Leveraging GWAS and the Transcriptome to Identify Potential Causal Genes in Osteoporosis

Abdullah "Arby" Abood

Baghdad بغداد, Iraq العراق

Associate of Science in Science, Northern Virginia Community College, Annandale, Virginia, 2011

Bachelor of Science in Biology, George Mason University, Fairfax, Virginia, 2013

Master of Science in Microbiology, Clemson University, Clemson, South Carolina, 2018

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

Center for Public Health Genomics

University of Virginia

May 2023

Dr. Charles Farber (Mentor)
Dr. Stefan Bekiranov (Committee head)
Dr. Gloria Sheynkman (Member)
Dr. Clint Miller (Member)
Dr. Stephen Turner (Member)
Dr. Alan Bergland (Member)

Table of Contents

Abstract

Osteoporosis is a prevalent bone disease that poses a significant health problem for millions of individuals globally. Genome-wide association studies (GWASs) have identified numerous associations that affect bone mineral density (BMD), the most reliable predictor of osteoporosis fracture. Further efforts are being made to identify the genes responsible for the effects of these associations. Most of these associations impact bone by altering gene regulation. In my dissertation work, I used innovative, unbiased approaches to prioritize previously identified genetic associations from humans. In the first chapter, I discuss how molecular "-omics" data and state-of-the-art analytical techniques are being employed to facilitate gene discovery from GWAS and provide meaning to these studies. I highlight the resources required in the bone field and novel approaches that I used in my graduate work, along with their potential for improvement in the coming years. In the second chapter, I focus on identifying potentially causal long non-coding RNAs (lncRNAs), which are understudied non-coding RNAs in the context of bone and osteoporosis. I identified 23 lncRNAs that may play a causal role in osteoporosis and are candidates for experimental follow-up studies. In the third chapter, I used long-read proteogenomics to identify potentially causal protein-coding isoforms in osteoporosis. I provide a list of potentially causal isoforms and validated *TPM2* functionally *in vitro*. Finally, I share my final thoughts on the current state of the field and future directions for the next generation of systems geneticists who seek to provide treatment for osteoporosis. Ultimately, my dissertation contributes to our comprehension of the genetic architecture of osteoporosis-related traits and presented new approaches for following up on GWAS studies.

Dedication

Above all, I dedicate this work to my wife, **Caroline**. I am grateful for her unwavering support and encouragement, which have made her my number one fan.

I would also like to express my deepest gratitude to my mentor, Dr. **Charles Farber**, who saw my potential despite my past failures. His never-ending support and guidance have shaped me into the scientist I am today and turned me into a better person. I am committed to paying forward all the valuable lessons he has taught me.

To my committee members, I am grateful for the time you dedicated to listening to me ramble away. **Stefan**, I appreciate your wisdom and answers to all my questions. **Clint**, thank you for inviting me to "Miller Time" and welcoming me as an honorary member of your lab. **Gloria**, thank you for teaching me project management skills. **Stephen**, I am proud to have recruited you to my committee as my outsider source to the secrets of science in "industry." Your experience has been invaluable to me. **Alan**, while you were a later addition to the party, I appreciate your contribution to my research; you have undoubtedly helped enhance my training and ability to advance science.

To my friends and family, I appreciate your continuous support and kindness during stressful times. Your patience during both good and bad times has made life easier.

To those who have made it all the way here, I encourage you to prioritize your mental health. Seek therapy, avoid dwelling on the past, and fight the stigma around mental illness.

Lastly, I dedicate this work to all the refugees worldwide. Life may be challenging, but it will always get better.
#refugee2PhD

List of Figures and Tables

List of Abbreviations

AI	Allelic imbalance
ALP	Alkaline phosphatase
AS	Alternative splicing
ASE	Allele-specific expression
ASMC	Airway smooth muscle cell
ASTS	Allele-specific transcript structure
ATACseq	Transposase-accessible chromatin with high-throughput sequencing
ATCC	American Type Culture Collection
B6	C57BL/6J mice
BMC	Boston medical center
BMD	Bone mineral density
Bp	Base pair
BUMC	Boston university medical campus
CAD	Coronary artery disease
caQTL	Chromatin accessibility quantitative trait locus
CAST	CAST/EiJ
CC	Collaborative cross
Chr	Chromosome
COPD	Chronic obstructive pulmonary disease
CPAT	Coding potential assessment tool
DE	Differential expression
DEXA	Dual-energy x-ray absorptiometry
DIU	Differential isoform usage
DLPFC	Dorsolateral prefrontal cortex
DO	Diversity outbred
eBMD	Estimated bone mineral density
ENCODE	ENCyclopedia Of DNA Elements
eQTL	Expression quantitative trait locus

FBS	Fetal bovine serum
FDR	False discovery rate
FN	Femoral neck
FSM	Full-splice match
GEFOS	GEnetic Factors for OSteoporosis
glm	Generalized linear model
GRN	Gene regulatory network
GTEx	Genotype-Tissue Expression
GWAS	Genome-wide association study
h^2	Narrow sense heritability
hFOB	Human fetal osteoblast
HMDP	Hybrid mouse diversity panel
Hob	Human osteoblasts
IMPC	International mouse phenotyping consortium
IRB	Institutional review board
ISM	Incomplete-splice match
Kbp	Kilo base pair
KDA	Key driver analysis
KOMP	Knockout mouse phenotyping consortium
LD	Linkage disequilibrium
lncRNA	Long non-coding RNA
LRP	Long-read proteogenomics
LRS	Long-read sequencing
LS	Lumbar spine
MAF	Minor allele frequency
Mbp	Mega base pair
meQTL	Methylation quantitative trait locus
MGI	Mouse genome informatics
MMnet	Macrophage multinucleation network
mRNA	Messenger RNA

MSC	Mesenchymal stem cell
NGS	Next generation sequencing
NIC	Novel in catalog
NMD	Nonsense mediated decay
NNC	Novel not in catalog
NS	Non-synonymous
OBCD	Origins of bone and cartilage disease
OFM	Osteoblast functional model
ONT	Oxford Nanopore
OR	Odds ratio
ORF	Open reading frame
PacBio	Pacific Biosciences
PBMC	Peripheral blood mononuclear cell
PIRB	Protein Isoform Resource for BMD
PPH4	Posterior probability of hypothesis 4
proQTL	protein quantitative trait loci
PSI	Percent spliced in
QTL	Quantitative trait locus
RACER	Regional Association ComparER
RCP	Regional colocalization probability
rMATS	Rapid Multivariate Analysis of Transcript Splicing
RNA	Ribonucleic acid
RNAseq	RNA sequencing
RRM	RNA recognition motifs
rRNA	Ribosomal RNA
sceQTLGen	Single-cell eQTL generation consortium
scRNAseq	Single-cell RNA sequencing
SF	Splice factor
siRNA	Short interfering RNA
SNP	Single nucleotide polymorphism

sQTL	Splicing quantitative trait locus
SRR	Serine racemase
T2D	Type 2 diabetes
TB	Total body
TF	Transcription factor
TFBS	Transcription factor binding sites
TPM	Transcripts per million
TPM2	Tropomyosin 2
GO	Gene ontology
TWAS	Transcriptome-wide association study
UTR	Untranslated region
WGCNA	Weighted gene co-expression network analysis
WKY	Wistar Kyoto

Chapter 1
Introduction

Published in:

Abood A and Farber CR. Using "-omics" Data to Inform Genome-wide Association Studies (GWASs) in the Osteoporosis Field. *Curr Osteoporos Rep*. 2021 Aug;19(4):369-380. doi: 10.1007/s11914-021-00684-w. 2021 Jun 14. PMID: 34125409; PMCID: PMC8767463.

Castaldi PJ, **Abood A**, Farber CR, Sheynkman GM. Bridging the splicing gap in human genetics with long-read RNA sequencing: finding the protein isoform drivers of disease. *Hum Mol Genet*. 2022 Aug 12:ddac196. doi: 10.1093/hmg/ddac196. PMID: 35960994.

1.1 Overview of the genetics of bone mineral density and osteoporosis

Osteoporosis is characterized by low bone mineral density (BMD) and deteriorated bone microarchitecture which leads to an increased risk of fracture [1,2]. In the USA, over 12 million individuals have been diagnosed with osteoporosis, leading to over 2 million fractures per year, a number expected to nearly double by 2025 [3]. Importantly, of the ~300,000 people that suffer from a hip fracture annually, 1 in 5 will die in the subsequent 12 months [4]. Osteoporotic fractures are also costly accounting for approximately $26 billion in health-care expenditures [3].

BMD is one of the strongest predictors of fracture [5]. It is also a highly heritable quantitative trait ($h^2 = 0.5$–0.8) that can be measured in large cohorts of individuals [6–9]. Although genome-wide association studies (GWASs) have become the mainstay of investigations into the genetic basis of BMD, genetic studies of bone traits began before the GWAS era [10]. Prior to GWASs, genetic studies of BMD and osteoporosis involved linkage in families and candidate gene association studies [9]. Linkage studies identified several loci for BMD; however, with notable exceptions (see [11] as an example), the challenges of gene discovery in the context of linkage studies limited their utility for unraveling complex traits such as BMD. Additionally, replication of loci identified by linkage has been low [12]. Similarly, candidate gene studies identified several associations for BMD, few of which have been replicated in large cohorts [13,14].

Most GWASs conducted for osteoporosis have focused on BMD. BMD can be measured using dual-energy X-ray absorptiometry (DEXA) or quantitative ultrasound (generates measures of estimated BMD (eBMD)). The largest GWAS for DEXA-derived lumbar spine and femoral neck BMD was performed on ~80K individuals and identified 56 loci [15,16]. The largest eBMD GWAS performed to date used the UK Biobank (N~420K) and identified 501 loci harboring 1,103 independent associations which explain 20.3% of the total variance of the trait [13].

Although GWASs have revolutionized the identification of BMD loci, few of the underlying causal genes have been identified. This is largely due to the fact that, unlike Mendelian disease, >90% of GWAS loci for common diseases are due to non-coding variants [17]. This suggests that most associations are caused by changes in gene regulation [18]. As a result, it is possible that variants in a GWAS locus may regulate a gene a considerable distance (100s of Kbps) up- or downstream. Additionally, extensive linkage disequilibrium (LD) adds to the difficulty in assigning target genes to loci and identifying the underlying causal variant(s) at each locus [19]. Together, these challenges have made it difficult to pinpoint causal genes highlighting the importance of developing novel approaches to inform BMD GWASs [20].

There are three primary reasons why causal gene discovery is important. First, the identification of new genes responsible for variation in BMD from GWAS [21,22] has already shed light on important new processes impacting bone [23,24]. This will only continue and increase in impact as the approaches we discuss below are utilized more widely to interrogate BMD GWAS. Second, the hopes of precision medicine for osteoporosis, which aims to tailor therapeutics based on individualized risk factors (i.e., an individual's genotype), rely on a comprehensive understanding of the genes impacting bone. Third, and possibly the most important, GWAS is a powerful approach to identify antiosteoporotic therapeutic targets. Historically, many drug targets from traditional studies have failed in clinical trials [25]. There are many reasons for these failures including that targets of investigation are often not causally linked to a disease. Recently, it has been shown that drug targets with evidence from genetic studies (including GWAS) are twice as likely to succeed in clinical trials [25,26]. Together, these factors are driving the focus on causal gene discovery.

Throughout this introduction, I will highlight how molecular "-omics" data and cutting-edge analytical approaches are being used to facilitate gene discovery from GWAS. My aim is to highlight specific studies that demonstrate how "-omics" data and analytical approaches can be used to "make sense" of GWAS. I also discuss resources that are needed in the bone field and novel approaches that will be used in the coming years.

1.2 Approaches for Causal Gene Discovery

Since the first GWAS for any disease in 2006 [27], several approaches have been developed with the goal of identifying causal genes. All of these approaches leverage the generation of molecular data and its analysis using genetic- and/or systems genetics–based analytical strategies. The concept is simple; genetic variants mediate their effects on a phenotype by altering molecular changes, such as differences in gene expression, alternative splicing, intron retention, protein levels, protein activities, and molecular interactions. As a result, the identification of disease-associated variants that influence molecular phenotypes allows one to identify causal genes and begin to unravel their mechanisms of action.

The advent of next-generation sequencing (NGS) has made profiling many of these changes straightforward and feasible to do in large human populations. This has led to a revolution in the use of molecular quantitative trait locus (QTL) data which are beginning to help us understand how BMD-associated variants impact molecular processes and, in turn, how these changes influence BMD and ultimately risk of fracture. The next logical step is to identify the causal genes in order to enhance our understanding of disease mechanisms. At the heart of this step are the approaches to correctly annotate causal genes and understand how they influence bone. This section will address current concepts related to establishing the cause-and-effect relationship between BMD GWAS reported genetic variants and their respective functional genes in bone.

1.2.1 Expression Quantitative Trait Loci (eQTL) Colocalization

One of the most widely used approaches to inform GWAS is through identification and colocalization of expression quantitative trait loci (eQTL) [28]. An eQTL is an association between a set of genetic variants and gene expression levels [28] (**Figures 1.1 and 1.2)**. eQTLs are divided into two categories based on their proximity to target genes: local (also referred to as cis) and distant (also referred to as trans). Local eQTLs are located in close proximity (typically defined as ± 1 Mbp) to the gene they regulate [28]. An example of a local eQTL would be a polymorphism in the promoter of a gene that leads to altered transcription factor binding and allele-specific expression. In contrast, distant eQTLs are located far from the genes they regulate and are often on different chromosomes [28]. A distant eQTL could manifest from a polymorphic transcription factor that influences its target genes differently based on its genotype. The first step in the identification of eQTLs consists of collecting and profiling the transcriptome of disease-relevant tissues or cell types using RNA-seq in a population of densely genotyped individuals. These data are then used to identify eQTLs by conducting association tests between millions of single-nucleotide polymorphisms (SNPs) and thousands of genes. The most direct and common way that local eQTLs are used to inform GWAS is through colocalization [29,30]. Colocalization is a set of statistical approaches that test the hypothesis that an eQTL and GWAS association (or any two associations) are driven by the same shared variant (**Table 1.1, Figure 1.2**). Essentially, colocalization is testing whether or not BMD-associated variants also influence gene expression. If so, then one can hypothesize that the BMD-associated variants influence gene expression and the change in gene expression alters BMD.

Several studies have demonstrated that many eQTLs are tissue or cell type specific, likely reflecting the cell type–specific nature of the epigenome [31,32]. As a result, eQTLs identified in disease-relevant tissues or cell types are likely to be the most informative for use in GWAS colocalization [28,31]. One of the largest projects for eQTL identification and analysis is the Genotype-Tissue Expression (GTEx) project [31]. GTEx is an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation. The project has profiled ~50 tissues in hundreds of individuals using RNA-seq and identified thousands of eQTLs [31,32]. In fact, GTEx has identified significant local eQTL for nearly 95% of all protein-coding genes in the human genome [32]. While GTEx has significantly increased our understanding of eQTLs and how they mediate the effects of GWAS, one of its limitations for the bone field is that bone or bone cells are not included in the 49 tissues investigated. Fortunately, studies are beginning to be conducted that identify eQTLs in bone and bone cells [33].

Bone tissue is primarily made up of three major cell types: osteoblasts, cells that form bone; osteoclasts, cells that break-down bone; and osteocytes, cells that coordinate the function of both osteoblasts and osteoclasts. A dynamic equilibrium between these cell types ensures that the skeleton is being properly maintained through bone remodeling. Ideally, we would have eQTL data (and other molecular data types) on all three important cell types.

The first and only osteoblast eQTL dataset was generated in 2009 from primary cultured human osteoblasts (HOb) derived from 95 unrelated donors of Swedish origin [34]. Global gene expression from these cells were profiled using microarray technology [34]. The authors then identified local eQTLs and used them to prioritize the genomic loci from one of the first BMD GWAS studies [35]. They identified serine racemase (SRR) as a novel BMD-associated gene [35]. Since its publication, this dataset has also been used by groups performing BMD GWAS to help identify causal genes [16,36,37]. However, its sample size, low-density genotyping, and use of microarray technology to profile expression have limited its effectiveness for gene discovery. The second population scale transcriptomic dataset was generated in 2010 on iliac crest bone biopsies from 84 postmenopausal women in Norway [38]. This study suffers the sample limitations as described for the study above, but it has been used by several groups performing GWAS to provide insight into potentially causal genes [16,39].

In 2018, Mullin et al. [33] generated a RNA-seq–based eQTL dataset using osteoclast-like cells differentiated in vitro from peripheral blood mononuclear cells (PBMCs) obtained from 158 female patients. These data were used to identify genes with eQTLs that colocalized with loci from two eBMD GWASs [13,37]. The authors used coloc [29], a widely used colocalization approach that tests whether association signals are driven by the same causal variant. In the first study [33], using 307 BMD GWAS significant SNPs, eight genes were reported to have a significantly colocalizing eQTL. In the second study [33], using 1,103 significant GWAS SNPs, evidence of colocalization of GWAS and eQTL association signals was identified for 21 genes. The low percentage of

GWAS loci with a colocalizing eQTL in osteoclasts may reflect the cross-sectional nature of the GWAS with differences in BMD being driven primarily by bone formation.

Studies support the use of eQTL data in aiding the interpretation of GWAS results in other disease fields such as Crohn's disease [40], bipolar disease [41], and diabetes [42]. Here we highlight the transcription factor *KLF14* and its role in type 2 diabetes (T2D) [43]. The genetic variants associated with T2D and other metabolic phenotypes map to a region of 3–48 kb upstream of *KLF14* [43]. The GWAS SNPs associated with the *KLF14* are colocalized with eQTLs only in adipose tissue despite *KLF14* being expressed in multiple tissues. Small et al. [43] showed that these SNPs act in adipose tissue to reduce *KLF14* expression and modulate, in trans, expression of 385 genes. The study also demonstrated the mechanism in which *KLF14* expression increases pre-adipocyte proliferation but disrupts lipogenesis [43]. Additionally, *in vivo* knockout in adipose tissue in mice partially recapitulated the human phenotype of insulin resistance, dyslipidemia, and T2D [43]. This is an excellent example of how eQTL data can inform GWAS and how such findings could similarly be used in the bone field, especially once large-scale bone-relevant datasets have been generated.

Most eQTL studies focus on "total" gene expression that is transcript levels summed over the exons of a gene. However, genetic variation can impact all aspects of transcriptional and posttranscriptional regulation [44]. For example, recent studies have identified splicing QTL (sQTL), which are loci influencing mRNA splicing [31] (**Figure 1.1**). No sQTL studies have been conducted in bone or bone cells; however, this approach has been used in other tissues [45]. For example, 8,966 sQTL were identified using dorsolateral prefrontal cortex (DLPFC) RNA-seq data from >200 individuals [46]. When they compared sQTL SNPs and GWAS SNPs (an approach similar to colocalization, but statistically less stringent), a significant overlap was observed for schizophrenia and other diseases, suggesting that part of the genetic risk for complex diseases is due to sQTL. Therefore, to facilitate comprehensive gene discovery, future eQTL studies in bone should address how BMD-associated variants impact all levels of gene regulation.

1.2.2 Approaches involving Splicing Quantitative Trait Loci (sQTL)

One of the themes of the evolution of multicellular organisms is the development of managed complexity, with alternative splicing serving as a prime example. As genes are partitioned into exons separated by intronic sequences of increasing length and sequence content, the complexity of gene protein products increases tremendously, with splicing serving as the control mechanism for this combinatorial protein complexity. Naturally occurring genetic variants can perturb this splicing control mechanism, and it is now clear that an appreciable proportion of GWAS loci contain variants that alter splicing [32,44,47,48].

GWAS has provided a wealth of novel insights into human biology, identifying genetic variants in over 55,000 loci associated with 5,000 complex traits and diseases [49]. In addition to GWAS, association studies across multiple tissues and cell types have mapped expression quantitative trait loci (eQTLs), whereby common genetic variants alter total gene expression levels [50]. Colocalization studies combining GWAS and eQTL results have successfully identified functional mechanisms for many GWAS loci [51], indicating that the functional of these GWAS loci is to alter gene regulation. However, the function of most GWAS loci remains uncharacterized, suggesting that other important regulatory mechanisms are involved (and that eQTL discovery is incomplete). It is increasingly clear that one of these other mechanisms is genetically influenced alternative splicing, measured by splicing quantitative trait loci (sQTLs, **Figure 1.3**), making the study of alternative splicing a priority for GWAS functional characterization of these loci [45,52,53].

The first observations of splicing in human immunoglobulin genes were made decades ago [54,55], but it was through the use of RNA-seq that the nearly ubiquitous nature of splicing in the human transcriptome was demonstrated [56]. Alternative splicing (AS) of mRNA molecules to produce distinct isoforms is a mechanism of gene regulation inherent to nearly every protein-coding gene (92–94%) [56,57]. Specific splicing events arise from the interplay of core splice factors, which are mandatory for splicing, and auxiliary splice factors, which regulate splicing [58,59] to form the 'splicing network' [45,60].

Aberrant splicing leads to a host of pathologies, from neurodegeneration to cancer [61–63]. Genetic variation can affect splice factors, their target binding sites or other regulatory elements to disrupt the balance of the splicing network. Splice-altering genetic variation is consistent with other quantitative traits whereby the effect size of variants on splicing is inversely correlated with their minor allele frequency [64]. Rare variants tend to have more dramatic effects on protein function, such as mis-splicing that leads to a truncated protein. On the other hand, common variants. contribute to complex genetic diseases through a continuum of effects on splicing, from dramatic loss of multiple exons to subtle shifts of splicing ratios [32,48]. Functional genomics approaches to map and functionally characterize sQTLs are rapidly advancing, and, as I discuss below, are ripe for integration with emerging long-read RNA sequencing approaches (**Chapter 3**). Past excellent reviews cover the topic of genetically regulated alternative splicing, with a focus on insights derivable from the short-read RNA-seq data available at the time [45,52,53]. My goal is to provide comments on the state of the field in terms of the study of splicing in the context of complex human disease (GWASs) with a focus on long-read sequencing.

1.2.2a Methods for sQTL discovery: central role of short-read RNA-seq-based splicing quantification

The earliest sQTL studies made creative use of exon arrays to identify sQTLs by comparing genetic effects on exon and gene level expression [65–67]; however, RNA-seq has revolutionized the identification of sQTLs, largely due to the fact that RNA-seq provides direct measurement of splicing through junctional reads [68,69]. A review of recent studies using sQTLs to functionally characterize GWAS loci may be found in **Table 1.2**. Previous reviews give an excellent overview of sQTL studies before 2018 [45,53].

In many cases, sQTL detection methods utilize the same regression-based software programs used for eQTL detection, such as MatrixeQTL [70], FastQTL [71], tensorQTL [72], and EMMAX [73], though other approaches such as transcriptome-wide association studies (TWAS) have been developed [74,75]. For eQTL studies, quantification of gene expression is fairly straightforward, but splicing differs from eQTLs in that alternative splicing is typically expressed in ratios in which the numerator is the count of a particular splice event, such as inclusion of an exon, and the denominator is the sum of the counts of all other linked splicing events. The most common metric used to quantify splice events is Ψ, 'percent spliced in' (PSI), which represents the rate at which a genic feature (such as an exon) is included in mature RNA transcripts. Thus, the foundation for most sQTL analyses depends on the concept of a splicing event which leads naturally to an 'event-based' approach to splicing quantification.

A wide variety of splice quantification methods have been developed [76–79]. Most event-based approaches take a 'local' approach in which the numerator and denominator values used to compute Ψ are calculated from directly observed short-read quantities, such as junctional reads or exon counts. Two widely used programs, rMATS [80] and LeafCutter [69], provide useful illustration of the central challenge in the event-based approach, namely the lack of a clear correct approach to calculating the denominator for Ψ. This denominator is meant to capture the set of splicing alternatives that is relevant for any given splicing event, but splicing often shows a complex pattern of dependency in which splice events in different parts of the gene body have a strong pattern of co-occurrence [81,82]. This complexity can make it challenging to unambiguously define the set of linked events that should constitute the denominator for calculating the Ψ of any specific event, as described in detail in the MISO publication [76]. rMATS addresses this issue by limiting its analysis to five well-defined classes of splicing events (exon skipping, retained intron, etc.), but it does not account for more complex splicing patterns. The txRevise approach addresses more complex splicing, but still limits analysis to events occurring in known isoforms (i.e. present in reference databases) [83]. In contrast, LeafCutter uses a data-driven approach to identify novel splicing events and to 'learn' patterns of splicing that may be quite complex, but this can come at the cost of shifting definitions of splicing events across or even within datasets. In addition, as Leafcutter relies solely on junction-spanning reads, it is unable to quantify changes in gene coverage such as intron retention or changes in UTRs. Overall, given the insufficiency of a single event-based approach to capture all possible transcript variations, selection of a tool represents the choice of which splicing features are most prioritized for sensitive and accurate detection.

An alternative to local, event-based quantification are isoform-based quantification methods, in which the abundances of full-length transcripts are first estimated from short reads, an approach that is used by tools such as kallisto [84], RSEM [85], Salmon [86], and StringTie [87]. A common approach is to calculate isoform ratios (count of one isoform/total isoform counts for the gene) which can be tested for association with genetic variants in transcript ratio QTL (trQTL) analyses, of which there are many examples [88,89]. Detection of changes in isoform usage is a multivariate problem, and methods like sQTLSeekeR/R2 [48], DRIMSeq [90], and THISTLE [91] implement statistical models that specifically account for the multivariate nature of isoform analysis. The advantage of isoform-based analysis is that by explicitly representing isoforms, which encode and are defined by a specific series of splice events, greater clarity can be achieved in the characterization of potentially complex splicing changes. The main drawback is the underlying inaccuracy of isoform estimation [92]. Even with state-of-the-art isoform inference methods, accuracy varies by expression level [93], is reduced for genes with many exons and a large number of expressed isoforms [94] and may lead to reduced performance to detect differential isoform expression between conditions [95].

Since neither event-based nor isoform inference-based approaches can fully recover missing information about the true isoform ratios, this leads to appreciable variability in splicing quantification [92] which also produces variability in the results from different sQTL-calling algorithms. A recent sQTL study using both event-based (Leafcutter) and the isoform-based (THISTLE) approach found the two methods produced overlapping but complementary results [91]. Another study found differences between sQTLSeekR and Leafcutter in the GTEx dataset [48].

1.2.2b Long-read RNA-seq to interpret genetically regulated splicing

By providing direct identification and quantification of full-length transcript isoforms, long-read sequencing—from technologies such as Pacific Biosciences (PacBio) and Oxford Nanopore (ONT)—can improve sQTL interpretation to provide a more direct link between genetic variants and their impact on transcript abundances [82,96–102].

One obvious way that long-read sequencing informs sQTL interpretation is to reveal effects of sQTLs on novel isoforms that would be undetectable or misinterpreted by analyses dependent entirely on reference transcriptome annotation (**Figure 1.3**). A significant number of sQTLs discovered through short-read RNA-seq are associated with novel junctions or exons [69]. These events are difficult to interpret, because they cannot be conclusively linked to a reference transcript. Long-read RNA-seq data from human cells or tissues routinely uncover tens of thousands of novel isoforms, indicating that human transcript isoform annotations are incomplete, estimated to only include roughly 33% of true isoforms [103–106].

Another way long-read sequencing improves sQTL interpretation is through resolution of sQTL effects on complex splicing phenomena, which we refer to here as simpler splicing events that tend to co-occur, such as distant exon inclusion events that occur within the same isoform [81,82]. In many cases, the identification of 'local' sQTL events is not sufficient to map the effect of a genetic variant to a specific isoform and then to its downstream effect on protein function. By providing accurate information on the isoform content within each sample, long-read sequencing can clarify genetic effects on novel splicing events and complex patterns of interrelated splicing. For example, even when all junctions are present in the reference, novel isoforms often arise from new combinations of known splicing events (i.e. junctions, exons), which would require long-read sequencing to resolve [103,107]. Though allele-specific expression of particular splicing events can be detected using short-read data [108,109], the use of long reads can reveal allele-specific expression of entirely full-length isoforms, even revealing variants that result in dramatic changes in isoform length [110–112]. Recently, tools have been developed to process long-read sequence reads to extract both splicing and allele information to trace the parental origin of isoforms [111–113].

Two illustrative examples of long-read sequencing to clarify genetically influenced splicing are the chronic obstructive pulmonary disease (COPD) and lung function GWAS association in *NPNT* and the body fat percentage GWAS association in *DUSP13*. In the case of *NPNT*, a genome-wide comparison of COPD GWAS peaks and leafcutter sQTLs from GTEx lung tissue identified *NPNT* as a locus harboring nearly identical genetic association signals for COPD and alternative splicing of multiple exons in *NPNT*. The A allele of the lead SNP rs34712979 introduces a novel splice acceptor site at the second exon, creating a NAGNAG motif. Analysis of short-read RNA-seq confirmed that the proximal acceptor site created by the A allele is strongly preferred to the canonical site. However, this variant also has unexplained sQTL associations with splicing in the second, third and fourth exons on *NPNT*, an observation that had no clear explanation in light of the reference isoforms. Targeted long-read sequencing in lung tissue from 10 subjects selected by rs34712979 genotype revealed the presence of multiple novel, truncated *NPNT* isoforms which were highly expressed. There were marked genotype-specific differences in the usage of these novel short isoforms that account for its pattern of sQTL associations [114]. In the case of *DUSP13*, Bayesian colocalization analysis between body fat percentage GWAS and sQTLs identified in muscle tissue from GTEx implicated three sQTL intron excision events [111]. Long-read sequencing coupled with allele-specific transcript structure (ASTS) analysis using LORALS [111] showed transcript ENST00000372700 (*DUSP13-202*) lacking four middle exons was more highly expressed from the risk (ALT) allele.

1.2.2c Using long-read RNA-seq to understand how genetic variants affect protein isoform functions through splicing

While the major functional consequence of eQTLs is to change RNA and protein expression levels, sQTLs can alter both the expression and sequence content of the resulting proteins. Using long-read data, one can predict the full-length encoded protein isoform product associated with an sQTL [115,116], bridging the gap between genetic variants and their functional consequences on proteins. Though protein QTLs may be measured through aptamer or mass spectrometry-based approaches, these modalities do not provide isoform-level protein quantification [117–119].

Knowledge of long-read-predicted protein isoforms opens up new possibilities for interpreting and prioritizing the function of disease-associated sQTLs [120,121]. Bioinformatic and experimental approaches can be used to relate protein isoform changes to protein functional changes. For example, isoforms associated with sQTLs can be bioinformatically analyzed to determine how splicing changes lead to disruption or modulation of protein functional features, such as structural domains [122–124] or other protein functional features [125,126]. Other approaches leverage isoform-specific expression correlation to derive isoform-specific networks [127,128], or to propagate gene-level annotations to the most likely functional isoform [129–131]. Knowledge of the predicted protein isoforms can also be a valuable guide to design experiments for functional validation [132], which can include high-throughput phenotypic screening of isoforms [133,134], and isoform-specific assays such as protein–protein interaction profiling [135–137].

To understand the molecular basis of sQTL associations, at the heart is the need to quantify the functional differences between alternative protein isoforms associated with sQTL genetic variants (i.e. the genetic ‘risk’ isoform). There are two possibilities here: (1) the alternative isoform has reduced stability or molecular activity, relative to the wild-type isoform, or (2) the alternative isoform is capable of a different set of molecular activities, relative to the wild-type isoform. Once the pairwise isoform functional effects are defined, one should then consider the cumulative protein functional capacity of the gene, which is directly computable from the collective quantities and functional activities of the individual protein isoforms. This cumulative gene-level protein output could make conceptualization of sQTL-effects more tractable in systems-scale analysis. A full description of various eQTL and sQTL relationships to protein consequences may be found in **Figure 1.4**.

1.2.3 Transcriptome-wide association studies (TWASs)

In recent years, transcriptome-wide association studies (TWASs) have become widely used approaches that utilize gene expression data to measure the association between genetically regulated gene expression and complex phenotypes [138] (**Table 1.1**). In an individual, gene expression (and as noted above, other aspects of gene regulation) is influenced by genetics and the environment [139,140]. The genetic contribution to gene expression can be quantified using eQTL and used to predict or impute expression in an individual based on genotype. For example, if a local eQTL explains 100% (no environmental contribution in this hypothetical example) of the variance in gene x, then all we need to know is an individual's eQTL genotype to know the expression of gene x in that individual. TWAS extends this example by estimating the genetic component of gene expression (using the advanced statistical analysis of eQTL data) across the genome in a reference population where gene expression and genotype have been measured (GTEx is an example) and then imputing (predicting) gene expression in a much larger population (such as those used in a BMD GWAS) [138]. Once gene expression is imputed, genetically regulated gene expression is associated with a disease or disease phenotype. Most genes identified by TWAS are located in GWAS associations for that disease (due to genetically regulated differences in gene expression being the basis of most GWAS associations). As a result, TWAS can pinpoint genes likely to be causal at GWAS loci.

TWASs for BMD are sparse but are starting to be performed. In one of the first studies, gene expression data from GTEx muscle and whole blood tissues in combination with the largest eBMD GWAS to date [141] identified 276 genes with significant gene-trait associations. To further pinpoint causal genes, the authors used colocalization to demonstrate that 142 of the 276 showed strong evidence for colocalization using GTEx data. Of the 142 genes, many were well-known regulators of BMD. Another study utilized 48 GTEx tissues and reported 88 significant genes, many of which were located in total body (TB) BMD GWAS [142]. Lastly, a recently published resource, PhenomeXcan [143], integrated TWAS gene- trait associations with colocalization to prioritize GWAS loci. A total of 675 genes were identified with both significant TWAS associations with BMD and colocalizing eQTLs.

1.2.4 Network Analysis

As mentioned above, GWASs have identified thousands of associations for BMD; however, the scarcity of population scale human RNA-seq datasets on bone or bone cells has hindered our ability to directly inform BMD GWAS. To address this limitation, it has recently been demonstrated that network approaches using transcriptomic (and in some cases other "-omic" data types) data can be used to provide information on which genes at a GWAS locus might be causal [144,145]. The general idea is simple; genes responsible for GWAS associations likely function in pathways that impact bone, such as osteoblast-mediated bone formation or osteoclast-mediated bone resorption. In turn, biological network reconstruction approaches can take molecular data and group genes, in an unbiased way, into groups (or pathways) based loosely on function (**Table 1.1**). As a result, it is possible to use networks to identify "network modules" that are enriched in GWAS genes and likely represent key pathways or biological processes regulating BMD. In other words, biological networks provide us with cellular wiring diagrams and GWAS points to "circuits" that when disrupted (by genetic variation) lead to disease. We can then use the knowledge of key circuits to inform GWAS.

In a series of studies, our lab has used co-expression networks to predict causal BMD GWAS genes. In a co-expression network, genes are connected based on the correlation of their expression [146,147]. Groups or "modules" of highly intercorrelated genes are identified by clustering [146,147], and modules have been shown in a number of studies across species to loosely group genes based on functional similarities [148,149]. Calabrese et al. [145] generated a co-expression network using transcriptomic data on marrow-free cortical bone from the Hybrid Mouse Diversity Panel (HMDP; a panel of 96 inbred mouse strains) [150]. We used mouse data to profile "pure" cortical bone since these data were not available in humans. To identify network modules that represented key "circuits," we mapped the mouse homologues of genes implicated by (i.e., located in a GWAS locus) the largest BMD GWAS at the time [16] onto the mouse bone network. We identified two modules with a significant enrichment of GWAS genes, collectively named the Osteoblast Functional Module (OFM). Based on a detailed characterization, we hypothesized that many of the OFM genes were causal GWAS genes and that they likely influenced BMD via a role in modulating osteoblast activity. The OFM allowed us to predict and infer the function of causal genes for 30 of 64 reported BMD GWAS loci. We further investigated two BMD loci on chromosomes 2p16.2 and 14q32.32. Based on the network analysis, we predicted that *Sptbn1* and *Mark3* were responsible for the effects of the two loci, respectively. In support of these predictions, we used GTEx to identify that both genes were regulated by eQTLs in multiple tissues that colocalized with their respective GWAS associations. We also showed for both genes that BMD was altered in mouse knockout models in the same direction predicted by the GWAS/eQTL data.

There have been two follow-up studies [144,151] that have refined the above approach. Sabik et al. [144] generated a cell type–specific (osteoblast) and time point-specific (mineralization) co-expression network using RNA-seq data on calvarial osteoblasts from a separate panel of inbred mouse strains. We identified a co-expression module enriched for genes implicated by BMD GWAS, correlated with in vitro osteoblast mineralization, and associated with skeletal phenotypes in human monogenic disease and mouse knockouts. We further investigated four loci and found that *Cadm1*, *B4galnt3*, *Dock9*, and *Gpr133* all had human colocalizing eQTL and altered BMD in knockout mice.

Our next refinement was the use of Bayesian networks. Bayesian networks differ from co-expression networks in that they use advanced statistical approaches to add directions to the network that allow one to infer causal relationships between genes [152]. Al-Barghouthi et al. [151] demonstrated the utility of this approach by generating a co-expression network from cortical bone RNA-seq data collected from outbred mice. We then created Bayesian networks for each co-expression module and performed what is called a "key driver analysis" (KDA) [151,153]. In a KDA, genes that are known to play important roles in bone (such as *RUNX2*, a transcription factor essential for osteoblastogenesis) are "seeded" onto the network. For each gene in the network, we then counted the number of "known" genes it was connected to and determined if this number was more than would be expected by chance. We then identified key drivers which were located in a GWAS locus and regulated by a colocalizing eQTL. This approach yielded 46 genes likely to be causal for human BMD GWAS associations. We further investigated two novel genes, *Sertad4* and *Glt8d2*, and demonstrated that BMD was altered in knockouts, further suggesting they were causal for their respective GWAS association. These data supported the idea that Bayesian networks provided a new perspective and approach to identify causal BMD GWAS genes.

Another study informed GWAS using a co-expression network generated on macrophages from a cross between Wistar Kyoto (WKY) and LEW rats [154]. MMnet (the macrophage multinucleation network) was a module in this network. Importantly, macrophages can fuse to form osteoclasts in bone, and WKY rats experience spontaneous macrophage fusing events, while LEW rats do not [155]. MMnet contained 190 genes regulated by trans eQTLs that were driven by the *Trem2* gene [155]. The authors reported that WKY rats show low bone mass, mineralization, and strength relative to LEW rats, suggesting that MMnet genes may control bone homeostasis. The authors show that the MMnet was enriched for genes located in BMD GWAS loci [154]. Given the evidence, the authors conducted in vivo knockout experiments of *Bcat1*, the most highly connected gene in the network, and showed that *Bcat1* deficiency results in high bone mass. Next, readily available knockout mice of 12 MMnet individual genes were obtained, half of which showed skeletal phenotype abnormalities. As the authors established a strong association between MMnet and osteoclast activity, they investigated whether the human orthologues share the same activity as their rat counterparts. Knockdown of 11 MMnet genes (three of which are GWAS hits) in vitro (TRAP+ human osteoclasts) and knockout of the same genes in vivo (mouse models) were concordant and strongly correlated. This study identified a physiologically important network that is highly conserved in rats, mice, and humans and was enriched in GWAS-implicated genes. This is a great representation of how network analysis can inform GWAS and provide important information on the function (in this case osteoclast multinucleation) of potentially causal genes.

1.2.5 Non-transcriptomics based data: Epigenomics

Another "-omics" data type that has proven useful for informing GWASs is epigenetic data [156]. The majority (>90%) of BMD GWAS loci are found within noncoding regions [17], indicating that they perturb gene regulation. Thus, it is likely that causal GWAS variants reside in regulatory elements, such as promoters, enhancers, and CpG islands, which can be identified using epigenomic data.

In the same way that eQTL can be identified for gene expression, QTL can be identified using epigenomic data [157]. Examples are chromatin accessibility QTLs (caQTLs), which are loci that influence chromatin accessibility [158] (**Figure 1.1**). Chromatin accessibility is a measure of the usage and activity of regulatory elements and is often assayed using ATAC- seq [159]. caQTLs can highlight potentially causal SNPs that may be driving genetically regulated changes in expression. In this way, they are most useful for identifying causal variants, not causal genes. However, they can also be integrated with eQTL information to link caQTLs to the genes they regulate, increasing confidence for a particular set of putatively causal genes [160].

To our knowledge, there are no studies using caQTLs to inform BMD GWAS. However, in T2D, a study profiled chromatin accessibility in pancreatic islet samples from 19 genotyped individuals and identified 2,949 caQTLs [161]. The authors performed a functional follow-up on 13 of the reported caQTLs using luciferase reporter assays in MIN6 β-cells and showed that more than half exhibited effects on enhancer activity that were consistent with in vivo chromatin accessibility changes. Importantly, islet caQTL analysis nominated putative causal SNPs in 13 T2D-associated GWAS loci, linking seven and six T2D risk alleles, respectively, to gain or loss of *in vivo* chromatin accessibility.

Lastly, accumulating evidence suggests that genetic variants may impact a complex disease by modulating DNA methylation levels (meQTLs) (**Figure 1.1**). To date, no meQTL analysis has been performed on bone samples. However, a study in depression cohorts has implicated gene targets by testing associations between SNPs and DNA methylation levels in whole blood [162]. Another study used meQTL to inform GWAS of asthma in exaggerated bronchoconstriction of airway smooth muscle cells (ASMCs) and showed that GWAS variants in asthma were significantly enriched for meQTLs [163].

1.3 Integrating "-omics" Data Types

So far, we have highlighted the use of single "-omic" data types. However, the use of multidimensional datasets (layering of various datasets) increases statistical power [164], potentially provides stronger evidence for causality, and captures biology that would not have been informed with any one modality. In a recent study, Qiu et al. [165] performed multi-"omics" analyses with expression, methylation, and metabolite QTLs to identify osteoporosis biomarkers. Their approach involved performing individual transcriptomic, methylomic, and metabolomic analysis in 119 European female subjects with high (n = 61) and low (n = 58) BMD. Using advanced statistical approaches, the authors identified gene-based biomarkers, some of which corresponded to genes located in BMD GWAS loci, suggesting they are causal.

Recently, Chesi et al. [166] took a different approach. Instead of generating population-level "-omics" data, the authors broadly profiled multiple "-omics" layers in human mesenchymal stem cell (MSC)–derived osteoblasts.

Their approach was anchored on promoter Capture-C [167]. Promoter Capture-C is a technique that uses promoter "baits" to "fish-out" interactions between promoters and the rest of the genome. These data provide links between regulatory elements (e.g., enhancers) and the promoters of their target genes that are presumably important for gene expression. In osteoblasts, they identified interactions that were in close proximity of BMD-associated variants identified by GWAS. They then used RNA- seq data to confirm target gene expression in osteoblasts as well as ATAC-seq data to confirm the region interacting with the promoter was a region of open chromatin and presumably an active regulatory element. They fine-mapped 273 BMD GWAS loci in primary osteoblasts. The authors report observing consistent contacts between candidate causal variants and putative target gene promoters in open chromatin for ~ 17% of the 273 BMD loci investigated. Knockdown of two novel implicated genes, *ING3* and *EPDR1*, inhibited osteoblastogenesis, while promoting adipogenesis.

In this introduction, I discuss the many ways that "-omics" data can be used to identify genes responsible for the effects of BMD GWAS loci. One limitation, as discussed above, is the paucity of "-omics" data on bone tissue and bone cells. While the number of such studies is growing, there is a need to generate population-scale transcriptomics data on the three main cell types in bone: osteoblasts, osteoclasts, and osteocytes. These data would significantly increase our ability to identify and characterize causal genes responsible for BMD GWAS associations. They would also be of significant use to the larger bone and human genetics communities to address other aspects of disease.

Another exciting approach that will impact our ability to use gene expression data to inform GWAS is single-cell RNA sequencing (scRNA-seq). scRNA-seq is emerging as a powerful tool to examine transcriptomes of individual cells. The clear benefit of this technology is its ability, when used in populations, to identify cell type–specific eQTLs, many of which are lost when using bulk methods that take into account average gene expression across all different cell subtypes [168]. In early 2020, the single-cell eQTLGen consortium (sceQTLGen) was founded, aimed at pinpointing the disease-causing genetic variants and their effect on gene expression [168]. The single-cell toolbox can be extended beyond the transcriptome to the epigenome with single-cell ATAC-seq (scATAC-seq). A recent study by Rai et al. [169] attempted to identify cell type–specific regulatory signatures underlying T2D in pancreatic islets. They reported that T2D GWAS SNPs were significantly enriched in the open chromatin of beta cells, but not in alpha or delta cell–specific open chromatin. In the bone field, scRNA-seq and scATAC-seq are just beginning to be performed, but have already demonstrated the extensive cellular heterogeneity of bone [170,171]. Using post-GWAS approaches described by this introduction on data generated by these two approaches will lead to an increase in our ability to inform GWAS.

Proteomic analysis offers another type of data that could be integrated into systems genetics approaches. Protein quantification studies have shown that transcript abundance is not highly correlated with protein translation [172,173]. Generally, proteomics technologies used in publications to study bone metabolism can be inherently divided into two categories: (i) expression screening and (ii) quantitative mass spectroscopy (MS) [173,174]. The main challenge facing proteomic work in bone is the efficient extraction of proteins from bone cells [175]. Therefore, most studies have instead used blood serum/plasma or PBMCs to study cellular signaling, secretory proteins, and differential protein expression between conditions [174,176]. However, the goal of GWAS follow-up at the proteome level is identifying genetic variants associated with protein concentrations (protein quantitative trait loci; proQTL) [177] (**Figure 1.1**). Integrating proQTL with GWAS variants using approaches such as colocalization may inform GWAS beyond what can be accomplished with transcriptomics data and bridge the knowledge gap regarding SNP-disease associations [177]. For example, using GWAS data from Framingham participants (long-term cardiovascular study cohort) reported 13 proteins harboring proQTL variants that match coronary disease-risk variants from GWAS, not all of which also had colocalizing eQTLs [178].

It should be noted that all the approaches discussed in this review provide hypotheses that must be tested in order to confirm gene discovery. It is impossible to confirm these hypotheses in humans, thus highlighting the importance of using model organisms such as mice. Available resources such as the International Mouse Phenotyping Consortium (IMPC) [179] and Knockout Mouse Project (KOMP) [180] which aim to identify the function of every protein-coding gene in the mouse genome, the Origins of Bone and Cartilage Disease (OBCD) initiative [181], and Bonebase project [182] that are providing detailed bone phenotyping of mice from these efforts are key components of this step. Additionally, in vitro (using human bone cell lines or human primary cells) or in vivo (using model organisms) testing of genes using gene editing technologies such as CRISPR/Cas9 [183] or its variations (CRISPRi, CRISPRa, etc. [184]) will be key in uncovering the biology identified by GWAS. Finally, the use of genome-scale CRISPR/Cas9 functional genetic screens could potentially uncover other genes beyond those directly implicated by GWAS.

1.4 Summary

In summary, over the past 12 years, GWASs have identified a large number of variants influencing BMD. Post-GWAS efforts are attempting to identify the genes responsible for their effects. We now know that most of the variants identified by GWAS exert their impact on bone by altering gene regulation. We also know that the discovery of causal genes has the potential to provide insight into osteoporosis etiology and identify novel therapeutic targets. The approaches and findings highlighted in this introduction will only continue to improve given rapid advances in statistical approaches and technologies to profile molecular phenotypes. As the field progresses and we continue to unlock the secrets of the human genome, it is our hope that we will be able to use this information to develop more effective therapies to treat and ultimately prevent osteoporosis. In my dissertation work, I aimed to further characterize previously identified genetic associations, from humans, using novel, unbiased approaches in the following studies:

(1) In **Chapter 2**, we focus on identifying potentially causal long non-coding RNAs, an understudied population of non-coding RNAs in the context of bone and osteoporosis. We were

able to identify 23 lncRNAs as potentially causal in osteoporosis and are candidates for experimental follow-up studies.

(2) In **Chapter 3**, we leveraged long-read proteogenomics to identify potentially causal protein-coding isoforms in osteoporosis. We provided a list of potentially causal isoforms and functionally validated the gene *TPM2 in vitro* as a novel causal gene in osteoporosis.

(3) In **Chapter 4**, I shared my final thoughts on the state of the field and future directions for the next generation of systems geneticists who aspire to provide treatment for osteoporosis.

Chapter 2

Identification of known and novel long non-coding RNAs potentially responsible for the effects of BMD GWAS loci

Published in:

Abood, A., Mesner, L., Rosenow, W., Al-Barghouthi, B.M., Horowitz, N., Morgan, E.F., Gerstenfeld, L.C. and Farber, C.R. (2022), Identification of Known and Novel Long Noncoding RNAs Potentially Responsible for the Effects of Bone Mineral Density (BMD) Genomewide Association Study (GWAS) Loci. J Bone Miner Res, 37: 1500-1510. https://doi.org/10.1002/jbmr.4622

This work has won the 2022 Raisz-Drezner JBMR® First Paper Award (https://tinyurl.com/raiszdrezner)

Prior to publication, this work was shared with the scientific community in the form of a poster presentation in 2021 (travel grant award)at The American Society of Bone and Mineral Research (ASBMR) and The American Society of Human Genetics (ASHG)

2.1 Abstract

Osteoporosis, characterized by low bone mineral density (BMD), is the most common complex disease affecting bone and constitutes a major societal health problem. Genome-wide association studies (GWASs) have identified over 1,100 associations influencing BMD. It has been shown that perturbations to long non-coding RNAs (lncRNAs) influence BMD and the activities of bone cells; however, the extent to which lncRNAs are involved in the genetic regulation of BMD is unknown. Here, we combined the analysis of allelic imbalance (AI) in human acetabular bone fragments with a transcriptome-wide association study (TWAS) and expression quantitative trait loci (eQTL) colocalization analysis using data from the Genotype-Tissue Expression (GTEx) project to identify lncRNAs potentially responsible for GWAS associations. We identified 27 lncRNAs in bone that are located in proximity to a BMD GWAS association and harbor SNPs demonstrating AI. Using GTEx data we identified an additional 31 lncRNAs whose expression was associated (FDR correction<0.05) with BMD through TWAS and had a colocalizing eQTL (regional colocalization probability (RCP)>0.1). The 58 lncRNAs are located in 43 BMD associations. To further support a causal role for the identified lncRNAs, we show that 23 of the 58 lncRNAs are differentially expressed as a function of osteoblast differentiation. Our approach identifies lncRNAs that are potentially responsible for BMD GWAS associations and suggest that lncRNAs play a role in the genetics of osteoporosis.

Keywords: Osteoporosis; Human association studies; Osteocytes; Osteoblasts

2.2 Introduction

Osteoporosis is characterized by low bone mineral density (BMD) and deteriorated structural integrity which leads to an increased risk of fracture [2,185]. In the U.S. alone, 12 million individuals have been diagnosed with osteoporosis, contributing to over 2 million fractures per year [3]. This number is expected to nearly double by 2025, resulting in approximately $26 billion in health care expenditures [3].

BMD is one of the strongest predictors of fracture [5] and is a highly heritable quantitative trait ($h^2 = 0.5$-0.8) [6–8,186]. As a result, the majority of genome-wide association studies (GWASs) conducted for osteoporosis have focused on BMD. The largest BMD GWAS performed to date used the UK BioBank (N~420K) and identified 1,103 associations influencing heel estimated BMD (eBMD) [13]. One of the main challenges of BMD GWAS is that the majority (>90%) of associations implicate non-coding variants that lie in intronic or intergenic regions suggesting they have a role in gene regulation. This has made it difficult to pinpoint causal genes and highlights the need for follow-up studies [20]. In addition, few studies have systematically evaluated non-coding transcripts as potential causal genes.

The largest and most functionally diverse family of non-coding transcripts are long non-coding RNAs (lncRNAs). LncRNAs are transcripts longer than 200 nucleotides and have no coding potential [187]. The majority of lncRNAs share sequence features with protein-coding genes including a 3' poly-A tail, a 5' methyl cap, and an open reading frame [188]. However, their expression is low and heterogenous, and they show intermediate to high tissue specificity [189]. Aberrant expression of lncRNAs has been linked to diseases such as osteoporosis [190]. Additionally, there is accumulating evidence suggesting their involvement in key regulatory pathways, including osteogenic differentiation [187,191].

Although understudied in the context of GWAS [189], there is increasing evidence suggesting that lncRNAs are causal for a subset of associations identified by GWAS. A recent analysis of data from the Genotype-Tissue

Expression (GTEx) project identified 690 potentially causal lncRNAs underlying associations influencing risk of a wide range of diseases [189]. Additionally, there is emerging evidence implicating lncRNAs in the genetics of BMD [192–194]. For example, a study reported 575 differentially expressed lncRNAs between high and low BMD groups in Caucasian women, 26 of which regulate protein-coding genes that are potentially causal in BMD GWAS [195]. Additionally, a recent BMD single nucleotide polymorphism (SNP) prioritization analysis implicated lncRNAs as potential effector transcripts [196]. Together these studies suggest that lncRNAs may play an important role in the genetic regulation of bone mass.

In recent years, a number of approaches have been developed that utilize transcriptomics data to inform GWAS, including the analysis of allelic imbalance (AI), transcriptome-wide association studies (TWASs), and expression quantitative trait loci (eQTL) colocalization [51]. AI results from the cis-regulatory effects (i.e., local eQTL) that can be tracked using heterozygous coding SNPs. In transcriptome-wide association studies (TWASs) the genetic component of gene expression in a reference population is estimated and then imputed in a much larger population. Once gene expression is imputed, genetically regulated gene expression is associated with a disease or disease phenotype [197]. Most genes identified by TWAS are located in GWAS associations for that disease and, as a result, TWAS can pinpoint genes likely to be causal at GWAS loci [198,199]. eQTLs are genetic variants associated with changes in gene expression and can be tissue-specific or shared across multiple tissues. eQTL colocalization tests whether the change in gene expression and the change in a trait of interest are driven by the same shared genetic variant(s). All three approaches, alone or in combination, have been successfully used to pinpoint potential causal disease genes at GWAS associations.

Here, we identified lncRNAs that are potentially responsible for the effects of BMD GWAS associations by first applying AI to bone samples and, next, applying TWAS and eQTL colocalization to gene expression data from GTEx. Through both approaches we identified 58 lncRNAs with evidence of being causal BMD GWAS genes. We further prioritized these lncRNAs by identifying those that were differentially expressed as a function of osteoblast differentiation. Together, these results highlight the potential importance of lncRNAs as candidate causal BMD GWAS genes.

2.3 Methods

Patient demographics

All human specimen collection was performed in accordance with institutional review board (IRB) approval from our institution (IRB number H-32517). Acetabular reaming from 17 Boston Medical Center (BMC) patients (ages 43–80 years) undergoing elective hip arthroplasty were collected: 12 females and 5 males; 8 black, 8 white, and 1 Hispanic. This demographic mix reflects the population serviced by Boston University Medical Campus (BUMC), which is an urban safety-net hospital.

RNA extraction

Bone fragments were isolated from the 17 patients. Total RNA was isolated from bone fragments as described in Sagi and colleagues [200]. Total RNA sequencing (RNAseq) libraries were constructed from bone as well as human fetal osteoblast (hFOB) RNA samples using Illumina TruSeq Stranded Total RNA with Ribo-Zero Gold sample prep kits (Illumina, San Diego, CA, USA). Constructed libraries contained all RNAs greater than 100 nucleotides (nt) (both unpolyadenylated and polyadenylated) minus cytoplasmic and mitochondrial ribosomal RNAs (rRNAs). Samples were sequenced to achieve a minimum of 50 million reads 2 × 75 base pair (bp) paired-end reads on an Illumina NextSeq500 (Illumina).

hFOB cell line culture

hFOB 1.19 cells (American Type Culture Collection [ATCC], Manassas, VA, USA; #CRL-11372) were cultured at 34C and differentiated at 39.5C as recommended with the following modifications. Growth media: minimal essential media (MEM; Gibco, Grand Island, NY, USA; 10370-021) supplemented with 10% fetal bovine serum (FBS; Atlantic Biologicals, Morrisville, NC, USA; S12450), 1% Glutamax (Gibco; 35050-061), 1% Pen Strep (Gibco; 15140-122). Differentiation media: MEM alpha (Gibco; 12571-063) supplemented with 10% FBS, 1% Glutamax, 1% Pen Strep, 50 µg/µL Ascorbic Acid (Sigma-Aldrich, St. Louis, MO, USA; A4544-25G), 10mM beta-Glycerophosphate (Sigma-Aldrich; G9422-100G), 10nM Dexamethasone (Sigma-Aldrich; D4902-25MG). RNA was isolated from ~0.5 × 106 cells at days 0, 2, 4, 6, 8, and 10 of differentiation as recommended (RNAeasy Minikit; QIAGEN, Valencia, CA, USA; 74106). Mineralized nodule formation was measured by staining cultures with Alizarin Red (40mM, pH 5.6; Sigma-Aldrich; A5533-25G). Reported results were obtained from three biological replicate experiments.

RNA sequencing and differential gene expression analysis

Computational analysis of RNA sequencing data for the 17 bone samples, Farr and colleagues [201] and the hFOB samples were performed using a custom bioinformatics pipeline. Briefly, FastqQC (Babraham Bioinformatics, Cambridge, UK; http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and RseQC [202] were used to assess the quality of raw reads. Adapter trimming was completed using Trimmomatic [203]. Sequences were aligned to the GENCODE v34 [204] reference genome using the SNP and splice aware aligner HISAT2 [205]. Genome assembly and abundances in transcripts per million (TPM) were quantified using StringTie [87]. Differential expression analysis for the hFOB differentiation experiment was performed using Deseq2 [206] across all six differentiation time points using analysis of deviance (ANODEV) which is conceptually similar to analysis of variance (ANOVA). Differential expression analysis for the comparison between this study's samples and the Farr and colleagues [201] samples was performed using Deseq2 [206] standard approach.

lncRNA discovery

The Coding Potential Assessment Tool (CPAT) [207] was used to assess the protein-coding potential of the novel transcripts assembled. In short, CPAT is a machine learning algorithm trained on a set of known human lncRNAs to identify novel putative lncRNAs based on shared sequence features. We used all known lncRNAs in the latest human genome assembly (GRCh38) as the training set. Novel transcripts with coding probability < 0.367 are regarded as lncRNAs in accordance with software authors. Novel lncRNAs with TPM < 1 were regarded as noise and discarded.

Allelic Imbalance (AI) analysis

Reads were aligned to the GENCODE v34 [204] reference genome using the SNP and splice aware aligner HISAT2 [205]. The resultant BAM files were then used as input for variant calling using the GATK pipeline [208]. Briefly, duplicate reads were identified using MarkDuplicates. Next, reads spanning introns were reformatted using SplitNCigarReads to match the DNA aligner conventions. Then base quality recalibration was performed to detect and correct for patterns of systematic errors in the base quality scores. Finally, the variant calling and filtration step was performed using HaplotypeCaller. The resultant VCF file included only known and novel SNPs and reference bias was corrected using WASP [209]. Briefly, mapped reads that overlap SNPs are identified. For each read that overlaps a SNP, its genotype is swapped with that of the other allele and it is re-mapped. If a re-mapped read fails to map to exactly the same location, it is discarded. The resultant corrected BAM and filtered VCF files were used as input for GATK ASEReadCounter to provide a table of filtered base counts at heterozygous sites for allele specific expression. Bases with a read depth less than 20 were discarded. In order to determine significance, a binomial test was performed and only heterozygous sites with false discovery rate (FDR)-corrected p value of < 0.05 were considered significant.

TWASs

We conducted a TWAS by integrating genome-wide SNP-level association summary statistics from a BMD GWAS [13] with GTEx version 8 gene expression QTL data from 49 tissue types. We used the S-MultiXcan [210] approach for this analysis, to correlate gene expression across tissues to increase power and identify candidate susceptibility genes. Gene-level associations were identified at FDR correction < 0.05 and were further filtered using fastENLOC (a faster implementation of ENLOC [30]) regional colocalization probability > 0.1 in at least one tissue type.

Bayesian colocalization analysis

We used fastENLOC to perform Bayesian colocalization analysis. We integrated summary statistics from the most recent (and largest) eBMD GWAS [13] and eQTL data from 49 GTEx tissues [31]. We used the recommended regional colocalization probability (RCP) threshold of > 0.1 as indication of significant overlap between SNP and eQTL.

2.4 Results

In this study, we used two approaches to identify lncRNAs that potentially underlie BMD GWAS associations. In the first approach, we quantified known (lncRNAs that have been reported in the GENCODE database) and novel lncRNAs using RNA sequencing (RNAseq) data from human bone fragments and identified lncRNAs located in proximity of a BMD GWAS association and harboring SNPs demonstrating AI. In the second approach, we leveraged GTEx to identify lncRNAs across a large number of tissues and cell-types whose expression was significantly associated with BMD by TWAS and regulated by an eQTL that colocalized with a BMD association. **Figure 2.1** provides an overview of our study.

2.4.1 Generation of bone expression data from bone fragments

To identify potentially casual lncRNAs in a BMD relevant tissue, we generated total RNAseq (ribo-depleted) data on bone fragments isolated from acetabular reamings from patients undergoing hip arthroplasty (n = 17; 5 males and 12 females; ages 43 to 80 years). The acetabular reamings were comprised primarily of bone and marrow with a small number of contaminating cartilage fragments. In contrast to most gene expression data generated on bone which are typically from biopsies that contain marrow, we were able to remove the marrow leaving purified trabecular and cortical bone. We hypothesized that the acetabular bone fragments consisted primarily of late-stage osteoblasts/osteocytes [211], allowing us to characterize lncRNAs enriched in these cell types. To confirm that the acetabular samples were enriched in osteocytes, we compared these data to published RNAseq data on bone biopsies [201]. Farr and colleagues [201] generated RNAseq data on 58 iliac crest needle biopsies from healthy women containing both bone and marrow. Average transcripts per million (TPM) across all samples in both experiments were highly correlated (**Figure 2.2A**, $r^2 = 0.845$, $p < 2.2 \times 10^{-16}$). Importantly, differential expression analysis between the two datasets showed that the top 1,000 genes with the largest fold change increase in the bone fragment samples compared to bone biopsy samples were enriched in Gene Ontology (GO) terms such as “skeletal system development” (FDR = 4.01×10^{-3}) and “extracellular matrix organization” (FDR = 4.11×10^{-5}).

To support the notion that our samples are unique in osteocyte enrichment, we used data from a recent study that identified an osteocyte gene signature consisting of 1,239 genes in mice and their orthologs in humans [212]. The ratio of expression (bone fragment samples/bone biopsy samples) was used. A ratio value > 1 indicates that gene expression is higher in the bone fragment samples relative to the bone biopsy samples. In contrast, a ratio value < 1 indicates that the gene is highly expressed in bone biopsy samples relative to bone fragment samples. We expect to see that osteocyte signature genes show ratio values > 1 and marrow enriched genes show ratio values < 1. The osteocyte signature genes showed a median ratio of 1.72 (62% of osteocyte signature genes ratio > 1). Additionally, the ratio of expression of genes enriched in bone marrow showed a median of 0.27 (91% of marrow enriched genes ratio < 1). The distribution of osteocyte signature genes ratio values showed a significant median shift (Wilcoxon test, $p < 2.2 \times 10^{-16}$) (**Figure 2.2D**), and the opposite pattern was observed for the bone marrow enriched genes (Wilcoxon test, $p < 2.2 \times 10^{-16}$) (**Figure 2.2E**). In addition, we compared the expression of osteocyte-specific genes reported in Bonewald et al. [211] (**Figure 2.2B**) and bone marrow enriched genes reported in (www.proteinatlas.org) (**Figure 2.2C**). In addition, during the isolation, care was also taken to remove the cartilage fragments. We repeated the analysis for cartilage marker genes and found a modest reduction ($p = 0.035$) of expression in our samples [213]. The difference was more modest, likely due to a significant overlap in the expression of these genes in both cartilage and bone/osteoblasts. Altogether, these data suggest that the purified acetabular bone fragments are enriched for late osteoblasts/osteocytes and are more marrow depleted compared to iliac crest biopsies.

2.4.2 Identifying novel lncRNAs in purified acetabular bone fragments

Given the paucity of bone transcriptomics data in the literature, and the tissue-specific nature of lncRNA expression, we hypothesized that many bone/osteocyte-specific lncRNAs would not be present in current sequence databases. Additionally, ~50% of lncRNAs do not possess a poly-A tail modification [214] and most RNAseq data is generated after poly-A selection. Therefore, in order to capture a more comprehensive profile of lncRNAs in bone, we implemented a lncRNA discovery step to identify putative "novel" lncRNA transcripts using the computational algorithm CPAT [207]. Across the 17 bone samples we identified 6,612 known lncRNAs and 2,440 novel lncRNAs (**Tables S2.1 and S2.2**). The mean length of novel lncRNAs was 30.3 kilobases (kb) and median length of 11.8 kb. These values were comparable to the mean length of known lncRNAs expressed in the bone samples (mean = 35.4 kb; median = 4.7 kb).

2.4.3 Identifying potentially casual lncRNAs in bone

For lncRNAs to be considered potentially causal in bone, we identified those that are both located in proximity of a BMD GWAS association and regulated by AI. We hypothesized that such genes may be causal for their respective associations because of the potential to be regulated by an eQTL which colocalizes with a BMD association. Of the 9,052 lncRNAs (2,440 novel and 6,612 known) we quantified in acetabular bone, 1,496 lncRNAs (~17% of expressed lncRNAs) were found within a 400 kb window (±200 kb from the lncRNA start site) of each of 1,103 GWAS associations previously identified by Morris and colleagues [13]. The rationale behind choosing this genomic distance was based on findings in Võsa and colleagues [215], where they showed that 92% of lead cis-eQTLs are within 100 kb of the transcription start site. Therefore, this window was extended to ensure we captured the majority of all cis-eQTL effects.

Next, we identified heterozygous coding variants that demonstrated significant evidence of AI within lncRNAs. None of the heterozygous coding SNPs used to assess AI were in linkage disequilibrium (LD) ($r^2 < 0.05$) with a lead BMD GWAS SNP, which is expected because these SNPs were only used to measure AI and not necessarily functionally associated with lead GWAS SNPs. Of the total number of lncRNAs we identified, 174 (47 known, 127 novel; ~2% of expressed lncRNAs) had at least one SNP demonstrating AI in at least one of the 17 bone fragment samples. Out of the 174, 27 (15.5%; 8 known, 19 novel) were located in proximity of a GWAS association (**Figure 2.3A, Table S2.3**). It is expected that we find a low number of lncRNAs (known or novel) under AI relative to the number of expressed lncRNAs within 400 kb of GWAS loci. Reasons for our expectation include the absence of an exonic heterozygous SNP because some lncRNAs that do not have an exonic heterozygous SNPs in LD with a regulatory SNP within the 17 acetabular bone samples will be missing from the intersection. Additionally, lncRNAs in general are lowly expressed; therefore, the power to identify lncRNAs under AI is lower than that of protein-coding genes.

2.4.4 Identifying putatively causal lncRNAs by leveraging GTEx

Next, we sought to leverage non-bone data to identify potentially causal lncRNAs. To do this, we integrated 1,103 BMD GWAS loci with GTEx (v8) eQTL data by coupling TWAS [138] using S-MultiXScan [210] and Bayesian colocalization analysis using fastENLOC [30]. The rationale behind using GTEx data is genes that are shared in multiple tissues and showing a colocalizing eQTL with BMD GWAS data can be potentially causal in bone tissue as well. Our TWAS analysis resulted in 333 significant lncRNA-BMD associations (FDR correction < 0.05), which constitute 5% of all known lncRNAs that are expressed in the acetabular samples. Our colocalization analysis yielded 48 lncRNAs with a colocalizing eQTL (regional colocalization probability [RCP] > 0.1) in at least one GTEx tissue. These lncRNAs with a colocalizing eQTL makeup < 1% of the known expressed lncRNAs in the acetabular bone samples. There were 31 lncRNAs (< 1%) significant in both the TWAS and eQTL colocalization analysis (**Table S2.4**).

2.4.5 Most identified lncRNAs are the only potential effector transcripts implicated by TWAS/eQTL colocalization in their respective GWAS associations

To determine if the lncRNAs listed in **Table S2.4** are the strongest candidates in their respective GWAS associations, we evaluated a recent report of protein coding genes that used the same approach [216]. Five out of the 31 lncRNAs (*LINC01116*, *LINC01117*, *SNHG15*, *LINC01290*, *LINC00665*) have a protein-coding gene with a colocalizing eQTL (*HOXD8*, *HOXD9*, *MYO1G*, *NACAD*, *EMP2*, *ZFP14*, *ZFP82*) within 1 megabase (Mb) of the lncRNA start site (**Table S2.5**). Upon further investigation of the RCP values, some of the lncRNAs showed higher RCP than their protein coding gene counterpart. For example, *LINC01290* had a higher RCP in lung tissue (0.4992) compared to its counterpart *EMP2* (0.2227). On the other hand, the same lncRNA has a lower RCP value (0.1498) than *EMP2* (0.6089) in breast and mammary gland tissue. However, for the remaining lncRNAs, this analysis provides support that the lncRNA alone is the potential effector transcript in the region because we show no evidence of protein coding colocalization within 1 Mb distance of the start site of the lncRNA.

2.4.6 Many identified lncRNAs are differentially expressed as a function of osteoblast differentiation

To provide further support for the hypothesis that these lncRNAs mediate GWAS associations, we measured their expression as a function of osteoblast differentiation in human fetal osteoblasts (hFOBs). We performed total RNAseq at six hFOB differentiation time-points (days 0, 2, 4, 6, 8, and 10). Of the 27 lncRNAs implicated in the analysis of AI, all eight known lncRNAs were differentially expressed (FDR < 0.05). On the other hand, none of the novel lncRNAs were differentially expressed (**Table S2.3**). Examples of the identified genes include *MALAT1* and *NEAT1* (**Figure 2.3B,C**), which were differentially expressed in hFOBs and showed evidence of AI in 8 and 10 of the 17 acetabular bone samples, respectively. There were four unique SNPs in the exonic regions of *MALAT1* (**Figure 2.3B**) that were heterozygous in at least one of the 17 individuals (with a maximum of eight individuals). All four SNPs showed higher expression in the alternative allele relative to the reference allele. The expression of *MALAT1* gene decreased as the cell differentiated into a mineralizing state (**Figure 2.3D**). Additionally, there were nine unique SNPs reported in the exonic regions of *NEAT1* that were heterozygous in at least one of the 17 individuals (with a maximum of 10 individuals). Of the nine, eight showed higher expression associated with the alternative allele compared to the reference allele. The remaining SNP was associated with the opposite pattern, and this was likely due to it being the only SNP not in high LD with the others ($r^2 = 0.0021$). *NEAT1* showed significant increase in expression around day 10 in hFOBs (**Figure 2.3E**).

We assessed the expression of lncRNAs identified by GTEx TWAS/eQTL colocalization in osteoblast differentiation using the same approach in the previous section. Out of the 31 lncRNAs identified by TWAS/eQTL colocalization, 15 were found to be differentially expressed (*LINC00184*, *SH3RF3-AS1*, *LINC01116*, *LINC01934*, *C3orf35*, *LINC01018*, *ARRDC3-AS1*, *LINC00472*, *SNHG15*, *GAS1RR*, *LINC00840*, *LINC01537*, *LINC00346*, *LINC01415*, *MIR155HG*). In general, the expression of those genes in hFOBs was low compared to the lncRNAs reported in the AI section. Examples include *SHR3F3-AS1* and *LINC00472*, which were regulated by colocalizing eQTL (**Figure 2.4B,D**) and were differentially expressed in hFOBs. (**Figure 2.4C,E**). *SH3RF3-AS1* was shown to have the highest RCP value overall (RCP = 0.72) and in only one

GTEx tissue (cultured fibroblasts) (**Figure 2.4A,D, Table S2.4**). Although the gene was differentially expressed across hFOB differentiation points, it had a very low overall level of expression (**Figure 2.4E**). The pattern of expression decreased during mid differentiation points with spikes in early and late points (**Figure 2.4E**). *LINC00472* was shown to have a colocalizing eQTL in four GTEx tissues with the highest RCP value in brain cerebellar hemisphere (RCP = 0.37) (**Figure 2.4A,B, Table S2.4**). The gene also showed a moderate level of expression in hFOBs with an average of 1.5 TPM (**Figure 2.4C**). The expression of *LINC00472* peaked at day 2 and then declined (**Figure 2.4C**).

2.5 Discussion

In this study, we interrogated BMD GWAS loci and identified known and novel lncRNAs as potential effector transcripts. We identified potentially important lncRNA using two different approaches. First, we identified novel and known lncRNAs in a unique transcriptomic bone dataset that were localized in GWAS loci and demonstrated AI. Second, we implicated additional lncRNAs by leveraging GTEx and identifying eQTLs in non-bone tissues that colocalized with eBMD GWAS loci whose expression was associated with eBMD via TWAS. We also assessed differential expression across the time course of hFOB differentiation to provide more evidence of a potential causal role for these lncRNAs.

In the first approach, we set out to perform transcriptomics on a unique set of bone samples in order to identify novel lncRNAs in bone, provide deeper coverage for known lncRNA identification, and apply AI analysis. The bone samples that exist in the literature are from bone biopsies, and as we show in the Results section, they are less enriched in bone-relevant genes compared to the dataset produced by the bone fragments used in this study.

A total of eight lncRNAs (*NEAT1*, *MALAT1*, *DLEU2*, *LINC01578*, *CARMN*, *AC011603.3*, *PXN-AS1*, *AC020656.1*) were found to be within a 400 kb window of an eBMD GWAS locus and were also differentially expressed across hFOB differentiation time points. Many of these lncRNAs have been demonstrated to play a role in bone. For example, *NEAT1* has been reported to stimulate osteoclastogenesis via sponging miR-7 [217] and the NEAT1/miR-29b-3p/BMP1 axis promotes osteogenic differentiation in human bone marrow–derived mesenchymal stem cells [218]. In addition, *MALAT1* has been shown to influence BMD [219]. *MALAT1* acts as a sponge of miR-34c to promote the expression of *SATB2*. *SATB2* then acts to reduce the alkaline phosphatase (ALP) activity of osteoblasts and mineralized nodules formation [219]. A recent study [220] has shown that *LINC01578* (referred to as CHASERR in this study) represses chromodomain Helicase DNA Binding Protein 2 (*Chd2*). A model for Chd2 loss of function by the International Mouse Phenotyping Consortium (IMPC) [221] reported that these mice exhibit significantly decreased body weight and length, skeletal abnormalities, abnormal bone structure, decreased fat levels, and BMD [220]. Last, *DLEU2* expression has been shown to be inversely correlated with BMD in a study involving postmenopausal white women [38]. The remaining four lncRNAs have not been reported to date to have a role in bone and should be further pursued.

In our second analysis, we reported 15 lncRNAs implicated jointly by colocalization, TWAS, and differential expression analysis. We show one example of the 15 lncRNAs reported in *SH3RF3-AS1* in **Figure 2.4A**. Most of these lncRNAs have not been shown previously in the literature to have a role in bone biology. However, *LINC00472* (**Figure 2.4B**) has been experimentally shown to influence osteogenic differentiation by sponging miR-300, which in turn increases the expression of *Fgfr2* in mice [222]. These preliminary results provide more evidence of the potential causal role of these lncRNAs in osteoporosis.

In this study, we were able to use multiple systems genetics approaches on two transcriptomic datasets (acetabular bone and GTEx) to identify lncRNAs that are potentially responsible for the effects of some BMD GWAS loci. This is the first study to our knowledge that evaluated the role of lncRNAs in mediating the effect of BMD GWAS loci from a genome wide perspective. We combined osteoblast differentiation samples and the literature to provide experimental evidence in previous studies to support the effector transcript list we generated from our analysis. These results highlight the importance of studying other aspects of the transcriptome to identify potential drug targets for osteoporosis and bone fragility.

Limitations of this study

This study is not meant to be comprehensive because we are limited by the number of samples and are not suitably powered to identify eQTLs and apply TWAS/colocalization analysis. However, due to the scarcity of population-level bone transcriptomic datasets, and the lack of bone cell or tissue data in GTEx, our study is an attempt to systematically leverage the available datasets to capture a subset of lncRNAs that we think are potentially causal. As mentioned, some of these lncRNAs have been implicated experimentally outside of this study. Moreover, lncRNAs under AI and within proximity of GWAS loci may not be causal as they could be false positives because they are not prioritized via a systems analysis such as colocalization. Another limitation of our study is that we evaluated their expression as a function of osteoblast differentiation; however, it is likely that some of the lncRNAs, if truly causal, impact BMD via a function in other cell-types (e.g., osteoclasts). Future studies should focus on enhancing these results by generating transcriptomic and eQTL datasets from bone and other bone cell types, using network approaches to aid in the prioritization of lncRNAs, and experimentally validating the role of specific lncRNAs.

2.6 Acknowledgements

Research reported in this publication was supported in part by the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health under Award Number AR071657 to CRF, LCG, and EFM, and AA was supported in part by a National Institutes of Health, Biomedical Data Sciences Training Grant (5T32LM012416). The authors acknowledge Emily Farber for generating RNAseq data on bone fragments. We thank the IMPC for accessibility to BMD data in knockout mice (www.mousephenotype.org). The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal on 6/30/20.

Authors' roles: AA: Conceptualization, Methodology, Formal analysis, Investigation, Validation, Visualization, Writing—original draft, Writing—review & editing. LM: Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing—review & editing. WR: Data curation, Formal analysis, Investigation, Methodology. BMA-B: Investigation, formal analysis. NH: Investigation. EFM: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing—review & editing. LCG: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing—review & editing. CRF: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing—review & editing.

The data that support the findings of this study are openly available in Gene Expression Omnibus (GEO) at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE186922, reference number GSE186922.

Chapter 3

Long read proteogenomics to connect disease-associated sQTLs to the protein isoform effectors in disease

In review at Cell Genomics:

Abood, A., Mesner, L., Jefferey, ED., Murali, M., Lehe, M., Saquing, J., Jordan, BM., Farber, CR., Sheynkman, GM. (2023), Long read proteogenomics to connect disease-associated sQTLs to the protein isoform effectors in disease. BioRxiv. https://doi.org/10.1101/2023.03.17.531557

Prior to submission, this work was shared with the scientific community in the form of an invited talk at The American Society of Human Genetics (ASHG) in 2022 and as a plenary poster presentation (travel grant award) in 2022 at The American Society of Bone and Mineral Research (ASBMR)

3.1 Abstract

A major fraction of loci identified by genome-wide association studies (GWASs) lead to alterations in alternative splicing, but interpretation of how such alterations impact protein is hindered by the technical limitations of short-read RNA-seq, which cannot directly link splicing events to full-length transcript or protein isoforms. Long-read RNA-seq represents a powerful tool to define and quantify transcript isoforms, and, recently, infer protein isoform existence. Here we present a novel approach that integrates information from GWAS, splicing QTL (sQTL), and PacBio long-read RNA-seq in a disease-relevant model to infer the effects of sQTLs on the ultimate protein isoform products they encode. We demonstrate the utility of our approach using bone mineral density (BMD) GWAS data. We identified 1,863 sQTLs from the Genotype-Tissue Expression (GTEx) project in 732 protein-coding genes which colocalized with BMD associations ($H_4PP \geq 0.75$). We generated deep coverage PacBio long-read RNA-seq data (N=~22 million full-length reads) on an in vitro model of relevance to BMD, human osteoblasts, identifying 68,326 protein-coding isoforms, of which 17,375 (25%) were novel. By casting the colocalized sQTLs directly onto protein isoforms, we connected 809 sQTLs to 2,029 protein isoforms from 441 genes expressed in osteoblasts. Using these data, we created one of the first proteome-scale resources defining full-length isoforms impacted by colocalized sQTLs. Overall, we found that 74 sQTLs influenced isoforms likely impacted by nonsense mediated decay (NMD) and 190 that potentially resulted in the expression of new protein isoforms. Finally, we identified colocalizing sQTLs in *TPM2* for splice junctions between two nearly mutually exclusive exons, and two different transcript termination sites, making it impossible to interpret without long-read RNA-seq data. siRNA mediated knockdown in osteoblasts showed two *TPM2* isoforms with opposing effects on mineralization. We expect our approach to be widely generalizable across diverse clinical traits and accelerate system-scale analyses of protein isoform activities modulated by GWAS loci.

3.2 Introduction

Genome-wide association studies (GWASs) have identified thousands of associations influencing complex diseases [49]; however, the main challenge limiting the use of GWAS data to uncover novel biology and new therapeutic targets is pinpointing causal genes. In recent years, it has become increasingly apparent that a substantial fraction of GWAS associations act by modulating alternative splicing (AS) [44,48,223]. Genetic variants influencing AS are identified as splice quantitative trait loci (sQTLs) and several studies have linked sQTL to disease associations through colocalization approaches [32,224–226]. However, it is not known to what extent the effects of GWAS loci are mediated by AS in general, and the identity of the downstream protein isoforms that mediate disease.

In a majority of functional genomics studies, splicing is characterized by algorithms such as LeafCutter [69] or rMATS [80], which quantifies local regions of splicing based on the relative abundance of introns (or exon-exon junctions) [69]. These approaches have proven reliable as a way to globally quantify individual splicing events, and even in a reference annotation-free manner to discover novel splicing events. But, this information is only a partial picture—a majority of human genes express isoforms with multiple, distinct splicing events that can influence each other in cis, creating dependencies of splicing choices within the same transcript [227,228]. Unfortunately, short-read RNA-seq datasets can only return a probabilistic, not definitive, knowledge of isoform expression, many times with inaccuracies [229].

The influence of splicing in the genetics of complex disease is clear; however, it is often difficult to connect the events influenced by sQTLs to the full-length transcript isoforms they impact. This disconnect complicates efforts to interpret the effect of sQTLs and the design of experiments to test the role of specific isoforms for functional validation. Furthermore, the true impact of an sQTL on protein isoform function is unknown without direct measurement of the protein. It is only with knowledge of the number and sequence of protein isoform expression changes that one can determine how a GWAS locus mediates protein function from simple loss of protein stability to the generation of an alternative protein isoform with different functional activities, such as differential protein-protein interactions.

In recent years, long-read RNA-seq technologies have been shown to generate better predictions of proteoforms [230] including candidate novel protein isoforms [111,115]. Here, we present a new approach that connects disease-associated sQTLs directly to the transcript and putative protein isoforms they impact. We accomplish this by integrating information from GWAS with large-scale sQTL datasets, which identifies hundreds of colocalized sQTLs, and directly cast such sQTLs onto relevant isoform models derived from long-read RNA-seq. In this process, sQTLs can be interpreted in terms of the disease-relevant protein isoforms that are correlated to a trait, enabling isoform-resolved studies from single-gene to systems scales. Such resolution facilitates hypothesis generation of individual or groups of isoforms playing a role in the clinical trait of interest.

As a proof of concept, we applied this sQTL-long-read contextualization approach to bone mineral density (BMD) GWAS [13] by leveraging sQTL from the Gene Tissue Expression (GTEx) project [31] and long-read proteogenomics in human fetal osteoblasts (hFOBs) [231] - a BMD relevant cell model. We identified 2,029 (643 novel) full-length protein isoforms from 441 protein-coding genes that are candidate effectors of BMD. To assess putative functions, we predicted complete open-reading frames and the effect of the associated protein isoforms on BMD. One of the genes identified through our approach was beta-tropomyosin (*TPM2*). Our analysis predicted that two different sets of isoforms characterized by the presence of two different nearly mutually exclusive exons had opposing effects on mineralization, which we confirmed through isoform specific knockdown in hFOBs.

Our approach facilitates the interpretation of the effects of sQTLs to implicate isoforms likely involved in the regulation of BMD. This approach can be used for biomarker and novel therapeutic target identification, as well as understanding the splicing determinants of clinical traits, across the spectrum of human diseases.

3.3 Results

An overview of our approach for systematic discovery of the transcript/protein isoforms potentially responsible for GWAS association is shown in **Figure 3.1**. The approach provides connections between GWAS associations, sQTLs, and most importantly, transcript/protein isoforms, and does so using a novel long read proteogenomics approach [115]; thereby, increasing the utility and interpretability of colocalized sQTLs.

3.3.1 Identification of genes potentially regulating BMD through splicing

To nominate splice events of potential relevance to BMD, we leveraged existing sQTLs across 49 tissues from the GTEx project [32]. Population-scale transcriptomic datasets of bone or bone cells are scarce. However, prior studies have reported extensive sharing of sQTLs across tissues, thus, we reasoned that many of the sQTL from GTEx would also be found in bone/bone cells and be of potential relevance to BMD [48]. We performed Bayesian colocalization analysis using coloc [29], GTEx sQTLs, and the largest BMD GWAS performed to date, which identified 1103 independent associations [13] (**Supplemental note 1**). Overall, we found 732 protein-coding genes with colocalizing sQTLs (H_rPP ≥ 0.75), denoted here as sGenes (**Figure 3.2A**). The colocalized sQTLs represent 1,863 distinct junctions with an average of 2.5 junctions per gene (**Figure 3.2A**). Over half of the sGenes (367, or 50%) have shared sQTLs across multiple tissues (**Table S3.1**). Examples of highly colocalizing sQTLs for *TCF7L2* and *FHL3* are shown in **Figure 3.2B-C**. Collectively, these results identify genes whose genetically regulated splicing alterations potentially mediate BMD associations.

3.3.2 Characterizing the full-length transcriptome across osteoblast differentiation

In order to provide molecular context for the individual splicing “events” associated with colocalized sQTLs, we generated an experimentally-derived and comprehensive map of relevant isoform expression. Our goal was to enable the identification of the full-length transcript isoforms impacted by each colocalized sQTL and to ensure such isoforms were expressed in a cell-type relevant to BMD. To accomplish this, we generated deep coverage long-read RNA-seq data across osteoblast differentiation in hFOBs. Osteoblasts build bone and are critical regulators of bone development and maintenance [232]. The hFOB cell-line is a well characterized model of osteoblast differentiation capable of *in vitro* mineralization [233]. Across 0, 2, 4, and 10 days of osteoblast differentiation, we generated a deep coverage full length transcriptome dataset (**Supplemental notes 2 & 4**), with collection of over 22 million full-length cDNA sequences using long-read RNA-seq on the PacBio Sequel II platform (**Figure S1A-B**) (see **Methods**). We detected 68,326 transcript isoforms from 12,068 genes. Transcripts of all lengths were evenly sampled (median: 2,074 nt, range: 87-8,787 nt). We found that 50,588 (74%) isoforms were known (annotated in GENCODE) and 17,738 (~26%) were novel. Of the novel isoforms, 10,793 (61%) arose from new combinations of known splice sites (**Figure S1C**); whereas 6,580 (39%) arose from at least one novel splice donor or acceptor. Overall, this map of transcript isoforms in human osteoblasts is both comprehensive and important for revealing cell-type-specific novel isoforms.

3.3.3 Connecting colocalized sQTLs to the transcript and protein isoforms they regulate through long read proteogenomics

Despite the wealth of global sQTL studies to date, few have connected sQTLs to the protein isoforms they impact. Here we directly mapped colocalized sQTLs onto the protein isoform models generated by long read proteogenomics of a relevant disease model. We found that 836 junctions exactly map (shared splice donor/acceptor sites) to 2,349 (700 novel; ~30%) in 459 protein-coding genes in the hFOB dataset (**Supplemental note 3**). Collectively, these sGenes are found within 362 of the 1,103 BMD GWAS associations (~33% of the total), with 221 lead associations harboring one sGene and 141 harboring more than one sGene (**Table S3.2 and Supplemental note 5**). To our knowledge, this is the first genome-scale map of full-length isoform candidates that contribute to a human disease.

Colocalized sQTLs impacting novel isoforms

It is now routine to discover hundreds of novel junctions corresponding to colocalized sQTLs, as we have done in this study. However, it is not possible to map sQTLs involving an unannotated junction to its source full-length isoform(s) without experimental knowledge of isoform expression, such as from long-read RNA-seq (**Figure 3.3**). In the long-read RNA-seq data, we detected 30 novel sQTL junctions, which were not found in the GENCODE reference, but mapped perfectly with one or more full-length novel isoforms detected in hFOBs (**Figure 3.3A**). For example, we identified a novel junction with a colocalizing sQTL ($H_4PP = 0.99$) in zinc finger protein 800 (*ZNF800*), a transcription factor expressed across a wide-range of cell-types that has been previously implicated in pancreatic beta cell development [234] and in cardio-metabolic traits [235], but not in the regulation of BMD (**Figure 3.3B**). We mapped this junction to a novel transcript in hFOBs; however, it does not have a match to any isoform in the GENCODE database (**Figure 3.3C**). The lead variant (rs62621812) for both the BMD locus and sQTL is a rare missense variant (global minor allele frequency in 1000G = 0.005). Although rare, hFOBs were heterozygous for the variant and we observed that the novel *ZNF800* isoforms almost exclusively originated (21 of 22 long reads) from the haplotype harboring the alternative allele, and its expression decreases during hFOB differentiation (**Figure 3.3D)**. The long *ZNF800* isoform containing the putative DNA binding domain is associated with increase in BMD, suggesting that its gene regulatory activities may be osteogenic.

Biological contextualization of isoforms corresponding to known sQTLs

Ostensibly, there is a clear path to contextualize sQTLs containing annotated junctions, as these junctions have a direct mapping to at least one isoform in the reference annotation. However, a single sQTL could map to multiple annotated isoforms and it is not known if all isoforms or only a subset may be relevant in mediating the trait of interest, a common case of isoform ambiguity in sQTL datasets (**Figure 3.4A**). We found a total of 614 sQTLs (73% of 836) mapping to multiple annotated isoforms in the GENCODE transcriptome (v38). To hone in on the most relevant isoforms for the trait of interest, we leverage information about which isoforms are expressed in the hFOB transcriptome map.

For example, the sQTL-to-isoform mapping of genes amplified in osteosarcoma 9 (*OS9*) illustrates the power of providing full-length isoform context. *OS9* has not been directly implicated in the regulation of BMD, but we observed one junction in *OS9* with a strongly supporting colocalizing sQTL ($H_4PP = 0.86$) (**Figure 3.4B**) that corresponds to the skipping of exon 13. This sQTL maps to multiple isoforms in the reference annotation—17 annotated transcripts in the GENCODE Basic set (24 in the GENCODE Comprehensive set)—leaving open the question of which isoform(s) may be specifically relevant to bone cells. Within the hFOB context, we found four *OS9* isoforms expressed in hFOBs: *OS9-203* and *OS9-220*, which both exclude exon 13, and *OS9-202* and *OS9-204*, which include exon 13 (**Figure 3.4C**). We found that *OS9-220* (excludes exon 13) and *OS9-202* (includes exon 13) are the dominantly expressed isoforms and exhibit an isoform switch during osteoblast differentiation with the exon 13-included form *OS9-202* dramatically increasing in abundance, relative to *OS9-220*, suggesting a direct role in driving potential osteogenic pathways (**Figure 3.4D**). Additionally, *OS9* is differentially expressed across osteoblast differentiation. Interestingly, independent evidence from the International Mouse Phenotyping Consortium (IMPC) [236] show that *Os9* knockout mice show significant changes in skeletal phenotypes including abnormal cranium morphology ($p = 1.40 \times 10^{-5}$; both sexes), abnormal tooth morphology ($p = 9.12 \times 10^{-5}$; both sexes), and vertebral fusions ($p = 3.92 \times 10^{-5}$; males only). Furthermore, the exon 13 skipping event is conserved between human and mouse. We hypothesize that the ratio of *OS9-202* to *OS9-220* is associated with decreased BMD, demonstrating how sQTLs connected with long-read RNA-seq data leads to isoform-resolved hypotheses.

Known sQTLs impacting novel isoforms in the biological context

Disease-relevant isoform maps can narrow the possibilities of isoform within which known sQTLs map. However, just as importantly, annotated junctions of colocalized sQTLs can actually be found to map to novel isoforms, meaning that the local splicing event is known, but the associated full-length protein isoform to which it is derived may be novel (**Figure 3.5A**). These "indirectly" novel sQTLs may be an underappreciated source of novel isoforms regulated by GWAS loci. We found that a total of 383 (46%) unique junctions were found in known isoforms only, but for 350 (42%) junctions, they map to both known and novel isoforms. Importantly, a portion of these events were only explained by novel isoforms (103 (12%) were found in novel isoforms only). Since these known sQTLs map to novel isoforms, we argue that they must be "recast" and "re- annotated" as candidate novel protein isoforms.

An example of a gene with annotated junction sQTLs mapping strictly to a novel isoform in hFOBs is dipeptidyl peptidase 8 (*DPP8)*. In our data, we found a colocalizing sQTL ($H_4PP = 0.83$) for the junction associated with inclusion of exon 17 (**Figure 3.5B**). This junction can be mapped in *DPP8-205* and *DPP8-215* in the GENCODE database (**Figure 3.5C**), however, there is no evidence of expression for these two isoforms in hFOBs (**Figure 3.5D**). Rather, the junction strictly maps to *PB.16541.32*, a novel isoform (**Figure 3.5D**). Skipping of exon 17 is found in isoform *DPP8-201*, which is associated with a decrease in BMD. Interestingly, in data from the IMPC, *Dpp8* knockout mice have decreased BMD.

3.3.4 Enrichment of splicing functions among BMD sQTLs and presence of a putative splicing regulatory network

Among all 459 genes with a colocalizing sQTL, we found an enrichment of UniProt keywords involved in alternative splicing (KW-0025 6.39×10^{-17}, RNA splicing). Splice factors (SFs) regulate splicing in trans (**Figure 3.6A**) and have been found to comprise 99 genes to date [237]. Of the 99 annotated SFs, 91 are found in hFOBs. We found 5 splice factors that show a colocalizing sQTL within our long-read hFOB data: *MBNL1*, *PCBP2*, *DDX5*, *PTBP1*, *HNRNPM*. Three of these (60%) show differential expression across hFOB differentiation: *DDX5*, *PTBP1*, *HNRNPM* and only one of them show differential isoform usage: *DDX5*

(**Supplemental note 6**).

One example of a putative splicing regulator in BMD is *PTBP1* which has been extensively studied as a global repressor of exons [238] and is implicated in complex diseases such as atherosclerosis [239]. In our colocalization data, we found that *PTBP1* contains a strongly supported colocalizing sQTL (H_4PP: 0.99; # tissues: 9), associated with the exclusion or inclusion of exon 9 (**Figure 3.6B-C**). Mapping of the junctions to GENCODE annotations reveals 11 isoforms, but in the context of the full-length transcriptome, we confirmed that the sQTL likely corresponds to the two major isoforms in hFOBs: *PTBP1-201*, which excludes exon 9, and *PTBP1-203*, which includes exon 9. Exon 9 resides in a linker region between RNA recognition motif (RRM) domains, and previous studies have shown that skipping of the exon 9 leads to altered polypeptide length between the RRM domains and reduced ability of *PTBP1* to repress exons, likely due to altered linker length between the RRM domains. Previous studies show that skipping of exon 9 reduces the ability of *PTBP1* to repress cryptic exons [240,241]. Therefore, the mechanism of action for this sQTL may be the altered ratio of repressive (*PTBP1-203*) and non-repressive (*PTBP1-201*), with higher ratios corresponding to decreased BMD (**Figure 3.6D**). This genetically-regulated isoform balance could, in turn, globally regulate the repression of *PTBP1* target exons during osteoblast differentiation. Interestingly, a global analysis of all lead sQTLs revealed a clear enrichment in *PTBP1* binding sites, suggesting both *PTBP1* trans-factor and its cis-regulatory targets are part of a splicing regulatory network that could mediate BMD.

3.3.5 Long read proteogenomics for characterizing the landscape of sQTL-associated protein isoforms

We have just described the mapping of sQTLs onto a comprehensive full-length transcriptome. However, even more information is encoded in this dataset, particularly properties of the predicted protein isoforms that are translated from the transcript isoforms. Furthermore, with sQTLs, risk variant-induced changes to the ratios of protein isoforms can have a diversity of molecular functional effects, from overall loss of protein abundance from NMD or degradation mechanisms, to gain of functions from production of multiple, functionally distinct or collaborating protein isoforms from the same gene [45,242]. In order to explore protein isoform-level consequences of colocalized sQTLs, we employed a recently developed "long read proteogenomics" pipeline [115], in which the hFOB transcriptome is used as a template for *in silico* translation of ORFs to generate osteoblast-relevant protein isoform models (**Supplemental notes 7 & 8**).

Despite the prevalence of sQTL datasets, little is known about how sQTLs result in changes to the protein isoform products of a gene, and, particularly, the nature of such changes (e.g., truncated protein, formation of a protein isoform with novel functions) In order to maintain the highest quality of the protein isoform models, we employed a stringent filtering schema (**Supplemental note 2**) and mapped colocalized sQTLs onto 2,029 protein isoforms from 441 genes, defined as all cases in which a junction is wholly residing within a predicted coding sequence (i.e., ORF).

Using a custom pipeline recently developed, Biosurfer (**Methods**), we devised a strategy to bioinformatically compare the groups of non-risk and risk isoforms. The clinical risk status was inferred by mapping the directionality of effect, normalized intron excision ratio (slope from GTEx), and isoform mappings, and for each sQTL, we designated one or more isoforms that are part of the risk versus non-risk group, for protein sequence comparison. In order to quantify the putative risk variant-induced changes, we implemented a custom hybrid alignment strategy, which aligns proteins both based on their exon-intron structure with consideration of the identity of the amino acid residues, and automatically detects regions of the protein that are deleted, inserted, or substituted. Note that these terms refer to changes in polypeptide sequence between two protein isoforms of the same gene, and do not refer to genetic variants. We performed this comparison pairwise between all risk and non-risk isoforms, weighted by their gene expression in hFOBs, and determined if, overall, the effect on the protein was one that led to increased NMD products, a dramatic reduction in protein length, or if the protein models indicate a potential change in two functionally relevant proteins, as indicated by those proteins that are full-length and preserved in their critical domain structure. We found that of the 809 sQTLs, only 74 (9%) likely represent a switch from a full-length protein product to a transcript undergoing NMD, by presence of a premature stop codon upstream of at least one junction [243] (**Table S3.3**). The remaining sQTL-protein-isoform groups involve a switch from one full-length protein to another, albeit with vast differences in length. We found 190 sQTLs that potentially lead to a truncated protein which may represent a sub functional form, or even a dominant negative [244] (**Table S4**). Although as previously observed [245] reduced length does not always correlate with reduced functional capacity [245]. To our knowledge, this is the first proteome-scale dataset of protein isoforms associated with a clinical trait by mapping directly from an sQTL dataset.

3.3.6 Distinct isoforms of *TPM2* have opposing effects on osteoblast differentiation and mineralization

Our analysis identified 441 genes and 2,029 proteoforms potentially linked to BMD. We next sought to select isoforms from this list to test their role in osteoblast differentiation and mineralization. We evaluated genes primarily based on the strength of colocalization (H_4PP) and the extent of sQTL sharing across tissues (reasoning that genes with sQTL observed in multiple tissues were more likely active in osteoblasts). We also evaluated additional data such as whether genes were the cause of a monogenic bone disease or had been implicated in the regulation of BMD in prior studies (**Tables S3.5 and S3.6**). We also prioritized genes not previously implicated in the regulation of BMD.

Based on our long-read RNA-seq data in hFOBs, four primary *TPM2* isoforms were expressed and characterized by two mutually exclusive exons (exons 6 and 7) and two alternative last exons (exons 10 and 11) (TPM2-6-10 (*TPM2* isoform containing exons 6 and 10) TPM2-6-11, TPM2-7-10, and TPM2-7-11) (**Figure 3.7A**). We developed a targeted mass spectrometry method [246] for peptides specific to shared and alternative exons of *TPM2*, including exons 6, 7, 10, and 11 (**Supplemental note 9**). Using these isoform-specific peptide signatures, we observed the same changes in isoform expression across hFOBs.

We identified multiple colocalizing sQTLs ($H_4PP > 0.75$) that converge on the two sets of splicing events. The sQTLs for junctions surrounding exons 6 and 7 were observed in a wide-range of tissues (13 to 33 tissues); whereas those around exons 10 and 11 were restricted to brain and testis (**Figure 3.7B**). The lead sQTL variant differed across tissues; however, the T allele of the lead GWAS SNP (rs2737273) in the locus was associated with a decrease in BMD and an increase in *TPM2* isoforms containing exon 6. Conversely, the C allele was associated with an increase in BMD and increase in isoforms containing exon 7. We also observed in brain and testis that the T allele was associated with increased exon 11 containing isoforms and a decrease in BMD; however, it is difficult to interpret the direction of effects of multiple sQTL due to the correlated nature of multiple splicing events and events occurring across tissues.

Based on the above information, the restriction of *Tpm2* expression to osteoblasts in the mouse (using BioGPS)[247], and its high expression in hFOBs (**Figure 3.7C**), we hypothesized that exon 6 containing *TPM2* isoforms would be negatively associated with mineralization in hFOBs, with the opposite effect found for exon 7 containing isoforms. To test this hypothesis, we performed isoform-specific siRNA knockdown experiments in hFOBs followed by quantification of mineralized nodules (**Figure 7D and E**). First, we knocked down all isoforms of *TPM2* by targeting constitutively expressed exon 2 and exon 9 (**Figure 7D**), and for both experiments we observed no significant differences in mineralization relative to the control (T-Test, $p = 0.51$ and $p = 0.09$ respectively), demonstrating that knock down of all *TPM2* isoforms simultaneously does not result in a phenotype (**Figure 7E**). Next, we targeted *TPM2* isoforms containing exon 6 (targeting the junction between exons 6-8) (**Figure 7D**), and, strikingly, observed a significant increase in mineralization relative to the control (T-Test, $p = 0.004$) (**Figure 7E**). This siRNA moderately decreased the expression of all isoforms, though it did attenuate expression of exon 6 containing isoforms to the greatest degree (**Figure 7D**). On the other hand, knockdown of isoforms containing exon 7 (**Figure 7D**) resulted in a significant decrease in mineralization in accordance with our hypothesis (T-Test, $p = 0.01$) (**Figure 7E**). To comprehensively assess the effect of all *TPM2* splice events, we also targeted exons 10 and 11. We observed no change in mineralization upon knockdown of exon 10 containing isoforms (T-Test, p = 0.82). Surprisingly, knockdown of exon 11 containing isoforms (**Figure 7D**) decreased mineralization (T-Test, p = 0.015) (**Figure 7E**). These data suggest that disruptions in the ratios of the four *TPM2* isoforms have significant impact on mineralization in hFOBs. Most importantly, our results clearly demonstrate that different *TPM2* isoforms have distinct functions with respect to osteoblast activity and are likely regulators of BMD.

3.4 Discussion

A wealth of GWAS and sQTL data has highlighted the widespread involvement of alternative splicing in patient disease risk [32,48], but how individual sQTLs affect downstream transcript and protein isoforms to mediate disease is largely unknown. Here, we present a broadly applicable method to increase the interpretability of such sQTLs. The method integrates sQTLs with long-read RNA-seq data to nominate full-length protein isoforms impacted by colocalized sQTLs, accelerating the path for functionally characterizing the molecular implications of sQTLs.

As a proof-of-principle, we applied our approach to BMD, the single strongest predictor of osteoporotic fracture [248]. We identified thousands of candidate full-length isoforms potentially involved in the regulation of BMD, providing strong support for the hypothesis that splicing is a major mediator of genetic variation impacting bone. Similar conclusions have been reported for other complex diseases such as Alzheimer's, schizophrenia, and cardiometabolic traits [46,224,249] . Our results provide the first resource to specifically connect transcript and protein isoform expression to biological processes and pathways impacting bone mass.

One of the genes with isoforms predicted to influence BMD through our approach was *TPM2*. *TPM2* splicing has not been directly studied in the context of osteoblast differentiation or bone disease; however, mutations have been connected to muscular diseases like the Escobar variant (characterized by skeletal defects including vertebral defects, bone fusion abnormalities and growth retardation) [250], nemaline myopathy [251], and atherosclerosis [252]. Additionally, loss of *Tpm2.1* (the isoform containing exons 6 and 11 together) in the mouse increases beta catenin levels [253], which is integral to osteoblast activity and bone formation [254]. We demonstrated that isoforms containing exon 6 and exon 7 show opposing effects on osteoblast mineralization, suggesting the differences in *TPM2* splicing lead to proteins with different functions in osteoblasts. Intriguingly, we also observed decreased mineralization upon knocking down isoforms containing exon 11. This was unexpected since the isoform with exons 6 and 11 was the most highly expressed and a relatively small decrease in exon 6 isoforms compared to all other isoforms significantly increased mineralization. These results suggest non-additive interactions among *TPM2* isoforms, and that osteoblast activity and mineralization is highly dependent on the precise isoform ratios present at a particular point in differentiation.

We identified an enrichment for genes involved in alternative splicing among all genes with colocalizing sQTLs, including many splicing factors. This led us to discover that colocalized sQTLs for both trans-acting splice factors and their cis-regulatory targets converge onto the same splicing network. For example, *PTBP1* is a splice factor that mediates alternative splicing by binding to polypyrimidine tract and other sites [255]. *PTBP1* and *TPM2* are regulated by colocalizing sQTLs and are potentially linked via a regulatory and biochemical interaction. *PTBP1* isoforms including exon 9 have been shown to bind upstream of and represses the inclusion of exon 7 in *TPM2*, leading to higher levels of isoforms containing exon 6 [238,241]. Intriguingly, the genetic associations indicate that higher levels of exon-9-containing *PTBP1* (lead SNP: rs2737273) and higher levels of exon-6-containing *TPM2* (lead SNP: rs3215700) are associated with lower BMD. Together, these data suggest the presence of a broader splicing network at play wherein sQTLs impact splicing factors that then impact their targets, which are themselves impacted by sQTLs.

Our work represents one “template” of integrating GWAS, sQTL, and long-read transcriptomics for colocalizing sQTL identification and interpretation. In our case there was a lack of short-read data on tissues directly relevant to BMD, therefore, we leveraged GTEx sQTLs and focused on putatively shared sQTLs [48]. By generating long-read RNA-seq in a model system directly relevant to bone, we were able to simultaneously use these data to identify isoforms and leverage this information to contextualize the GTEx sQTLs within a bone-relevant cell. In this sense the long-read RNA-seq data provided a biological context "filter" to nominate the most relevant isoforms for functional validation.

We envision applying different variations of the approach, based on the ability to generate long-read data from particular tissues or cell-types. Another variation of our approach is generating a comprehensive long-read transcriptome reference from a subset of samples, which can serve as a higher accuracy isoform reference for aligning and assembling short-read RNA-seq reads. However, this approach is still subject to well-known limitations of short-read-based isoform characterization. Finally, as the throughput of long-read RNA-seq reaches that of traditional short-read methods, with recent developments from PacBio and ONT platforms, among others, it will become feasible to generate high quality long-read data at population scales that would enable direct discovery of isoform-level QTLs. It is also likely that variations of our approach could use diverse methods of detecting genetically-regulated splicing signals, beyond colocalization approaches, such as transcriptome-wide association studies (TWAS).

Through this study, we were able to identify hundreds of potentially causal isoforms in osteoporosis. However, our study does have limitations, such as the paucity of population-scale bone relevant transcriptomics data, prompting us to transfer putative sQTLs from an independent dataset (GTEx) [48]. Therefore, our analysis missed sQTL associations that are specific in bone. While studies have shown that sQTLs are often shared across tissues, the lack of sQTL data in bone and bone cells means that bone-specific sQTL were missed. In addition, the use of sQTL from multiple non-bone tissues may have inflated the number of false positives due to colocalization signals that have no biological impact on bone. We do believe that the overall false positive rate was reduced based on the requirement that isoforms were expressed in human osteoblasts. Finally, we demonstrated the utility of our approach on a single cell type, but future work should include tissues representing a more comprehensive collection of cell-types of relevance to bone (e.g., osteoclasts, osteocytes, and other cells of mesenchymal origin).

Central to our method is the usage of long-read-driven isoform models serving as a scaffold for interpreting candidate sQTLs; therefore, the quality of transcriptome-wide maps remains critical for appropriate biological interpretation. Long-read RNA-seq technologies continue to evolve at a rapid pace and there is a need for continued evaluation of sequencing metrics such as comprehensiveness, accuracy, and quantitative precision. Recently, the Long Read Genome Annotation Consortium (LRGASP) has undertaken this as a community effort, with the conclusion that novel isoforms are hard to evaluate given the lack of orthogonal, full-length isoform sequencing methods. Sources of biochemical, sequencing, and bioinformatic artifacts have been previously discussed. As it stands, no method yet exists for large-scale detection of isoforms at the protein level, therefore, for tractability, we employed a bioinformatic proteogenomics approach in which the transcript models are used as a proxy for inferring the putative protein isoform products, similarly to the reference annotation [105]. Given the state of protein isoform information, all isoforms of analysis should be fully validated, as we have done with *TPM2*. And, even with full knowledge of protein isoforms expressed in disease relevant models, limitations arising from the nature of complex loci containing many distinct sQTLs and highly complex splicing, in terms of number of events and dependencies between distal splicing, may still be out of reach for straightforward interpretation without extensive functional validation.

In this study, we developed an integrative systems genetics approach to identify isoforms that are potentially responsible for the effects of BMD GWAS loci [13]. This is the first study to our knowledge that directly incorporates long-read RNA-seq to systematically increase the interpretability of colocalized sQTL, providing a catalog of full-length isoforms that are potentially causal effectors of BMD. These results highlight the importance of genetically regulated variation in alternative splicing to BMD and identifies hundreds of potential drug targets for osteoporosis and bone fragility. Our work should serve as a model for other researchers to increase the clinical utility of sQTL analysis across the disease spectrum. This includes usage of sQTL data from publicly available resources like GTEx [31], disease-relevant cohorts, and emerging long-read sequencing approaches.

3.6 Methods

GWAS and sQTL analysis

eBMD GWAS summary statistics were downloaded from the GEnetic Factors for OSteoporosis Consortium (GEFOS) website using http://www.gefos.org/?q=content/data-release-2018 (accessed July 2021). The coordinates of the GWAS SNPs were updated from hg19 to hg38 using LiftOver within Bioconductor. GTEx V8 sQTL all association summary statistics data were downloaded from the GTEx Google Cloud Platform (GCP) using https://console.cloud.google.com/storage/browser/gtex-resources (accessed July 2021). The data were prepared as input for coloc [29] as follows: A list of genes with a start site that is within ∓200 Kb of each eBMD GWAS SNP was created. Using this list, a list of sQTL associations within ∓200 Kb of each of these genes was created for each GTEx tissue. coloc.abf was used with this input (all the GWAS SNPs and all the sQTL data within ∓200 Kb of each gene start site). In order for a junction to be considered significant (junction with a colocalizing sQTL), it must have coloc [29] H_4PP of ≥ 0.75.

Osteoblast differentiation

hFOB 1.19 (American type culture center [ATCC], Manassas, VA; CRL-11372) were grown and differentiated with RNA isolated on days 0, 2, 4 and 10 exactly as outlined in Abood et al. [256]. Briefly, hFOB 1.19 cells

(American Type Culture Collection [ATCC], Manassas, VA, USA; #CRL-11372) were cultured at 34°C and differentiated at 39.5°C as recommended with the following modifications. Growth media: minimal essential media (MEM; Gibco, Grand Island, NY, USA; 10370-021) supplemented with 10% fetal bovine serum (FBS; Atlantic Biologicals, Morrisville, NC, USA; S12450), 1% Glutamax (Gibco; 35050-061), 1% Pen Strep (Gibco; 15140-122). Differentiation media: MEM alpha (Gibco; 12571-063) supplemented with 10% FBS, 1% Glutamax, 1% Pen Strep, 50 µg/µL Ascorbic Acid (Sigma-Aldrich, St. Louis, MO, USA; A4544-25G), 10mM beta-Glycerophosphate (Sigma-Aldrich; G9422-100G), 10 nM Dexamethasone (Sigma-Aldrich; D4902-25MG). RNA was isolated from ~0.5 × 10^6 cells at days 0, 2, 4, and 10 of differentiation as recommended (RNAeasy Minikit; QIAGEN, Valencia, CA, USA; 74106). Mineralized nodule formation was measured by staining cultures with Alizarin Red (40 mM, pH 5.6; Sigma-Aldrich; A5533-25G). Reported results were obtained from three biological replicate experiments for days 2, 4, and 10, and two biological replicates for day 0 of differentiation.

Long-read RNA-seq data collection and sequencing

RNA was extracted from three biological replicates of hFOB cells at days (0, 2, 4, and 10) using Qiagen RNeasy mini kit. Extracted RNA was used to create full length cDNA using the NEBNext® Single Cell/Low Input cDNA Synthesis & Amplification kit (New England Biolabs Ipswich MA Lot#10078130). Following the PacBio Iso-seq protocol, first and second strand cDNA were synthesized using NEB dT oligo and the PacBio Template Switching Oligo. For the cDNA amplification 14 cycles of PCR were performed followed by a Pronex bead cleanup (Promega Corporation Madison WI, Lot #NG103A). The amplified and purified cDNA were QCed using the Bioanalyzer DNA 12000 kit. The samples were then sent to the Smith lab at the University of Louisville for long-read PacBio sequencing.

Approximately 300 ng of cDNA was converted into a SMRTbell library using the Iso-Seq Express Kit SMRT Bell Express Template prep kit 2.0 (Pacific Biosciences). This protocol employs bead-based size selection to remove low mass cDNA, specifically using an 86:100 bead-to-sample ratio (Pronex Beads, Promega). Library preparations were performed in technical duplicate. We sequenced libraries using 11 SMRT cells on the Sequel II system using polymerase v2.1 with a loading concentration of 85pM. A 2-hour extension and 30-hour movie collection time was used for data collection. The "ccs" command from the PacBio SMRTLink suite (SMRTLink version 9) was used to convert raw reads (~22 million) into circular consensus sequence (CCS) reads. CCS reads with a minimum of three full passes and a 99% minimum predicted accuracy (QV20) were kept for further analysis.

Quantitative mass spectrometry-based proteomics

Protein extraction, quantitation, and digestion

hFOB cells used for proteomic mass spec analysis were isolated by removing media from a 10 cm plate of cells (~5x10^6) and subsequently washing with 5 ml of PBS. The cells at day 0, were treated with 2 ml 0.05% trypsin/EDTA for 5 minutes at 37°C, triturated, pelleted at 1000xg for 5 minutes, washed with PBS, pelleted, resuspended in 1.5 ml PBS, transferred to a 1.5 ml microfuge tube, pelleted with the pellet snap frozen and stored at -80°C. The cells at day 2 and 4 were treated with 2.5 ml, 8 mg collagenase (Gibco 17018-029) / ml HBSS (Hanks Balanced Salt Solution [HBSS], Gibco 14025-092, supplemented with 4mM CaCl) for 15 minutes at 37°C followed by the addition of 2.5 ml 0.25% trypsin/EDTA with incubation at 37°C continued for an additional 15 minutes. Cells were subsequently treated as outlined for day 0 cells. The cells at day 10 were initially incubated with 5 ml 60 mM EDTA, ph 7.4 (prepared in PBS) for 15 minutes at RT. The EDTA solution was removed and cells washed with HBSS. The cells were then subsequently treated as outlined for day 2 and day 4 cells

Harvested hFOB cells, approximately 8-10 million cell count each, were pelleted in triplicate for four different time points: day 0, day 2, day 4, day 10. The twelve samples were frozen at -80°C until lysis. Each pellet was lysed according to the Filter Aided Sample Preparation (FASP) protocol adapted from [257]. Lysis buffer was changed to 6% SDS, 150 mM DTT, 75 mM Tris-HCl. To each pellet, an aliquot lysis buffer equal to 2x the pellet volume was added and probe-sonicated to lyse the cells and shear the nucleotide material. Sonication continued for 1-5 minutes until the sample was clear and no longer viscous. The lysates were then incubated at 95°C for 2.5 minutes. Protein quantitation was estimated by BCA assay to be approximately 500-4000 ug per lysate. Aliquots equivalent to 80 ug per sample were used for FASP and buffer exchanged into 200 mM EPPS pH 8.5. A technical replicate of day 10 replicate C was prepared as well as an unrelated Jurkat lysate sample, resulting in a total of 14 samples for proteolytic digestion. Digestion was performed as per Navarrete-Perea et al. [258] with Lys-C overnight, followed by trypsin for six hours, using a 1:100 enzyme-to-protein ratio.

Tandem Mass Tag (TMTpro) labeling

Reagents from the TMTpro 16plex isobaric label reagent set A44522 (ThermoFisher, Waltham, MA) were used for labeling each of the 14 samples (TMTpro-133C and TMTpro-134N were not needed). Per protocol, 20 uL of anhydrous acetonitrile was added to each tube of 0.5 mg dry TMTpro reagent and allowed to incubate at room temperature for 5 minutes with occasional vortexing. The entirety of each TMTpro vial was added to its corresponding digest sample and allowed to incubate at room temperature for one hour, vortexing every 10 minutes, followed by a final centrifugation of the tubes.

C18 desalting, label efficiency check

Aliquots of 2 uL (1.5%) of each digest were combined and desalted using EasyPep Mini C18 desalting resin (ThermoFisher, Waltham, MA). The remaining sample was kept frozen at -80°C until the next day. Eluted sample was dried via speed vac and reconstituted in 6 uL of 0.1% formic acid. The entire sample was injected for LC-MS/MS analysis to check labeling efficiency and mixing ratios among the 14 labeled samples. After check analysis, the remaining sample from all digests were brought back to room temperature, quenched according to TMTpro protocol, and mixed according to normalized ratios determined from this check analysis (see below).

LC-MS/MS check analysis

Desalted sample was analyzed by nanoLC-MS/MS using a Dionex Ultimate 3000 (Thermo Fisher Scientific, Bremen, Germany) coupled to an Orbitrap Eclipse Tribrid mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). Six microliters (estimated 1ug) were loaded onto an Acclaim PepMap 100 trap column (300 um x 5 mm x 5 um C18) and gradient-eluted from an Acclaim PepMap 100 analytical column (75 um x 25 cm, 3 um C18) equilibrated in 96% solvent A (0.1% formic acid in water) and 4% solvent B (80% acetonitrile in 0.1% formic acid). The peptides were eluted at 300 nL/min using the following gradient: 4% B from 0-5 minutes, 4-10% B from 5-10 minutes, 10-35% B from 10-60 minutes, 35-55% B from 60-70 minutes, 55-90% B from 70-71 minutes, and 90% from 71-73 minutes.

The Orbitrap Eclipse was operated in positive ion mode with 2.1kV at the spray source, RF lens at 30% and data dependent MS/MS acquisition with XCalibur version 4.3.73.11. MS data acquisition was set up according to the existing method template, “TMT SPS-MS3 RTS”. Positive ion Full MS scans were acquired in the Orbitrap from 400-1600 m/z with 120,000 resolution. Data dependent selection of precursor ions was performed in Cycle Time mode, with 2.5 seconds in between Master Scans, using an intensity threshold of 5 x 10^3 on counts and applying dynamic exclusion (n=1 scans for an exclusion duration of 60 seconds and ∓10 ppm mass tolerance). Monoisotopic peak determination was applied and charge states 2-8 were included for CID scans (quadrupole isolation mode; rapid scan rate, 0.7 m/z isolation window, 32% collision energy, AGC standard). MS3 quantification scans were performed when triggered by the real-time search (RTS) algorithm. MS3 (HCD) scans were collected in the Orbitrap with 50,000 resolution, 50% collision energy, AGC target of 300%, and automatic maximum inject time mode for a maximum of 10 SPS precursors per cycle.

Sample preparation for targeting *TPM2* peptides

Replicates “A” and “C” of each of the following TMTpro-labeled samples were mixed according to normalized ratios determined from the check analysis: Days 0, 2, 4, 10. In addition, one technical replicate for 10°C was included, as well as one Jurkat sample (used as a negative control). An estimated 6.6 ug from each sample was mixed together, for a total of 66 ug. The pooled sample was subjected to C18 desalting using EasyPep Mini C18 desalting resin (ThermoFisher, Waltham, MA), reconstituted with 66 uL 0.1% formic acid for 1 ug/uL concentration, and used for tMS2 targeting analysis without further fractionation.

LC-MS/MS: DDA and tMS2 analysis

The resulting peptides were analyzed by nanoLC-MS/MS using a Dionex Ultimate 3000 (Thermo Fisher Scientific, Bremen, Germany) coupled to an Orbitrap Eclipse Tribrid mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). One microliter (estimated 1 ug) was loaded onto an Acclaim PepMap 100 trap column (300 um x 5 mm x 5 um C18) and gradient-eluted from an Acclaim PepMap 100 analytical column (75 um x 25 cm, 3 um C18) equilibrated in 96% solvent A (0.1% formic acid in water) and 4% solvent B (80% acetonitrile in 0.1% formic acid). The peptides were eluted at 300 nL/minute using the following gradient: 4% B from 0-5 minutes, 4- 28% B from 5-210 minutes, 28-50% B from 210-240 minutes, 50-95% B from 240-245 minutes and 95% B from 245-250 minutes.

DDA mode: The Orbitrap Eclipse was operated in positive ion mode with 2.1kV at the spray source and RF lens at 30% with XCalibur version 4.5. MS data acquisition was set up based on the existing method template, “TMT SPS-MS3 RTS”. The details of the method are identical to that described above, except that MS3 (HCD) scans were collected in the Orbitrap with 50,000 resolution, 50% collision energy, and AGC target of 300%.

tMS2 mode: A total of 22 peptides from source protein *TPM2* were selected for tMS2 targeting. Charge states z = 2 and z = 3 were selected for each peptide for a total of 44 target m/z values (**Table S3.7**). The Orbitrap Eclipse was operated in positive ion mode with 2.1kV at the spray source and RF lens at 30% with XCalibur version 4.5. In tMS2 mode, no MS1 scans were acquired and MS2 scans were isolated in the quadrupole with a

1.6 m/z isolation window and fragmented using HCD with 30% collision energy. Standard AGC was used and fragment ions were detected in the Orbitrap with 30,000 resolution. Retention time scheduling was not used.

Mass spectrometric data analysis

A) Database Searching for offline HPLC fractions

"DDA mass spec.raw" files were searched using the Proteome Discoverer software suite (PD 2.4.0.305) [259] with a processing workflow for SPS MS3 reporter ion quantification using SequestHT search algorithm and Percolator validation node. A sample-specific protein database, was generated as described in "Long-read RNA-seq analysis pipeline" (below) and used for Sequest HT searching. Precursor ion tolerance was set to 10 ppm and fragment ion tolerance was set to 0.1 Da. Static modifications of TMTpro (+304.207 Da) on peptide N-terminus and lysine residues as well as carbamidomethylation (+57.021 Da) on cysteine residues were used. The following dynamic modifications for peptides were used: oxidation (+15.995 Da) of methionine and phosphorylation (+79.966 Da) of serine, threonine, and tyrosine. The following static modifications for proteins were used: N-terminal acetylation (+42.011 Da), N-terminal methionine loss (-131.040 Da), and methionine loss plus acetylation at the N-terminus (-89.030 Da). A target FDR of 1% for peptide and protein validation was used.

B) Quantification

Impurity correction factors were included for reporter ion intensities, based on the quality control information provided with the TMTpro reagent kit. Reporter ion abundances above a minimum S/N of 2 and below a co-isolation threshold of 70% were summed across peptide spectral matches (PSM) to calculate peptide abundance. Normalization mode was set to "Total Peptide Amount" and scaling mode was "On All Average" so that the average abundance per protein and peptide was normalized to 100. Quantification rollup parameters were set to "Protein Abundance Based" protein ratio calculation, allowing for a maximum 100-fold change in abundance, with low abundance resampling imputation and ANOVA (individual proteins) hypothesis testing.

C) Manual inspection of targeted *TPM2* peptides

Skyline viewer program [260] was used to view and validate MS2 spectra for target peptides. For quantitative analysis, the ion counts of individual reporters ions were recorded for each target peptide MS2 spectrum using FreeStyle raw file viewer.

Long-read RNA-seq analysis pipeline

Long-read RNAseq processing was performed using the Isoseq3 workflow (https://github.com/PacificBiosciences/IsoSeq). Primer removal and demultiplexing of cells was performed using "lima". Next, the demultiplexed samples were refined by keeping only transcripts with poly-A-tails and removing any concatemers using the module "refine". Following that the full-length non-chimeric poly-A containing reads were clustered into isoforms using the module "cluster" and then aligned using the PacBio read compatible minimap2 [261] aligner. Finally the clusters of isoforms were collapsed into non-redundant isoforms using the "collapse" module. A raw matrix of expression was generated using cDNA Cupcake's "demux_isoseq_with_genome.py" module. Post Isoseq3, SQANTI3 [262] was used to classify isoforms into five categories: FSM (Full splice match), ISM (Incomplete splice match), NIC (Novel in catalog), NNC (Novel not in catalog), and genic. Following that, IsoAnnot was used to generate a full-length transcriptome reference from long-read data.

Long-read proteogenomics analysis pipeline

We relied heavily on the published long-read proteogenomics pipeline [115]. In short, open-reading frame identification steps used CPAT [207] to assess coding potential of isoforms, followed by generation of predicted ORFs using the "ORF_calling" module. Next, we generated the CDS GTF file to obtain the list of proteoforms using "make_pacbio_CDS" and "refinement" modules. We performed NMD and truncation analyses using Biosurfer (https://github.com/sheynkman-lab/Biosurfer_BMD_analysis). Briefly, each ORF containing an sQTL junction is compared against all ORFs not containing this sQTL and both length differences and the NMD rule are applied.

Experimental validation of *TPM2* in hFOBs

siRNA Knock Down; general information

hFOB cells used in the siRNA knockdown experiments were transfected within five days of thawing from liquid nitrogen storage. Briefly, cells were seeded at 75,000 cell/well of a 24 well plate and transfected within 24 hours after plating. In a typical experiment, 10 wells were seeded for each siRNA treatment (7 and a No Target control, see **Table S3.8**) plus a 'not transfected' control as well as 4 wells for RNA collected on the day of transfection (94 wells in toto). Four wells of each treatment were used for determining the amount of mineral formed at differentiation Day 10 with the remaining six wells used for RNA which was isolated at various times during the course of the experiment.

siRNA transfection procedure and conditions

Within 24 hours after plating, cells were transfected with siRNAs using Lipofectamine LTX reagent (Invitrogen ref# 15338100) following the manufacturer's recommended procedure, with minor modifications. Briefly, between 16-18 hours after plating, 2.5 ul Lipofectamine LTX was mixed with 37.5 ul pre-warmed Opti-MEM (Gibco, ref # 31985-070) per well of cells transfected. In parallel, 1 ul of 5 uM siRNA was mixed with 37.5 ul pre-warmed Opti-MEM per well. After a minimum of 5 minutes incubation time at room temperature (RT), the Lipofectamine mix and siRNA mix were combined, mixed and allowed to incubate for a minimum of 15 minutes at RT. After this time period, the ~75 ul Lipofectamine/siRNA mixture was diluted into 500ul pre-warmed Opti-MEM/well, mixed, and applied to a well of cells after growth media was removed. Cells were placed back in the incubator (temperature=34°C) for 5-7 hours at which time the Opti-MEM transfection mix was removed and replaced with 500 ul/well of pre-warmed growth media after the cells were washed with pre-warmed DPBS (Dulbeco's Phosphate Buffered Saline, Gibco, ref# 14190-144).

Osteoblast Differentiation

Twenty-four hours after transfection (48 hours after plating) growth media was removed, cells washed with 0.5 ml/well pre-warmed DPBS, 0.5ml/well pre-warmed differentiation media added and cells placed at 39.5°C. Differentiation media was replaced every other day after cells were washed with DPBS. On the tenth day, cells were washed 3x with 1 ml DPBS/well, fixed with 0.5 ml 10% Buffered Formalin Phosphate (Fisher #SF100-4) for 15 minutes at RT, washed three times with 1 ml/well water and stained with 400 ul/well 40 mM Alizarin Red, pH 5.6 (with NH4OH; Sigma A5533-25G) for 25 minutes at RT after which the stain was removed and cells washed 10x for 10 minutes each with 1 ml/well water. After images were scanned, the amount of alizarin red bound to the mineral formed was quantified by eluting in 2 ml/well 5% (v/v) Perchloric Acid (HClO4, Sigma-Aldrich 311413-500ML) and incubated for 20 minutes with shaking at RT. The amount of alizarin red bound for each sample/treatment was determined from the absorbance at 405 nm wavelength of the eluent along with standards prepared from the alizarin red staining solution appropriately diluted into 5% HClO4 and was expressed in units of nmol alizarin red bound.

RNA isolation and cDNA preparation

RNA was isolated using the Qiagen RNase minikit (cat# 74106) following the manufacturer's protocol. Briefly, media was removed from cells and washed 3x 1 ml/well DPBS and subsequently lysed in 400 ul/well RLT buffer containing 40 mM dithiothreitol (DTT) with mild shaking for 10 minutes, transferred to a microfuge tube and stored at -80°C until processing. At the time of processing, RNA was isolated from thawed lysates exactly as outlined. RNA was immediately DNased after isolation Applied Biosystem's TURBO DNA-free kit (Invitrogen ref# AM1907) following the manufacturer's protocol. DNased samples were stored -80°C. cDNA was prepared from thawed samples by initially determining the RNA concentration with a Qubit 4 fluorometer and the RNA HS assay kit (Thermo Fisher ref# Q33226 and Q32855, respectively) following the manufacturer's protocol. Random primed cDNA was synthesized from 1 ug DNased RNA using Applied Biosystems High Capacity Reverse Transcription kit (cat# 4368813).

Determining the extent and duration of siRNA knock down with real time PCR.

The relative abundance of different transcriptional isoforms of the Tropomyosin 2 gene (*TPM2*) was determined in duplicates for each sample/treatment from four separate experiments two days after siRNA transfection. Each reaction contained ~10 ng cDNA, 800 nM each primer (see **Table S3.9**), 0.5X Power Up SYBR Green Master mix (Applied

Biosystems ref# 100029285) in a 10ul reaction and amplified in a QuantStudio 5 Real-Time PCR system (Applied Biosystems, A28135) under these cycle conditions (50°C (2 minutes), 95°C (2 minutes); ([95°C (1 second), 60°C (30 seconds)] for 40 cycles); melt curve (95°C (1 second), 60°C (20 seconds), 95°C (1 second)). Relative quantification of the different transcriptional isoforms was determined by the 2 exp (–delta delta C(T)) method [263] using the geometric mean of the C(T) values of *CCDC47* and *CHMP2A* as the reference genes. Briefly, the cycle number in which half of the final amount of product produced (Cq/C(T)) is determined following the manufacturer's recommendations. The C(T) for each sample/reaction for each *TPM2* primer pair is subtracted from the geometric mean of the C(T) of primer pairs for the genes *CCDC47* and *CHMP2A* for the same sample (ΔC(T)). Finally, the delta C(T) for a given primer pair for the 'No Target Control' sample is subtracted from the siRNA knockdown samples C(T)s for that particular primer pair resulting in the ΔΔC(T) (ddCT). The reported values are $2^{(-)ddC(T)}$.

3.6 Acknowledgements

Research reported in this publication was supported in part by the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health under Award Number AR071657 to CRF, LCG, and EFM, and AA was supported in part by a National Institutes of Health, Biomedical Data Sciences Training Grant (5T32LM012416). We thank the IMPC for accessibility to BMD data in knockout mice (www.mousephenotype.org). We thank the donors in the Genotype-Tissue Expression (GTEx) Project and their families. GTEx 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. We also acknowledge Leon Sheynkman for his advice and technical knowledge.

Author Contributions: Abdullah Abood: Conceptualization; formal analysis; investigation; methodology; validation; writing – original draft; writing – review and editing. Larry Mesner: Data curation; formal analysis; investigation; methodology; visualization. Erin Jeffery: Data curation; formal analysis; investigation; methodology. Mayank Murali: Formal analysis. Micah Lehe: Formal analysis. Jamie Saquing: Consultation.

Gloria Sheynkman: Funding acquisition; writing – review and editing. Charles R Farber: Funding acquisition; writing – review and editing.

Data Availability Statement: The data that support the findings of this study are openly available in Gene Expression Omnibus (GEO) [GSE224588]. All analysis code is available on GitHub [https://github.com/aa9gj/Bone_proteogenomics_manuscript], relevant data is stored in Zenodo [DOI: 10.5281/zenodo.7603851]

Chapter 4

Concluding Remarks and Future Directions

The field of complex disease genetics surely is an exciting endeavor. A critical milestone in the field was the development of Genome-wide association studies (GWASs), which since their inception in 2005, have identified thousands of associations in hundreds of complex diseases and traits [49]. Optimism surrounding this technology at the time soon turned into a looming uncertainty as over 90% of the associations were found in intronic and intergenic regions suggesting a gene regulatory role, a more complicated consequence than preliminarily anticipated [31]. However, the benefits of this systematic method to discovery far outweighed technological parallels (i.e., linkage studies) in the field. Subsequently, cross disciplinary teams took on the challenge to follow-up on GWASs (dissecting GWAS associations), both computationally and experimentally, in order to pinpoint the causal genes within the associations. Herein lies my contribution to the field which involved identification of potentially causal genes impacting bone mineral density (BMD), the strongest predictor of osteoporotic fracture, and therefore osteoporosis.

The first GWAS interrogating BMD was published in 2008 [264], three years after the first ever GWAS [265]. Since then, scientists have identified hundreds of associations aimed at uncovering the genetic basis of osteoporosis, however, our understanding of the genes and mechanisms driving these genetic associations has been poor. Given that the majority of associations are implicated in gene regulation, follow-up studies integrated multi-omics data (described extensively in **Chapter 1**) to enhance our ability to systematically identify potential causal genes and subsequently validate them experimentally. My research focused on integrating transcriptomics data with already published GWAS data. At the start of my PhD, BMD GWAS follow-up had contributed improvements in our understanding of the genetic basis of BMD and osteoporosis. However, the focus of the majority of these follow-up studies was on protein-coding genes, turning a blind eye on major areas of research including a role for non-coding RNAs and the isoform-specific role of protein-coding genes in disease. In my work, I delved into these aspects of the transcriptome that have not been investigated in the context of osteoporosis.

In **Chapter 2**, we were interested in identifying a role for long non-coding RNAs (lncRNAs) as potential causal genes in osteoporosis. LncRNAs are transcripts longer than 200 nucleotides and have no coding potential [187]. The majority of lncRNAs share sequence features with protein-coding genes including a 3' poly-A tail, a 5' methyl cap, and an open reading frame [188]. However, their expression is low and heterogenous, and they show intermediate to high tissue specificity [189]. Aberrant expression of lncRNAs has been linked to diseases including osteoporosis [190]. Additionally, there is accumulating evidence suggesting their involvement in key regulatory pathways, including osteogenic differentiation [187,191]. We were able to integrate transcriptomics data and GWAS data in a small population of bone samples using Allelic Imbalance (AI) in genes found within 400 Kb of BMD GWAS associations. Due to the paucity of bone-relevant transcriptomics data, we were limited in our ability to apply molecular QTL methods. Therefore, we used AI to prioritize lncRNAs in bone fragment samples obtained from 17 patients undergoing hip replacement surgery. In conjunction, we performed expression quantitative trait loci (eQTL) Bayesian colocalization, and Transcriptome-Wide Association Study (TWAS) to identify potentially causal lncRNAs in the Genotype-Tissue Expression (GTEx) tissues, which do not contain bone relevant samples. Finally, we confirmed the expression of the lncRNAs resulting from both approaches in human fetal osteoblasts (hFOBs) during the process of differentiation into a fully mineralized cell. We applied this unbiased, systematic approach to discover a total of 23 lncRNAs, many of which are novel in the context of osteoporosis and bone biology and could potentially mediate the link between BMD and associations identified in BMD GWAS. Aside from the novelty, the work is based on human samples and includes validation of lncRNAs expression in osteoblast cultures which makes the conclusions highly relevant for possible translation. However, this study was not meant to be comprehensive because we were limited by the number of samples and are not suitably powered to identify eQTLs and apply TWAS/colocalization analysis. Nonetheless, due to the scarcity of population-level bone transcriptomic datasets, and the lack of bone cell or tissue data in GTEx, our study is an attempt to systematically leverage the available datasets to capture a subset of lncRNAs that we think are potentially causal. Some of these lncRNAs have been implicated experimentally outside of this study. Moreover, lncRNAs under AI and within proximity of GWAS loci may not be causal as they could be false positives because they are not prioritized via a systems analysis such as colocalization. Another limitation of our study is that we evaluated their expression as a function of osteoblast differentiation; however, it is likely that some of the lncRNAs, if truly causal, impact BMD via a function in other cell-types (eg, osteoclasts).

In **Chapter 3**, we investigated alternative splicing (AS) in the context of GWAS associations. One major task is to understand the functional consequences of AS by identifying how AS affects pathways and processes that impact human disease. A wealth of GWAS and sQTL functional genomics data exists among hundreds of complex traits, and it has become clear that AS variation has an effect on patients' disease risk. The field has yielded vast catalogs of splicing Quantitative Trait Loci (sQTLs), with some characterization of the mechanisms of the implicated Single Nucleotide Polymorphisms (SNPs, e.g., effect of SNP on RNA binding proteins or splice donor/acceptor modification), but knowledge of the downstream transcript and protein isoforms affected remain unknown. We presented a novel method that should increase the utility and interpretability of sQTLs colocalized to disease. Our method integrates sQTL analysis directly with long-read RNA-seq data, and a recently developed "long-read proteogenomics" pipeline [115], to identify full-length functional protein isoforms impacted by colocalized sQTLs, and thus the molecular implications of altered splicing in patients. We demonstrated this approach on BMD, but our work should serve as a model for other areas of research to increase the clinical utility of sQTL analysis across the disease spectrum. This includes usage of sQTL data from publicly available resources like GTEx or disease-relevant cohorts.

In this study, we identified a comprehensive list of full-length isoforms for BMD as potential effector in osteoporosis. More importantly, we did this by leveraging sQTL junctions generated from short-read RNAseq data and interpreting their effects on full-length isoforms from an independent long-read RNAseq sample in a disease relevant model. To date, we are unaware of studies that systematically investigated the role of sQTLs in bone mass. Our work shows that splicing is a major mechanism influencing BMD. Alternative splicing tends to be a distinct mechanism, compared to gene expression, with little overlap between biological and genetically regulated splicing. Indeed, we identified many genes with colocalizing sQTLs that do not show any enrichment in known bone monogenic disease genes, International Mouse Phenotyping Consortium (IMPC) BMD relevant genes, or known bone processes, demonstrating that studies focused on detecting isoform-driven signals are uncovering new biological and potentially clinical processes.

One of the more exciting findings from our study is the gene *TPM2* and its isoforms. *TPM2* splicing has not been directly studied in the context of osteoblast differentiation or bone disease, however, studies have implicated splice isoforms of the gene in muscular diseases like the Escobar variant (characterized by skeletal defects including vertebral defects, bone fusion abnormalities and growth retardation) [250], nemaline myopathy [251], and atherosclerosis [252]. Additionally, There is evidence from the literature suggesting that loss of Tpm2.1 (isoform containing exons 6 and 11 together) leads to increase in beta catenin levels [253] which is integral to bone formation [254]. In our results, isoforms containing exon 6 and exon 7 show opposing effects on hFOB mineralization, however, we should also note that knocking down *TPM2* isoforms containing exon 11 have shown a significant decrease in mineralization. One possible explanation is that the ratio of two isoforms (i.e. those containing exons 6 and 11 with those containing exons 7 and 11) and not the absolute amount is necessary for the cells to mineralize.

We found several lines of evidence showing convergence of BMD sQTLs on both trans-acting splice factors as well as their cis-regulatory targets. For example, we found that both *PTBP1* (a splice factor) and *TPM2* harbor strongly colocalizing sQTLs and are potentially linked via regulatory and biochemical interactions. Studies have shown that *PTBP1* isoforms containing exon 9 bind the unprocessed *TPM2* transcript in intron 6 resulting in higher levels of exon 6 containing isoforms [238,241]. Conversely, higher levels of *PTBP1* isoforms that lack exon 9 has reduced repressive activity and leads to derepression of exon 7 and thus higher levels of exon-7-containing TPM2. Intriguingly, the genetic associations indicate that higher levels of exon-9-containing *PTBP1* (lead SNP: rs2737273) and higher levels of exon-6-containing *TPM2* (lead SNP: rs3215700), independently, are associated with lower BMD, suggesting a core splicing axis involving *PTBP1* and *TPM2*. The lead sQTL within *TPM2* (in intron 6) is a SNP that extends the polypyrimidine tract by a single nucleotide which potentially leads to reduced PTBP1 binding, suggesting this SNP might be the causal SNP in that region. The mean H_4PP (Probability of colocalization) in events where rs3215700 is the lead SNP is 0.94 (median 0.97) while the mean H_4PP of other lead SNPs for the junctions surrounding exons 6 and 7 is 0.90. Furthermore, unpublished data provide evidence for this lead SNP to have a high impact in osteoblasts. Overall, future studies interested in this splicing network should consider functionally annotating this SNP.

We speculate that this approach can be extended to integrate other GWAS follow-up methods such as Transcriptome-wide association studies (TWAS) which was extensively described in **Chapter 1**. Nonetheless, we understand that our study has limitations. We were not statistically powered to identify sQTLs in a bone relevant tissue, therefore we inferred the effect of shared sQTLs in different tissues on a bone relevant cell line. Ideally, population studies should leverage long-read RNAseq to identify isoforms ratios associated with genotypes rather than junctions. However, our approach takes advantage of the wealth of publicly available junction sQTL data. At a molecular level, we were able to validate our hypothesis and suggest a mechanism for *TPM2* isoforms leading to an opposing effect on mineralization, but we were not able to fully validate this mechanism experimentally. We consider our work a resource to generate and validate similar hypotheses to provide more evidence for an isoform specific role in osteoporosis.

Future directions

The osteoporosis community has been taking baby steps, relative to other fields (i.e., brain & cardiovascular), into the generation of a bone-specific population level transcriptome. While the progress has been very slow, it might be wise to take advantage of the limited scope of discovery produced by consortia (such as GTEx) using bulk transcriptomics. For example, bulk transcriptomics hindered any progress in identifying cell-type specific

effects on complex disease and/or isoform level implications. Moreover, we are currently in the single-cell sequencing (2019 method of the year [266]) and long-read sequencing (2022 method of the year [267]) revolutions. This yields a great opportunity for the osteoporosis community to make a leap into the present [111,268] by generating a single-cell long-read based population level transcriptomics. As long-read technologies rapidly improve in throughput and cost, their use in population genetics studies, either alone or in an integrative strategy, will likely increase. For example, current technologies like PacBio MAS-seq can deliver the potential of the single-cell-isoform defined transcriptome. Additionally, this technology can open the door wide into knowledge of the repertoire of expressed protein isoforms and could also guide drug development.
Studies have shown that drug targets against proteins with underlying genetic evidence are twice as likely to succeed in clinical trials [25].

Although precision medicine holds great promise, it currently faces several gaps in research that prevent it from reaching all potential patients. One major obstacle is the lack of diversity among participants in biomedical research [269]. This limitation reduces the generalizability and availability of genomic-based treatments and prevention strategies. The serious under-representation of diverse populations in genetic/genomic studies is highly problematic since genetic information obtained from one population may not be applicable to other populations [270]. This is due to differences in linkage disequilibrium (LD), allele frequencies, and genetic architecture. Without a diverse sample, important signals revealing powerful insights into genetic associations and drug response may go unnoticed. I believe the research presented in this dissertation can be extended to begin addressing this issue as the field of personalized medicine is incomplete without meeting medical needs of diverse individuals. As long-read technologies rapidly improve in throughput and cost, their use in population genetics studies, either alone or in an integrative strategy, will likely increase. In the near term, integrative strategies that combine long-read and short-read data will likely remain popular. Therefore, our strategy is to leverage resources such as GTEx or sQTL Compendia, which provides candidate sQTLs in normal tissues and determine how such sQTLs may map onto full-length isoforms expressed in representative disease models (cell-line or tissue). This integrative in silico/in vitro approach could allow for inference of the effects of sQTLs that are detected in an independent population or meta-analysis (in which there is sufficient sample size and power), but are placed within the isoform-relevant context of the disease model. Currently, short-read sequencing is approximately 10-fold higher in throughput at the same cost, but due to developments such as MAS-Iso-Seq (PacBio) and the PromethION system (ONT), long-read data may become comparable in throughput and cost in the next 5 years. Long-read pipelines and long-read/short-read integrative strategies are maturing, with recent consortia such as the Long Read Genome Annotation Consortium conducting comprehensive comparisons across model organisms, library preparations, platforms and bioinformatic pipelines. Strategies to integrate long-read RNA-seq with proteomics are also emerging, pointing to the potential to obtain pQTL data that are protein isoform-resolved. Long-read RNA-seq at the single cell level could also open the door toward isoform-resolved sQTLs that are specific to certain cellular contexts and populations.

Finally, "good science" comes through extensive collaboration and biomedical science is not an exception to this rule. It should go without saying that the field of computational biology and biological data science has made major strides and contributed major improvements to our biological understanding of disease and discovery. However, these discoveries must go hand in hand with extensive collaborations at the bench. The need for experimental validation should be a top priority for all computational biologists, and it comes with its own experimental design at the top. A computational biologist should provide a hypothesis, a feasible target for experimental validation, and should work closely with the bench scientist to address any issues that arise during the process. This includes generating visualizations, analyzing data (including qPCR), and interpreting the results correctly.

Overall, our work was able to identify candidate biomarkers and therapeutic drug targets for osteoporosis. This PhD work provides different methodologies for data integration to generate hypotheses and validate them in the context of bone disease, a growing field of discovery.

Main figures and tables

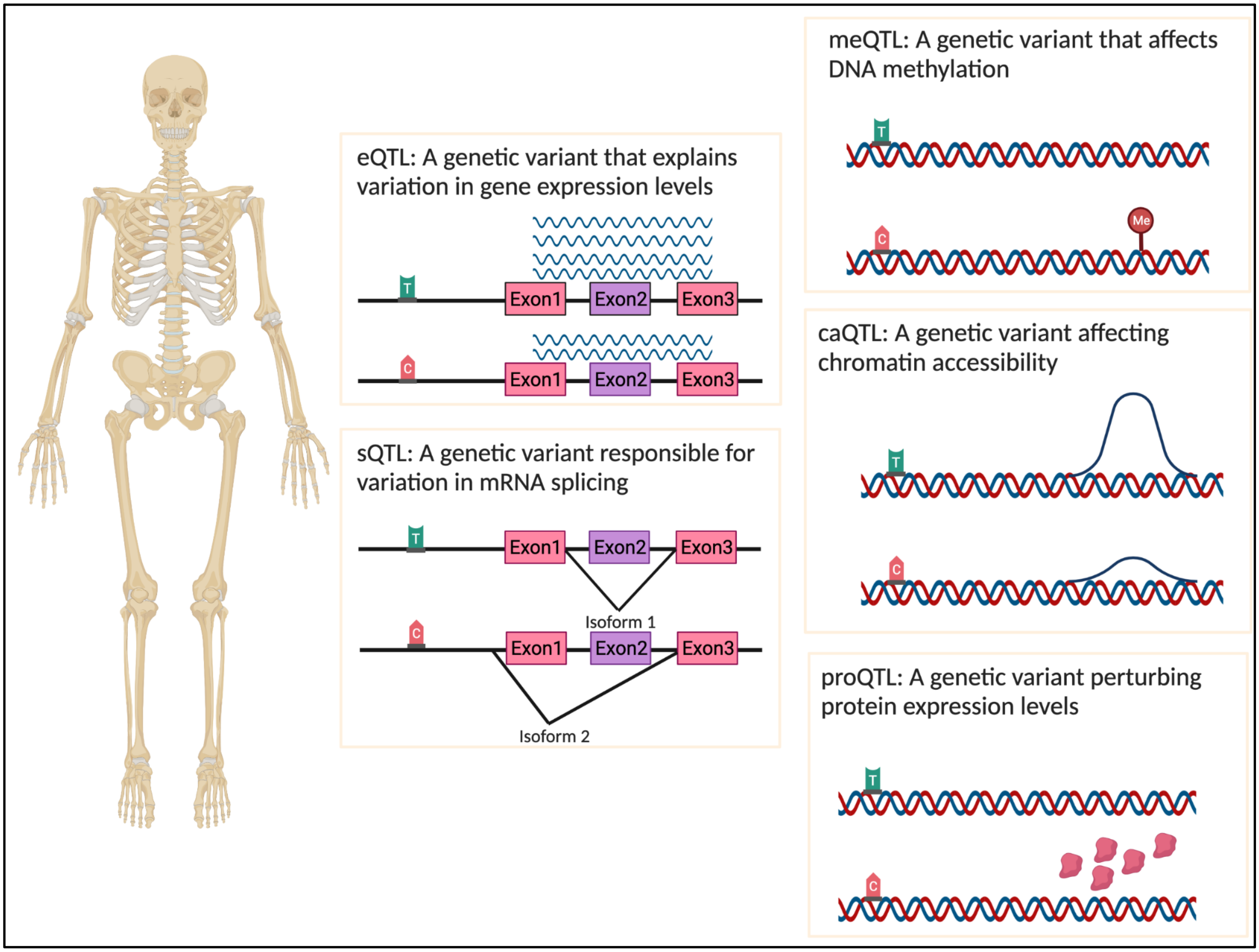

eQTL: A genetic variant that explains variation in gene expression levels
T
C
Exon1
Exon2
Exon3
sQTL: A genetic variant responsible for variation in mRNA splicing
Isoform 1
Isoform 2
meQTL: A genetic variant that affects DNA methylation
Me
caQTL: A genetic variant affecting chromatin accessibility
proQTL: A genetic variant perturbing protein expression levels

Figure 1.1: Examples of quantitative trait loci (QTL) for molecular phenotypes such as gene expression (eQTL), alternative splicing (sQTL), DNA methylation (meQTL), chromatin accessibility (caQTL), and protein expression (proQTL). QTL approaches can generally be applied to any molecular trait quantifiable in a population of individuals.

Approach	**Description**	**References**
Genome-wide association study (GWAS)	Hypothesis-free statistical approach that identifies associations of genetic variants and diseases or disease-associated traits. GWASs are divided into two types: case-control and quantitative	[13,16,37]
Transcriptome-wide association study (TWAS)	Statistical approach that leverages gene expression imputation to identify significant gene-trait associations by estimating the genetic component of gene expression in a reference population where gene expression and genotype have been measured (GTEx is an example) and then imputing (predicting) gene expression in a much larger population (such as those used in a BMD GWAS). Once gene expression is imputed, genetically regulated gene expression is associated with a disease or disease phenotype	[138,143]
Colocalization	Statistical test to determine whether a single variant is responsible for both GWAS and molecular QTL signal (i.e., eQTL) in a genomic region. One can draw a parallel to fluorescence microscopy where colocalization refers to the observation of overlap between different fluorescent labels that tag different "targets" located in the same area of the cell	[29,143]
Network analysis	Approaches using "-omics" data (e.g., RNA-seq, proteomics, etc.) to partition genes into groups based on functional similarities in an unbiased manner. A growing number of studies have demonstrated the ability of networks to predict causal genes at GWAS loci	[144,145,151]

Table 1.1: Defining concepts used in this introduction

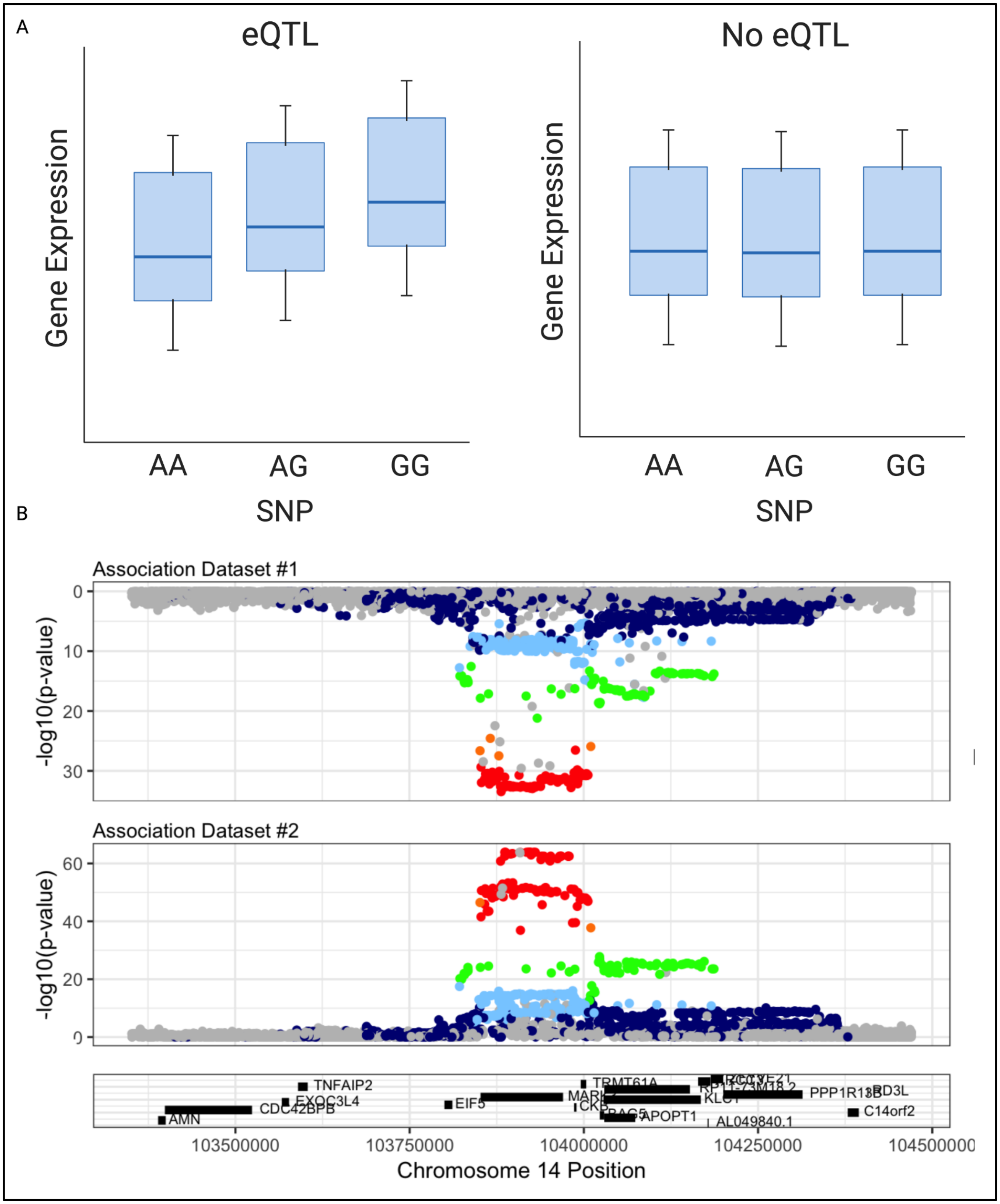


Figure 1.2: eQTL discovery and colocalization. A Two examples of an eQTL with box plots showing that gene expression is (left) or is not (right) correlated with the genotype of a single-nucleotide polymorphism (SNP). B An example of a colocalizing eQTL for *MARK3* visualized using RACER [271]. Every circle represents a SNP. An eQTL for *MARK3* is shown on the top panel of the mirror plot. The BMD GWAS association is plotted on the bottom panel. Note that the same SNPs associated with BMD are associated with the expression of *MARK3*. The colors signify r^2, a measure of linkage disequilibrium.

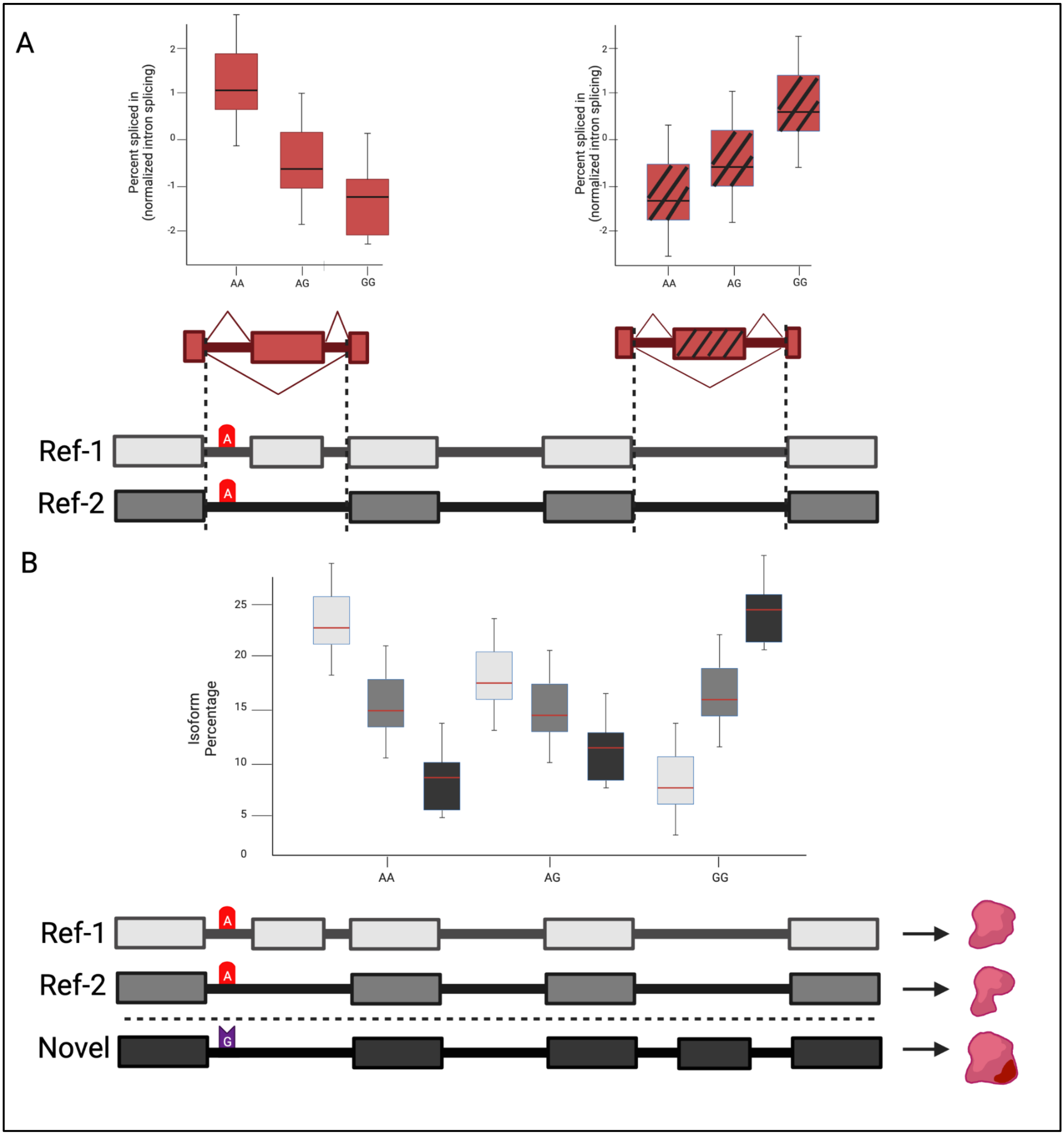


Figure 1.3: Long-read sequencing provides isoform-level characterization of sQTL effects. To illustrate, a hypothetical example may be considered. Panel A shows the event-based characterization of two exon inclusion/exclusion events, one of which involves a cryptic exon (represented by the dashed lines). The first event can be explained by the reference transcript annotation, but the second event indicates the presence of a novel isoform and the identity of this isoform cannot be defined from short-read data alone. Panel B shows the results of long-read sequencing which identify the novel isoform. The pattern of isoform usage by genotype confirms that this pattern of exon/inclusion events is driven by increased usage of the novel isoform in subjects

with the G allele of the causal sQTL variant. Proper characterization of the isoform also provides more accurate information on protein sequence and functional potential (as seen in different shapes of the pink proteins).

Disease/Trait	Year	Sample size	Tissue	Junction/isoform Quantification	sQTL calling	Study
Various traits	2022	GTEx	GTEx (49 tissues)	LeafCutter*	FastQTL + PEER*	Rouhana et al. [272]
Various neurological and psychiatric disorders	2022	100	255 primary human microglial samples from multiple brain regions	LeafCutter	TensorQTL	Lopez et al. [273]
Alzheimer's disease and related dementias	2022	Various including GTEx	Various including GTEx	Previously reported	Previously reported	Bellenguez et al. [274]
Pancreatic Cancer	2022	TCGASpliceSeq ($N = 176$)	Pancreatic ductal adenocarcinoma (PDAC)	SpliceSeq	MatrixeQTL + regression analysis	Tian et al. [275]
Various complex traits	2022	N/A	Various immune cells	In-house package	N/A	Yamaguchi et al. [276]
BD (Bipolar Disorder)	2022	511 total samples from 295 unique donors	Subgenual anterior cingulate cortex and amygdala samples	LeafCutter	FastQTL + PEER	Zandi et al. [277]
Cardiometabolic traits	2022	426 Finnish men from the METSIM study	Subcutaneous adipose tissue	LeafCutter	QTLtools + PEER	Brotman et al. [224].

Prostate Cancer (PrCa)	2022	467	Benign prostate tissue	RSEM + sQTLseekeR	sQTLseekeR	Tian et al. [278]
Coronary Artery Disease (CAD)	2022	151	Cultured smooth muscle cells	LeafCutter	FastQTL + PEER	Aherrahrou et al. [279]
Various brain related traits	2021	PsyENCODE cohort ($N = 1073$)	Brain	THISTLE + LeafCutter	FastQTL + PEER	Yang et al. [91]
Meta-analysis of Various traits	2021	Varied	Varied	LeafCutter	QTLtools + PEER	Kerimov et al. [280]
Developing cortical wall or adult cortex	2021	Primary human neural progenitors ($n = 85$) and their sorted neuronal progeny ($n = 74$)	Primary human neural progenitors and their sorted neuronal progeny	LeafCutter	EMMAX	Aygun et al. [281]
Meta-analysis of multiple cohorts for human immune traits	2021	Multiple cohorts: 1) DGN: $N = 922$, 2) BLUEPRINT: $N = 197$, 3) GEUVADIS $N = 462$	Various immune cells	LeafCutter	FastQTL + PEER	Mu et al. [282]
Various traits	2021	GTEx	GTEx (All tissues)	sQTLseekeR2	sQTLseekeR2	Garrido-Martín et al. [48]
Various traits	2021	GTEx	GTEx (49 tissues)	LeafCutter*	FastQTL + PEER*	Barbeira et al. [47]
Type 2 diabetes (T2D)	2021	GTEx	GTEx (48 tissues)	LeafCutter*	FastQTL + PEER*	Chen et al. [283]

Kidney function	2021	GTEx	GTEx (Kidney)	LeafCutter*	FastQTL + PEER	Stanzick et al. [284]
Type 1 Diabetes (T1D)	2021	Genotype-Tissue Expression (GTEx)	GTEx (All tissues)	LeafCutter*	FastQTL + PEER	Gao et al. [285]
Glioma	2021	CommonMind Consortium (CMC) and GTEx	Multiple brain tissues	LeafCutter	Matrixeqtl + PEER	Patro et al. [286]
Amyotrophic lateral sclerosis (ALS)	2021	154 ALS cases and 49 control individuals	Cervical, thoracic, and lumbar spinal cord segments	LeafCutter	TensorQTL + PEER	Humphrey et al. [249]
Complex disease in Colon	2021	485	Colonic mucosal biopsy	LeafCutter	FastQTL + PEER	Díez-Obrero et al. [287]
Parkinson’s disease (PD)	2021	230	Monocytes	LeafCutter	QTLtools + PEER	Navarro et al. [288]
Mental illness (bipolar disorder, schizophrenia, major depression)	2021	200	Postmortem subgenual anterior cingulate cortex (sgACC)	SQTLseekeR	sQTLseekeR	Akula et al. [289]
Schizophrenia	2021	151	Prefrontal cortical samples	LeafCutter	QTLtools + PEER	Liu et al. [290]

Melanoma	2021	106	Human primary melanocytes	LeafCutter	FastQTL + PEER	Zhang et al. [291]
Aging human brain	2020	Religious Order Study (ROS) and Memory and Aging Project (MAP) cohorts (*N* = 450)	Brain	LeafCutter*	FastQTL + PEER*	Yang et al. [292]
Chronic obstructive pulmonary disease (COPD)	2020	GTEx + Lung Tissue Research Consortium (LTRC)	GTEx (Lung) + LTRC	LeafCutter	FastQTL + PEER	Saferali et al. [114]
Bladder cancer	2020	580 cases/1101 controls (GTEx, TCGA, GEO, CancerSplicingQTL, 1000 Genomes Project)	Bladder	LeafCutter and SpliceSeq*	FastQTL + PEER + sQTLSeekeR*	Guo et al.[293])
Cancer	2020	19 257 cases and 30 208 controls (71 studies from 52 publications)	Various tissues	LeafCutter*	FastQTL + PEER*	Yuan et al. [294]
CAD, stroke, migraine, abdominal aortic aneurysm	2020	19 paired primary human coronary artery smooth muscle and endothelial cells	HCASMCs and HCAECs	MAJIQ	In-house regression analysis	Nurnberg et al. [295]
Various traits	2020	838	Various tissues	LeafCutter	FastQTL + PEER	GTEx consortium [31]

Parkinson’s disease (PD)	2019	ROS + MAP+CMC (*N* = 902)	Brain	LeafCutter*	FastQTL + PEER*	Li et al. [296]
Immune activation	2019	970 RNA-seq from 200 individuals of African- and European-descent	Resting and stimulated monocytes	LeafCutter	MatrixeQTL + PEER	Rotival et al. [297]
Chronic obstructive pulmonary disease (COPD)	2019	376	Whole Blood	LeafCutter	MatrixeQTL + PEER	Saferali et al. [298]
Schizophrenia	2019	201	Mid-gestational human brains	LeafCutter	FastQTL + PEER	Walker et al. [75]
Cardiovascular disease	2019	83	Induced pluripotent stem cell (iPSC), hepatocyte-like cell (HLC), primary liver tissues	LeafCutter	QTLtools + PEER	Gawronski et al. [299]
Alzeheimer’s disease (AD)	2018	450	Dorsolateral prefrontal cortex (DLPFC)	LeafCutter	FastQTL + PEER	Raj et al.[300]
Coronary Artery Disease (CAD)	2018	52	HCASMC	LeafCutter	FastQTL + PEER	Liu et al. [301]

Table 1.2: An overview of sQTL studies that examine GWAS loci in the context of complex traits (published 2018–2022). * denotes previously reported sQTL dataset.

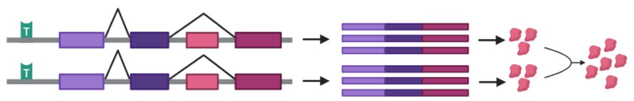
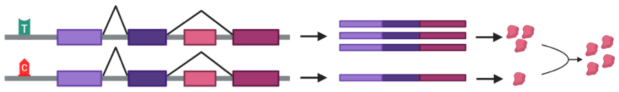
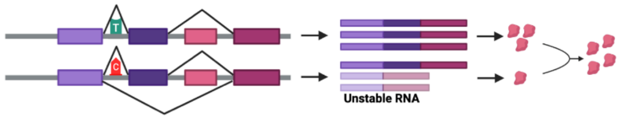
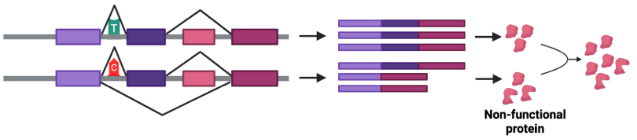
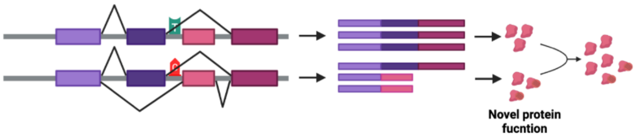
Scenario schematic (Genome → Transcripts → Protein isoforms → Cumulative protein output for the gene)	**Scenario description**	**QTL type**	**Gene-level effect**
Wild-type gene function (i.e., physiologically normal activities)			
	Wild type	No QTL	Physiologically normal
eQTL - Change in transcriptional abundance (non-sQTL)			
	Variant (allele) leads to down-regulation of transcription.	eQTL	Reduced protein output
sQTL - attenuation of the protein functional capacity of the gene			
	Alternative splicing can lead to an unstable mRNA or mRNA subjected to nonsense mediated decay, in which the protein is not translated at appreciable levels.	eQTL and/or sQTL	Reduced protein output
	Protein isoform is translated and properly folded, but has lost some molecular activity, such as an affected enzymatic site or binding affinity of a transcription factor to its target sites.	sQTL	Reduced protein activity and/or output
sQTL - augmentation or alteration of the protein functional capacity of the gene			
	Protein isoform exhibits different or novel functions (gain or loss of functional domains, protein binding partners, enzymatic activity, localization, etc.). There is also the possibility of increased protein isoform stability to produce more protein.	sQTL	Gene has a changed or gain of function
Note that the relationship between protein isoform functional differences and overall gene-level function can be non-linear. For example, a risk isoform can have attenuated molecular function, compared to its wild-type counterpart, but, the overall effect on gene function can be more extreme, such as in the case of a dominant negative (88). In other cases, the attenuated function of an isoform has a small effect because it relates to a subtle shift in splicing ratios, and the overall effect on gene-level function is subtle, playing a “fine tuning” role. Also, note that sometimes the loss of amino acid sequence leads to a gain of function, such as when an autoinhibitory domain is lost.			

Figure 1.4: Protein molecular consequences of eQTLs and sQTLs

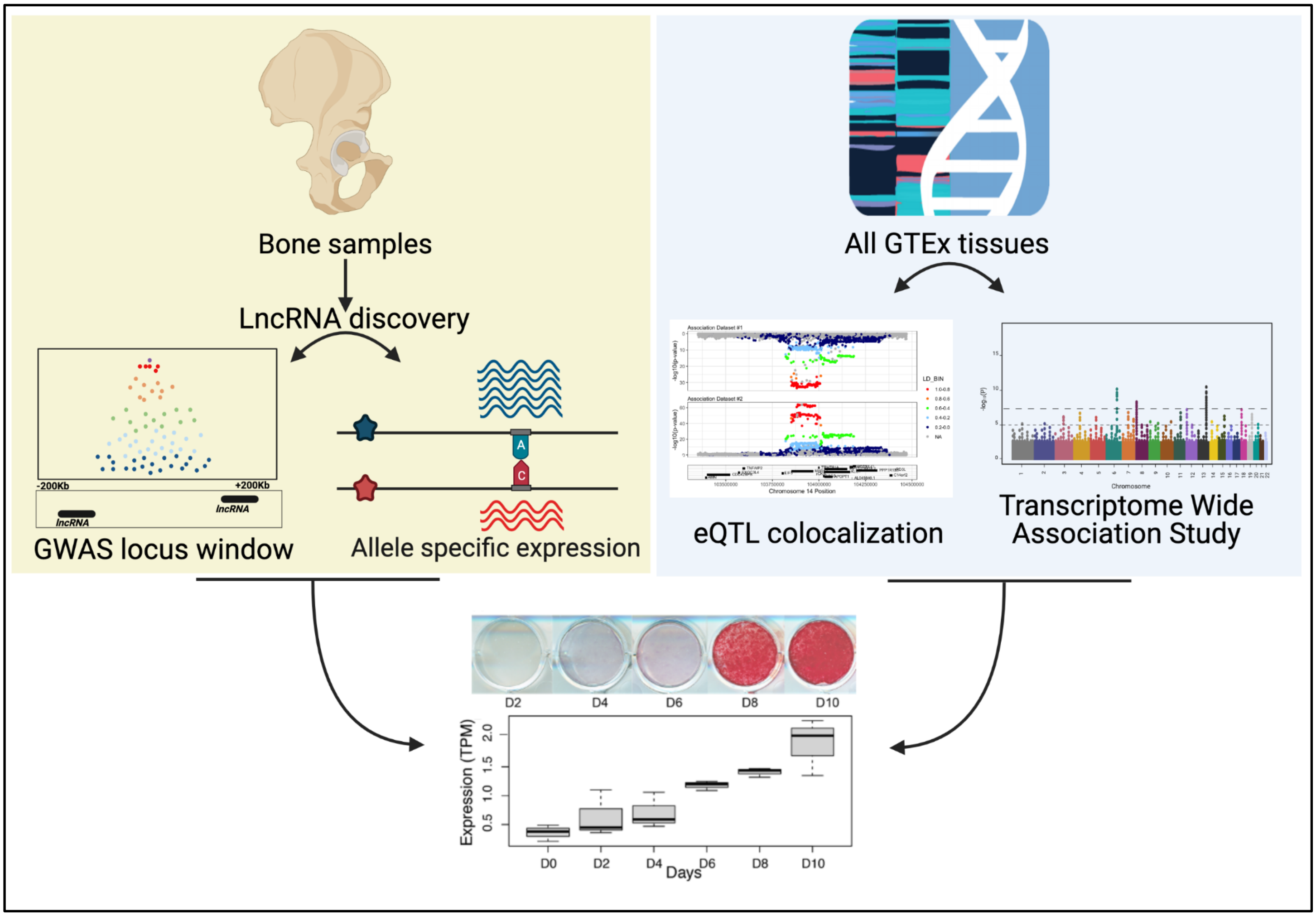


Figure 2.1: Overview of the study. We conducted de novo lncRNA discovery using RNAseq data on human acetabular bone fragments from 17 patients. We then identified known and novel lncRNAs located in GWAS associations that were influenced by AI (yellow box). We applied TWAS and colocalization on eQTL data from 49 GTEx project tissues (blue box). We assessed the role of lncRNAs reported by both approaches in osteogenic differentiation using RNAseq data from the hFOB cell line at six time points across differentiation (bottom panel). AI = allelic imbalance; GTEx = genotype-tissue expression; hFOB = human fetal osteoblast; TWAS = transcriptome-wide association study.

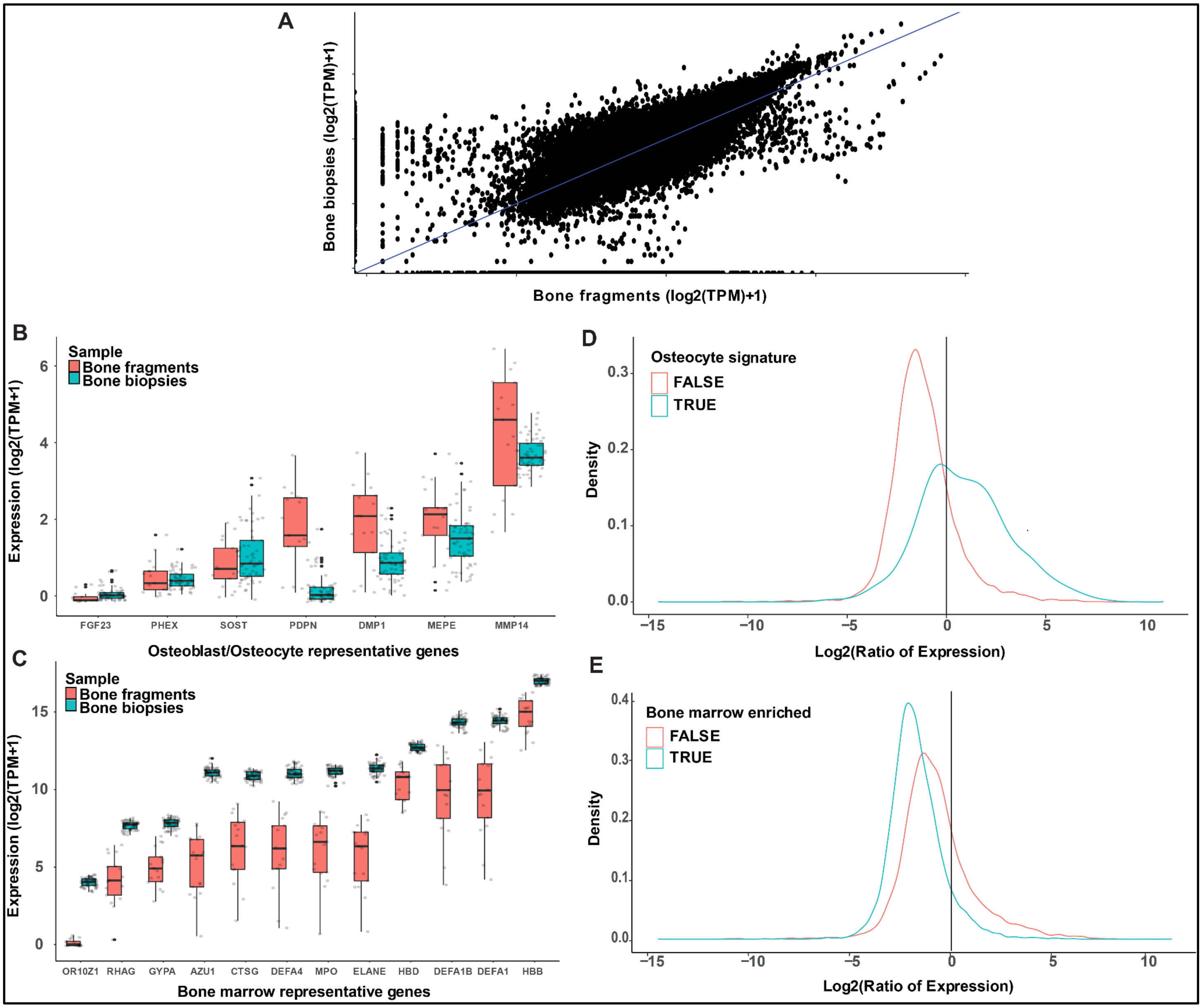


Figure 2.2: Enrichment of osteocyte marker genes in bone fragment samples (used in this study) compared to bone biopsy samples in the literature. (A) Overall gene expression is highly correlated between the RNAseq data generated in both studies ($r^2 = 0.845$, $p < 2.2 \times 10^{-16}$); Farr and colleagues [201] (B) Gene expression of osteocyte marker genes reported in Bonewald [211] showing enrichment in the bone fragments samples (this study) relevant to bone biopsies. (C) Gene expression of bone marrow enriched genes reported in The Human Protein Atlas (www.proteinatlas.org/) showing higher expression in bone biopsy samples. (D) Osteocyte signature genes reported in Youlten and colleagues [212] are highly expressed in bone fragment samples relative to bone biopsies (Wilcoxon test, $p < 2.2 \times 10^{-16)}$ (E) Bone marrow enriched genes reported in Youlten

and colleagues [212] are highly expressed in bone biopsy samples compared to bone fragment samples (Wilcoxon test, $p < 2.2 \times 10^{-16}$).

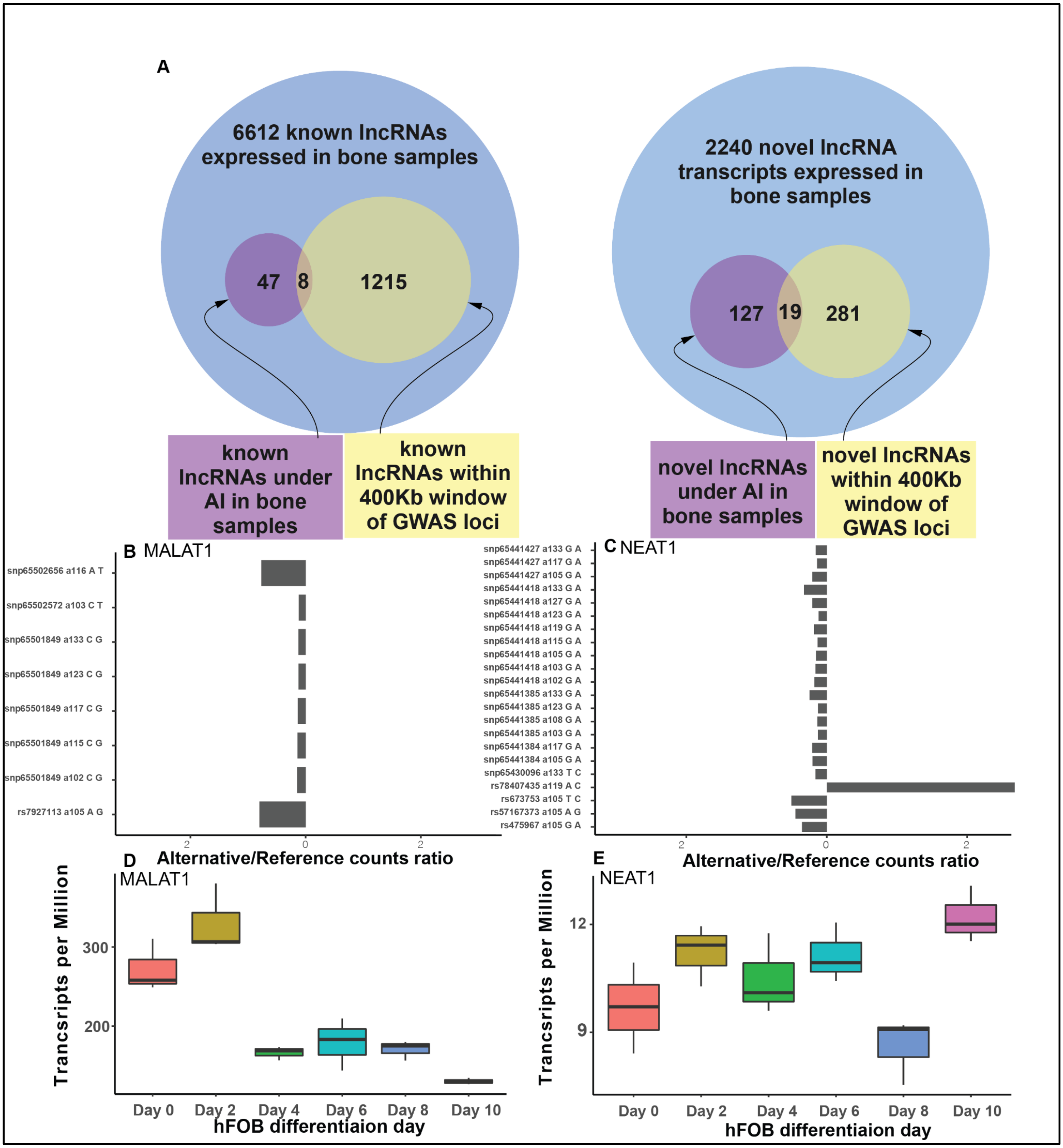


Figure 2.3: Identification of lncRNAs located within eBMD GWAS associations, are under AI in acetabular bone, and are differentially expressed in hFOBs. (A) Venn diagram showing the number of known and novel lncRNAs within proximity of GWAS loci, implicated by AI, and implicated by both approaches. (B) lncRNA *MALAT1* AI plot showing the ratio of reads aligning to the alternative SNP relative to the reference SNP in eight of the bone fragments samples where the gene is under AI. (C) lncRNA *NEAT1* AI plot showing the ratio of reads aligning to the alternative SNP relative to the reference SNP in 10 of the bone fragments samples where the gene is under AI. rs78407435 is not in LD with the rest of the SNPs in the region and this is likely the reason it shows a different direction of effect. (D) Expression of MALAT1 across hFOB differentiation points. (E) Expression of *NEAT1* across hFOB differentiation points. AI = allelic imbalance; hFOB = human fetal osteoblast.

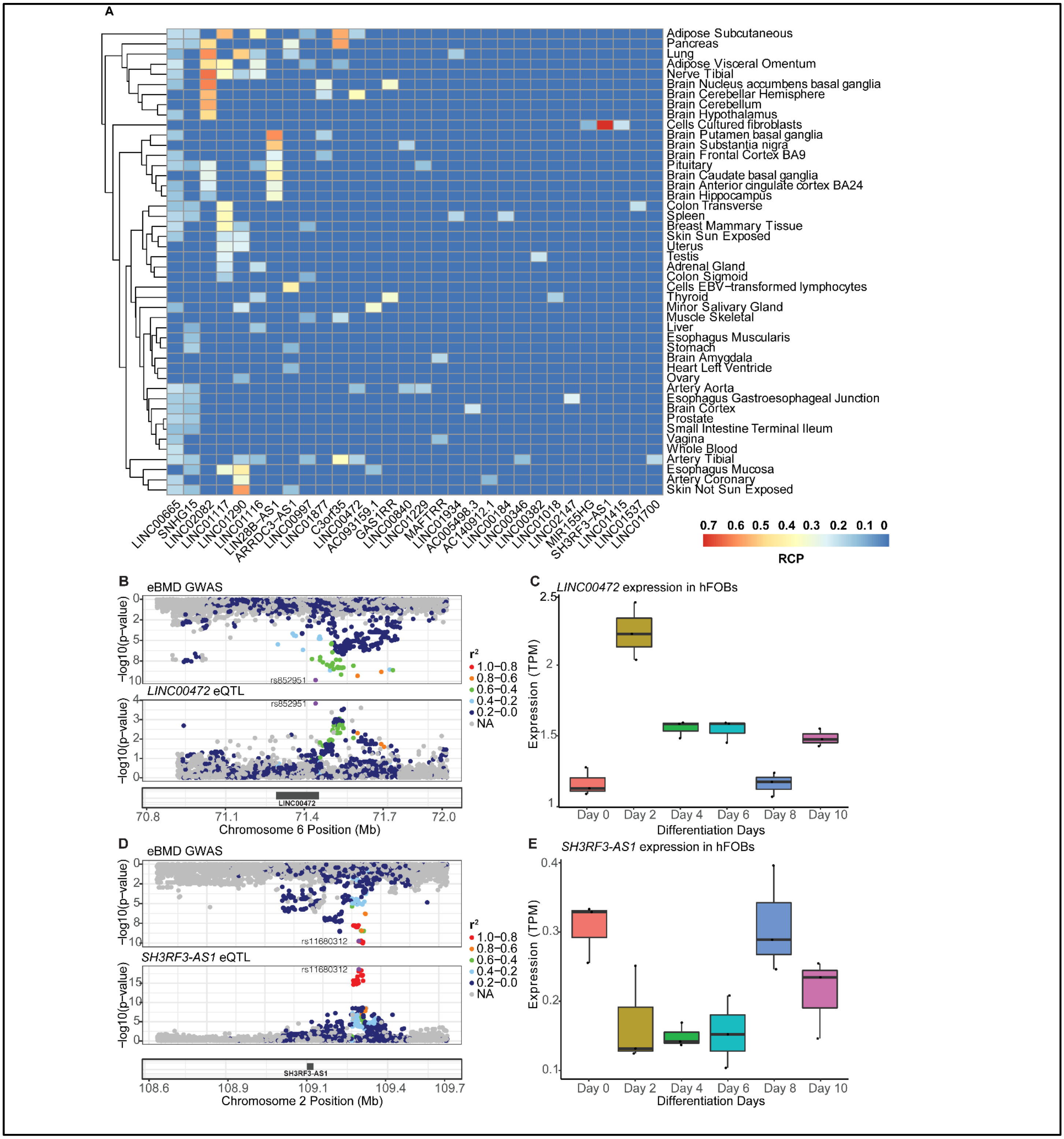


Figure 2.4: lncRNAs implicated by eQTL colocalization and TWAS are potential effector transcripts of BMD GWAS loci. (A) Heat map showing colocalization events in GTEx tissues. (B) lncRNA *LINC00472* colocalization plot showing colocalization of eBMD GWAS locus with eQTL from brain cerebellar hemisphere with RCP of 0.37 (C) Differential expression of *LINC00472* across hFOB differentiation points (D) lncRNA *SH3RF3-AS1* colocalization plot showing colocalization of eBMD GWAS locus with GTEx fibroblasts eQTL

data with RCP of 0.72 (E) Differential expression of *SH3RF3-AS1* across hFOB differentiation points. hFOB = human fetal osteoblast. RCP = Regional colocalization probability.

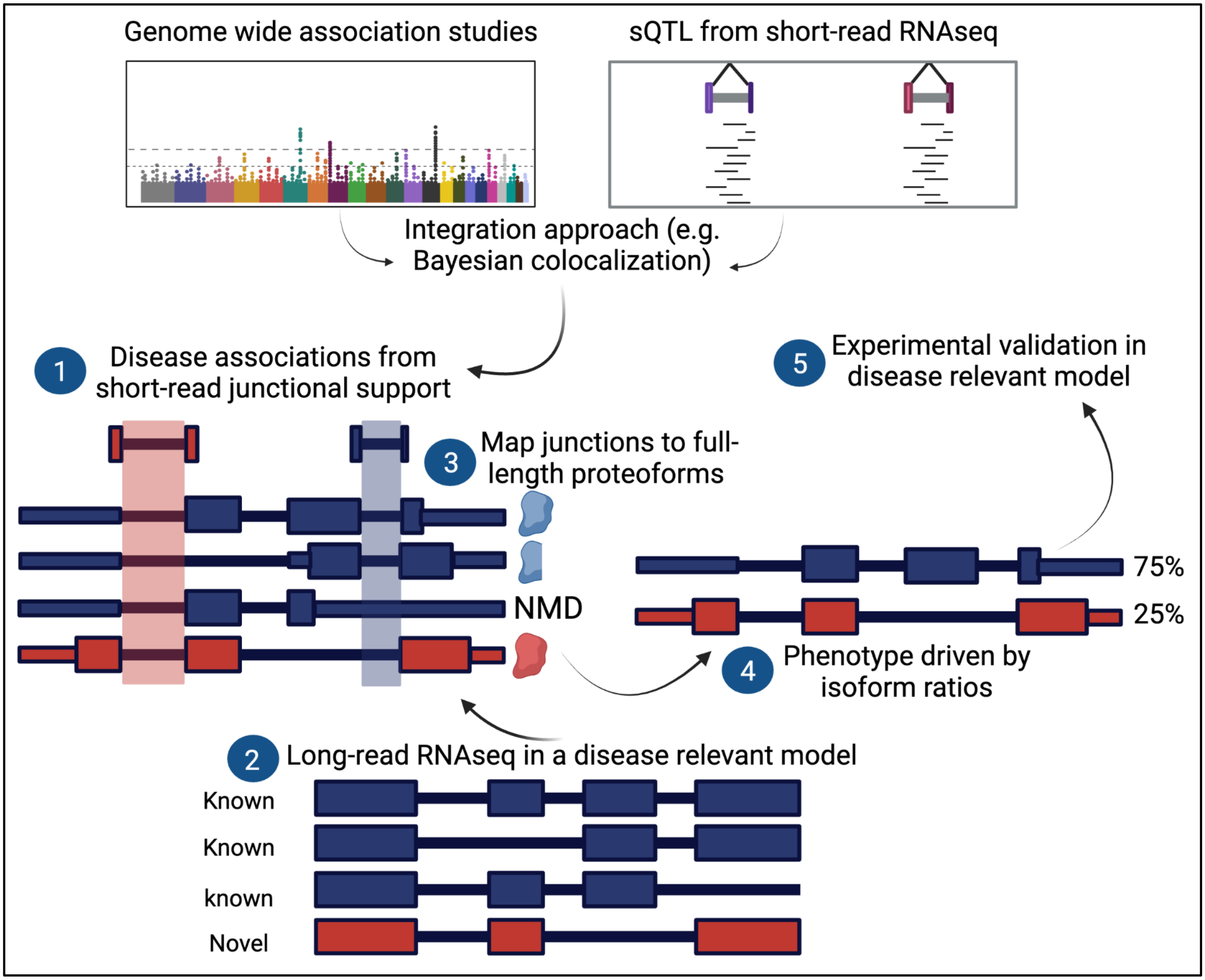

Genome wide association studies
sQTL from short-read RNAseq
Integration approach (e.g. Bayesian colocalization)
1 Disease associations from short-read junctional support
3 Map junctions to full-length proteoforms
NMD
5 Experimental validation in disease relevant model
75%
25%
4 Phenotype driven by isoform ratios
2 Long-read RNAseq in a disease relevant model
Known
Known
known
Novel

Figure 3.1: **Overview of the approach to link genetically regulated splicing (sQTLs) to candidate protein isoform effectors.** In step 1, disease associations were identified by integrating data from the latest BMD GWAS with sQTL data from 49 GTEx tissues using Bayesian colocalization analysis. In step 2, long-read RNAseq data were generated from a disease relevant model identifying both known (blue) and novel (red) isoforms and their predicted open reading frames (ORFs), which in turn were used to map (step 3) the junctions identified in step 1 (red = novel, blue = known). Additionally, the impact of these junctions on ORFs was predicted (e.g truncation, nonsense mediated decay (NMD), or novel protein). Hypotheses generated from data in step 4 are then experimentally validated in the same disease model in step 5.

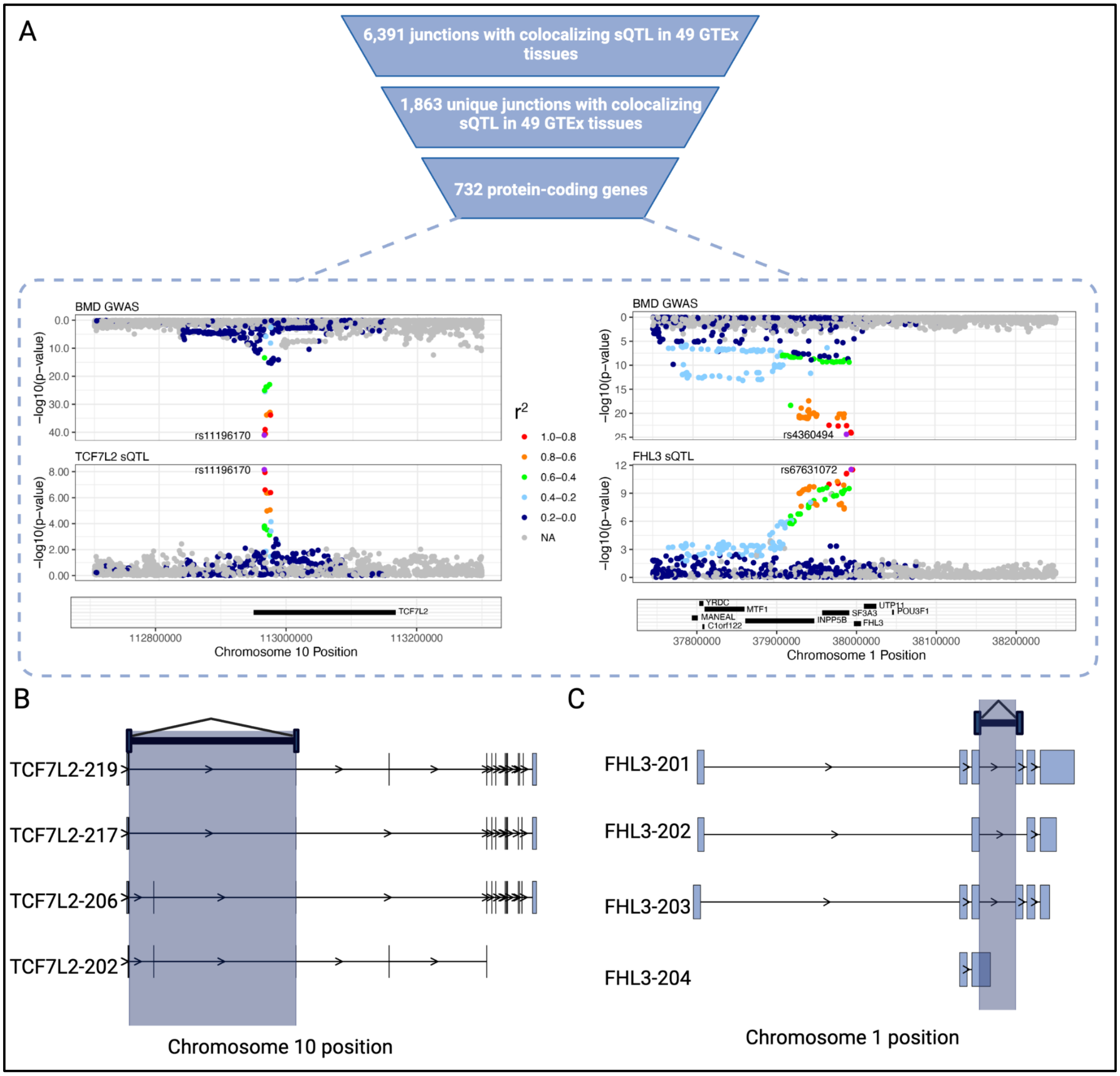


Figure 3.2: **Identification of sQTLs colocalizing with BMD GWAS associations**. A) Mirrorplots representing examples of highly colocalizing sQTLs in *TCF7L2* (H_4PP = 0.99) and *FHL3* (H_4PP = 0.99). B) Mapping the junction with colocalizing sQTL to the reference transcriptome (e.g. GENCODE) reveals multiple candidate isoforms in *TCF7L2*. C) As in B, mapping the junction with colocalizing sQTL in *FHL3* reveals multiple potential isoform candidates that could be impacted by the sQTL.

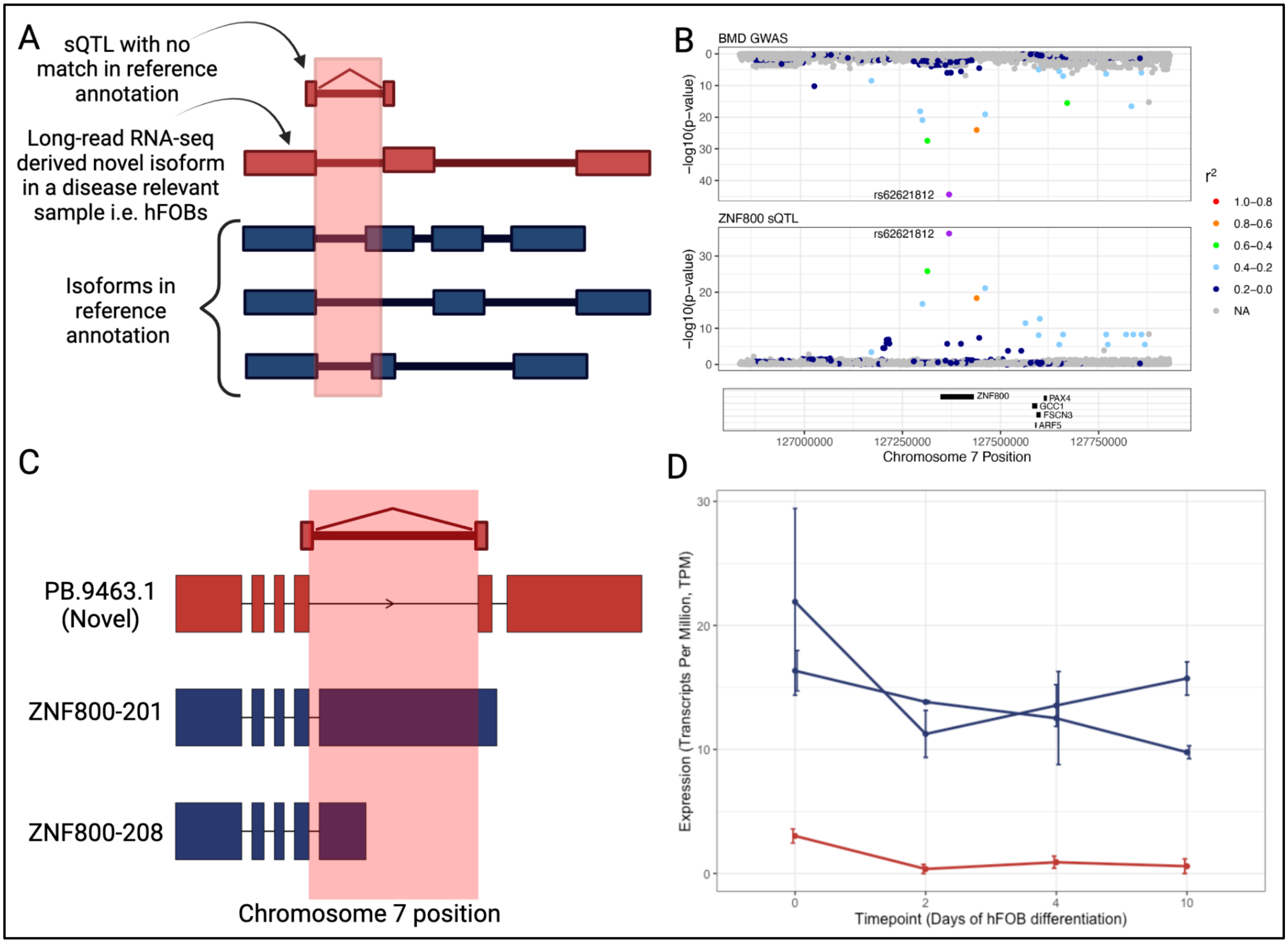


Figure 3.3: **Colocalized sQTLs impact novel isoforms**. A) Scenario in which a novel sQTL (with no match in isoforms within reference annotation (blue) is mediating its' effect through a novel isoform (red) identified via long-read RNAseq in hFOBs. B) Mirrorplot representing a highly colocalizing sQTL (H4PP = 0.99) in *ZNF800* where the most significant BMD GWAS SNP and the lead sQTL SNP are the same (rs62621812). C) Isoform models for *ZNF800* identified via long-read RNAseq in hFOBs where the junction with colocalizing sQTL maps only to the novel isoform *PB.9463.1*. D) Expression of *ZNF800* isoforms across hFOB differentiation timepoints where red represents the novel isoform *PB.9463.1* and blue represents *ZNF800-201* and *ZNF800-208*.

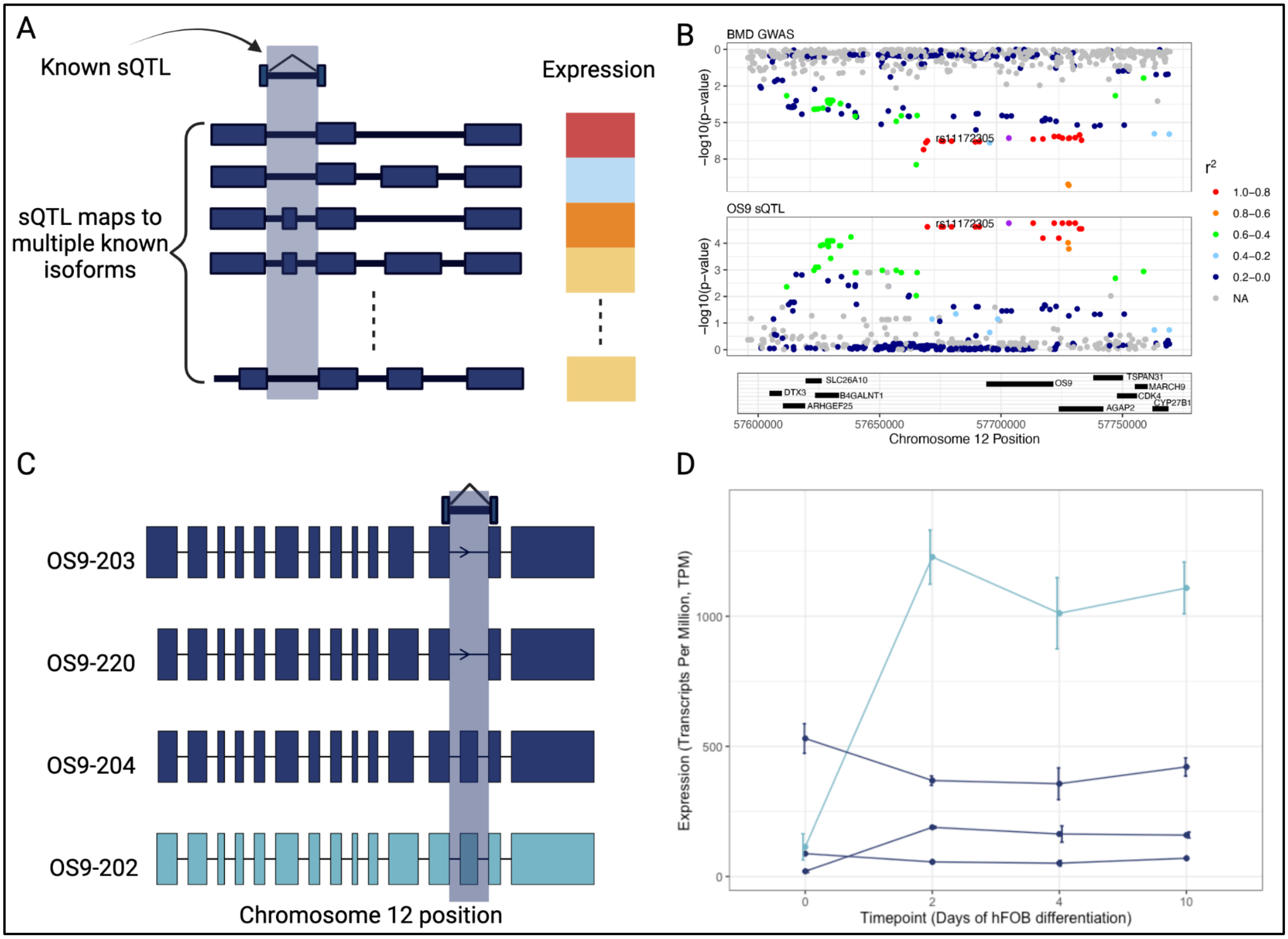


Figure 3.4: **Contextualization of isoforms corresponding to known sQTLs:** A) Scenario in which a known sQTL (match in isoforms within reference annotation) shown in blue map to multiple annotated isoforms (blue) and it is not known if all isoforms or only a subset may be relevant in mediating the trait of interest therefore full-length expression can be leveraged. B) Mirrorplot representing a colocalizing sQTL ($H_4PP = 0.86$) in *OS9*. C) Isoform models for *OS9* identified via long-read RNA-seq in hFOBs where the junction with colocalizing sQTL maps to two isoforms *OS9-220* and *OS9-203*. D) Expression of *OS9* isoforms across hFOB differentiation timepoints where light blue represents isoform *OS9-202* and blue represents the rest of the isoforms. This example highlights that putting these sQTLs within a biological framework can provide insights into potential causal isoforms.

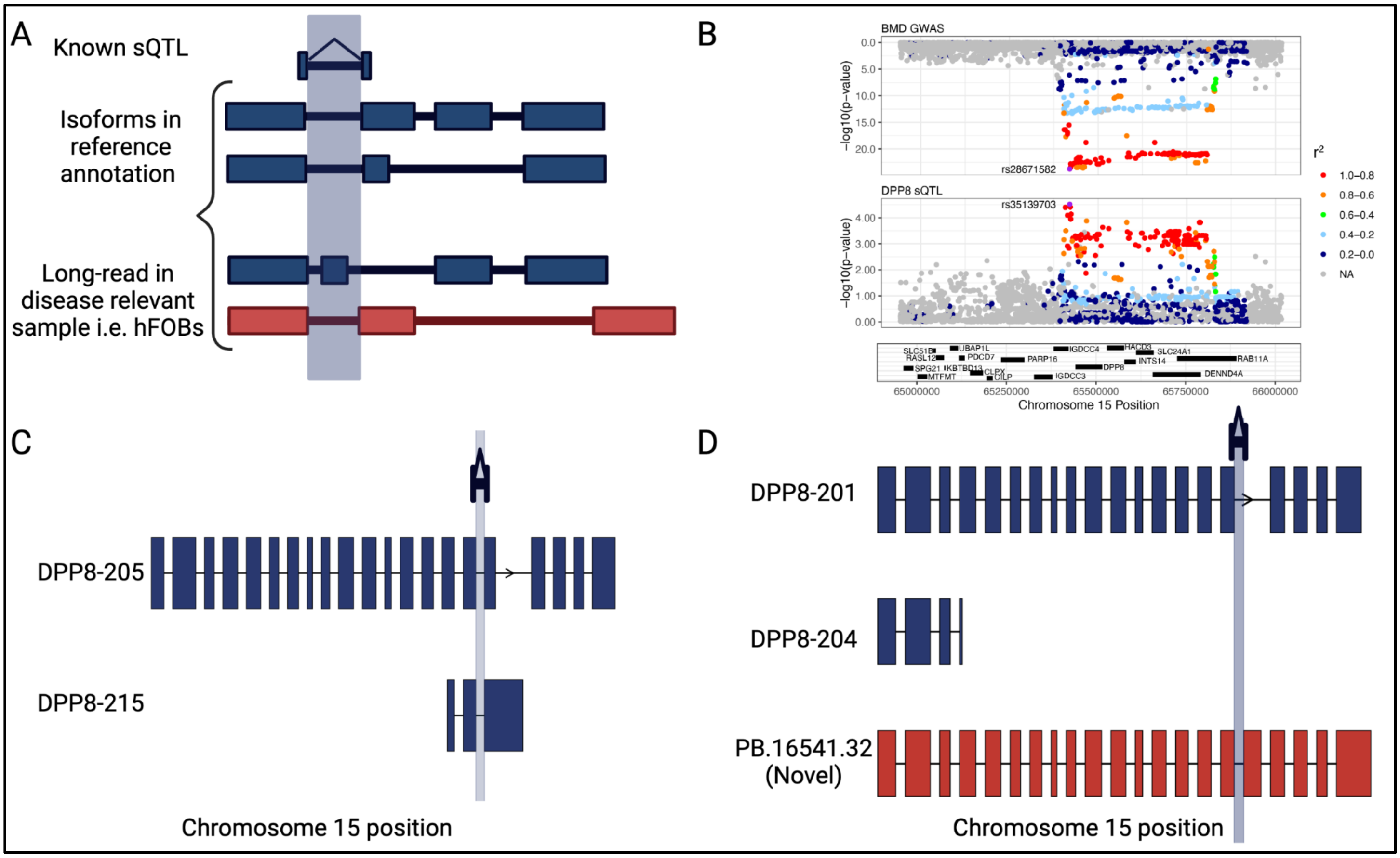


Figure 3.5: **Known sQTLs impact novel isoforms in the biological context.** A) Scenario in which annotated junctions of colocalized sQTLs can actually be found to map to novel isoforms, meaning that the local splicing event is known (blue), but the associated full-length protein isoform to which it is derived may be novel (red). B) Mirrorplot representing a colocalizing sQTL ($H_4PP = 0.83$) in *DPP8*. C) Isoform models for *DPP8* in GENCODE v38 containing the junction (blue). D) Isoform models for *DPP8* identified via long-read RNAseq in hFOBs where the junction with colocalizing sQTL maps strictly to a novel isoform *PB.16541.32*.

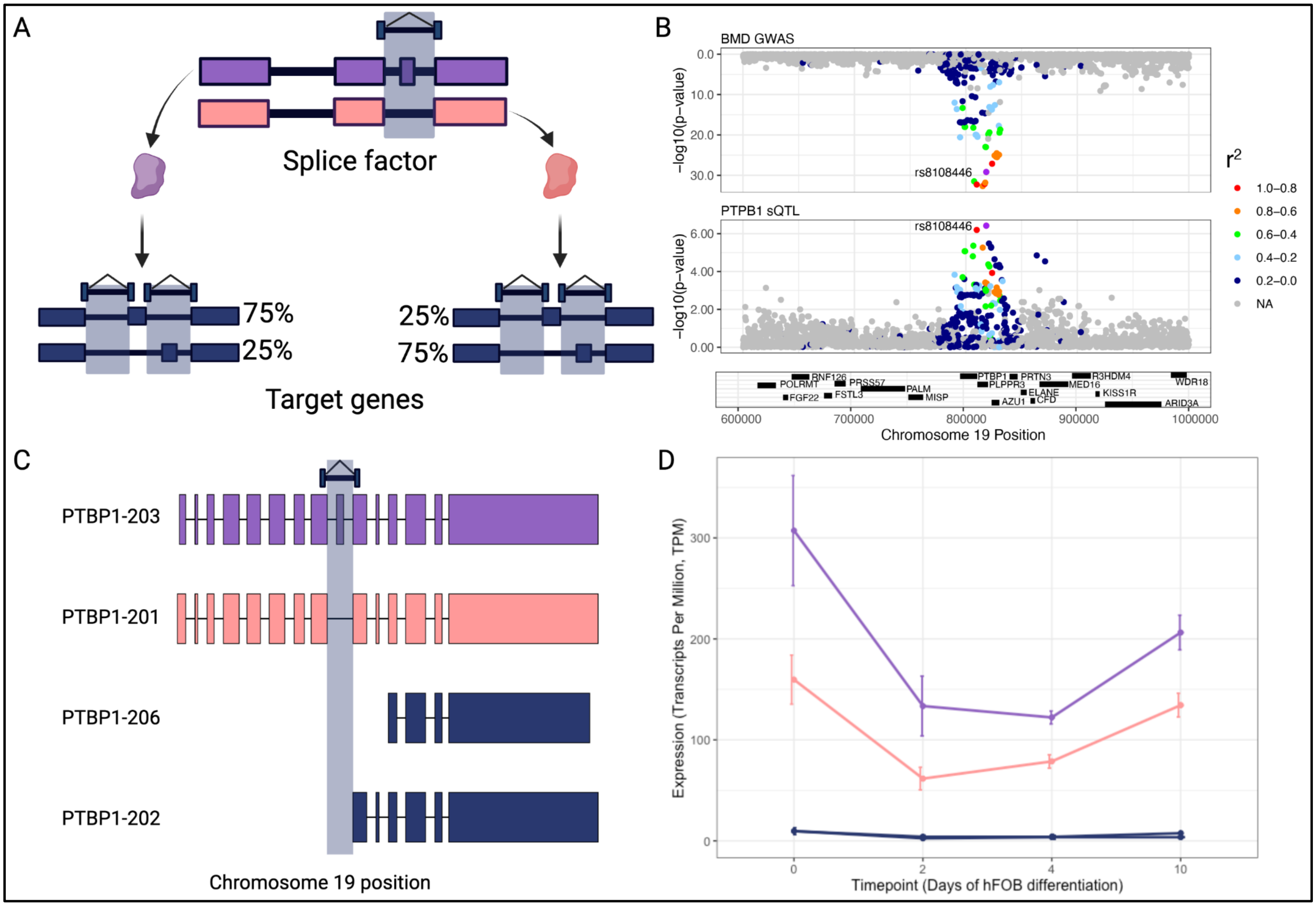


Figure 3.6: **Enrichment of lead sQTLs in splice factor binding sites**. A) Splice factors are regulated by sQTLs and in turn regulate the splicing of target genes (also containing sQTLs). B) Mirrorplot showing a significantly colocalizing sQTL around exon 9 in the splice factor *PTBP1*. C) Isoform models of *PTBP1* with the junction coordinates highlighted. D) Expression of all *PTBP1* isoforms including *PTBP1-203* (purple) and *PTBP1-201* (pink).

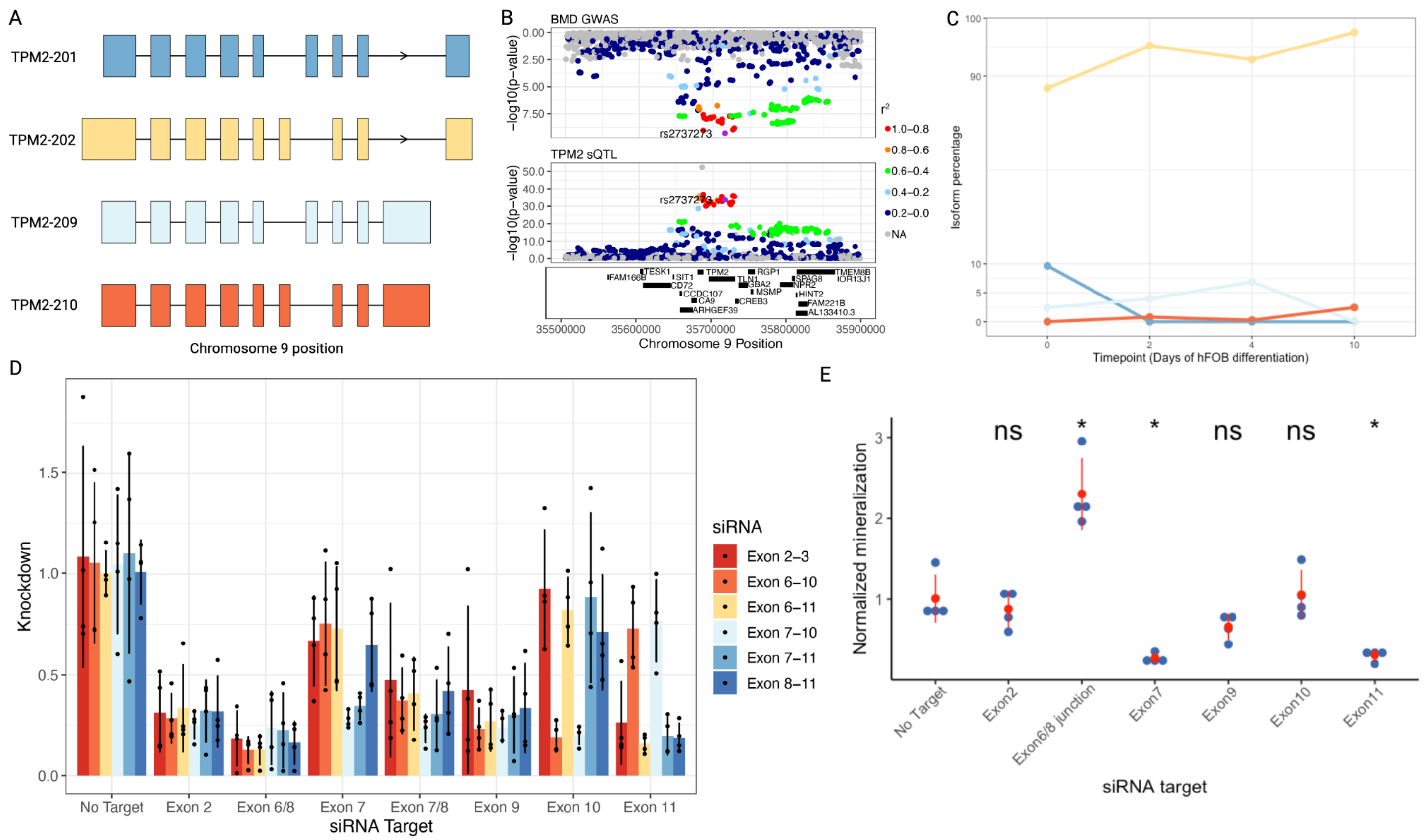


Figure 3.7: **Different TPM2 isoforms show opposing effects on mineralization**. A) Isoform models for *TPM2* isoforms expressed in hFOBs. B) Mirrorplot showing a significantly colocalizing sQTL around exons 6 and 7 in *TPM2* (H_4PP = 0.98). C) Isoform percentages across hFOB differentiation timepoints highlighting *TPM2-202* as the major isoform. D) siRNA knockdown of isoforms containing target exon (x-axis). Each color represents an siRNA target (red = isoforms containing exon 2, orange = isoform containing exons 6 and 10, yellow = isoforms containing exons 6 and 11, sky blue = isoforms containing exons 7 and 10, light blue = isoforms containing exon 7 and 11, dark blue = isoforms containing exon 11). E) Quantification of nodule mineralization using alizarin red staining in hFOBs (*) represents significant T-Test comparison between no target control and the target exon while (ns) represents no significant difference.

Appendix A

Supplemental Data

All supplemental data are available at:

https://doi.org/10.5281/zenodo.7672230

Supplemental Data 2.1 Known lncRNAs expressed in bone fragment samples

Supplemental Data 2.2 Unknown lncRNAs expressed in bone fragment samples

Supplemental Data 2.3 lncRNAs within 400 Kb of BMD GWAS loci and are differentially expressed across differentiation time points

Supplemental Data 2.4 lncRNAs that are implicated by TWAS and colocalization analysis in GTEx tissues

Supplemental Data 2.5 Colocalized lncRNAs and their Protein-coding counterparts

Supplemental Data 2.6 Coding SNPs under ASE and LD with GWAS lead SNPs

Supplemental Data 3.1 Tissue sharing for genes with colocalizing sQTLs by integrating all 49 GTEx tissues with BMD GWAS data

Supplemental Data 3.2 Distribution of sGenes within BMD GWAS lead associations

Supplemental Data 3.3 Prediction of NMD status based on long-read RNAseq data

Supplemental Data 3.4 Evaluation of transcript truncation based on long-read RNAseq data

Supplemental Data 3.5 Prioritization scheme to nominate isoforms for experimental validation

Supplemental Data 3.6 Effect sizes obtained from BMD GWAS summary statistics and slope values from GTEx tissue summary statistics for every lead sQTL

Supplemental Data 3.7 Targeted proteomics peptides for *TPM2*

Supplemental Data 3.8 *TPM2* siRNA sequences

Supplemental Data 3.9 *TPM2* qPCR primer sequences

Supplemental Data 3.10 List of junctions with significantly colocalizing sQTLs (H4PP > 0.75) for all gene types

Supplemental Data 3.11 Distribution of significant sQTL colocalizations (H4PP >= 0.75) by tissue for all gene types

Supplemental Data 3.12 List of junctions with significantly colocalizing sQTLs (H4PP > 0.75) in protein-coding genes

Supplemental Data 3.13 Distribution of significant sQTL colocalizations (H4PP >= 0.75) by tissue for protein-coding genes

Supplemental Data 3.14 Lead sQTLs in hFOBs distribution by chromosome and which are within the intron they regulate

Supplemental Data 3.15 Lead sQTLs and their proxy affecting canonical splice sites in hFOBs

Supplemental Data 3.16 Enrichment of lead sQTLs in splice factor binding sites from eCLIP data

Supplemental Data 3.17 Potential effects of proteoforms on BMD

Appendix B

Supplemental text

1. Bayesian colocalization identified potential causal junctions in BMD

We leveraged Bayesian colocalization analysis with coloc [29] using the largest BMD GWAS [13] and existing sQTLs across 49 tissues from the GTEx project [31]. Overall, we found 820 genes with colocalizing sQTLs ($H_4PP \geq 0.75$) (total number of associations = 6,889) (**Table S3.10 and S3.11**). The number of unique junctions which we define as junctions with a unique donor-acceptor site pair regardless of the tissue where colocalization is reported is 2,043. We focused our study on protein-coding genes and their isoforms. The majority of sQTLs (6,391 of 6,889; ~93%) affect protein-coding genes (732 of 820; 89%) (**Table S3.12 and S3.13**), a finding similar to other studies [48]. The number of unique junctions is 1,863 found in 732 protein-coding genes (89% of total genes). Over half the junctions (51%) with colocalizing sQTL showed an H_4PP of 0.90 or above.

2. Characterization of the full-length transcriptome in hFOBs using long-read RNAseq

We generated deep coverage long-read RNA-seq data across osteoblast differentiation in hFOBs. Long reads were collected in biological duplicate on day 0 and biological triplicate for days 2, 4, and 10, with a total of 22 million full-length reads obtained. A gradual increase in mineralization confirmed differentiation into functionally mature osteoblasts (see **Methods**, **Figure S1A-B**). We applied a stringent filtering strategy where isoforms are considered detected if expressed in all replicates of at least one time-point and making up a minimum of 1% isoform fractional abundance for the gene. We detected 68,326 transcript isoforms from 12,068 genes. The number of isoforms with a full-splice match (FSM) when compared to the GENCODE database is 33,106 (48% of total) while those with an incomplete splice match (ISM) is 17,482 (~26% of total). The number of novel isoforms is 17,738 (~26%) (**Figure S1C**). Of the novel isoforms, 10,793 (61%) arose from new combinations of known splice sites (NIC; novel in-catalog), whereas 6,580 (39%) arose from at least one novel splice donor or acceptor (NNC; novel not in-catalog) (**Figure S1C**). The median length of novel isoforms is 2,123 nt (mean = 2,268 nt). The median length of known isoforms is 2,055 nt (mean = 2,214 nt).

To characterize expression and splicing changes occuring during osteoblast differentiation, we used tappAS [122] and found that 2,034 genes (~17% of all expressed genes) were differentially expressed (DE) and 3,539 (29% of all expressed genes) were differentially spliced, or undergoing differential isoform usage (DIU) (**Figure S1D**). Interestingly, DE and DIU genes were related to distinct GO terms. DE genes associated with bone-relevant processes such as positive regulation of bone mineralization (GO:0030501; FDR = 0.004), extracellular matrix organization (GO:0030198; FDR=8.88 x 10^{-8}), and collagen-containing extracellular matrix (GO:0062023; FDR = 5.06 x 10^{-11}) (**Figure S1D**). DIU genes, however, did not associate with annotated bone-relevant processes, but were enriched in terms related to the regulation of AS, including mRNA splicing and the spliceosome (GO:0000398; FDR = 0.0002) (**Figure S1D**). Together, these results suggest the presence of a splicing program acting independently of gene regulation during osteoblast differentiation, suggesting that genetic determinants of BMD could be acting through such splice-specific pathways.

3. Mapping sQTLs colocalized with BMD onto the osteoblast-specific full-length transcript reference

The GTEx sQTLs are identified from non-bone tissues. Therefore, we sought to place the BMD-associated splicing events within the context of bone isoforms for the proper interpretation of putative molecular hypotheses. Accordingly, we proceeded to map the colocalized sQTLs onto the transcriptome of differentiating osteoblasts. The number of unique junctions with colocalizing sQTLs that are also observed in hFOBs is 836 (45% of colocalized junction in protein-coding genes), these junctions exactly map (donor-acceptor coordinates) to 459 protein-coding genes which in total have 2,349 isoforms (700 novel; ~30%). These genes are found within 362 associations (~33% of all BMD GWAS associations), with 221 lead associations harboring one sGene and 141 harboring more than one sGene (**Table S3.2**). Majority of these junctions are found in known isoforms, specifically 383 (46%) are found in known isoforms only, and 350 (42%) are found in both known and novel isoforms. But we also have a portion of these events that are only explained by novel isoforms (103 (12%) are found in novel isoforms only). In order to confirm that the novel sQTL are in fact novel and are not a product of isoforms not expressed highly in hFOBs, we also mapped them to GENCODE v38. There were 73 sQTL (of 103; ~71%) that can be explained by known isoforms within GENCODE v38. Consequently, the genes reported (and their isoforms) are potential causal candidates in osteoporosis.

4. Identifying full-length isoform from colocalized sQTLs implicated in osteoblast differentiation

We next asked whether we see an enrichment of the genes with colocalizing sQTL for those undergoing splicing regulation in hFOBs differentiation. Over one third of the genes with colocalizing sQTLs show differential isoform usage (164/459; ~36% of genes with a colocalizing sQTL). Genes that show DIU are enriched in colocalizing sQTLs (Fisher's exact test, p = 0.003), indicating the relevance of the disease model (osteoblasts) and studying splicing in this process. On the other hand, we report 82 genes with colocalizing sQTL were differentially expressed (82/459; ~18%). However, genes that show DE are not enriched in colocalizing sQTLs (Fisher's exact test, p = 0.57). The number of genes with colocalizing sQTL that show both DE and DIU is 36 genes (**Figure S1E**).

5. Identification of high priority potential causal variants relevant to BMD

We aimed to provide a potential mechanism by which variants mediate splicing of introns. One way to investigate this is by dissecting the lead variants within junctions with colocalizing sQTLs. The total number of unique lead variants (lowest p value in the locus) associated with colocalized junction in hFOBs is 1,573 potentially regulating the splicing of these genes (as some lead SNPs have varying p-values due to tissue differences).

Among the colocalized sQTLs, we found that for over half of them (875/1,573, or 56%), the lead SNP or their proxy SNP in high linkage disequilibrium (LD ≥ 0.8) resides directly within the associated intron **(Table S3.14)**, suggesting a potential regulatory mechanism associated likely to splice machinery. These proportions are similar to those reported for other complex diseases [44].

Next, we investigated whether these lead SNPs (or their proxy) are found within canonical splice sites (5’ splice-donors and 3’ splice-acceptors) in known and novel isoforms within hFOBs. We investigated whether these sQTLs are impacting unannotated splice sites, and two novel isoforms belonging to *DHRS12* (rs2296028: exon 8 acceptor site) and *PGS1* (rs11656568; exon 2 donor site) respectively showing disruptions to these sites (**Table S3.15**). This low number of SNPs that lie within these sites is expected as their disruption may be strongly deleterious [302].

6. Splice factors with colocalizing sQTLs

We obtained splice factor binding sites from CLIPdb in ENCODE [303]. We performed enrichment analysis of the lead sQTLs within the splice factor binding sites. We see significant enrichment of lead sQTLs within 34 of the 44 splice factors (**Table S3.16**). Of these, eight show differential expression across hFOB differentiation time points and 13 show differential isoform usage.

The splice factor hnRNPM shows differential expression across differentiation time points and a colocalizing sQTL. This gene has not been implicated previously in the regulation of BMD in human or mouse studies. However, evidence suggests that hnRNPM is implicated in an alternative splicing program Ewing sarcoma cells, which are aggressive tumors of bone and soft tissues [304]. Our results show three junctions with colocalizing sQTLs leading to the production of long and short isoforms of the gene.

7. Mock example to illustrate connecting colocalized sQTLs to isoforms

We illustrate a toy example in Figure S2, To determine how each isoform impacted BMD, we cross referenced the effect size of the lead GWAS SNP with the directionality and magnitude of the slope in the same SNP from the GTEx sQTL data (**Figure S2**). In this example, the TT genotype is associated with a decrease in BMD (a result obtained from GWAS summary statistics). The same genotype is also associated with increased normalized intron excision ratio (from sQTL summary statistics). An Increase in normalized intron excision ratio can be interpreted as an increase in the presence of the exon-exon junction. Therefore, we can hypothesize that genotype TT is associated with isoforms not containing exon 2 in this example which in turn is associated with a decrease in BMD. On the other hand, we can conclude that isoforms with exon 2 are associated with an increase in BMD.

8. Resource of candidate protein isoforms that mediate BMD

We compiled all the information on sQTL-linked protein isoforms, including their predicted risk status, creating the Proteoform for BMD Resource (PBR) (**Table S3.17**). We supplemented each isoform with additional evidence that may be pertinent to follow up studies. In terms of the relevance of the isoform in BMD and bone traits, we considered two major categories of evidence: i) Literature evidence pertaining to a novel role of the gene in bone processes and ii) data-driven evidence to contextualize our results. For the first category, we investigated whether the genes reported were implicated previously in bone monogenic disease (26 genes of 1,088; ~3%), have been previously identified as genes with colocalizing eQTLs in Al-Barghouthi et al. [305] (147 genes of 512; 32%), have been shown to influence bone strength in Diversity Outbred mice [306] (32 genes of 1,370; ~3%) or have been shown to disrupt BMD in IMPC [236] (14 genes of 371; ~4%). For our data-driven evidence, we leveraged differential isoform usage across hFOB differentiation (164 genes), extent of sQTL sharing across GTEx tissues, the strength of H_4PP, and the strength of the effect size (results are reported in **Tables S3.5 and S3.6**).

9. For every containing-lacking isoform pair, we check whether their open-reading frames (ORFs) contain a stop codon and whether the length of the ORF is divisible by 3. If the ORF has a stop codon located at least 50 base pairs upstream of the last splice site in the mature transcript (i.e. at the beginning of the last exon), it is considered a candidate for nonsense-mediated decay. Based on this identification, we assign an 'NMD status' of either 0 or the sum of all weights of the isoform pairs corresponding to the colocalized sQTL, which are identified as candidates for NMD. Similar to the NMD analysis, all the lacking-containing isoform pairs associated with the sQTLs were subjected to Biosurfer's hybrid alignment. From the alignment, blocks of amino-acids are generated based on identification of deletions, insertions, changes in coding status, frameshits

etc. This is used to calculate the average change in the length of the amino acid sequence between two transcript isoforms. First the blocks of the two transcripts are grouped together that correspond to the same exonic regions, and then the difference in length for each block is calculated. The resulting pairs of delta length and weight values are then used to compute the weighted average of the delta length, which is stored in the 'Average Delta Amino Acid' field. If there are no delta length and weight pairs, then the value is set to 0.

10. Experimental validation of *TPM2* in hFOBs

We observed 7 colocalizing sQTLs in total, 5 of which are around exons 5-8 observed in 13-33 tissues depending on the junction. The other two are around exons 9-11 observed specifically in the brain and testis. In order to ensure that all isoforms of *TPM2* are found, regardless of filtering criteria, we decided to cluster all the isoforms based on each time point and create a relative abundance percentage rather than absolute abundance. We were able to capture all four isoforms, albeit, at a low level of expression in certain time points.

In order to confirm that these isoforms are being translated. We were able to identify 12 unique *TPM2* peptides (**Table S3.7**). Of those, 4 are unique to exon 6, 4 unique to exon 7, 2 unique to exon 10, and 2 unique to exon 11. Our proteomics results suggest that all four isoforms of *TPM2* identified are being translated and concurs with expression data suggesting a decrease of *TPM2* abundance as hFOBs mature. The ratios of exon 6 and exon 7 in our proteomics analysis suggest an equal presence, which is not consistent with the transcript abundances reported in our long-read data. These results highlight previously reported discordance in transcript-protein correlations [307]. Taken together, we speculate that the ratios of isoforms containing exon 6 or exon 7 are associated with changes in BMD.

Methods

Long-read RNA-seq differential analysis

All differential statistical analyses in long-read data were performed using tappAS [122]. The input files for tappAS are the raw expression matrix obtained from cDNA Cupcake, the full-length transcriptome reference file generated from IsoAnnot , and a design matrix for time-series analysis. Within tappAS, maSigPro [308] was chosen for differential transcript expression using the following parameters: polynomial degree = 3, alpha = 0.05, R^2 cutoff = 0.7, and max K clusters = 10 however "mclust" [309] was used to ensure an optimal number of clusters. Differential isoform usage analysis was performed within tappAS using maSigPro with the following parameters: polynomial degree = 3, alpha = 0.05.

Functional annotation of sQTLs and enrichment regulatory regions

BioMart [310] was used to obtain the genomic positions of lead sQTLs associated with protein-coding genes using GRCh38. Ensembl REST API (https://rest.ensembl.org/) was used to obtain the variants in LD with the lead sQTLs ($r^2 \geq 0.80$; high LD). GenomicRanges [311] package in R was used to identify overlaps of the lead sQTL or those in proxy within the intron (junction) they regulate. We constructed introns from the SQANTI3 GTF file using the package "gread" followed by overlap with splice site acceptor (ssa) or splice site donor (ssd) using GenomicRanges. We used SNPsnap [312] within the package "VSEA" in R to obtain a background set of SNPs with similar minor allele frequencies, distance to nearest genes, and LD patterns. To test enrichment of lead sQTLs splice factor binding sites, we used "fisher.test" within R with a significance threshold alpha < 0.05). The splice factor information was obtained from Van Nostrand et al. [237] and splice factor binding sites were obtained from the eCLIP database within ENCODE [303].

Appendix C

Supplemental figures

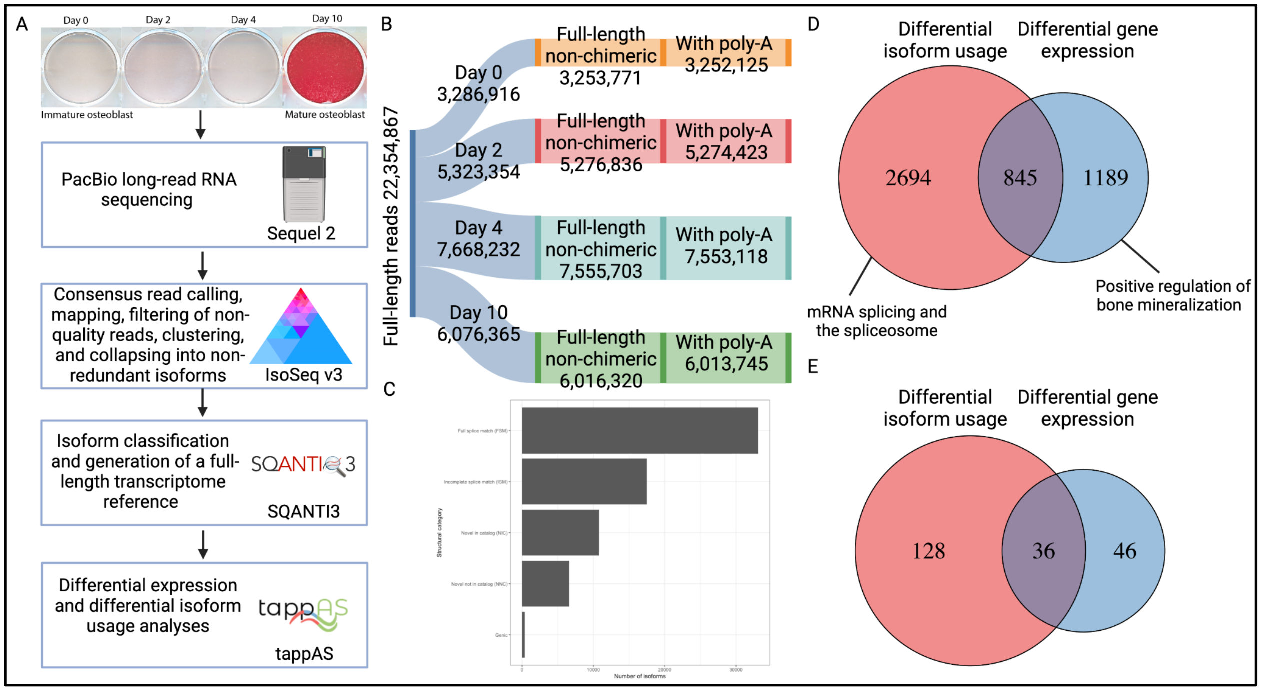


Figure S1: **Long-read RNAsequencing**. A) Analysis pipeline for long-read RNAseq performed in hFOBs across 4 differentiation timepoints (0,2,4,10). Red color indicates mineralized nodules stained with alizarin red. B) Distribution of long-read RNAseq across differentiation timepoints. C) Isoform classification. D) Venn diagram showing the number of genes showing differential isoform usage and genes that are differentially expressed across hFOB differentiation along with representative Gene Ontology (GO) terms. E) Venn diagram showing genes with colocalizing sQTLs showing differential isoform usage and differential expression across hFOB differentiation.

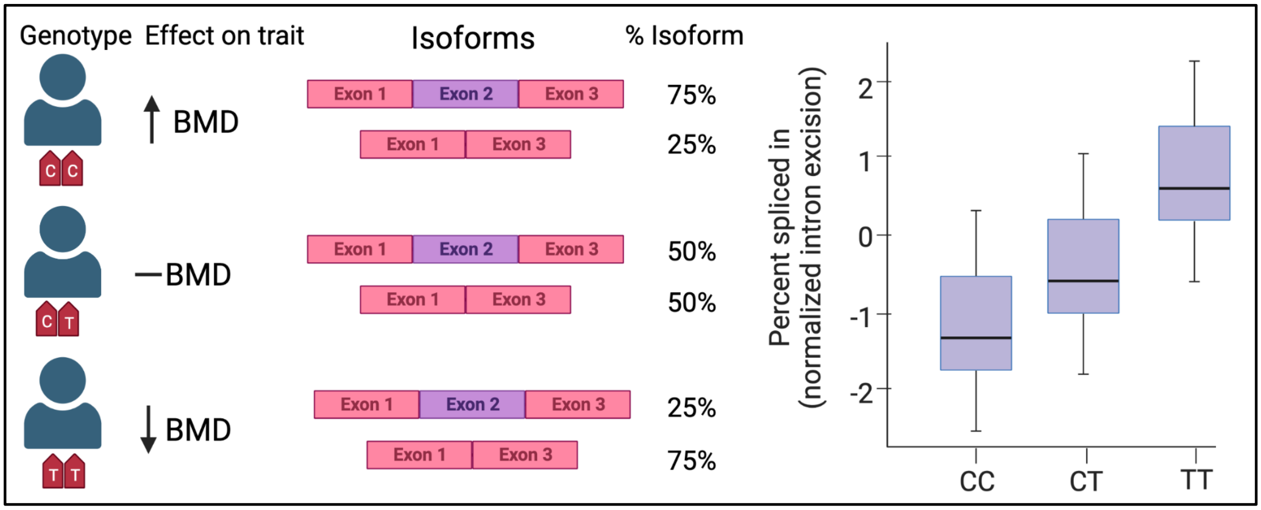


Figure S2: **Mock illustration to determine how each isoform impacted BMD**. The TT genotype is associated with a decrease in BMD (a result obtained from GWAS summary statistics). The same genotype is also

associated with increased normalized intron excision ratio (from sQTL summary statistics). An Increase in normalized intron excision ratio can be interpreted as an increase in the presence of the exon-exon junction. Therefore, we can hypothesize that genotype TT is associated with isoforms not containing exon 2 in this example which in turn is associated with a decrease in BMD. On the other hand, we can conclude that isoforms with exon 2 are associated with an increase in BMD.

References

1. Office of the Surgeon General (US). *The Burden of Bone Disease*. (Office of the Surgeon General (US), 2004).
2. NIH Consensus Development Panel on Osteoporosis Prevention, Diagnosis, and Therapy. Osteoporosis prevention, diagnosis, and therapy. *JAMA* **285**, 785–795 (2001).
3. Burge, R. *et al.* Incidence and Economic Burden of Osteoporosis-Related Fractures in the United States, 2005-2025. *Journal of Bone and Mineral Research* vol. 22 465–475 Preprint at https://doi.org/10.1359/jbmr.061113 (2007).
4. Center, J. R., Nguyen, T. V., Schneider, D., Sambrook, P. N. & Eisman, J. A. Mortality after all major types of osteoporotic fracture in men and women: an observational study. *Lancet* **353**, 878–882 (1999).
5. Johnell, O. *et al.* Predictive value of BMD for hip and other fractures. *J. Bone Miner. Res.* **20**, 1185–1194 (2005).
6. Smith, D. M., Nance, W. E., Kang, K. W., Christian, J. C. & Johnston, C. C., Jr. Genetic factors in determining bone mass. *J. Clin. Invest.* **52**, 2800–2808 (1973).
7. Arden, N. K., Baker, J., Hogg, C., Baan, K. & Spector, T. D. The heritability of bone mineral density, ultrasound of the calcaneus and hip axis length: a study of postmenopausal twins. *J. Bone Miner. Res.* **11**, 530–534 (1996).
8. Slemenda, C. W. *et al.* The genetics of proximal femur geometry, distribution of bone mass and bone mineral density. *Osteoporos. Int.* **6**, 178–182 (1996).
9. 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).
10. Ralston, S. H. & Uitterlinden, A. G. Genetics of osteoporosis. *Endocr. Rev.* **31**, 629–662 (2010).
11. Styrkarsdottir, U. *et al.* Linkage of osteoporosis to chromosome 20p12 and association to BMP2. *PLoS Biol.* **1**, E69 (2003).
12. Ioannidis, J. P. *et al.* Meta-analysis of genome-wide scans provides evidence for sex- and site-specific

regulation of bone mass. *J. Bone Miner. Res.* **22**, 173–183 (2007).

13. Morris, J. A. *et al.* An atlas of genetic influences on osteoporosis in humans and mice. *Nat. Genet.* **51**, 258–266 (2019).
14. Richards, J. B. *et al.* Collaborative meta-analysis: associations of 150 candidate genes with osteoporosis and osteoporotic fracture. *Ann. Intern. Med.* **151**, 528–537 (2009).
15. Zheng, H. *et al.* Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. *Nature* **526**, 112–117 (2015).
16. 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).
17. Maurano, M. T., Humbert, R. & Rynes, E. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **2012;337(6099):1190-1195**, 1222794 (1126).
18. 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).
19. Reich, D. E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
20. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).
21. 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).
22. Zheng, H.-F. *et al.* WNT16 influences bone mineral density, cortical bone thickness, bone strength, and osteoporotic fracture risk. *PLoS Genet.* **8**, e1002745 (2012).
23. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
24. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
25. King, E. A., Davis, J. W. & Degner, J. F. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of

drug approval. *PLoS Genet.* **15**, e1008489 (2019).

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

27. Dewan, A. *et al.* HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science* **314**, 989–992 (2006).

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

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

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

31. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).

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

33. Mullin, B. H. *et al.* Characterisation of genetic regulatory effects for osteoporosis risk variants in human osteoclasts. *Genome Biol.* **21**, 80 (2020).

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

35. Styrkarsdottir, U. *et al.* Multiple genetic loci for bone mineral density and fractures. *N. Engl. J. Med.* **358**, 2355–2365 (2008).

36. 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).

37. 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).

38. Reppe, S. *et al.* Eight genes are highly associated with BMD variation in postmenopausal Caucasian women. *Bone* **46**, 604–612 (2010).

39. Medina-Gomez, C. *et al.* Meta-analysis of genome-wide scans for total body BMD in children and adults reveals allelic heterogeneity and age-specific effects at the WNT16 locus. *PLoS Genet.* **8**, e1002718 (2012).

40. van Heyningen, V. Faculty Opinions recommendation of Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature* Preprint at https://doi.org/10.3410/f.1082972.535947 (2007).

41. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).

42. Hakonarson, H. *et al.* A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. *Nature* **448**, 591–594 (2007).

43. Small, K. S. *et al.* Regulatory variants at KLF14 influence type 2 diabetes risk via a female-specific effect on adipocyte size and body composition. *Nat. Genet.* **50**, 572–580 (2018).

44. Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604 (2016).

45. Park, E., Pan, Z., Zhang, Z., Lin, L. & Xing, Y. The Expanding Landscape of Alternative Splicing Variation in Human Populations. *Am. J. Hum. Genet.* **102**, 11–26 (2018).

46. Takata, A., Matsumoto, N. & Kato, T. Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. *Nat. Commun.* **8**, 14519 (2017).

47. Barbeira, A. N. *et al.* Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome Biol.* **22**, 49 (2021).

48. Garrido-Martín, D., Borsari, B., Calvo, M., Reverter, F. & Guigó, R. Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. *Nat. Commun.* **12**, 727 (2021).

49. Loos, R. J. F. 15 years of genome-wide association studies and no signs of slowing down. *Nat. Commun.* **11**, 5900 (2020).

50. 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).

51. Hukku, A. *et al.* Probabilistic Colocalization of Genetic Variants from Complex and Molecular Traits: Promise and Limitations. *Cold Spring Harbor Laboratory* 2020.07.01.182097 (2020) doi:10.1101/2020.07.01.182097.

52. Gamazon, E. R. & Stranger, B. E. Genomics of alternative splicing: evolution, development and pathophysiology. *Hum. Genet.* **133**, 679–687 (2014).

53. Lu, Z.-X., Jiang, P. & Xing, Y. Genetic variation of pre-mRNA alternative splicing in human populations. *Wiley Interdisciplinary Reviews: RNA* vol. 3 581–592 Preprint at https://doi.org/10.1002/wrna.120 (2012).

54. Early, P. *et al.* Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. *Cell* **20**, 313–319 (1980).

55. Choi, E., Kuehl, M. & Wall, R. RNA splicing generates a variant light chain from an aberrantly rearranged kappa gene. *Nature* **286**, 776–779 (1980).

56. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).

57. Berget, S. M., Moore, C. & Sharp, P. A. Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 3171–3175 (1977).

58. Nilsen, T. W. & Graveley, B. R. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**, 457–463 (2010).

59. Kalsotra, A. & Cooper, T. A. Functional consequences of developmentally regulated alternative splicing. *Nat. Rev. Genet.* **12**, 715–729 (2011).

60. Barash, Y. *et al.* Deciphering the splicing code. *Nature* **465**, 53–59 (2010).

61. Scotti, M. M. & Swanson, M. S. RNA mis-splicing in disease. *Nat. Rev. Genet.* **17**, 19–32 (2016).

62. Cooper, T. A., Wan, L. & Dreyfuss, G. RNA and disease. *Cell* **136**, 777–793 (2009).

63. Cieply, B. & Carstens, R. P. Functional roles of alternative splicing factors in human disease. *Wiley Interdiscip. Rev. RNA* **6**, 311–326 (2015).

64. Ferraro, N. M. *et al.* Transcriptomic signatures across human tissues identify functional rare genetic variation. *Science* **369**, (2020).

65. Coulombe-Huntington, J., Lam, K. C. L., Dias, C. & Majewski, J. Fine-scale variation and genetic determinants of alternative splicing across individuals. *PLoS Genet.* **5**, e1000766 (2009).

66. Zhang, X. *et al.* Identification of common genetic variants controlling transcript isoform variation in human whole blood. *Nat. Genet.* **47**, 345–352 (2015).

67. Kwan, T. *et al.* Genome-wide analysis of transcript isoform variation in humans. *Nat. Genet.* **40**, 225–231 (2008).

68. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).

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

70. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).

71. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).

72. Taylor-Weiner, A. *et al.* Scaling computational genomics to millions of individuals with GPUs. *Genome Biol.* **20**, 228 (2019).

73. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* vol. 42 348–354 Preprint at https://doi.org/10.1038/ng.548 (2010).

74. Gusev, A. *et al.* A transcriptome-wide association study of high-grade serous epithelial ovarian cancer identifies new susceptibility genes and splice variants. *Nat. Genet.* **51**, 815–823 (2019).

75. Walker, R. L. *et al.* Genetic Control of Expression and Splicing in Developing Human Brain Informs

Disease Mechanisms. *Cell* **181**, 745 (2020).

76. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7**, 1009–1015 (2010).

77. Trincado, J. L. *et al.* SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.* **19**, 40 (2018).

78. Mehmood, A. *et al.* Systematic evaluation of differential splicing tools for RNA-seq studies. *Brief. Bioinform.* **21**, 2052–2065 (2020).

79. Sebestyén, E., Zawisza, M. & Eyras, E. Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. *Nucleic Acids Res.* **43**, 1345–1356 (2015).

80. Shen, S. *et al.* rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E5593–601 (2014).

81. Anvar, S. Y. *et al.* Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing. *Genome Biology* vol. 19 Preprint at https://doi.org/10.1186/s13059-018-1418-0 (2018).

82. Tilgner, H. *et al.* Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat. Biotechnol.* **33**, 736–742 (2015).

83. Alasoo, K. *et al.* Genetic effects on promoter usage are highly context-specific and contribute to complex traits. *Elife* **8**, (2019).

84. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Erratum: Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 888 (2016).

85. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* vol. 12 Preprint at https://doi.org/10.1186/1471-2105-12-323 (2011).

86. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

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

88. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).

89. Ye, C. J. *et al.* Genetic analysis of isoform usage in the human anti-viral response reveals influenza-specific regulation of ERAP2 transcripts under balancing selection. *Genome Res.* **28**, 1812–1825 (2018).

90. Nowicka, M. & Robinson, M. D. DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Res.* **5**, 1356 (2016).

91. Yang, J., Qi, T., Wu, Y., Zhang, F. & Zeng, J. Genetic control of RNA splicing and its distinctive role in complex trait variation. Preprint at https://doi.org/10.21203/rs.3.rs-155233/v1.

92. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10**, 1177–1184 (2013).

93. Kanitz, A. *et al.* Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biol.* **16**, 150 (2015).

94. Zhang, C., Zhang, B., Lin, L.-L. & Zhao, S. Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics* **18**, 583 (2017).

95. Teng, M. *et al.* A benchmark for RNA-seq quantification pipelines. *Genome Biology* vol. 17 Preprint at https://doi.org/10.1186/s13059-016-0940-1 (2016).

96. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **21**, 30 (2020).

97. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).

98. Mantere, T., Kersten, S. & Hoischen, A. Long-Read Sequencing Emerging in Medical Genetics. *Front. Genet.* **10**, 426 (2019).

99. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nat. Rev. Genet.* (2019)

doi:10.1038/s41576-019-0150-2.

100. Volden, R. *et al.* Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 9726–9731 (2018).

101. Eid, J. *et al.* Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* vol. 323 133–138 Preprint at https://doi.org/10.1126/science.1162986 (2009).

102. Liu, S. *et al.* Targeted transcriptome analysis using synthetic long read sequencing uncovers isoform reprograming in the progression of colon cancer. *Commun Biol* **4**, 506 (2021).

103. Sharon, D., Tilgner, H., Grubert, F. & Snyder, M. A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* **31**, 1009–1014 (2013).

104. Workman, R. E. *et al.* Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat. Methods* **16**, 1297–1305 (2019).

105. Mudge, J. M. & Harrow, J. The state of play in higher eukaryote gene annotation. *Nat. Rev. Genet.* **17**, 758–772 (2016).

106. Deveson, I. W. *et al.* Universal Alternative Splicing of Noncoding Exons. *Cell Systems* vol. 6 245–255.e5 Preprint at https://doi.org/10.1016/j.cels.2017.12.005 (2018).

107. Sheynkman, G. M. *et al.* ORF Capture-Seq as a versatile method for targeted identification of full-length isoforms. *Nat. Commun.* **11**, 2326 (2020).

108. Amoah, K. *et al.* Allele-specific alternative splicing and its functional genetic variants in human tissues. *Genome Res.* **31**, 359–371 (2021).

109. Li, G. *et al.* Identification of allele-specific alternative mRNA processing via transcriptome sequencing. *Nucleic Acids Research* vol. 40 e104–e104 Preprint at https://doi.org/10.1093/nar/gks280 (2012).

110. Tilgner, H., Grubert, F., Sharon, D. & Snyder, M. P. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proceedings of the National Academy of Sciences* vol. 111 9869–9874 Preprint at https://doi.org/10.1073/pnas.1400447111 (2014).

111. Glinos, D. A. *et al.* Transcriptome variation in human tissues revealed by long-read sequencing. *Nature*

608, 353–359 (2022).

112. Deonovic, B., Wang, Y., Weirather, J., Wang, X.-J. & Au, K. F. IDP-ASE: haplotyping and quantifying allele-specific expression at the gene and gene isoform level by hybrid sequencing. *Nucleic Acids Research* vol. 45 e32–e32 Preprint at https://doi.org/10.1093/nar/gkw1076 (2017).
113. Souza, V. B. C. de *et al.* Transformation of alignment files improves performance of variant callers for long-read RNA sequencing data. Preprint at https://doi.org/10.1101/2022.02.08.479579.
114. Saferali, A. *et al.* Characterization of a COPD-Associated *NPNT* Functional Splicing Genetic Variant in Human Lung Tissue via Long-Read Sequencing. Preprint at https://doi.org/10.1101/2020.10.20.20203927.
115. Miller, R. M. *et al.* Enhanced protein isoform characterization through long-read proteogenomics. *Genome Biol.* **23**, 69 (2022).
116. Mays, A. D. *et al.* Single-Molecule Real-Time (SMRT) Full-Length RNA-Sequencing Reveals Novel and Distinct mRNA Isoforms in Human Bone Marrow Cell Subpopulations. *Genes* vol. 10 253 Preprint at https://doi.org/10.3390/genes10040253 (2019).
117. Pietzner, M. *et al.* Mapping the proteo-genomic convergence of human diseases. *Science* vol. 374 Preprint at https://doi.org/10.1126/science.abj1541 (2021).
118. Chick, J. M. *et al.* Defining the consequences of genetic variation on a proteome-wide scale. *Nature* **534**, 500–505 (2016).
119. Wu, L. *et al.* Variation and genetic control of protein abundance in humans. *Nature* **499**, 79–82 (2013).
120. Reixachs-Solé, M. & Eyras, E. Uncovering the impacts of alternative splicing on the proteome with current omics techniques. *Wiley Interdiscip. Rev. RNA* **13**, e1707 (2022).
121. Li, H.-D., Menon, R., Omenn, G. S. & Guan, Y. The emerging era of genomic data integration for analyzing splice isoform function. *Trends in Genetics* vol. 30 340–347 Preprint at https://doi.org/10.1016/j.tig.2014.05.005 (2014).
122. de la Fuente, L. *et al.* tappAS: a comprehensive computational framework for the analysis of the functional impact of differential splicing. *Genome Biol.* **21**, 119 (2020).

123. Tapial, J. *et al.* An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. *Genome Research* vol. 27 1759–1768 Preprint at https://doi.org/10.1101/gr.220962.117 (2017).

124. Tranchevent, L.-C. *et al.* Identification of protein features encoded by alternative exons using Exon Ontology. *Genome Research* vol. 27 1087–1097 Preprint at https://doi.org/10.1101/gr.212696.116 (2017).

125. Ghadie, M. A., Lambourne, L., Vidal, M. & Xia, Y. Domain-based prediction of the human isoform interactome provides insights into the functional impact of alternative splicing. *PLoS Comput. Biol.* **13**, e1005717 (2017).

126. Narykov, O., Johnson, N. T. & Korkin, D. Predicting protein interaction network perturbation by alternative splicing with semi-supervised learning. *Cell Rep.* **37**, 110045 (2021).

127. Iancu, O. D. *et al.* Cosplicing network analysis of mammalian brain RNA-Seq data utilizing WGCNA and Mantel correlations. *Front. Genet.* **6**, 174 (2015).

128. Saha, A. *et al.* Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome Res.* **27**, 1843–1858 (2017).

129. Eksi, R. *et al.* Systematically Differentiating Functions for Alternatively Spliced Isoforms through Integrating RNA-seq Data. *PLoS Computational Biology* vol. 9 e1003314 Preprint at https://doi.org/10.1371/journal.pcbi.1003314 (2013).

130. Li, W. *et al.* High-resolution functional annotation of human transcriptome: predicting isoform functions by a novel multiple instance-based label propagation method. *Nucleic Acids Research* vol. 42 e39–e39 Preprint at https://doi.org/10.1093/nar/gkt1362 (2014).

131. Li, W. *et al.* Pushing the annotation of cellular activities to a higher resolution: Predicting functions at the isoform level. *Methods* vol. 93 110–118 Preprint at https://doi.org/10.1016/j.ymeth.2015.07.016 (2016).

132. Möröy, T. & Heyd, F. The impact of alternative splicing in vivo: Mouse models show the way. *RNA* vol. 13 1155–1171 Preprint at https://doi.org/10.1261/rna.554607 (2007).

133. Bertomeu, T. *et al.* A High-Resolution Genome-Wide CRISPR/Cas9 Viability Screen Reveals Structural

Features and Contextual Diversity of the Human Cell-Essential Proteome. *Molecular and Cellular Biology* vol. 38 Preprint at https://doi.org/10.1128/mcb.00302-17 (2018).

134. Prinos, P. *et al.* Alternative splicing of SYK regulates mitosis and cell survival. *Nature Structural & Molecular Biology* vol. 18 673–679 Preprint at https://doi.org/10.1038/nsmb.2040 (2011).

135. Buljan, M. *et al.* Tissue-Specific Splicing of Disordered Segments that Embed Binding Motifs Rewires Protein Interaction Networks. *Molecular Cell* vol. 46 871–883 Preprint at https://doi.org/10.1016/j.molcel.2012.05.039 (2012).

136. Ellis, J. D. *et al.* Tissue-Specific Alternative Splicing Remodels Protein-Protein Interaction Networks. *Molecular Cell* vol. 46 884–892 Preprint at https://doi.org/10.1016/j.molcel.2012.05.037 (2012).

137. Yang, X. *et al.* Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell* vol. 164 805–817 Preprint at https://doi.org/10.1016/j.cell.2016.01.029 (2016).

138. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).

139. Choi, J. K. & Kim, S. C. Environmental effects on gene expression phenotype have regional biases in the human genome. *Genetics* **175**, 1607–1613 (2007).

140. Hunter, D. J. Gene–environment interactions in human diseases. *Nat. Rev. Genet.* **6**, 287–298 (2005).

141. Yin, P. *et al.* Integrating genome-wide association and transcriptome predicted model identify novel target genes with osteoporosis. 771543 (2019) doi:10.1101/771543.

142. Liu, Y. *et al.* Gene Expression and RNA Splicing Imputation Identifies Novel Candidate Genes Associated with Osteoporosis. *J. Clin. Endocrinol. Metab.* **105**, (2020).

143. Pividori, M. *et al.* PhenomeXcan: Mapping the genome to the phenome through the transcriptome. *Sci. Adv.* **6**, eaba2083 (2020).

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

145. 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).

146. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, Article17 (2005).

147. Fuller, T. F. *et al.* Weighted gene coexpression network analysis strategies applied to mouse weight. *Mamm. Genome* **18**, 463–472 (2007).

148. Zhang, S., Zhao, H. & Ng, M. K. Functional Module Analysis for Gene Coexpression Networks with Network Integration. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **12**, 1146–1160 (2015).

149. Gaiteri, C., Ding, Y., French, B., Tseng, G. C. & Sibille, E. Beyond modules and hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders. *Genes Brain Behav.* **13**, 13–24 (2014).

150. Bennett, B. J. *et al.* A high-resolution association mapping panel for the dissection of complex traits in mice. *Genome Res.* **20**, 281–290 (2010).

151. Al-Barghouthi, B. M. *et al.* Systems genetics analyses in Diversity Outbred mice inform human bone mineral density GWAS and identify *Qsox1* as a novel determinant of bone strength. Preprint at https://doi.org/10.1101/2020.06.24.169839.

152. Charniak, E. Bayesian networks without tears. *AI magazine* **12**, 50–50 (1991).

153. Zhao, Y. *et al.* Network-Based Identification and Prioritization of Key Regulators of Coronary Artery Disease Loci. *Arterioscler. Thromb. Vasc. Biol.* **36**, 928–941 (2016).

154. Pereira, M. *et al.* A trans-eQTL network regulates osteoclast multinucleation and bone mass. *Elife* **9**, (2020).

155. Kang, H. *et al.* Kcnn4 is a regulator of macrophage multinucleation in bone homeostasis and inflammatory disease. *Cell Rep.* **8**, 1210–1224 (2014).

156. Tak, Y. G. & Farnham, P. J. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics &*

Chromatin vol. 8 Preprint at https://doi.org/10.1186/s13072-015-0050-4 (2015).

157. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398–1414.e24 (2016).
158. Kumasaka, N., Knights, A. J. & Gaffney, D. J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.* **48**, 206–213 (2016).
159. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.* **109**, 21.29.1–21.29.9 (2015).
160. Kumasaka, N., Knights, A. J. & Gaffney, D. J. High-resolution genetic mapping of putative causal interactions between regions of open chromatin. *Nat. Genet.* **51**, 128–137 (2019).
161. Khetan, S. *et al.* Type 2 Diabetes–Associated Genetic Variants Regulate Chromatin Accessibility in Human Islets. *Diabetes* **67**, 2466–2477 (2018).
162. Ciuculete, D. M. *et al.* meQTL and ncRNA functional analyses of 102 GWAS-SNPs associated with depression implicate HACE1 and SHANK2 genes. *Clin. Epigenetics* **12**, 99 (2020).
163. Thompson, E. E. *et al.* Cytokine-induced molecular responses in airway smooth muscle cells inform genome-wide association studies of asthma. *Genome Med.* **12**, 64 (2020).
164. Xie, Y. & Ahn, C. Statistical methods for integrating multiple types of high-throughput data. *Methods Mol. Biol.* **620**, 511–529 (2010).
165. Qiu, C. *et al.* Multi-omics Data Integration for Identifying Osteoporosis Biomarkers and Their Biological Interaction and Causal Mechanisms. *iScience* **23**, 100847 (2020).
166. Chesi, A. *et al.* Genome-scale Capture C promoter interactions implicate effector genes at GWAS loci for bone mineral density. *Nat. Commun.* **10**, 1260 (2019).
167. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47**, 598–606 (2015).
168. van der Wijst, M. *et al.* The single-cell eQTLGen consortium. *Elife* **9**, (2020).
169. Rai, V. *et al.* Single-cell ATAC-Seq in human pancreatic islets and deep learning upscaling of rare cells

reveals cell-specific type 2 diabetes regulatory signatures. *Mol Metab* **32**, 109–121 (2020).

170. Wang, Z. *et al.* Single-cell RNA sequencing deconvolutes the in vivo heterogeneity of human bone marrow-derived mesenchymal stem cells. *bioRxiv* (2020).

171. Yang, J. *et al.* A systematic dissection of human primary osteoblasts in vivo at single-cell resolution. *bioRxiv* (2020).

172. Breker, M. & Schuldiner, M. The emergence of proteome-wide technologies: systematic analysis of proteins comes of age. *Nat. Rev. Mol. Cell Biol.* **15**, 453–464 (2014).

173. Nielson, C. M., Jacobs, J. M. & Orwoll, E. S. Proteomic studies of bone and skeletal health outcomes. *Bone* **126**, 18–26 (2019).

174. Lee, J.-H. & Cho, J.-Y. Proteomics approaches for the studies of bone metabolism. *BMB Rep.* **47**, 141–148 (2014).

175. Jiang, X. *et al.* Method development of efficient protein extraction in bone tissue for proteome analysis. *J. Proteome Res.* **6**, 2287–2294 (2007).

176. Hennrich, M. L. *et al.* Cell-specific proteome analyses of human bone marrow reveal molecular features of age-dependent functional decline. *Nat. Commun.* **9**, 4004 (2018).

177. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).

178. Yao, C. *et al.* Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat. Commun.* **9**, 1–11 (2018).

179. International Mouse Knockout Consortium, Collins, F. S., Rossant, J. & Wurst, W. A mouse for all reasons. *Cell* **128**, 9–13 (2007).

180. Austin, C. P. *et al.* The knockout mouse project. *Nat. Genet.* **36**, 921 (2004).

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

182. Maynard, R. D. & Ackert-Bicknell, C. L. Mouse Models and Online Resources for Functional Analysis of Osteoporosis Genome-Wide Association Studies. *Front. Endocrinol.* **10**, 277 (2019).

183. Jinek, M. *et al.* A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337**, 816–821 (2012).

184. Adli, M. The CRISPR tool kit for genome editing and beyond. *Nat. Commun.* **9**, 1911 (2018).

185. United States Public Health Service & Surgeon General of the United States. *Bone Health and Osteoporosis: A Report of the Surgeon General*. (University Press of the Pacific, 2004).

186. Richards, J. B., Zheng, H.-F. & Spector, T. D. Erratum: Genetics of osteoporosis from genome-wide association studies: advances and challenges. *Nat. Rev. Genet.* **13**, 672–672 (2012).

187. Zhang, J., Hao, X., Yin, M., Xu, T. & Guo, F. Long non-coding RNA in osteogenesis: A new world to be explored. *Bone Joint Res.* **8**, 73–80 (2019).

188. Marchese, F. P., Raimondi, I. & Huarte, M. The multidimensional mechanisms of long noncoding RNA function. *Genome Biol.* **18**, 206 (2017).

189. de Goede, O. M. *et al.* Population-scale tissue transcriptomics maps long non-coding RNAs to complex disease. *Cell* **184**, 2633–2648.e19 (2021).

190. Silva, A. M. *et al.* Long noncoding RNAs: a missing link in osteoporosis. *Bone Res* **7**, 10 (2019).

191. Nardocci, G. *et al.* Identification of a novel long noncoding RNA that promotes osteoblast differentiation. *J. Cell. Biochem.* **119**, 7657–7666 (2018).

192. Chen, X.-F. *et al.* An Osteoporosis Risk SNP at 1p36.12 Acts as an Allele-Specific Enhancer to Modulate LINC00339 Expression via Long-Range Loop Formation. *Am. J. Hum. Genet.* **102**, 776–793 (2018).

193. Roca-Ayats, N. *et al.* Functional characterization of the C7ORF76 genomic region, a prominent GWAS signal for osteoporosis in 7q21.3. *Bone* **123**, 39–47 (2019).

194. Mei, B. *et al.* LncRNA ZBTB40-IT1 modulated by osteoporosis GWAS risk SNPs suppresses osteogenesis. *Hum. Genet.* **138**, 151–166 (2019).

195. Zhou, Y. *et al.* Long Noncoding RNA Analyses for Osteoporosis Risk in Caucasian Women. *Calcif. Tissue Int.* **105**, 183–192 (2019).

196. Zhang, X., Deng, H.-W., Shen, H. & Ehrlich, M. Prioritization of osteoporosis-associated genome-wide

association study (GWAS) single-nucleotide polymorphisms (SNPs) using epigenomics and transcriptomics. *JBMR Plus* **5**, e10481 (2021).

197. Abood, A. & Farber, C. R. Using '-omics' Data to Inform Genome-wide Association Studies (GWASs) in the Osteoporosis Field. *Curr. Osteoporos. Rep.* (2021) doi:10.1007/s11914-021-00684-w.

198. Li, D., Liu, Q. & Schnable, P. S. TWAS results are complementary to and less affected by linkage disequilibrium than GWAS. *Plant Physiol.* **186**, 1800–1811 (2021).

199. Wainberg, M. *et al.* Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* **51**, 592–599 (2019).

200. Sagi, H. C., Young, M. L., Gerstenfeld, L., Einhorn, T. A. & Tornetta, P. Qualitative and quantitative differences between bone graft obtained from the medullary canal (with a Reamer/Irrigator/Aspirator) and the iliac crest of the same patient. *J. Bone Joint Surg. Am.* **94**, 2128–2135 (2012).

201. Farr, J. N. *et al.* Effects of Age and Estrogen on Skeletal Gene Expression in Humans as Assessed by RNA Sequencing. *PLoS One* **10**, e0138347 (2015).

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

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

204. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).

205. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).

206. 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).

207. Wang, L. *et al.* CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **41**, e74 (2013).

208. Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* **16**, 195 (2015).

209. van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–1063 (2015).

210. Barbeira, A. N. *et al.* Integrating Predicted Transcriptome From Multiple Tissues Improves Association Detection. Preprint at https://doi.org/10.1101/292649.

211. Bonewald, L. F. The amazing osteocyte. *J. Bone Miner. Res.* **26**, 229–238 (2011).

212. Youlten, S. E. *et al.* Osteocyte transcriptome mapping identifies a molecular landscape controlling skeletal homeostasis and susceptibility to skeletal disease. *Nat. Commun.* **12**, 2444 (2021).

213. Funari, V. A. *et al.* Cartilage-selective genes identified in genome-scale analysis of non-cartilage and cartilage gene expression. *BMC Genomics* **8**, 165 (2007).

214. Cheng, J. *et al.* Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308**, 1149–1154 (2005).

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

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

217. Zhang, Y. *et al.* lncRNA Neat1 Stimulates Osteoclastogenesis Via Sponging miR-7. *J. Bone Miner. Res.* **35**, 1772–1781 (2020).

218. Zhang, Y., Chen, B., Li, D., Zhou, X. & Chen, Z. LncRNA NEAT1/miR-29b-3p/BMP1 axis promotes osteogenic differentiation in human bone marrow-derived mesenchymal stem cells. *Pathol. Res. Pract.* **215**, 525–531 (2019).

219. Yang, X., Yang, J., Lei, P. & Wen, T. LncRNA MALAT1 shuttled by bone marrow-derived mesenchymal stem cells-secreted exosomes alleviates osteoporosis through mediating microRNA-34c/SATB2 axis. *Aging* vol. 11 8777–8791 Preprint at https://doi.org/10.18632/aging.102264 (2019).

220. Rom, A. *et al.* Regulation of CHD2 expression by the Chaserr long noncoding RNA gene is essential for viability. *Nature Communications* vol. 10 Preprint at https://doi.org/10.1038/s41467-019-13075-8 (2019).

221. Muñoz-Fuentes, V. *et al.* The International Mouse Phenotyping Consortium (IMPC): a functional catalogue of the mammalian genome that informs conservation. *Conserv. Genet.* **19**, 995–1005 (2018).

222. Guo, H.-L. *et al.* LINC00472 promotes osteogenic differentiation and alleviates osteoporosis by sponging miR-300 to upregulate the expression of FGFR2. *Eur. Rev. Med. Pharmacol. Sci.* **24**, 4652–4664 (2020).

223. Modrek, B. & Lee, C. A genomic view of alternative splicing. *Nat. Genet.* **30**, 13–19 (2002).

224. Brotman, S. M. *et al.* Subcutaneous adipose tissue splice quantitative trait loci reveal differences in isoform usage associated with cardiometabolic traits. *Am. J. Hum. Genet.* **109**, 66–80 (2022).

225. Thom, C. S. & Voight, B. F. Genetic colocalization atlas points to common regulatory sites and genes for hematopoietic traits and hematopoietic contributions to disease phenotypes. *BMC Med. Genomics* **13**, 89 (2020).

226. Qi, T. *et al.* Genetic control of RNA splicing and its distinct role in complex trait variation. *Nat. Genet.* **54**, 1355–1363 (2022).

227. Anvar, S. Y. *et al.* Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing. *Genome Biol.* **19**, (2018).

228. Foord, C. *et al.* The variables on RNA molecules: concert or cacophony? Answers in long-read sequencing. *Nat. Methods* **20**, 20–24 (2023).

229. Sarantopoulou, D. *et al.* Comparative evaluation of full-length isoform quantification from RNA-Seq. *BMC Bioinformatics* **22**, 266 (2021).

230. Smith, L. M., Kelleher, N. L. & Consortium for Top Down Proteomics. Proteoform: a single term describing protein complexity. *Nat. Methods* **10**, 186–187 (2013).

231. Harris, S. A. Enger RJ, Riggs BL, Spelsberg TC. *conditionally immortalized human fetal osteoblastic cell* ….

232. Caetano-Lopes, J., Canhão, H. & Fonseca, J. E. Osteoblasts and bone formation. *Acta Reumatol. Port.* **32**,

103–110 (2007).

233. Yen, M.-L. *et al.* Multilineage differentiation and characterization of the human fetal osteoblastic 1.19 cell line: a possible in vitro model of human mesenchymal progenitors. *Stem Cells* **25**, 125–131 (2007).

234. Osipovich, A. B. *et al.* A developmental lineage-based gene co-expression network for mouse pancreatic β-cells reveals a role for Zfp800 in pancreas development. *Development* **148**, (2021).

235. Civelek, M. *et al.* Genetic Regulation of Adipose Gene Expression and Cardio-Metabolic Traits. *Am. J. Hum. Genet.* **100**, 428–443 (2017).

236. Groza, T. *et al.* The International Mouse Phenotyping Consortium: comprehensive knockout phenotyping underpinning the study of human disease. *Nucleic Acids Res.* (2022) doi:10.1093/nar/gkac972.

237. Van Nostrand, E. L. *et al.* A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020).

238. Ling, J. P. *et al.* PTBP1 and PTBP2 Repress Nonconserved Cryptic Exons. *Cell Rep.* **17**, 104–113 (2016).

239. Hensel, J. A. *et al.* Splice factor polypyrimidine tract-binding protein 1 (Ptbp1) primes endothelial inflammation in atherogenic disturbed flow conditions. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2122227119 (2022).

240. Gueroussov, S. *et al.* An alternative splicing event amplifies evolutionary differences between vertebrates. *Science* **349**, 868–873 (2015).

241. Belanger, K., Nutter, C. A., Li, J., Yu, P. & Kuyumcu-Martinez, M. N. A developmentally regulated spliced variant of PTBP1 is upregulated in type 1 diabetic hearts. *Biochem. Biophys. Res. Commun.* **509**, 384–389 (2019).

242. Castaldi, P. J., Abood, A., Farber, C. R. & Sheynkman, G. M. Bridging the splicing gap in human genetics with long-read RNA sequencing: finding the protein isoform drivers of disease. *Hum. Mol. Genet.* **31**, R123–R136 (2022).

243. Green, R. E. *et al.* Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes. *Bioinformatics* **19**, i118–i121 (2003).

244. Wilkie, A. O. M. Dominance and Recessivity. *eLS* 1–10 Preprint at https://doi.org/10.1002/9780470015902.a0005475.pub2 (2018).

245. Ng, A. H. M. *et al.* A comprehensive library of human transcription factors for cell fate engineering. *Nat. Biotechnol.* **39**, 510–519 (2021).

246. Peterson, A. C., Russell, J. D., Bailey, D. J., Westphall, M. S. & Coon, J. J. Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. *Mol. Cell. Proteomics* **11**, 1475–1488 (2012).

247. Wu, C. *et al.* BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.* **10**, R130 (2009).

248. Lips, P. Epidemiology and predictors of fractures associated with osteoporosis. *Am. J. Med.* **103**, 3S–8S; discussion 8S–11S (1997).

249. Humphrey, J. *et al.* Integrative transcriptomic analysis of the amyotrophic lateral sclerosis spinal cord implicates glial activation and suggests new risk genes. *Nature Neuroscience* vol. 26 150–162 Preprint at https://doi.org/10.1038/s41593-022-01205-3 (2023).

250. Schirwani, S. *et al.* Homozygous intronic variants in TPM2 cause recessively inherited Escobar variant of multiple pterygium syndrome and congenital myopathy. *Neuromuscul. Disord.* **31**, 359–366 (2021).

251. Mokbel, N. *et al.* K7del is a common TPM2 gene mutation associated with nemaline myopathy and raised myofibre calcium sensitivity. *Brain* **136**, 494–507 (2013).

252. Meng, L.-B. *et al.* TPM2 as a potential predictive biomarker for atherosclerosis. *Aging* **11**, 6960–6982 (2019).

253. Shin, H., Kim, D. & Helfman, D. M. Tropomyosin isoform Tpm2.1 regulates collective and amoeboid cell migration and cell aggregation in breast epithelial cells. *Oncotarget* **8**, 95192–95205 (2017).

254. Chen, J. & Long, F. β-catenin promotes bone formation and suppresses bone resorption in postnatal growing mice. *J. Bone Miner. Res.* **28**, 1160–1169 (2013).

255. Pina, J. M., Hernandez, L. A. & Keppetipola, N. M. Polypyrimidine tract binding proteins PTBP1 and

PTBP2 interact with distinct proteins under splicing conditions. *PLoS One* **17**, e0263287 (2022).

256. Abood, A. *et al.* Identification of Known and Novel Long Noncoding RNAs Potentially Responsible for the Effects of Bone Mineral Density (BMD) Genomewide Association Study (GWAS) Loci. *J. Bone Miner. Res.* **37**, 1500–1510 (2022).

257. Wiśniewski, J. R. Filter-Aided Sample Preparation for Proteome Analysis. in *Microbial Proteomics: Methods and Protocols* (ed. Becher, D.) 3–10 (Springer New York, 2018). doi:10.1007/978-1-4939-8695-8_1.

258. Navarrete-Perea, J., Yu, Q., Gygi, S. P. & Paulo, J. A. Streamlined Tandem Mass Tag (SL-TMT) Protocol: An Efficient Strategy for Quantitative (Phospho)proteome Profiling Using Tandem Mass Tag-Synchronous Precursor Selection-MS3. *J. Proteome Res.* **17**, 2226–2236 (2018).

259. Orsburn, B. C. Proteome Discoverer-A Community Enhanced Data Processing Suite for Protein Informatics. *Proteomes* **9**, (2021).

260. MacLean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).

261. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

262. Tardaguila, M. *et al.* SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res.* (2018) doi:10.1101/gr.222976.117.

263. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**, 402–408 (2001).

264. Richards, J. B. *et al.* Bone mineral density, osteoporosis, and osteoporotic fractures: a genome-wide association study. *Lancet* **371**, 1505–1512 (2008).

265. Haines, J. L. *et al.* Complement factor H variant increases the risk of age-related macular degeneration. *Science* **308**, 419–421 (2005).

266. Method of the Year 2019: Single-cell multimodal omics. *Nat. Methods* **17**, 1 (2020).

267. Method of the Year 2022: long-read sequencing. *Nat. Methods* **20**, 1 (2023).

268. Salz, R. *et al.* SUsPECT: A pipeline for variant effect prediction based on custom long-read transcriptomes for improved clinical variant annotation. *bioRxiv* 2022.10.23.513417 (2022) doi:10.1101/2022.10.23.513417.

269. Oh, S. S. *et al.* Diversity in Clinical and Biomedical Research: A Promise Yet to Be Fulfilled. *PLoS Med.* **12**, e1001918 (2015).

270. Graham, S. E. *et al.* The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600**, 675–679 (2021).

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

272. Rouhana, J. M. *et al.* ECLIPSER: identifying causal cell types and genes for complex traits through single cell enrichment of e/sQTL-mapped genes in GWAS loci. *bioRxiv* 2021.11.24.469720 (2021) doi:10.1101/2021.11.24.469720.

273. Lopes, K. de P. *et al.* Genetic analysis of the human microglial transcriptome across brain regions, aging and disease pathologies. *Nat. Genet.* (2022) doi:10.1038/s41588-021-00976-y.

274. Bellenguez, C. *et al.* New insights into the genetic etiology of Alzheimer’s disease and related dementias. *Nat. Genet.* **54**, 412–436 (2022).

275. Tian, J. *et al.* Aberrant RNA splicing is a primary link between genetic variation and pancreatic cancer risk. *Cancer Res.* (2022) doi:10.1158/0008-5472.CAN-21-4367.

276. Yamaguchi, K. *et al.* Splicing QTL analysis focusing on coding sequences reveals pathogenicity of disease susceptibility loci. *bioRxiv* 2021.12.30.474578 (2022) doi:10.1101/2021.12.30.474578.

277. Zandi, P. P. *et al.* Amygdala and anterior cingulate transcriptomes from individuals with bipolar disorder reveal downregulated neuroimmune and synaptic pathways. *Nat. Neurosci.* **25**, 381–389 (2022).

278. Tian, Y. *et al.* Novel role of prostate cancer risk variant rs7247241 on PPP1R14A isoform transition through allelic TF binding and CpG methylation. *Hum. Mol. Genet.* **31**, 1610–1621 (2022).

279. Aherrahrou, R. *et al.* Genetic regulation of human aortic smooth muscle cell gene expression and splicing predict causal coronary artery disease genes. *bioRxiv* 2022.01.24.477536 (2022) doi:10.1101/2022.01.24.477536.

280. Kerimov, N. *et al.* A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat. Genet.* **53**, 1290–1299 (2021).

281. Aygün, N. *et al.* Brain-trait-associated variants impact cell-type-specific gene regulation during neurogenesis. *Am. J. Hum. Genet.* **108**, 1647–1668 (2021).

282. Mu, Z. *et al.* The impact of cell type and context-dependent regulatory variants on human immune traits. *Genome Biol.* **22**, 122 (2021).

283. Chen, B. Y. *et al.* ColocQuiaL: a QTL-GWAS colocalization pipeline. *Bioinformatics* **38**, 4409–4411 (2022).

284. Stanzick, K. J. *et al.* Discovery and prioritization of variants and genes for kidney function in >1.2 million individuals. *Nat. Commun.* **12**, 4350 (2021).

285. Gao, Y. *et al.* Genome-wide association study reveals novel loci for adult type 1 diabetes in a 5-year nested case-control study. *World J. Diabetes* **12**, 2073–2086 (2021).

286. Patro, C. P. K., Nousome, D., Glioma International Case Control Study (GICC) & Lai, R. K. Meta-Analyses of Splicing and Expression Quantitative Trait Loci Identified Susceptibility Genes of Glioma. *Front. Genet.* **12**, 609657 (2021).

287. Díez-Obrero, V. *et al.* Genetic Effects on Transcriptome Profiles in Colon Epithelium Provide Functional Insights for Genetic Risk Loci. *Cell Mol Gastroenterol Hepatol* **12**, 181–197 (2021).

288. Navarro, E. *et al.* Dysregulation of mitochondrial and proteolysosomal genes in Parkinson's disease myeloid cells. *Nature Aging* vol. 1 850–863 Preprint at https://doi.org/10.1038/s43587-021-00110-x (2021).

289. Akula, N. *et al.* Deep transcriptome sequencing of subgenual anterior cingulate cortex reveals cross-diagnostic and diagnosis-specific RNA expression changes in major psychiatric disorders.

Neuropsychopharmacology vol. 46 1364–1372 Preprint at https://doi.org/10.1038/s41386-020-00949-5 (2021).

290. Liu, S. *et al.* Brain transcriptional regulatory architecture and schizophrenia etiology converge between East Asian and European ancestral populations. *bioRxiv* 2021.02.04.922880 (2021) doi:10.1101/2021.02.04.922880.

291. Zhang, T. *et al.* Cell-type-specific meQTLs extend melanoma GWAS annotation beyond eQTLs and inform melanocyte gene-regulatory mechanisms. *The American Journal of Human Genetics* vol. 108 1631–1646 Preprint at https://doi.org/10.1016/j.ajhg.2021.06.018 (2021).

292. Yang, H.-S. *et al.* Genetics of Gene Expression in the Aging Human Brain Reveal TDP-43 Proteinopathy Pathophysiology. *Neuron* **107**, 496–508.e6 (2020).

293. Guo, Z. *et al.* Alternative splicing related genetic variants contribute to bladder cancer risk. *Mol. Carcinog.* **59**, 923–929 (2020).

294. Yuan, M., Yu, C. & Yu, K. Association of human XPA rs1800975 polymorphism and cancer susceptibility: an integrative analysis of 71 case–control studies. *Cancer Cell Int.* **20**, 1–19 (2020).

295. Nurnberg, S. T. *et al.* Genomic profiling of human vascular cells identifies TWIST1 as a causal gene for common vascular diseases. *PLoS Genet.* **16**, e1008538 (2020).

296. Li, Y. I., Wong, G., Humphrey, J. & Raj, T. Prioritizing Parkinson's disease genes using population-scale transcriptomic data. *Nat. Commun.* **10**, 994 (2019).

297. Rotival, M., Quach, H. & Quintana-Murci, L. Defining the genetic and evolutionary architecture of alternative splicing in response to infection. *Nat. Commun.* **10**, 1671 (2019).

298. Saferali, A. *et al.* Analysis of genetically driven alternative splicing identifies FBXO38 as a novel COPD susceptibility gene. *PLoS Genet.* **15**, e1008229 (2019).

299. Gawronski, K. A. B. *et al.* Evaluating the contribution of cell-type specific alternative splicing to variation in lipid levels. Preprint at https://doi.org/10.1101/659326.

300. Raj, T. *et al.* Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's

disease susceptibility. *Nature Genetics* vol. 50 1584–1592 Preprint at https://doi.org/10.1038/s41588-018-0238-1 (2018).

301. Liu, B. *et al.* Genetic Regulatory Mechanisms of Smooth Muscle Cells Map to Coronary Artery Disease Risk Loci. *Am. J. Hum. Genet.* **103**, 377–388 (2018).

302. Kurmangaliyev, Y. Z., Sutormin, R. A., Naumenko, S. A., Bazykin, G. A. & Gelfand, M. S. Functional implications of splicing polymorphisms in the human genome. *Hum. Mol. Genet.* **22**, 3449–3459 (2013).

303. Yang, Y.-C. T. *et al.* CLIPdb: a CLIP-seq database for protein-RNA interactions. *BMC Genomics* **16**, 51 (2015).

304. Passacantilli, I., Frisone, P., De Paola, E., Fidaleo, M. & Paronetto, M. P. hnRNPM guides an alternative splicing program in response to inhibition of the PI3K/AKT/mTOR pathway in Ewing sarcoma cells. *Nucleic Acids Res.* **45**, 12270–12284 (2017).

305. Al-Barghouthi, B. M. *et al.* Transcriptome-wide Association Study and eQTL colocalization identify potentially causal genes responsible for bone mineral density GWAS associations. *bioRxiv* 2021.10.12.464046 (2022) doi:10.1101/2021.10.12.464046.

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

307. Gry, M. *et al.* Correlations between RNA and protein expression profiles in 23 human cell lines. *BMC Genomics* **10**, 365 (2009).

308. Conesa, A., Nueda, M. J., Ferrer, A. & Talón, M. maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics* **22**, 1096–1102 (2006).

309. Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *R J.* **8**, 289–317 (2016).

310. Smedley, D. *et al.* BioMart--biological queries made easy. *BMC Genomics* **10**, 22 (2009).

311. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).

312. Pers, T. H., Timshel, P. & Hirschhorn, J. N. SNPsnap: a Web-based tool for identification and annotation of matched SNPs. *Bioinformatics* **31**, 418–420 (2015).