PORTLAND PRESS

Research Article

Integrative genomics analysis of eQTL and GWAS summary data identifies *PPP1CB* as a novel bone mineral density risk genes

Yu Zhai[1,2,*], Lu Yu[1,*], Yang Shao[2] and ⓘ Jianwei Wang[2]

[1]Changzhou Hospital Affiliated to Nanjing University of Chinese Medicine, Changzhou 213003, Jiangsu, China; [2]Wuxi Hospital Affiliated to Nanjing University of Chinese Medicine, Wuxi 214071, Jiangsu, China

Correspondence: Jianwei Wang (172195614@qq.com)

OPEN ACCESS

In recent years, multiple genome-wide association studies (GWAS) have identified numerous susceptibility variants and risk genes that demonstrate significant associations with bone mineral density (BMD). However, exploring how these genetic variants contribute risk to BMD remains a major challenge. We systematically integrated two independent expression quantitative trait loci (eQTL) data ($N = 1890$) and GWAS summary statistical data of BMD ($N = 142,487$) using *Sherlock* integrative analysis to reveal whether expression-associated variants confer risk to BMD. By using *Sherlock* integrative analysis and MAGMA gene-based analysis, we found there existed 36 promising genes, for example, *PPP1CB*, *XBP1*, and *FDFT1*, whose expression alterations may contribute susceptibility to BMD. Through a protein–protein interaction (PPI) network analysis, we further prioritized the *PPP1CB* as a hub that has interactions with predicted genes and BMD-associated genes. Two eSNPs of rs9309664 ($P_{eQTL} = 1.42 \times 10^{-17}$ and $P_{GWAS} = 1.40 \times 10^{-11}$) and rs7475 ($P_{eQTL} = 2.10 \times 10^{-6}$ and $P_{GWAS} = 1.70 \times 10^{-7}$) in *PPP1CB* were identified to be significantly associated with BMD risk. Consistently, differential gene expression analysis found that the *PPP1CB* gene showed significantly higher expression in low BMD samples than that in high BMD samples based on two independent expression datasets ($P = 0.0026$ and $P = 0.043$, respectively). Together, we provide a convergent line of evidence to support that the *PPP1CB* gene involves in the etiology of osteoporosis.

Introduction

Osteoporosis is a common age-related complex disease. The disease leads to a huge economic burden on health care systems with a cost of about $17 billion per year in the United States, and it is expected to be $25.5 billion annually by the year of 2025 [1]. Bone mineral density (BMD) is the most commonly used indicator to assess the increased risk of a fracture that is the characteristics of osteoporosis [2]. BMD is highly influenced by genetic factors, and the narrow-sense heritability of BMD has been estimated to be approximately 85% [3–5]. Consequently, there exists a considerable interest in identifying the genetic basis of osteoporosis for developing effective methods for its treatment and prevention.

Genome-wide association study (GWAS) is a good approach that is capable of simultaneously examining the genetic links between millions of variants and phenotypes of interest. In the recent decade, numerous GWASs of osteoporosis and BMD have successfully reported multiple genetic variants and related susceptible genes that show robust associations [4,6–10]. These identified genetic loci from GWAS contain a wealth of information with the potential to explore new risk genes and molecular pathways that have critical roles in bone biology and inform drug discovery [6,11]. Nevertheless, it remains elusive on how these genetic loci confer risk to osteoporosis and BMD. Additionally, many genetic loci with weak

or modest effects were hard to be detected in a single GWAS study due to their genome-wide threshold of strict statistical significance. Thus, more studies are warranted to uncover the underlying effects of the small-to-modest GWAS association signals on BMD that may be conducive to understand the missing heritability of this trait.

Although more BMD-associated genetic loci warrant to be identified, growing evidence has strongly demonstrated that the altered expression of gene has a crucial part in the pathogenesis of osteoporosis [12–14]. Furthermore, recent studies [15–18] have employed systematically integrative approaches to integrate the GWAS summary data and expression quantitative trait loci (eQTL) data to explore the potent regulatory effects of the risk variants in reported GWAS. He et al. [15] introduced a Bayesian statistical approach of *Sherlock* to systematically reveal the cis- and trans-regulatory effects of risk genes on complex diseases based on GWAS summary data and eQTL data. By using this bioinformatics tool, numerous studies have identified many novel risk genes, which cannot be found with the use of the GWAS approach alone, for different complex traits, such as schizophrenia [19], gout disease [20], and major depressive disorders [21,22].

The primary goal of the current investigation is to explore whether genetic variants linked with gene expression confer risk to BMD and identify BMD-associated genes based on combining both GWAS and eQTL data by using a Bayesian method of *Sherlock*. To further validate the potent roles of these identified BMD-associated genes in osteoporosis etiology, we re-performed the integrative analysis in an independent eQTL dataset and applied additional genomic analyses based on RNA expression data.

Materials and methods
Bone mineral density GWAS data
We employed a large-scale GWAS summary statistics data on BMD [10] for searching risk genetic variants. Briefly, this BMD-associated GWAS study was reported recently by the Genetic Factors for Osteoporosis (GEFOS) Consortium, which comprises 142,487 individuals in total. All individuals have signed the informed consent and the ethical approval was obtained from the Northwest Multi-centre Research Ethics Committee. Data were imputed centrally based on the UK10K/1000G combined imputation panel (hg19). Only SNPs down to a MAF of 0.1% and with an info-score threshold of >0.4 were included for analysis. The association information of genetic variants in GWAS including the name of each SNP and related *P* value was employed as input in *Sherlock* analysis and MAGMA analysis. There were a total of 17,166,350 SNPs in the chosen GWAS employed as input in the current investigation. For more detailed information, please refer to the original published article [10] and the GEFOS website (http://www.gefos.org/).

Discovery eQTL data
We used the monocyte eQTL data that were from a single-center cohort study of the Gutenberg Heart Study (GHS) as discovery eQTL, including a number of 1490 unrelated study participants with both DNA and RNA available. Informed consent was signed from all subjects. The *Affymetrix* Genome-Wide Human SNP Array 6.0 (http://www.affymetrix.com) was applied to conduct the genotyping for each sample. Based on the standard criteria of quality control for SNPs in the Affymetrix SNP Array 6.0 including the *P* value for Hardy–Weinberg equilibrium $>1 \times 10^{-4}$, the calling rate of genotype >98%, and the frequency of minor allele > 0.01, we removed 225,042 low-quality SNPs and obtained 675,350 SNPs for further analysis. With the use of the *Illumina* HT-12 v3 BeadChip (http://www.Illumina.com), genome-wide expression analysis was conducted based on RNA samples from monocytes. There were a total of 37,804 genes included in the *Illumina* HT-12 BeadChip. Of these genes, a number of 22,305 genes were treated as being prominently expressed. We removed 8058 not well-characterized genes and used 12,808 well-characterized genes to carry out a subsequent eQTL analysis. For more information, please refer to the original published article [23].

Independent validation eQTL data
We also employed an independent eQTL dataset published by Dixon and coworkers [24], which have created a global map of the effects of polymorphism on gene expression. Ethical approval was given by the Multicentre Research Ethics Committees (U.K.), and written informed consent was given by all participants. Whole-genome genotyping was conducted according to manufacturers' instructions using the Human Hap300 Genotyping BeadChip (Illumina) and the Sentrix Human-1 Genotyping BeadChip in a BeadChip with full automation. Genome-wide expression analysis was carried out on lymphoblastoid RNA samples using the U133 Plus 2.0 GeneChips (Affymetrix), according to the manufacturers' protocol. After strict quality control, genotypes and gene expression of 400 participants were used to generate eQTL resources. Please refer to the original paper for more detailed information [24].

Sherlock integrative analysis using discovery eQTL data

Given the vast majority of the GWAS-identified variants associated with traits of interest are located in non-coding genomic regions [25], it is likely to infer that the identified susceptibility variants influence the expression level of the relevant gene rather than the function of its protein. Based on the putative assumption that specific gene expression may convey susceptibility to BMD, we employed the approach of *Sherlock* integrative analysis to integrate GWAS summary statistics on BMD from Kemp et al. [10] and circulating monocyte eQTL data from Zeller et al. [23]. The statistical inference procedures of the *Sherlock* method are described as following: *Sherlock* first searches expression-associated SNPs (as called eSNPs) in the human monocyte samples with the use of the eQTL data from the study of Zeller et al. [23]. Second, *Sherlock* tool assesses the potent association of eSNPs with BMD using the GWAS summary data from Kemp et al. [10]. There exists three scenarios: (1) A positive score would be recorded based on a specific eSNP of a gene in the chosen GWAS shows a significant association with BMD; (2) A negative score would be assigned based on a specific eSNP of this gene shows a non-significant association with BMD; (3) No score would be assigned based on an SNP was not eSNP but shows a significant association with BMD. The summed score of a specific gene was depended on the number of SNPs with the evidence combined from both GWAS and eQTL data. Concerning each gene, *Sherlock* software conducted a Bayesian statistical inference with the use of the integrated data of the potent eSNPs of the gene to examine whether the alteration in the expression of the specific gene has any influence on BMD risk. Through computing the logarithm of Bayes factor, *Sherlock* tool predicts BMD-associated genes by combining the data from eQTL and GWAS summary statistics. With a comparison of existing traditional analysis, which commonly neglects moderate SNPs, *Sherlock* tool based on an effective Bayesian model employs SNPs in GWAS with moderately and strongly genetic association signals. For *Sherlock* analysis, the method of Bonferroni correction was applied to correct the *P*-values of genes.

Sherlock integrative analysis using independent eQTL data

Subsequently, to replicate whether these identified genes are genuine BMD-associated risk genes, we conducted a further *Sherlock* integrative analysis based on an independent eQTL data (biological validation), which were reported by a study of Dixon and his colleagues [24]. All the parameter settings of *Sherlock* analysis were the same with that of discovery eQTL data. Consistently, the *P*-value of each gene was corrected by utilized the Bonferroni correction method.

MAGMA gene-based analysis

We used the software of Multi-marker Analysis of GenoMic Annotation (MAGMA) [26] as an independent bioinformatics tool to perform a gene-based enrichment analysis for technically validate the identified genes from *Sherlock* Bayesian analysis. For the MAGMA tool, we could extract the name of each SNP with its *P*-value from BMD GWAS summary statistics as input to identify the significant association signals at a gene level. SNPs mapped into a specific gene or the genomic region extended ± 20 kb downstream or upstream of the gene were chosen to identify multiple variants convergent effects and collectively calculate the *P*-value of the gene risk to BMD. The linkage disequilibrium (LD) information between chosen SNPs was computed based on the 1000 Genome European panel. The Bonferroni correction method was used for multiple testing.

Null GWAS data

To make sure these identified BMD-associated genes because of genetic biology instead of random events, we carried out a MAGMA gene-based analysis with the same parameters by using an artificial GWAS dataset (as a negative control dataset). The control GWAS was derived from a real genetic dataset with a number of 3960 participants published by Landi and colleagues [27]. We randomly assigned the phenotype of BMD to each participant to construct a Null trait. In view of the Null GWAS was supposed to be no true genetic effect, the power of this analysis based on relatively small sample size is not an issue.

Pathway enrichment analysis

To reveal the biological interpretation of these prioritized BMD-associated genes from the discovery eQTL data, we used the ClueGO [28], an easy use Cytoscape plug-in [29], to create a functionally organized pathway term network. We used gene-set data from the two most recent sources of KEGG [30] and GO [31] (including biological process, molecular function, and cellular component), ensuring an up-to-date functional analysis. We also employed the method of "*GO Term Fusion*", which fusion of GO parent-child terms based on similar associated genes, to reduce the redundancies of GO terms. Based on the two-sided test of hypergeometric distribution, the *P* values were

calculated by depletion or enrichment for groups and terms. The Bonferroni step down correction was applied to correct P values for multiple testing.

Computer-based permutation analysis

To reveal whether *Sherlock*-identified risk genes in the discovery stage (corrected $P < 0.05$: $N = 147$ genes; raw $P < 0.05$: $N = 2064$ genes) were significantly overlapped with genes identified from Sherlock validation analysis (corrected $P < 0.05$: $N = 98$ genes; raw $P < 0.05$: $N = 2031$ genes) or MAGMA analysis (corrected $P < 0.05$: $N = 1106$ genes), respectively, we conducted a computer-based permutation analysis [32]. For this permutation analysis, we randomly selected the number of genes as same as significantly identified genes from background genes for 100,000 times and documented the overlapped rate with genes from the *Sherlock* discovery stage. Then, we counted how many times the counts of overlapped genes were larger than the number detected from real data 100,000 times. The probability of the observed number considered as the empirical P value.

Protein–protein interaction (PPI) network

Multiple lines of evidence have documented that genes associated with complicated disorders are more tend to be interacted [14,33,34]. Thus, PPI network-based analysis has been extensively applied to search for groups of functionally related genes that may collectively contribute risk to complex disease. We conducted a PPI network-based analysis of these identified BMD-associated risk genes based on the GeneMANIA database (http://www.genemania.org) [35]. The GeneMANIA is a flexible and user-friendly tool for analyzing a list of genes to infer the functions of inputted genes and narrowing down the number of genes for molecular experiments. If inputting a query list of genes, the GeneMANIA tool extends the inputted genes with functionally similar genes based on current available proteomics and genomics data.

Identification of BMD-associated genes expression profiles

Sherlock analysis identifies the disease-associated gene based on the assumption that risk gene abnormal expression may implicate in the pathogenesis of disease of interest. To reveal whether the identified susceptibility genes are differentially expressed in low BMD compared with high BMD group, we obtained two existing RNA expression datasets available in the NCBI's GEO database (Accession Nos. GES2208 and GSE7429). GSE2208 contains transcriptome data of 10 low BMD subjects and 10 high BMD subjects. GSE7429 contains RNA expression data of 9 low BMD subjects and 10 high BMD subjects. For more detailed information on the sample collections and other procedures of quality control, please find two original articles [24,36]. The R package of *corrplot* was utilized to show the differential co-expression patterns of these identified risk genes between high BMD and low BMD group. The difference between high BMD and low BMD groups was compared for significance with the Student's T-test. A P value < 0.05 was considered statistically significant.

Results

Prioritization of BMD-associated risk genes

The workflow of the present study design is shown in Figure 1. To determine whether these genes with abnormal expression may convey risk to BMD, we applied the *Sherlock* tool to integrate GWAS summary data with a total of 17,166,350 SNPs ($N = 142,487$ samples) with eQTL data based on 1490 samples. The *Sherlock* integrative analysis, which is often used to find disease-associated genes by matching patterns of GWAS and eQTL, was carried out for the Bayesian statistical inference. Based on this Bayesian method, we identified 147 top BMD-associated genes whose abnormal expression may influence BMD risk at the threshold of $P < 1.57 \times 10^{-6}$ (Bonferroni corrected $P < 0.05$; Supplemental Table S1) and 2064 suggestive BMD-associated genes ($P < 0.05$). For each of these identified significant genes, for example, *NR1H3*, *MVK*, *ACP2*, *EPB41L2*, *ZEP57*, *PBX1*, and *PPP1CB*, at least one SNP showed significant association with the expression of this gene and BMD risk simultaneously, inferring that these SNPs are promising functional candidates with potential cis- and/or trans-regulatory effects on risk genes. Furthermore, we performed a pathway-based analysis for these 147 significant genes and there were 11 pathways showing significant enrichment (Table 1); such as the pathway of negative regulation of type 2 immune response ($P = 0.00018$), inactivation of MAPK activity ($P = 0.0011$), and PPAR signaling pathway ($P = 0.0021$).

Validation of BMD-associated risk genes

To further validate our results in the discovery stage, we replicated the findings of *Sherlock* integrative analysis with the use of an independent of Dixon et al. eQTL dataset (Replication dataset: $N = 400$). By using the same parameter
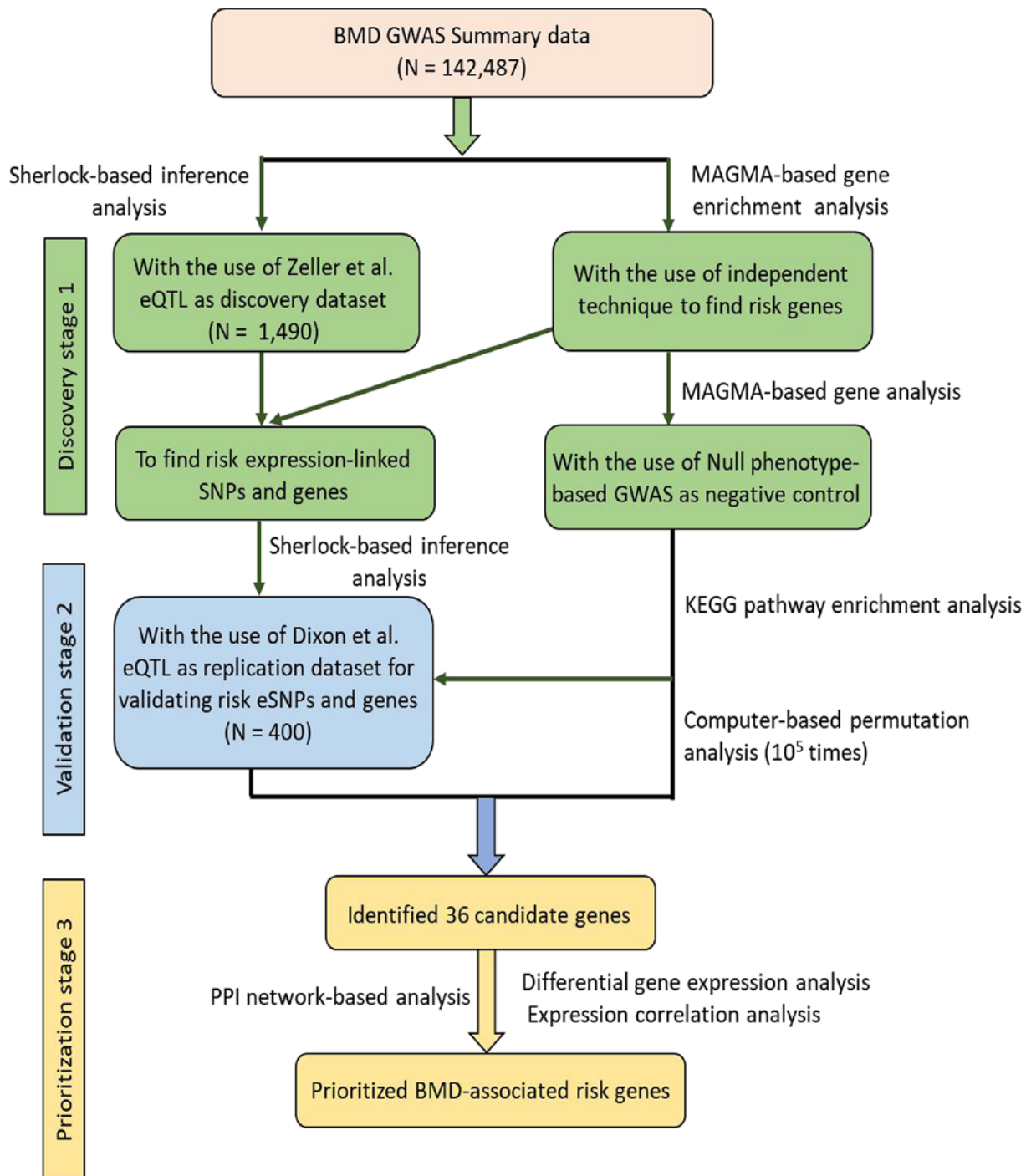
Figure 1. The workflow of prioritizing the BMD-associated genes

settings, *Sherlock* integrative analysis based on independent eQTL data identified 98 significant genes (Supplementary Table S3; Bonferroni corrected $P < 0.05$) and 2031 suggestive genes ($P < 0.05$). Among these 147 significant genes identified from the discovery stage, there were 48 genes validated in the replication stage ($P < 0.05$). In both the discovery and validation stage, there were 11 BMD-associated genes reached Bonferroni correction significance, namely *CPT1A*, *CKB*, *MRPL21*, *ZFP57*, *ACP2*, *PPP1CB*, *ST7L*, *EPB41L2*, *PBX1*, *AMT*, and *VNN3* (corrected *P* values < 0.05; Supplementary Figure S1). Multiple cis-SNPs in these identified genes convey risk to BMD (Supplementary Tables S2 and S4).

Table 1 Significant pathways enriched by identified BMD-associated genes based on discovery dataset

Pathway ID	Pathway terms	Associated genes proportion	*P* value	Corrected *P* value	Associated genes
GO:0002829	Negative regulation of type 2 immune response	0.200	0.00018	0.0018	*ARG2, SOCS5, STAT6*
GO:1904294	Positive regulation of ERAD pathway	0.200	0.00018	0.0018	*ATXN3, EDEM2, XBP1*
GO:0071498	Cellular response to fluid shear stress	0.143	0.00050	0.0045	*KLF2, SOCS5, XBP1*
GO:0000188	Inactivation of MAPK activity	0.111	0.00106	0.0085	*DUSP1, DUSP12, DUSP3*
GO:1901800	Positive regulation of proteasomal protein catabolic process	0.047	0.00122	0.0085	*ATXN3, DAB2, EDEM2, SOCS5, XBP1*
KEGG:03320	PPAR signaling pathway	0.056	0.00210	0.0126	*CPT1A, FABP7, FADS2, NR1H3*
KEGG:04142	Lysosome	0.041	0.00235	0.0118	*ACP2, AP3D1, CTSB, GUSB, IDUA*
GO:0048246	Macrophage chemotaxis	0.079	0.00287	0.0115	*LGALS3, MMP28, S100A8*
GO:1904375	Regulation of protein localization to cell periphery	0.043	0.00529	0.0159	*CIB1, DAB2, EPB41L2, SPTBN1*
GO:0030170	Pyridoxal phosphate binding	0.050	0.01034	0.0207	*GCAT, KYAT3, PYGM*
GO:1901880	Negative regulation of protein depolymerization	0.048	0.01131	0.0113	*CIB1, SPTBN1, TRIOBP*

MAGMA-based gene enrichment analysis of BMD

To ensure the reliability of these identified BMD-risk genes, we carried out an independent technical approach of MAGMA gene-based analysis. We found a number of 1106 genes were significantly associated with the phenotype of BMD based on Bonferroni correction for multiple testing (corrected $P < 0.05$). The top-ranked association signals were of *CPED1* (corrected $P = 7.98 \times 10^{-202}$), *CCDC170* (corrected $P = 1.78 \times 10^{-182}$), and *WNT16* (corrected $P = 3.96 \times 10^{-137}$). There were 36 significant genes from MAGMA analysis overlapped with genes identified from *Sherlock* analysis (correlated $P < 0.05$ of genes from the discovery stage; Figure 2A and Table 2). We observed that none of these 36 identified genes show significant associations in the MAGMA analysis of Null GWAS (Table 2). In addition, a number of 86 MAGMA-identified significant genes were overlapped with *Sherlock*-identified genes ($P < 0.05$ of genes from the discovery stage; Supplementary Figure S2).

Furthermore, we found that no matter significant or suggestive genes identified in the *Sherlock* discovery stage were significantly higher overlapped with genes identified in the *Sherlock* validation stage and MAGMA validation (permutation $P < 1.0 \times 10^{-5}$; see Figure 2B,C and Supplementary Figure S3A,B). More interestingly, comparing the MAGMA results of Null GWAS, we observed that *Sherlock*-identified genes from the discovery and validation stage have significantly high overlap rates with MAGMA-identified genes from BMD GWAS than that from Null GWAS (Figure 3A,B).

PPI analysis of 36 BMD-associated risk genes identified by Sherlock analysis

To determine whether these 36 BMD-associated risk genes identified by *Sherlock* and MAGMA analysis interact with each other, we conducted a PPI network-based analysis using interactions of genetic interaction, prediction links, co-expression, co-localization, and physical interactions based on the GeneMANIA database [35]. Figure 4 shows that these BMD-associated risk genes are constructed a biological network, indicating there exist highly biological connections between these identified genes. For example, the hub genes of *PPP1CB, XBP1, FDFT1*, and *SPTBN1* have the most number of interactions with both predicted genes and BMD-associated genes (Figure 4). Additionally, the hub gene of *XBP1* shows strong evidence of physical interactions with BMD-associated genes of *H1F0* and predicted genes of *ZNF440, SRSF1*, and *H1FX* based on the interactions from BIOGRID and IREF databases [37,38]. The hub gene of *PPP1CB* has genetic interactions with BMD-associated genes of *SPTBN1, TRIOP*, and *SSH2* (Figure 4) based on a genome-wide human genetic interaction map [39].
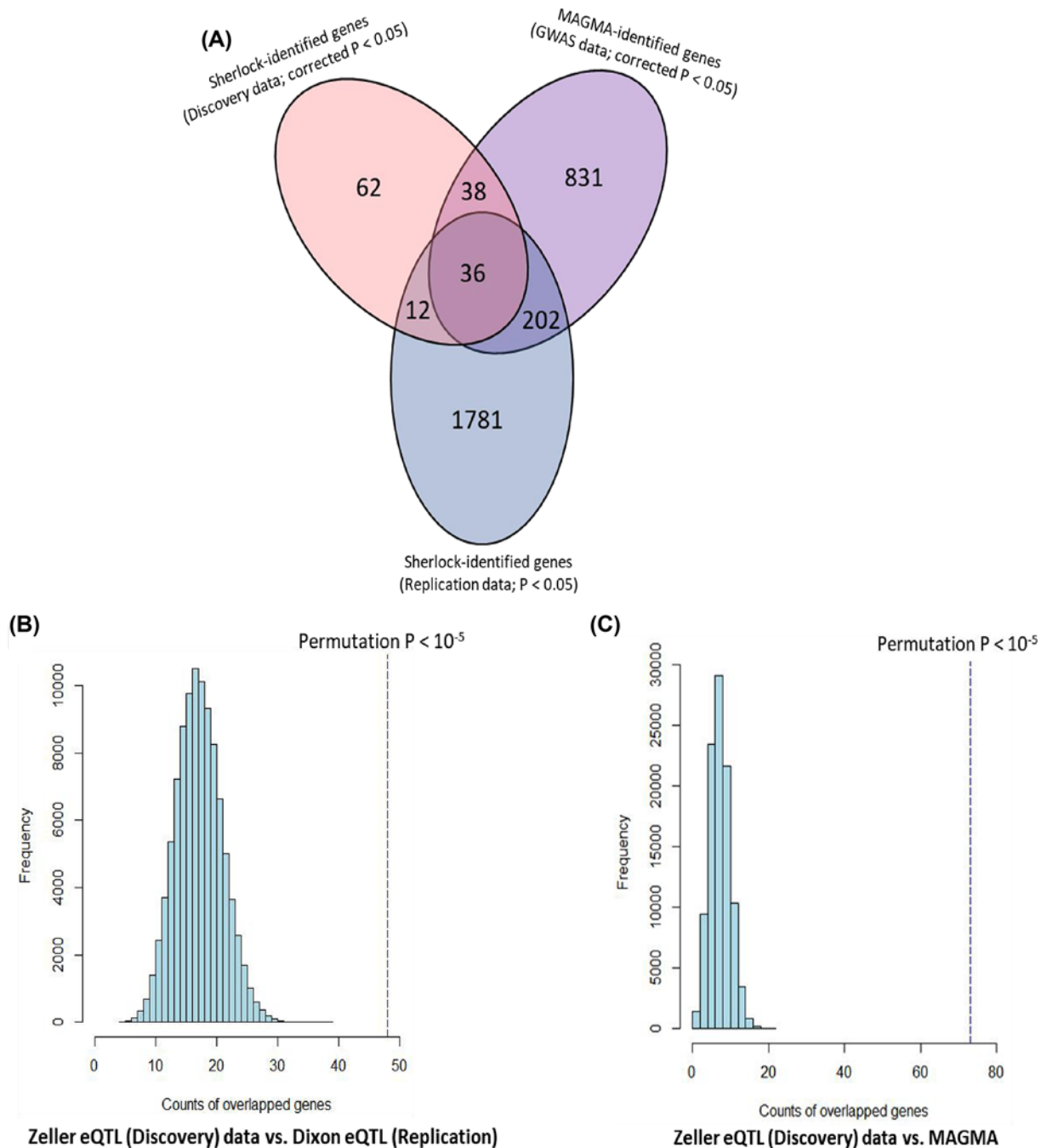
Figure 2. Convergent evidence of BMD-associated genes from independent datasets

(**A**) Venn plot of BMD-relevant genes based on three datasets: *Sherlock*-identified genes from the discovery dataset (corrected $P<0.05$); *Sherlock*-identified genes from the validation dataset ($P<0.05$); MAGMA-identified genes from the GWAS dataset (corrected $P<0.05$). (**B**) Computer-based permutation analysis (100,000 times) of the counts of risk genes from Zeller eQTL dataset (corrected $P<0.05$; discovery stage) overlapped with that from Dixon eQTL dataset ($P<0.05$; Replication stage). (**C**) Computer-based permutation analysis (100,000 times) of the counts of risk genes from Zeller eQTL dataset (corrected $P<0.05$; discovery stage) overlapped with that from MAGMA-based dataset (corrected $P<0.05$; technical validation stage).

Table 2 Identification of BMD-associated risk genes from multiple omics datasets

Gene	LBF	Sherlock-identified P value (discovery dataset; corrected $P<0.05$)	Sherlock-identified P value (replication dataset; $P<0.05$)	MAGMA-identified P value (BMD GWAS; corrected $P<0.05$)	MAGMA-identified P value (Null GWAS; corrected $P<0.05$)
MSRA	10.59	7.87E-07	1.52E-06	5.29E-46	Non-significance
BLK	9.14	7.87E-07	9.26E-05	4.61E-35	Non-significance
ACP2	7.50	7.87E-07	7.59E-07	4.35E-14	Non-significance
EPB41L2	7.48	7.87E-07	7.59E-07	1.96E-18	Non-significance
ACSS2	6.84	7.87E-07	1.58E-03	6.81E-12	Non-significance
ZFP57	6.82	7.87E-07	7.59E-07	2.18E-06	Non-significance
MARK3	6.76	7.87E-07	1.82E-05	1.66E-37	Non-significance
FDFT1	6.66	7.87E-07	6.68E-05	5.00E-18	Non-significance
PPP1CB	6.62	7.87E-07	7.59E-07	4.01E-13	Non-significance
SPTBN1	6.61	7.87E-07	9.72E-05	9.60E-72	Non-significance
AMT	6.59	7.87E-07	7.59E-07	3.43E-11	Non-significance
MPHOSPH9	6.53	7.87E-07	8.81E-05	2.85E-09	Non-significance
CPT1A	6.49	7.87E-07	7.59E-07	1.92E-15	Non-significance
XBP1	6.43	7.87E-07	1.52E-06	1.49E-13	Non-significance
CLDN23	6.38	7.87E-07	1.67E-05	8.60E-23	Non-significance
H1F0	6.36	7.87E-07	4.28E-04	7.94E-08	Non-significance
SSH2	6.27	7.87E-07	9.11E-05	4.85E-16	Non-significance
LACTB2	6.27	7.87E-07	2.43E-05	2.19E-16	Non-significance
VNN3	6.25	7.87E-07	7.59E-07	8.52E-07	Non-significance
DGKQ	6.12	7.87E-07	6.83E-05	3.96E-16	Non-significance
LAMB2	6.08	7.87E-07	2.83E-02	6.99E-12	Non-significance
TRIOBP	5.96	7.87E-07	1.03E-04	2.42E-06	Non-significance
TPCN2	5.96	7.87E-07	1.52E-05	1.57E-21	Non-significance
BAG5	5.94	7.87E-07	1.52E-06	1.95E-25	Non-significance
MRPL21	5.84	1.57E-06	7.59E-07	1.67E-10	Non-significance
CTSB	5.79	1.57E-06	2.43E-05	8.22E-19	Non-significance
NME4	5.71	1.57E-06	2.58E-05	6.83E-13	Non-significance
TRIP11	5.70	1.57E-06	1.52E-06	1.11E-08	Non-significance
PIGN	5.55	1.57E-06	1.52E-06	2.99E-07	Non-significance
ATXN3	5.46	1.57E-06	4.56E-04	8.28E-09	Non-significance
ST7L	5.40	1.57E-06	7.59E-07	2.38E-17	Non-significance
WARS2	5.35	1.57E-06	1.52E-05	2.64E-11	Non-significance
CLMN	5.14	1.57E-06	2.38E-03	1.55E-08	Non-significance
CKB	5.14	1.57E-06	7.59E-07	9.61E-38	Non-significance
UBE2E3	5.06	1.57E-06	1.93E-04	2.85E-08	Non-significance
CIB1	5.06	1.57E-06	1.52E-06	5.56E-07	Non-significance

Differential expression analysis of identified genes in high BMD and low BMD subjects

Furthermore, we tested the co-expression patterns of these identified genes between high BMD and low BMD subjects based on two independent RNA expression datasets (i.e. GSE2208 and GSE7429). Based on the Pearson correlation analysis, we found that the co-expression patterns of these genes showed obvious differences between high and low BMD (Figures 5A,B and 6A,B). Subsequently, we first performed a differential gene expression (DGE) analysis to test the expression level of these genes in high BMD and low BMD subjects in the GSE2208 dataset. We found the genes of $PPP1CB$ ($P = 0.0026$), $FDFT1$ ($P = 0.0057$), and $XBP1$ ($P = 0.0023$) showed significant up-regulated expression in low BMD subjects compared with high BMD subjects (Figure 5C–E). Consistently, we found the $PPP1CB$ gene showed significantly higher expression in low BMD subjects than that in high BMD subjects with the use of an independent dataset of GSE7429 ($P = 0.043$; Figure 6C). In the dataset of GSE7429, we also detected that $LAMB2$ ($P = 0.022$) and $ATXN3$ ($P = 0.038$) were significantly differential expressed between low BMD and high BMD subjects (Figure 6D,E). In addition, several genes showed marginal evidence of differences between low BMD and high BMD subjects in both datasets (Supplementary Figures S4 and S5).
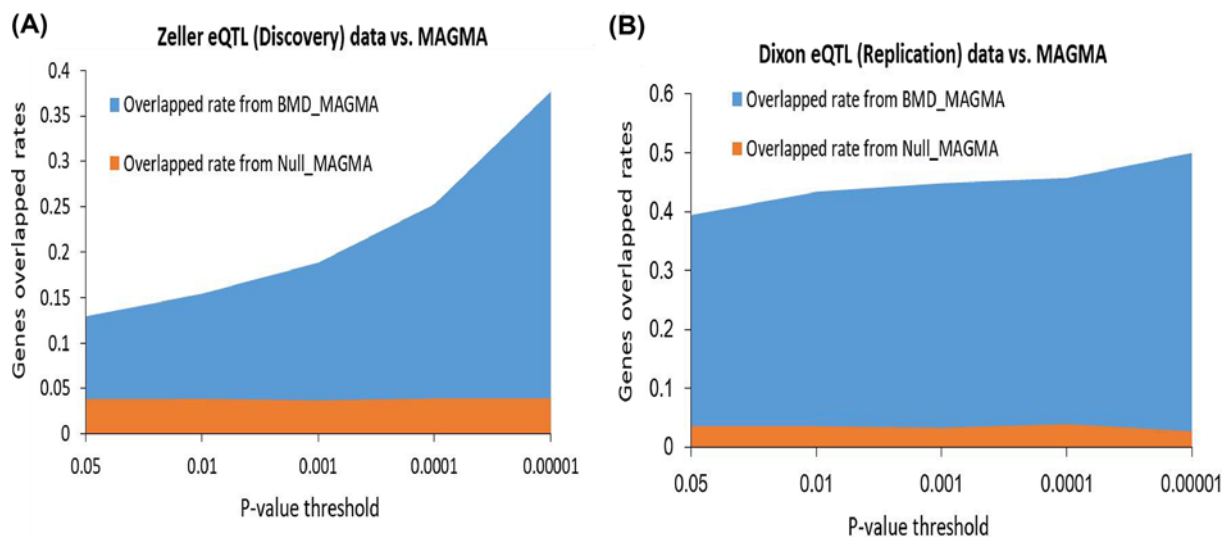
Figure 3. Genes identified from Sherlock Bayesian analysis validated by MAGMA-identified genes

(**A**) Sherlock-identified genes from Zeller eQTL dataset (discovery stage) were higher overlapped with MAGMA-identified genes from BMD GWAS (Dataset #1) than that from Null-based GAWS (Dataset #2) at five different *P*-value thresholds of 0.05, 0.01, 0.001, 0.0001, and 0.00001. (**B**) Sherlock-identified genes from Dixon eQTL dataset (validation stage) were higher overlapped with MAGMA-identified genes from BMD GWAS (Dataset #1) than that from Null-based GAWS (Dataset #2) at five different *P*-value thresholds of 0.05, 0.01, 0.001, 0.0001, and 0.00001.

Replication and refinement of eQTL and GWAS results for PPP1CB

Furthermore, we concentrated our analysis on the hub gene of *PPP1CB*, which is a hub gene in our PPI network analysis. There were two eSNPs of rs9309664 ($P_{eQTL} = 1.42 \times 10^{-17}$ and $P_{GWAS} = 1.40 \times 10^{-11}$; see Supplementary Table S2) and rs7475 ($P_{eQTL} = 2.10 \times 10^{-6}$ and $P_{GWAS} = 1.70 \times 10^{-7}$; see Supplementary Table S4) in *PPP1CB* identified. To validate our results, we used the tool of HaploReg [40] to examine the association of two eSNPs (rs9309664 and rs7475) with the expression of *PPP1CB* in independent datasets. The result confirmed that rs9309664 is significantly associated with *PPP1CB* expression in blood cells from two independent datasets ($P = 6.19 \times 10^{-6}$ and $P = 3.4 \times 10^{-26}$; Supplementary Table S5), as well as rs7475 showed significant association with *PPP1CB* expression in a glioblastoma cell line from an independent dataset ($P = 0.00024$; Supplementary Table S5).

Discussion

Hitherto, more than 200 BMD-associated loci have been identified by GWAS studies [6–8,10]. However, the underlying biological mechanism of these risk SNPs and genes are still intractability.

Due to the influence of genetic LD among SNPs, these documented BMD-associated loci usually contain numerous highly LD SNPs with similar significant association signals. Thus, to confirm the exact causal SNPs of BMD-associated loci appeared to be a tough job. In view of a great number of detected susceptibility SNPs were mapped into the non-coding genomic regions, we reasonably inferred that these reported SNPs confer susceptibility to BMD via regulation of gene expression level. To address this issue, an effective and powerful inference method is warranted to be developed to combine the multi-omics information from genetic associations of existing GWAS summary data and independent eQTL data. In the current investigation, we employed a Bayesian integrative approach to combine the genetic associations from the large-scale GWAS of BMD ($N = 142,487$) and two independent eQTL datasets. In total, we identified 36 promising genes with eSNPs potentially implicated in BMD risk. To further prioritize the most convincing gene, we carried out several bioinformatics analyses, including MAGMA gene-based analysis, pathway-based enrichment analysis, *in silico* permutation analysis, PPI network-based analysis, and DGE analysis. Our results revealed that several genes, especially the *PPP1CB* gene, may represent authentic susceptibility genes for BMD.

The standard process of GWAS-based data analysis commonly uses millions of genetic SNPs to identify a number of common SNPs that showed significant associations with traits of interest such as osteoporosis. Nevertheless, the power of GWAS has been conspicuously limited by the enormous number of examined variants. GWAS may ignore the comprehensively integrated effects of modest variants or genes, which not reach a genome-wide significance but still
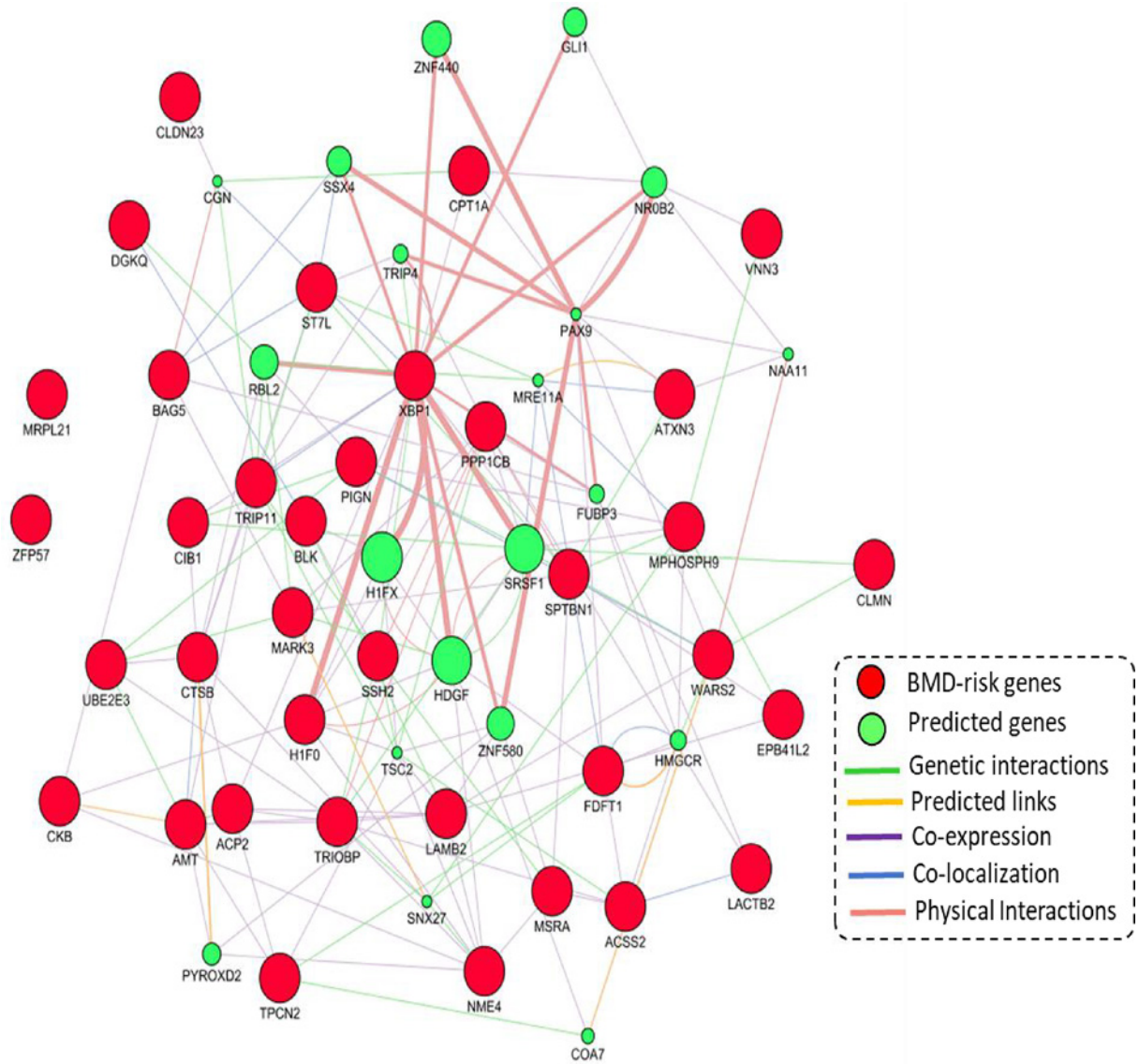
Figure 4. Protein–protein interactions between proteins encoded by 36 identified BMD-risk genes based on the application of the GeneMANIA database

The 36 BMD-risk genes were provided as query (red nodes) and a number of additional genes were predicted to be linked (green nodes). The predicted attributes included co-localization, genetic interactions, predicted links, co-expression, and physical interactions.

importantly implicated in the pathology of osteoporosis. Furthermore, only based on the findings of GWAS analysis is impossible to infer whether the detected disease-associated SNPs contain regulatory functions. Thus, *Sherlock* integrative analysis is a good and effective method for combining the information of GWAS with eQTL data and has been applied to identify many novel risk genes of many complex diseases [15,19–22], which cannot be detected by GWAS alone.

With the use of *Sherlock* integrative analysis and MAGMA gene-based analysis, there were 36 BMD-associated risk genes identified with eSNPs at a genome-wide significance; for example, *ACP2* (cis-rs11039035, $P_{\text{GWAS}} = 1.10 \times 10^{-11}$; cis-rs2290148, $P_{\text{GWAS}} = 2.40 \times 10^{-8}$), *EPB4L2* (cis-rs9375797, $P_{\text{GWAS}} = 1.70 \times 10^{-8}$; cis-rs4897473, $P_{\text{GWAS}} = 2.40 \times 10^{-15}$), *VNN3* (cis-rs9402490, $P_{\text{GWAS}} = 9.40 \times 10^{-67}$; cis-rs1856293, $P_{\text{GWAS}} = 2.30 \times 10^{-10}$), and *CKB* (cis-rs12894275, $P_{\text{GWAS}} = 7.20 \times 10^{-9}$; cis-rs2071407, $P_{\text{GWAS}} = 1.70 \times 10^{-12}$). Growing evidence indicates that disease-related genes encode functional proteins, which commonly generate a high interaction network [14,33,34,41].
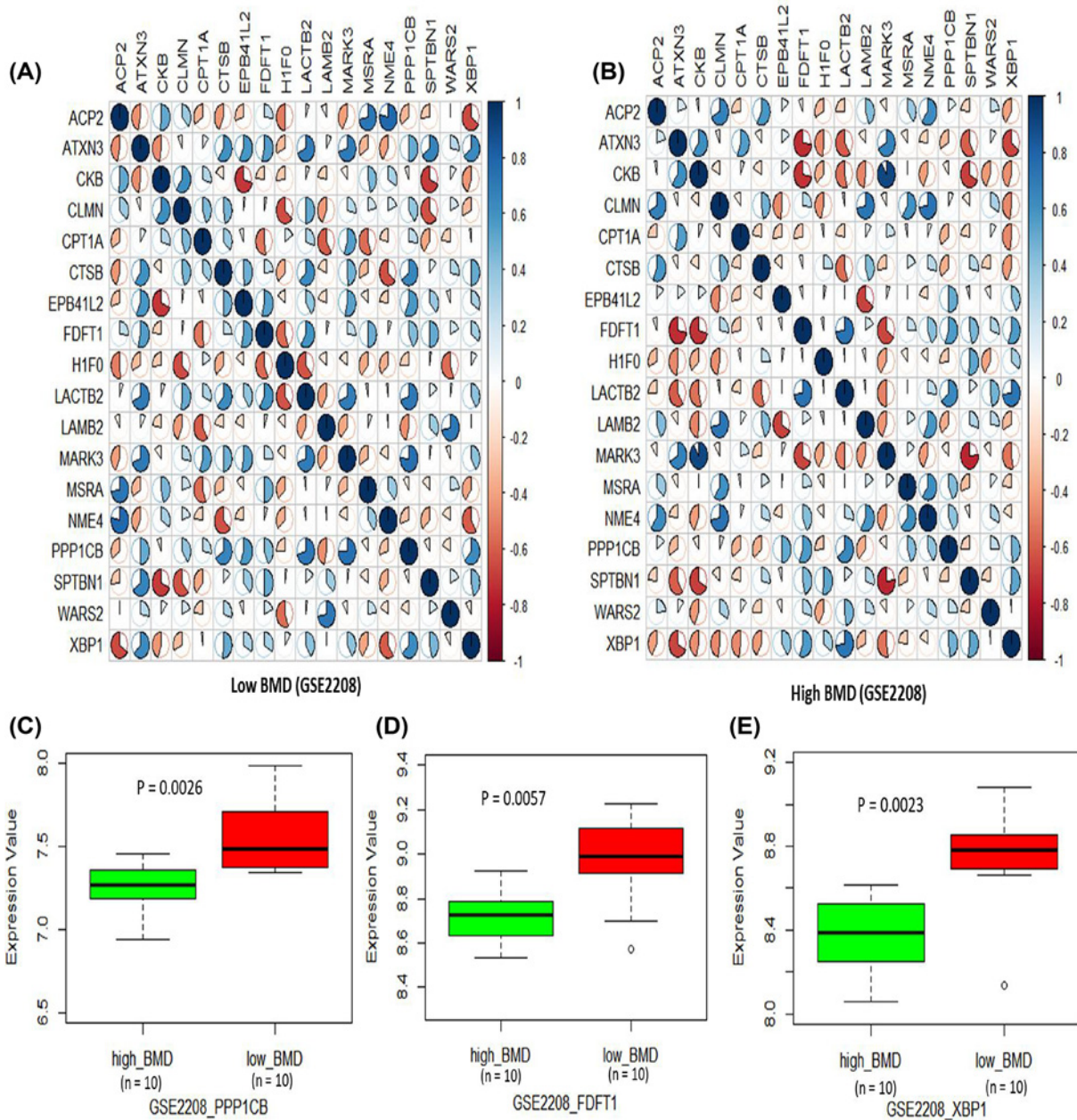
Figure 5. Differential expression profiles of 36 genes between high BMD and low BMD subjects in the GSE2208 dataset
(**A**) Co-expression patterns of 36 risk genes in low BMD subjects in the GSE2208 dataset. (**B**) Co-expression patterns of 36 risk genes in high BMD subjects in the GSE2208 dataset. (**C**) Boxplot of the different gene expression of *PPP1CB* between high BMD and low BMD subjects in the GSE2208 dataset. (**D**) Boxplot of the different gene expression of *FDFT1* between high BMD and low BMD subjects in the GSE2208 dataset. (**E**) Boxplot of the different gene expression of *XBP1* between high BMD and low BMD subjects in the GSE2208 dataset.

By conducting a PPI network-based analysis, we found that these 36 identified genes appeared to be highly connected to each other. There were several genes of *PPP1CB*, *XBP1*, *FDFT1*, and *SPTBN1* with the most number of interactions as the hub genes of the enriched network. For the hub gene of *PPP1CB*, there were two significant eSNPs of rs9309644 ($P_{eQTL} = 1.42 \times 10^{-17}$ and $P_{GWAS} = 1.40 \times 10^{-11}$) and rs7475 ($P_{eQTL} = 2.10 \times 10^{-6}$ and $P_{GWAS} = 1.70 \times 10^{-7}$) identified. Consistently, these two eSNPs were reported to be significantly associated with *PPP1CB* expression in the other three previous studies [42–44]. Furthermore, based on two independent expression datasets, we found *PPP1CB* showed significantly higher expression in the low BMD group than that in the high BMD group. These
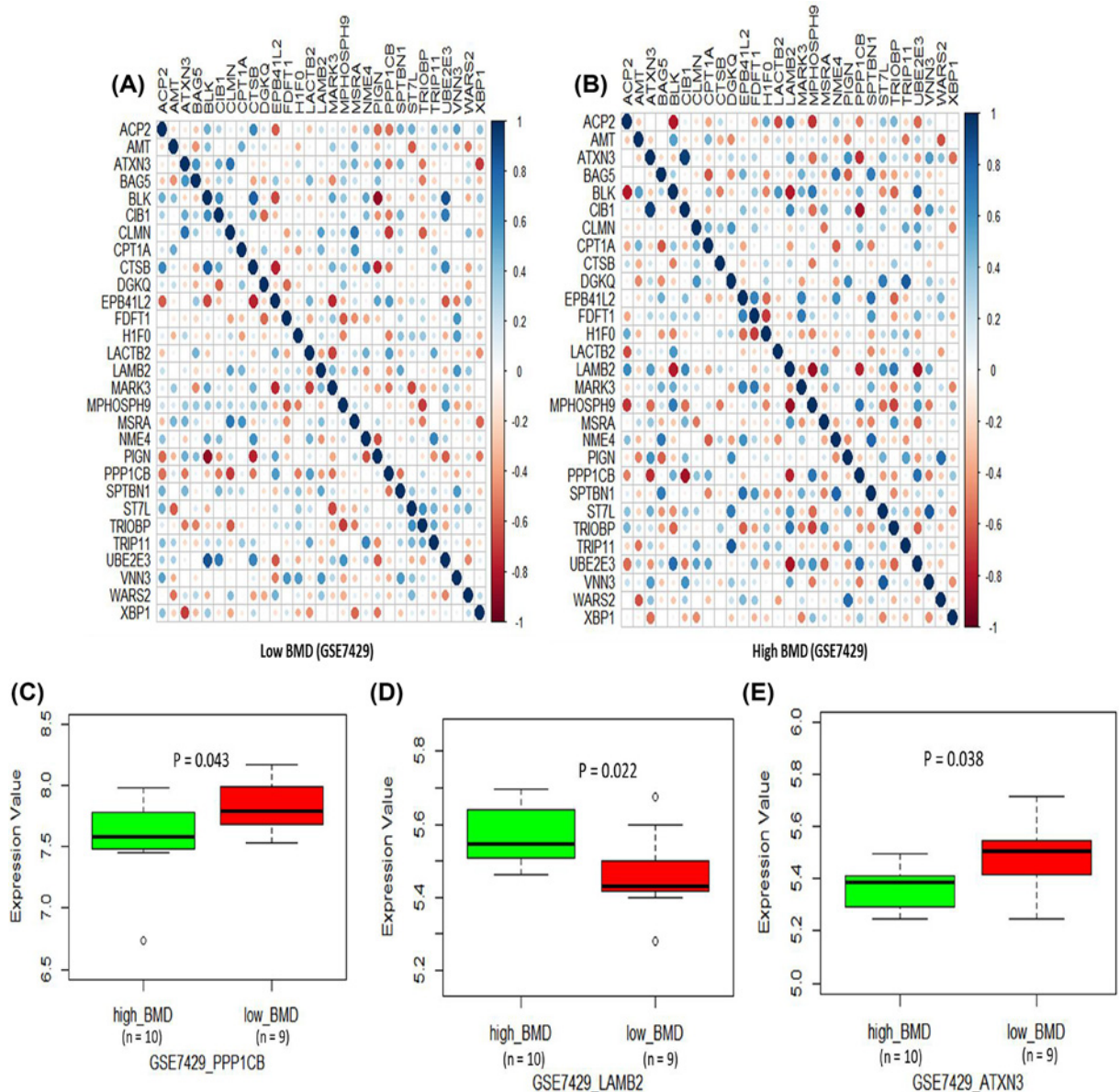
Figure 6. Differential expression profiles of 36 genes between high BMD and low BMD subjects in the GSE7429 dataset
(**A**) Co-expression patterns of 36 risk genes in low BMD subjects in the GSE7429 dataset. (**B**) Co-expression patterns of 36 risk genes in high BMD subjects in the GSE7429 dataset. (**C**) Boxplot of the different gene expression of *PPP1CB* between high BMD and low BMD subjects in the GSE7429 dataset. (**D**) Boxplot of the different gene expression of *FDFT1* between high BMD and low BMD subjects in the GSE7429 dataset. (**E**) Boxplot of the different gene expression of *XBP1* between high BMD and low BMD subjects in the GSE7429 dataset.

findings indicate that the *PPP1CB* gene plays an important role in low BMD subjects, which is a strong risk factor for osteoporosis and a key factor for its diagnosis and treatment [2].

The protein encoded by *PPP1CB* gene, which is located on the chromosome of 2p23.2, is one of the three catalytic subunits of protein phosphatase 1 (PP1). Phosphoprotein phosphatase 1 (PPP1), a major type 1 serine/threonine phosphatase, is ubiquitously expressed and regulates various cellular functions including glycogen metabolism, cell division, and muscle contractility [45–48]. In recent, a study [49] showed a role for *PPP1CB* that it is the myosin light chain phosphatase responsible for Ca^{2+}-transient rise and enhanced cell shortening in cardiomyocytes. Additionally, Cho et al. [45] have reported that PPP1CB regulates adipocyte differentiation partially by targeting the transcription factor C/EBPδ. Protein phosphatases are pivotal enzymes in diverse eukaryotic physiological processes including

cell differentiation and proliferation, cell–cell communication, and regulation of transmembrane and intracellular signaling pathways. Dysregulation of these proteins expression or activity has been documented to be linked with numerous human diseases [45,50,51], including cancer, obesity, and osteoporosis.

In sum, our present study highlights 36 convincing genes associated with BMD risk and further provides strong evidence to support that *PPP1CB* represents a genuine BMD-associated risk gene with eSNPs conferring susceptibility to osteoporosis. Through integrating the GWAS-based genetic associations with gene expression data, we demonstrated a plausible elucidation of the biological function of genetic variants on osteoporosis susceptibility. Current integrative genomics analysis offers an effective pipeline for incorporating SNPs across the whole genome to genes via cis- and/or trans-regulatory effects on transcriptional abundance, which is more powerful for interpreting the biological mechanism of functional SNPs than a pure GWAS method. Identification of risk genes from our results give a group of new molecular targets for subsequent functional experiments and therapeutic testing in future, which will improve the translation of pharmacogenomics testing into clinical practice. Furthermore, more related molecular studies are warranted to reveal the biological mechanism of *PPP1CB* gene implicated in osteoporosis.

Competing Interests
The authors declare that there are no competing interests associated with the manuscript.

Funding

Author Contribution
Y.Z., L.Y., and Y.S. conducted data collection and bioinformatics analysis and wrote the manuscript. J.W. conceived the study and helped write the manuscript.

Ethics Approval
This study does not contain any studies with human subjects or animals performed by any of the authors.

Consent for Publication
The authors declare that we gave consent for publication.

Abbreviations
BMD, bone mineral density; DGE, differential gene expression; eQTL, expression quantitative trait loci; eSNP, expression-associated single-nucleotide polymorphism; GEFOS, the Genetic Factors for Osteoporosis; GHS, the Gutenberg Heart study; GO, Gene Ontology; GWAS, genome-wide association study; KEGG, Kyoto Encyclopedia of Genes and Genomes; LD, linkage disequilibrium; MAGMA, Multi-marker Analysis of GenoMic Annotation; PPI, protein–protein interaction.

References
1 Burge, R. et al. (2007) Incidence and economic burden of osteoporosis-related fractures in the United States, 2005-2025. *J. Bone Miner. Res.* **22**, 465–475, https://doi.org/10.1359/jbmr.061113

2 Cauley, J.A. et al. (2007) Long-term risk of incident vertebral fractures. *JAMA* **298**, 2761–2767, https://doi.org/10.1001/jama.298.23.2761

3 Liu, C.T. et al. (2012) Heritability of prevalent vertebral fracture and volumetric bone mineral density and geometry at the lumbar spine in three generations of the Framingham study. *J. Bone Miner. Res.* **27**, 954–958, https://doi.org/10.1002/jbmr.1537

4 Zheng, H.F. et al. (2015) Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. *Nature* **526**, 112–117, https://doi.org/10.1038/nature14878

5 Peacock, M. et al. (2002) Genetics of osteoporosis. *Endocr. Rev.* **23**, 303–326, https://doi.org/10.1210/edrv.23.3.0464

6 Richards, J.B., Zheng, H.F. and Spector, T.D. (2012) Genetics of osteoporosis from genome-wide association studies: advances and challenges. *Nat. Rev. Genet.* **13**, 576–588, https://doi.org/10.1038/nrg3228

7 Estrada, K. et al. (2012) Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nat. Genet.* **44**, 491–501, https://doi.org/10.1038/ng.2249

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

9 Chesi, A. et al. (2019) Genome-scale Capture C promoter interactions implicate effector genes at GWAS loci for bone mineral density. *Nat. Commun.* **10**, 1260, https://doi.org/10.1038/s41467-019-09302-x

10 Kemp, J.P. et al. (2017) Identification of 153 new loci associated with heel bone mineral density and functional involvement of GPC6 in osteoporosis. *Nat. Genet.* **49**, 1468–1475, https://doi.org/10.1038/ng.3949

11 Nelson, M.R. et al. (2015) The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860, https://doi.org/10.1038/ng.3314

12 Choi, Y.J. et al. (2017) Transcriptional profiling of human femoral mesenchymal stem cells in osteoporosis and its association with adipogenesis. *Gene* **632**, 7–15, https://doi.org/10.1016/j.gene.2017.08.015

13 Wang, W. et al. (2017) Integrative analysis of GWAS, eQTLs and meQTLs data suggests that multiple gene sets are associated with bone mineral density. *Bone Joint Res.* **6**, 572–576, https://doi.org/10.1302/2046-3758.610.BJR-2017-0113.R1

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

15 He, X. et al. (2013) Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. *Am. J. Hum. Genet.* **92**, 667–680, https://doi.org/10.1016/j.ajhg.2013.03.022

16 Ayalew, M. et al. (2012) Convergent functional genomics of schizophrenia: from comprehensive understanding to genetic risk prediction. *Mol. Psychiatry* **17**, 887–905, https://doi.org/10.1038/mp.2012.37

17 Schadt, E.E. et al. (2005) An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* **37**, 710–717, https://doi.org/10.1038/ng1589

18 Zhu, Z. et al. (2016) Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487, https://doi.org/10.1038/ng.3538

19 Yang, C.P. et al. (2018) Comprehensive integrative analyses identify GLT8D1 and CSNK2B as schizophrenia risk genes. *Nat. Commun.* **9**, 838, https://doi.org/10.1038/s41467-018-03247-3

20 Lee, M.G. et al. (2019) Integrative Genome-Wide Association Studies of eQTL and GWAS Data for Gout Disease Susceptibility. *Sci. Rep.* **9**, 4981, https://doi.org/10.1038/s41598-019-41434-4

21 Huo, Y.X. et al. (2016) Identification of SLC25A37 as a major depressive disorder risk gene. *J. Psychiatr. Res.* **83**, 168–175, https://doi.org/10.1016/j.jpsychires.2016.09.011

22 Zhong, J. et al. (2019) Integration of GWAS and brain eQTL identifies FLOT1 as a risk gene for major depressive disorder. *Neuropsychopharmacology* **44**, 1542–1551

23 Zeller, T. et al. (2010) Genetics and beyond–the transcriptome of human monocytes and disease susceptibility. *PLoS ONE* **5**, e10693, https://doi.org/10.1371/journal.pone.0010693

24 Dixon, A.L. et al. (2007) A genome-wide association study of global gene expression. *Nat. Genet.* **39**, 1202–1207, https://doi.org/10.1038/ng2109

25 Hindorff, L.A. et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 9362–9367, https://doi.org/10.1073/pnas.0903103106

26 de Leeuw, C.A. et al. (2015) MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219, https://doi.org/10.1371/journal.pcbi.1004219

27 Landi, M.T. et al. (2009) A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am. J. Human Genet.* **85**, 679–691, https://doi.org/10.1016/j.ajhg.2009.09.012

28 Bindea, G. et al. (2009) ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091–1093, https://doi.org/10.1093/bioinformatics/btp101

29 Shannon, P. et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504, https://doi.org/10.1101/gr.1239303

30 Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30, https://doi.org/10.1093/nar/28.1.27

31 Ashburner, M. et al. (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25, https://doi.org/10.1038/75556

32 Akula, N. et al. (2016) An Integrative Genomic Study Implicates the Postsynaptic Density in the Pathogenesis of Bipolar Disorder. *Neuropsychopharmacology* **41**, 886–895, https://doi.org/10.1038/npp.2015.218

33 Ma, Y. and Li, M.D. (2017) Establishment of a Strong Link Between Smoking and Cancer Pathogenesis through DNA Methylation Analysis. *Sci. Rep.* **7**, 1811, https://doi.org/10.1038/s41598-017-01856-4

34 Gulsuner, S. et al. (2013) Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* **154**, 518–529, https://doi.org/10.1016/j.cell.2013.06.049

35 Warde-Farley, D. et al. (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* **38**, W214–W220, https://doi.org/10.1093/nar/gkq537

36 Xiao, P. et al. (2008) In vivo genome-wide expression study on human circulating B cells suggests a novel ESR1 and MAPK3 network for postmenopausal osteoporosis. *J. Bone Miner. Res.* **23**, 644–654, https://doi.org/10.1359/jbmr.080105

37 Ravasi, T. et al. (2010) An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* **140**, 744–752, https://doi.org/10.1016/j.cell.2010.01.044

38 Hein, M.Y. et al. (2015) A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* **163**, 712–723, https://doi.org/10.1016/j.cell.2015.09.053

39 Lin, A. et al. (2010) A genome-wide map of human genetic interactions inferred from radiation hybrid genotypes. *Genome Res.* **20**, 1122–1132, https://doi.org/10.1101/gr.104216.109

40 Ward, L.D. and Kellis, M. (2016) HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res.* **44**, D877–D881, https://doi.org/10.1093/nar/gkv1340

41 Chen, M., Cho, J. and Zhao, H. (2011) Incorporating biological pathways via a Markov random field model in genome-wide association studies. *PLoS Genet.* **7**, e1001353, https://doi.org/10.1371/journal.pgen.1001353

42 Heap, G.A. et al. (2009) Complex nature of SNP genotype effects on gene expression in primary human leucocytes. *BMC Med. Genomics* **2**, 1,
    https://doi.org/10.1186/1755-8794-2-1

43 Fehrmann, R.S. et al. (2011) Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate
    genes, with a major role for the HLA. *PLos Genet.* **7**, e1002197, https://doi.org/10.1371/journal.pgen.1002197

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

45 Cho, Y.L. et al. (2015) Phosphoprotein phosphatase 1CB (PPP1CB), a novel adipogenic activator, promotes 3T3-L1 adipogenesis. *Biochem. Biophys.
    Res. Commun.* **467**, 211–217, https://doi.org/10.1016/j.bbrc.2015.10.004

46 Korrodi-Gregorio, L., Esteves, S.L. and Fardilha, M. (2014) Protein phosphatase 1 catalytic isoforms: specificity toward interacting proteins. *Transl. Res.*
    164, 366–391, https://doi.org/10.1016/j.trsl.2014.07.001

47 Printen, J.A., Brady, M.J. and Saltiel, A.R. (1997) PTG, a protein phosphatase 1-binding protein with a role in glycogen metabolism. *Science* **275**,
    1475–1478, https://doi.org/10.1126/science.275.5305.1475

48 Alessi, D. et al. (1992) The control of protein phosphatase-1 by targetting subunits. The major myosin phosphatase in avian smooth muscle is a novel
    form of protein phosphatase-1. *Eur. J. Biochem.* **210**, 1023–1035, https://doi.org/10.1111/j.1432-1033.1992.tb17508.x

49 Aoyama, H. et al. (2011) Isoform-specific roles of protein phosphatase 1 catalytic subunits in sarcoplasmic reticulum-mediated Ca(2+) cycling.
    Cardiovasc. Res. **89**, 79–88, https://doi.org/10.1093/cvr/cvq252

50 Broome, D.T. and Datta, N.S. (2016) Mitogen-activated protein kinase phosphatase-1: function and regulation in bone and related tissues. *Connect.
    Tissue Res.* **57**, 175–189, https://doi.org/10.3109/03008207.2015.1125480

51 Meeusen, B. and Janssens, V. (2018) Tumor suppressive protein phosphatases in human cancer: Emerging targets for therapeutic intervention and
    tumor stratification. *Int. J. Biochem. Cell Biol.* **96**, 98–134, https://doi.org/10.1016/j.biocel.2017.10.002

Discovery data (N = 136 genes)

PHLDB1, TRIP6, PODNL1, MSRA, GGT7, LGALS3, PARG, CCBL2, SFMBT1, STEAP3, DGKQ, TRPC4AP, FIBP, DUSP12, MELK, H2AFY, LRP3, CLMN, KLF1, ARG2, CTPS, CXCL14, OR51T1, WDR17, NAT14, EDEM2, PYGM, ZNF605, HNRPUL1, PDE3A, MMP28, AXIN1, MON1A, CBWD5, SLC25A13, CTSB, CHRNB1, IFNA10, P8, DFNB31, ZNF460, XRCC1, RRP1B, DDX5, STAT6, D21S2056E, BAG5, LAMB2, CIB1, PCDP1, CBY1, KLF11, PLEKHG3, MFHAS1, CLDN23, EBAG9, SPTBN1, DAB2, FABP7, SSH2, NUDT17, LDOC1, MPHOSPH9, GATAD1, FALZ, PIGN, KLF2, TMEM53, PICK1, SENP7, NSF, DUSP3, CAMSAP1L1, OR2L3, UBE2E3, IGFBP7, GCAT, HOXB2, KCNK17, DKFZP761P0423, IL8, IDUA, TRIOBP, CCND2, B3GALNT1, MVK, TBRG1, CYB5D1, GUSB, GYPE, ADCY4, SLC47A1, SNORA33, NR1H3, RWDD2A, TPCN2, NUDT11, BLK, GLI1, HTR3C, NFATC1, LNX1, AP3D1, MARK3, TMUB2, QRICH1, SOCS5, FDFT1, FADS2, TFG, DUSP1, TMEM8, MS4A4A, NME4, PLB1, EIF2S2, ATXN3, POLR2K, ULK3, PH-4, DYSF, CENTG3, HCFC1, H1F0, NCKIPSD, XBP1, RIPK4, WARS2, LACTB2, RERE, TRIP11, MS4A6A, MRPS2, CADM1, ACSS2, UNC84B, FGFRL1
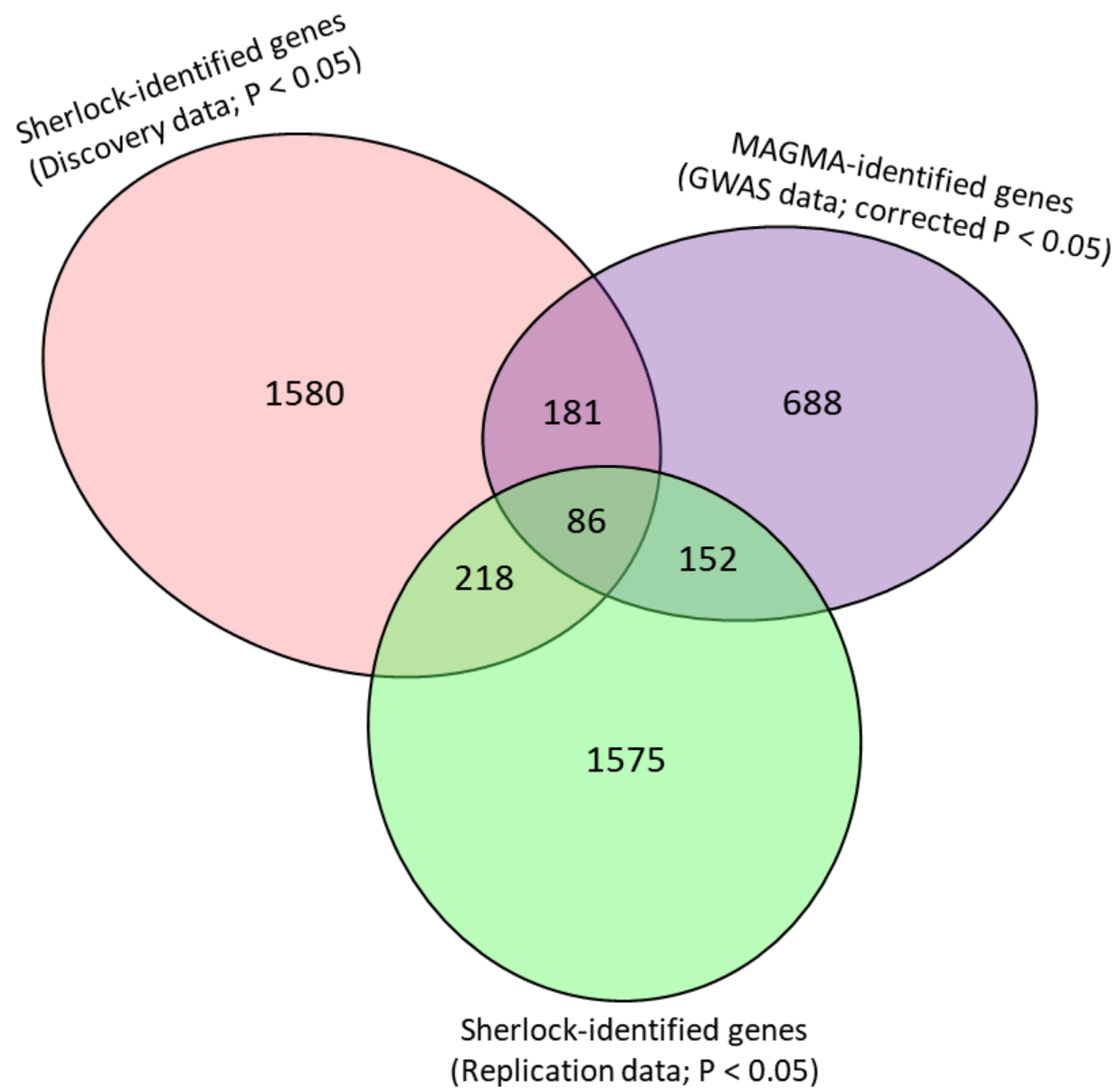
11 validated genes

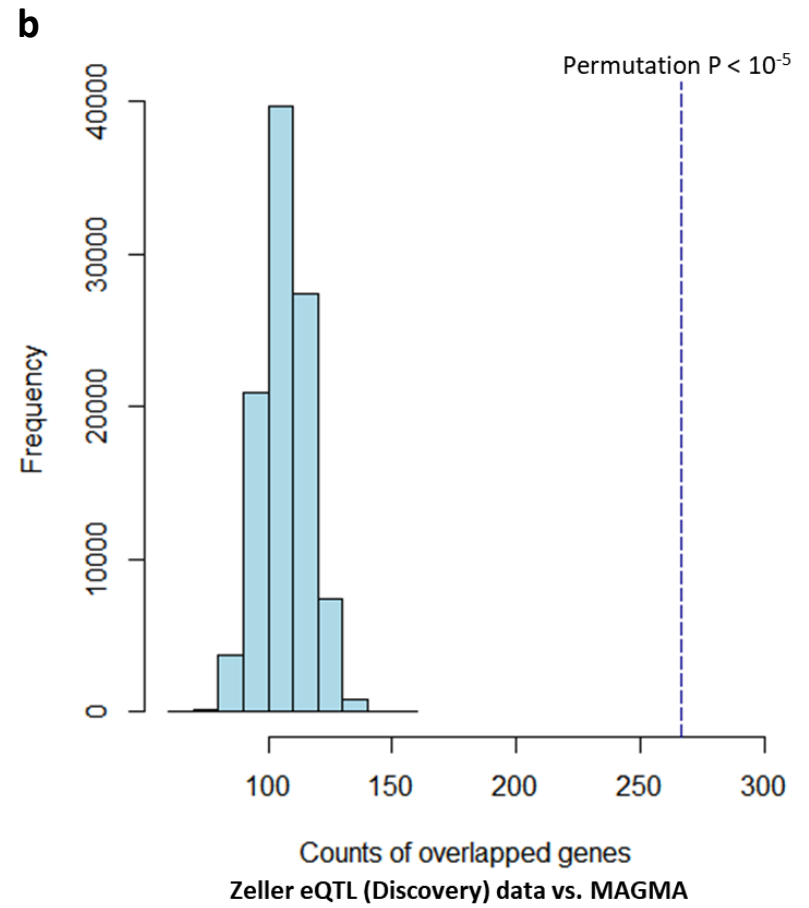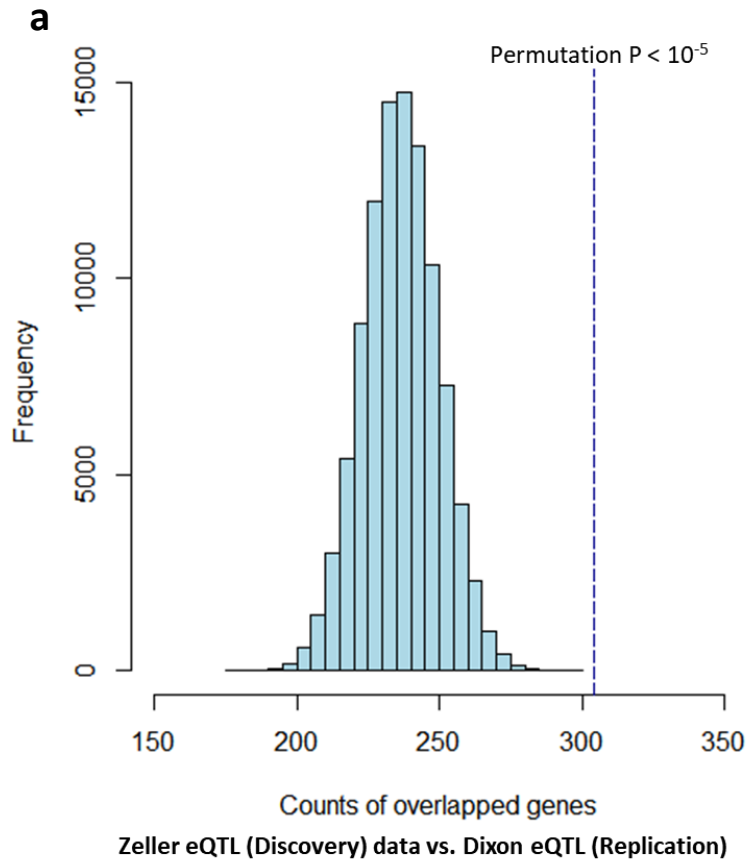*CPT1A , CKB , MRPL21, ZFP57, ACP2 , PPP1CB, ST7L, EPB41L2, PBX1, AMT, VNN3*

Validation data (N = 87 genes)

TINAG, ABCA3, GPR112, SH3RF2, AGBL2, BRP44L, TAX1BP1, IL1RAP, SEC24D, EYA1, CDK6, THADA, CSNK1G3, CDS1, ZNF560, OR8B2, MYLIP, ZXDA, UTY , SAMD10, IMP4, NEIL2, KLHL8, LRRTM1, RGS4, SPINK6, CLDN20, ZIM2, LETM1, HLA-DRB1, FMNL2, FYB, TRPS1, NUP214, AKR1C1, MBL2, SPIN3, NR4A2, BOLA2, TTC12, LGR4, FADS1, F5, C21orf81, QKI ITCH, XRCC3, WNT3, AGTR2, PCGF3, KIRREL, C18orf19, BTN3A3, MYH10, SOCS1, KLF8, RPS23, CLTC, INSIG1, RFT1, SDHD, RRAGB, MBD6, DDB2, TNFSF11, PARP6, CTNNB1, ILDR1, KCNJ13, LSS, SHOC2, AHNAK, SSBP4, CYP2C9, AXIN2, WBSCR16, DOK1, KCNA2, GALNT12, CLCN5, MAPK8IP1, TRAT1, FAM48A, NRF1, HDAC5, SCN3B, SRR

Supplemental Figure S1

Supplemental Figure S1. Venn plot of BMD-associated genes identified by Sherlock Bayesian analysis based on two independent datasets. There were 11 validated genes conferring risk to BMD. The P values of all genes were corrected by the method of Bonferroni correction.

Supplemental Figure S2

Supplemental Figure S2. Venn plot of BMD-relevant genes based on three datasets: Sherlock-identified genes from the discovery dataset (P < 0.05); Sherlock-identified genes from the validation dataset (P < 0.05); MAGMA-identified genes from the GWAS dataset (corrected P < 0.05).

a

Permutation P < 10^{-5}

Frequency

Counts of overlapped genes
Zeller eQTL (Discovery) data vs. Dixon eQTL (Replication)

b

Permutation P < 10^{-5}

Frequency

Counts of overlapped genes
Zeller eQTL (Discovery) data vs. MAGMA

Supplemental Figure S3

Supplemental Figure S3. Computer-based permutation analysis. a) Computer-based permutation analysis (100,000 times) of the counts of risk genes from Zeller eQTL dataset (P < 0.05; discovery stage) overlapped with that from Dixon eQTL dataset (P < 0.05; Replication stage). b) Computer-based permutation analysis (100,000 times) of the counts of risk genes from Zeller eQTL dataset (P < 0.05; discovery stage) overlapped with that from MAGMA-based dataset (corrected P < 0.05; technical validation stage).

Supplemental Figure S4
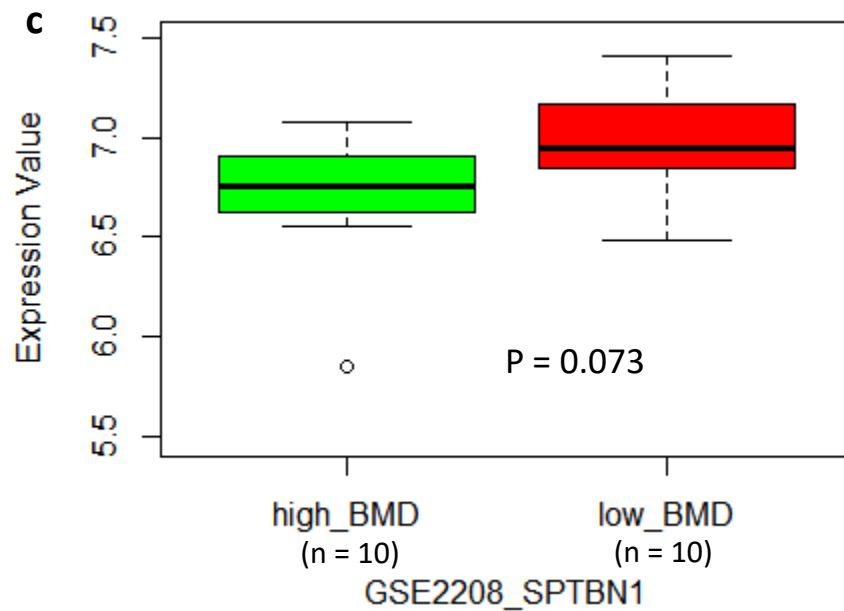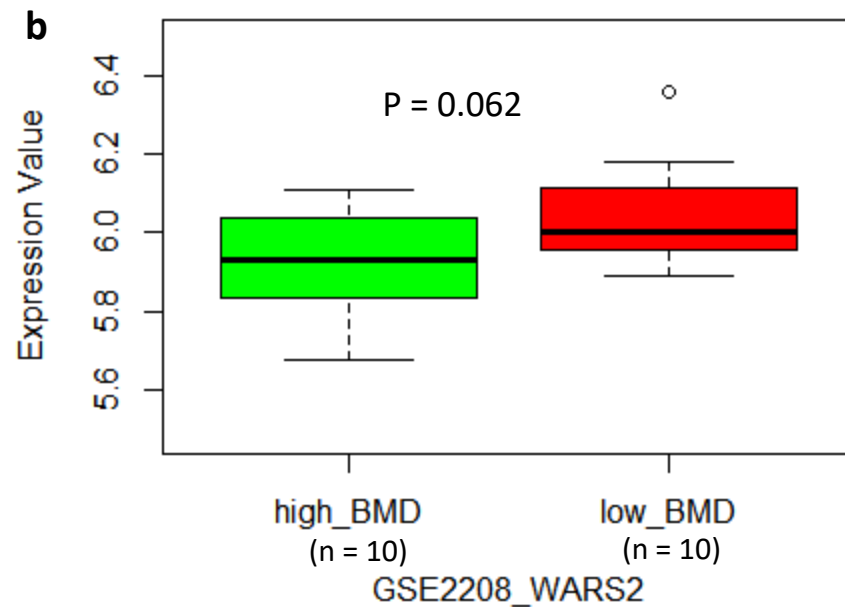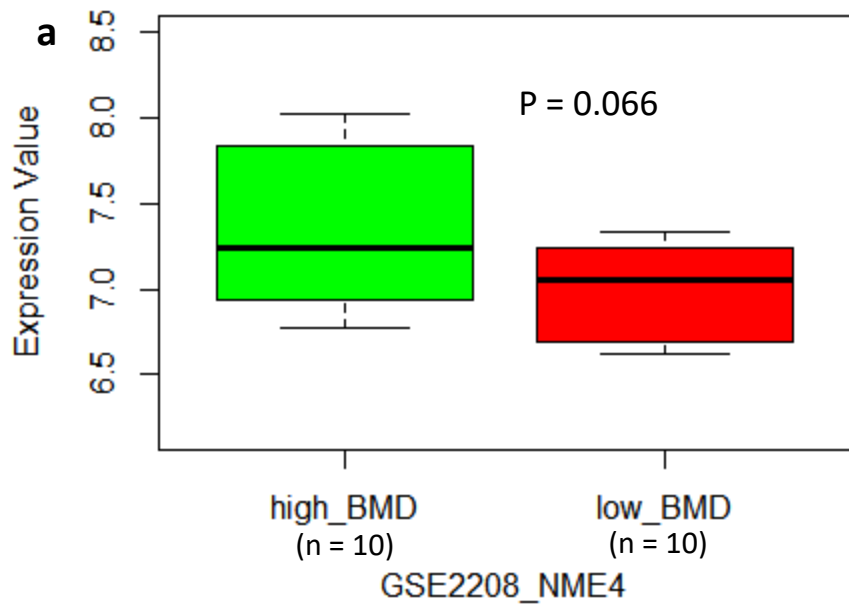
Supplemental Figure S4. Boxplots show the expression profile of identified genes between high BMD and low BMD in GSE2208 dataset. a) Boxplot of the different gene expression of *NME4*between high BMD and low BMD subjects in GSE2208 dataset. b) Boxplot of the different gene expression of *WARS2*between high BMD and low BMD subjects in GSE2208 dataset. c) Boxplot of the different gene expression of *SPTBN1*between high BMD and low BMD subjects in GSE2208 dataset.

Supplemental Figure S5
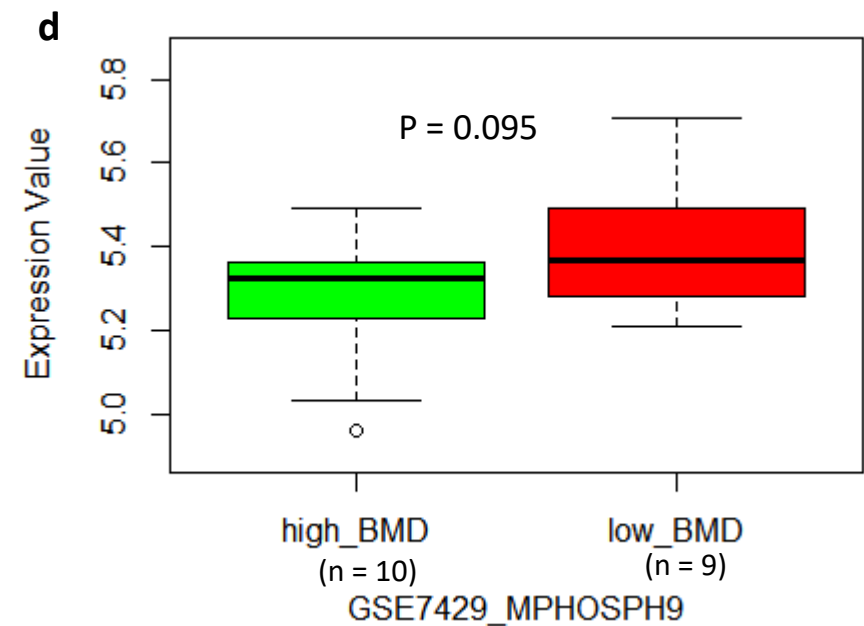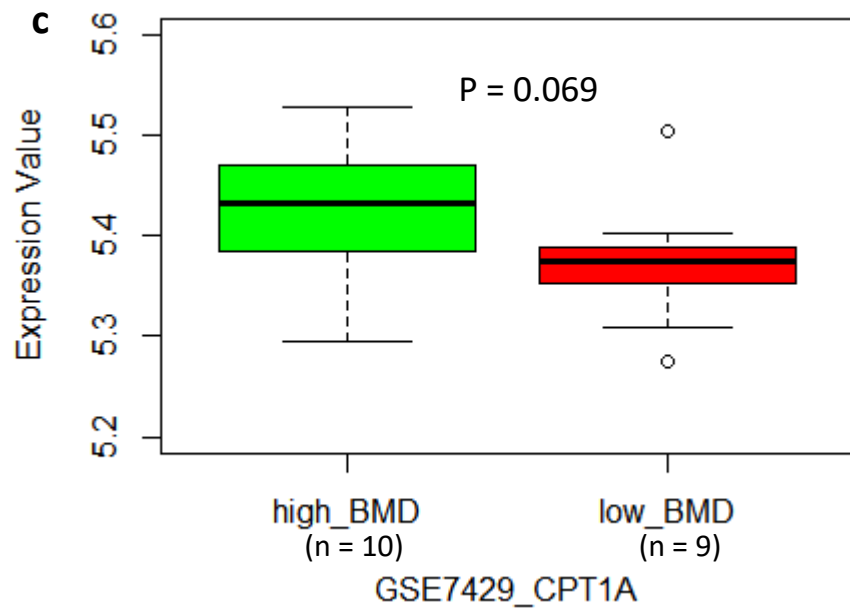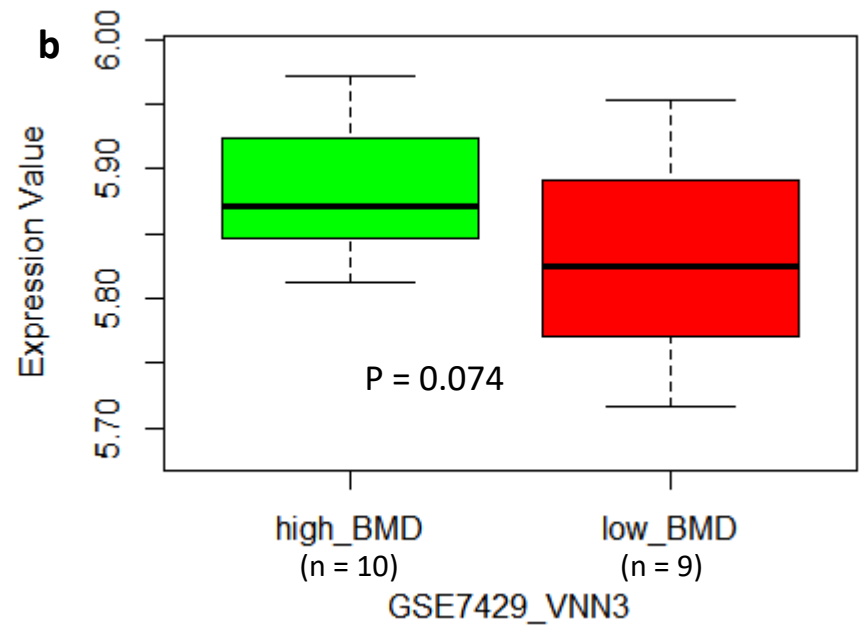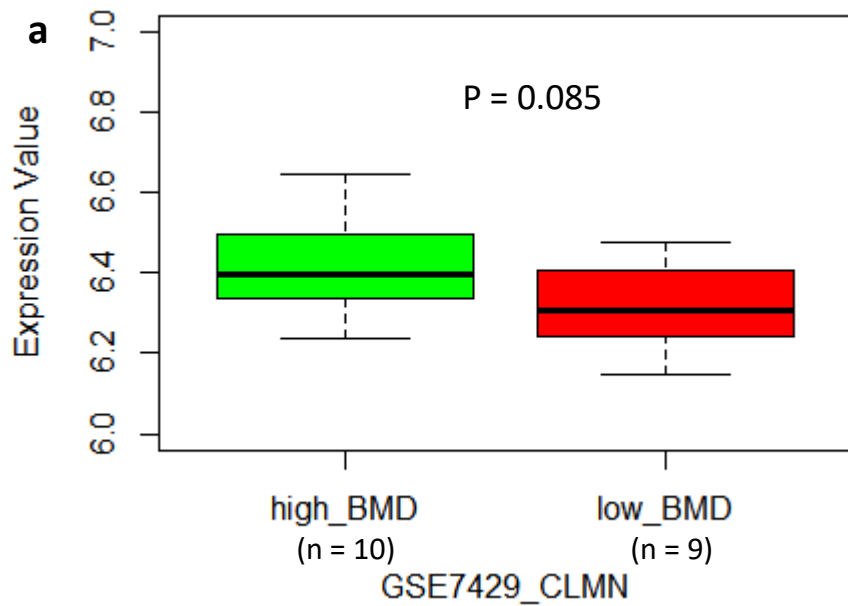
Supplemental Figure S5. Boxplots show the expression profile of identified genes between high BMD and low BMD in GSE2208 dataset. a) Boxplot of the different gene expression of *CLMN* between high BMD and low BMD subjects in GSE7429 dataset. b) Boxplot of the different gene expression of *VNN3* between high BMD and low BMD subjects in GSE7429 dataset. c) Boxplot of the different gene expression of *CPT1A* between high BMD and low BMD subjects in GSE7429 dataset. d) Boxplot of the different gene expression of *MPHOSPH9* between high BMD and low BMD subjects in GSE7429 dataset.

Supplemental Table S1. *Sherlock* integrative analysis identifies 147 genes as BMD risk genes in the discovery stage

Gene	LBF	Gene P value
NR1H3	12.2999	7.87E-07
MVK	11.5158	7.87E-07
TBRG1	11.5094	7.87E-07
PDE3A	11.315	7.87E-07
FALZ	11.1695	7.87E-07
EBAG9	10.8478	7.87E-07
MSRA	10.5948	7.87E-07
ULK3	9.99511	7.87E-07
HOXB2	9.76989	7.87E-07
GATAD1	9.5143	7.87E-07
BLK	9.14053	7.87E-07
MS4A4A	9.04221	7.87E-07
LDOC1	9.01994	7.87E-07
P8	8.71756	7.87E-07
WDR17	8.64742	7.87E-07
NAT14	8.58634	7.87E-07
MFHAS1	8.57354	7.87E-07
TMEM53	8.52609	7.87E-07
ADCY4	8.49684	7.87E-07
SENP7	8.47952	7.87E-07
RIPK4	8.19958	7.87E-07
PLEKHG3	7.91651	7.87E-07
PICK1	7.78789	7.87E-07
PARG	7.53614	7.87E-07
ACP2	7.49565	7.87E-07
GCAT	7.49433	7.87E-07
EPB41L2	7.47978	7.87E-07
CXCL14	7.30761	7.87E-07
PODNL1	7.22259	7.87E-07
GUSB	7.11892	7.87E-07
PH-4	7.10204	7.87E-07
CBY1	7.05232	7.87E-07
CBWD5	6.9474	7.87E-07
UNC84B	6.90261	7.87E-07
TMUB2	6.88889	7.87E-07
RERE	6.88072	7.87E-07
DUSP1	6.84233	7.87E-07
ACSS2	6.83981	7.87E-07
ZFP57	6.8172	7.87E-07
HNRPUL1	6.78973	7.87E-07
EDEM2	6.78454	7.87E-07
MARK3	6.75701	7.87E-07
PBX1	6.68498	7.87E-07
FDFT1	6.65806	7.87E-07
CADM1	6.64867	7.87E-07
FABP7	6.64681	7.87E-07
AP3D1	6.6278	7.87E-07
PPP1CB	6.61774	7.87E-07
SPTBN1	6.61328	7.87E-07
AMT	6.59051	7.87E-07

DUSP12	6.5719	7.87E-07
LRP3	6.57174	7.87E-07
PLB1	6.55676	7.87E-07
ARG2	6.53133	7.87E-07
MPHOSPH9	6.53093	7.87E-07
STEAP3	6.5064	7.87E-07
CPT1A	6.49275	7.87E-07
XBP1	6.43078	7.87E-07
CCBL2	6.42554	7.87E-07
TRIP6	6.42465	7.87E-07
CLDN23	6.38306	7.87E-07
AXIN1	6.38213	7.87E-07
H1F0	6.35558	7.87E-07
DAB2	6.316	7.87E-07
SLC47A1	6.29566	7.87E-07
FGFRL1	6.29175	7.87E-07
SOCS5	6.2903	7.87E-07
OR2L3	6.28478	7.87E-07
SSH2	6.27166	7.87E-07
LACTB2	6.26756	7.87E-07
KLF11	6.25736	7.87E-07
MS4A6A	6.25532	7.87E-07
VNN3	6.24743	7.87E-07
B3GALNT1	6.20561	7.87E-07
NFATC1	6.20172	7.87464E-07
MRPS2	6.1761	7.87E-07
RWDD2A	6.15794	7.87E-07
FIBP	6.13312	7.87E-07
CYB5D1	6.12814	7.87E-07
DGKQ	6.12386	7.87E-07
KLF1	6.11453	7.87E-07
LAMB2	6.07699	7.87E-07
IL8	5.99556	7.87E-07
OR51T1	5.99219	7.87E-07
TRIOBP	5.95978	7.87E-07
NSF	5.95953	7.87E-07
TPCN2	5.95545	7.87E-07
TMEM8	5.95372	7.87E-07
QRICH1	5.94704	7.87E-07
BAG5	5.9431	7.87E-07
STAT6	5.8953	7.87E-07
CAMSAP1L1	5.88891	7.87E-07
MON1A	5.87306	1.57E-06
MRPL21	5.83995	1.57E-06
CTSB	5.78827	1.57E-06
TRPC4AP	5.78065	1.57E-06
DDX5	5.76447	1.57E-06
GGT7	5.71233	1.57E-06
SNORA33	5.71046	1.57E-06
KLF2	5.71025	1.57E-06
NME4	5.70779	1.57E-06
TRIP11	5.70165	1.57E-06
GLI1	5.69007	1.57E-06
IFNA10	5.68755	1.57E-06
IDUA	5.62864	1.57E-06

EIF2S2	5.62145	1.57E-06
MMP28	5.62022	1.57E-06
ZNF605	5.61031	1.57E-06
SLC25A13	5.58744	1.57E-06
PHLDB1	5.58561	1.57E-06
DUSP3	5.5704	1.57E-06
DFNB31	5.56622	1.57E-06
PIGN	5.55191	1.57E-06
CTPS	5.48714	1.57E-06
LNX1	5.48174	1.57E-06
ATXN3	5.45739	1.57E-06
NCKIPSD	5.44234	1.57E-06
GYPE	5.41442	1.57E-06
ST7L	5.40094	1.57E-06
NUDT17	5.3943	1.57E-06
WARS2	5.35449	1.57E-06
POLR2K	5.33098	1.57E-06
DYSF	5.29734	1.57E-06
PCDP1	5.29729	1.57E-06
CHRNB1	5.27137	1.57E-06
NUDT11	5.26999	1.57E-06
ZNF460	5.26737	1.57E-06
FADS2	5.26155	1.57E-06
PYGM	5.24992	1.57E-06
HTR3C	5.23737	1.57E-06
SFMBT1	5.22316	1.57E-06
D21S2056E	5.22062	1.57E-06
LGALS3	5.21765	1.57E-06
MELK	5.19724	1.57E-06
TFG	5.18289	1.57E-06
RRP1B	5.1736	1.57E-06
CLMN	5.14477	1.57E-06
CKB	5.14417	1.57E-06
CENTG3	5.11152	1.57E-06
KCNK17	5.08362	1.57E-06
IGFBP7	5.08341	1.57E-06
XRCC1	5.0827	1.57E-06
UBE2E3	5.0613	1.57E-06
CIB1	5.05829	1.57E-06
H2AFY	5.04571	1.57E-06
CCND2	5.03881	1.57E-06
HCFC1	5.02937	1.57E-06

Supplemental Table S2. The results of 11 BMD-associated genes in the discovery stage

Gene name	LBF[a]	P value[b]	Cis or trans SNP[c]	P value from eQTL[d]	P value from GWAS[e]	Correction-based P value[f]
ACP2	7.50	7.87×10^{-7}	rs11039035 (cis)	3.71×10^{-12}	1.10×10^{-11}	< 0.05
EPB41L2	7.48	7.87×10^{-7}	rs9375797 (cis)	3.47×10^{-8}	1.70×10^{-8}	< 0.05
ZFP57	6.82	7.87×10^{-7}	rs3129063 (cis)	1.46×10^{-29}	4.50×10^{-6}	< 0.05
PBX1	6.68	7.87×10^{-7}	rs4581300 (trans)	1.57×10^{-6}	2.70×10^{-7}	< 0.05
PPP1CB	6.62	7.87×10^{-7}	rs9309664 (cis)	1.42×10^{-17}	1.40×10^{-11}	< 0.05
AMT	6.59	7.87×10^{-7}	rs3749237 (cis)	2.48×10^{-10}	4.60×10^{-11}	< 0.05
CPT1A	6.49	7.87×10^{-7}	rs3736228 (cis)	3.93×10^{-17}	6.50×10^{-19}	< 0.05
VNN3	6.25	7.87×10^{-7}	rs9402490 (cis)	5.76×10^{-12}	9.40×10^{-67}	< 0.05
MRPL21	5.84	1.57×10^{-6}	rs646586 (cis)	1.37×10^{-86}	8.40×10^{-9}	< 0.05
ST7L	5.40	1.57×10^{-6}	rs3790608 (cis)	8.16×10^{-22}	1.20×10^{-18}	< 0.05
CKB	5.14	1.57×10^{-6}	rs12894275 (cis)	2.14×10^{-8}	7.20×10^{-9}	< 0.05

Note: [a] LBF:The abbreviation of the logarithm of the Bayes factor. LBF considers as a vital indicator to evaluate the association of a risk gene and complicate disease such as BMD in current study via pooling the eQTL data and GWAS summary statistics. The large value of LBF highly reflects the high links between the gene and BMD; e.g., a number of 7.5 LBF value (*exp*(7.50) = 1,808 times) represents the gene of interest preferably shows prominent association with BMD than non-significant association underlying the null hypothesis.

[b]P–value from calculation based on the statistical inferences method of *Sherlock* integrative analysis. Considering *Sherlock* employs finite times of permutation test to calculate the P value for each gene, some of these top-ranked genes (e.g., *ACP2* and *EPB41L2*) have distinct LBF values but

obtain the same rankings (namely, their P values are same).

[c]We only showed the SNP has the highest score of LBF in each gene.

[d]P-value from expression quantitative trait analysis of Zeller et al.

[e]P-value from GWAS on BMD of Kemp et al.

[f]The method of Bonferroni correction was employed to correct the P-values from *Sherlock* tool.

Supplemental Table S3. *Sherlock* integrative analysis identifies 98 genes as BMD risk genes in the validation stage

Gene	LBF	Gene P value
AMT	13.5691	7.59E-07
MAPK8IP1	11.7318	7.59E-07
SPIN3	11.5091	7.59E-07
MYLIP	10.9087	7.59E-07
MYH10	9.95042	7.59E-07
TTC12	9.80843	7.59E-07
SSBP4	9.76063	7.59E-07
BOLA2	9.50038	7.59E-07
NR4A2	9.29855	7.59E-07
ZFP57	8.9893	7.59E-07
NUP214	8.60844	7.59E-07
AGTR2	8.41611	7.59E-07
PBX1	8.26044	7.59E-07
ILDR1	7.97657	7.59E-07
IMP4	7.96895	7.59E-07
OR8B2	7.94505	7.59E-07
WNT3	7.85747	7.59E-07
EYA1	7.77769	7.59E-07
KCNJ13	7.51417	7.59E-07
LSS	7.4792	7.59E-07
CDK6	7.42466	7.59E-07
CDS1	7.4131	7.59E-07
RGS4	7.30141	7.59E-07
F5	7.20892	7.59E-07
DOK1	7.19648	7.59E-07
FADS1	7.16901	7.59E-07
SDHD	7.11542	7.59E-07
FMNL2	7.04845	7.59E-07
ITCH	6.84407	7.59E-07
MRPL21	6.82275	7.59E-07
HDAC5	6.81755	7.59E-07
C18orf19	6.79965	7.59E-07
CSNK1G3	6.64201	7.59E-07
NRF1	6.59861	7.59E-07
CLDN20	6.57949	7.59E-07
SH3RF2	6.43827	7.59E-07
CLCN5	6.42869	7.59E-07
WBSCR16	6.35304	7.59E-07
KLF8	6.31301	7.59E-07
MBL2	6.30824	7.59E-07
KIRREL	6.28072	7.59E-07
THADA	6.27297	7.59E-07
CKB	6.26649	7.59E-07
SCN3B	6.16817	7.59E-07
UTY	6.15071	7.59E-07
TNFSF11	6.14619	7.59E-07
AKR1C1	6.14427	7.59E-07

KLHL8	6.12499	7.59E-07
ZIM2	6.10742	7.59E-07
SHOC2	6.10708	7.59E-07
XRCC3	6.07382	7.59E-07
LRRTM1	6.04204	7.59E-07
AGBL2	5.9564	7.59E-07
LETM1	5.94412	7.59E-07
SRR	5.92776	7.59E-07
VNN3	5.92444	7.59E-07
EPB41L2	5.9032	7.59E-07
PPP1CB	5.8688	7.59E-07
SEC24D	5.86731	7.59E-07
TAX1BP1	5.84235	7.59E-07
SOCS1	5.8379	7.59E-07
RRAGB	5.7971	7.59E-07
RPS23	5.79692	7.59E-07
CTNNB1	5.77722	7.59E-07
SAMD10	5.6892	7.59E-07
QKI	5.66379	7.59E-07
PARP6	5.65955	7.59E-07
IL1RAP	5.58381	7.59E-07
ACP2	5.57894	7.59E-07
GALNT12	5.50698	7.59E-07
MBD6	5.49491	7.59E-07
KCNA2	5.46585	7.59E-07
ABCA3	5.45385	7.59E-07
PCGF3	5.37444	7.59E-07
ZNF560	5.37433	7.59E-07
FAM48A	5.37108	7.59E-07
CPT1A	5.35921	7.59E-07
ST7L	5.35852	7.59E-07
SPINK6	5.35312	7.59E-07
ZXDA	5.34714	7.59E-07
AXIN2	5.34304	7.59E-07
AHNAK	5.32541	7.59E-07
TRPS1	5.29976	7.59E-07
NEIL2	5.28915	7.59E-07
FYB	5.24885	7.59E-07
RFT1	5.23427	7.59E-07
BRP44L	5.20072	7.59E-07
CYP2C9	5.19215	7.59E-07
INSIG1	5.19194	7.59E-07
DDB2	5.15843	7.59E-07
GPR112	5.14337	7.59E-07
C21orf81	5.13445	7.59E-07
CLTC	5.07912	7.59E-07
BTN3A3	5.07595	7.59E-07
LGR4	5.05598	7.59E-07
HLA-DRB1	5.02535	1.52E-06
TINAG	5.01643	1.52E-06
TRAT1	5.0079	1.52E-06

Supplemental Table S4. The results of 11 BMD-associated genes in the validation stage

Gene name	LBF[a]	P value[b]	Cis or trans SNP[c]	P value from eQTL[d]	P value from GWAS[e]	Correction-based P value[f]
ACP2	5.58	7.59×10^{-7}	rs2290148 (cis)	5.00×10^{-5}	2.40×10^{-8}	< 0.05
EPB41L2	5.90	7.59×10^{-7}	rs4897473 (cis)	2.70×10^{-6}	2.40×10^{-15}	< 0.05
ZFP57	8.99	7.59×10^{-7}	rs2747430 (cis)	5.10×10^{-29}	5.30×10^{-7}	< 0.05
PBX1	8.26	7.59×10^{-7}	rs1487628 (trans)	2.3×10^{-6}	1.20×10^{-7}	< 0.05
PPP1CB	5.87	7.59×10^{-7}	rs7475 (cis)	2.10×10^{-6}	1.70×10^{-7}	< 0.05
AMT	13.57	7.59×10^{-7}	rs1464569 (cis)	6.70×10^{-6}	8.20×10^{-7}	< 0.05
CPT1A	5.36	7.59×10^{-7}	rs2156464 (cis)	6.60×10^{-8}	2.50×10^{-10}	< 0.05
VNN3	5.92	7.59×10^{-7}	rs1856293 (cis)	4.00×10^{-5}	2.30×10^{-10}	< 0.05
MRPL21	6.82	7.59×10^{-7}	rs629426 (cis)	2.80×10^{-9}	4.40×10^{-8}	< 0.05
ST7L	5.36	7.59×10^{-7}	rs1106287 (cis)	3.80×10^{-7}	4.70×10^{-15}	< 0.05
CKB	6.27	7.59×10^{-7}	rs2071407 (cis)	9.00×10^{-7}	1.70×10^{-12}	< 0.05

Note: [a] LBF:The abbreviation of the logarithm of the Bayes factor. LBF considers as a vital indicator to evaluate the association of a risk gene and complicate disease such as BMD in current study via pooling the eQTL data and GWAS summary statistics. The large value of LBF highly reflects the high links between the gene and BMD; e.g., a number of 5.58 LBF value (*exp*(5.58) = 265 times) represents the gene of interest preferably shows prominent association with BMD than non-significant association underlying the null hypothesis.

[b] P–value from calculation based on the statistical inferences method of *Sherlock* integrative analysis. Considering *Sherlock* employs finite times of permutation test to calculate the P value for each gene, some of these top-ranked genes (e.g., *ACP2* and *EPB41L2*) have distinct LBF values but obtain the same rankings (namely, their P values are same).

[c]We only showed the SNP has the highest score of LBF in each gene.

[d]P-value from expression quantitative trait analysis of Dixon et al.

[e]P-value from GWAS on BMD of Kemp et al.

[f]The method of Bonferroni correction was employed to correct the P-values from *Sherlock* tool.

Supplemental Table S5. Additional evidence of eQTL findings for *PPP1CB* gene

rs# of eSNP	Traits	P values	References (PMID)
rs9309664	Gene expression of SPDY1 in blood cells in Celiac disease	6.19×10^{-6}	[1](19128478)
rs9309664	Gene expression of PPP1CB in blood	3.4×10^{-26}	[2](21829388)
rs7475	Allele-specific Expression Patterns in human glioblastoma cell line U87MG	0.00024	[3](22467206)

References:

1.   Heap, G.A., et al., *Complex nature of SNP genotype effects on gene expression in primary human leucocytes.* BMC Med Genomics, 2009. **2**: p. 1.

2.   Fehrmann, R.S., et al., *Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA.* PLoS Genet, 2011. **7**(8): p. e1002197.

3.   Li, G., et al., *Identification of allele-specific alternative mRNA processing via transcriptome sequencing.* Nucleic Acids Res, 2012. **40**(13): p. e104.