Integrating Genetic and Genomic Analyses to Identify Novel Genes with a Role in the Complex Genetic Architecture of Osteoporosis

Olivia Lauren Sabik

Cleveland, Ohio

Bachelor of Arts in Chemistry, Kenyon College, Gambier, Ohio, 2014

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

Department of Biochemistry and Molecular Genetics

University of Virginia

August 2019

Dr. Charles Farber

Dr. Stefan Bekiranov

Dr. Mete Civelek

Dr. Xiaowei Lu

Dr. Todd Stukenberg

Abstract

Osteoporosis is an increasingly prevalent global health burden characterized by decreased bone strength and increased risk of fracture. Despite its significant impact on human health, there is currently a lack of highly effective treatments free of side effects for osteoporosis. Genetic discovery has been shown to be effective method for the unbiased identification of novel drug targets and genome-wide association studies (GWASs) have begun to provide insight. Over the last decade, osteoporosis-related GWASs have led to the identification of approximately 1100 associations for bone mineral density (BMD) and other bone traits related to risk of fracture. However, there have been limited efforts to identify the causal genes and mechanisms underlying these GWAS associations and there is much left to learn from these studies. Additionally, osteoporosis is influenced not only by the regulation of BMD, but also by gross bone geometry and bone microarchitecture. These parameters are difficult to study using human GWAS, but instead rely on model organism studies, largely in the mouse.

Here, we explore the genetic determinants of osteoporosis, regulating both bone geometry and BMD. We used allele specific expression analysis in inbred mice to identify novel genes potentially influencing bone geometry. Additionally, we built new tools that produce visualizations of the colocalization of expression quantitative trait loci (eQTL) with GWAS associations. Finally, to investigate BMD GWAS, we integrated gene co-expression network analysis with the results of BMD GWAS to identify novel genes influencing BMD. Using both computational and experimental methods, we discovered novel genes influencing osteoporosis and have developed a platform from which we can better understand the mechanisms that underlie bone fragility and increased risk of fracture.

Dedication

To my parents, Cindy and Mark, my sisters, Lindsay and Natalie, and their families: I could not have produced this work without your love, support, and encouragement. I am forever grateful for everything you have gifted me with: mom, for imbuing in me an intense appreciation for education and modeling lifelong learning; dad, for instilling curiosity and confidence in me and helping me learn how to persevere; Linds and Nat, for always picking up the phone, being there for me, and for your guidance throughout this process; and to Matt, Justin, Ellery, Rowan, and Lucy, thank you for helping me have fun and stay grounded.

To Jamie, for enduring endless descriptions of the bugs in my code and the experimental mishaps. You never fail to brighten my day and for that I am forever grateful.

To my mentor, Charles Farber, for providing the guidance and support I needed to grow as a scientist. Thank you for pushing me to develop new skills, investigate leads in new areas, and become an independent computational biologist. To the members of the Farber lab, thank you for helping me both technically and emotionally. To Gina, thank you for being a great coworker and friend. To Larry, thank you for being a sounding board for my experiments. To Basel, thank you for sharing your scripts and ideas. To Eric, thank you for your dedication. And to Atum and Arby, thanks for making the lab fun.

And to my high school, college, CUDO, and BIMS friends, thank you for all the fun we've had these past five years. You've brought so much joy and light into my life.

Thank you all.

Table of Contents

List of Figures

List of Abbreviations

BMD	bone mineral density
GWAS	genome wide association studies
RACER	Regional Association ComparER
SNP	single nucleotide polymorphism
TFBS	transcription factor binding sites
DEXA	dual-energy x-ray absorptiometry
GEFOS	GEnetic Factors for OSteoporosis
LS	lumbar spine
FN	femoral neck
eBMD	estimate bone mineral density
ONJ	osteonecrosis of the jaw
NS	non-synonymous
eQTL	expression quantitative trait locus
ENCODE	ENCyclopedia Of DNA Elements
GTEx	Gene Tissues EXpression
CAD	coronary artery disease
GRN	gene regulatory network
TF	transcription factor
uCT	micro computed tomography
QTL	quantitative trait locus
CC	collaborative cross
DO	diversity outbred
HG	high growth
Feml2	femur length 2
Chr	chromosome
CAST	CAST/EiJ
B6	C57BL/6J
HG9	congenic mouse, chr. 9 CAST on B6 background
LOD	logarithm of odds
UTR	untranslated region
FDR	false discovery rate

Mbp	mega base pairs
cM	centi Morgans
RNA	ribonucleic acid
ASE	allele-specific expression
glm	general linear model
GNU	GNU's Not Unix!
LD	linkage disequilibrium
PPH4	posterior probability of hypothesis 4
LDC	late differentiation cluster
OFM	osteoblast functional module
WGCNA	weighted gene co-expression network analysis
EDC	early differentiation cluster
GO	gene ontology
BMC	bone mineral content
OR	odds ratio
h_g^2	SNP-heritability
MSC	mesenchymal stem cells
IMPC	International Mouse Phenotyping Consortium
PPH4	posterior probability of hypothesis 4, colocalizing QTL
MGI	Mouse Genome Informatics
OBCD	Origins of Bone and Cartilage Disease
BV/TV	bone volume fraction (bone volume/total volume)
KOMP	Knockout Mouse Phenotyping Consortium
iPSCs	induced pluripotent stem cells
GS	gene significance
MM	module membership

Chapter 1
Introduction

Published in part in: Sabik, O. L. & Farber, C. R. Using GWAS to identify novel therapeutic targets for osteoporosis. *Translational Research* (2016). doi: 10.1016/j.trsl.2016.10.009

1.1 Epidemiology and current therapeutics for osteoporosis

Osteoporosis is a disease of weakened bone, clinically characterized by low bone mineral density (BMD) and an increased risk for fracture [1]. Osteoporotic fractures are a major public health burden [2,3] and as a larger fraction of the population reaches old age, the annual rate of fractures and associated costs in the United States are projected to rise as much as 48% by 2025, resulting in approximately 3 million fractures and $25.3 billion in health care costs annually [4]. This bleak outlook has led to increased efforts to develop a more effective means of treating and preventing bone disease.

The majority of existing therapeutics for osteoporosis are antiresorptives, such as bisphosphonates [5], which inhibit osteoclast-mediated bone resorption [6]. While these drugs are effective in halting bone loss and further increases in fracture risk, they are prescribed upon diagnosis, after significant bone loss has already occurred [7]. As a result, anabolic agents that build new bone are needed. Teriparatide, an injected peptide that targets the parathyroid hormone receptor, has been shown to induce bone formation [8]. However, the need for a daily injection makes this a difficult course of treatment [9]. Romosozumab, an antibody that targets the Wnt signaling inhibitor sclerostin, was just approved for clinical use [10]. The drug has been shown to reduce the incidence of fracture, however side effects have been observed for romosozumab [11]. Side effects associated with existing therapeutics, such as atypical femoral fracture and osteonecrosis of the jaw [12,13], though rare, have led to a marked decrease in preventative use [14]. Given these many disincentives for using the current treatments, the identification of novel anabolic therapeutic targets that can be affected via orally active drugs is a major goal in the field.

Table 1.1 Anti-osteoporotic drug targets that have been linked to changes in BMD by GWAS.

Drug Class	Drug Target	Target gene implicated by GWAS	Refs
Denosumab	RANKL	RANKL	15
Sclerostin inhibitors	Sclerostin (SOST)	SOST	16
Selective oestrogen receptor modulators	Oestrogen receptor	ESR1	17
Parathyroid hormone analogues	Parathyroid hormone (PTH) receptor	Not identified, but pathway highlighted by PTH-like hormone and PTH-related protein	8,18
Bisphosphonates	Farnesyl pyrophosphate	Not identified	19
Oestrogen	Oestrogen receptor	ESR1	20
Cathepsin K inhibitors	Cathepsin K	Not identified	21
Dickkopf 1 (DKK1) inhibitors	DKK1	DKK1	22

Table adapted from [23].

Most of the current anti-osteoporotic therapies (**Table 1.1**) were identified using traditional molecular approaches and mouse knockout screens [9,23,24]. Specific genes and pathways known to play a role in bone maintenance were tested for effects on bone cell function *in vitro* and bone mass *in vivo* [25]. While this method has been effective, in today's world of vast genetic and genomic tools there may be more efficient ways to identify novel drug targets. In fact, a recent retrospective analysis revealed that drugs targeting a wide range of diseases that had been implicated through genetic studies were almost twice as likely to succeed in the drug development pipeline than those not identified using genetic approaches [26]. The increased success rate of targets supported by genetic evidence may be due to the fact that genetic studies provide a way to identify genes that, when modified, lead to an observable clinical effect not compensated for by other genes. Thus, genetic and genomic approaches provide an avenue for the unbiased discovery of novel drivers and regulators of specific biological processes and diseases. As described below, genome-wide association studies (GWAS) for BMD and mouse studies of other osteoporosis-related traits have given us a wealth of potential new drug targets.

1.2 Genetic studies are potentially powerful approaches to identify anti-osteoporotic therapeutics

The utility of genetics studies, in particular GWAS, are often called into question, primarily because GWASs do not directly identify causal genes and mechanisms regulating the trait and the effects that are identified are often small [27,28]. In light of such criticism it is useful to define why genetic studies are important. There are three general ways in which information from genetic studies can provide important biological and clinical insight. First,

genetic information can be used to "personalize" medicine. In theory, once we define the genetic architecture of a disease, this information could be used to identify at-risk individuals, for example by calculating polygenic risk scores [29]. Information on variants that impact an individual's response to treatment would also be invaluable as a clinical decision-making tool [30]. For instance, genetic diagnostics that identified individuals more likely to develop osteonecrosis of the jaw (ONJ) or atypical femoral fractures upon taking bisphosphonates would be of enormous clinical utility. Second, genetic studies inform biology and identifying novel genes is important to develop a comprehensive understanding of osteoporosis and other diseases. The utility of GWAS for this purpose will only grow as we develop robust methods for moving from loci to genes to disease mechanisms. It is also important to highlight that genetic studies differ from more traditional molecular gene discovery approaches in that they are unbiased. The importance of the unbiased nature of such studies is underscored by the fact that the majority of loci identified by most GWASs do not contain known genes related to the phenotype of interest, highlighting that these studies lead to the identification of novel biological insight. Third, possibly the most important use of genetic studies, is in the search for new anti-osteoporotic therapeutics. Current drug discovery paradigms have been hindered by the high attrition rate in the development pipeline. Most targets have been identified by non-genetic studies and as described above, evidence suggests that drug targets implicated by GWAS are twice as likely to succeed in clinical trials [26]. Importantly, the success rate may be even higher for osteoporosis given that five out of the eight (63%) anti-osteoporosis therapeutics currently approved or in advanced clinical trials are supported by genetic data (**Table 1.1**) [23]. Thus, genetic and genomic approaches to identifying drug targets are not only feasible, but are likely more successful than other avenues of anti-osteoporotic target identification.

1.3 Our current understanding of the genetic architecture of osteoporosis

Osteoporosis is a complex disease, influenced by numerous traits that determine bone strength, including bone mineral density (BMD), bone geometry, and bone microarchitecture [31]. While these traits are influenced in part by the environment, they are also among the most heritable disease-associated quantitative traits (h^2>0.50) [31–34]. Osteoporosis-associated traits are highly polygenic [35–37] and genetic studies have been carried to identify genetic variants and genes that affect osteoporosis-related traits; however, identifying the underlying causal variants and genes has proven difficult [38].

Recently, the omnigenic model of the genetic architecture underlying complex traits was developed to explain the observations made in human GWAS and provide a framework for understanding the results of GWAS [39,40]. The main principle underlying the omnigenic model is that all genes expressed in disease-relevant tissues will contribute to disease risk, resulting in a GWAS association; however, only a subset of these, termed "core" genes, play a direct role in disease etiology. The rest, termed "peripheral" genes, while statistically associated with the phenotype via GWAS, do not play a direct role in regulating the phenotype. The effect of peripheral genes on the phenotype is mediated by core genes. This theory of "core" and "peripheral" genes is supported by observations made in GWASs. Though core genes are the most directly related to the phenotype biologically, variants associated with disease-relevant genes do not harbor the majority of the heritability of a trait. The majority of the heritability for complex traits is spread across the genome, generally enriched in regions of active transcription and depleted in regions that are repressed in relevant cell-types. This result underlies the hypothesis that all actively expressed genes in disease-relevant tissues have an impact on the phenotype. However, this also indicates that not all causal GWAS-implicated genes are equally important, from a mechanistic and a

therapeutic standpoint. Given their direct role in regulating a phenotype of interest, we are particularly interested in identifying core genes.

Core genes typically have a biologically interpretable role in the processes involved in disease, as they likely play a role in the biological processes that determine the disease phenotype. Moreover, perturbations of core genes are expected to produce profound effects on phenotypes. Given the expected link between core gene function and the phenotype of interest, it is also reasonable to assume that these genes may be better drug targets [39]. Thus, identifying core genes for osteoporosis-related traits, such as BMD, may be critical to developing novel, effective therapeutics for osteoporosis. However, genetic studies alone do not provide enough information to distinguish core genes from peripheral genes. Therefore, new approaches for the identification of core genes are necessary.

The utility of the core gene designation has been debated. Some believe that the definition of core genes, as genes whose "product (protein, or RNA for a noncoding gene) has a direct effect--not mediated through regulation of another gene--on cellular and organismal processes leading to a change in the expected value of a particular phenotype" [40], is too narrow and that limiting our study of complex traits to a discrete set of core genes will restrict our understanding of the genetic architecture of complex disease [41]. Others argue that, if we have not yet identified the cellular and organismal processes that influence a disease, we are biased in our search for core genes, ignoring biology we do not yet understand [42]. However, core genes are not simply defined by their membership in a key biological pathway known to influence a trait, or by their statistically unconditional effect on the trait of interest. Core genes have other predicted properties, for example, perturbation of a core gene is predicted to have a large effect on the trait of interest [39,40]. Thus, it may be possible to leverage the predicted biological properties of core genes, without applying a

restrictive definition or bias toward known biology, to further identifying the genes, mechanisms, and pathways that drive osteoporosis.

1.4 Genetic Approaches for studying osteoporosis-related traits

As described above, current therapeutics for osteoporosis are less than optimal [14] and genetic studies have been shown to be effective in identifying drug targets with a greater than average chance of making it through the development pipeline [26]. Thus, numerous genetic studies have been conducted to identify novel biology underlying the etiology of osteoporosis. Here, we describe GWAS in humans and genetic approaches in the mouse used to study osteoporosis.

1.4.1 Genome-wide association studies for BMD

Before the development of GWAS, disease-influencing loci were mapped in humans using family-based linkage studies [31]. Linkage studies track the co-segregation of genetic markers with genetic variants that influence disease in families. The advantage of this approach is that large chromosomal regions are co-inherited in families, allowing disease alleles to be "tagged" using just a few hundred genetic markers spaced across the genome. This was critical because at the time, only a small number of genetic markers were known and only a limited number of variants could be easily genotyped, so linkage was the state of the art. The disadvantage of linkage is that, due to the small number of recombinations breaking up chromosomes in families, the loci identified were large, containing hundreds of genes. Two critical advances allowed for the development of more high-resolution genetic approaches; (1) the completion of the Human Genome Project in 2003 [43] and the International HapMap Project [44] in 2005, which provided a large list of reference single

nucleotide polymorphisms (SNPs) that could be used as genetic markers [45] and (2) the development of massively parallel genotyping assays which allowed for the rapid and cost-effective collection of hundreds of thousands of SNPs in large numbers of samples [46]. These two advances made it possible to use GWAS to "scan" for associations between millions of SNPs across the genome and phenotypes in large populations. In unrelated individuals, historical recombinations have "chopped" the genome into small blocks of co-inherited variants, and as a result, GWAS associations typically identify small genomic regions, implicating just a few genes. Thus, as genotyping large cohorts became technologically and economically feasible, GWAS have become the most common and effective means of investigating the genetic basis of common diseases.

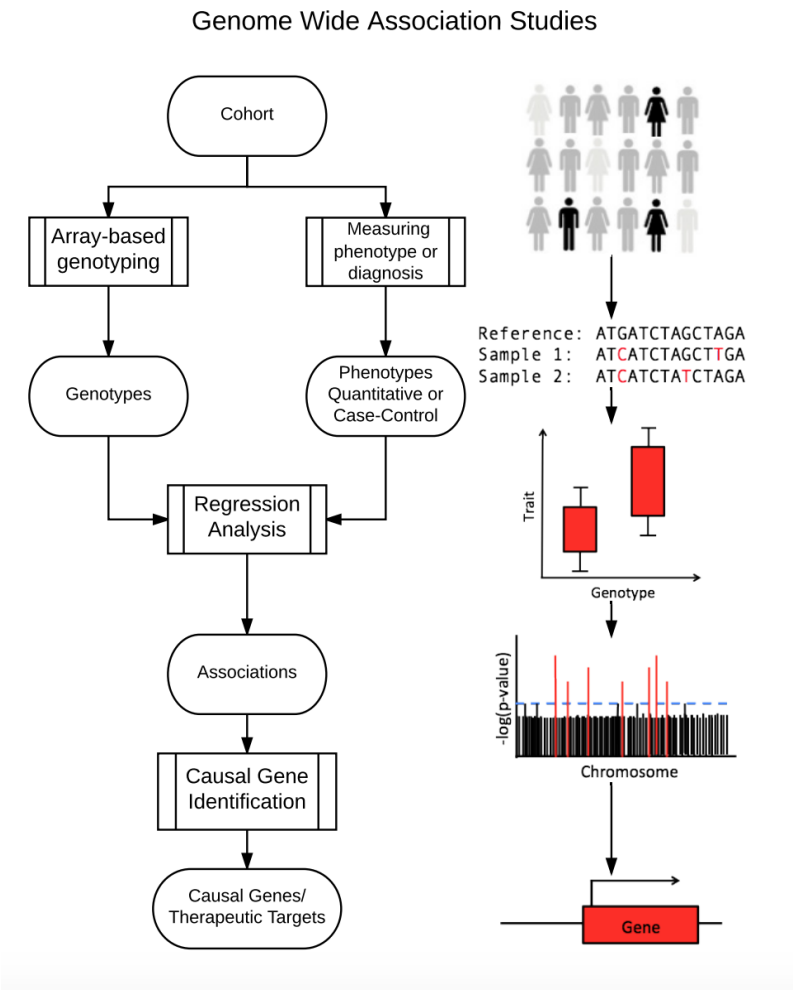


Figure 1.1 *Workflow for genome wide association studies (GWAS). A cohort of subjects, either cases, who are diagnosed with disease, and controls, who are healthy, or a group of people who vary in a quantitative trait, are genotyped at a large number of SNPs. Next, for a quantitative trait, a regression analysis is used to identify differences in phenotype as a function of genotype for millions of genetic variants. In case-control studies, the allele frequencies of variants between disease-diagnosed and disease-free subjects are compared. Of these comparisons, those that are statistically significant are called associations; these represent genetic loci that are associated with changes in the quantitative trait or the disease phenotype. Finally, the genes within these associated regions are studied in order to identify the causal genes that impact the phenotype. This process is outlined in Fig. 1.2.*

GWAS are performed by genotyping hundreds of thousands of SNPs, often in cohorts of tens or hundreds of thousands of individuals [47] (**Figure 1.1**). In addition to genotyped SNPs, non-genotyped SNPs are often imputed into a cohort [48]. Imputation is the process of predicting the genotype of untyped SNPs and is possible because we have detailed maps of how variants are co-inherited in specific populations [44,49]. Thus, it is common for a GWAS to test approximately 5-10 million common SNPs for disease associations [50]. GWAS are performed using either a case-control or quantitative trait design [51]. In the former, allele frequencies of SNPs are compared between genotyped cases and controls. For disease-related quantitative traits, SNP genotypes are tested using regression-based approaches, which determine if SNP allele dosage is associated with a change in phenotype. As of June 2019, the GWAS Catalog currently holds ~3989 GWAS publications, identifying ~138,000 associations for various traits [52]. This expansion will only continue as the cost of genotyping and the methods for phenotyping continue to improve, and as resources such as the UK BioBank continue to collect data in ever-growing cohorts [53].

The end result of a GWAS is a set of loci that harbor genetic variants that influence a disease. This is an important distinction as it is often assumed that GWAS identifies specific genes. In actuality, GWAS is just the first step in uncovering genes and variants contributing to a disease. GWAS associations pinpoint sets of SNPs in linkage disequilibrium (SNPs that are frequently co-inherited in a population). The resulting regions of association vary in size and typically implicate multiple genes, only a subset of which drive the observed phenotypic differences. Thus, the challenge of GWAS lies in identifying the causal genes and understanding the mechanism by which these genes affect the phenotype. Some of these associated SNPs are located in exons, or coding regions of genes and can impact protein function. It is often easy to identify the mechanism by which coding SNPs affect the

phenotype, however the majority of loci identified through GWAS implicate non-coding SNPs [54], suggesting they alter gene regulation [55]. Regulatory variation is particularly challenging to mechanistically characterize for several reasons. For instance, SNPs can affect regulation by altering transcription factor binding sites (TFBSs) [56,57], the 3-dimensional structure of the genome [58], or alternative splicing [59]. Additionally, the field has yet to develop robust approaches to identify regulatory sequences and methods to study the differences in function between two alternative regulatory sequences [60–62]. Thus, it remains challenging dissect how GWAS associations impact disease and the functional entities that mediate the effects.

Since 2007, several GWAS have been performed for bone phenotypes. Most notably, BMD has been the trait of choice for GWAS, primarily due to its high heritability (h^2>0.50) [63,64], association with fracture [65] and ease of measurement in large cohorts. There have been over 30 GWAS conducted for BMD leading to the identification of over 1100 independent associations [23,35–37,66]. The largest GWAS for BMD measured by dual-energy x-ray absorptiometry (DEXA) was conducted by GEnetic Factors for OSteoporosis (GEFOS) Consortium [35]. The GEFOSII study was a meta-analysis of lumbar spine (LS) and femoral neck (FN) BMD in 17 separate cohorts, identifying 56 associations for BMD. GEFOSII used a two-stage design that included both a discovery and replication cohort. The discovery cohort consisted of ~32,000 individuals and the effects of the most significant SNPs were then replicated in ~50,000 individuals. More recent GWAS for "estimated" BMD (eBMD), measured using heel ultrasound as a part of the UK BioBank project, have larger sample sizes, and thus have identified many more associations. The Kemp *et al.* GWAS analyzed eBMD data from 142,487 individuals and identified 307 conditionally independent associations [36] and the Morris *et al.* GWAS included 426,824 individuals and identified 1103

independent associations [37]. The results of these eBMD studies encompass 84% of loci previously identified in BMD GWAS using DEXA [35–37].

Several important observations were made in the Estrada *et al.* study and confirmed in the Kemp *et al.* and Morris *et al.* studies. First, the effects of individual loci on BMD and eBMD are small. In aggregate, the 64 loci identified in the Estrada *et al.* study explained less than 5% of the phenotypic variance in LSBMD and FNBMD and the 1103 identified in the Morris *et al.* study explained 20% of the variation in eBMD [35,37]. The results of these studies indicate that BMD is a complex trait, influenced by thousands of loci, each contributing a small effect on BMD. The complex genetic architecture of BMD is further supported in mouse studies demonstrating that roughly 10% of random gene knockouts have BMD or other skeletal phenotypes [67], and by GWAS for human height in >250,000 people which identified nearly 700 independent loci [68]. Second, approximately half of the 64 BMD loci identified in the Estrada *et al.* study harbored genes known to be involved in BMD, while the rest harbored only genes not previously implicated in the regulation of BMD. In the Morris *et al.* study, 1103 associations were identified and the majority contain no gene known to influence BMD. As indicated above, it can be difficult to identify which genes are truly causal, but for the loci harboring known genes we expect many of these to play a role. It is also possible that a subset of loci contain more than one causal gene. Generally, genes in GWAS loci that are known to play a role in BMD include: (1) members of the beta-catenin/Wnt signaling pathway which regulates osteoblasotgenesis, osteoblast proliferation, and apoptosis of osteoblasts and osteoclasts [69,70], (2) the receptor activator of nuclear factor-κB (RANK), RANK ligand (RANKL), and osteoprotegerin (OPG) pathway, which regulates the relationship between osteoblast and osteoclast activity in bone remodeling [71,72], and (3)

developmental genes involved in the process of endochondral ossification, namely transcription factors which induce expression of key genes in the ossification process [23,73].

The magnitude of effects of genetic variants fall on a continuous spectrum ranging from single "Mendelian" mutations of large effect size that cause diseases like osteogenesis imperfecta [74], to the small effects identified by GWAS. Though still a contentious matter of debate, it is likely that most diseases and quantitative traits are influenced by variants along the entire spectrum from small to large. For example, rare, large-effect variants have been identified in Wnt Family Member 1 (WNT1) in individuals with very low BMD [75]. Additionally, genome sequencing and association studies in large cohorts have identified variants near the Engrailed Homeobox 1 (EN1), Leucine-Rich Repeat Containing G Protein-Coupled Receptor 4 (LGR4) and Collagen Type I Alpha 2 (COL1A2) genes that are rare and have relatively large effects on BMD [76–78]. However, the evidence is mounting that most of the genetic component of BMD is due to large number of common variants of small effects [35], and the omnigenic model posits that the contributions may be largely peripheral to the relevant biology, with a subset of variants regulating core genes.

While GWAS have identified over 1100 associations for BMD, very few of these associations have been linked to a gene or mechanism influencing BMD. There is still much information to be gleaned from these studies, however it will require novel approaches to follow up on these results.

1.4.2 Murine studies of osteoporosis related-traits

While the results of BMD GWAS promise to open new doors of investigation in the bone field and uncover novel therapeutic targets, it is important to point out that BMD is not an "ideal" osteoporosis phenotype. For example, some patients who suffer from low

BMD do not experience osteoporotic fracture, and others who have normal BMD do experience fractures [31]. Additionally, it has been shown that ~50% of the variance in bone strength, the main determinant of an individual's risk of fracture, is due to BMD, whereas the other half of the variance is due to parameters such as bone size, geometry and tissue-level properties [79]. Thus, there are a number of groups that are expanding to alternative phenotypes, such as trabecular and cortical microarchitecture defined by micro computed tomography (uCT), to capture genetic influences on bone strength that are independent of BMD [80,81]. As mentioned above, samples sizes of tens of thousands of subjects are needed to identify the small effects of common genetic variants. One of the reasons that GWAS has been successful for BMD is the ability to assemble very large cohorts. Therefore, it is unclear how successful GWAS for traits other than BMD will be due to the difficulty in measuring most bone traits in the large number of subjects (N>10,000) needed for sufficient statistical power to detect genetic effects. It is also likely that case-control GWAS for osteoporotic fracture will help to fill the gap, though the small number of studies performed to date have identified few loci, likely because fracture is a noisy phenotype [82,83]. Fracture is a noisy phenotype for GWAS because frequently, fractures from all sites are lumped together in a single analysis, introducing more noise. Thus, many researchers have turned to mice in order to identify novel genes influencing bone strength-related traits other than BMD. A number of reviews have been written on this topic [9,84,85], however, briefly, the approaches fall into just a few categories: (1) using congenic mouse strains, (2) mapping bone phenotypes using recombinant inbred strains, and (3) using mutagenesis and gene knockout screens to identify models of bone disease.

Congenic mouse strains are strains in which a quantitative trait locus (QTL) harboring variants influencing a phenotype are bred from a donor strain onto a recipient

strain. This is achieved by backcrossing the progeny of a cross between two divergent strains to one of the original strains, isolating the effect of a single QTL from the donor strain [86]. For example, QTLs for high bone mass identified in C3H/HeJ mice have been transferred onto the low bone mass C57BL/6J line, leading to the identification of four loci harboring variation that influences BMD [87]. By continuously backcrossing to isolate smaller and smaller subsets of the region, the minimal region required to produce the phenotype could be identified, thus narrowing the search space within the QTL. Though congenic strains were crucial in the discovery of genetic regions influencing bone phenotypes, the processes of breeding such strains is time consuming, and more advanced methods, utilizing high throughput genotyping and phenotyping to map traits, expedited the process.

There have been numerous examples of mapping experiments using crosses of two classic inbred strains that have led to the identification of novel genes influencing BMD [88–91] and disease associated traits other than BMD as well, for example microarchitectural traits [91]. Recently, more complex genetic reference populations have been leveraged in studies of the genetic basis of bone microarchitecture [81], including the collaborative cross (CC) and the diversity outbred (DO) populations, which are multi-parental recombinant inbred lines. However, these approaches require profiling hundreds of mice, and are also costly.

Finally, high-throughput phenotyping assays paired with advances in gene knockout technology have enabled projects like the International Knockout Mouse Consortium and the International Mouse Phenotyping Consortiums [67,92]. These studies have identified hundreds of genes that produced a bone phenotype with knocked out and publicly reported the results in online databases. This data can be used to follow up on human and mouse genetic studies of bone mineral density and other osteoporosis-related traits. Mice are extremely useful for testing the impact of changes to a single gene, experiments that are not

possible to conduct in humans, making them an ideal complementary model system for the study of complex traits in humans.

1.5 Causal gene discovery

Given the importance of translating genetic associations into biological knowledge, how does one go from locus to gene to mechanism? Below we outline state-of-the-art approaches applicable in both mouse and human studies and discuss how they can be used to inform genetic associations for osteoporosis-related traits.

1.5.1 "Direct" approaches – fine-mapping, annotating SNPs, eQTLs

There is not a standard "pipeline" one can use to go from genetic association to causal variants/genes (**Figure 1.2**). Instead, many different approaches may be taken depending on the disease, characteristics of the locus under investigation, and the resources available. In general, though, it is desirable to both decrease the number of potentially causal variants and link those variants to either changes in protein function or, as is more often the case, an alteration in gene regulation.

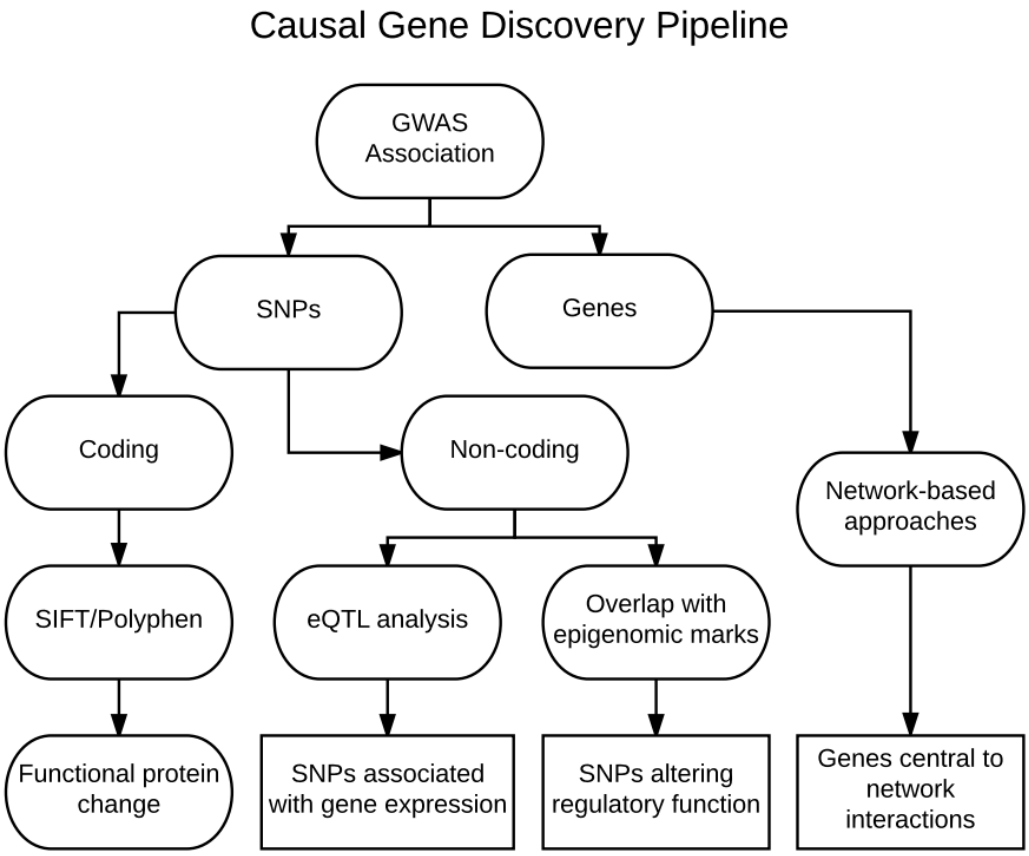


Figure 1.2 *Workflow for causal gene discovery. As described in Fig. 1.1, GWAS result in the identification of genomic loci associated with a trait or disease. Both the significantly associated SNPs and the genes within the region can be identified. If a SNP is identified within the coding region of a gene, computational techniques can be employed to determine the likelihood that the SNP causes a functional change in the protein it produces. If the SNP lies in a non-coding region, it could play a role in regulating gene expression. This hypothesis can be supported by relating the SNP to gene expression using expression quantitative trait locus (eQTL) analysis, or by SNP colocalization with regulatory epigenomic marks. The genes within the region can also be integrated with network information to identify potentially causal genes.*

Reducing the number of potential causal variants is referred to as fine-mapping. An example of fine-mapping is performing targeted complete re-sequencing of a GWAS locus in subset of study participants and then genotyping the identified variants in the entire cohort. By increasing the number of SNPs investigated in a region it may be possible to identify the truly causal variant that would be more statistically associated with the disease. It is also possible to fine-map a large number of loci at once by typing dense SNP sets using custom genotyping arrays [93]. It is also commonplace for the follow-up population to be larger than the original GWAS population, which allows for increased power to include or exclude individual variants as likely causal. Statistical fine-mapping is another approach that can be used to refine the list of potentially causal variants. These methods are similar in concept to "direct" fine-mapping, however, instead of targeted variant discovery by sequencing, resources such as the complete genome sequences generated by the 1000 Genomes project are used to identify a nearly comprehensive list of variants that can then be "imputed" statistically into the study population across previously identified loci and tested for association [49]. Imputation is also available in mouse mapping studies, as many classical inbred strains have been fully sequenced [94]. No matter the method used, fine-mapping can be utilized in many cases to reduce the search space for causal SNPs and variants within a genetic locus.

Once a high-confidence set of variants for a locus is identified the next approach typically taken to identify the causal entity in a GWAS locus is to annotate the SNPs within the locus. First, this will lead to the identification of non-synonymous (NS) SNPs that lie within genes. These altered gene sequences, and their protein products, are often strong causal candidates, especially if they are predicted to alter protein function using computational methods [95,96]. Second, if coding variants are not implicated, which is usually

the case, SNPs can be identified that overlap epigenetic marks often associated with poised or active regulatory elements, such as DNase I hypersensitivity sites and histone modifications. The Encyclopedia of DNA Elements (ENCODE) [97] and Epigenomics Roadmap [98] projects have generated epigenomics data on a wide-range of tissues and cell-types, including cell-types relevant to skeletal biology such as primary human osteoblasts, chondrocytes and mesenchymal stem cells. Coupled with these data are a number of computational approaches that can be used to inform GWAS, including the integration of information about pleiotropy [99], regulation [100], and epigenetics [101]. For example, Bayesian approaches have been developed that rank SNPs based on functional annotation data [102–104]. These approaches are generally limited to human studies, where large publicly available databases of epigenetic data are available, however similar approaches may be taken in the mouse if epigenetic data is available.

Once the most likely causal variants have been identified, the next step is to link variants to their target gene(s). In the case of NS SNPs, this is immediately evident. For non-coding variants identifying their target is more difficult. This can be especially challenging in light of the observation that distal regulatory elements, such as enhancers, can be located up to 1 Mbp away from the gene promoter they act upon and it is often the case that a single enhancer works to fine-tune the expression of more than one gene [105]. The most direct route of linking non-coding variants to their target gene(s) is to use population-scale expression data to identify expression quantitative trait loci (eQTL). eQTL are variants that regulate transcription or post-transcriptional processing (stability, splicing, etc.) [106,107]. There are two types of eQTL, distal and local. Distal eQTL are variants that influence the expression of genes in trans, typically on different chromosomes [108]. In contrast, local, or cis, eQTL influence transcript levels of genes in close proximity. In the context of GWAS, we are

interested in identifying local eQTL for genes that may be causal for a particular locus. eQTL discovery consists of collecting and profiling disease-relevant tissues or cell-types in a population of densely genotyped individuals or across a population of inbred mice, using either gene expression microarrays [106] or RNA-seq [109]. Ideally, these individuals or strains would be a subset of the GWAS study population. Variants within a locus can then be tested for association with all the genes in proximity of the original locus. There are now resources, such as the data generated by the Gene Tissue Expression (GTEx) project [109], that have population-scale expression data and eQTL results for a large number of tissues that can be used to inform GWAS in the absence of disease-relevant samples from the disease GWAS. All of these methods of directly interrogating SNPs can aid in the prioritization of candidate genes and SNPs in the region, both through alteration of protein-coding sequence, and via regulatory mechanisms.

While these direct approaches to find the causal variants narrow down the list of candidates, they do not provide biological context for the potential effectors of the phenotype. Additionally, these approaches are generally based on the statistical significance of the association between the genotype of a particular SNP and the phenotype, which can be influenced by experimental design, and does not always reflect biology. In order to gain a mechanistic understanding of the drivers of these associations, genes and SNPs need to be prioritized based on biological information, rather than statistical ranking.

1.5.2 Network-based approaches

An additional framework for following up on GWAS associations is to use network-based approaches to biologically contextualize loci [110,111]. Network-based strategies have been implemented in order to predict causal genes at GWAS loci, and to implicate network modules in disease (as examples [112–117]). Especially in the case of regulatory variation, disease-

associated SNPs may act via subtle changes that are propagated through entire cellular networks. Therefore, by approaching GWAS from a global, systems perspective, we can better connect associations with their physiological impacts.

One of the most widely used types of networks for informing GWAS are co-expression networks. Co-expression networks are modular, meaning each distinct module represents a group of highly co-expressed genes [118]. These modules tend to contain genes involved in similar biological processes, e.g. the function of bone-forming osteoblasts or bone-resorbing osteoclasts [119]. In practice, co-expression network analysis takes all the genes in the genome and, in a relatively unbiased manner, organizes them into functionally coherent groups. These properties are helpful for causal gene discovery because we know that complex traits are typically influenced by functionally similar genes. Therefore, by performing an unbiased, biologically driven grouping of genes and identifying modules that are enriched for those implicated in genetic studies it is possible to prioritize which genes may be causal. Additionally, due to the functional similarity of genes within modules, the mechanism by which a novel gene affects disease can be inferred by the function of the other genes it is connected to within a module. Furthermore, networks have the property of being scale-free, meaning that they contain a small number of highly interconnected genes and an increasingly large number of less connected genes [120]. Studies have demonstrated that in some modules highly interconnected "hub" genes are more likely to be key genes affecting a disease-related trait [114,120,121]. As a result, analyzing GWAS data in the context of biological networks has the potential to identify causal genes and pathways relevant to disease.

One group, studying coronary artery disease (CAD), utilized both databases of known metabolic and signaling pathways and novel, tissue-specific gene co-expression networks generated from their own RNA-seq data [117]. First, both "knowledge-driven

pathways" and "data-driven modules" were used to generate a comprehensive list of gene sets potentially involved in CAD. Gene sets included "knowledge-based" biological pathways from the Reactome [122], Biocarta [123] and Kyoto Encyclopedia of Genes and Genomes (KEGG) [124] and "data-driven" co-expression modules from ten different human and mouse co-expression studies of various CAD-related tissues, including adipose, blood, liver, muscle, heart, and kidney. Next, using gene expression data, eQTL were identified and correlated with results from the CARDIoGRAM GWAS [125]. Only SNPs that were both associated with CAD by the CARDIoGRAM GWAS and identified as influencing gene expression by eQTL analysis were included in downstream analyses. Using prioritized SNPs in conjunction with the comprehensive list of knowledge-based pathways and data-driven modules, specific gene sets that were enriched for GWAS/eQTL SNPs were identified. In order to identify the most influential genes, the resulting gene sets were then overlaid on a causal network derived from Bayesian network models of gene-gene interactions [126,127]. Genes central to the gene-gene interaction network that were highly connected with CAD-associated genes and were identified in more than one gene set were termed "key drivers" of CAD. Finally, key drivers were perturbed using siRNA treatment, and their regulatory role was characterized. This work led to the association of novel genes with CAD, for example glyoxalase 1 (GLO1), which aids in defending against improperly glycated forms of proteins. This analysis was made possible by the large amount of available disease relevant and tissue-specific data, but could be applied to many other phenotypes with a wealth of high dimension data. However, this is not currently a feasible approach to study osteoporosis and other bone diseases, as we lack large-scale genomic data from disease-relevant tissues. As more genomic data are generated in bone, this could become a more feasible approach for the study of osteoporosis.

Despite the success of these network-based methods in some models, many of them rely on genomic data from relevant primary tissues. Efforts to further understand the results of BMD GWAS have been stymied by a lack of genomic data from bone tissue and specific bone cell types. Thus, many have employed gene expression data from mice have to successfully identify novel genes involved in complex traits [128], including the immune response [129], response to viral infection [130], and bone mineral density [131]. There is great promise in using genetic and genomic approaches in the mouse to understand human disease.

1.6 Summary

In summary, over the past decade, genetic studies in the mouse and human GWASs have provided an unprecedented understanding of how genetic variation influences osteoporosis and fracture. We now know that most of the variation in BMD at the population level is due to thousands of variants with subtle effects on BMD and most of these variants exert their impact on bone by altering gene regulation. While the initial GWAS results are promising, there is much to be done both in terms of comprehensively defining the genetic architecture of osteoporosis and fracture and converting genetic data into biological knowledge. In this work, we aim to further characterize previously identified genetic associations, from both mouse and human, using novel, unbiased approaches in the following studies:

(1) In Chapter 2, we follow up on *Feml2,* an association for femur length identified in a cross between two mouse strains: C57BL6/J, which has long femurs, and CAST/EiJ, which has short femurs. Using RNA sequencing data from C57BL6/J x CAST/EiJ F1s, we identified 6 genes exhibiting allele-

specific expression, which may be novel determinants of femur length and bone strength.

(2) In Chapter 3, we present a novel tool for visualizing the relationship between genetic associations and eQTL. Testing whether a genetic association and eQTL colocalize is an effective method for predicting the causal gene driving an association; however, until now there had not been an effective data visualization strategy for these relationships.

(3) In Chapter 4, we integrate a cell-type and biological-process specific, mouse co-expression network with the results of BMD GWAS and use the predicted properties of core genes to identify a module enriched for core genes for mineralization. We then identify four novel genes that likely underlie BMD GWAS associations.

Ultimately, this work contributes to our understanding of the genetic architecture of osteoporosis-related traits and presents new tools to follow up on GWAS studies more generally.

Chapter 2
Genetic Dissection of a QTL Affecting Bone Geometry

Olivia L. Sabik, Juan F. Medrano, and Charles R. Farber

Published in: Sabik, O. L., Medrano, J. F. & Farber, C. R. Genetic Dissection of a QTL Affecting Bone Geometry. *G3* (2017). doi:10.1534/g3.116.037424

2.1 Abstract

Parameters of bone geometry such as width, length, and cross-sectional area are major determinants of bone strength. Though these traits are highly heritable, few genes influencing bone geometry have been identified. Here, we dissect a major quantitative trait locus (QTL) influencing femur size. This QTL was originally identified in an F2 cross between the C57BL/6J-hg/hg (HG) and CAST/EiJ strains and was referred to as femur length in high growth mice 2 (*Feml2*). *Feml2* was located on Chromosome (Chr.) 9 at ~20 cM. Here, we show that the HG.CAST-(*D9Mit249-D9Mit133*)/Ucd congenic strain captures *Feml2*. In an F2 congenic cross, we fine-mapped the location of *Feml2* to an ~6 Mbp region extending from 57.3 to 63.3 Mbp on Chr. 9. We have identified candidates by mining the complete genome sequence of CAST/EiJ and through allele specific expression analysis of growth plates in C57BL/6J x CAST/EiJ F1 hybrids. Interestingly, we also find that the refined location of *Feml2* overlaps a cluster of six independent genome-wide associations for human height. This work provides the foundation to identify novel genes affecting bone geometry.

2.2 Introduction

Osteoporosis is a disease of severe bone loss that leads to skeletal fragility and an increased risk of fracture [7]. In the U.S., osteoporosis affects over 12 million people and is directly responsible for 1.5 million fractures annually [7]. Although fracture is not commonly associated with mortality, of the ~300,000 people each year that suffer a hip fracture, one in five will die in the subsequent 12 months [132].

Bone geometry is one of the many factors that contribute to bone strength [133]. Studies in mice have demonstrated that up to 50% of the variance in bone strength is due to bone size [134]. Furthermore, the relationship between bone geometry and fracture risk in

humans has been demonstrated in both the wrist and the spine, as decreased cross-sectional area of the radius and vertebrae are associated with increased risk of fracture [133]. In addition, similar to most other characteristics of bone, bone geometry is highly heritable (h^2>0.50) and amenable to genetic analysis [135]. Therefore, increasing our understanding of the genes influencing bone geometry using genetic analyses has the potential to inform strategies for the treatment and prevention of bone fragility.

To identify quantitative trait loci (QTL) affecting body composition, Corva *et al.* generated an F2 cross between the C57BL/6J-*hg*/*hg* (HG) and CAST/EiJ (CAST) strains [136]. The HG strain is a C57BL/6J (B6) mouse that is homozygous for a deletion encompassing the *Socs2* gene (the high-growth (*hg*) locus), a negative regulator of growth hormone signaling, resulting in increased growth and body size [137,138]. In contrast, CAST mice are genetically divergent wild-derived inbred mice that are small and lean. One QTL identified in the *hg* x CAST cross was femur length in high growth mice 2 (*Feml2*). *Feml2* was located on Chr. 9 at ~20 cM, explained 10.7% of the variance in femur length and was independent of the *hg* locus [136]. In order to identify the gene(s) responsible for *Feml2*, we generated the HG.CAST-(*D9Mit249-D9Mit133*)/Ucd (HG9) congenic strain that possessed CAST alleles from 9 to 84 Mbp on an HG background [139]. The HG9 strain has previously been used to fine-map a distinct QTL affecting adiposity [140].

In the current study, two HG9 F2 intercrosses were used to fine-map *Feml2*. We used the complete genome sequences of the B6 and CAST strains and allele-specific expression analysis of B6 x CAST F1 mice to identify candidate genes driving the effect on femur length. Interestingly, the human syntenic region contains a cluster of six genome-wide associations with human height. These data provide the foundation to identify genes contributing to bone size and geometry.

2.3 Results

2.3.1 Feml2 is captured in HG9 mice

We characterized femur geometry in HG9 and HG male mice. As shown in **Table 2.1** femur length was decreased by 6% ($P=1.3 \times 10^{-5}$) in HG9 males, consistent with the effects of *Feml2* in the original F2 cross [136]. In addition, medio-lateral and anterior-posterior femur widths were also decreased by 6-7% ($P=2.1 \times 10^{-3}$ and $P=3.7 \times 10^{-2}$) in HG9 males (**Table 2.1**). These data confirm the capture of *Feml2* in the HG9 congenic.

Table 2.1 *Characterization of femur geometry in male HG and HG9 mice*

	HG (N=14)	***HG9 (N=7)***	***Difference (%) (HG9-HG/HG)***	***P value***
Femur length (mm)	*17.2 ± 0.1*	*16.2 ± 0.1*	*-6*	*1.3×10^{-5}*
Medio-Lateral femur width (mm)	*2.41 ± 0.04*	*2.28 ± 0.02*	*-6*	*2.1×10^{-3}*
Anterior-Posterior femur width (mm)	*1.55 ± 0.05*	*1.45 ± 0.02*	*-7*	*3.7×10^{-2}*

2.3.2 High-resolution mapping of Feml2

To refine the location of *Feml2* within the HG9 congenic interval, we generated two F2 crosses, HG9 X HG (N=283) and HG9 X B6 (N=457). Femur length was first mapped in each cross separately. No differences in the LOD score profile of peak positions were observed (**Figure 2.1A, B**). As a result, both crosses were combined to increase mapping resolution (**Figure 2.1C**). *Feml2* mapped to 30.1 cM with a peak LOD score of 40.0 (**Figure 2.1C**). The effects of *Feml2* on femur length were additive with each CAST allele decreasing femur length by 0.23 mm (**Figure 2.1D**). *Feml2* explained 7.8% of the total variance in femur

length. The 95% confidence interval for *Feml2* extended from 28.0 to 31.7 cM, which in physical distance equated to the 5.7 Mbp region extending from 57.6 to 63.3 Mbp (GRCm38/mm10).

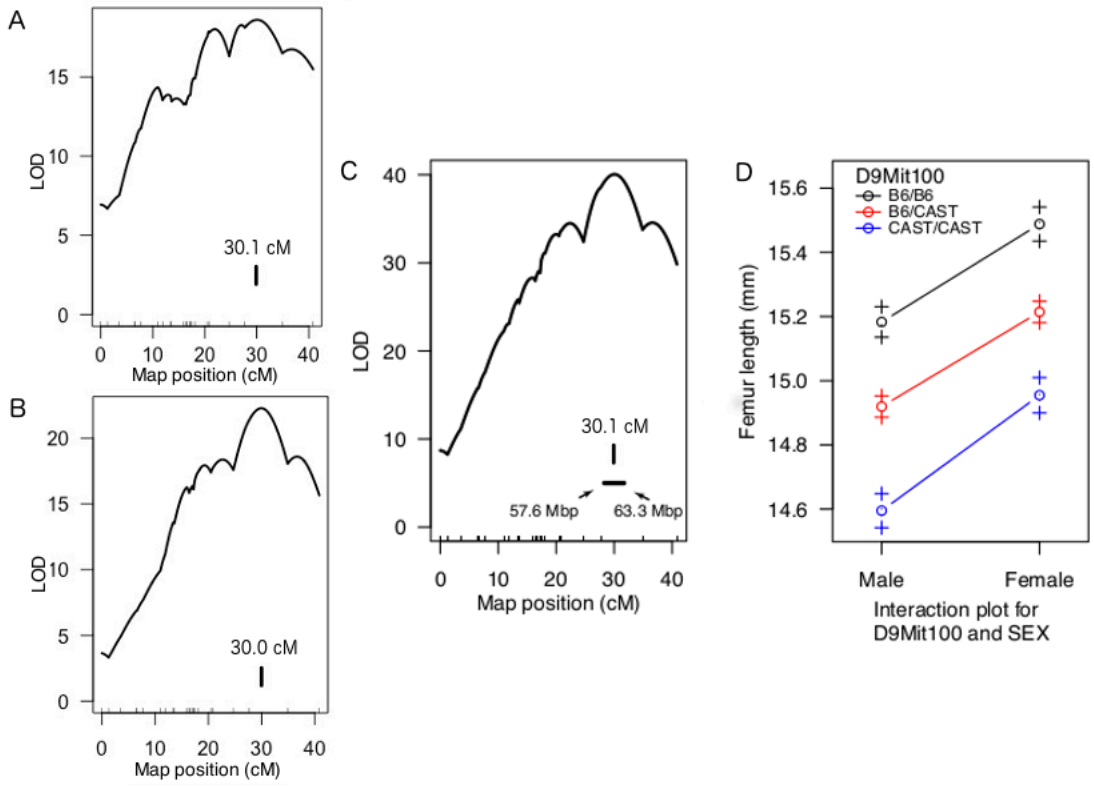


Figure 2.1 *Feml2 LOD score profiles. Vertical lines indicate peak LOD scores. The peak of Feml2 was located at 30.1 cM in the HG9 x HG (N=283) cross (A), 30.0 cM in the HG9 x B6 (N=457) cross (B), and 30.1 cM in both crosses combined (N = 740). The 95% CI for the location of Feml2 in the combined cross was 28.3-31.7 cM (57.6-63.3 Mbp) (C). Feml2 had highly significant effects (P<0.001) on femur length in both male and female mice.*

2.3.3 Characterizing Feml2 variants between CAST and B6

Feml2 contains a total of 69 RefSeq protein-coding genes (**Supplemental Table 2.1**). Based on the sequenced CAST genome [141], *Feml2* contains 46,624 high-confidence single nucleotide polymorphisms (SNPs) between CAST and B6. Most (45,664; 97.9%) of the SNPs are noncoding (**Table 2.2**). There are 960 (2.1%) coding variants of which 86 were

potentially "high-impact". Of the 86, there are 81 missense, one initiator codon, three stop-gained and one stop-retained variants (**Table 2.2**). Potentially high-impact variants were found in 37 of the 69 *Feml2* genes. The initiator codon variant is in the enhancer of mRNA decapping 3 (*Edc3)* gene; however, a second in-frame 'ATG' is located 6 bp downstream. Two of the stop-gain and the stop-retained variants were found in the "unclassified" gene *1700036A12Rik*. The other stop-gain variant was in another "unclassified" gene *Gm10657* (**Supplemental Table 2.1**)

Table 2.2 *List of SNPs located within Feml2*

Variant	**Number**
Non-Coding	
downstream gene variant	1993
upstream gene variant	2230
intergenic variant	20949
intron variant	20452
splice region variant	40
total	*45664*
Coding	
3 prime UTR variant	523
5 prime UTR variant	51
synonymous variant	246
missense variant	81
initiator codon variant	1
stop gained	3
stop retained	1
mature miRNA variant	1
non-coding exon variant	53
total	*960*

In addition to SNPs, there were 8803 small insertions/deletions (INDELs). There were 189 INDELs in untranslated exons (UTR), two frameshifts and one in-frame insertion. The rest were intergenic. The two frameshift variants were found in *Arid3b* and *Nptn*; however, both occurred in exons predicted by Ensembl, that were not part of the RefSeq transcript for either gene. The in-frame insertion was also found in the *Arid3b* gene and did occur with a RefSeq exon.

2.3.4 Characterizing Feml2 allele-specific expression using CAST X B6 F1 RNA-seq data

To identify *Feml2* genes whose expression is under genetic regulation, we quantified allele-specific expression in growth plate tissue in CAST x B6 F1 mice. Six of the 69 genes from the *Feml2* region were found to be expressed in an allele-specific manner, demonstrating higher transcript levels originating from either the B6 or CAST chromosomes in growth plate samples at an FDR<0.20 (**Table 2.3**, **Figure 2.2**). These genes are GRAM domain containing 2 (*Gramd2*), La Ribonucleoprotein Domain Family Member 6 or Acheron (*Larp6*), ADP-Dependent Glucokinase (*Adpgk*), Bone Marrow Stromal Cell-Derived Ubiquitin-Like 7(*Ubl7*), Meiotic Recombination Protein REC114-Like (*Rec114*), and Heparin/Heparan Sulfate:Glucuronic Acid C5-Epimerase (*Glce*). All of these genes, except for Adpgk, were preferentially expressed from the CAST allele as compared to the B6 allele (**Figure 2.2**).

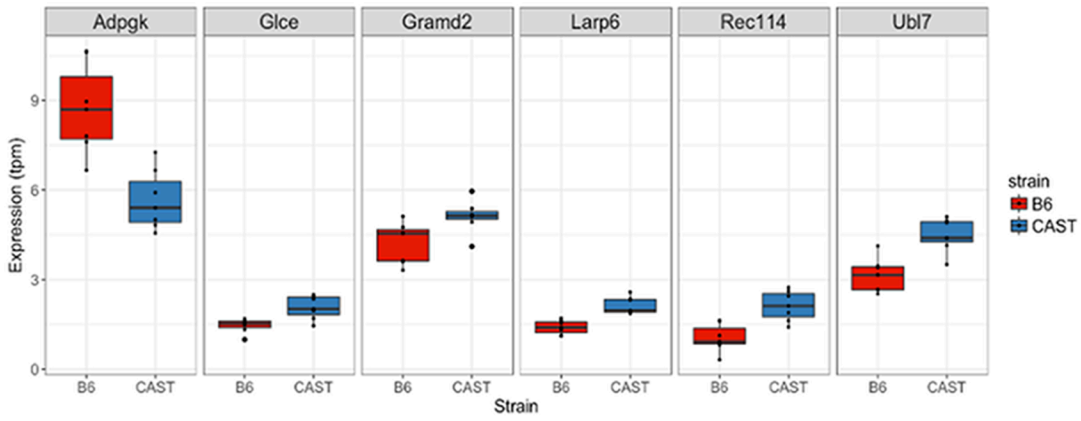


Figure 2.2 *Boxplots of the expression of six genes demonstrating significant (FDR<0.20) allele-specific expression differences. Expression is expressed in transcripts per million (tpm) and binned by strain of origin.*

Table 2.3 *Allele-specifically expressed genes from the Eml2 region*

Gene Name	Chr	CAST/EiJ mean tpm	C57BL/6J mean tpm	CAST/EiJ mean counts	C57BL/6J mean counts	logFC[1]	p-value	FDR
Gramd2	9	5.11	4.22	534	339	0.69	0.002	0.11
Larp6	9	2.13	1.40	216	142	0.64	0.005	0.11
Adpgk	9	5.66	8.72	586	893	-0.60	0.013	0.13
Ubl7	9	4.49	3.14	247	174	0.54	0.012	0.13
Rec114	9	2.12	1.04	43	19	1.09	0.008	0.13
Glce	9	2.06	1.46	341	254	0.49	0.021	0.17

[1]Positive logFC values correspond to favored CAST/EiJ expression, while negative logFC values correspond to favored C57BL/6J expression.

2.3.5 Feml2 overlaps with a cluster of human height genome-wide associations

Given the significant impact of *Feml2* on femur length, it is possible that *Feml2* harbors multiple independent variants impacting skeletal dimensions. To determine if there is evidence that *Feml2* represents a "hot-spot" of genes influencing long bone size, we analyzed the syntenic region of the human genome for human height associations (which are often driven by changes in skeletal dimensions [142]) identified by GWAS [68]. *Feml2* is syntenic with human Chromosome 15 from 67.5 to 75.5 Mbp. This region contains six independent associations as identified by GWAS for human height [68](**Figure 2.3**). Through permutation, there was suggestive evidence that the human region syntenic with *Feml2* contained more associations for height than would be expected by chance (P=0.06). Additionally, these SNPs were found to be near the human homologs of all six genes exhibiting allelic expression (**Table 2.4**).

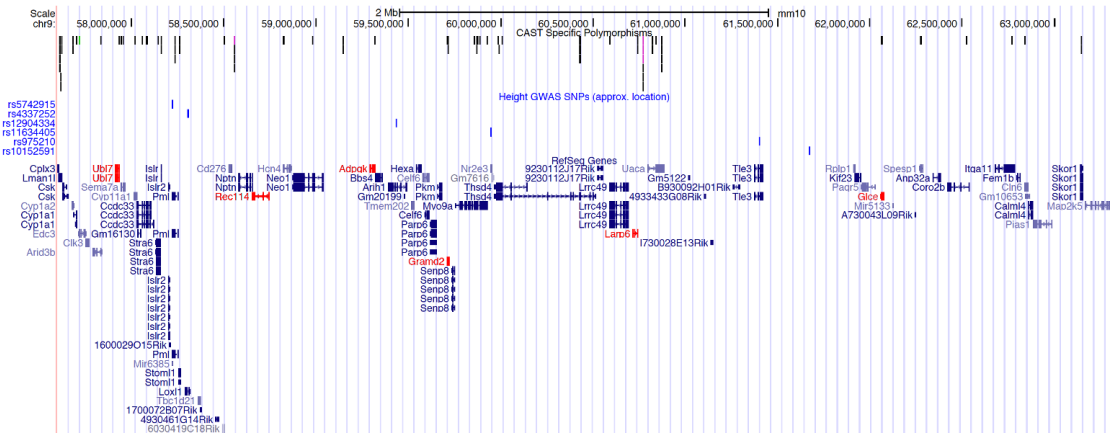


Figure 2.3 UCSC Genome browser view of mm10 chr9:57300000-63300000, which comprises the *Feml2* locus. The uppermost track displays potential high-impact predicted coding polymorphisms between CAST and B6 (black = missense, green = initiator codon variant, purple = stop gained and stop retained variants). The second uppermost track (blue) contains SNPs homologous to those identified in a GWAS for human height [68], which may play a regulatory role in gene expression that influences femur length. Finally, the bottom-most track displays the genes within the *Feml2* region. In red are genes that are exhibited significant (FDR<0.20) differences in allele-specific expression.

Table 2.4 *Human homologs of genes in Feml2 with observed allele specific expression in CAST/EiJ x C57BL/6J F1 mice and the lead height-associated SNPs nearest the human genes.*

Gene	Start (Mbp)[1]	End (Mbp)[1]	Nearest rsID#	Chr	SNP coord (Mbp)[1]	Dist to TSS (Kbp)
Gramd2	72.159807	72.197785	rs12904334	15	72.550363	390.556
Larp6	70.829130	70.854159	rs975210	15	70.072012	757.118
Rec114	73.443158	73.560014	rs4337252	15	73.934423	491.265
Ubl7	74.445977	74.461182	rs5742915	15	74.044291	401.686
Adpgk	72.751369	72.783785	rs12904334	15	72.550363	201.006
Glce	69.160634	69.272199	rs10152591	15	69.755817	595.183

[1]Human genome build hg38

2.4 Discussion

As a first step in identifying the gene(s) responsible for a QTL with a major effect on bone geometry, we developed a congenic strain containing the *Feml2* QTL and conducted an F2 intercross to refine the location of *Feml2*. Furthermore, we used the CAST and B6 genome sequences to identify genes within *Feml2* potential impact by coding variation. Consistent with the high polymorphism rate between CAST and B6 [141], 37 of the 69 *Feml2* genes contained coding variants predicted to potentially impact protein function (**Table 2.2**). Additionally, we hypothesized that non-coding variants between CAST and B6 could be driving expression differences between the two strains and that this change in expression could be responsible for the reduced femur length in CAST mice. Using RNA-seq data from tibial growth plates of CAST x B6 F1s, we identified six of the 69 genes in the *Feml2* region that are preferentially expressed from one parental allele.

Of the genes influenced by allele-specific expression, we note two that have the potential to be driving the observed femur length phenotype based on their known biological function. *Larp6*, also known as Acheron, is an RNA-binding protein, which is specific to collagen mRNAs [143]. *Larp6* regulates the translation of type I collagen subunits through sequence-specific binding to conserved stem loops in the 5' UTR of collagen mRNAs [144]. Collagens, predominantly type I collagen, comprise 90% of the bone matrix and its processing is critical for proper skeletal development [145] and it has been observed that overexpression of *Larp6* blocks ribosomal loading onto collagen mRNAs, reducing translation of collagen and bone matrix formation [146]. In CAST x B6 F1s the expression of *Larp6* was higher from the CAST allele, potentially consistent with shorter femurs in HG9 mice. *Adpgk,* also known as ADP-dependent glucokinase, catalyzes the phosphorylation of D-glucose to D-glucose 6-phosphate using ADP as the phosphate donor [147]. Knockouts of

Adpgk result in short stature and slender bones [148]. In CAST x B6 F1s the expression of *Adpgk* was lower from the CAST allele, consistent with shorter and more slender femurs in HG9 mice. The allele-specific expression of these genes suggests they may be involved in the regulation of femur length.

We identified the human syntenic region for *Feml2* and found that it contained a number of associations for human height identified through GWAS. Human height is, in part, influenced by long bone length [149] and this well-powered GWAS, conducted by Wood *et al.,* provides a robust collection of loci associated with human height to compare with the *Feml2* region's influence on mouse long bone length. This human region appears to be a hotspot for associations with height, further implicating the murine region as a region influencing femur length and, more generally, skeletal dimension (**Figure 2.3**). Further characterization of the genes in the *Feml2* region in mouse is potentially applicable to the study of human height.

In conclusion, this study identified *Feml2*, a region of the murine genome influencing femur length and identified genes with coding and regulatory alterations, a subset of which may be responsible for the effects of *Feml2*. Additionally, comparisons with its human syntenic region support the notion that *Feml2* may contain multiple polymorphic genes, which in aggregate are responsible for its effect on bone geometry. Further characterization of these candidate genes and the identification of the mechanism by which they alter femur dimension will aid in the study of bone geometry and bone strength.

2.5 Acknowledgments

The authors thank Alma Islas-Trejo (UC-Davis), Vincent de Vera (UC-Davis) and Rodrigo Gularte (UC-Davis) for technical assistance. Funding was provided by NRI-USDA-

CSREES 2005-35205-15453 and grant R01DK69978 from the National Institute of Diabetes and Digestive and Kidney Diseases to JFM. CRF was supported by grants R01AR057759, R01AR064790 and R01AR068345 from the National Institute of Arthritis and Musculoskeletal and Skin Diseases. OLS is funded by an institutional training grant T32-GM008136 at the University of Virginia from the National Institute of General Medical Sciences.

CRF developed and characterized the HG9 mouse, mapped the *Feml2* locus, and characterized the *Feml2* variants between B6 and CAST in the lab of JFM. OLS collected and processed the growth plate tissue for RNA sequencing, conducted the allele-specific expression analysis, and carried out the comparison of the *Feml2* locus and human height GWAS results in the lab of CRF.

2.6 Methods

Mouse strains and husbandry:

The mouse strains used in the study were B6, HG and the HG9 (MGI:3771219) congenic strain. The development of the HG and HG9 congenic strains have been previously described [139,150]. B6 and HG mice were obtained from vivarium stock. Mice were provided a normal chow diet (Purina 5008; 23.5% protein, 6.5% fat, 3.3 Kcal/g) and water *ad libitum* and housed in groups of 2-5 in polycarbonate cages bedded with a 2 to 1 mixture of CareFRESH (Absorption Corp, Ferndale, WA) and soft paper chips (Canbrands Int, Moncton, Canada). Mice were maintained under controlled conditions of temperature (21°C±2°C), humidity (40–70%), and lighting (14 h light, 10 h dark, lights on at 7 AM). All mouse protocols were managed according to the guidelines of the American Association for Accreditation of Laboratory Animal Care (AAALAC) and approved by the Institutional Animal Care and Use Committee at the University of California, Davis.

Characterization of bone geometry in HG9 congenic mice:

At 9 weeks of age (± 5 days) male HG9 congenic (N=7) and HG control mice (N=14) mice were anesthetized and body weights and lengths were measured to the nearest decigram and centimeter, respectively. Mice were then euthanized and femurs were removed and cleaned of soft tissue. For each femur, we measured the length and width, in the mediolateral and anterior-posterior orientations, using digital calipers (Mitutoyo, Corporation, Takatsu-ku, Japan). Experimenters were blinded to the genotype of the mice.

Development and characterization of the HG9F2 mapping populations:

The congenic F2 mouse populations used in this study have been described in [140]. Briefly, HG9 X HG (N=283) and HG9 X B6 (N=457) male and female F2 mice were generated by intercrossing F1 mice. F2 mice were phenotyped as described above for HG9 congenics. Mice were genotyped using microsatellite markers. HG9 x B6 F2 mice were genotyped for the *hg* locus as described in [139].

Feml2 fine-mapping:

All statistical analyses were performed using the R Language and Environment for Statistical Computing [151]. The R/qtl package was used to perform the linkage analysis [152]. Sex-averaged genetic maps were generated and conditional genotype probabilities were estimated, using the calc.genoprob function, along the length of the congenic donor region at 0.1 cM intervals. Both F2 crosses were combined for the linkage analysis. The scanone function, using the Haley-Knott regression algorithm, was used to perform interval mapping using a model that included sex, body weight, and cross type (HG9 X HG=1 or HG9 X B6=2) terms as additive covariates. LOD significance for all models tested were empirically determined using 1000 permutations. We converted genetic to physical distance by regressing Mbp onto cM for all markers.

Growth Plate RNA collection:

Seven male F1s from a cross between B6 and CAST were euthanized at 21 days of age by isoflurane anesthesia followed by cervical dislocation. Proximal and distal tibial growth plates were rapidly dissected, placed in TRIzol (Ambion by Life Technologies) and pulverized using the Tissue Tearor homogenizer (BioSpec Products). Total RNA was extracted from homogenized tissue (mirVana miRNA Isolation Kit, Ambion by Life Technologies). RNA concentration was measured by fluorometry (Qubit 2.0 Fluorometer, Life Technologies).

RNA-sequencing sample preparation:

RNA-Seq libraries were constructed from 200 ng of total RNA using Illumina TruSeq Stranded Total RNA with Ribo-Zero Gold sample prep kits (Illumina, Carlsbad, CA). Constructed libraries contained RNAs >200 nt (both unpolyadenylated and polyadenylated) and were depleted of cytoplasmic and mitochondrial rRNAs. An average of 6.7 million 2 x 75 bp paired-end reads were generated for each sample on an Illumina NextSeq 500 (Illumina, Carlsbad, CA).

RNA-seq alignment strategy and allele specific expression (ASE) analysis:

Using g2gtools [153] a transcriptome containing both B6 (mm10) and CAST (mm10 with version 4 SNPs and indels from the Mouse Genome Project, http://www.sanger.ac.uk/science/data/mouse-genomes-project) alleles was generated. Reads from each F1 hybrid were aligned to the joint transcriptome using Bowtie, allowing no more than 3 mismatches and reporting all alignments of the stratum containing the fewest number of mismatches [154]. Next, EMASE [155] was used to quantify the number of reads from both the maternal and paternal allele. Each end of the paired-end samples was processed separately and only reads aligning in both samples were included. Similarly, each

lane was processed separately in order to correct for lane-specific effects. Post-quantification, all samples passed quality control checks based on the expected global proportions of reads aligning to each parental strain. All samples showed nearly equal reads mapping to each parental strain. In order to identify genes showing allelic expression, edgeR was used to compare the quantity of each transcript by strain across the seven samples. Only those transcripts with measured expression in at least one haplotype in all of the samples were included in ASE analysis. The glmFit and glmLRT functions were used to statistically compare the expression of each transcript between the CAST and B6 alleles.

Additional statistical analyses:

Bone geometry measures in HG9 congenic and HG control mice were compared using a student's T-test in R [151]. Comparisons at a $P<0.05$ were deemed significant. Permutation analysis was used to determine the probability of six GWAS associations for height occurring within the 7.5 Mbp human region syntenic with *Feml2* by randomly selecting 1000 7.5 Mbp regions from the genome and counting the number of associations within each region. Genome-wide significant SNPs identified through a GWAS for human height were mapped to mm10 from hg18 using the liftover tool from UCSC.

Data Availability

Both C57BL/6J and CAST/EiJ strains are commercially available from Jackson Labs. **Supplementary file 2.1** contains QTL mapping information, including mouse IDs, phenotype information, and SNP marker identifiers, locations, and genotypes. Phenotype information can be found in the accompanying README file. Gene expression data from chondrocytes are available at GEO with the accession number: GSE90055.

Chapter 3

RACER: A data visualization strategy for exploring multiple genetic associations

Olivia L. Sabik and Charles R. Farber

Preprint: Sabik, O. L. & Farber, C. R. RACER: A data visualization strategy for exploring multiple genetic associations. *bioRxiv* 495366 (2018). doi:10.1101/495366

3.1 Abstract

Genome-wide association studies (GWASs) have identified thousands of loci associated with risk of various diseases; however, the genes responsible for the majority of loci have not been identified. One means of uncovering potential causal genes is the identification of expression quantitative trait loci (eQTL) that colocalize with disease loci. Statistical methods have been developed to assess the likelihood that two associations (e.g. disease locus and eQTL) share a common causal variant, however, visualization of the two loci is often a crucial step in determining if a locus is pleiotropic. While the current convention is to plot two associations side-by-side, it is difficult to compare across two x-axes, even if they are identical. Thus, we have developed the Regional Association ComparER (RACER) package, which creates "mirror plots", in which the two associations are plotted on a shared x-axis. Mirror plots provide an effective tool for the visual exploration and presentation of the relationship between two genetic associations.

Availability and Implementation RACER is provided under the GNU General Public License version 3 (GPL-3.0). Source code is available at https://github.com/oliviasabik/RACER.

3.2 Introduction

Genome-wide association studies (GWASs) have identified thousands of loci associated with disease risk; however, the genes responsible for the majority of these disease-associated loci remain largely unknown [38]. A common approach to identify causal genes is to determine if disease-associated variants also influence molecular phenotypes, such as gene expression [156]. This approach has become more widely implemented as expression

quantitative trait loci (eQTL) across many tissues have become available from projects such as the Genotype-Tissue Expression Project (GTEx) [157]. Several statistical approaches that provide formal evidence of colocalization between two associations (e.g. a disease locus and eQTL) have been developed [158–161]; however, effective visualization is often an important component of colocalization analyses to ensure the presence of a single pleiotropic association. A common convention is to plot two associations separately using LocusZoom or LocusCompare and present them side by side, though it is often difficult to compare associations plotted on two different x-axes [162]. To address this issue, we built the Regional Association ComparER (RACER) package, which creates "mirror plots" for two individual associations. Mirror plots illustrate two associations, one inverted, on a shared x-axis, allowing for the direct comparison of the associated variants for two phenotypes.

3.3 Results

3.3.1 RACER Features

RACER was developed as a data visualization tool for the comparison of two sets of association data that share a common locus. With RACER, users can plot association data, minimally containing columns for chromosome, genomic coordinates, and p-values for an association. RACER contains a formatting function which can take any association data as input and format it for compatibility with plotting functions. RACER also contains a function for annotating association data with population-specific linkage disequilibrium (LD) data from the 1000 genomes project using LD Link using reference SNP IDs or formatting existing linkage disequilibrium provided by the user for a specific study population [43,163]. Once the association data has been formatted and annotated, RACER can produce three different types of plots. (1) a plot of a single association (**Supplemental Figure 3.1**), (2) a

scatter plot of the p-values from two different association data sets (**Supplemental Figure 3.2**), or (3) a mirror plot for two associations (**Figure 3.1**).

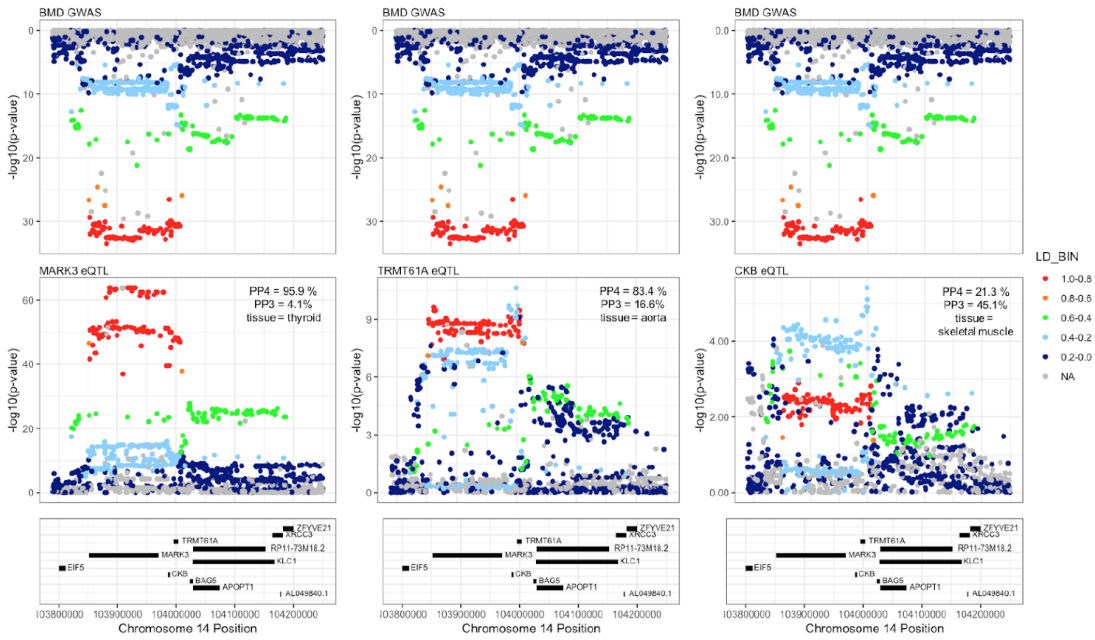


Figure 3.1. Mirror plots for *MARK3*, *TRMT61A* and *CKB* eQTL and a BMD GWAS locus. The mirror plots illustrate the similarity of the BMD association and *MARK3* eQTL, the complexity of the *TRMT61A* eQTL, and the dominance of the secondary association in the *CKB* eQTL.

A vignette illustrating how to create the *MARK3* eQTL/BMD association mirror plot described below can be found at https://oliviasabik.github.io/RACERweb/articles/IntroToRACER.html.

3.3.2 RACER Application

As a demonstration of the utility of RACER, we present a case using GTEx eQTL data to interrogate a locus on Chr. 14q32.32 associated with bone mineral density (BMD). The Chr. 14q32.32 locus spanned approximately 160 Kbp and included three genes: *MARK3, CKB* and *TRMT61A*. We previously demonstrated that the expression of all three

genes were influenced by significant eQTL ($p < 1.0\times10^{-5}$) in at least one GTEx tissue [131]. In the original paper, we analyzed these relationships using GTEx release v6 and BMD GWAS data from a 2012 study [35]. To demonstrate the use of RACER, we performed a new analysis using GTEx release v7 and BMD GWAS data from a 2017 study (**Supplemental File 3.1, 3.2**) [36,164]. First, we used the coloc R package to estimate the posterior probability (PPH4) that each pair of associations were due to single causal variants [158]. Using coloc, we observed that both *MARK3* and *TRMT61A* were likely to share a causal variant (PPH4 = 95.9% and PPH4 = 83.4%, respectively). The likelihood that the *CKB* eQTL colocalizes with the Chr. 14q32.32 locus was low (PPH4 = 21.3%)

We used RACER to create mirror plots comparing the BMD association with each of the three eQTL. This visualization of the *MARK3* and *TRMT61A* eQTL in direct comparison with the BMD association indicate that the *MARK3* eQTL has an architecture more similar to the BMD association than the *TRMT61A* eQTL. The *MARK3* eQTL is nearly identical to the BMD association; the same variants are the most significantly associated with both *MARK3* expression and BMD and the pattern of association is similar across SNPs of decreasing LD. While the *TRMT61A* eQTL and BMD association have a PPH4>75%, which is considered sufficient evidence of a shared causal variant, it appears to be influenced by multiple associations in this region [158]. The variants that are the most significantly associated with *TRMT61A* expression only exhibit low linkage disequilibrium with the SNPs that are the most significantly associated with BMD. However, the most significant BMD variants do see to be represented in the association, albeit at a lower level of significance. As observed in the coloc results, the *CKB* eQTL signal is dominated by an alternative signal, similar in architecture to the strongest signal in the *TRMT61A* eQTL. Using RACER, we confirmed the coloc results for *CKB* and gained a more nuanced view of

the *TRMT61A* and *MARK3* results. Though this analysis does not exclude the involvement of *TRMT61A* or *CKB*, it does provide further evidence that *MARK3* is responsible for the association.

While we demonstrated the comparison between a disease association and an eQTL, RACER can be used to visualize the comparison between any two associations at a common locus, including associations for different phenotypes which may arise from a pleiotropic variant, or comparable associations arising from studies carried out in populations of different ethnicities.

3.4 Conclusions

We have developed RACER, an R package to produce mirror plots, which allow for the direct comparison of two different associations within the same locus. Mirror plots provide an effective tool for the visual exploration and presentation of the relationship between two genomic associations.

3.5 Acknowledgments

We would like to thank Nathan Sheffield (University of Virginia) and John Lawson (University of Virginia) for their advice in the development of RACER, and Basel Al-Barghouthi (University of Virginia), Eric Taleghani (University of Virginia), and Catherine Robertson (University of Virginia), all of whom provided critical testing and input throughout the development of RACER. OLS was supported by a Wagner Fellowship from the University of Virginia and the University of Virginia Cell and Molecular Biology Training Grant funded by the National Institute of General Medical Sciences (T32 GM8136-31A1). This work was also supported by the National Institute of Arthritis and Musculoskeletal and

Skin Diseases of the National Institutes of Health [AR071657, AR064790, and AR068345 to CRF] The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Conflicts of Interest: none declared.

OLS developed the package and the publication was written by OLS with guidance from CRF.

Chapter 4

Identification of a core module for bone mineral density through the integration of a co-expression network and GWAS

Olivia L. Sabik, Gina M. Calabrese, Eric R. Taleghani, Cheryl L. Ackert-Bicknell, and Charles R. Farber

In preparation

4.1 Abstract

Genome-wide association studies (GWASs) for bone mineral density (BMD), one of the most significant predictors of osteoporotic fracture, have identified over 1100 independent associations; however, few of the causal genes have been identified. Recently, the "omnigenic" model of the genetic architecture of complex traits proposed two general categories of causal genes, core and peripheral. Core genes play a direct role in regulating traits; thus, their identification is key to revealing critical regulators and potential therapeutic targets. Here, we identified a co-expression module enriched for genes exhibiting properties consistent with core genes for BMD by analyzing GWAS data through the lens of a cell-type and timepoint-specific gene co-expression network for mineralizing osteoblasts. We identified multiple co-expression modules enriched for genes implicated by BMD GWAS and prioritized modules based on their enrichment for genes with core-like properties. Only one module, the purple module, was enriched for genes correlated with *in vitro* mineralization ($r = 0.49$; FDR = 0.012), with known roles in skeletal development ($P < 2.2 \times 10^{-16}$), that when perturbed produce a bone phenotype in mice (Odds Ratio (OR) = 4.1; $P = 2.14 \times 10^{-9}$), and are monogenic bone disease genes in humans (OR = 21.3; $P = 6.94 \times 10^{-9}$). Furthermore, the purple module contained genes from two distinct transcriptional profiles with regards to osteoblast differentiation, one of which, termed the late differentiation cluster (LDC), was more highly enriched for genes with core-like properties. Within the LDC, we found that the most highly connected genes were more likely to overlap a BMD GWAS association and associations that contained LDC genes overlapped enhancers and promoters in osteoblasts. Finally, we identified four LDC genes (*B4GALNT3*, *CADM1*, *DOCK9*, and *GPR133*) with colocalizing expression quantitative trait loci (eQTL) and altered BMD in mouse knockouts. Our network-based approach identified a "core" module for

BMD and has provided a resource for expanding our understanding of the genetics of bone mass.

4.2 Introduction

Osteoporosis is a disease characterized by low bone mineral density (BMD) and an increased risk of fracture [165]. Worldwide, osteoporosis is estimated to affect over 200 million individuals, directly resulting in 8.9 million fractures [166]. Osteoporosis is a multifactorial disease, influenced by both environmental and genetic variation. While environmental impacts on fracture are significant, fracture-related traits, such as BMD, are among the most heritable disease-associated quantitative traits (h^2>0.50) [31–33]. Due to its high heritability, clinical importance, and ease of measurement in large cohorts, nearly all genome-wide association studies (GWAS) for osteoporosis have focused on BMD [167]. These studies have been tremendously successful, identifying over 1100 independent BMD associations [35,36,168]. However, despite the wealth of genetic signals, the genes and mechanisms through which these associations impact bone remain largely unknown. As a result, there is a critical need for new approaches to identify causal genes [167].

Recently, the "omnigenic model" was proposed as a framework for understanding the genetic architecture of complex traits such as BMD [39,40]. The model posits that all genes expressed in disease-relevant cell types have the potential to contribute to disease variation. One of the key concepts of the omnigenic model is the classification of causal genes as either "core" or "peripheral". Core genes directly modulate traits, independent of all other genes. In contrast, peripheral genes impact traits through their effects on core genes [40]. Given the direct role that core genes play in the regulation of disease-related traits, there is great interest in identifying them, however GWASs alone are incapable of distinguishing between core and peripheral genes.

The precise definition of a core gene is open to debate [40–42,169]; however, as direct mediators of disease-related phenotypes, core genes are expected to exhibit certain characteristics. For example, core genes are expected to participate in trait-related biological processes and have expression levels that correlate with disease. Additionally, severe perturbation of a core gene is anticipated to have a large impact on a disease (e.g. monogenic disease genes). In the omnigenic model, it is hypothesized that peripheral genes account for a substantial component of the heritability of a trait because their effects are amplified by interactions with networks of co-regulated core genes, which, if co-regulated transcriptionally, will be co-expressed [40]. Given the expectation that core genes will be co-expressed, integrating the results of GWAS with co-expression networks that reflect the transcriptional programs associated with the trait of interest is a logical approach to identify modules of core genes. In fact, a number of studies have successfully used co-expression networks to inform GWAS [117,170–173]. For example, we previously used a bone co-expression network to identify the osteoblast functional module (OFM), a group of co-expressed genes related to osteoblast activity, and used it to predict causal genes underlying BMD GWAS loci and infer their function [131].

Here, we extend and refine our previous approach with the goal of identifying core genes for BMD. We used weighted gene co-expression network analysis (WGCNA) to generate a co-expression network for mature, mineralizing osteoblasts and identified modules enriched for genes implicated by BMD GWASs. We then used the following biologically motivated filters to identify "core" modules: (1) correlation with *in vitro* mineralization (a process of fundamental importance to BMD), (2) enrichment for genes when knocked-out in mice alter BMD, and (3) enrichment for genes involved in monogenic skeletal disease. Of the 65 network modules, the purple module was highlighted by all filters

and contained many genes with well-known roles in osteoblast activity and bone formation. Furthermore, using gene expression data collected in purified osteoblasts throughout differentiation, we were able to identify two clusters of genes within the purple module that follow distinct patterns of expression, an early and a late differentiation cluster (EDC and LDC). We found that the LDC was more enriched for all genes with known core-like properties. Within the LDC, we observed that the most highly connected genes were more likely to overlap a GWAS association and were more strongly correlated with *in vitro* mineralization. We identified four highly connected genes from the LDC that had colocalizing human eQTL and altered BMD in mouse knockout studies: *B4GALNT3*, *CADM1*, *DOCK9*, and *GPR133*. We anticipate that this integrative approach, utilizing cell-type and biological process-specific transcriptomic profiles, filters reflecting the properties of core genes, and the results of GWAS, will aid in the search for critical core genes and pathways underlying complex phenotypes and disease.

4.3 Results

4.3.1 Construction of a co-expression network reflecting transcriptional programs in mineralizing osteoblasts

The goal of this work was to use a cell- and stage-specific co-expression network to identify osteoblast "core" genes underlying BMD GWAS associations. To identify core genes related to BMD, we chose to focus on a single cell type at a single-time point during differentiation: mature, mineralizing osteoblasts. We began by using WGCNA to construct a co-expression network using transcriptomic profiles generated from mineralizing primary calvarial osteoblasts from 42 strains of Collaborative Cross (CC) mice [174]. The resulting network consisted of 65 modules of genes, with an average of 292 genes per module (**Figure**

4.1 and Supplemental File 4.1). Each co-expression module was distinguished by its assigned color, e.g. the purple module.

To confirm that modules of genes produced by the co-expression analysis represented transcriptional programs reflecting specific biological processes, we assessed whether modules were enriched for genes associated with specific gene ontology (GO) terms [175]. Most network modules were enriched for general biological processes, such as the immune response ($P_{adj} = 6.6 \times 10^{-36}$) in the blue module, mRNA metabolism ($P_{adj} = 7.8 \times 10^{-9}$) in the darkolivegreen module, and chromatin remodeling ($P_{adj} = 1.9 \times 10^{-4}$) in the grey60 module (**Figure 4.1 and Supplemental File 4.2**). However, as would be expected, there were a subset of modules enriched for genes involved in the activity of osteoblasts. For example, the cyan module was enriched for members of the Wnt signaling pathway (a key regulator of osteoblast activity) ($P_{adj} = 2.3 \times 10^{-4}$), the turquoise module was enriched for genes encoding extracellular matrix proteins ($P_{adj} = 3.5 \times 10^{-25}$) (such as genes encoding for collagens ($P_{adj} = 0.4.2 \times 10^{-10}$)), and the purple module was enriched for genes involved in skeletal system development ($P_{adj} = 2.3 \times 10^{-10}$) and osteoblast differentiation ($P_{adj} = 2.0 \times 10^{-6}$) (**Figure 4.1 and Supplemental File 4.2**). Given that our network modules represented distinct biological processes, including those involved in mineralization and osteoblast activity, we were confident it would provide a platform for identifying core genes related to mineralization that potentially underlie BMD GWAS associations.

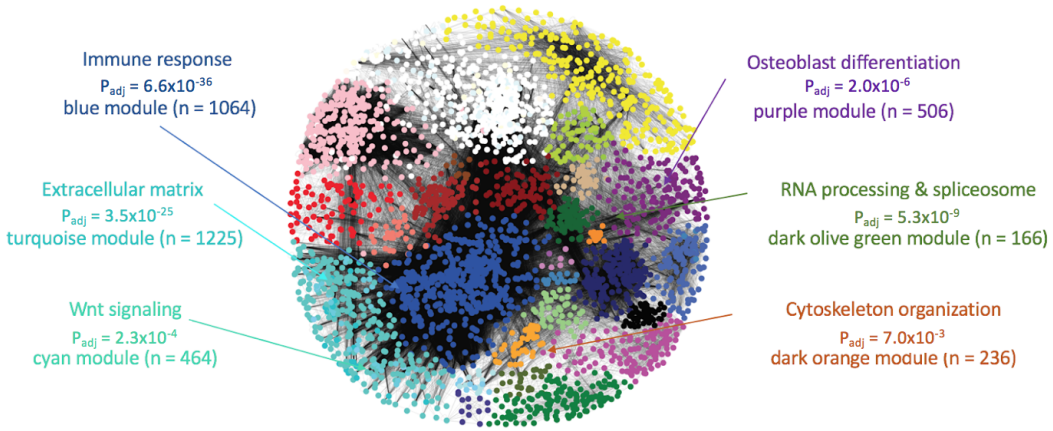


Figure 4.1 *Weighted gene co-expression network generated using transcriptomic profiles from mineralizing osteoblasts.* The network was composed of 65 modules of co-expressed genes, many of which were enriched for specific biological processes relevant to osteoblasts.

4.3.2 Identification of co-expression modules enriched for genes implicated by GWAS

To identify modules of co-expressed genes informative for GWAS, we first determined if any of the 65 modules were enriched for genes that overlapped GWAS associations. Using data from the two largest GWASs performed at the time, one study of Dual Energy X-Absorptiometry (DEXA)-derived areal BMD measures at the lumbar spine and femoral neck [35] ("Estrada *et al.* GWAS"; N=32,961) and one study of ultrasound determined heel estimated BMD (eBMD) [36] ("Kemp *et al.* GWAS", N=142,487), we developed a list of 789 human genes ($N_{Estrada}$ = 179, N_{Kemp} = 701, (91 shared genes)) intersecting BMD GWAS loci. A total of 723 (92%) of these had mouse homologs in the network (**Supplemental File 4.3&4.4**). Of the 65 modules in the network, 13 were enriched for mouse homologs of human genes implicated by GWAS (Fisher's exact test, $P_{adj} < 0.05$) (**Supplemental File 4.5 and Figure 4.2A**). Additionally, we performed stratified linkage disequilibrium (LD) score regression by calculating the BMD heritability partitioned by

SNPs surrounding genes in each module using the Kemp *et al.* GWAS [36,176]. We found 16 modules enriched for partitioned BMD heritability, including nine of the 13 enriched for BMD GWAS implicated genes (**Figure 4.2B and Supplemental File 4.6**).

4.3.3 The purple module is enriched for core genes

Next, we focused on identifying which of the 13 modules identified above contained genes with core-like properties. Instead of using the strict statistical definition proposed by [39,40], we selected genes using biologically motivated criteria. First, we compared the 13 module eigengenes with *in vitro* mineralization across the same 42 CC strains used in the construction of the co-expression network (**Supplemental Figure 4.1)**. Only one, the purple module, had a pattern of expression that was significantly correlated with mineralization ($r = 0.49$, $P_{adj} = 0.012$), suggesting the purple module was enriched for genes with a direct role in mineralization (**Figure 4.2C and Supplemental Figure 4.2**).

Core genes are defined by their direct influence on disease-relevant biological processes [39,40]. Thus, perturbation of core genes are more likely to result in a significant impact on a phenotype, as in the case of a mouse knockout or human monogenic disease. We identified all gene knockouts that produced a bone phenotype, defined as either a change in BMD, bone mineral content (BMC), abnormal bone morphology, or abnormal bone cell activity by utilizing mouse knockout phenotype data from several databases [92,177–179] (**Supplemental File 4.7**). Of the 13 modules enriched for BMD GWAS genes, two were enriched for genes whose deficiency impacted bone in mice (**Figure 4.2D**). The purple module was the most significantly enriched (Odds Ratio (OR) =5.4, $P_{adj} = 1.61 \times 10^{-34}$). We also compiled a list of 35 known drivers of monogenic bone diseases associated with osteoblast dysfunction, including osteogenesis imperfecta, hyperostosis, and osteosclerosis

(**Supplemental File 4.8**) [180–184]. Again, the purple module, containing 11 of 35 (31.4%) monogenic disease genes, was the most significantly enriched (OR = 21.3, P_{adj} = 6.94 x 10^{-9}) (**Figure 4.2E**). Together, these independent lines of evidence suggested the purple module was enriched for BMD core genes.

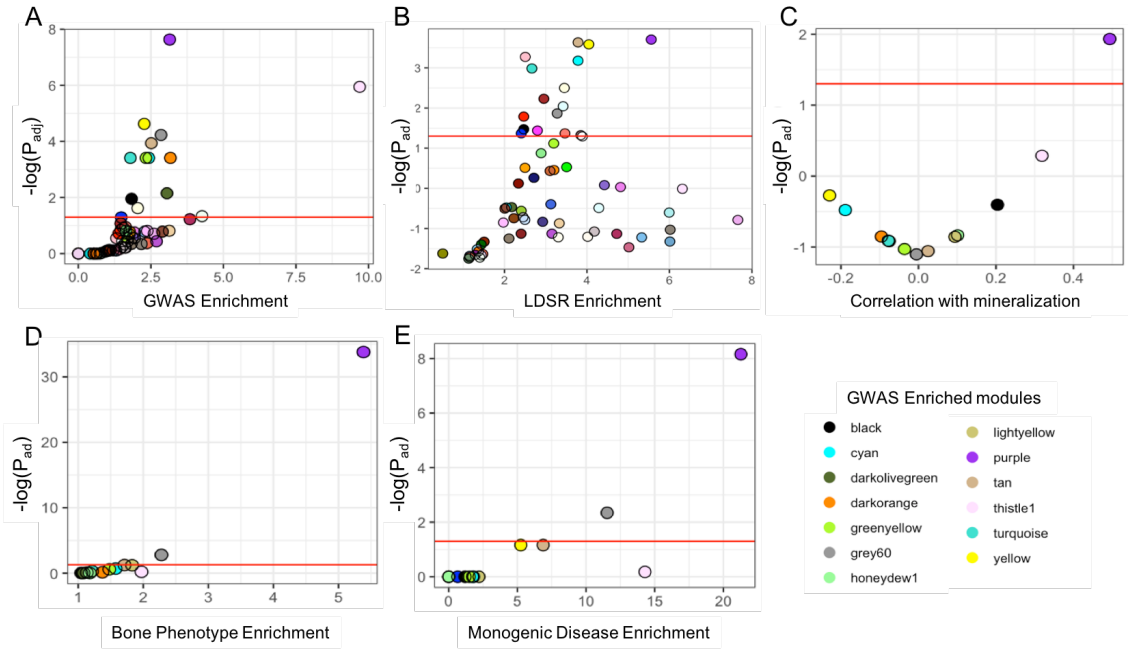


Figure 4.2 *The purple module is enriched for genes with core-like properties.* (A) Module enrichments for genes overlapping a BMD GWAS association. (B) Enrichments for partitioned BMD heritability for each module determined using stratified LD score regression. (C) Correlation between each module eigengene and *in vitro* mineralization. (D) Module enrichments for genes that, when knocked out, produced a bone phenotype and (E) human monogenic bone disease genes. Red line in each panel represents $P_{adj} < 0.05$.

4.3.4 New BMD GWAS associations further support the purple module as a core gene module

While we were analyzing the Kemp *et al.* GWAS data, a second eBMD GWAS was conducted (Morris *et al.* GWAS) [37]. The Estrada *et al.* (N=32,961) and Kemp *et al.* (N=142,487) GWASs identified 56 and 307 conditionally independent associations,

respectively [35,36]. In comparison, the Morris *et al.* GWAS (N=426,824) identified 1103 eBMD associations; an increase of over 3.5-fold [37]. The associations identified by the Morris *et al.* GWAS overlapped 1581 genes, as compared to 789 by the Estrada *et al.* and Kemp *et al.* GWASs (**Supplemental File 4.9**). Assuming the genetic architecture of BMD is consistent with the omnigenic model, we expected the inclusion of the Morris *et al.* GWAS data would increase the number of modules enriched for GWAS implicated genes. Consistent with this hypothesis, the number of modules enriched for GWAS-implicated genes doubled (N_{Kemp} = 13, N_{Morris} = 26) using the Morris *et al.* GWAS (**Figure 4.3A**, **Supplemental File 4.10**). As observed in the first analysis, most (18/26, 69%) of the new modules enriched for GWAS-implicated genes were also enriched for partitioned BMD heritability (**Supplemental File 4.11 & Figure 4.3C**). These new modules were enriched for genes involved in general biological processes such as RNA splicing (brown module, $P_{adj} = 4.04x10^{-11}$), cell junctions (floralwhite module, $P_{adj} = 6.16x10^{-3}$), cell motor activity (orange, $P_{adj} = 6.61x10^{-3}$), the cell cycle (lightgreen, $P_{adj} = 3.17x10^{-4}$), ER to Golgi trafficking (salmon, $P_{adj} = 1.80x10^{-2}$), the glycolytic process (red, $P_{adj} = 1.08x10^{-13}$), and not processes specific to osteoblast activity and/or mineralization **(Supplemental File 4.2)**.

Similar to our first analysis, the purple module was among the most enriched for GWAS implicated genes (OR = 2.67, $P_{adj} = 3.4 \times 10^{-11}$) (**Figures 4.3A**) and BMD heritability captured (OR = 5.8, $P_{adj} = 4.7 \times 10^{-6}$) (**Figures 4.3B**). Using the Estrada *et al.* and Kemp *et al.* GWAS, the purple module contained 45 genes implicated by GWAS (OR = 3.15, $P_{adj} = 2.3 \times 10^{-8}$) (5.7% of GWAS genes; 8.9% of purple module genes) and explained 27% of the SNP-heritability (h_g^2) in the study, or 4.6% of the total heritability. Using the Morris *et al.* GWAS, the number of purple module genes implicated by GWAS increased to 77 (OR = 2.7, $P_{adj} = 3.4 \times 10^{11}$) (4.9% of GWAS genes; 15.2% of purple module genes) explaining 25.3% of the

h_g^2, or 5.4% of the total heritability. Additionally, the purple module was still the only one correlated with *in vitro* mineralization (**Figure 4.3D**), the most significantly enriched for genes eliciting a bone phenotype when knocked-out in mice **(Figure 4.3E**), and human monogenic bone disease genes (**Figure 4.3F**). These data indicate that even with a significant increase in the number of GWAS-implicated genes included in the analysis, the purple module is still the only one enriched for genes with core properties.

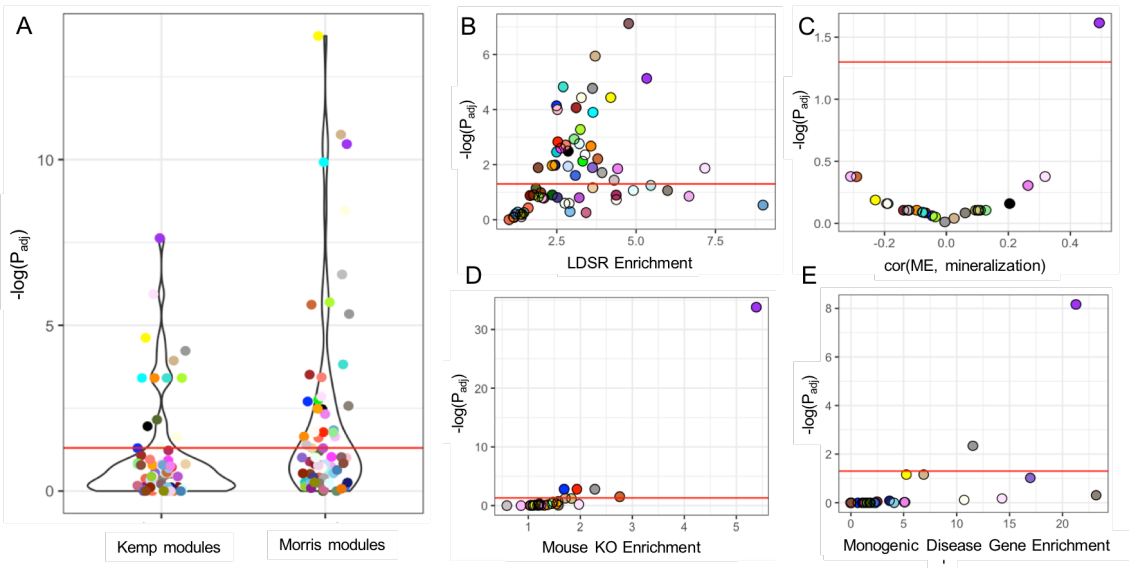


Figure 4.3 *The purple module was the only core module even after increasing the number of analyzed GWAS associations by 3.5-fold.* (A) Modules identified as enriched for GWAS implicated genes in the Kemp *et al.* GWAS versus the Morris *et al.* GWAS using all genes within the boundaries of the association. (B) Module enrichments for partitioned BMD heritability for each module determined using stratified LD score regression. (C) Correlation between each module eigengene and *in vitro* mineralization. (D) Module enrichments for genes that, when knocked out, produced a bone phenotype and (E) human monogenic bone disease genes. Red line in each panel represents $P_{adj} < 0.05$.

4.3.5 The purple module contains genes belonging to one of two distinct transcriptional programs across osteoblast differentiation

As described above, the purple module was enriched for GO categories important for the function of osteoblasts. Consistent with this, it contained many genes known to play a role in osteoblast differentiation and mineralization, including *Runx2* [185], *Sp7* [186], *Sost* [187,188], *Bglap* [189], and *Alpl* [190] (**Supplemental File 4.12**). Thus, to further investigate the purple module, we evaluated the expression of its genes with regards to osteoblast differentiation. To do this, we utilized transcriptomic profiles collected from purified osteoblasts at multiple time points across differentiation (GSE54461). Using k-means clustering, we found that the genes within the purple module clearly partitioned into two distinct transcriptional profiles with regards to differentiation (**Figure 4.4A, B**). We have termed these groups the Early Differentiation Cluster (EDC; high expression early and low expression late) (N=192 transcripts; 175 unique genes) and the Late Differentiation Cluster (LDC; low expression early and high expression late) (N=423 transcripts; 323 unique genes).

We assessed whether there were differences between the EDC and the LDC with regard to network parameters and their enrichment for functional annotations seen in the purple module. We first looked at intramodular connectivity, measured by module membership (correlation between the expression of each gene and the module eigengene). On average, LDC genes had higher module membership scores than EDC genes ($P = 3.0 \times 10^{-4}$) (**Figure 4.4C**). Additionally, the LDC was more significantly enriched than the EDC for genes implicated by GWAS (OR = 3.0, $P_{adj} = 5.2 \times 10^{-10}$), osteoblast relevant GO terms (e.g. "skeletal development" ($P_{adj} = 9.6 \times 10^{-11}$) and "osteoblast differentiation" ($P_{adj} = 1.4 \times 10^{-4}$)), genes that when knocked-out result in a bone phenotype (OR = 7.3, $P_{adj} = 1.1 \times 10^{-33}$) and monogenic bone disease genes (OR = 33.2, $P_{adj} = 8.4 \times 10^{-11}$) (**Figure 4.4D**). The

fact that LDC genes are expressed at high levels during late differentiation, coincident with when the osteoblasts are actively mineralizing, suggests that LDC contains core genes specific for the process of mineralization. For all downstream analyses we focused on the LDC.

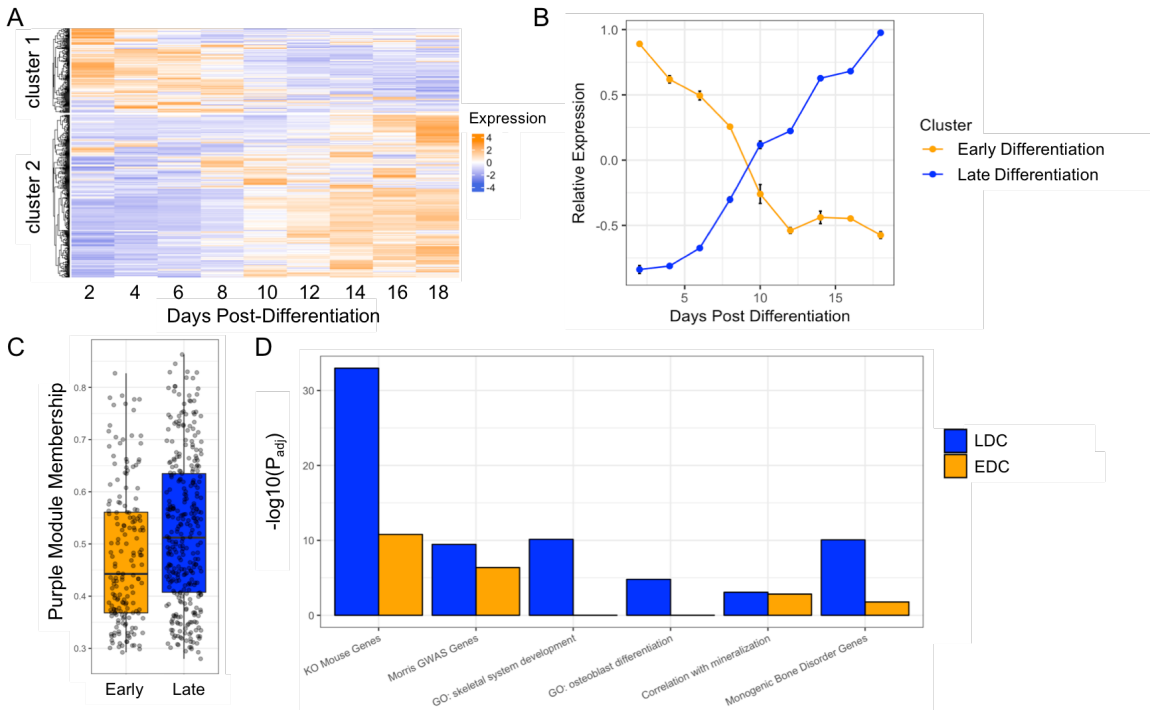


Figure 4.4 *The purple module consists of genes representing two distinct transcriptional profiles across osteoblast differentiation, one of which, the late differentiation cluster (LDC), is more enriched for genes with properties consistent with core genes for mineralization.* (A) Purple module genes show two distinct patterns of expression across differentiation, (B) Genes in cluster 1 (or the early differentiation cluster; EDC; N=192 genes) are expressed high early in osteoblast differentiation. Genes in cluster 2 (or the late differentiation cluster; LDC; N=423 genes) are expressed high late in osteoblast differentiation. (C) LDC genes have a significantly higher purple module membership score ($P= 3.0x10^{-4}$). (D) The LDC is more significantly enriched than the EDC for genes implicated by BMD GWAS in humans, associated with GO terms for bone development, for genes that when knocked out, produce a bone phenotype, and for genes involved in monogenic bone disorders.

4.3.6 BMD-associated variants in GWAS loci harboring LDC genes overlap active regulatory elements in osteoblasts

Based on the fact that the LDC is enriched for genes involved in osteoblast differentiation and that mineralization is fundamental in the regulation of BMD, we anticipate that many of the genes in the LDC are true core genes and causal genetic drivers of BMD. If true, then BMD-associated variants in associations harboring LDC genes should regulate the expression of LDC genes in osteoblasts. To test this, we utilized histone modification data from the Roadmap Epigenome Project [98]. In the Morris *et al.* BMD GWAS, 48 LDC genes overlapped 84 (7.6% of the 1103 total) associations (a subset of LDC genes overlapped multiple clustered associations). For each of the 84 independently associated lead SNPs, we analyzed histone modifications across the osteoblast genome and observed that they were more likely to overlap regions marked by modifications associated with active regulatory elements such as H3K4me1 (2.8x enrichment, $P < 1x10^{-3}$), H3Kme2 (3.2x enrichment, $P < 1x10^{-3}$), H3K4me3 (3.8x enrichment, $P < 1x10^{-3}$), and H3K27ac (2.6x enrichment, $P < 1x10^{-3}$) relative to 1000 sets of random SNPs matched for allele frequency and distance from a transcription start site (**Figure 4.5**). Additionally, we observed depletion of LDC SNPs in heterochromatic regions marked by H3K9me3 (0.14x depletion, $P < 1x10^{-3}$). To determine if the enrichments were specific to osteoblasts, we calculated the ratio between the LDC BMD set overlap and the mean random set overlap across all 129 Roadmap tissues and cell-types. For all activating marks (H3K27ac, H3K4me1, H3K4me2, H3K4me3) osteoblasts were in the top 10% when tissues were ranked based on the overlap ratio (**Supplemental File 4.13**). The tissues for which the random sets had a higher ratio included cell types related to osteoblasts, such as mesenchymal stem cell (MSC) derived chondrocytes and other MSC-derived tissues including adipose and skeletal muscle. These

data support the premise that loci harboring LDC genes impact BMD through the regulation of gene expression in osteoblasts, further supporting the causality of LDC genes.

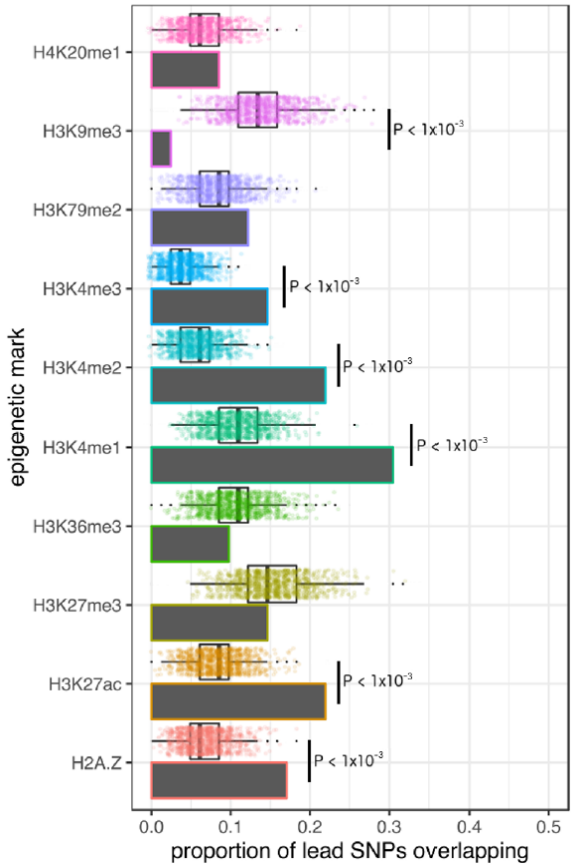


Figure 4.5 *Lead SNPs for GWAS associations harboring LDC genes overlap active regulatory elements in osteoblasts.* Grey bars represent the proportion of LDC SNPs (n = 84) that overlap each of the epigenetic marks measured in osteoblasts. Box and dot plots represent the proportion of each set of random SNPs (N = 1000) (matched to the LDC SNPs for MAF and distance from TSS) overlapping each epigenetic mark measured in osteoblasts.

4.3.7 The LDC genes *CADM1*, *B4GALNT3*, *DOCK9*, and *GPR133* are novel genetic determinants of BMD

The overarching goal of this study was to identify causal genes for BMD GWAS associations. As described above, 48 (14.9%) LDC genes overlapped an eBMD GWAS association from the Morris *et al.* study. To further identify those with strong evidence of

being causal, we utilized expression quantitative trait locus (eQTL) data from the Gene Tissue Expression (GTEx) project to identify local eQTL colocalizing with BMD associations [164]. We also used total body BMD data on LDC gene knockouts collected as part of the International Mouse Phenotyping Consortium (IMPC) [178]. Together, these data allowed us to directly link BMD associated variants to LDC genes and LDC genes to pathways regulating BMD. We performed a colocalization analysis for each eQTL/BMD association pair for all 48 genes in all tissues and identified 12 LDC genes with colocalizing eQTL with a significant posterior probability of colocalization (PPH4>0.7) (**Supplemental File 4.14 and Figures 4.6A, B, C, and D**). The IMPC has measured total body BMD via DEXA scan on a large collection of mouse knockouts. We queried each of 12 LDC genes with a colocalizing eQTL and found that 5 (41.7%) mutants had been analyzed for BMD. Of these, four genes (*Cadm1*, *B4galnt3*, *Dock9*, and *Adgrd1*) had significantly altered total body BMD ($P_{adj} < 0.05$) (**Supplemental File 4.15 and Figures 4.6E, F, G and H**). For *Cadm1* and *Dock9* the direction of effect inferred from the eQTL/BMD association matched the direction of the effect observed in the mouse knockout; however, for *B4galnt3* and *Adgrd1* the directions did not match (**Supplemental File 4.15**). Together, these data strongly support *Cadm1*, *B4galnt3*, *Dock9* and *Adgrd1* as core genes and causal regulators of BMD.

Lastly, we evaluated network parameters of *Cadm1*, *B4galnt3*, *Dock9* and *Adgrd1*. We observed that *Cadm1* and *B4galnt3* were ranked in the top 20 based on LDC connectivity (**Supplemental File 4.12**). In fact, *Cadm1* was the 2nd most highly connected gene. Together, the four genes had, on average, higher module membership than the average LDC gene (0.72 vs. 0.52; P = 0.002). In support of the importance of connectivity in the LDC, we observed that more highly connected LDC genes were more likely (P=0.008) to overlap a BMD GWAS locus (**Supplemental Figure 4.3A**) and there was a strong positive correlation

between connectivity and *in vitro* mineralization for all LDC genes (r = 0.71, P<2.2x10^{-16}) (**Supplemental Figure 4.3B**). These data suggest that connectivity is an important feature of the LDC and a strong proxy for biological importance.

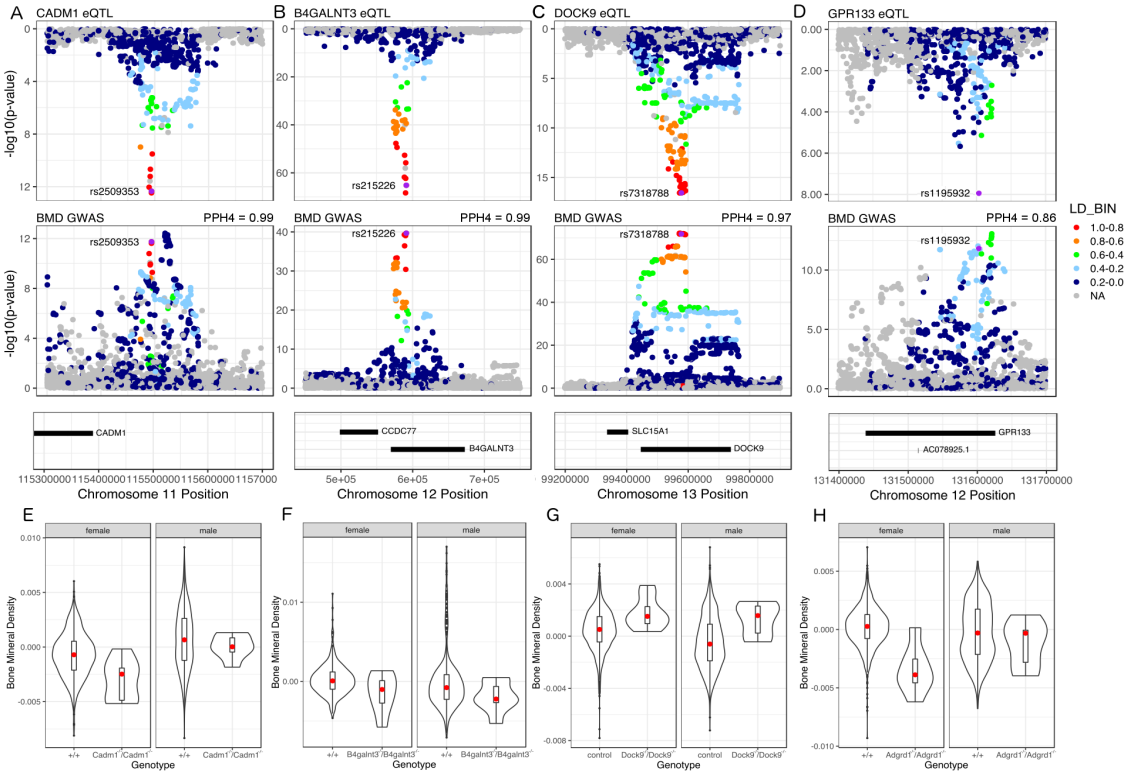


Figure 4.6 *Adgrd1, B4galnt3, Cadm1, and Dock9 are novel regulators of BMD.* (A-D) All four genes have an eQTL in at least one tissue in the GTEx database that colocalizes with a proximal BMD GWAS association. (E-H) Knockout mice from the KOMP for each gene exhibit altered BMD.

4.4 Discussion

Osteoporosis is an increasingly common disease associated with reduced BMD and negative health outcomes, namely fracture [165]. Despite the prevalence of the disease, we still do not fully understand the genes and mechanisms that influence its determinants, such as BMD. Moreover, current therapeutics for osteoporosis have been associated with rare, but severe side effects, leading to decreased compliance [14]. Identification of the causal core genes

that regulate BMD will help us to further understand the etiology of the disease and lead to the development of novel therapeutics. In this study, we identified the LDC, a module enriched for core genes influencing BMD, by integrating a cell- and timepoint-specific co-expression network with the results of BMD GWAS. Furthermore, we identified four LDC genes that overlap a GWAS locus, have colocalizing eQTL, and exhibit altered BMD in knockouts, suggesting they are causal for their respective BMD GWAS association.

Many have debated the utility of the core designation proposed in the omnigenic model [39,41,42,169], as the definition is quite narrow, including only genes whose "product (protein, or RNA for a noncoding gene) has a direct effect--not mediated through regulation of another gene--on cellular and organismal processes leading to a change in the expected value of a particular phenotype" [40]. We found, however, that employing the properties of core genes was helpful in identifying a core module enriched for causal genes underlying BMD GWAS loci. We leveraged the ideas that core genes would be related to key biological processes related to BMD regulation and that perturbation of core genes would lead to significant changes in the phenotype [39,40] and identified modules enriched for genes exhibiting these characteristics. We used gene ontology analysis, the correlation between gene expression and *in vitro* mineralization, and enrichment for genes that produced a bone phenotype in mice when knocked out and genes that drive monogenic bone disease in order to identify the purple module as a core module for BMD. The purple module was enriched for genes that have a demonstrated role in regulating BMD, many of which have been implicated by previous BMD GWAS, including *RUNX2* [191], *ESR1* [192], and *SOST* [193].

We compared the Kemp *et al.* study with the Morris *et al.* study, which identified twice as many candidate modules enriched for BMD GWAS genes [36,37]. Despite the increased number of candidate modules, we found that the purple module was consistently

exceptional among the modules enriched for GWAS genes. We also observed that increased GWAS sample size led to an increase in diversity of the biological processes represented by the modules enriched for GWAS genes. These results support the premise of the omnigenic model that any gene that is expressed in a cell-type of interest will be associated with a phenotype, given enough power [39]. Overall, these results indicate that our method of utilizing a cell-type specific co-expression network and core-related *in vitro* and *in vivo* phenotypes effectively led to the identification of a module of core genes.

Using GTEx eQTL data and *in vivo* mouse phenotypes, we provided strong supporting evidence that *CADM1*, *B4GALNT3*, *DOCK9* and *GPR133* are novel regulators of BMD and causal for their respective GWAS association. None of these genes had been previously directly connected to the regulation of BMD. *CADM1* (Cell Adhesion Molecule 1) is a ubiquitously expressed cell adhesion molecule involved in many biological processes, including cancer, spermatogenesis, and neuronal/mast/epithelial cell function [194–196] that had been implicated in osteoclast proliferation and activity [197] and as an osteoblast-specific marker in the context of osteosarcoma [198,199]. *B4GALNT3* (Beta-1,4-N-Acetyl-Galactosaminyltransferase 3) is a glycosyltransferase that transfers N-acetylgalactosaine (GalNAc) onto glucosyl residues, thus forming N,N-prime-diacetyllactoseadiamine (LacdiNAc), which serves as a terminal structure of cell surface N-glycans that contributes to cell signaling [200,201]. *B4GALNT3* is expressed in bone and associated with circulating levels of sclerostin [16,202,203]. *DOCK9* (Dedicator of Cytokinesis 9) is a guanine nucleotide-exchange factor that activates *Cdc42* [204], which has been shown to regulate osteoclast differentiation and ossification [205,206]. *GPR133* (Adhesion G Protein-Coupled Receptor D1) is a G protein-coupled receptor that participates in cell-cell and cell-matrix interactions [207]. Our results demonstrate the utility of the LDC in increasing our understanding of the molecular and

genetic basis of BMD. In addition to identifying core genes potentially responsible for GWAS associations, the use of networks to inform GWAS has the added benefit of providing insight into novel gene function. By using GO enrichments, gene expression across differentiation, and membership of key lineage genes, it was clear the LDC contained genes important for the process of mineralization. Also, using epigenomics data on human osteoblasts from the ROADMAP project [98], we were able to show that the lead BMD GWAS SNPs for associations overlapping LDC genes were more likely to fall in active regions of the genome in osteoblasts. The implicated SNPs overlapped activating marks including H3K4me2, which marks transcription factor binding sites [208] , and transcription start sites [209], and H3K27ac, H3K4me2, and H3K4me3, which all mark active enhancers [210–214]. Additionally, we observed depletion of LDC SNPs in heterochromatic regions, marked by H3K9me3 [215,216]. Overall, these data supported the role of the underlying LDC genes as core genes with a causal role in the process of osteoblast-mediated mineralization.

While these findings are promising, additional data would help to clarify our results. For example, the eQTL used in this study were not derived from expression data in bone tissue, as bone tissue expression was not measured in the GTEx project. While we identified colocalizing eQTL in other tissues, only two of these eQTL/GWAS associations relationships match the results of the mouse knockout phenotypes, however this may be because the direction of effect of the eQTL in the surrogate tissue does not reflect the direction of effect in osteoblasts. Thus, a comprehensive characterization of eQTL in bone cells would be beneficial for future studies. Additionally, there is not complete concordance in the direction of effect between the Estrada *et al.* GWAS study of BMD measured via DEXA scan and the Kemp *et al.* and Morris *et al.* studies of eBMD [35–37]. In the Kemp *et al.* study, six specific cases are outlined in which the opposite direction of effect is observed for

BMD and eBMD for the same variants [36]. These differences could also contribute to our observation that the direction of effect between the eQTL/GWAS associations is not consistent with the mouse knockout for two of our candidate genes. Moreover, this is not a comprehensive study of the determinants of BMD, as we used gene expression data from the mouse as a discovery platform. Using mouse cells to generate our network may limit the translational applications of the work due to missing homologs between mouse and human. Furthermore, more detailed mechanistic studies of the four genes we identified will be needed to definitively determine the mechanism by which they regulate osteoblast activity.

While we identified four novel regulators of bone mineral density, there is still much to be gleaned from the late differentiation cluster. For example, *SLC8A3* (solute carrier family 8 member 3, aka *NCX3*) is a sodium/calcium exchanger that controls calcium homeostasis [217] and is a highly interconnected gene within the module. *Slc8a3* overlapped a BMD locus identified in a meta-analysis [218] however, summary statistics for the region were not reported, so we could not conduct colocalization analysis between this association and the local *SLC8A3* eQTL. Additionally, *SLC8A3* knockout mice have been generated and extensively phenotyped for bone microarchitecture and histomorphometry and results indicate that *SLC8A3* likely plays a role in bone, though much remains to be studied. *SLC8A3* is just one example of the wealth of information in the LDC that we have yet to tap into. There are still many genes with no known connection to BMD in the LDC that are likely very important to osteoblast biology and mineralization. However, the LDC is not just a list of candidate genes; it also provides insight into the molecular hierarchy driving osteoblast differentiation and mineralization, which can provide biological context that can help lead to the identification of key drivers of these processes. Furthermore, we can utilize the property of LDC membership to infer that candidate genes within the LDC likely play a

role in the process of mineralization. Moving forward, the LDC can serve as a platform for the identification of novel determinants of BMD.

Finally, this approach could also be applied to other bone cell types. For example, one could use *in vitro* measures of bone resorbing osteoclast activity as a filter to identify groups of genes influencing osteoclast activity and ultimately, BMD. This workflow could be useful for determining the cell-type specific contributions of other isolated cell types in complex, tissue-level phenotypes, in particular those for which there is an illustrative *in vitro* model.

Overall, we have used an integrative, network-based method to identify core genes for the process of mineralization and BMD. While the definition of a core gene is still open to debate, we found the expected properties of core genes are effective lenses through which to contextualize GWAS associations. Integrating gene co-expression networks, GWAS data, *in vitro* and *in vivo* phenotypic data, and eQTL information has led us to a more complete understanding of the biology and genetics of BMD.

4.5 Acknowledgements

Research reported in this publication was supported by the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health under Award Number R01AR064790 to CLA-B and CRF. OLS was supported by a Wagner Fellowship from the University of Virginia. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript (v7) were obtained from: the GTEx Portal on 01/11/2018. The International Mouse Phenotyping Consortium is partially funded by the

NIH Knockout Mouse Programme (KOMP) project and the IMPC informatics and the data portal are supported by NIH grant U54 HG006370.

This project was designed by CRF, CLA-B, and OLS. Sample processing was carried out by GMC. All statistical analysis was conducted by OLS and ERT, overseen by OLS. This manuscript was prepared by OLS with guidance and input from CRF.

4.6 Methods

RNA Sequencing

Neonatal collaborative cross heads were received from the University of North Carolina. At UNC, neonatal (3-5 days) collaborative cross mice were euthanized by CO_2, decapitated onto paper towels soaked in 70% ethanol, and placed in cold PBS on ice for overnight shipping. Once received, calvaria were dissected, paying special attention to brain and interparietal bone removal. Isolated calvaria were placed in 24 well plates containing 0.5 mL of digest solution (0.05% trypsin and 1.5 U/ml collagenase P) and incubated on a rocking platform at 37 degrees during six, fifteen-minute digestions in 0.5 mL of digestion solution. Fraction 1 is discarded and fractions 2-6 are collected. Fractions 2-6 are added to an equal volume of cold plating media (89 mL DMEM, 1 mL 100x Pen/Strep solution, and 10 mL Lot tested FBS). The resulting cells are filtered using a 70-100 mm cell strainer to remove clots, centrifuged at 1000 rpm for 5 minutes and re-suspended in 0.5 ml plating media. The resulting cells are plated in a T25 flask. 24 hours later, cells are washed with PBS, treated with trypsin, counted, and plated at a density of $1.5x10^5$ cells per well in a 12-well plate, and allowed to grow to confluence for 48 hours. After 48 hours of growth, cells are switched to differentiation media (10 mL lot tested FCS, 1 mL 100x Pen/Strep solution, 283.8 uL ascorbic acid (0.1 M), 400 uL B-glycerol phosphate (1 M), and 88.3 mL alpha-

MEM per 100 mL) and allowed to differentiate for 10 days. On day 10, total RNA was extracted from the mineralized cultures using *mir*Vana RNA isolation kit (ThermoFisher Scientific).

RNA-Seq libraries were constructed from 200 ng of total RNA using Illumina TruSeq Stranded Total RNA with Ribo-Zero Gold sample prep kits (Illumina, Carlsbad, CA). Constructed libraries contained RNAs >200 nt (both unpolyadenylated and polyadenylated) and were depleted of cytoplasmic and mitochondrial rRNAs. An average of 39.7 million 2 x 75 bp paired-end reads were generated for each sample on an Illumina NextSeq 500 (Illumina, Carlsbad, CA). FastQC was used to evaluate the quality of the reads, and all samples passed the QC stage [219]. Reads were mapped to the eight collaborative cross founder transcriptomes using Bowtie, and quantified using EMASE [220]. EMASE output transcript level expression estimates calculated by assigning multi-mapping reads across the genome using and expectation-maximization algorithm to allocate reads that differentiate between genes, then isoforms of a gene, and then alleles (GSE134081).

WGCNA network construction

Estimated transcript count data was used as the basis for co-expression network construction. We removed transcripts with less than an average tpm <= 0.3 tpm across all samples, resulting in 29,000 transcripts used to construct the network. We used a variance stabilizing transformation from the DESeq2 package that decouples the variance from the mean [221]. Next, we used PEER in order to remove latent confounding batch effects from our data [222]. As per PEER recommendations, we estimated PEER factors equal to one quarter of the number of samples (N = 24) and included covariates in the calculation. We carried out the downstream analysis with the residual values from PEER transformation.

Finally, we used quantile normalization to match the distribution of each of the samples in the analysis.

The resulting expression data was used to construct a signed, weighted gene co-expression network using the weighted gene co-expression network analysis (WGCNA) package [223]. There were no evident outliers from the hierarchical clustering analysis. The pickSoftThreshold() function from the WGCNA package was used to determine the power used to calculate the network. The minimum power value that had an $R^2 >= 0.9$ for the scale-free topology model fit was used, and the network was calculated using a power of 9. We then used the blockwiseModules() function to construct a signed network with a merge cut height of 0.15, and a minimum module size of 20 genes. Using WGCNA, we constructed a signed network composed of 65 modules of co-expressed genes, with an average of 292 genes per module.

Gene Ontology Analysis

For those modules that were enriched for BMD GWAS genes, we conducted gene ontology analysis to identify the functional categories represented by each module. Using the ToppFun tool on the ToppGene site, we identified the significantly enriched categories for GO molecular functions, GO biological processes, GO cellular components, human and mouse phenotypes, and pathways [224]. The significance cutoffs reported for these enrichments were Benjamini & Hochberg corrected FDR q-values.

Creating BMD GWAS list

In order to identify co-expression modules enriched for BMD GWAS genes, we identified all genes overlapping a BMD GWAS locus using the 2012 and 2017 BMD GWAS [35,36]. For each BMD locus, a bin was defined by the furthest upstream and downstream SNPs with LD >= 0.7 as calculated from the European populations in the 1000 genomes phase III

data identified using the LDLink LDProxy tool [163]. Then, using the Genomic Ranges tool, we identified all genes from the GRCh37/hg19 Ensembl gene set overlapping a BMD GWAS bin [225,226]. If no gene intersected a bin, we identified the nearest upstream and downstream genes from the bin. The Estrada GWAS resulted in 179 genes and the eBMD GWAS resulted in 701 genes, resulting in a list of 731 unique genes. We converted the list of human genes to mouse homologs.

BMD GWAS gene enrichment

In order to identify modules of genes enriched for GWAS genes, we used a fisher's exact test to measure the statistical significances of the representation of GWAS genes in each module. We then applied a Bonferroni correction to correct for testing the enrichment of all 65 modules, and applied a significance cutoff of 0.05 to the adjusted p-values, resulting in 13 modules of genes enriched for 2012 and 2017 GWAS genes, and 26 modules of genes enriched for 2012, 2017, and 2018 GWAS genes.

LD Score Regression

In order to evaluate the relevance of the BMD GWAS gene enriched modules we calculated the partitioned heritability of the SNPs in the regions surrounding the genes in each module. We used the LD score regression method, which takes gene lists as an input and returns the enrichment of the associated SNP set for heritability for the tested trait. For each set of modules, we tested using this method, we corrected the enrichment p-values for multiple testing using a Bonferroni correction, and applied a p-value cutoff of 0.05 to the adjusted p-values.

In vitro mineralization measurement and correlation

In order to identify the modules of co-expressed genes with patterns of expression correlated with mineralization, we measured *in vitro* mineralization in osteogenic cells from

the calvaria of 42 strain of collaborative cross mice. After 10 days of differentiation and mineral production, cells are washed with PBS and treated with 10% NBF (1 mL per well) and incubated at room temperature for 15 minutes. The NBF is removed and cells are washed with H_2O (1mL x 2). Next, wells are stained with alizarin red (0.5 mLs, 40 mM @ pH 5.6) for 20 minutes on a shake plate at 120 rpm. Alizarin red stain is then removed, and cells are washed 5 times with deionized H_2O for 5 minutes on a shake plate at 180 rpm. Once rinsed, the mineralized wells are scanned, and .tiff images are retained to extract geometric parameters of the mineral deposits. After imaging, the wells are de-stained by incubation with 5% perchloric acid (1 mL) at room temperature for 5 minutes while shaking at 120 rpm. Eluent is collected and read at 405 nm. The levels of *in vitro* mineralization varied significantly across the population, with a 63-fold change from the highest to lowest mineralization samples (max_mmAR = 2.995993, min_mmAR = 0.04719, mmAR = millimolar alizarin red).

In this population, *in vitro* mineralization had a heritability of 47.8% ($p=1.8x10^{-46}$), indicating that the between-strain variation is larger than the within strain variation and that there is a genetic contribution to the process of mineralization. Using the WGCNA package, the eigengene of each module was calculated, and the correlation between the eigengene and the in vitro mineralization phenotype was calculated using the cor() function in R. The p-values associated with the correlation between the module eigengenes and *in vitro* mineralization were corrected for multiple testing using a Bonferroni correction and a p-value cutoff of 0.05 was applied to the adjusted p-values.

Module enrichment for genes with associated bone phenotypes and monogenic bone disease

In order to identify modules of co-expressed BMD GWAS genes that are enriched for genes with bone phenotype annotations, we curated a list of genes which produce a bone

phenotype when knocked out. We used four databases of gene perturbations that result in bone phenotypes, including genes annotated with a bone phenotype in the Mouse Genome Informatics database (MGI), the Origins of Bone and Cartilage Disease (OBCD) database, the International Mouse Phenotyping Consortium (IMPC), and the Bonebase Database [92,177–179]. Specifically, we pulled BMD, altered bone morphology, altered bone cell activity, changes in ossification or mineralization, or association with a known bone disease from the MGI database. The OBCD database contained genes with changes in bone mineral content (BMC), bone volume fraction (BV/TV), and BMD of the femur and BMD of the vertebra. We mined the IMPC database for any genes with altered BMD, and we pulled all Bonebase genes with altered BV/TV in the femur or vertebra. This resulted in a list of 923 unique "bone" genes (**Supplemental File 4.7**).

We also curated a list of genes associated with monogenic bone disorders using a literature review, specifically focusing on genes that disrupt osteoblast function, leading to monogenic bone disorders [180–184]) (**Supplemental File 4.8**).

We used a fisher's exact test to measure the statistical significance of the representation of genes with associated mouse knockout bone phenotypes and monogenic bone disease in each module. We then applied a Bonferroni correction to correct for testing the enrichment of all 13 or 26 modules and applied a significance cutoff of 0.05 to the adjusted p-values.

Clustering analysis in osteoblast differentiation gene expression data

We investigated the expression profiles of all purple module genes in the context of differentiation. Using gene expression data from osteoblasts throughout differentiation (Series GSE54461), we used k-means clustering to identify differentiation-related transcriptional programs in the purple module. We tested k = 1:5, and found two robust

clusters of genes within the purple module. Enrichment analysis of the two cluster were assessed for enrichment in all function categories as described above.

Epigenetic enrichment analysis for LDC BMD GWAS associations

For BMD GWAS lead SNP (and proxies with LD >= 0.7) overlapping an LDC gene (n = 84), GenomicRanges [225] was used to calculate the proportion of lead SNPs overlapping regions marked by epigenetic modifications, including H3K4me1, H3K4me2, H3K4me3, H3K9me3, H3K27ac, H3K27me3, H3K26me3, H3K79me2 and H4K20me1, and histone H2AZ from the Roadmap Project [98]. Using the GenomicRanges function findOverlaps(), we quantified the overlap between the LDC-associated lead SNPs and each epigenetic mark. To assess the enrichment of this overlap, we compared against 1000 sets of control SNPs (n = 84). We chose sets of control SNPs that were within +/- 20% of the mean distance from a transcription start site for the BMD GWAS lead SNPs, and within +/- 20% of the mean minor allele frequency of the BMD GWAS lead SNPs. P-values were calculated by taking the proportion of random sets of SNPs with a more extreme enrichment in the tail of the distribution with which we are comparing our experimental proportion. If the experimental proportion is more extreme than any measured random set, the p-value is reported as < $1x10^{-3}$. This same procedure was used to evaluate the tissue specificity for each mark. For each mark, the overlap with the LDC BMD SNP set and the 1000 random SNP sets were computed and the ratio between the proportion of overlapping LDC BMD SNPs and the mean proportion of overlapping random SNPs was computed. Higher ratios indicated greater enrichment of the LDC BMD SNPs over random SNPs with a given mark in a given tissue.

Colocalization analysis

For each gene in the LDC that overlapped a BMD GWAS association from the Morris *et al.* study, eQTL from all GTEx tissues were identified [164,168]. Using the coloc package, we assessed the potential for colocalization between the QTL for BMD and the proximal cis-eQTL [158]. Two associations were considered to colocalize if the posterior probability of hypothesis four (PPH4), which is the probability of colocalization, is > 0.7. The RACER package to plot the two associations in a mirrorPlot [227].

Mouse phenotype statistical comparisons

Using the International Mouse Phenotyping Consortium (IMPC) database, we identified genes from the purple module that have eQTL that colocalize with BMD QTL and exhibit a difference in BMD when knocked out in mouse [178]. Using the PhenStat package, we analyzed the differences between control and knockout animals using a mixed model framework [228]. The specific equation used for each analysis are in **Supplemental File 4.15**.

Network Topology Analysis

A t-test was used to compare the module membership of the four causal genes and the remainder of the LDC genes and the connectivity of the LDC genes overlapping a BMD GWAS locus as opposed to those that do not. A linear model was used to assess the relationship between gene connectivity and gene correlation with *in vitro* mineralization.

Chapter 5

Concluding Remarks and Future Directions

Despite the success of a decade of GWAS aimed at uncovering the genetic basis of osteoporosis, our understanding of the genes and mechanisms driving these genetic associations has been modest. Until now, the majority of follow up studies have focused on three key areas: (1) candidate gene studies related to known pathways involved in regulating bone mineral density, which are limited in scope, (2) fine-mapping approaches, which help narrow the set of potential causal variants, but do not provide biological context, and (3) network-based approaches, which rely on molecular data in tissues of interest [167]. These approaches have provided limited improvements in our understanding of the genetic basis of BMD and thus, in this work, we set out to develop new approaches to identifying causal genes underlying genetic associations for osteoporosis-related traits and follow up on previously identified associations with the goal of identifying novel genes influencing BMD.

We began by investigating a QTL for femoral length (*Feml2*), identified in a cross between C57BL/6J-*bg*/*bg* (HG) and CAST/EiJ mice [136]. *Feml2* was located on Chromosome 9 and had been captured in a congenic strain (HG9) [139]. We utilized full genome sequence data from the C57BL/6J and CAST/EiJ strains to identify variants potentially responsible for the effects of *Feml2*. We then performed RNA-seq on growth plates from HG9 mice and identified 6 genes exhibiting allele-specific expression in C57BL/6J x CAST/EiJ F1 mice. We also identified that the human genomic region syntenic to *Feml2* was a hotspot for associations with height, indicating that this region may harbor multiple regulators of femur length [68]. This study demonstrated the power of mouse genetics in the identification of novel genes underlying associations for osteoporosis traits other than BMD.

Our most exciting findings were from the integration of a co-expression network and BMD GWAS data to identify novel causal genes underlying BMD GWAS. To identify novel genes influencing BMD, we focused on one of the fundamental processes regulating

BMD—mineralization. By measuring gene expression in bone-forming osteoblasts, we were able to focus on identifying genes involved in mineralization. We constructed a co-expression network using gene expression data from mineralizing osteoblasts and identified the co-expression modules enriched for genes implicated by BMD GWAS. Next, we utilized the concept of core genes, as described in the omnigenic model [39,40]. Core genes are statistically defined as causal GWAS genes that have a direct effect on a trait, conditionally independent of all other genes. Identifying core genes by this definition is difficult, requiring a full understanding of the mechanisms through which a gene impacts a phenotype. We defined core genes based on a biological rather than statistical definition [41,42]. We found utility in applying the anticipated properties of core genes to the modules enriched for genes implicated by BMD GWAS, identifying one core module for BMD, from which we could identify novel genes influencing BMD. The four genes we identified, *Adgrd1*, *B4galnt3*, *Cadm1*, and *Dock9*, have the potential to serve as novel therapeutic targets for osteoporosis.

Moving forward, increasing the genomic resources available from human bone tissue will be critical to improving this approach. One of the obstacles hindering causal gene discovery for bone GWAS is the paucity of population-scale transcriptomic and epigenomic data from bone tissue or primary bone cells. There are a number of large repositories of expression and epigenetic data from particular tissue types; however, there is no such resource for the bone field yet. For example, the Gene Tissue Expression (GTEx) project is an NIH funded effort to generate RNA-seq expression profiles (and soon epigenetics and proteomics data) from multiple tissues (>40) in a large genotyped human cohort [109]. The resulting resource is extraordinarily powerful and provides the opportunity to understand how genetic variation influences expression on a genome-wide basis. One of the primary efforts of GTEx to date is to provide the genetics and genomics community with eQTL

results, allowing investigators to use these data to inform GWAS. Unfortunately, GTEx is not collecting bone tissues and primary bone cells. Our group and others are using the GTEx resource, along with data from two small bone eQTL studies [229,230] to inform BMD GWAS; however, this only works for signals shared between bone and non-bone tissues or genetic effects on bone that arise from expression changes in non-bone tissues. Related resources, such as the ENCODE project [97], which focuses on cataloging functional elements found in the genome, does contain histone modification and DNase I hypersensitivity site data on primary osteoblasts and related cells such as mesenchymal stem cells and chondrocytes, though not from other bone cells. Thus, there is a significant need in the bone field for generating "-omics" data on bone and primary bone cells that can be used for causal gene discovery and as an independent discovery platform.

As we have discussed above, one of the major bottlenecks in our understanding of how genetic variation leads to differences in traits, such as BMD, is identifying causal genes for existing GWAS loci. However, that does not imply that the utility of GWAS has been exhausted. To the contrary, there is much more to discover, even for BMD, especially in light of the observation that the most recent BMD GWAS only explains ~20% of the heritability in BMD [37]. Heritability estimates for BMD are generally >50%, so to date GWASs have explained roughly 40% of its genetic component [37]. Additional GWASs are ongoing and there will no doubt be larger meta-analyses for BMD that will yield even more loci. Additionally, GWAS conducted in understudied ethnic populations, such as the meta-analysis that identified a genome-wide significant locus for LDC gene *Slc8a3* conducted in East Asians [231], will also be important in the study of the genetics of BMD.

There is also a need for continued investigation of phenotypes beyond BMD. In particular, GWASs for traits that account for aspects of bone strength independent of BMD

would be ideal, for example bone size and microarchitecture. Though many of these traits are more difficult to measure than BMD, their interrogation would provide significant insight into the genetics of bone strength and fracture risk. Such studies are already starting to be performed in large-scale cohorts. For example, trabecular microarchitecture as measured by quantitative computed tomography (QCT) was recently investigated by GWAS in a cohort of ~15,000 individuals [232]. We anticipate the trend of GWAS for a more diverse set of bone-strength related traits will increase in the coming years. In addition to more diverse phenotypes, a powerful application of GWAS would be to separate the genetic analysis of bone accrual and bone loss, especially in light of the observation that the genetic correlation between BMD in pre- and postmenopausal women is modest (r=0.30) [233]. Bone is accrued until peak bone mass between the ages 20 and 25. GWAS in pediatric populations could therefore be used to identify loci specifically affecting the attainment of peak bone mass [234]. Small GWASs for BMD in pediatric populations have already started to provide insight [235]. We are also interested in developing a much better understanding of genetic loci affecting bone loss due to aging in both sexes, or more importantly, bone loss in females after menopause. The dramatic loss of bone after menopause in women is the primary reason why 80% of the 12 million Americans with osteoporosis are female [7]. To date, GWAS has not been used to study bone loss in postmenopausal women, although it is the single strongest determinant of poor bone health in women. Lastly, with the exception of GWAS studies for BMD in Asian populations (as examples [231,236]), most GWAS for bone traits have been performed in individuals of European ancestry [35]. In order for GWAS results to inform drug discovery that is applicable to all populations of people it is imperative that GWASs for all bone traits be performed in populations with diverse ethnic backgrounds.

As discussed above, in the context of causal gene discovery it is imperative that transcriptomic, and other "-omics" data, on bone and bone cells be collected in large genotyped populations. Without these resources it will be difficult to definitively identify causal variants and genes. However, in combination with these resources there is a need for better methods for defining causal variants and genes. Ideally such methods will include computational strategies for defining cis-acting regulatory sequences and accurately predicting the effects of genetic variation on these sequences [237]. Furthermore, experimental methods are needed to query large number of variants on regulatory sequence function [238]. Currently, most experimental strategies rely on reporter assays that assess regulatory function of variants using artificial systems that investigate sequences outside of their native chromosomal context [60]. There are, however, new methods, such as using lentiviral based reporter assays that integrate into the genome that are very promising [239]. Additionally, CRISPR/Cas9 based genome-editing approaches that allow one to modify individual variants in human cells are becoming more efficient and will likely play a large role in causal variant and gene discovery moving forward [240]. Such experiments will need to be performed in human cells, thus cell lines that mimic their *in vivo* counterparts are urgently needed. An attractive alternative to cell lines is bone cells derived from induced pluripotent stem cells (iPSCs). These would be the ideal resources for the necessary human cell studies and groups have already demonstrated that iPSCs can be used to derive osteoblasts [241] and osteoclasts [242].

In conclusion, BMD GWAS has provided a rich resource for the study of the genetic determinants of BMD, however alternative approaches are required to understand the genetic basis of osteoporosis more fully. Mouse studies of osteoporosis-related traits that are difficult or impossible to measure in humans will provide a greater understanding of the full set of parameters that govern bone strength and may unlock alternative approaches to

increasing strength, independent of BMD. New methods for the identification of novel genes underlying BMD GWAS associations have the potential to unlock new insight into the etiology and treatment of the disease. And as our understanding of the genetic architecture of complex traits evolves, so too should our approach to identifying novel genes driving complex phenotypes. This work has contributed to this goal by identifying novel genes influencing bone geometry and BMD and by developing new tools and approaches for identifying and contextualizing genes underlying GWAS associations, laying the foundation for future studies of complex traits.

Appendix A

Supplemental Data

All supplemental data are available at:

https://github.com/oliviasabik/supplemental_dissertation_data

Supplemental file 2.1 QTL mapping information, including mouse IDs, phenotype information, and SNP marker identifiers, locations, and genotypes.

Supplemental Data 3.1 Example BMD GWAS association data from the Chr. 14q32.32 locus.

Supplemental Data 3.2 Example GTEx eQTL data for MARK3 from thyroid.

Supplemental File 4.1 Co-expression network data. Contains summary data, gene significance (GS) and module membership (MM) information for each module for each transcript.

Supplemental File 4.2 Gene ontology results for all 65 co-expression modules.

Supplemental File 4.3 Mouse and human homologs of the genes overlapping BMD GWAS associations from the Estrada *et al.* study [35].

Supplemental File 4.4 Mouse and human homologs of the genes overlapping BMD GWAS associations from the Kemp *et al.* study [36].

Supplemental File 4.5 Results of enrichment analysis for Kemp and Estrada BMD GWAS genes.

Supplemental File 4.6 Results of LD score regression for the Kemp *et al.* GWAS for all 64 modules.

Supplemental File 4.7 Mouse genes that produce a bone phenotype when knocked out, curated from [92,177–179].

Supplemental File 4.8 Human genes that drive monogenic bone disorders, curated from [180–184].

Supplemental File 4.9 Mouse and human homologs of the genes overlapping BMD GWAS associations from the Morris *et al.* study [37].

Supplemental File 4.10 Results of enrichment analysis for Morris BMD GWAS genes.

Supplemental File 4.11 Results of LD score regression for the Morris *et al.* GWAS for all 64 modules.

Supplemental File 4.12 Purple module transcript annotations, including purple module membership, *in vitro* mineralization gene significance calculations, and osteoblast differentiation cluster membership.

Supplemental File 4.13 Tissue enrichments for epigenetic overlap analysis.

Supplemental File 4.14 Results of colocalization tests for all LDC genes and proximal BMD GWAS signals.

Supplemental File 4.15 Directions of effect for eQTL and proximal BMD GWAS signals and mouse knockout data for *Cadm1, B4galnt3, Dock9,* and *Adgrd1*.

Appendix B

Supplemental Figures and Tables

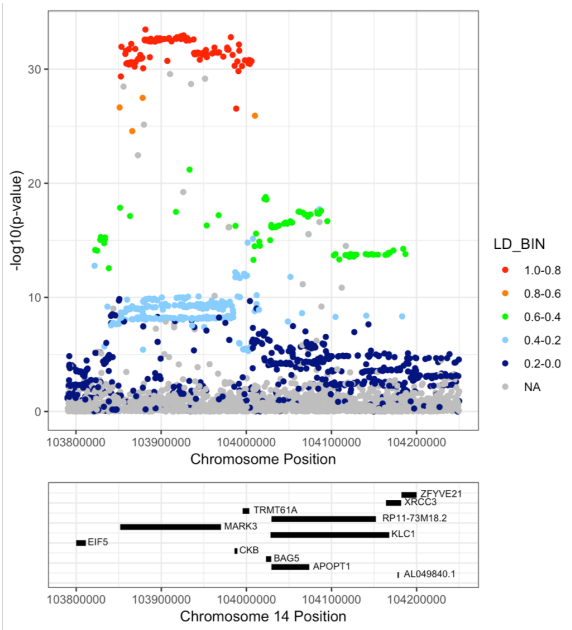
Supplemental Table 2.1 List of potentially high-impact CAST/EiJ variants within Feml2 genes.
*expressed in mouse growth plated as determined by RNA sequencing.

Chr	Position	Gene	dbSNP	Ref	CAST /EiJ	Consequence
9	57537750	2310046O06Rik	rs260748872	C	G	missense variant
9	57538878	2310046O06Rik	rs38947963	G	A	missense variant
9	57544878	*Mpi	-	T	C	missense variant
9	57545287	*Mpi	-	T	C	missense variant
9	57552713	*Mpi	-	T	C	missense variant
9	57609252	*Lman1l	rs39219434	G	T	missense variant
9	57611095	*Lman1l	rs38081297	T	C	missense variant
9	57611815	*Lman1l	rs36883533	C	T	missense variant
9	57611844	*Lman1l	rs235424109	A	T	missense variant
9	57612578	*Lman1l	rs36560024	G	A	missense variant
9	57613650	*Lman1l	rs223545612	A	G	missense variant
9	57620573	*Lman1l	rs265481243	C	T	missense variant
9	57620661	*Lman1l	rs255256004	T	C	missense variant
9	57681790	Cyp1a2	rs8236810	G	C	missense variant
9	57681997	Cyp1a2	rs8236815	G	A	missense variant
9	57700121	Cyp1a1	rs8250141	G	A	missense variant
9	57713398	*Edc3	rs36786250	A	G	initiator codon variant
9	57833545	*Gm17231	rs258660237	G	A	missense variant
9	57929714	*Ubl7	rs30066215	T	G	missense variant
9	57940299	*Sema7a	rs226337950	G	T	missense variant
9	57954532	*Sema7a	rs48774862	G	A	missense variant
9	58015195	*Cyp11a1	rs45867326	G	A	missense variant
9	58058276	*Ccdc33	rs29643506	T	C	missense variant
9	58076623	*Ccdc33	rs51621524	C	A	missense variant
9	58081975	*Ccdc33	rs29690624	T	G	missense variant

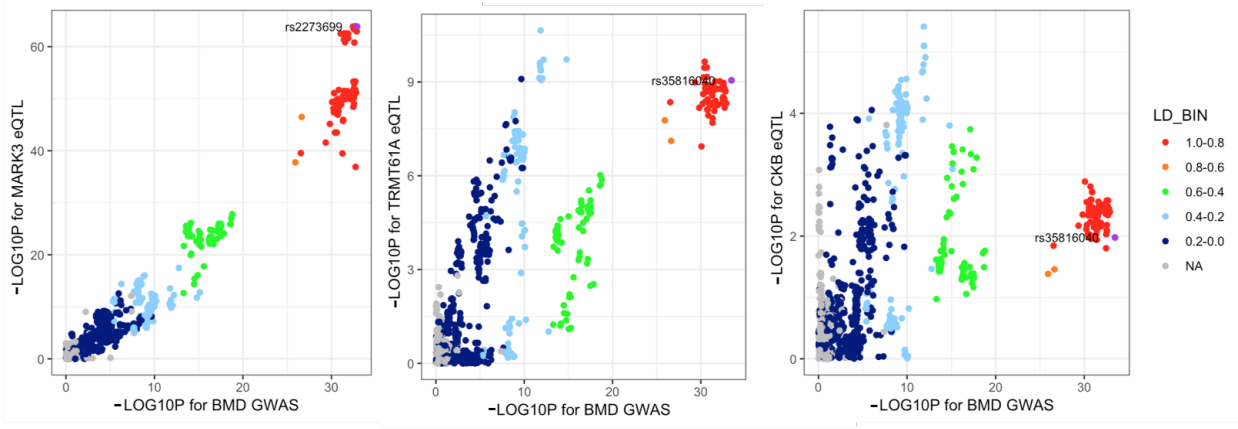
9	58143018	*Stra6	rs50728118	C	G	missense variant
9	58157937	*Islr	rs212222891	G	A	missense variant
9	58157952	*Islr	rs230884510	T	C	missense variant
9	58234760	*Pml	rs237113441	G	T	missense variant
9	58234761	*Pml	rs256533111	C	T	missense variant
9	58234914	*Pml	rs232090362	C	T	missense variant
9	58260402	*Stoml1	rs255702081	A	G	missense variant
9	58260854	*Stoml1	rs224651736	C	G	missense variant
9	58499197	*6030419C18Rik	rs252396713	C	T	missense variant
9	58530599	*Cd276	-	C	A	missense variant
9	58554844	*Gm10657	rs216192681	A	C	missense variant
9	58554863	*Gm10657	rs256521291	C	T	missense variant
9	58554899	*Gm10657	rs252702259	G	C	missense variant
9	58555164	*Gm10657	rs48739782	G	A	stop gained
9	58824166	*Hcn4	rs258098063	G	C	missense variant
9	58957975	*Neo1	rs48614416	T	A	missense variant
9	58978730	*Neo1	rs48614416	A	C	missense variant
9	59146096	*Gm7589	rs219826815	C	T	missense variant
9	59146165	*Gm7589	rs49703292	A	G	missense variant
9	59314743	*Adpgk	rs13480222	A	G	missense variant
9	59396591	*Arih1	-	C	T	missense variant
9	59709498	*Gramd2	rs237544748	A	G	missense variant
9	59713832	*Gramd2	rs220530268	T	C	missense variant
9	59713884	*Gramd2	rs37111673	A	G	missense variant
9	59855424	*Myo9a	rs36786203	G	A	missense variant
9	59870922	*Myo9a	rs223983082	T	G	missense variant
9	59871291	*Myo9a	rs242891687	T	A	missense variant
9	59871941	*Myo9a	rs249734798	C	A	missense variant
9	59884573	*Myo9a	rs261498011	G	T	missense variant

9	59921846	*Myo9a	rs37457286	C	A	missense variant
9	59982794	*Thsd4	rs37565778	A	G	missense variant
9	59987366	*Thsd4	rs6224703	T	C	missense variant
9	60002916	*Thsd4	rs38077312	C	T	missense variant
9	60428188	*Thsd4	rs39206755	C	T	missense variant
9	60428239	*Thsd4	rs38716900	G	A	missense variant
9	60428477	*Thsd4	rs38047306	T	C	missense variant
9	60587859	*Lrrc49	rs241323558	C	A	missense variant
9	60737152	*Larp6	rs38513944	C	T	missense variant
9	60737422	*Larp6	rs37357660	A	G	missense variant
9	60769608	*1700036A12Rik	rs263607526	C	G	missense variant
9	60769781	*1700036A12Rik	rs254161566	C	A	stop gained
9	60769804	*1700036A12Rik	rs224669653	C	T	missense variant
9	60769851	*1700036A12Rik	rs254164990	G	T	stop gained
9	60769942	*1700036A12Rik	rs257856766	C	T	missense variant
9	60769955	*1700036A12Rik	rs213290293	A	G	stop retained variant
9	60821859	*Gm9869	rs227683477	T	C	missense variant
9	60821898	*Gm9869	rs258591700	G	C	missense variant
9	60838083	*Gm9869	rs36651271	C	T	missense variant
9	60869590	*Uaca	rs234670406	A	G	missense variant
9	60869607	*Uaca	rs3667578	T	G	missense variant
9	60870316	*Uaca	rs215156422	C	T	missense variant
9	60870608	*Uaca	rs37765393	A	G	missense variant
9	62060441	*Glce	rs255376057	G	C	missense variant
9	62273017	Spesp1	rs240098849	T	A	missense variant
9	62273131	Spesp1	-	G	C	missense variant
9	62520391	*Coro2b	rs33689334	A	G	missense variant
9	62767727	*Itga11	rs30437685	G	A	missense variant
9	62838912	*Cln6	rs232500018	T	C	missense variant

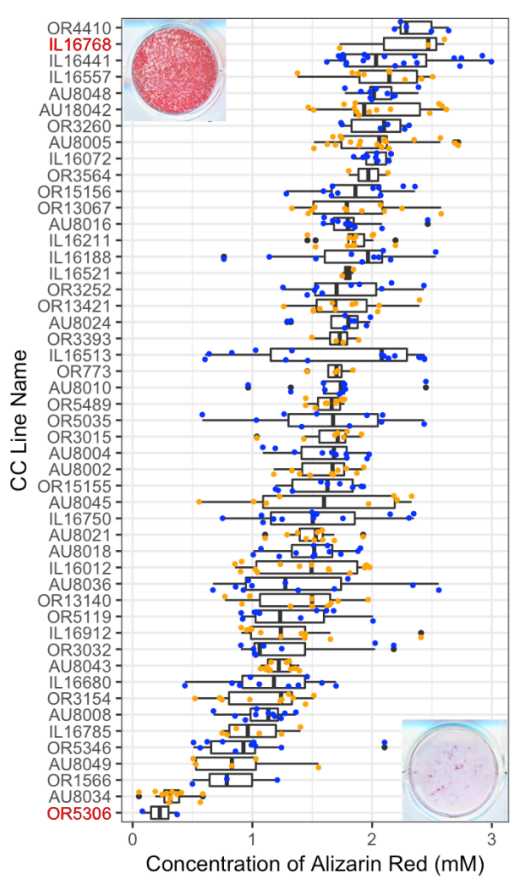
9	62849051	*Cln6	rs233474328	G	A	missense variant
9	63144390	*Skor1	rs215460083	C	T	missense variant
9	63145482	*Skor1	rs252858589	C	T	missense variant



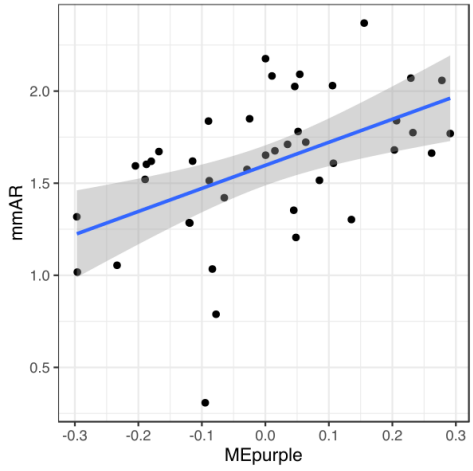
Supplemental Figure 3.1 Single association plot for the BMD GWAS locus. RACER can also be used to create a plot of a single association, for example, this plot of the Chr. 14q32.32 association for BMD.



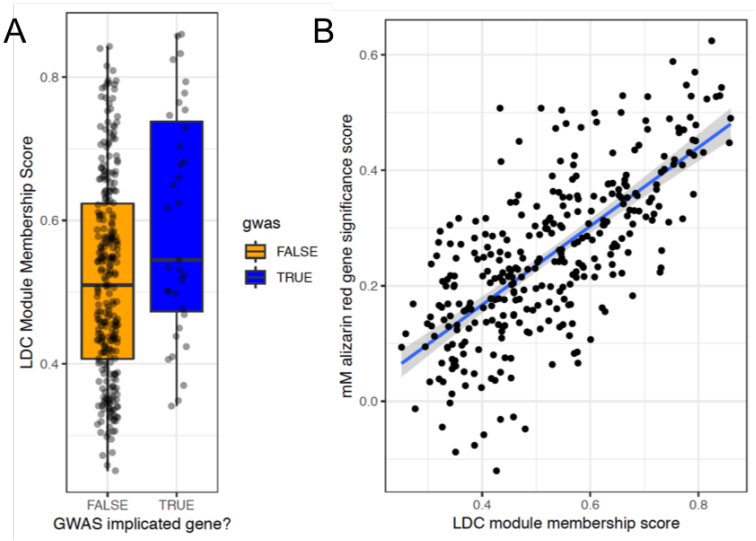
Supplemental Figure 3.2 Scatter plots for MARK3, TRMT61A, and CKB eQTL and a BMD GWAS locus. These scatter plots illustrate the similarity of the BMD association and MARK3 eQTL, and the complex relationship between the CKB eQTL, the TRMT61A eQTL, and the BMD association.



Supplemental Figure 4.1 *In vitro* mineralization varied across the 42 strains of CC mice. Osteoblast cultures were stained ten days post-induction of differentiation with alizarin red stain. Inset images are examples of high mineralizing (strain = IL16768) and low mineralizing (strain = OR5306) strains.



Supplemental Figure 4.2 The purple module eigengene was correlated with *in vitro* mineralization across the population.



Supplemental Figure 4.3 More highly interconnected purple module genes (A) are more likely to overlap GWAS associations and (B) have patterns of expression more highly correlated with in vitro mineralization.

References

1. Kanis, J. A. Diagnosis of osteoporosis. *Osteoporosis International* **7**, 108–116 (1997).
2. Cauley, J. A. Public health impact of osteoporosis. *J. Gerontol. A Biol. Sci. Med. Sci.* **68**, 1243–1251 (2013).
3. Office of the Surgeon General (US). *Bone Health and Osteoporosis: A Report of the Surgeon General*. (2010).
4. 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).
5. Maraka, S. & Kennel, K. A. Bisphosphonates for the prevention and treatment of osteoporosis. *BMJ* **351**, h3783 (2015).
6. Carano, A., Teitelbaum, S. L., Konsek, J. D., Schlesinger, P. H. & Blair, H. C. Bisphosphonates directly inhibit the bone resorption activity of isolated avian osteoclasts in vitro. *J. Clin. Invest.* **85**, 456–461 (1990).
7. Black, D. M. & Rosen, C. J. Postmenopausal Osteoporosis. *N. Engl. J. Med.* **374**, 2096–2097 (2016).
8. Neer, R. M. *et al.* Effect of parathyroid hormone (1-34) on fractures and bone mineral density in postmenopausal women with osteoporosis. *N. Engl. J. Med.* **344**, 1434–1441 (2001).
9. Brommage, R. Genetic Approaches To Identifying Novel Osteoporosis Drug Targets. *J. Cell. Biochem.* **116**, 2139–2145 (2015).
10. Markham, A. Romosozumab: First Global Approval. *Drugs* **79**, 471–476 (2019).
11. Cosman, F. *et al.* Romosozumab Treatment in Postmenopausal Women with Osteoporosis. *N. Engl. J. Med.* **375**, 1532–1543 (2016).

12. Durie, B. G. M., Katz, M. & Crowley, J. Osteonecrosis of the jaw and bisphosphonates. *N. Engl. J. Med.* **353**, 99–102; discussion 99–102 (2005).

13. Schilcher, J., Koeppen, V., Aspenberg, P. & Michaëlsson, K. Risk of Atypical Femoral Fracture during and after Bisphosphonate Use. *New England Journal of Medicine* **371**, 974–976 (2014).

14. Kolata, G. Fearing drugs' rare side effects, millions take their chances with osteoporosis. *NY Times*. (2016).

15. Cummings, S. R. *et al.* Denosumab for prevention of fractures in postmenopausal women with osteoporosis. *N. Engl. J. Med.* **361**, 756–765 (2009).

16. Lewiecki, E. M. Sclerostin: a novel target for intervention in the treatment of osteoporosis. *Discov. Med.* **12**, 263–273 (2011).

17. Ettinger, B. *et al.* Reduction of Vertebral Fracture Risk in Postmenopausal Women With Osteoporosis Treated With Raloxifene: Results From a 3-Year Randomized Clinical Trial. *JAMA* **282**, 637–645 (1999).

18. Greenspan, S. L. *et al.* Effect of recombinant human parathyroid hormone (1-84) on vertebral fracture and bone mineral density in postmenopausal women with osteoporosis: a randomized trial. *Ann. Intern. Med.* **146**, 326–339 (2007).

19. Liberman, U. A. *et al.* Effect of Oral Alendronate on Bone Mineral Density and the Incidence of Fractures in Postmenopausal Osteoporosis. *Obstetrical & Gynecological Survey* **51**, 238–241 (1996).

20. Writing Group for the Women's Health Initiative Investigators. Risks and Benefits of Estrogen Plus Progestin in Healthy Postmenopausal Women: Principal Results From the Women's Health Initiative Randomized Controlled Trial. *JAMA: The Journal of the American Medical Association* **288**, 321–333 (2002).

21. Gauthier, J. Y. *et al.* The discovery of odanacatib (MK-0822), a selective inhibitor of cathepsin K. *Bioorg. Med. Chem. Lett.* **18**, 923–928 (2008).

22. Canalis, E. Update in new anabolic therapies for osteoporosis. *J. Clin. Endocrinol. Metab.* **95**, 1496–1504 (2010).

23. Richards, J. B., Zheng, H.-F. & Spector, T. D. Genetics of osteoporosis from genome-wide association studies: advances and challenges. *Nat. Rev. Genet.* **13**, 576–588 (2012).

24. Rissanen, J. P. & Halleen, J. M. Models and screening assays for drug discovery in osteoporosis. *Expert Opin. Drug Discov.* **5**, 1163–1174 (2010).

25. Lacey, D. L. *et al.* Bench to bedside: elucidation of the OPG–RANK–RANKL pathway and the development of denosumab. *Nat. Rev. Drug Discov.* **11**, 401 (2012).

26. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).

27. McClellan, J. & King, M.-C. Genetic heterogeneity in human disease. *Cell* **141**, 210–217 (2010).

28. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).

29. Lewis, C. M. & Vassos, E. Prospects for using risk scores in polygenic medicine. *Genome Med.* **9**, 96 (2017).

30. Gibson, G. On the utilization of polygenic risk scores for therapeutic targeting. *PLoS Genet.* **15**, e1008060 (2019).

31. Ralston, S. H. & Uitterlinden, A. G. Genetics of osteoporosis. *Endocr. Rev.* **31**, 629–662 (2010).

32. Zheng, H.-F., Spector, T. D. & Richards, J. B. Insights into the genetics of osteoporosis from recent genome-wide association studies. *Expert Rev. Mol. Med.* **13**, e28 (2011).

33. Ralston, S. H. & de Crombrugghe, B. Genetic regulation of bone mass and susceptibility to osteoporosis. *Genes Dev.* **20**, 2492–2506 (2006).

34. Karasik, D. *et al.* Heritability and Genetic Correlations for Bone Microarchitecture: The Framingham Study Families. *J. Bone Miner. Res.* **32**, 106–114 (2017).

35. Estrada, K. *et al.* Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nat. Genet.* **44**, 491–501 (2012).

36. Kemp, J. P. *et al.* Identification of 153 new loci associated with heel bone mineral density and functional involvement of GPC6 in osteoporosis. *Nat. Genet.* **49**, 1468–1475 (2017).

37. Morris, J. A. *et al.* An atlas of genetic influences on osteoporosis in humans and mice. *Nat. Genet.* **51**, 258–266 (2019).

38. Gallagher, M. D. & Chen-Plotkin, A. S. The Post-GWAS Era: From Association to Function. *Am. J. Hum. Genet.* **102**, 717–730 (2018).

39. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).

40. Liu, X., Li, Y. I. & Pritchard, J. K. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell* **177**, 1022–1034.e6 (2019).

41. Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J. & Visscher, P. M. Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model. *Cell* **173**, 1573–1580 (2018).

42. Cox, N. J. Comments on Pritchard Paper. *Journal of Psychiatry and Brain Science* **2**, S5 (2017).

43. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

44. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).

45. Risch, N. & Merikangas, K. The Future of Genetic Studies of Complex Human Diseases. *Science* **273**, 1516–1517 (1996).

46. Shalon, D., Smith, S. J. & Brown, P. O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* **6**, 639–645 (1996).

47. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881–888 (2008).

48. Porcu, E., Sanna, S., Fuchsberger, C. & Fritsche, L. G. Genotype Imputation in Genome-Wide Association Studies. *Current Protocols in Human Genetics* (2013). doi:10.1002/0471142905.hg0125s78

49. Consortium, T. 1000 G. P. & The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).

50. LaFramboise, T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res.* **37**, 4181–4193 (2009).

51. Pearson, T. A. & Manolio, T. A. How to interpret a genome-wide association study. *JAMA* **299**, 1335–1344 (2008).

52. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

53. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).

54. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation

in Regulatory DNA. *Science* **337**, 1190–1195 (2012).

55. Freedman, M. L. *et al.* Principles for the post-GWAS functional characterization of cancer risk loci. *Nature Genetics* **43**, 513–518 (2011).
56. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 9362–9367 (2009).
57. Kilpinen, H. *et al.* Coordinated Effects of Sequence Variation on DNA Binding, Chromatin Structure, and Transcription. *Science* **342**, 744–747 (2013).
58. Darrow, E. M. *et al.* Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E4504–12 (2016).
59. Jakubczik, F. *et al.* A SNP in the Immunoregulatory Molecule CTLA-4 Controls mRNA Splicing In Vivo but Does Not Alter Diabetes Susceptibility in the NOD Mouse. *Diabetes* **65**, 120–128 (2016).
60. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
61. Peterson, T. A. *et al.* Regulatory Single-Nucleotide Variant Predictor Increases Predictive Performance of Functional Regulatory Variants. *Hum. Mutat.* **37**, 1137–1143 (2016).
62. Chen, J. & Tian, W. Explaining the disease phenotype of intergenic SNP through predicted long range regulation. *Nucleic Acids Res.* **44**, 8641–8654 (2016).
63. Pocock, N. A. *et al.* Genetic determinants of bone mass in adults. A twin study. *J. Clin. Invest.* **80**, 706–710 (1987).
64. Krall, E. A. & Dawson-Hughes, B. Heritable and life-style determinants of bone mineral density. *J. Bone Miner. Res.* **8**, 1–9 (1993).

65. Trémollieres, F. A. *et al.* Fracture risk prediction using BMD and clinical risk factors in early postmenopausal women: sensitivity of the WHO FRAX tool. *J. Bone Miner. Res.* **25**, 1002–1009 (2010).

66. Rivadeneira, F. *et al.* Twenty bone-mineral-density loci identified by large-scale meta-analysis of genome-wide association studies. *Nat. Genet.* **41**, 1199–1206 (2009).

67. Brommage, R. *et al.* High-throughput screening of mouse gene knockouts identifies established and novel skeletal phenotypes. *Bone Res* **2**, 14034 (2014).

68. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).

69. Krishnan, V., Bryant, H. U. & Macdougald, O. A. Regulation of bone mass by Wnt signaling. *J. Clin. Invest.* **116**, 1202–1209 (2006).

70. Baron, R. & Kneissel, M. WNT signaling in bone homeostasis and disease: from human mutations to treatments. *Nat. Med.* **19**, 179–192 (2013).

71. Boyce, B. F. & Xing, L. Functions of RANKL/RANK/OPG in bone modeling and remodeling. *Arch. Biochem. Biophys.* **473**, 139–146 (2008).

72. Wittrant, Y. *et al.* RANKL/RANK/OPG: new therapeutic targets in bone tumours and associated osteolysis. *Biochim. Biophys. Acta* **1704**, 49–57 (2004).

73. Nishimura, R. *et al.* Regulation of endochondral ossification by transcription factors. *Front. Biosci.* **17**, 2657–2666 (2012).

74. Van Dijk, F. S. & Sillence, D. O. Osteogenesis imperfecta: clinical diagnosis, nomenclature and severity assessment. *Am. J. Med. Genet. A* **164A**, 1470–1481 (2014).

75. Laine, C. M. *et al.* WNT1 mutations in early-onset osteoporosis and osteogenesis imperfecta. *N. Engl. J. Med.* **368**, 1809–1816 (2013).

76. Styrkarsdottir, U. *et al.* Nonsense mutation in the LGR4 gene is associated with several

human diseases and other traits. *Nature* **497**, 517–520 (2013).

77. Styrkarsdottir, U. *et al.* Two Rare Mutations in the COL1A2 Gene Associate With Low Bone Mineral Density and Fractures in Iceland. *J. Bone Miner. Res.* **31**, 173–179 (2016).
78. Zheng, H.-F. *et al.* Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. *Nature* **526**, 112–117 (2015).
79. Jepsen, K. J. Functional interactions among morphologic and tissue quality traits define bone quality. *Clin. Orthop. Relat. Res.* **469**, 2150–2159 (2011).
80. Paternoster, L. *et al.* Genetic determinants of trabecular and cortical volumetric bone mineral densities and bone microstructure. *PLoS Genet.* **9**, e1003247 (2013).
81. Levy, R., Mott, R. F., Iraqi, F. A. & Gabet, Y. Collaborative cross mice in a genetic association study reveal new candidate genes for bone microarchitecture. *BMC Genomics* **16**, 1013 (2015).
82. Guo, Y. *et al.* Genome-wide association study identifies ALDH7A1 as a novel susceptibility gene for osteoporosis. *PLoS Genet.* **6**, e1000806 (2010).
83. Hwang, J.-Y. *et al.* Meta-analysis identifies a MECOM gene as a novel predisposing factor of osteoporotic fracture. *J. Med. Genet.* **50**, 212–219 (2013).
84. Rosen, C. J., Beamer, W. G. & Donahue, L. R. Defining the genetics of osteoporosis: using the mouse to understand man. *Osteoporos. Int.* **12**, 803–810 (2001).
85. Brommage, R. & Ohlsson, C. Translational studies provide insights for the etiology and treatment of cortical bone osteoporosis. *Best Pract. Res. Clin. Endocrinol. Metab.* **32**, 329–340 (2018).
86. Markel, P. *et al.* Theoretical and empirical issues for marker-assisted breeding of congenic mouse strains. *Nat. Genet.* **17**, 280–284 (1997).
87. Turner, C. H. *et al.* Improved bone strength in low bone density C57BL/6J mice

carrying donated QTLs from C3H/HeJ mice. in *Bone* **28**, S85–S85 (2001).

88. Klein, R. F., Mitchell, S. R., Phillips, T. J., Belknap, J. K. & Orwoll, E. S. Quantitative trait loci affecting peak bone mineral density in mice. *J. Bone Miner. Res.* **13**, 1648–1656 (1998).

89. Beamer, W. G. *et al.* Quantitative Trait Loci for Femoral and Lumbar Vertebral Bone Mineral Density in C57BL/6J and C3H/HeJ Inbred Strains of Mice. *Journal of Bone and Mineral Research* **20**, 1700–1712 (2005).

90. Shimizu, M. *et al.* Identification of peak bone mass QTL in a spontaneously osteoporotic mouse strain. *Mamm. Genome* **10**, 81–87 (1999).

91. Bouxsein, M. L. *et al.* Chromosomal location of genes that contribute to vertebral trabecular bone density and microarchitecture in mice. in *Bone* **28**, S72–S73 (2001).

92. Freudenthal, B. *et al.* Rapid phenotyping of knockout mice to identify genetic determinants of bone strength. *J. Endocrinol.* **231**, R31–46 (2016).

93. Törn, C. *et al.* Complement gene variants in relation to autoantibodies to beta cell specific antigens and type 1 diabetes in the TEDDY Study. *Sci. Rep.* **6**, 27887 (2016).

94. Kirby, A. *et al.* Fine mapping in 94 inbred mouse strains using a high-density haplotype resource. *Genetics* **185**, 1081–1095 (2010).

95. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).

96. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **Chapter 7**, Unit7.20 (2013).

97. Sloan, C. A. *et al.* ENCODE data at the ENCODE portal. *Nucleic Acids Res.* **44**, D726–32 (2016).

98. Chadwick, L. H. The NIH Roadmap Epigenomics Program data resource. *Epigenomics* **4**, 317–324 (2012).

99. Chung, D., Yang, C., Li, C., Gelernter, J. & Zhao, H. GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. *PLoS Genet.* **10**, e1004787 (2014).

100. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res.* **22**, 1748–1759 (2012).

101. Hardison, R. C. Genome-wide epigenetic data facilitate understanding of disease susceptibility association studies. *J. Biol. Chem.* **287**, 30932–30940 (2012).

102. Spain, S. L. & Barrett, J. C. Strategies for fine-mapping complex traits. *Hum. Mol. Genet.* **24**, R111–9 (2015).

103. Stephens, M. & Balding, D. J. Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* **10**, 681–690 (2009).

104. Wellcome Trust Case Control Consortium *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44**, 1294–1301 (2012).

105. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38**, 626–635 (2006).

106. Farber, C. R. & Lusis, A. J. Integrating global gene expression analysis and genetics. *Adv. Genet.* **60**, 571–601 (2008).

107. Rockman, M. V. & Kruglyak, L. Genetics of global gene expression. *Nat. Rev. Genet.* **7**, 862–872 (2006).

108. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* **16**, 197–212 (2015).

109. Melé, M., Ferreira, P. G., Reverter, F. & DeLuca, D. S. The human transcriptome across tissues and individuals. (2015).

110. Jia, P. & Zhao, Z. Network.assisted analysis to prioritize GWAS results: principles, methods and perspectives. *Hum. Genet.* **133**, 125–138 (2014).

111. Leiserson, M. D. M., Eldridge, J. V., Ramachandran, S. & Raphael, B. J. Network analysis of GWAS data. *Curr. Opin. Genet. Dev.* **23**, 602–610 (2013).

112. Califano, A., Butte, A. J., Friend, S., Ideker, T. & Schadt, E. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nat. Genet.* **44**, 841–847 (2012).

113. Farber, C. R. Systems-level analysis of genome-wide association data. *G3* **3**, 119–129 (2013).

114. Farber, C. R. Identification of a gene module associated with BMD through the integration of network analysis and genome-wide association data. *J. Bone Miner. Res.* **25**, 2359–2367 (2010).

115. Gustafsson, M. *et al.* A validated gene regulatory network and GWAS identifies early regulators of T cell–associated diseases. *Sci. Transl. Med.* **7**, 313ra178–313ra178 (2015).

116. Huan, T. *et al.* Integrative network analysis reveals molecular mechanisms of blood pressure regulation. *Mol. Syst. Biol.* **11**, 799 (2015).

117. Mäkinen, V.-P. *et al.* Integrative genomics reveals novel molecular pathways and gene networks for coronary artery disease. *PLoS Genet.* **10**, e1004502 (2014).

118. Horvath, S. & Dong, J. Geometric interpretation of gene coexpression network analysis. *PLoS Comput. Biol.* **4**, e1000117 (2008).

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

120. Horvath, S. *Weighted Network Analysis: Applications in Genomics and Systems Biology*. (Springer Science & Business Media, 2011).

121. Horvath, S. *et al.* Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 17402–17407 (2006).

122. Fabregat, A. *et al.* The Reactome pathway Knowledgebase. *Nucleic Acids Res.* **44**, D481–7 (2016).

123. Nishimura, D. BioCarta. *Biotech Software & Internet Report* **2**, 117–120 (2001).

124. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

125. Preuss, M. *et al.* Design of the Coronary ARtery DIsease Genome-Wide Replication And Meta-Analysis (CARDIoGRAM) Study. *Circulation: Cardiovascular Genetics* **3**, 475–483 (2010).

126. Zhu, J. *et al.* Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLoS Comput. Biol.* **3**, e69 (2007).

127. Zhu, J. *et al.* Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat. Genet.* **40**, 854–861 (2008).

128. Gupta, P. Translating Mouse Systems Genetics to Discovery in Human Disease. *Dissertation.* (UCLA, 2017).

129. Graham, J. B. *et al.* Extensive Homeostatic T Cell Phenotypic Variation within the Collaborative Cross. *Cell Rep.* **21**, 2313–2325 (2017).

130. Ferris, M. T. *et al.* Modeling host genetic regulation of influenza pathogenesis in the collaborative cross. *PLoS Pathog.* **9**, e1003196 (2013).

131. 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).

132. Gass, M. & Dawson-Hughes, B. Preventing osteoporosis-related fractures: an overview. *Am. J. Med.* **119**, S3–S11 (2006).

133. Bouxsein, M. L. & Karasik, D. Bone geometry and skeletal fragility. *Curr. Osteoporos. Rep.* **4**, 49–56 (2006).

134. Ng, A. H. M., Wang, S. X., Turner, C. H., Beamer, W. G. & Grynpas, M. D. Bone quality and bone strength in BXH recombinant inbred mice. *Calcif. Tissue Int.* **81**, 215–223 (2007).

135. Norgard, E. A. *et al.* Identification of quantitative trait loci affecting murine long bone length in a two-generation intercross of LG/J and SM/J Mice. *J. Bone Miner. Res.* **23**, 887–895 (2008).

136. Corva, P. M., Horvat, S. & Medrano, J. F. Quantitative trait loci affecting growth in high growth (hg) mice. *Mamm. Genome* **12**, 284–290 (2001).

137. Horvat, S. & Medrano, J. F. Lack of Socs2 expression causes the high-growth phenotype in mice. *Genomics* **72**, 209–212 (2001).

138. Metcalf, D. *et al.* Gigantism in mice lacking suppressor of cytokine signalling-2. *Nature* **405**, 1069–1073 (2000).

139. Farber, C. R., Corva, P. M. & Medrano, J. F. Genome-wide isolation of growth and obesity QTL using mouse speed congenic strains. *BMC Genomics* **7**, 102 (2006).

140. Farber, C. R. *et al.* Genetic dissection of a major mouse obesity QTL (Carfhg2): integration of gene expression and causality modeling. *Physiol. Genomics* **37**, 294–302 (2009).

141. Keane, T. M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289–294 (2011).

142. Soranzo, N. *et al.* Meta-analysis of genome-wide scans for human adult stature identifies novel Loci and associations with measures of skeletal frame size. *PLoS Genet.* **5**, e1000445 (2009).

143. Stefanovic, L., Longo, L., Zhang, Y. & Stefanovic, B. Characterization of binding of LARP6 to the 5' stem-loop of collagen mRNAs: implications for synthesis of type I collagen. *RNA Biol.* **11**, 1386–1401 (2014).

144. Zhang, Y. & Stefanovic, B. LARP6 Meets Collagen mRNA: Specific Regulation of Type I Collagen Expression. *Int. J. Mol. Sci.* **17**, 419 (2016).

145. Young, M. F. Bone matrix proteins: their function, regulation, and relationship to osteoporosis. *Osteoporos. Int.* **14 Suppl 3**, S35–42 (2003).

146. Cai, L., Fritz, D., Stefanovic, L. & Stefanovic, B. Binding of LARP6 to the Conserved 5' Stem–Loop Regulates Translation of mRNAs Encoding Type I Collagen. *J. Mol. Biol.* **395**, 309–326 (2010).

147. Richter, S. *et al.* Expression and role in glycolysis of human ADP-dependent glucokinase. *Mol. Cell. Biochem.* **364**, 131–145 (2012).

148. Skarnes, W. C. *et al.* A conditional knockout resource for the genome-wide study of mouse gene function. *Nature* **474**, 337–342 (2011).

149. Lui, J. C. *et al.* EZH1 and EZH2 promote skeletal growth by repressing inhibitors of chondrocyte proliferation and hypertrophy. *Nat. Commun.* **7**, 13685 (2016).

150. Horvat, S. & Medrano, J. F. Interval mapping of high growth (hg), a major locus that increases weight gain in mice. *Genetics* **139**, 1737–1748 (1995).

151. R Core Team. R: A language and environment for statistical computing. (2013).

152. Broman, K. W., Wu, H., Sen, S. & Churchill, G. A. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**, 889–890 (2003).

153. Choi, K. B. g2gtools 0.1.29 usage. *g2gtools 0.1.29 documentation* Available at: https://g2gtools.readthedocs.io/en/latest/. (Accessed: 24th October 2016)

154. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).

155. Choi, K. B. EMASE 0.10.16 usage. *EMASE 0.10.16 documentation* Available at: http://emase.readthedocs.io/en/latest/. (Accessed: 14th November 2016)

156. Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6**, e1000888 (2010).

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

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

159. Hormozdiari, F. *et al.* Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nat. Genet.* **50**, 1041–1047 (2018).

160. Nica, A. C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* **6**, e1000895 (2010).

161. Sieber, K. *PICCOLO*. https://github.com/Ksieber/piccolo (Github).

162. Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).

163. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555–3557 (2015).

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

165. Black, D. M. & Rosen, C. J. Clinical Practice. Postmenopausal Osteoporosis. *N. Engl. J. Med.* **374**, 254–262 (2016).

166. Johnell, O. & Kanis, J. A. An estimate of the worldwide prevalence and disability associated with osteoporotic fractures. *Osteoporos. Int.* **17**, 1726–1733 (2006).

167. Sabik, O. L. & Farber, C. R. Using GWAS to identify novel therapeutic targets for osteoporosis. *Transl. Res.* (2016). doi:10.1016/j.trsl.2016.10.009

168. Morris, J. A. *et al.* An Atlas of Human and Murine Genetic Influences on Osteoporosis. *bioRxiv* 338863 (2018). doi:10.1101/338863

169. Boyle, E. A., Li, Y. & Pritchard, J. K. The omnigenic model: Response from the authors. *J. Psychiatry Brain Sci.* **2**, S8 (2017).

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

171. Rau, C. D. *et al.* Systems Genetics Approach Identifies Gene Pathways and Adamts2 as Drivers of Isoproterenol-Induced Cardiac Hypertrophy and Cardiomyopathy in Mice. *Cell Syst* **4**, 121–128.e4 (2017).

172. Kogelman, L. J. A. *et al.* Identification of co-expression gene networks, regulatory genes and pathways for obesity based on adipose tissue RNA Sequencing in a porcine model. *BMC Med. Genomics* **7**, 57 (2014).

173. Eising, E. *et al.* Gene co-expression analysis identifies brain regions and cell types involved in migraine pathophysiology: a GWAS-based study using the Allen Human Brain Atlas. *Hum. Genet.* **135**, 425–439 (2016).

174. Churchill, G. A. *et al.* The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nat. Genet.* **36**, 1133–1137 (2004).

175. 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).

176. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
177. Bolser, D. Mouse Genome Informatics (MGI, Mouse Genome Database, MGD). *Dictionary of Bioinformatics and Computational Biology* (2004). doi:10.1002/9780471650126.dob1002
178. Koscielny, G. *et al.* The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. *Nucleic Acids Res.* **42**, D802–9 (2014).
179. Dyment, N. A. *et al.* High-Throughput, Multi-Image Cryohistology of Mineralized Tissues. *J. Vis. Exp.* (2016). doi:10.3791/54468
180. Johnson, M. L. How rare bone diseases have informed our knowledge of complex diseases. *Bonekey Rep* **5**, 839 (2016).
181. Boudin, E. & Van Hul, W. MECHANISMS IN ENDOCRINOLOGY: Genetics of human bone formation. *Eur. J. Endocrinol.* **177**, R69–R83 (2017).
182. Robinson, M.-E. & Rauch, F. Mendelian bone fragility disorders. *Bone* (2019).
183. Marini, F. & Brandi, M. L. Genetic determinants of osteoporosis: common bases to cardiovascular diseases? *Int. J. Hypertens.* **2010**, (2010).
184. Rocha-Braz, M. G. M. & Ferraz-de-Souza, B. Genetics of osteoporosis: searching for candidate genes for bone fragility. *Arch Endocrinol Metab* **60**, 391–401 (2016).
185. Komori, T. Regulation of Osteoblast Differentiation by Runx2. *Advances in Experimental Medicine and Biology* 43–49 (2009).
186. Yoshida, C. A. *et al.* SP7 inhibits osteoblast differentiation at a late stage in mice. *PLoS*

One **7**, e32364 (2012).

187. Semënov, M., Tamai, K. & He, X. SOST is a ligand for LRP5/LRP6 and a Wnt signaling inhibitor. *J. Biol. Chem.* **280**, 26770–26775 (2005).

188. Atkins, G. J. *et al.* Sclerostin is a locally acting regulator of late-osteoblast/preosteocyte differentiation and regulates mineralization through a MEPE-ASARM-dependent mechanism. *Journal of Bone and Mineral Research* **26**, 1425–1436 (2011).

189. Nakamura, A. *et al.* Osteocalcin secretion as an early marker of in vitro osteogenic differentiation of rat mesenchymal stem cells. *Tissue Eng. Part C Methods* **15**, 169–180 (2009).

190. Golub, E. E. & Boesze-Battaglia, K. The role of alkaline phosphatase in mineralization. *Curr. Opin. Orthop.* **18**, 444 (2007).

191. Lee, H.-J. *et al.* Association of a RUNX2 promoter polymorphism with bone mineral density in postmenopausal Korean women. *Calcif. Tissue Int.* **84**, 439–445 (2009).

192. Koller, D. L. *et al.* Meta-analysis of genome-wide studies identifies WNT16 and ESR1 SNPs associated with bone mineral density in premenopausal women. *J. Bone Miner. Res.* **28**, 547–558 (2013).

193. Mencej-Bedrač, S., Preželj, J., Kocjan, T., Komadina, R. & Marc, J. Analysis of Association of LRP5, LRP6, SOST, DKK1, and CTNNB1 Genes with Bone Mineral Density in a Slovenian Population. *Calcif. Tissue Int.* **85**, 501–506 (2009).

194. Wakayama, T. & Iseki, S. Role of the spermatogenic–Sertoli cell interaction through cell adhesion molecule-1 (CADM1) in spermatogenesis. *Anat. Sci. Int.* **84**, 112–121 (2009).

195. Zhang, W. *et al.* CADM1 regulates the G1/S transition and represses tumorigenicity through the Rb-E2F pathway in hepatocellular carcinoma. *Hepatobiliary Pancreat. Dis. Int*

15, 289–296 (2016).

196. Cao, W., Shi, P. & Ge, J.-J. miR-21 enhances cardiac fibrotic remodeling and fibroblast proliferation via CADM1/STAT3 pathway. *BMC Cardiovasc. Disord.* **17**, 88 (2017).
197. Nakamura, S. *et al.* Negative feedback loop of bone resorption by NFATc1-dependent induction of Cadm1. *PLoS One* **12**, e0175632 (2017).
198. Inoue, T. *et al.* Cell adhesion molecule 1 is a new osteoblastic cell adhesion molecule and a diagnostic marker for osteosarcoma. *Life Sci.* **92**, 91–99 (2013).
199. Mentink, A. *et al.* Predicting the therapeutic efficacy of MSC in bone tissue engineering using the molecular marker CADM1. *Biomaterials* **34**, 4592–4601 (2013).
200. Sato, T. *et al.* Molecular Cloning and Characterization of a Novel Human β1,4-N-Acetylgalactosaminyltransferase, β4GalNAc-T3, Responsible for the Synthesis ofN,N'-Diacetyllactosediamine, GalNAcβ1–4GlcNAc. *Journal of Biological Chemistry* **278**, 47534–47544 (2003).
201. Ikehara, Y. *et al.* Apical Golgi localization of N,N'-diacetyllactosediamine synthase, β4GalNAc-T3, is responsible for LacdiNAc expression on gastric mucosa. *Glycobiology* **16**, 777–785 (2006).
202. Sasaki, N., Shinomi, M., Hirano, K., Ui-Tei, K. & Nishihara, S. LacdiNAc (GalNAcβ1-4GlcNAc) contributes to self-renewal of mouse embryonic stem cells by regulating leukemia inhibitory factor/STAT3 signaling. *Stem Cells* **29**, 641–650 (2011).
203. Zheng, J. *et al.* Genome-wide mapping identifies beta-1,4-N-acetyl-galactosaminyl-transferase as a novel determinant of sclerostin levels and bone mineral density. *bioRxiv* 455386 (2018).
204. Meller, N., Irani-Tehrani, M., Kiosses, W. B., Del Pozo, M. A. & Schwartz, M. A. Zizimin1, a novel Cdc42 activator, reveals a new GEF domain for Rho proteins. *Nat.*

Cell Biol. **4**, 639–647 (2002).

205. Park, S. J., Lee, J. Y., Lee, S. H., Koh, J.-M. & Kim, B.-J. SLIT2 inhibits osteoclastogenesis and bone resorption by suppression of Cdc42 activity. *Biochem. Biophys. Res. Commun.* **514**, 868–874 (2019).

206. Aizawa, R. *et al.* Cdc42 regulates cranial suture morphogenesis and ossification. *Biochem. Biophys. Res. Commun.* **512**, 145–149 (2019).

207. Bohnekamp, J. & Schöneberg, T. Cell adhesion receptor GPR133 couples to Gs protein. *J. Biol. Chem.* **286**, 41912–41916 (2011).

208. Wang, Y., Li, X. & Hu, H. H3K4me2 reliably defines transcription factor binding regions in different cells. *Genomics* **103**, 222–228 (2014).

209. Pekowska, A., Benoukraf, T., Ferrier, P. & Spicuglia, S. A unique H3K4me2 profile marks tissue-specific gene regulation. *Genome Res.* **20**, 1493–1502 (2010).

210. Pradeepa, M. M. Causal role of histone acetylations in enhancer function. *Transcription* **8**, 40–47 (2017).

211. Calo, E. & Wysocka, J. Modification of enhancer chromatin: what, how, and why? *Mol. Cell* **49**, 825–837 (2013).

212. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 21931–21936 (2010).

213. Local, A. *et al.* Identification of H3K4me1-associated proteins at mammalian enhancers. *Nat. Genet.* **50**, 73–82 (2018).

214. Russ, B. E. *et al.* Regulation of H3K4me3 at Transcriptional Enhancers Characterizes Acquisition of Virus-Specific CD8+ T Cell-Lineage-Specific Function. *Cell Rep.* **21**, 3624–3636 (2017).

215. Nicetto, D. & Zaret, K. S. Role of H3K9me3 heterochromatin in cell identity

establishment and maintenance. *Curr. Opin. Genet. Dev.* **55**, 1–10 (2019).

216. Becker, J. S., Nicetto, D. & Zaret, K. S. H3K9me3-Dependent Heterochromatin: Barrier to Cell Fate Changes. *Trends Genet.* **32**, 29–41 (2016).
217. Gabellini, N., Bortoluzzi, S., Danieli, G. A. & Carafoli, E. The human SLC8A3 gene and the tissue-specific Na+/Ca2+ exchanger 3 isoforms. *Gene* **298**, 1–7 (2002).
218. Zhang, L. *et al.* Multistage genome-wide association meta-analyses identified two new loci for bone mineral density. *Hum. Mol. Genet.* **23**, 1923–1933 (2014).
219. Babraham Bioinformatics. FastQC: a quality control tool for high throughput sequence data. *Cambridge, UK: Babraham Institute* (2011).
220. Raghupathy, N. *et al.* Hierarchical analysis of RNA-seq reads improves the accuracy of allele-specific expression. *Bioinformatics* **34**, 2177–2184 (2018).
221. 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).
222. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
223. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
224. Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* **37**, W305–11 (2009).
225. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
226. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res.* **30**, 38–41

(2002).

227. Sabik, O. L. & Farber, C. R. RACER: A data visualization strategy for exploring multiple genetic associations. *bioRxiv* 495366 (2018). doi:10.1101/495366

228. Kurbatova, N., Mason, J. C., Morgan, H., Meehan, T. F. & Karp, N. A. PhenStat: A Tool Kit for Standardized Analysis of High Throughput Phenotypic Data. *PLoS One* **10**, e0131274 (2015).

229. Grundberg, E. *et al.* Population genomics in a disease targeted primary cell model. *Genome Res.* **19**, 1942–1952 (2009).

230. Grundberg, E. *et al.* Global analysis of the impact of environmental perturbation on cis-regulation of gene expression. *PLoS Genet.* **7**, e1001279 (2011).

231. Choi, H. J. *et al.* Genome-wide association study in East Asians suggests UHMK1 as a novel bone mineral density susceptibility gene. *Bone* **91**, 113–121 (2016).

232. Nielson, C. M. *et al.* Novel Genetic Variants Associated With Increased Vertebral Volumetric BMD, Reduced Vertebral Fracture Risk, and Increased Expression of SLC1A3 and EPHB2. *J. Bone Miner. Res.* **31**, 2085–2097 (2016).

233. Shaffer, J. R., Kammerer, C. M., Bruder, J. M., Bauer, R. L. & Mitchell, B. D. Different Genes Contribute to Variation in Peak Bone Density and Bone Loss. American Society for Bone and Mineral Research Abstract (2014).

234. Kemp, J. P., Medina-Gomez, C., Tobias, J. H., Rivadeneira, F. & Evans, D. M. The case for genome-wide association studies of bone acquisition in paediatric and adolescent populations. *Bonekey Rep* **5**, 796 (2016).

235. Chesi, A. *et al.* A trans-ethnic genome-wide association study identifies gender-specific loci influencing pediatric aBMD and BMC at the distal radius. *Hum. Mol. Genet.* **24**, 5053–5059 (2015).

236. Cho, Y. S. *et al.* A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nature Genetics* **41**, 527–534 (2009).

237. Knight, J. C. Approaches for establishing the function of regulatory genetic variants involved in disease. *Genome Med.* **6**, 92 (2014).

238. Dailey, L. High throughput technologies for the functional discovery of mammalian enhancers: new approaches for understanding transcriptional regulatory network dynamics. *Genomics* **106**, 151–158 (2015).

239. Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Research* **27**, 38–52 (2017).

240. Soldner, F. *et al.* Parkinson-associated risk variant in distal enhancer of α-synuclein modulates target gene expression. *Nature* **533**, 95–99 (2016).

241. Bilousova, G. *et al.* Osteoblasts derived from induced pluripotent stem cells form calcified structures in scaffolds both in vitro and in vivo. *Stem Cells* **29**, 206–216 (2011).

242. Jeon, O. H. *et al.* Human iPSC-derived osteoblasts and osteoclasts together promote bone regeneration in 3D biomaterials. *Sci. Rep.* **6**, 26761 (2016).