated networks, we identify two DDGs that putatively impact human BMD via their roles in LMP differentiation along either an adipogenic (*Fgfrl1*) or osteogenic (*Tpx2*) trajectory. Based on our previous work, both *Fgfrl1* (fibroblast growth factor receptor-like 1) and *Tpx2* (TPX2 microtubule nucleation factor) were identified as genes with significant human BMD GWAS associations that also colocalized with eQTL identified in the cultured fibroblast and Testis GTEx tissues, respectively[29]. The *Fgfrl1* network was enriched for tradeSeq-identified genes (N = 6 genes, P_{adj} = 7.5 x 10^{-3}) for LMPs along an adipogenic trajectory (**Fig. 5A**). The *Tpx2* network was enriched for tradeSeq-identified genes (N = 9 genes, P_{adj} = 5.7 x 10^{-7}) for LMPs along an osteogenic trajectory (**Fig. 5B**). An increase in the expression of all tradeSeq-identified genes was identified for each network (**Fig. 5C-D, Supplementary Table S16, S18**); their expression patterns were consistent with the cell type boundaries in which they were identified via tradeSeq (**Fig. 5C-D**). Additionally, *Tpx2* exhibited a significant alteration of BMD in both male and female mutant mice (Genotype P-value = 1.03 x 10^{-3}) from IMPC (**Fig. 5E**). Four of the genes in the *Tpx2* network were kinesin family (*Kif*) motor protein genes[191]: *Kif4, Kif11, Kif15, Kif23*. In the *Fgfrl1* network, many genes can be associated with adipocyte function (e.g., *Lpl*, *Plpp3*, *Igfbp4*)[192–194] and the maintenance of cilia (e.g., *Cfap100*, *St5* (*Denn2b*), *Mark1*)[195–197].

3.4. Discussion

BMD GWAS has been successful at identifying thousands of SNPs associated with disease; however, the identification of the causal genes and defining their functional role in disease remains challenging. The integration of "-omics" data, particularly transcriptomics, can assist in overcoming this challenge. Leveraging transcriptomics data has proven invaluable to informing GWAS, as demonstrated in studies that use this data to perform eQTL mapping, transcriptome-wide association studies (TWASs), and co-expression/gene-regulatory network prediction. Genetic variation impacting causal genes are often colocalized with GWAS associations, thus providing a potential mechanism through which disease manifests via perturbations in causal gene function or expression. While bulk RNA-seq data has been the foundation of such analyses, leveraging scRNA-seq data can provide valuable biological context by predicting the cell type in which causal genes are affected. To inform BMD GWAS, the generation of population-scale transcriptomics data at single-cell resolution in bone-relevant cell types can assist in the discovery of novel gene targets. Here, we leverage our previously established BMSC-OB model and perform scRNA-seq on 80 DO mice to generate single-cell transcriptomics data of mesenchymal cell types relevant to bone. Using this scRNA-seq data, our goal was to prioritize putative causal genes and provide a biological context in which GWAS-implicated genes potentially cause disease, at cell type-specific resolution. Through our temporal gene expression and network analyses, we identified 21 networks governed by predicted differentiation driver genes (DDGs) with a corresponding human BMD GWAS association colocalizing with eQTL/sQTL in a GTEx tissue.

We demonstrate that the BMSC-OB model serves as an effective method to enrich mesenchymal lineage cells, particularly bone-relevant cells. We characterized cells from 80 mice and identified both osteogenic and adipogenic cells derived from the mesenchymal lineage, such as two populations of osteoblasts (OB1 and OB2), osteocyte-like cells (Ocy), and MALPs. Our trajectory inference analysis identified three distinct trajectories in which mesenchymal progenitor cells give rise to both osteogenic and adipogenic trajectories, thus portraying biologically relevant and known paths of differentiation of mesenchymal progenitor cells. Temporal gene expression was analyzed along each trajectory, in a cell type-specific fashion, to identify genes that were changing the most as a function of pseudotime (tradeSeq-identified genes). Subsequent *cis*-eQTL analysis indicated that the expression of some tradeSeq-identified genes were also associated with DO haplotype at the eGene locus, such as *Pkm* and *S100a1*. Further, the DO haplotype responsible for the eQTL for these two eGenes could also be associated with the relative proportion of cell types. While instances such as these were rare to identify, they illustrate that the potential consequence of genetic variation impacting the expression of tradeSeq-identified genes may be observed in the abundances of certain cell types. Nevertheless, these results indicate a role of tradeSeq-identified genes in the process of differentiation.

To inform BMD GWAS, we utilized the scRNA-seq data in a network analysis to contextualize causal genes (and their associated network) by predicting the cell types through which these genes are likely acting. Towards this goal, we generated cell type-specific Bayesian networks from our BMSC-OB scRNA-seq data. Our approach was similar to our previous network analyses where bulk RNA-seq data was leveraged to identify genes with strong evidence of playing central roles in networks[19,43,44]. In contrast,

here we utilized scRNA-seq data to identify DDGs and prioritize networks based on the likelihood ($P_{adj} < 0.05$) that they are involved in the differentiation of mesenchymal lineage cells (based on network connections enriched for tradeSeq-identified genes determined from inferred trajectories). Leveraging our previous work[29,37], we prioritized DDGs if they were genes with BMD GWAS associations colocalizing with human eQTL/sQTL in a GTEx tissue. Together, a gene being both a DDG and having BMD GWAS associations that colocalize with eQTL/sQTL is strong support of causality.

We identified 21 DDGs and associated networks, some of which with little to no known prior connection to bone. We contextualize these causal genes and their networks by not only providing cell type predictions in which they likely operate, but also providing information regarding the biological processes they likely affect. For example, the *Tpx2* network was identified in LMPs differentiating along an osteogenic trajectory. *Tpx2* is a microtubule nucleation factor that interacts with spindle microtubules during cellular division[198]. In our previous study, *Tpx2* was identified as a gene that has BMD GWAS associations that colocalize with eQTL in the Testis GTEx tissue[29]. While GTEx does not maintain bone tissue, eQTL are shared across many tissues; therefore, non-bone eQTL may exert their effects in cell types associated with bone, such as LMPs, and evidence of a human eQTL effect indicates that genetic variation can modulate the expression of *Tpx2*. Additionally, when knocked out by IMPC, *Tpx2* exhibited a significant increase in whole body BMD (excluding the skull); therefore, the regulation of this gene may have a role in BMD. In the surrounding neighborhood of the *Tpx2* network, other genes are associated with cellular division as well, such as Topoisomerase 2A (*Top2a*) and the kinesin family (*Kif*) genes[191,199]. Taken together, these results indicate a potential role of *Tpx2* as a

mediator of BMD via its role in microtubule maintenance during the expansion of osteogenic cell populations.

Additionally, the *Fgfrl1* network was identified in LMPs differentiating along an adipogenic trajectory. Fibroblast growth factor receptor-like 1 (*Fgfrl1*) is presumed to function as a decoy receptor that regulates FGF ligands[200]. Our previous study also identified *Fgfrl1*, which has BMD GWAS associations that colocalize with eQTL in the cultured fibroblasts GTEx tissue[19]. In the surrounding neighborhood of the *Fgfrl1* network, *Lpl*, *Plpp3*, *Igfbp4* have well-established roles in adipocyte function and metabolism[192–194]; however, other genes can be associated with cilia function, such as *Cfap100*, *St5* (*Denn2b*), *Mark1*[195–197]. Interestingly, the maintenance of cilia is essential to the function of both LMPs and pre-adipocytes (MALPs) while mature adipocytes lack cilia[201]. Therefore, the modulation of ciliogenesis and/or cilia function may coincide with *Fgfrl1* signaling. Additionally, given the prioritization of MALPs in the CELLECT analysis and the well-established inverse relationship between marrow adiposity and BMD[131,132], skewed balance of LMP differentiation toward adipogenic cell fates may affect BMD. In summary, the *Fgfrl1* network harbors genes involved in adipogenic function, including cilia, which may contribute to LMP differentiation along an adipogenic trajectory. Together, these results indicate a potential role of *Fgfrl1* as a mediator of BMD via its role in adipogenic differentiation and maintenance of cilia.

Analyses performed here are not without limitations to consider. A pitfall of scRNA-seq is the sparsity of the resulting data, which yields an increased frequency of zero values for the expression of some genes in a proportion of cells, also known as “drop-outs”[97]. While statistical approaches can be employed to impute missing data, the accuracy

of such methods and whether or not the resulting improvement in transcriptomic signal recovery is enough to warrant such intervention is contentious[99,100]. However, this issue may be partially offset given the larger scale of the scRNA-seq performed in this study and the average expression approach performed for network and eQTL analysis. An additional limitation is that the BMSC-OB model does not capture osteoclasts, another cell type associated with bone tissue. Importantly, results from our CELLECT analysis indicate that BMD heritability was not enriched for genes whose expression was more specific to osteoclast-like cells. Although we likely captured immature osteoclasts in our model, as mature cells would be too large to be captured for sequencing, the low prioritization of the osteoclast-like cells may due to BMD being a product of osteoblast-mediated bone accrual than bone loss via osteoclasts. Lastly, while our study employed 80 DO mice, the issue of statistical power is still a limitation; however, we demonstrate that the BMSC-OB model is amenable to high throughput and the inclusion of hundreds of mice, thus statistical power will be improved in future studies.

In summary, we showcase the use of large-scale scRNA-seq data to inform GWAS by performing a network analysis to contextualize BMD GWAS associations. Through the use of various single-cell analyses, we have expanded upon our understanding of the genetics of BMD. Our work exemplifies the power of single-cell transcriptomics coupled with systems genetics to discover the genetic determinants of disease.

3.5. Methods

3.5.1. Sample preparation and scRNA-seq

We prepared our samples in the same fashion as performed previously in Al-Barghouthi and colleagues[19]. In brief, bone marrow was extracted from the femurs of initially 80 DO mice. BMSCs were grown to confluence after 3 days of incubation in 48-well plates and then underwent *in vitro* osteoblast differentiation for 10 days with osteogenic differentiation media (alpha MEM, 10% FBS, 1% pen/strep, 1% glutamax, 50 μg/μL ascorbic acid [Sigma, St. Louis, MO, USA], 10 nM B-glycerophosphate [Sigma], 10 nM dexamethasome [Sigma]). After differentiation, single cells were liberated from mineralizing cultures via incubations with 60 mM ethylenediaminetetraacetic acid pH 7.4 (EDTA [Thermo Fisher Scientific], made in DPBS), 8 mg/mL collagenase (Gibco) in HBSS/4 mM CaCl2 (Fisher), and 0.25% trypsin–EDTA (Gibco). After single-cell isolation, cells from mice were pooled into groups containing cells from five mice total and concentrated to 800 cells/μL in PBS supplemented with 0.1% BSA (bovine serum albumin). Pooled single cells were prepared for sequencing using the 10× Chromium Controller (10× Genomics, Pleasanton, CA, USA) with the Single Cell 3' v2 reagent kit, according to the manufacturer's protocol. Libraries were sequenced on the NextSeq500 (Illumina, San Diego, CA, USA).

3.5.2. scRNA-seq analysis pipeline

The data was subsequently processed using the 10× Genomics Cell Ranger toolkit (version 5.0.0) using the GRCm38 reference genome[160]. Using Seurat[161] (version 4.1.0), a combined Seurat object containing all cells was generated with the inclusion of features

detected in at least three cells and cells with at least 200 features detected. We used Souporcell[162] (version 2.0.0) to deconvolve the genotypes of all mice and to remove doublet cells. Cells were assigned to their associated DO mouse by making a pairwise comparison between allele calls made by the shared variants captured between Souporcell and GigaMUGA genotype arrays generated for all mice in the cohort, as previous performed in Dillard and colleagues[136]. We filtered out cells with more than 6200 reads and less than 400 reads, as well as those cells with more than 10% mitochondrial reads. Further, cells were removed if they expressed greater than 20% *Rpl* and 15% *Rps* reads, which equates to cells approximately exceeding the 98 percentile. After filtering, 139,392 cells remained and the resulting object underwent standard normalization, scaling, and the top 3000 features were modeled from a variance stabilizing transformation (VST) using the Seurat. Cell-cycle markers based on Tirosh and colleagues[163] were regressed out using the "CellCycleScoring" and scaling functions. For subsequent dimensionality reduction, 15 principal components (PCs) were summarized. Then, a kNN ($k = 20$) graph was created and the Louvain algorithm was used to cluster cells at a resolution of 0.5. Annotation of cell-type clusters was performed manually based on differential gene expression analysis using the Seurat "FindAllMarkers" function (**Supplementary Table S1**).

For subsequent WGCNA and eQTL mapping, transcriptomic profiles for each cell type cluster were generated for each sample using a mean expression approach, as performed similarly by others[202,203]. For each sample contributing at least 5 cells to a given cluster, unnormalized unique molecular identifier (UMI) counts of gene expression for all cells in the cluster for the sample were averaged and then rounded to the nearest hundredth decimal place. A total of 80, 80, 77, 67, 50, 76, 80 mice contributed enough cells to the

MPC, LMP, OBP, OB1, OB2, Ocy, and MALP cell type clusters, respectively. Genes with non-zero expression values in fewer than 15 samples were removed. A total of 11971, 15162, 14857, 13674, 13825, 14136, and 14534 genes remained for the MPC, LMP, OBP, OB1, OB2, Ocy, and MALP clusters, respectively. Samples were normalized by computing CPMs (counts per million) without log transformation for each gene using edgeR[204] (version 4.0.7), then transformed via VST using DESeq2[170] (version 1.42.0), and quantile normalized using preprocessCore (version 1.60.2).

3.5.3. Trajectory and tradeSeq Analysis

Trajectory inference analysis was performed using Slingshot[87] (version 1.8.0) on the mesenchymal lineage cell clusters (seven total) of the BMSC-OB scRNA-seq data. The starting cluster was set as the MPC cluster and trajectories were inferred using 15 PCs. TradeSeq[93] (version 1.4.0) was used to analyze gene expression along the trajectories by fitting a negative binomial generalized additive model (NB-GAM) to each gene using the "fitGAM" function with nknots = 10, which was determined by using the "evaluateK" function. Prior to performing the tradeSeq analysis, the scRNA-seq data was downsampled to reduce the size of the dataset to approximately 10,000 cells (sampled at random across all seven clusters).

All cell type boundaries were established to encompass on average 78% of cells of a cell cluster (**Supplementary Table S14**). To identify genes significantly changing between boundaries, we first performed a global test with tradeSeq to compare gene expression between lineages (two osteogenic, one adipogenic) in order to highlight genes exhibiting a significant difference in expression using the "startVsEndTest" function

(**Supplementary Table S14, S15**). Next, we performed tradeSeq to compare gene expression within each trajectory to highlight genes with a significant difference in expression between boundaries in a trajectory-specific fashion using the "startVsEndTest" function (**Supplementary Table S14, S16-18**). All tests were performed with the $\log_2$ fold change threshold (l2fc) = 0.5. For all global and lineage-specific tests, the P-values associated with each gene were adjusted to control the false discovery rate using the "p.adjust" function from the stats (version 4.2.1) R package and genes were filtered to include those with a $P_{adj} < 0.05$.

3.5.4. CELLECT Analysis

CELLECT[178] (CELL-type Expression-specific integration for Complex Traits) (version 1.1.0) was used to identify likely etiologic cell types underlying complex traits of both the BMSC-OBs scRNA-seq data (**Fig. 1E, Supplementary Table S11**). CELLECT quantifies the association between the GWAS signal and cell type expression specificity using the S-LDSC genetic prioritization model[179]. Summary statistics from the UK Biobank eBMD and Fracture GWAS (Data Release 2018) and cell type annotations from each scRNA-seq data set were used as input. Cell type expression specificities were estimated using CELLEX[178] (CELL-type EXpression-specificity) (version 1.2.1) (**Supplementary Table S4**).

3.5.5. WGCNA

Cell type-specific mean expression matrices (as obtained above) were used as input to generate signed co-expression network modules (**Supplementary Table S12-13**). IterativeWGCNA[187] (version 1.1.6) was used from a Singularity container built from a Docker hub image[205]. A soft threshold (power) of 14, which exceeded a R^2 threshold of 0.85 for all cell type clusters, was selected for module construction (**Supplementary Fig. 2**). Modules were generated using iterativeWGCNA with default parameters for the "blockwiseModules" function, a minimum module size of 20 genes, minCoreKME = 0.7, and minKMEtoStay = 0.5.

3.5.6. Bayesian network learning

Bayesian networks were learned from each of the cell type-specific modules of co-expressed genes with the bnlearn (version 4.8.3). Gene expression matrices containing the genes for each module were used as input to the "mmhc" function which employs the Max-Min Hill Climbing algorithm (MMHC) algorithm[52] to learn the underlying structure of the Bayesian network. From the generated networks, igraph (version 1.6.0) was used to resolve 3-step neighborhoods[206]. Nodes (genes) that were unconnected to a neighborhood or connected to only one neighbor were removed. Neighborhoods were filtered to include those with a size greater than 1 standard deviation from the mean across all neighborhoods generated for the network.

DDGs (differentiation driver genes) are genes that yield large 3-step neighborhoods that are enriched ($P_{adj} < 0.05$) with tradeSeq-identified genes for a given cell type boundary. We calculated whether each neighborhood contained more tradeSeq-identified genes (for

the neighborhoods' associated cell type boundary) than would be expected by chance using the hypergeometric distribution ("phyper" function) from the stats (version 4.2.1) R package. The arguments were as follows: *q*: (number of neighbors in a neighborhood that are also tradeSeq-identified genes for a given cell type boundary) – 1; *m*: total number of tradeSeq-identified genes for a given cell type boundary; *n*: (total number of identified neighborhoods) – *m*; *k*: neighborhood size (total number of neighbors); lower.tail = false. P-values were adjusted to control the false discovery rate using the "p.adjust" function from the stats (version 4.2.1) R package. These pruning steps resulted in a total of 408 DDGs and associated networks for all cell types (**Supplementary Table S21, S22-24**).

3.5.7. DO eQTL mapping

Prior to performing the eQTL analysis, DNA was extracted from the tails of the 80 DO mice, using the PureLink Genomic DNA mini kit (Invitrogen) and genotyped using the GigaMUGA array by Neogen Genomics (GeneSeek; Lincoln, NE). Processing and quality control of genotype data, including calculation of genotype/allele probabilities, was performed as previously described in Al-Barghouthi and colleagues[19]. Cell type-specific mean expression matrices (as obtained above) for mesenchymal lineage clusters were used as input for the eQTL mapping, which was performed using a linear mixed model (LMM) via the "scan1" function from the qtl2[117] (version 0.30) R package with allowances for the following covariates: sex, age at sacrifice (in days), weight, length, and DO mouse generation. To identify significant eQTL, we calculated a LOD (logarithm of the odds) threshold; for each cell type cluster, we chose 50 genes at random and then permuted them 1000 times using the "scan1perm" function from qtl2. We established the LOD threshold

of 9.68 and 9.49 for the autosomal chromosomes and X chromosome, respectively, by taking the average of the median LOD across each cell type. A total of 563 eQTL that exceeded the LOD thresholds and were no more than 1 Mbp from the transcription start site of the associated eGene (**Supplementary Table S19**).

3.5.8. Cell type proportion analysis

To account for technical sources of variation often retained in scRNA-seq, cell type proportions were transformed using the arcsin (asin) square root transformation from the speckle[207] R package (version 0.0.3). Tests of statistical significance were performed using the propeller t-test and ANOVA functions with default parameters. Sex of the mice and the batch each mouse was associated with for sequencing were modeled as covariates. Transformed values were used as input for computing tests of statistical differences of cell type proportions between mice, as well as correlation to phenotypic traits (**Supplementary Table S8, S9-10**).

3.6. Acknowledgements

Research reported in this publication was supported in part by the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health under award number R01AR68345 to CRF, MAH, and CJR and R01AR077992 to CRF.

3.7. Author Contributions

Luke J Dillard: Writing - review & editing; Visualization; Conceptualization; Methodology; Formal analysis; Data curation; Writing - original draft. Gina M Calabrese: Methodology. Larry D Mesner: Methodology. Charles R Farber: Writing - review & editing; Funding acquisition; Supervision; Conceptualization.

3.8. Disclosures

The authors declare that they have no conflicts of interest with the contents of this article.

3.9. Chapter 3 - Main Figures

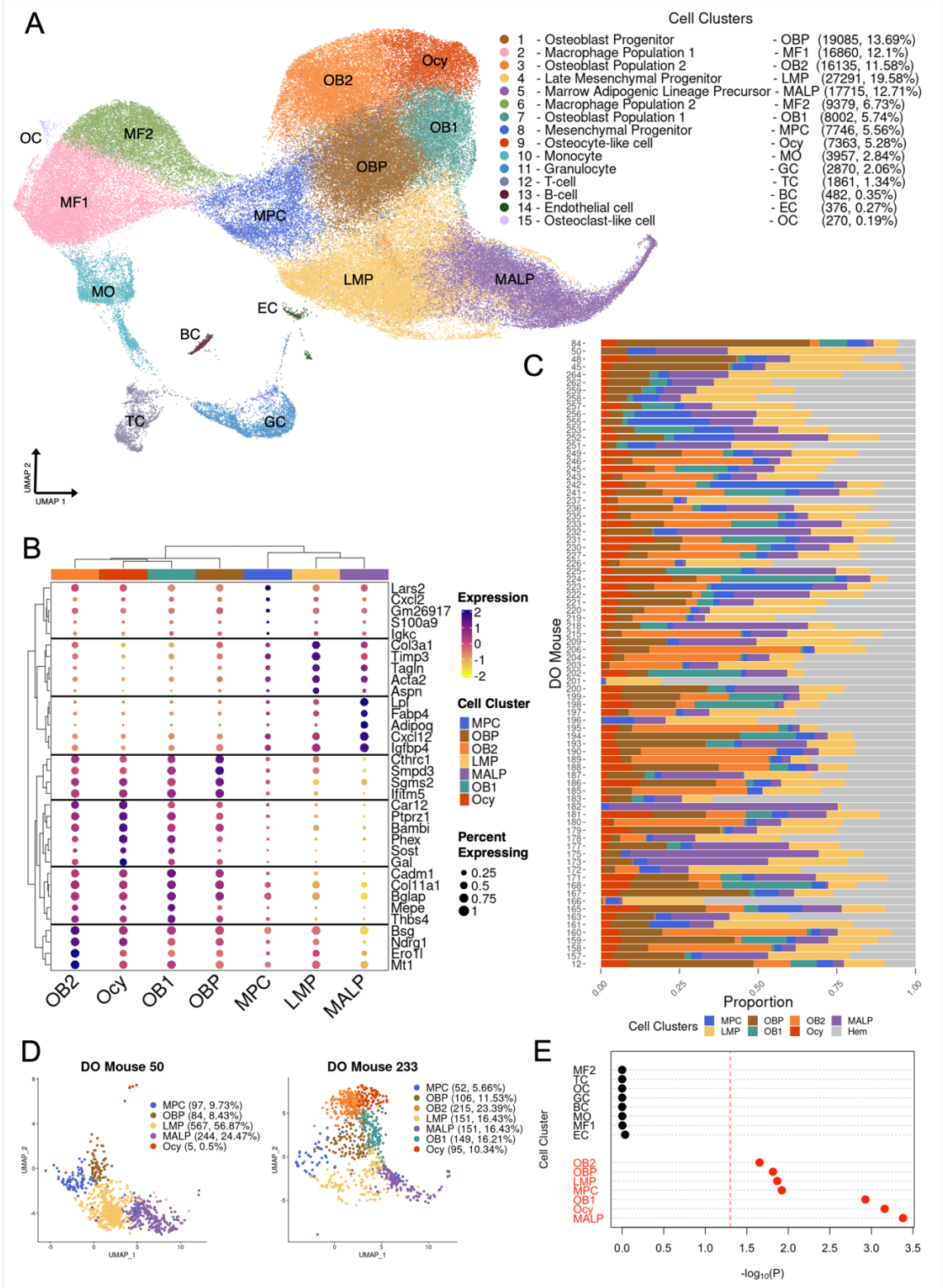


Figure 1. Analysis of single cell RNA-seq (scRNA-seq) data for BMSC-OBs derived from 80 Diversity Outbred (DO)

(A) Uniform Manifold Approximation and Projection (UMAP) of 139,392 single cells (BMSC-OBs). Cell numbers and corresponding percentages for the fifteen (15) annotated cell clusters are listed in parenthesis to the right of the annotated cluster name. **(B)** Dot plot[185] portraying representative and highly expressed genes for all annotated cell clusters. Dot color indicates the scaled average gene expression while the size of the dot corresponds to the percentage of cells of a given cluster that express a given gene. **(C)** The raw proportional abundances of seven (7) mesenchymal cell clusters and one (1) cluster (Hem) representing the remain cells (i.e., hematopoietic immune cells) across all 80 DO mice. **(D)** UMAP plots for mesenchymal lineage cell clusters for DO mouse 50 and DO mouse 233. **(E)** CELLECT (CELL-type Expression-specific integration for Complex Traits) cell type prioritization results. The seven (7) mesenchymal cell clusters were significantly enriched ($P < 0.05$) for BMD heritability and the MALP, Ocy, and OB1 cell clusters were significantly enriched after correcting P-value for multiple testing. All remaining cell clusters (i.e., hematopoietic immune cells) were not significant ($P > 0.05$)

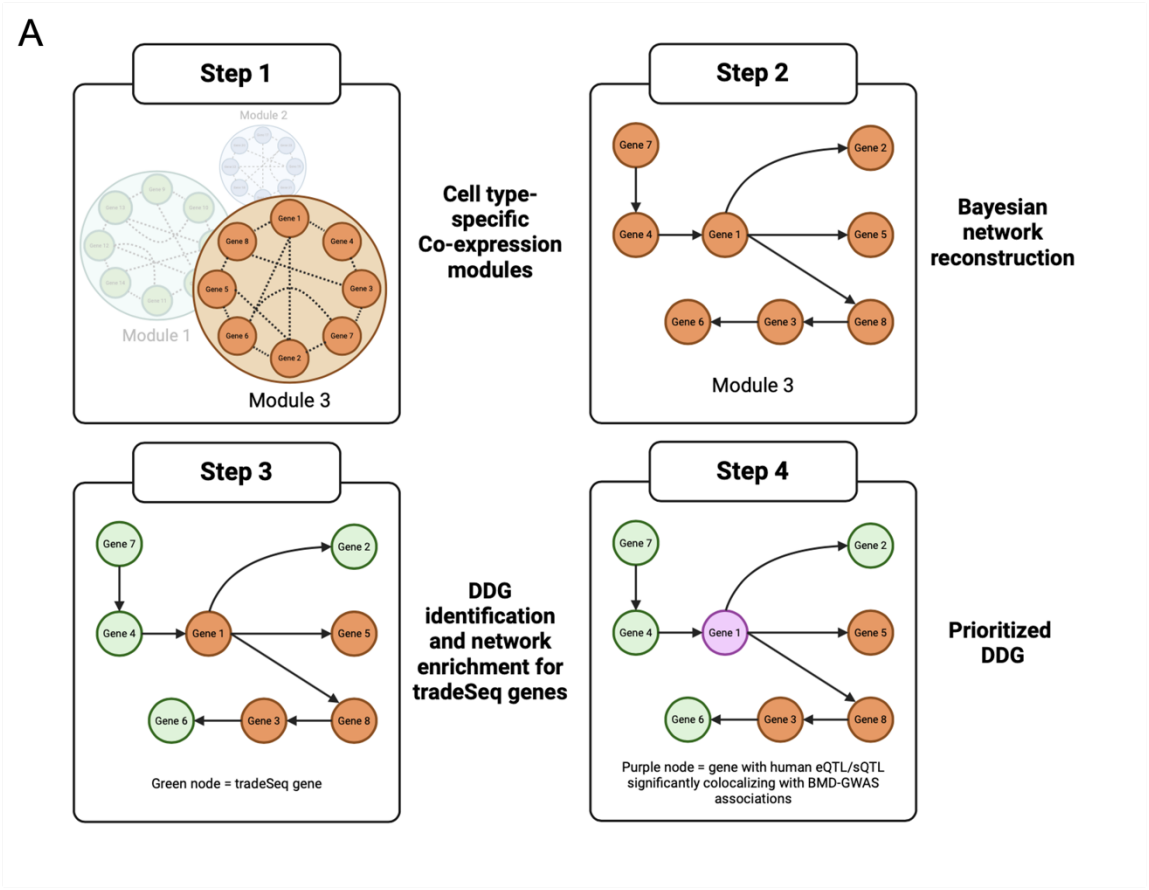


Figure 2. Overview of the network analysis pipeline

(A) *Step 1:* For all seven (7) of the mesenchymal lineage cell clusters (MPC, LMP, OBP, OB1, OB2, Ocy, MALP), cell type-specific co-expression modules were generated using iterative Weighted Gene Co-expression Network Analysis (iterativeWGCNA). *Step 2:* Bayesian networks were learned to generate directed networks and model causal interactions between co-expressed genes. *Step 3:* Differentiation Driver Genes (DDGs) were identified by extracting subnetworks (i.e., large 3-step neighborhood) for each gene in each cell type-specific Bayesian network and highlighting those subnetworks that were enriched ($P_{adj} < 0.05$) for lineage-specific tradeSeq genes for the cell type boundary. *Step 4:* DDGs (and associated networks) were prioritized if the DDG was identified previously as an expression/splicing quantitative trait loci (eQTL/sQTL) that colocalized with BMD GWAS associations. Made with Biorender.

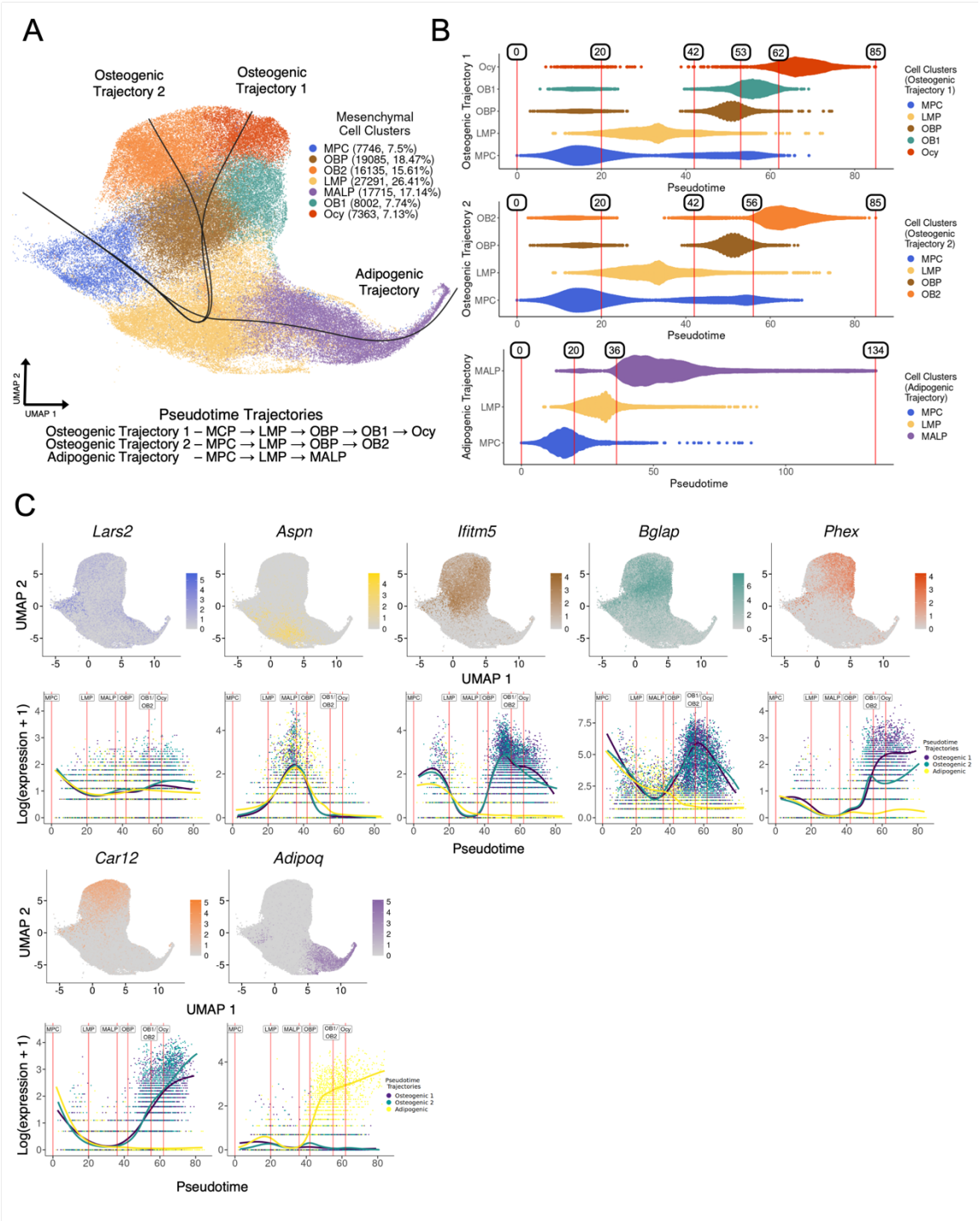

A
Osteogenic Trajectory 2
Osteogenic Trajectory 1
Mesenchymal Cell Clusters
MPC (7746, 7.5%)
OBP (19085, 18.47%)
OB2 (16135, 15.61%)
LMP (27291, 26.41%)
MALP (17715, 17.14%)
OB1 (8002, 7.74%)
Ocy (7363, 7.13%)
Adipogenic Trajectory
UMAP 2
UMAP 1
Pseudotime Trajectories
Osteogenic Trajectory 1 – MCP → LMP → OBP → OB1 → Ocy
Osteogenic Trajectory 2 – MPC → LMP → OBP → OB2
Adipogenic Trajectory – MPC → LMP → MALP
B
Osteogenic Trajectory 1
Cell Clusters (Osteogenic Trajectory 1)
Osteogenic Trajectory 2
Cell Clusters (Osteogenic Trajectory 2)
Adipogenic Trajectory
Cell Clusters (Adipogenic Trajectory)
Pseudotime
C
Lars2
Aspn
Ifitm5
Bglap
Phex
Car12
Adipoq
UMAP 2
UMAP 1
Log(expression + 1)
Pseudotime
Pseudotime Trajectories
Osteogenic 1
Osteogenic 2
Adipogenic

Figure 3. Pseudotime Trajectory Inference analysis and establishment of cell type boundaries for tradeSeq analysis

(A) Three (3) trajectories (two adipogenic, one adipogenic) were inferred from the mesenchymal cell clusters of the BMSC-OB scRNA-seq data using Slingshot. All trajectories originate from the MPC and end in either osteogenic (Ocy, OB2) or adipogenic (MALP) cell fates. **(B)** For each of the trajectories, cell type boundaries were generated using pseudotime values along the trajectories, which encompass the majority of cells of a cell type mapping to their respective trajectory. **(C)** Normalized gene expression of select genes associated with each cluster are represented in feature plots (*top*) and each gene plotted as a function of pseudotime (*bottom*) for all pseudotime trajectories (color corresponds to cell type annotation observed throughout). Vertical lines (red) represent the cell type (pseudotime) boundaries established for each cell type (label). The cell type boundary for OB1 and OB2 are represented as one red line/label for visualization purposes.

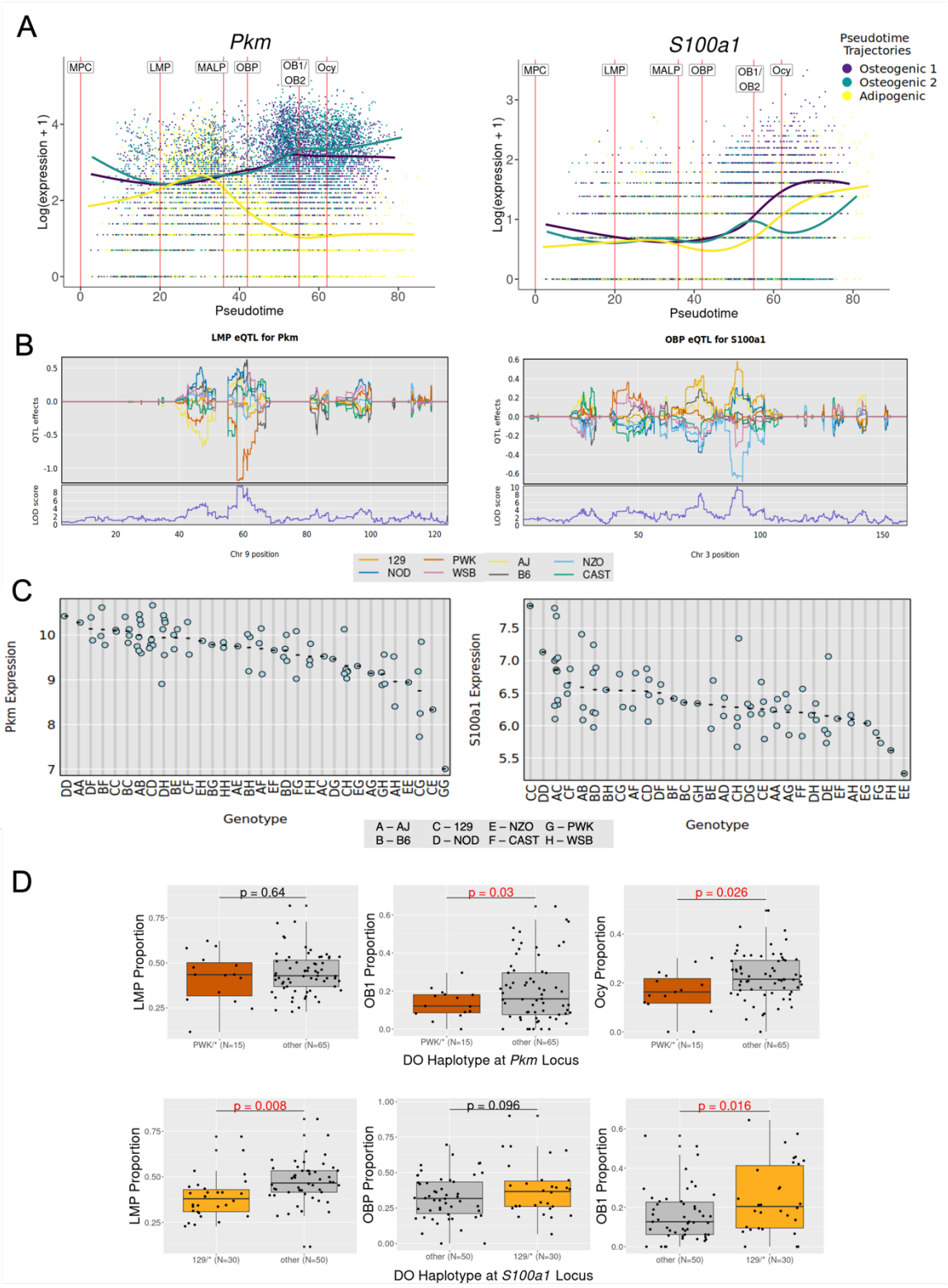

A
Pkm
S100a1
MPC
LMP
MALP
OBP
OB1/
OB2
Ocy
Log(expression + 1)
Pseudotime
Pseudotime
Trajectories
Osteogenic 1
Osteogenic 2
Adipogenic
B
LMP eQTL for Pkm
OBP eQTL for S100a1
QTL effects
LOD score
Chr 9 position
Chr 3 position
129
NOD
PWK
WSB
AJ
B6
NZO
CAST
C
Pkm Expression
S100a1 Expression
Genotype
Genotype
A – AJ
B – B6
C – 129
D – NOD
E – NZO
F – CAST
G – PWK
H – WSB
D
p = 0.64
p = 0.03
p = 0.026
LMP Proportion
OB1 Proportion
Ocy Proportion
PWK/* (N=15)
other (N=65)
DO Haplotype at Pkm Locus
p = 0.008
p = 0.096
p = 0.016
OBP Proportion
other (N=50)
129/* (N=30)
DO Haplotype at S100a1 Locus

Figure 4. TradeSeq-identified genes associated with BMSC-OB differentiation exhibit eQTL effects.
(A) *Pkm* was identified as a significant global tradeSeq-identified gene ($P_{adj} = 8.35 \times 10^{-8}$) for LMP cells along an osteogenic trajectory (LMP_to_OBP) (*left*). *S100a1* was identified as a significant global tradeSeq-identified gene ($P_{adj} = 0.023$) for OBP cells along an osteogenic trajectory 1 (OBP_to_OB1) (*right*). **(B)** Plots indicating the cell type-specific expression quantitative trait loci (eQTL) signal for both *Pkm* and *S100a1*. A negative eQTL effect on *Pkm* expression was observed in LMPs for Diversity Outbred (DO) mice with a PWK haplotype background at the *Pkm* locus (*left*). A positive eQTL effect on the expression of *S100a1* was observed in OBPs for DO mice with a 129 haplotype background at the *S100a1* locus, while a negative effect was observed for NZO mice (*right*). **(C)** The expression of *Pkm* and *S100a1* based on DO mouse (expression values transformed via variance stabilizing transformation (VST), as described in Methods). Genotype abbreviations correspond to DO haplotype background (legend) at the respective gene locus. Mice with at least one PWK allele (genotype abbreviation G) tend to have decreased expression of *Pkm* (*left*). Mice with at least one 129 allele (genotype abbreviation C) tend to have increased expression of *S100a1*, while NZO mice (genotype abbreviation E) have decreased expression (*right*). **(D)** PWK mice had a significant reduction in mature osteoblasts (OB1) and osteocyte-like cells (Ocy) proportions relative to other mice ($P = 0.030$ and $P = 0.026$, respectively; t-test), while LMP proportions were unaffected. Asterisks represent any of the other haplotype backgrounds. 129 mice showed a significant decrease in LMP proportion and increase in OB1 proportion ($P = 0.008$ and $P = 0.016$, respectively; t-test), but OBP proportions were unaffected. No significant effects on cell type proportions were observed in NZO mice (**Supplementary Fig. 3**).

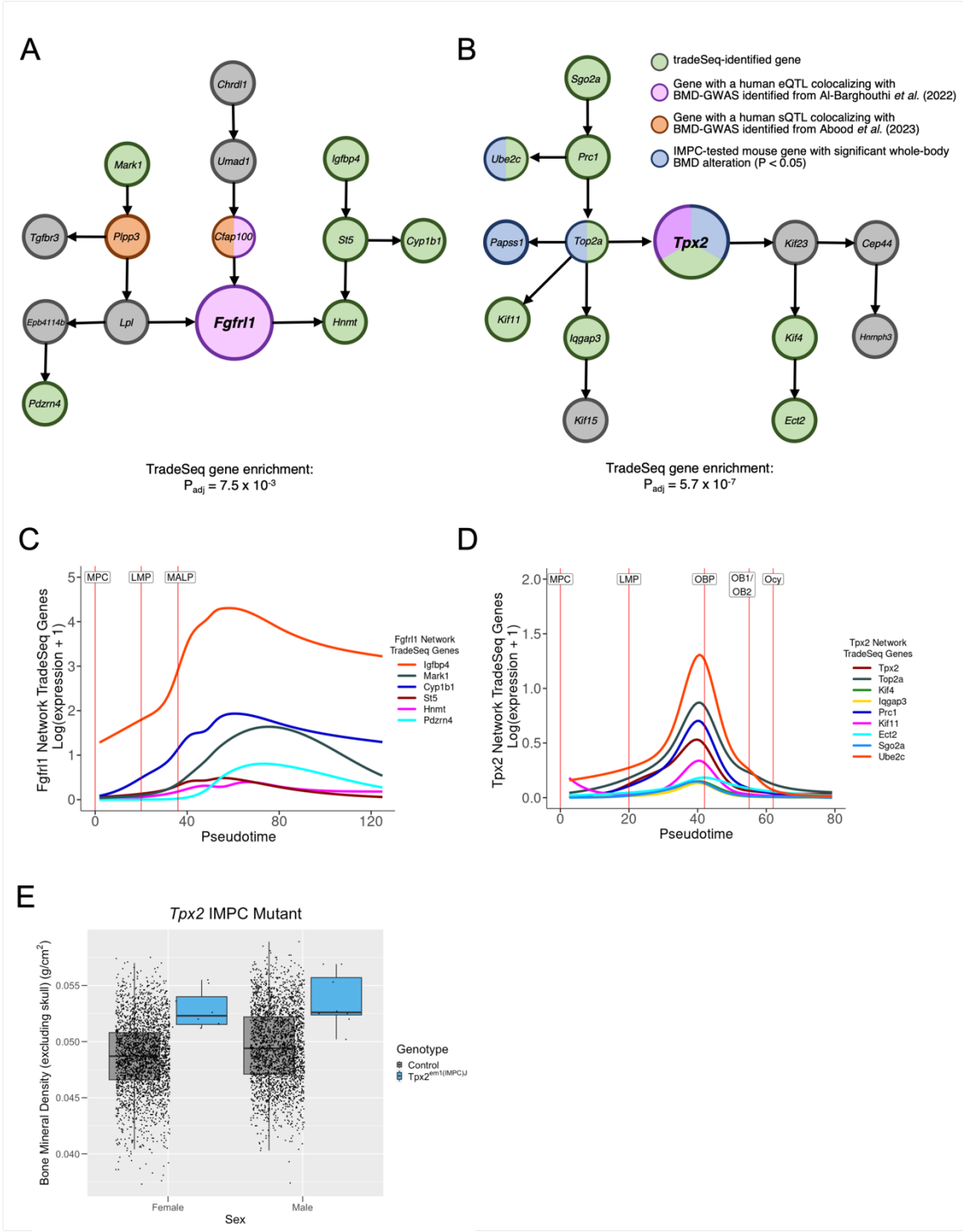

A
Chrdl1
Umad1
Igfbp4
Mark1
Tgfbr3
Plpp3
Cfap100
St5
Cyp1b1
Epb4114b
Lpl
Fgfrl1
Hnmt
Pdzrn4
TradeSeq gene enrichment:
P adj = 7.5 x 10-3
B
Sgo2a
Ube2c
Prc1
Papss1
Top2a
Tpx2
Kif23
Cep44
Kif11
Iqgap3
Kif4
Hnmph3
Kif15
Ect2
TradeSeq gene enrichment:
P adj = 5.7 x 10-7
tradeSeq-identified gene
Gene with a human eQTL colocalizing with BMD-GWAS identified from Al-Barghouthi et al. (2022)
Gene with a human sQTL colocalizing with BMD-GWAS identified from Abood et al. (2023)
IMPC-tested mouse gene with significant whole-body BMD alteration (P < 0.05)
C
MPC
LMP
MALP
Fgfrl1 Network TradeSeq Genes Log(expression + 1)
Pseudotime
Fgfrl1 Network TradeSeq Genes
Igfbp4
Mark1
Cyp1b1
St5
Hnmt
Pdzrn4
D
MPC
LMP
OBP
OB1/ OB2
Ocy
Tpx2 Network TradeSeq Genes Log(expression + 1)
Pseudotime
Tpx2 Network TradeSeq Genes
Tpx2
Top2a
Kif4
Iqgap3
Prc1
Kif11
Ect2
Sgo2a
Ube2c
E
Tpx2 IMPC Mutant
Bone Mineral Density (excluding skull) (g/cm2)
Female
Male
Sex
Genotype
Control
Tpx2em1(IMPC)J

Figure 5. Fgfrl1 and Tpx2 are prioritized DDGs and putative drivers of mesenchymal differentiation.

(A) *Fgfrl1* was identified as a Differentiation Driver Gene (DDG) of a network for LMPs differentiating along an adipogenic trajectory. The network is enriched (P_{adj} = 7.5 x 10^{-3}) for trajectory-specific tradeSeq-identified genes for the LMP_to_MALP cell type boundary (*Hnmt, St5, Igfbp4, Cyp1b1, Pdzrn4, Mark1*). *Fgfrl1* was previous identified as a gene that has BMD GWAS associations that colocalize with an eQTL in the cultured fibroblast GTEx tissue. **(B)** *Tpx2* was identified as a DDG of a network for LMPs differentiating along an osteogenic trajectory. The network is enriched (P_{adj} = 5.7 x 10^{-7}) for tradeSeq-identified genes for the LMP_to_OBP cell type boundary (*Tpx2, Top2a, Kif4, Iqgap3, Prc1, Kif11, Ect2, Sgo2a, Ube2c*). *Tpx2* is both a tradeSeq gene and previous identified as a gene that has BMD GWAS associations that colocalize with an eQTL in the Testis GTEx tissue. **(C-D)** An increase in the expression of all tradeSeq-identified genes coincides with the cell type boundary in which they were identified as significant along their respective trajectories. **(E)** Box plot displaying whole-body bone mineral density (BMD) measurements (excluding skull) from the International Mouse Knockout Consortium (IMPC) for *Tpx2* mutant mice, which exhibited a significant increase in BMD (Genotype P-value = 1.03 x 10^{-3}) in both male and female mice (N = 8 (M) and 8 (F) mutants; N = 2574 (M) and 2633 (F) controls)

Chapter 4

Concluding Remarks and Future Directions

4.1. Summary and Conclusions

Genome-wide associations studies (GWAS) have been successful at identifying thousands of SNPs associated with disease; however, defining the role of putative causal genes identified from GWAS remains challenging. The integration of "-omics" data, particularly transcriptomics, can assist in overcoming this challenge. Leveraging transcriptomics data has proven invaluable to informing GWAS, as demonstrated in studies that use this data modality to perform eQTL mapping and co-expression/gene-regulatory network predictions. Historically, bulk RNA-seq data has been the foundation of such analyses; however, leveraging scRNA-seq data can provide valuable biological context by predicting the cell types in which causal genes operate and influence disease. To inform BMD GWAS, the generation of population-scale transcriptomics data at single-cell resolution in bone relevant cell types can assist in the discovery of novel gene targets. Ultimately, this dissertation showcases the utility of an *in vitro* approach to generating single-cell transcriptomics for bone-relevant cell types derived from genetically distinct mice (i.e., the BMSC-OB model using DO mice) to inform human BMD GWAS.

While the drug discovery pipeline is lengthy, taking dozens of years for an identified target to become clinically tested and approved for human use, novel strategies for target identification will continually be advanced upon. Drug targets supported by findings from GWAS studies tend to be more successful in their candidacy for treatment. Undertaking population scale genomic studies is essential to advancing medicine, understanding risk for disease, and discovering novel therapeutics. Additionally, the inclusion of other data, such as transcriptomics data, can further assist in the prioritization

of targets identified from GWAS. Here, the integration of these large-scale data enabled the contextualization of BMD causal genes through the use of single-cell transcriptomics (scRNA-seq). These high-priority target genes, now with newly provided biological context (e.g., predicted cell type context, biological process through which they operate), can be further prioritized and downstream efforts can be pursued to fully characterize or validate their predicted impact on BMD.

Further, the work described here can be extended to contribute to research outside the context of osteoporosis and BMD. Analytical strategies taking single-cell transcriptomics data as input, such as the network analysis described in **Chapter 3**, can be used as a framework to assist in prioritizing other causal genes identified in other complex disease GWAS studies, such as coronary artery disease (CAD), neurological conditions/disorders, or diabetes.

4.1.1. Assessing the utility of the BMSC-OB model

In **Chapter 2**, we demonstrate that bone marrow-derived stromal cells cultured under osteogenic conditions (BMSC-OBs) from the Diversity Outbred (DO) mouse population can be used as an effective *in vitro* model to generate population-scale scRNA-seq data for mesenchymal lineage cells in large numbers of mice. Subsequent characterization of the cell types captured in the scRNA-seq data and other single-cell analyses (e.g., CELLECT and SCENIC) validate that the BMSC-OB model not only enriches for osteogenic cells, but yields biologically informative transcriptomic profiles of cell types relevant to informing BMD GWAS.

4.1.2. Leveraging scRNA-seq data from the BMSC-OB model to inform GWAS

In **Chapter 3**, we showcase the scalability of the BMSC-OB model and perform scRNA-seq on 80 DO mice. The inclusion of more samples enabled robust characterization of the mesenchymal lineage cells captured by the BMSC-OB model. Further, we performed additional single-cell analyses, such as cell type-specific *cis*-eQTL analysis, temporal gene expression analysis, and a network analysis using the scRNA-seq data. Our goal was to prioritize putative causal genes and provide a biological context in which GWAS-implicated genes potentially cause disease, at cell type specific resolution. Coupling both the temporal gene expression, which identified putative genes driving differentiation, and network analyses, we identified many networks that had central genes with a corresponding human BMD GWAS association colocalizing with eQTL/sQTL in a GTEx tissue.

4.2. Future Directions

4.2.1. In vitro investigation of prioritized targets

An overarching goal of performing multi-omic studies and other computational analyses is to generate a list of high priority targets with the most evidence of being causal to a disease or phenotype to investigate further. Subsequent investigations are ideally taken from a computational setting, to a wet laboratory setting and *in vitro* studies are often pursued as a first step. Wet laboratory strategies for target validation do not feature significantly in this dissertation, nevertheless, our laboratory is experienced in molecular biology techniques that can enable preliminary validations. In fact, much of our recently published work involved first using computational strategies to identify or

prioritize genetic targets (using various data modalities), then leverage a variety of approaches to observe or perturb the genetic target *in vitro*.

4.2.2. RNAi-mediated knock-down

Performing knock-down studies *in vitro* is a molecular biology strategy often employed to study the effects of perturbed or inhibited expression of a specific gene. RNA interference (RNAi) can be performed via short interfering RNA (siRNA), for example, to serve as an effective and reproducible approach to performing knock-down studies *in vitro*[208]. In brief, one approach to siRNA-mediated knock-down involves the introduction of siRNAs containing sequence that is complementary to the target mRNA. The RNA-induced silencing complex (RISC) binds the siRNAs, then, together they bind to the complementary target mRNA transcripts. The transcripts are subsequently cleaved and degraded, thus resulting in inhibited translation and decreased protein product of the target gene[208].

The idealized outcome of such studies could be increased/decreased expression of other genes, or potentially an observable cellular phenotype as a result of knock-down of the target. Before pursuing siRNA-mediated knock-down studies, however, other preliminary studies are performed to provide a more granular understanding of the expression of the target, such as investigating the isoform-specific (spliced transcripts) expression pattern of the gene of interest. For example, RT-qPCR (reverse transcription-quantitative PCR) can be used to analyze the expression of various spliced isoforms of a target, however, a comprehensive understanding of the repertoire of potentially several spliced isoforms of the target would be required. Alternatively, in our more recent work, Abood and colleagues[37] used long-read RNA-seq to delineate the isoform-specific

expression of various genes. Importantly, hFOBs were used as the cellular input for the study, an immortalized pre-osteoblast cell line, which were cultured under osteogenic culture medium *in vitro*. Samples for sequencing were taken across multiple points during their differentiation (days 0, 2, 4, and 10). The resulting long-read RNA-seq data captured the diverse array of spliced transcripts for many genes, such as *Tpm2*, which had primarily four spliced isoforms. They showed that upon the siRNA-mediated knock-down of *Tpm2* isoforms, altered accumulation of mineralized nodules occurred *in vitro*, an indication of the activity of mature osteoblasts during culturing. Knock-down of prioritized targets *in vitro*, followed by mineralization assays, is a pipeline leveraged frequently in our group; it serves as convincing preliminary evidence of causality of our prioritized genetic targets in the process of bone formation.

In **Chapter 3**, we use a network analysis leveraging scRNA-seq to contextualize genes with BMD GWAS associations. In doing so, we predict the cell type in which these genes putatively act, along with the other network constituents that are co-expressed. Of the two high-priority targets identified, both *Fgfrl1* and *Tpx2* may have a role in BMD or osteogenic cell differentiation in the LMP cell cluster. In order to investigate the role of these genetic targets further, the aforementioned *in vitro* assays can be employed. Assessing *Fgfrl1* and *Tpx2* expression during osteogenic cell differentiation, perhaps in a time course experiment (12 days) where cell cultures are sampled at multiple time points, would provide a more granular perspective into their expression profiles. Characterizing the precise isoforms of targets may provide additional insight into the defined expression patterns of these targets during differentiation. Finally, siRNA studies could be performed to assess how perturbed expression of these targets, at the isoform-specific

level, can impact the ability of LMP differentiation or mineralization of osteogenic cells. The outcome of these studies would confirm the proposed roles of both *Fgfrl1* and *Tpx2* in differentiation. Additionally, assessing the expression of the constituents of their respective networks would provide a systems-level understanding of the network as a result of siRNA-mediated knockdown of *Fgfrl1* and *Tpx2.* Nevertheless, the aforementioned *in vitro* studies are essential and provide support for subsequent *in vivo* studies, which is another capability of our lab.

4.2.3. Prime-editing

Since the emergence of CRISPR-Cas System (clustered regularly interspaced short palindromic repeats (CRISPR)-Cas (CRISPR-associated proteins), a multitude of subtypes and derivatives of this molecular assay have become wildly popular in the biomedical research community[209]. While the CRISPR-Cas System can take many functional forms, its most noteworthy capability is applied to genetically modify specific sites across the genome. For example, our lab is currently establishing expertise in performing Prime-editing, which supposably improves the accuracy of genome editing. The Prime-editing systems employ a modified Cas9 nickase fused with a reverse transcriptase, along with a guide RNA (pegRNA)[210]. Together, the Prime-editing system enables single base pair gene editing while improving upon shortcomings of traditional editing strategies, such as error-prone repair of double-strand breaks and propensity for indels[210]. The ability to efficiently and accurately edit specific nucleotide sites across the genome makes Prime-editing extremely relevant to many biomedical research investigations.

In **Chapter 3**, we perform a cell type-specific eQTL analysis to identify genetic loci associated with the expression of specific genes. In the context of our BMSC-OB model which employs DO mice, each of which have a unique genomic background derived from eight specific founder mice, our eQTL analysis identified founder backgrounds that were associated with the expression of certain genes. Importantly, we highlighted *cis*-eQTLs, which constrained the identification of eQTL signal that was within 1 Mbp of the transcription start site of the associated gene (eGene). Given that we leveraged scRNA-seq data for the basis of this eQTL analysis, we highlighted many eGenes at cell type-specific resolution. Therefore, an abundance of information was gained by performing this eQTL analysis, such as: 1) genomic loci of interest, 2) genes with expression levels associated with the genomic loci of interest, 3) the founder strains from which the genomic loci are derived, and 4) cell types in which the effects of identified eQTL may be observed. Together, these insights can make a compelling story pertaining to a specific target gene, one that may warrant follow-up laboratory investigations, such as the aforementioned Prime-editing.

Among those identified eGenes (identified in **Chapter 3**) were *Pkm* and *S100a1*, which exhibit strong *cis*-eQTL effects on their expression and associated with specific DO mouse haplotype background at their respective loci. Further fine-mapping of the loci can potentially identify the SNPs that are responsible for the eQTL effect. Given that these eQTLs are predicted to exert their effects on their target eGenes in *cis,* fine-mapped SNPs likely reside in regulatory elements in close proximity to the eGene. A variety of follow-up investigations can be employed to validate the eQTL effect on the prospective eGene, such as Prime-editing.

Prior to pursuing avenues of research employing Prime-editing, preliminary studies to further investigate the eGenes of interest, such as *Pkm*, could include quantifying the baseline expression level of an eGene in a control mouse population, such as B6, and subsequently comparing baseline to the expression in a mouse population with a genome that is comprised entirely of the DO founder background of interest, such as PWK (which is putatively associated with a negative effect on the expression of *Pkm*). These studies should be done in a homogenous population of LMPs, which was the cell type in which the eQTL was identified. The hypothesized outcome of this study should be decreased levels of *Pkm* in LMPs from PWK mice, relative to B6 control mice. Additionally, fine-mapping studies could identify the SNPs driving the eQTL in PWK mice, which should discover causal SNPs located in either a known promotor region for the associated gene, or a novel regulatory element. Finally, after these preliminary studies validate our hypotheses, Prime-editing studies could be warranted. Editing the causal SNPs in the PWK mice could "rescue" the expression levels of *Pkm* to those observed in B6 control mice. Additionally, introducing the causal SNP via Prime-editing into the B6 control mice could further validate the genetic effect of the eQTL on *Pkm* expression. These experiments serve as a suggested framework for follow-up *in vitro* experiments to perform after prioritized genetic targets are identified.

4.3. Long-read, single-cell transcriptomics

One of the themes of this dissertation was to showcase the capabilities of scRNA-seq, specifically as it applies to informing BMD GWAS. As mentioned in **Chapter 1**, scRNA-seq has significant advantages over traditional, bulk RNA-seq, such as being able to attribute transcriptomic signal to single cells. Upon downstream clustering and

analyzing the expression across thousands of cells, a more refined understanding of the transcriptome underlying specific cell types can be gained, such as osteoblasts or MALPs. However, some scRNA-seq methods (e.g., some droplet-based protocols), are not capable of capturing isoform-specific gene expression. As we continue to venture towards a more granular understanding of the transcriptome, neglecting the impact of splicing on a biological system or involvement in disease would hinder biomedical research. Therefore, to overcome this looming challenge, recent advancements in scRNA-seq technology have strived to couple both highly desired cell type specificity with long-read sequencing. For example, the biotechnology company, PacBio (Menlo Park, CA), has developed MAS-seq (Multiplexed Arrays Sequencing) to capture full-length RNA transcripts, at single cell resolution.

While MAS-seq protocols have yet to be established in our lab, its implementation could dramatically improve our understanding of the isoform-specific expression profiles of bone-relevant cell types. In **Chapter 3**, we highlight two high-priority eGenes (*Pkm* and *S100a1*), along with the associated DO haplotype background potentially impacting their expression. Further, mice with the associated DO haplotype background for these two targets exhibited abundance/proportion differences in various cell types captured in the BMSC-OB scRNA-seq data. These analyses attempt to connect DO haplotype, expression of specific targets, and the abundance of certain cell types. Osteogenic cell types, namely the OB1 population (which exhibited canonical markers of mature, mineralizing osteoblasts) were among the cell types that were observed to have significant ($P < 0.05$) differences in their abundance in mice harboring the DO haplotype background of interest.

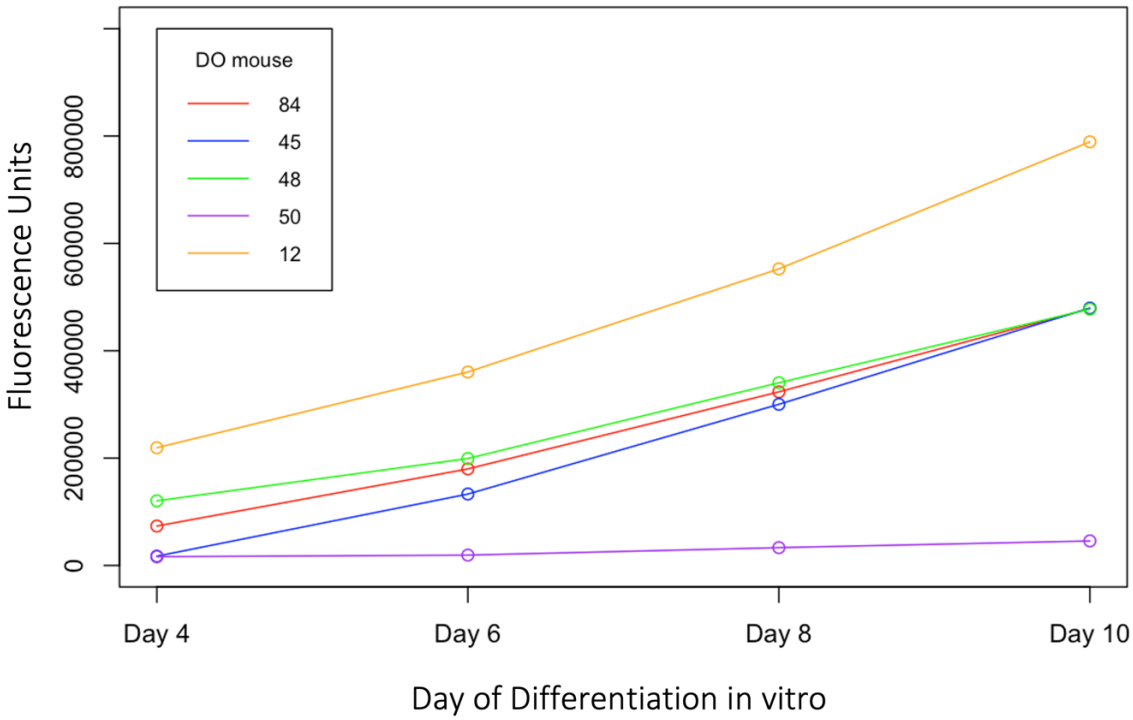
Understanding how genetic background can potentially impact the abundance of relevant cell types is at the forefront of biomedical research that perform scRNA-seq studies. Given the importance of osteoblasts in bone biology and BMD, follow-up studies to investigate the aforementioned findings may be warranted. In terms of follow-up investigation of these two targets, MAS-seq could be worthwhile to pursue.

For example, in the case of *Pkm*, we show that mice with the PWK background have decreased expression of *Pkm*. Further, these same mice also have decreased proportion of both OB1 and Ocy. Many approaches could be pursued *in vitro* to further validate the putative connection between PWK, *Pkm*, and decreased osteoblast proportion. In a pilot study, MAS-seq could be employed to feasibly gather long-read, scRNA-seq data on BMSC-OBs derived from mice with entirely PWK background; this would capture 1. expression levels of *Pkm* (at isoform-level) and 2. the proportions of OB1/Ocy for each PWK mouse, both of which are necessary features to capture. Importantly, this experiment would be performed in a control mouse population as well, such as B6, for comparison purposes. The ideal outcome of this study would be validation of the aforenoted trend (PWK background, decreased *Pkm* expression, and reduced proportion of osteogenic cells) and the trend would be significant upon comparison to control (B6). Aside from validation of the hypothesis, other novel insights could be gleaned from this proposed experiment, such as elucidating the isoform-specific expression patterns for *Pkm* across BMSC-OBs (in both PWK and B6 mouse background).

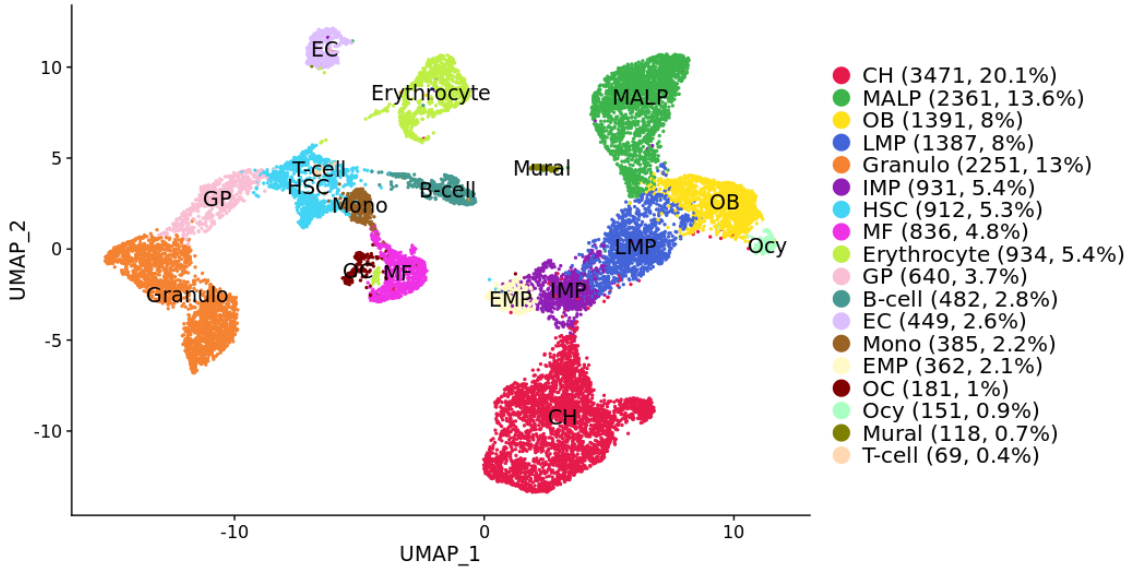
Appendix A

Supplementary Figures

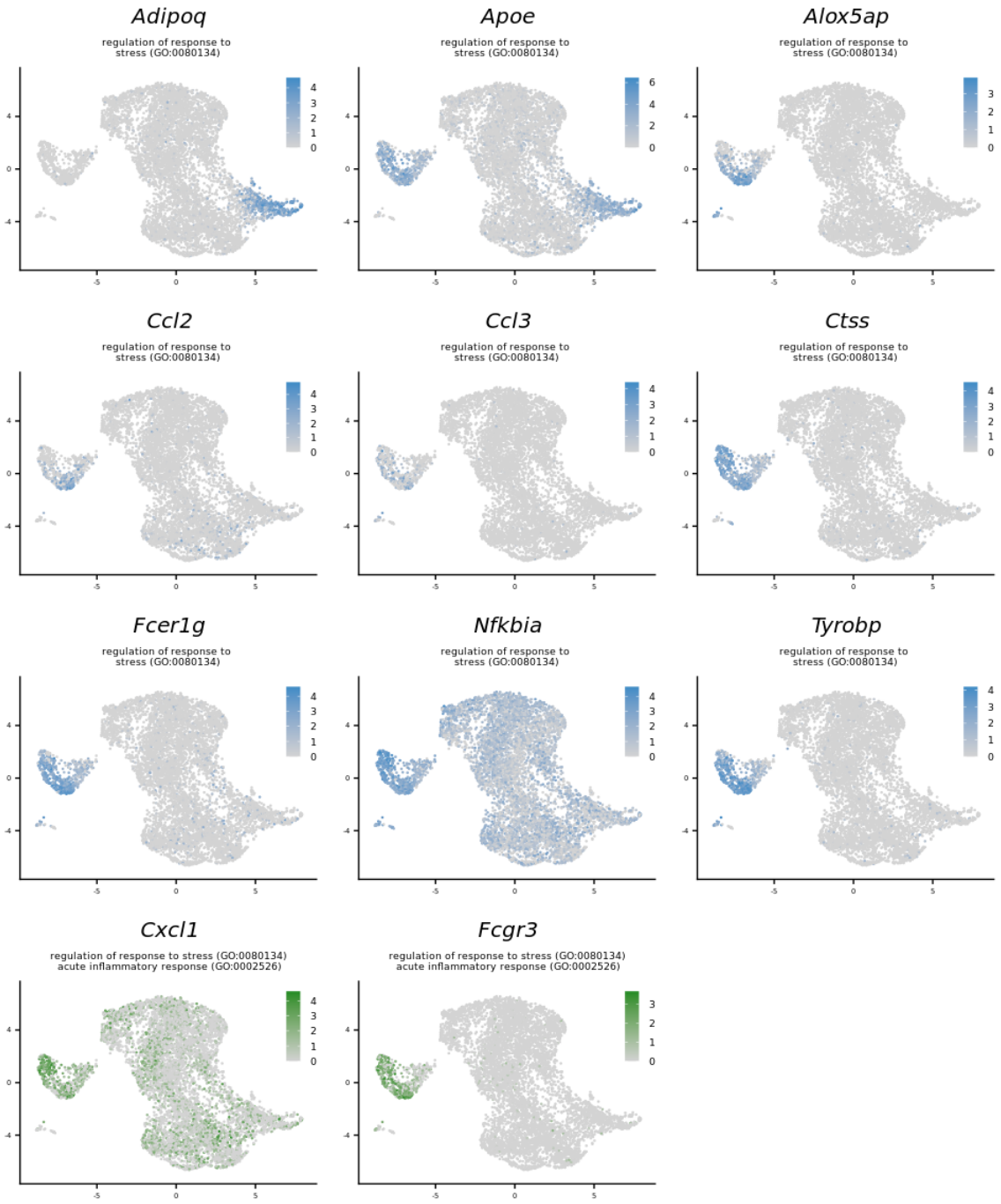
A. Chapter 2 - Supplementary Figures



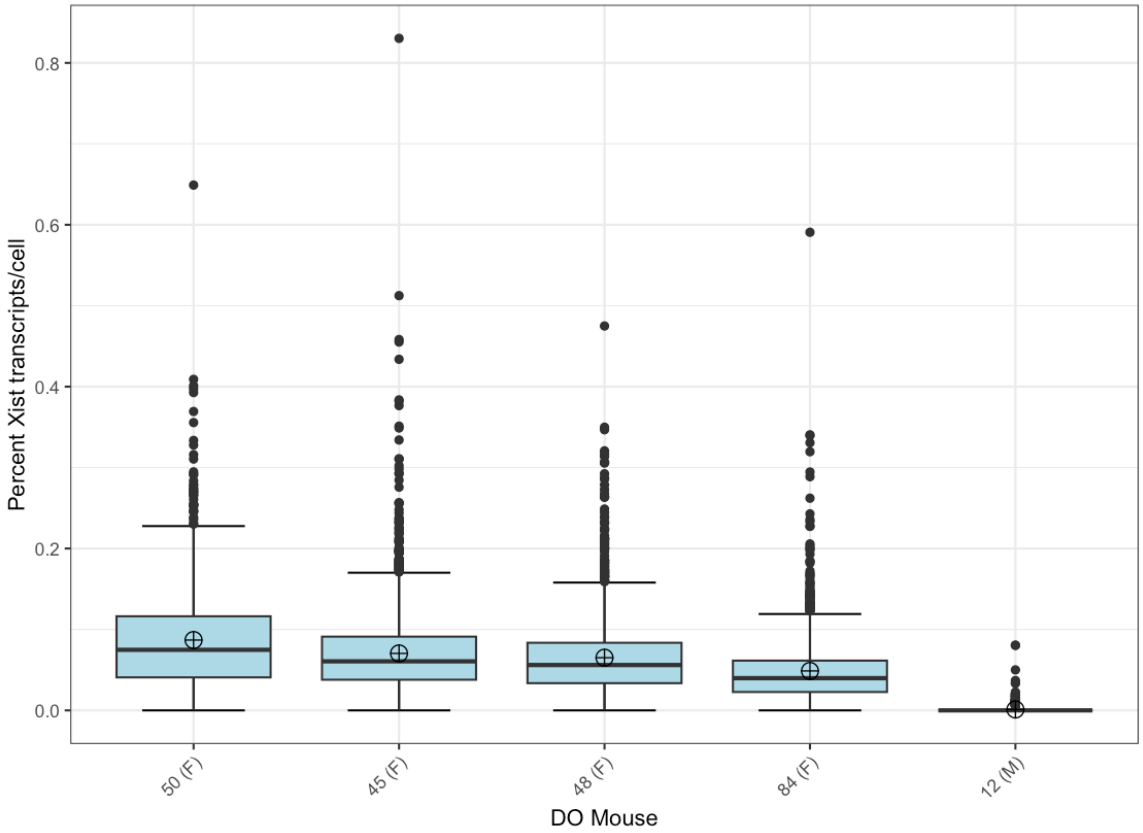
Supplemental Figure 1: Mineralization of BMSC-OBs after in vitro osteogenic differentiation. Mineralized deposits were quantified via IRDye 680 BoneTag Optical Probe incorporation. Fluorescence units were calculated by subtracting the average number of units recorded in background wells from the units recorded in the DO mouse sample wells.



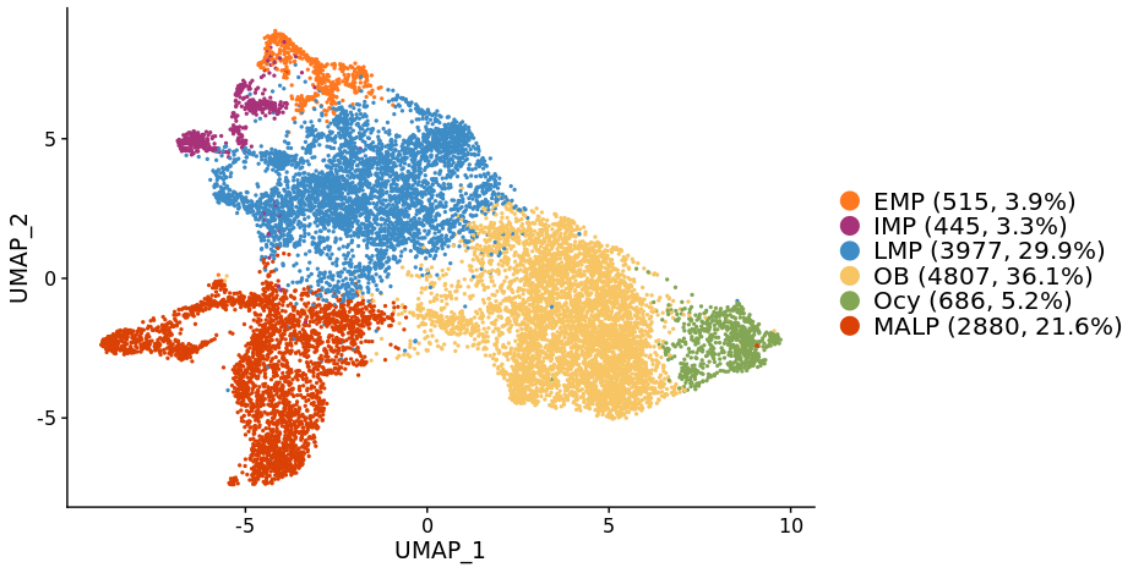
Supplemental Figure 2: Uniform Manifold Approximation and Projection (UMAP) of cell clusters for the 17,311 cells from the Zhong et al. (2020) scRNA-seq dataset. Endosteal Td+ bone marrow cells were sequenced from 1-month-old (n=2), 1.5-month-old (n=3), 3-month-old (n=3) male Col2/Td mice. The Zhong et al. (2020) scRNA-seq data was processed in the same fashion as the BMSC-OBs scRNA-seq data (Methods) and clustered at a resolution of 0.675. Cell count numbers and corresponding percentage of the entire population are listed in parentheses to the right of the annotated cluster name: OB: osteoblast; Ocy: osteocyte; EMP: early mesenchymal progenitor; IMP: intermediate mesenchymal progenitor; LMP: late mesenchymal progenitor; MALP: marrow adipogenic lineage precursors; CH: chondrocyte; HSC: hematopoietic stem cell; EC: endothelial cell; GP: granulocyte progenitor; OC: osteoclast; Granulo: granulocyte; MF: macrophage; Mono: monocyte; Mural: mural cells; T-cell: T-lymphocyte; B-cell: B-lymphocyte.



Supplemental Figure 3: Feature plots portraying the normalized expression of select Differentially Expressed Genes (DEGs). The selected DEGs had an average log2 Fold Change (avg_log2FC) greater than 2.0 in any given cluster of the BMSC-OB scRNA-seq and that belong to either Gene Ontology (GO) Terms: regulation of response to stress (GO:0080134), acute inflammatory response (GO:0002526), or both.

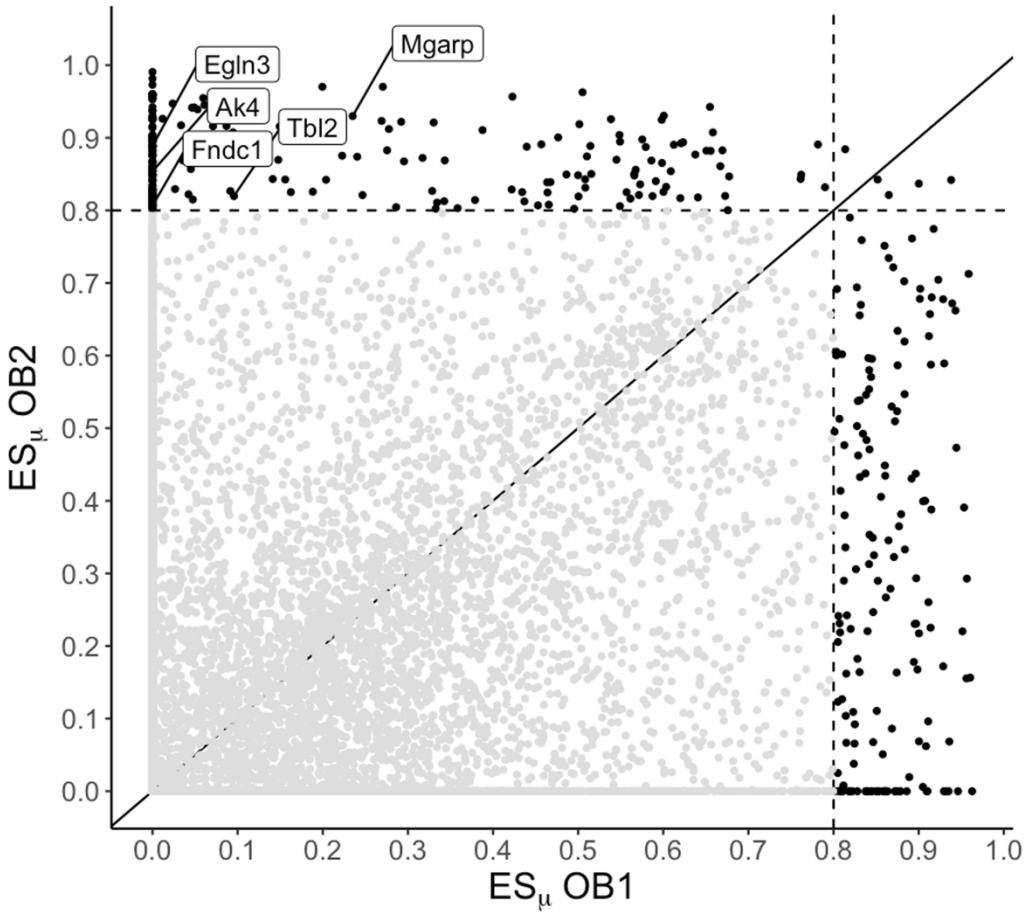


Supplemental Figure 4. Boxplots for each Diversity Outbred (DO) mouse portraying *Xist* (X-inactive specific transcript) expression (as a percentage of all reads) for each barcoded cell belonging to each of the 5 Diversity Outbred (DO) mice (M = male, F = female). Souporcell genotype deconvolution of the BMSC-OB scRNA-seq dataset resulted in grouping of individual cells based on genotype. Genetic variants captured for each genotype cluster were compared to the same variants captured by GigaMUGA genotype microarrays (performed previously on each mouse). In a pairwise comparison between Souporcell genotype clusters and GigaMUGA microarray data for each DO mouse, a mouse was assigned a genotype cluster based on the highest percentage of matching allele calls made for genetic variants identified between Souporcell and GIGAMUGA genotype microarrays. DO mouse 12 was confirmed to be male based on low Xist expression in all cells, further confirming accurate genotype clustering of cells via Souporcell.

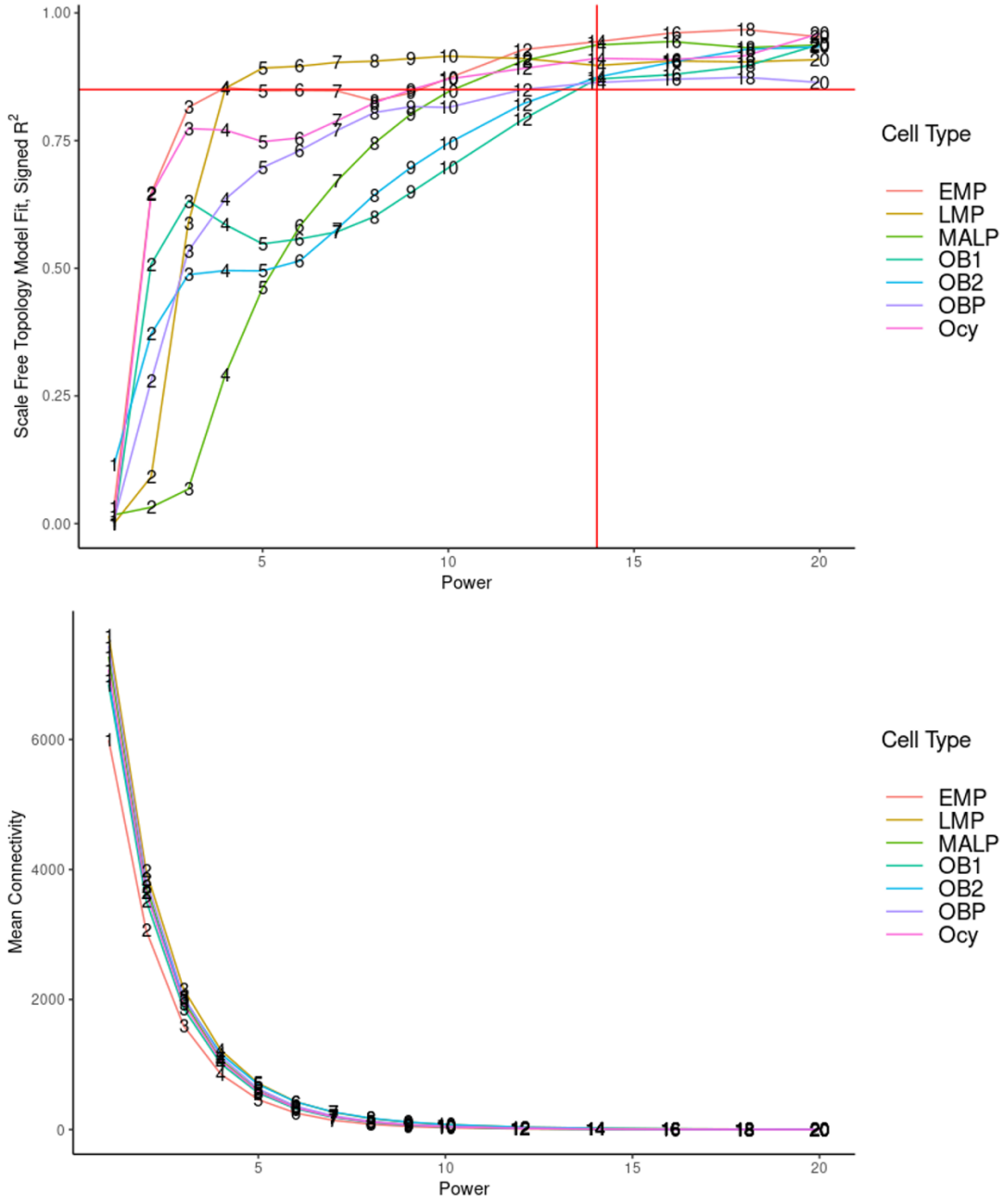


Supplemental Figure 5: Uniform Manifold Approximation and Projection (UMAP) of cell clusters for the 13,310 cells from the integrated scRNA-seq data (BMSC-OB and Zhong et al. (2020) datasets). ScRNA-seq data integration was performed using Canonical Correlation Analysis (CCA) and using only the osteogenic and adipogenic lineage cells as input. The integrated data was processed in the same fashion as the BMSC-OBs scRNA-seq data (Methods) and clustered at a resolution of 0.22. Cell count numbers and corresponding percentage of the entire population are listed in parentheses to the right of the annotated cluster name: EMP: early mesenchymal progenitor; IMP: intermediate mesenchymal progenitor; LMP: late mesenchymal progenitor; OB: osteoblast; Ocy: osteocyte; MALP: marrow adipogenic lineage precursors.

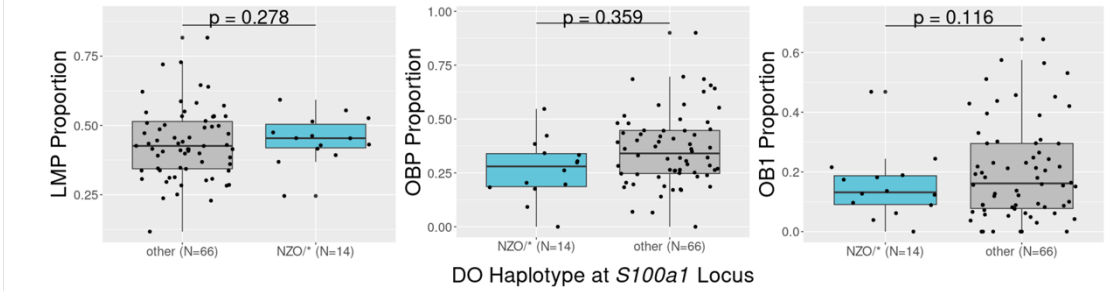
B. Chapter 3 - Supplementary Figures



Supplementary Figure 1: Comparison of gene expression specificity scores (ESμ) for two osteoblast populations identified in the BMSC-OB scRNA-seq data. ESμ scores are a continuous values between 0 (not specific) and 1 (very specific). Black dots exceeding the dotted line are those genes with ESμ scores greater than 0.8. The five (5) labeled genes (*Egln3, Ak4, Fndc1, Tbl2, Mgarp*) were associated with the Gene Ontology (GO) term cellular response to hypoxia (GO:0071456), more highly expressed and more specifically expressed in OB2 (relative to OB1).



Supplementary Figure 2: Scale Free Topology and Mean Connectivity graphs for the cell type-specific iterativeWGCNA analysis. A soft thresholding power of 14 was selected for the generation of all co-expression modules for all clusters, which was the point at which R^2 exceeded a threshold of 0.85



Supplementary Figure 3: Tests of significance for cell type proportions for NZO mice. Mice with at least one NZO allele at the *S100a1* locus (N = 14) had no significant difference in cell type proportions (P > 0.05; t-test) as compared mice with other DO haplotype background at this locus. Asterisks represent any of the other haplotype backgrounds.

Appendix B

Supplementary Tables

A. Chapter 2 - Supplementary Tables

Supplementary Tables and Description of Supplementary Tables associated with **Chapter 2** can be found online with the corresponding published article (PMID: 37436066) at https://doi.org/10.1002/jbmr.4882

B. Chapter 3 - Supplementary Tables

Supplementary Tables and Description of Supplementary Tables associated with **Chapter 3** can be found online at the following Zenodo link: https://doi.org/10.5281/zenodo.11066753

References

1. Marini, F. & Brandi, M. L. Genetic determinants of osteoporosis: common bases to cardiovascular diseases? *Int. J. Hypertens.* **2010**, (2010).
2. Haseltine, K. N. *et al.* Bone Mineral Density: Clinical Relevance and Quantitative Assessment. *J. Nucl. Med.* **62**, 446–454 (2021).
3. Burge, R. *et al.* Incidence and economic burden of osteoporosis-related fractures in the United States, 2005-2025. *J. Bone Miner. Res.* **22**, 465–475 (2007).
4. Johnell, O. *et al.* Predictive value of BMD for hip and other fractures. *J. Bone Miner. Res.* **20**, 1185–1194 (2005).
5. Boudin, E., Fijalkowski, I., Hendrickx, G. & Van Hul, W. Genetic control of bone mass. *Mol. Cell. Endocrinol.* **432**, 3–13 (2016).
6. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era--concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255–266 (2008).
7. Morrison, N. A. *et al.* Prediction of bone density from vitamin D receptor alleles. *Nature* **367**, 284–287 (1994).
8. Pluijm, S. M. F. *et al.* Collagen type I alpha1 Sp1 polymorphism, osteoporosis, and intervertebral disc degeneration in older men and women. *Ann. Rheum. Dis.* **63**, 71–77 (2004).
9. Uffelmann, E. *et al.* Genome-wide association studies. *Nature Reviews Methods Primers* **1**, 1–21 (2021).
10. Morris, J. A. *et al.* An atlas of genetic influences on osteoporosis in humans and mice. *Nat. Genet.* **51**, 258–266 (2019).

11. Kim, S. K. Identification of 613 new loci associated with heel bone mineral density and a polygenic risk score for bone mineral density, osteoporosis and fracture. *PLoS One* **13**, e0200785 (2018).

12. Peter, I. & Seddon, J. M. Genetic epidemiology: successes and challenges of genome-wide association studies using the example of age-related macular degeneration. *Am. J. Ophthalmol.* **150**, 450-452.e2 (2010).

13. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).

14. Qin, D. Next-generation sequencing and its clinical application. *Cancer Biol Med* **16**, 4–10 (2019).

15. Guttmacher, A. E. & Collins, F. S. Genomic medicine--a primer. *N. Engl. J. Med.* **347**, 1512–1520 (2002).

16. Tyers, M. & Mann, M. From genomics to proteomics. *Nature* **422**, 193–197 (2003).

17. Civelek, M. & Lusis, A. J. Systems genetics approaches to understand complex traits. *Nat. Rev. Genet.* **15**, 34–48 (2014).

18. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).

19. Al-Barghouthi, B. M. *et al.* Systems genetics in diversity outbred mice inform BMD GWAS and identify determinants of bone strength. *Nat. Commun.* **12**, 3408 (2021).

20. Doane, A. S. & Elemento, O. Regulatory elements in molecular networks. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **9**, (2017).

21. Sul, J. H. *et al.* Accurate and fast multiple-testing correction in eQTL studies. *Am. J. Hum. Genet.* **96**, 857–868 (2015).

22. Võsa, U. *et al*. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, 1300–1310 (2021).

23. Zhang, J. & Zhao, H. eQTL studies: from bulk tissues to single cells. *J. Genet. Genomics* (2023) doi:10.1016/j.jgg.2023.05.003.

24. Yap, C. X. *et al*. Trans-eQTLs identified in whole blood have limited influence on complex disease biology. *Eur. J. Hum. Genet.* **26**, 1361–1368 (2018).

25. Dutta, D. *et al*. Aggregative trans-eQTL analysis detects trait-specific target gene sets in whole blood. *Nat. Commun.* **13**, 4323 (2022).

26. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

27. GTEx Portal. https://www.gtexportal.org/home/.

28. GTEx Consortium *et al*. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

29. Al-Barghouthi, B. M. *et al*. Transcriptome-wide association study and eQTL colocalization identify potentially causal genes responsible for human bone mineral density GWAS associations. *Elife* **11**, (2022).

30. Mai, J., Lu, M., Gao, Q., Zeng, J. & Xiao, J. Transcriptome-wide association studies: recent advances in methods, applications and available databases. *Commun Biol* **6**, 899 (2023).

31. Marden, J. H. Quantitative and evolutionary biology of alternative splicing: how changing the mix of alternative transcripts affects phenotypic plasticity and reaction norms. *Heredity* **100**, 111–120 (2006).

32. Marasco, L. E. & Kornblihtt, A. R. The physiology of alternative splicing. *Nat. Rev. Mol. Cell Biol.* **24**, 242–254 (2023).

33. Mucaki, E. J., Shirley, B. C. & Rogan, P. K. Expression Changes Confirm Genomic Variants Predicted to Result in Allele-Specific, Alternative mRNA Splicing. *Front. Genet.* **11**, 109 (2020).

34. Wright, C. J., Smith, C. W. J. & Jiggins, C. D. Alternative splicing as a source of phenotypic diversity. *Nat. Rev. Genet.* **23**, 697–710 (2022).

35. Li, Y. I. *et al.* Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).

36. Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* **12**, 87–98 (2011).

37. Abood, A. *et al.* Long-read proteogenomics to connect disease-associated sQTLs to the protein isoform effectors of disease. *bioRxiv* 2023.03.17.531557 (2023) doi:10.1101/2023.03.17.531557.

38. Visscher, P. M., Yengo, L., Cox, N. J. & Wray, N. R. Discovery and implications of polygenicity of common diseases. *Science* **373**, 1468–1473 (2021).

39. Serin, E. A. R., Nijveen, H., Hilhorst, H. W. M. & Ligterink, W. Learning from Co-expression Networks: Possibilities and Challenges. *Front. Plant Sci.* **7**, 444 (2016).

40. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

41. Goh, K.-I. *et al.* The human disease network. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 8685–8690 (2007).

42. Li, J. *et al*. Application of Weighted Gene Co-expression Network Analysis for Data from Paired Design. *Sci*. *Rep*. **8**, 622 (2018).

43. Calabrese, G. M. *et al*. Integrating GWAS and Co-expression Network Data Identifies Bone Mineral Density Genes SPTBN1 and MARK3 and an Osteoblast Functional Module. *Cell Syst* **4**, 46-59.e4 (2017).

44. Sabik, O. L., Calabrese, G. M., Taleghani, E., Ackert-Bicknell, C. L. & Farber, C. R. Identification of a Core Module for Bone Mineral Density through the Integration of a Co-expression Network and GWAS Data. *Cell Rep*. **32**, 108145 (2020).

45. Groza, T. *et al*. The International Mouse Phenotyping Consortium: comprehensive knockout phenotyping underpinning the study of human disease. *Nucleic Acids Res*. **51**, D1038–D1045 (2023).

46. van Dam, S., Võsa, U., van der Graaf, A., Franke, L. & de Magalhães, J. P. Gene co-expression analysis for functional classification and gene-disease predictions. *Brief*. *Bioinform*. **19**, 575–592 (2018).

47. De Smet, R. & Marchal, K. Advantages and limitations of current network inference methods. *Nat*. *Rev*. *Microbiol*. **8**, 717–729 (2010).

48. Thompson, D., Regev, A. & Roy, S. Comparative analysis of gene regulatory networks: from network reconstruction to evolution. *Annu*. *Rev*. *Cell Dev*. *Biol*. **31**, 399–428 (2015).

49. Needham, C. J., Bradford, J. R., Bulpitt, A. J. & Westhead, D. R. A Primer on Learning in Bayesian Networks for Computational Biology. *PLoS Comput*. *Biol*. **3**, (2007).

50. Reed, J. N. *et al*. Systems genetics analysis of human body fat distribution genes identifies Wnt signaling and mitochondrial activity in adipocytes. *bioRxiv* 2023.09.06.556534 (2023) doi:10.1101/2023.09.06.556534.

51. Puga, J. L., Krzywinski, M. & Altman, N. Points of Significance. Bayesian networks. *Nat. Methods* **12**, 799–800 (2015).

52. Tsamardinos, I., Brown, L. E. & Aliferis, C. F. The max-min hill-climbing Bayesian network structure learning algorithm. *Mach. Learn.* **65**, 31–78 (2006).

53. Chang, J. *et al*. Dynamic Bayesian networks with application in environmental modeling and management: A review. *Environmental Modelling & Software* **170**, 105835 (2023).

54. Zou, M. & Conzen, S. D. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics* **21**, 71–79 (2005).

55. Li, X. & Wang, C.-Y. From bulk, single-cell to spatial RNA sequencing. *Int. J. Oral Sci.* **13**, 1–6 (2021).

56. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **50**, 1–14 (2018).

57. Newman, A. M. *et al*. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).

58. Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & De Preter, K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Commun.* **11**, 5650 (2020).

59. Kukurba, K. R. & Montgomery, S. B. RNA Sequencing and Analysis. *Cold Spring Harb. Protoc.* **2015**, 951–969 (2015).

60. Hegenbarth, J.-C., Lezzoche, G., De Windt, L. J. & Stoll, M. Perspectives on Bulk-Tissue RNA Sequencing and Single-Cell RNA Sequencing for Cardiac Transcriptomics. *Frontiers in Molecular Medicine* **2**, (2022).

61. Decoding noncoding RNAs. *Nat. Methods* **19**, 1147–1148 (2022).

62. Mattick, J. S. *et al.* Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat. Rev. Mol. Cell Biol.* **24**, 430–447 (2023).

63. Fang, S. *et al.* NONCODEV5: a comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Res.* **46**, D308–D314 (2018).

64. Winkle, M., El-Daly, S. M., Fabbri, M. & Calin, G. A. Noncoding RNA therapeutics - challenges and potential solutions. *Nat. Rev. Drug Discov.* **20**, 629–651 (2021).

65. Teufel, M. & Sobetzko, P. Reducing costs for DNA and RNA sequencing by sample pooling using a metagenomic approach. *BMC Genomics* **23**, 613 (2022).

66. Andrews, T. S., Kiselev, V. Y., McCarthy, D. & Hemberg, M. Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. *Nat. Protoc.* **16**, 1–9 (2021).

67. Genomics, 10x. *Chromium Next GEM Single Cell 3' Reagent Kits v3.1 User Guide*. https://cdn.10xgenomics.com/image/upload/v1660261285/support-documents/CG000204_ChromiumNextGEMSingleCell3_v3.1_Rev_D.pdf (2019).

68. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).

69. You, Y. *et al*. Benchmarking UMI-based single-cell RNA-seq preprocessing workflows. *Genome Biol*. **22**, 339 (2021).

70. UC Davis Genome Center. What are UMIs and why are they used in high-throughput sequencing? *https://dnatech.genomecenter.ucdavis.edu/* https://dnatech.genomecenter.ucdavis.edu/faqs/what-are-umis-and-why-are-they-used-in-high-throughput-sequencing/.

71. Wu, Y. & Zhang, K. Tools for the analysis of high-dimensional single-cell RNA sequencing data. *Nat. Rev. Nephrol*. **16**, 408–421 (2020).

72. Seurat - Guided Clustering Tutorial. *https://satijalab.org/* https://satijalab.org/seurat/articles/pbmc3k_tutorial (2023).

73. Su, M. *et al*. Data analysis guidelines for single-cell RNA-seq in biomedical studies and clinical applications. *Mil Med Res* **9**, 68 (2022).

74. Stuart, T. *et al*. Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

75. Sun, S., Zhu, J., Ma, Y. & Zhou, X. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biol*. **20**, 269 (2019).

76. Lever, J., Krzywinski, M. & Altman, N. Principal component analysis. *Nat. Methods* **14**, 641–642 (2017).

77. Migenda, N., Möller, R. & Schenck, W. Adaptive dimensionality reduction for neural network-based online principal component analysis. *PLoS One* **16**, e0248896 (2021).

78. Levine, J. H. *et al*. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162**, 184–197 (2015).

79. Kiselev, V. Y., Andrews, T. S. & Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.* **20**, 273–282 (2019).

80. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605 (11/2008).

81. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018).

82. Roca, C. P. *et al*. A cross entropy test allows quantitative statistical comparison of t-SNE and UMAP representations. *Cell Rep Methods* **3**, 100390 (2023).

83. Becht, E. *et al*. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* (2018) doi:10.1038/nbt.4314.

84. Kobak, D. & Linderman, G. C. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nature biotechnology* vol. 39 156–157 (2021).

85. Amezquita, R., Lun, A., Hicks, S. & Gottardo, R. *Multi-Sample Single-Cell Analyses with Bioconductor*. https://bioconductor.org/books/3.14/OSCA.multisample/index.html (2021).

86. Ahlmann-Eltze, C. *Pseudobulk and Differential Expression*. https://bioconductor.org/packages/devel/bioc/vignettes/glmGamPoi/inst/doc/pseudobulk.html (2024).

87. Street, K. *et al*. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).

88. Trapnell, C. *et al*. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol*. **32**, 381–386 (2014).

89. Lin, C. & Bar-Joseph, Z. Continuous-state HMMs for modeling time-series single-cell RNA-Seq data. *Bioinformatics* **35**, 4707–4715 (2019).

90. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol*. **38**, 1408–1414 (2020).

91. Ding, J., Sharon, N. & Bar-Joseph, Z. Temporal modelling using single-cell transcriptomics. *Nat. Rev. Genet*. **23**, 355–368 (2022).

92. Tewarie, P., van Dellen, E., Hillebrand, A. & Stam, C. J. The minimum spanning tree: an unbiased method for brain network analysis. *Neuroimage* **104**, 177–188 (2015).

93. Van den Berge, K. *et al*. Trajectory-based differential expression analysis for single-cell sequencing data. *Nat. Commun*. **11**, 1201 (2020).

94. Song, D. & Li, J. J. PseudotimeDE: inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell RNA sequencing data. *Genome Biol*. **22**, 124 (2021).

95. Leary, J. R. & Bacher, R. Interpretable trajectory inference with single-cell Linear Adaptive Negative-binomial Expression (scLANE) testing. *bioRxiv* 2023.12.19.572477 (2023) doi:10.1101/2023.12.19.572477.

96. Larsen, K. GAM : The Predictive Modeling Silver Bullet. (2015).

97. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med*. **9**, 75 (2017).

98. Adil, A., Kumar, V., Jan, A. T. & Asger, M. Single-Cell Transcriptomics: Current Methods and Challenges in Data Acquisition and Analysis. *Front. Neurosci*. **15**, 591122 (2021).

99. Yu, X., Abbas-Aghababazadeh, F., Chen, Y. A. & Fridley, B. L. Statistical and Bioinformatics Analysis of Data from Bulk and Single-Cell RNA Sequencing Experiments. *Methods Mol. Biol*. **2194**, 143–175 (2021).

100. Cheng, Y., Ma, X., Yuan, L., Sun, Z. & Wang, P. Evaluating imputation methods for single-cell RNA-seq data. *BMC Bioinformatics* **24**, 302 (2023).

101. Huang, D. *et al*. Advances in single-cell RNA sequencing and its applications in cancer research. *J. Hematol. Oncol*. **16**, 98 (2023).

102. Lafzi, A., Moutinho, C., Picelli, S. & Heyn, H. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. *Nat. Protoc*. **13**, 2742–2757 (2018).

103. Zee, A. *et al*. Sequencing Illumina libraries at high accuracy on the ONT MinION using R2C2. *Genome Res*. **32**, 2092–2106 (2022).

104. Molloy, E. J. & Bearer, C. F. Translational research is all-encompassing and lets everyone be a researcher. *Pediatr. Res*. **90**, 2–3 (2021).

105. Vandenbergh, J. G. Use of House Mice in Biomedical Research. *ILAR J*. **41**, 133–135 (2000).

106. Mukherjee, P., Roy, S., Ghosh, D. & Nandi, S. K. Role of animal models in biomedical research: a review. *Lab. Anim. Res.* **38**, 18 (2022).

107. Jilka, R. L. The relevance of mouse models for investigating age-related bone loss in humans. *J. Gerontol. A Biol. Sci. Med. Sci.* **68**, 1209–1217 (2013).

108. Ruberte, J. *et al*. Bridging mouse and human anatomies; a knowledge-based approach to comparative anatomy for disease model phenotyping. *Mamm. Genome* **34**, 389–407 (2023).

109. Bryda, E. C. The Mighty Mouse: the impact of rodents on advances in biomedical research. *Mo. Med.* **110**, 207–211 (2013).

110. Gurumurthy, C. B. & Lloyd, K. C. K. Generating mouse models for biomedical research: technological advances. *Dis. Model. Mech.* **12**, (2019).

111. What is a mouse model? *The Jackson Laboratory* https://www.jax.org/why-the-mouse/model.

112. Saul, M. C., Philip, V. M., Reinholdt, L. G., Center for Systems Neurogenetics of Addiction & Chesler, E. J. High-Diversity Mouse Populations for Complex Traits. *Trends Genet.* **35**, 501–514 (2019).

113. Brekke, T. D., Steele, K. A. & Mulley, J. F. Inbred or Outbred? Genetic Diversity in Laboratory Rodent Colonies. *G3* **8**, 679–686 (2018).

114. Bogue, M. A., Churchill, G. A. & Chesler, E. J. Collaborative Cross and Diversity Outbred data resources in the Mouse Phenome Database. *Mamm. Genome* **26**, 511–520 (2015).

115. Churchill, G. A., Gatti, D. M., Munger, S. C. & Svenson, K. L. The Diversity Outbred mouse population. *Mamm. Genome* **23**, 713–718 (2012).

116. Morgan, A. P. *et al*. The Mouse Universal Genotyping Array: From substrains to subspecies. *G3 (Bethesda)* **6**, 263–279 (2015).

117. Broman, K. W. *et al*. R/qtl2: Software for Mapping Quantitative Trait Loci with High-Dimensional Data and Multiparent Populations. *Genetics* **211**, 495–502 (2019).

118. Li, Q., Xu, R., Lei, K. & Yuan, Q. Insights into skeletal stem cells. *Bone Res* **10**, 61 (2022).

119. Seita, J. & Weissman, I. L. Hematopoietic stem cell: self-renewal versus differentiation. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **2**, 640–653 (2010).

120. Zhao, E. *et al*. Bone marrow and the control of immunity. *Cell. Mol. Immunol.* **9**, 11–19 (2012).

121. Comazzetto, S., Shen, B. & Morrison, S. J. Niches that regulate stem cells and hematopoiesis in adult bone marrow. *Dev. Cell* **56**, 1848–1860 (2021).

122. Bianco, P. & Robey, P. G. Skeletal stem cells. *Development* **142**, 1023–1027 (2015).

123. Salhotra, A., Shah, H. N., Levi, B. & Longaker, M. T. Mechanisms of bone development and repair. *Nat. Rev. Mol. Cell Biol.* **21**, 696–711 (2020).

124. Boyle, W. J., Simonet, W. S. & Lacey, D. L. Osteoclast differentiation and activation. *Nature* **423**, 337–342 (2003).

125. Kronenberg, H. M. Developmental regulation of the growth plate. *Nature* **423**, 332–336 (2003).

126. Church, C., Horowitz, M. & Rodeheffer, M. WAT is a functional adipocyte? *Adipocyte* **1**, 38–45 (2012).

127. Suchacki, K. J. *et al*. Bone marrow adipose tissue is a unique adipose subtype with distinct roles in glucose homeostasis. *Nat. Commun.* **11**, 3097 (2020).

128. Li, Z. & MacDougald, O. A. Preclinical models for investigating how bone marrow adipocytes influence bone and hematopoietic cellularity. *Best Pract. Res. Clin. Endocrinol. Metab.* **35**, 101547 (2021).

129. Woods, K. & Guezguez, B. Dynamic Changes of the Bone Marrow Niche: Mesenchymal Stromal Cells and Their Progeny During Aging and Leukemia. *Front Cell Dev Biol* **9**, 714716 (2021).

130. Chen, Q. *et al*. Fate decision of mesenchymal stem cells: adipocytes or osteoblasts? *Cell Death Differ.* **23**, 1128–1139 (2016).

131. Fazeli, P. K. *et al*. Marrow fat and bone--new perspectives. *J. Clin. Endocrinol. Metab.* **98**, 935–945 (2013).

132. Veldhuis-Vlug, A. G. & Rosen, C. J. Clinical implications of bone marrow adiposity. *J. Intern. Med.* **283**, 121–139 (2018).

133. Abdallah, B. M., Alzahrani, A. M., Abdel-Moneim, A. M., Ditzel, N. & Kassem, M. A simple and reliable protocol for long-term culture of murine bone marrow stromal (mesenchymal) stem cells that retained their in vitro and in vivo stemness in long-term culture. *Biol. Proced. Online* **21**, 3 (2019).

134. Bhat, S., Viswanathan, P., Chandanala, S., Prasanna, S. J. & Seetharam, R. N. Expansion and characterization of bone marrow derived human mesenchymal stromal cells in serum-free conditions. *Sci. Rep.* **11**, 3403 (2021).

135. Gimble, J. M. *et al*. In vitro Differentiation Potential of Mesenchymal Stem Cells. *Transfus. Med. Hemother.* **35**, 228–238 (2008).

136. Dillard, L. J. *et al*. Single-Cell Transcriptomics of Bone Marrow Stromal Cells in Diversity Outbred Mice: A Model for Population-Level scRNA-Seq Studies. *J. Bone Miner. Res.* **38**, 1350–1363 (2023).

137. Feng, K. *et al*. Multi-omics analysis of bone marrow mesenchymal stem cell differentiation differences in osteoporosis. *Genomics* **115**, 110668 (2023).

138. Ilas, D. C. *et al*. The osteogenic commitment of CD271+CD56+ bone marrow stromal cells (BMSCs) in osteoarthritic femoral head bone. *Sci. Rep.* **10**, 11145 (2020).

139. Huang, T. *et al*. Inhibition of osteogenic and adipogenic potential in bone marrow-derived mesenchymal stem cells under osteoporosis. *Biochem. Biophys. Res. Commun.* **525**, 902–908 (2020).

140. Lin, J. T. & Lane, J. M. Osteoporosis: a review. *Clin. Orthop. Relat. Res.* 126–134 (2004).

141. Peacock, M., Turner, C. H., Econs, M. J. & Foroud, T. Genetics of osteoporosis. *Endocr. Rev.* **23**, 303–326 (2002).

142. Zhu, X., Bai, W. & Zheng, H. Twelve years of GWAS discoveries for osteoporosis and related traits: advances, challenges and applications. *Bone Res* **9**, 23 (2021).

143. Tam, V. *et al*. Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).

144. Mahajan, A. *et al*. Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat. Genet.* **50**, 559–571 (2018).

145. Akiyama, M. Multi-omics study for interpretation of genome-wide association study. *J. Hum. Genet.* **66**, 3–10 (2020).

146. van der Sijde, M. R., Ng, A. & Fu, J. Systems genetics: From GWAS to disease pathways. *Biochim. Biophys. Acta* **1842**, 1903–1909 (2014).

147. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M. Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.* **10**, 184–194 (2009).

148. Giambartolomei, C. *et al*. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).

149. Wen, X., Pique-Regi, R. & Luca, F. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS Genet.* **13**, e1006646 (2017).

150. Li, B. & Ritchie, M. D. From GWAS to Gene: Transcriptome-Wide Association Studies and Other Methods to Functionally Understand GWAS Discoveries. *Front. Genet.* **12**, 713230 (2021).

151. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

152. Nguyen, A., Khoo, W. H., Moran, I., Croucher, P. I. & Phan, T. G. Single Cell RNA Sequencing of Rare Immune Cell Populations. *Front. Immunol.* **9**, 1553 (2018).

153. Papatheodorou, I. *et al*. Expression Atlas update: from tissues to single cells. *Nucleic Acids Res.* **48**, D77–D83 (2020).

154. Debnath, S. *et al*. Discovery of a periosteal stem cell mediating intramembranous bone formation. *Nature* **562**, 133–139 (2018).

155. Tikhonova, A. N. *et al*. The bone marrow microenvironment at single-cell resolution. *Nature* **569**, 222–228 (2019).

156. Zhong, L. *et al*. Single cell transcriptomics identifies a unique adipose lineage cell population that regulates bone marrow environment. *Elife* **9**, (2020).

157. Debnath, S. & Greenblatt, M. B. Specimen preparation for single-cell sequencing analysis of skeletal cells. *Methods Mol. Biol.* **2221**, 89–100 (2021).

158. Chai, R. C. Single-cell RNA sequencing: Unravelling the bone one cell at a time. *Curr. Osteoporos. Rep.* **20**, 356–362 (2022).

159. Hanna, H., Mir, L. M. & Andre, F. M. In vitro osteoblastic differentiation of mesenchymal stem cells generates cell layers with distinct properties. *Stem Cell Res. Ther.* **9**, 203 (2018).

160. Church, D. M. *et al*. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol.* **7**, e1000112 (2009).

161. Hao, Y. *et al*. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).

162. Heaton, H. *et al*. Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat. Methods* **17**, 615–620 (2020).

163. Tirosh, I. *et al*. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).

164. Piper, M., Mistry, M., Liu, J., Gammerdinger, W. & Khetani, R. *Hbctraining/ScRNA-Seq_online: ScRNA-Seq Lessons from HCBC (First Release)*. (2022). doi:10.5281/zenodo.5826256.

165. Andrews, S. *FastQC: A Quality Control Analysis Tool for High Throughput Sequencing Data*. (Github).

166. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).

167. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

168. Zhang, Y., Park, C., Bennett, C., Thornton, M. & Kim, D. Rapid and accurate alignment of nucleotide conversion sequencing reads with HISAT-3N. *Genome Res*. (2021) doi:10.1101/gr.275193.120.

169. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol*. **33**, 290–295 (2015).

170. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. **15**, 550 (2014).

171. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol*. **36**, 411–420 (2018).

172. Van de Sande, B. *et al*. A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat. Protoc*. **15**, 2247–2276 (2020).

173. Hoffman, P. *Seurat-Disk: Interfaces for HDF5-Based Single Cell File Formats*. (Github).

174. Moerman, T. *et al*. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics* **35**, 2159–2161 (2019).

175. Imrichová, H., Hulselmans, G., Atak, Z. K., Potier, D. & Aerts, S. i-cisTarget 2015 update: generalized cis-regulatory enrichment analysis in human, mouse and fly. *Nucleic Acids Res*. **43**, W57-64 (2015).

176. Aibar, S. *et al*. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).

177. Suo, S. *et al*. Revealing the Critical Regulators of Cell Identity in the Mouse Cell Atlas. *Cell Rep*. **25**, 1436-1445.e3 (2018).

178. Timshel, P. N., Thompson, J. J. & Pers, T. H. Genetic mapping of etiologic brain cell types for obesity. *Elife* **9**, (2020).

179. Finucane, H. K. *et al*. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet*. **47**, 1228–1235 (2015).

180. Thomas, P. D. *et al*. PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Sci*. **31**, 8–22 (2022).

181. Fiers, M. W. E. J. *et al*. Mapping gene regulatory networks from single-cell omics data. *Brief. Funct. Genomics* **17**, 246–254 (2018).

182. Dolgalev, I. & Tikhonova, A. N. Connecting the Dots: Resolving the Bone Marrow Niche Heterogeneity. *Front Cell Dev Biol* **9**, 622519 (2021).

183. Omatsu, Y. *et al*. The essential functions of adipo-osteogenic progenitors as the hematopoietic stem and progenitor cell niche. *Immunity* **33**, 387–399 (2010).

184. Zhong, L., Yao, L., Seale, P. & Qin, L. Marrow adipogenic lineage precursor: A new cellular component of marrow adipose tissue. *Best Pract. Res. Clin. Endocrinol. Metab*. **35**, 101518 (2021).

185. Marsh, S., Salmon, M. & Hoffman, P. *Custom Visualizations & Functions for Streamlined Analyses of Single Cell Sequencing*. (2023). doi:10.5281/zenodo.7534950.

186. Alquicira-Hernandez, J. & Powell, J. E. Nebulosa recovers single cell gene expression signals by kernel density estimation. *Bioinformatics* (2021) doi:10.1093/bioinformatics/btab003.

187. Greenfest-Allen, E., Cartailler, J.-P., Magnuson, M. A. & Stoeckert, C. J. iterativeWGCNA: iterative refinement to improve module detection from WGCNA co-expression networks. *bioRxiv* 234062 (2017) doi:10.1101/234062.

188. Chan, C. K. F. *et al*. Identification of the Human Skeletal Stem Cell. *Cell* **175**, 43-56.e21 (2018).

189. Mizuhashi, K. *et al*. Resting zone of the growth plate houses a unique class of skeletal stem cells. *Nature* **563**, 254–258 (2018).

190. Matsushita, Y. *et al*. A Wnt-mediated transformation of the bone marrow stromal cell identity orchestrates skeletal regeneration. *Nat. Commun.* **11**, 332 (2020).

191. Miki, H., Setou, M., Kaneshiro, K. & Hirokawa, N. All kinesin superfamily protein, KIF, genes in mouse and human. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 7004–7011 (2001).

192. Enerbäck, S., Ohlsson, B. G., Samuelsson, L. & Bjursell, G. Characterization of the human lipoprotein lipase (LPL) promoter: evidence of two cis-regulatory regions, LP-alpha and LP-beta, of importance for the differentiation-linked induction of the LPL gene during adipogenesis. *Mol. Cell. Biol.* **12**, 4622–4633 (1992).

193. Federico, L. *et al*. Lipid phosphate phosphatase 3 regulates adipocyte sphingolipid synthesis, but not developmental adipogenesis or diet-induced obesity in mice. *PLoS One* **13**, e0198063 (2018).

194. Maridas, D. E., DeMambro, V. E., Le, P. T., Mohan, S. & Rosen, C. J. IGFBP4 Is Required for Adipogenesis and Influences the Distribution of Adipose Depots. *Endocrinology* **158**, 3488–3500 (2017).

195. Sigg, M. A. *et al*. Evolutionary Proteomics Uncovers Ancient Associations of Cilia with Signaling Pathways. *Dev. Cell* **43**, 744-762.e11 (2017).

196. Kumar, R. *et al*. A cell-based GEF assay reveals new substrates for DENN domains and a role for DENND2B in primary ciliogenesis. *Sci Adv* **8**, eabk3088 (2022).

197. Fumoto, K. *et al*. Mark1 regulates distal airspace expansion through type I pneumocyte flattening in lung development. *J. Cell Sci.* **132**, (2019).

198. Zhang, R., Roostalu, J., Surrey, T. & Nogales, E. Structural insight into TPX2-stimulated microtubule assembly. *Elife* **6**, (2017).

199. Uusküla-Reimand, L. & Wilson, M. D. Untangling the roles of TOP2A and TOP2B in transcription and cancer. *Sci Adv* **8**, eadd4920 (2022).

200. Trueb, B. Biology of FGFRL1, the fifth fibroblast growth factor receptor. *Cell. Mol. Life Sci.* **68**, 951–964 (2011).

201. Hilgendorf, K. I. Primary Cilia Are Critical Regulators of White Adipose Tissue Expansion. *Front. Physiol.* **12**, 769367 (2021).

202. Neavin, D. *et al*. Single cell eQTL analysis identifies cell type-specific genetic control of gene expression in fibroblasts and reprogrammed induced pluripotent stem cells. *Genome Biol.* **22**, 76 (2021).

203. van der Wijst, M. G. P. *et al*. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* **50**, 493–497 (2018).

204. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

205. Cartailler, J. P. *Iterativewgcna*. (2022).

206. Porter, M. D. & Smith, R. Network neighborhood analysis. in *2010 IEEE International Conference on Intelligence and Security Informatics* (IEEE, 2010). doi:10.1109/isi.2010.5484781.

207. Phipson, B. *et al*. propeller: testing for differences in cell type proportions in single cell data. *Bioinformatics* **38**, 4720–4726 (2022).

208. Kim, D. & Rossi, J. RNAi mechanisms and applications. *Biotechniques* **44**, 613–616 (2008).

209. Xu, Y. & Li, Z. CRISPR-Cas systems: Overview, innovations and applications in human disease research and gene therapy. *Comput. Struct. Biotechnol. J.* **18**, 2401–2415 (2020).

210. Scholefield, J. & Harrison, P. T. Prime editing - an update on the field. *Gene Ther*. **28**, 396–401 (2021).