Integrating genome-wide association and transcriptome predicted model identify novel target genes with osteoporosis

Peng Yin[1,*,†], Muchun Zhu[1,†], Fan Hu[1], Jiaxin Jiang[1], Li Yin[1], Shuqiang Wang[1]and Yingxiang Li[2,3]

[1] Joint Engineering Research Center for Health Big Data Intelligent Analysis Technology, Shenzhen Institute of Advanced Technology, Chinese Academy of Science, 1068 Xueyuan Avenue, Shenzhen University Town, Shenzhen, China. [2] WeGene, Shenzhen, China. [3] Department of Anthropology and Ethnology, Xiamen University, Xiamen, China.

[*] Correspondence: Peng.Yin@siat.ac.cn

[†] Muchun Zhu, Peng Yin contributed equally to this work.

Abstract

Osteoporosis (OP) is a highly polygenetic disease which is usually characterized by low bone mineral density. Genome-wide association studies (GWAS) have identified hundreds of genetic loci associated with bone mineral density. However, the biological mechanisms of these loci remain elusive. To identify potential causal genes of the associated loci, we detected trait-gene expression associations by transcriptome-wide association study (TWAS) method. It directly imputes gene expression effects from GWAS data, using a statistical prediction model trained on GTEx reference transcriptome data, with blood and skeletal tissues data. Then we performed a colocalization analysis to evaluate the posterior probability of biological patterns: association characterized by a single shared causal variant or two distinct causal variants. The ultimate analysis identified 276 candidate genes, including 3 novel loci, 204 novel candidate genes and 69 replicated from GWAS. The 3 novel loci located at chr6: 72417543, chr15: 69601206, chr21: 30530692, mapping to gene *RIMS1*, *SPESP1*, *MAP3K7CL*. The results of colocalization analysis indicated that 142 of them showing strong evidence of a single shared causal variant and 134 of them showing evidence of joint causal variants. Their biological function was directly or indirectly associated with the occurrence of OP validated by VarElect tool. Several important OP-associated pathways were detected by protein-protein interaction and pathway enrichment analysis. Target genes were further enriched for differential expression genes in osteoblasts expression profiles, e.g. *IBSP*, affecting calcium and hydroxyapatite binding, and *CD44*, regulating alternative splicing of gene transcription. Transcriptome fine-mapping identifies more disease-related genes and provide additional insight into the development of novel targeted therapeutics to treat OP.

Keywords: Osteoporosis; Transcriptome-wide association study; Colocalization method; Gene expression

Background

Osteoporosis(OP) is a highly polygenetic disease which has been studied intensively on the genetic level, resulting in abundant detections associated with gene loci and polymorphisms (Peacock et al. 2002; Clark and Duncan 2015). Osteoporosis is defined clinically that bone mineral density is 2.5 standard deviations or more below the young adult mean and remains the single golden standard predictor of primary osteoporotic fractures (Nguyen et al. 2007; Duncan and Brown 2010; Rachner et al. 2011). Bone mineral density is highly heritable, with evidence showing that the heritability of bone mineral density ranges from 50% to 80% (Duncan and Brown 2010). The recent large genome-wide association study (GWAS) to date estimated bone mineral density at the heel in 426,824 individuals and identified 1,103 independent genome-wide significant associations at 518 loci (Morris et al. 2019). They explain about 20% phenotypic variance in estimated bone mineral density. However, the majority of GWAS hits are in non-coding regions and their biological methnasims are difficult to understand (Moonesinghe et al. 2008; Nicolae et al. 2010).

The effect of genetic variation on phenotype is complex, where it may alter the abundance of one or more proteins by regulating gene expression and then affects the trait (SNP-Expression-Phenotype) (Musunuru et al. 2010; Lappalainen et al. 2013; Westra et al. 2013; Albert and Kruglyak 2015; Zhang et al. 2015). Gene expression is arguably the most impactful and well-studied effect of regulatory genetic variation. GWAS loci are enriched for expression quantitative trait loci (eQTL), rendering it a potential link between genetic variant and biology of disease (Stranger et al. 2007; Gusev et al. 2014; Lee et al. 2015). While most GWAS studies do not concomitantly measure gene expression, the influence of genetic variation on gene expression allows us to use gene expression reference datasets to predict gene expression given a set of genotypes, and subsequently identify new disease-associated genes (Nica et al. 2010; Nicolae et al. 2010; Albert and Kruglyak 2015). Transcriptome-wide association study (TWAS) approach has been implemented to identify genes with expression associated with complex traits by integrating genetic and transcriptional variation (Gusev et al. 2016; Barbeira et al. 2018). Instead of testing millions of SNPs in GWAS, TWAS evaluates the association of predicted expression for thousands of genes, greatly reducing the burden of multiple comparisons in statistical inference. This approach has been shown to have the potential to identify the genes responsible for GWAS-identified associations for complex traits and provide mechanistic insight regarding genes being regulated via disease-associated genetic variants (Mancuso et al. 2017; Gusev et al. 2018; Lu et al. 2018; Wu et al. 2018; Atkins et al. 2019).

In this paper, we conducted transcriptome-wide association study to identify genes associated with OP by integrating gene expression from the Genotype-Tissue Expression (GTEx) and GWAS summary data from the Genetic Factors for Osteoporosis (GEFOS) Consortium, and then evaluated the biological patterns of expression-trait association by COLOC method. Next, we performed VarElect to understand the biological function of association between the TWAS-significant genes and OP. Comparing with the results of differential analysis of the two mRNA expression

profiles for OP, we further verified the causal associations between OP and TWAS-significant genes.

Results

TWAS identified candidate genes for OP

We performed TWAS method. Two gene-expression reference panels in muscle-skeletal and whole blood were used, with totally 13,416 genes towards GWAS summary data from GEFOS consortium to identify novel genes associated with OP. TWAS identified 276 significant associated genes at p-value < 3.7E-6, as shown in Figure 1.

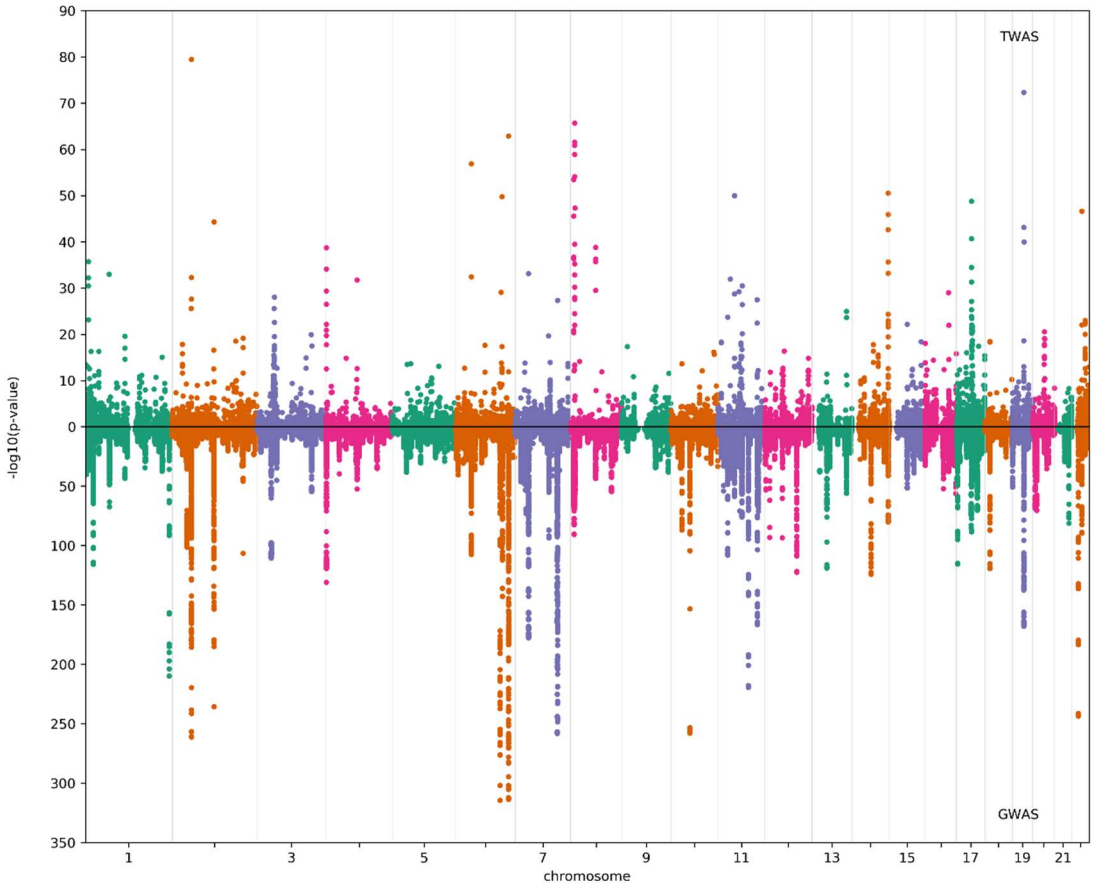
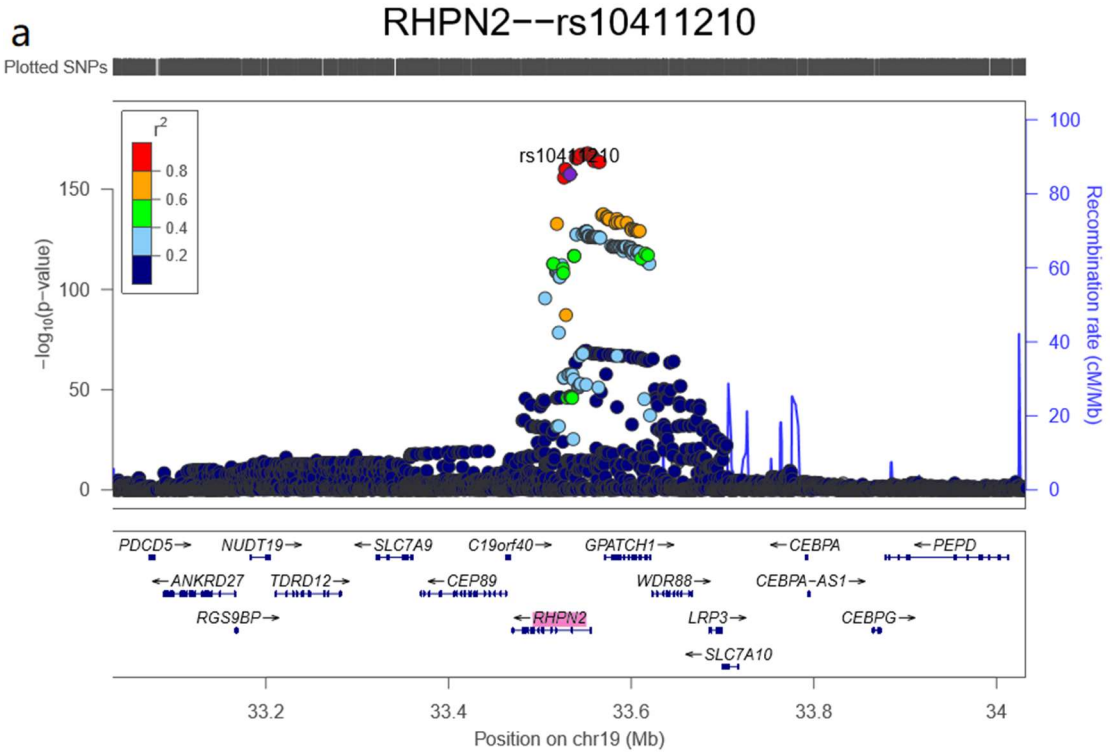
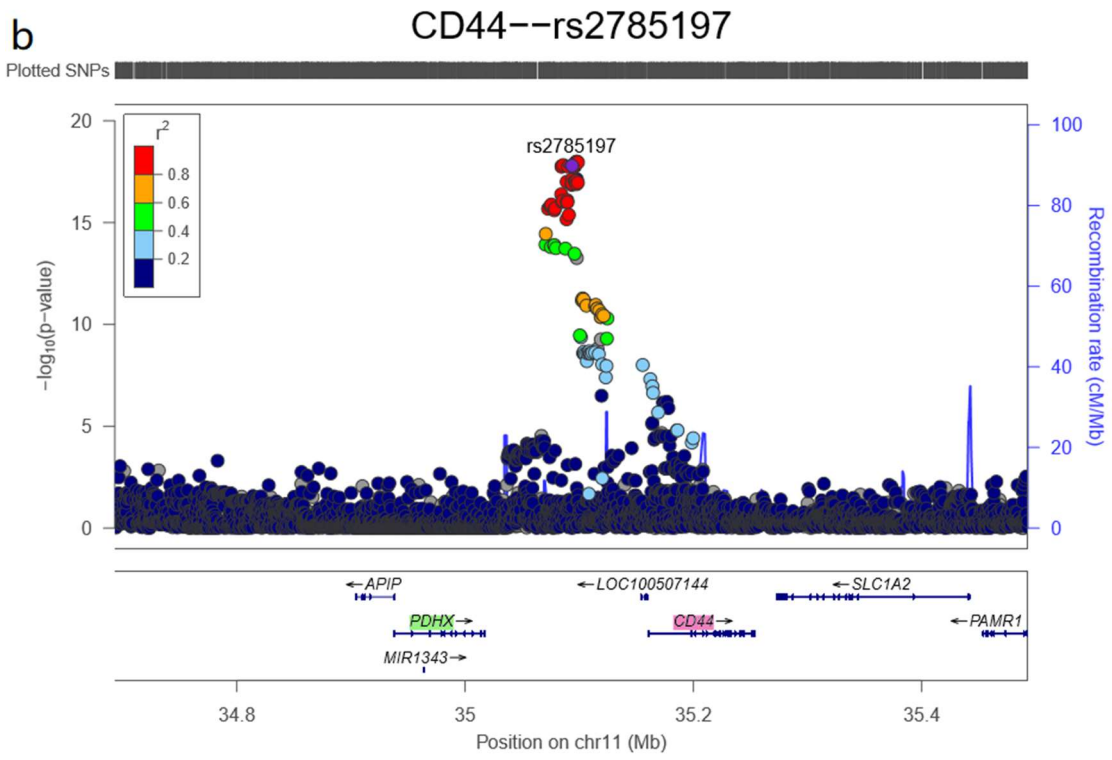


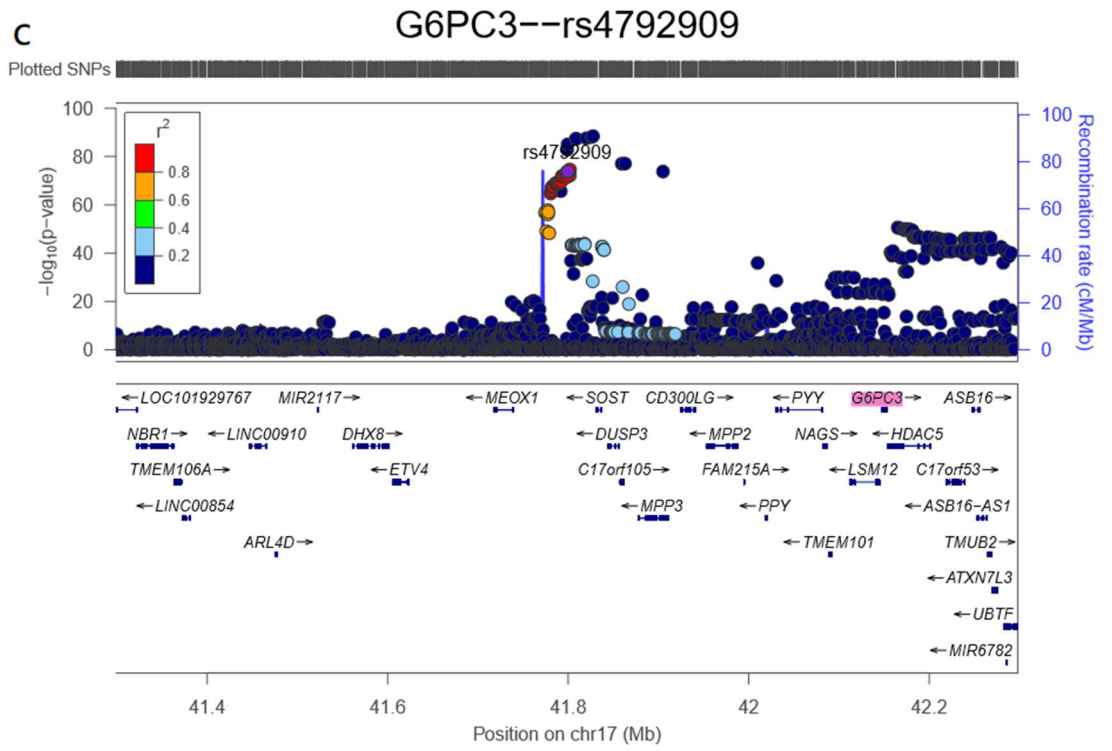
Figure 1. A manhattan plot of the results from TWAS analysis and GWAS analysis for OP. The transcriptome−wide significance threshold is p-value = 3.7E-6; The genome−wide significance threshold is p-value = 6.6E-9. There are 1,103 conditionally independent SNPs at 515 loci passed the criteria for genome-wide significance in n = 426,824 UK Biobank participants.

TWAS method can detect causal genes by introducing the effect prediction of genetic variants on the gene expression. There are the following four biological patterns identified by TWAS (Figure 2). First, for significantly associated SNPs with OP in the coding regions (introns and extons), the causal genes identified by GWAS and TWAS are more likely to be consistent, as shown in Figure 2a. The effect size of rs10411210 (P_{GWAS} = 1.6E-119) on OP in GWAS is corresponding with that of rs10411210 on *RHPN2* (P_{TWAS} = 4.4E-73) gene expression in TWAS. Second, for SNPs in non-coding regions, the candidate genes may be close to the significant eQTLs but different from

the GWAS hits, as shown in Figure 2b. Variant rs2785197 (P_{GWAS} = 6.5E-44) in 11p13 mapping to *PDHX* in GWAS, but the causal gene for rs2785197 is more likely to be *CD44* (P_{TWAS} = 1.1E-32) in our TWAS. The colocalization analysis showed *CD44* (PP4=0.99 in Supplementary Table 3) gene expression was regulated by single variant rs2785197, which may be regarded as its expression regulation element. Third, the candidate genes may be regulated by relatively distant significant SNP in non-coding regions, as shown in 2c. Our TWAS results indicated that rs4792909 (P_{GWAS} = 1.5E-74) in 17q21.31 may be associated with *G6PC3* (P_{TWAS} = 4.2E-26). The distance between rs12478002 and *G6PC3* was 349kb, but we did not found gene reported by GWAS near rs4792909. Fourth, the candidate genes were discovered in non-significantly associated SNPs with OP. There GWAS non-significants regions as novel loci: rs1003260 (P_{GWAS} = 3.6E-08) in the 6q13 associated with *RIMS1* (P_{TWAS} = 2.1E-8) shown in Figure 2d, rs12917011 (P_{GWAS} = 2.1E-06) in the 15q23 associated with *SPESP1* (P_{TWAS} =3.3E-8) shown in Supplementary Figure 6a, rs2251381 (P_{GWAS} = 1.4E-06) in the 21q21.3 associated with *MAP3K7CL* (P_{TWAS} = 1.1E-9)shown in Supplementary Figure 6b. These 3 novel discoveries are firstly reported to be associated with BMD and further investigation can be performed.







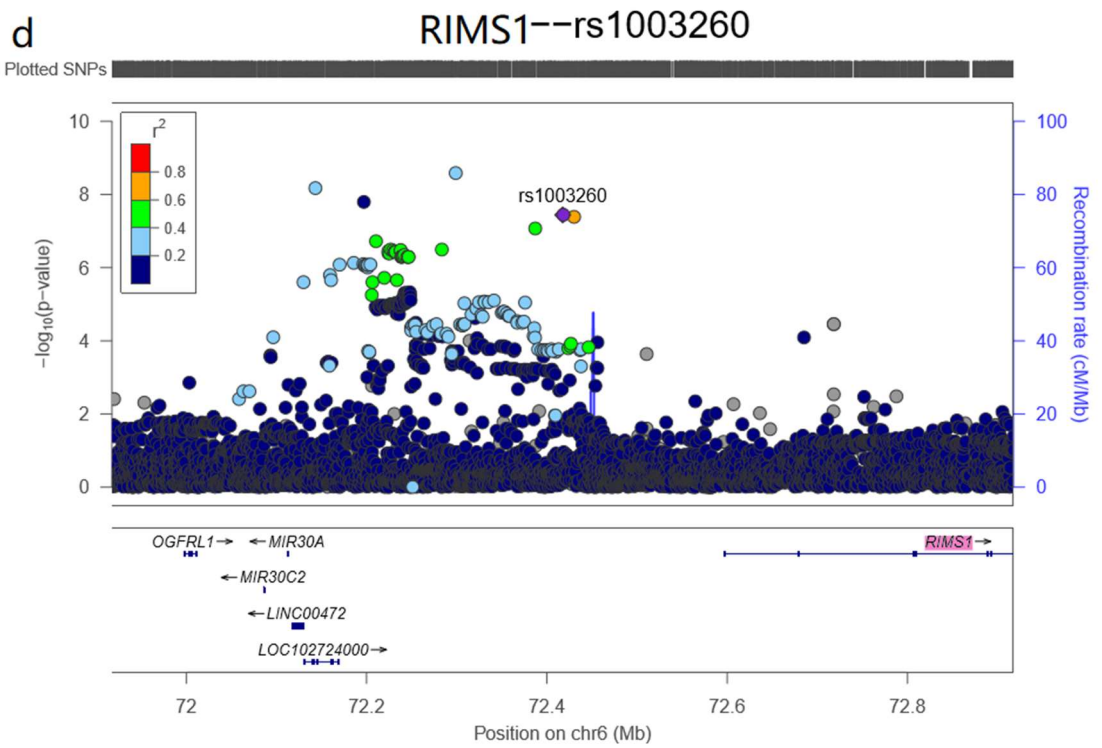
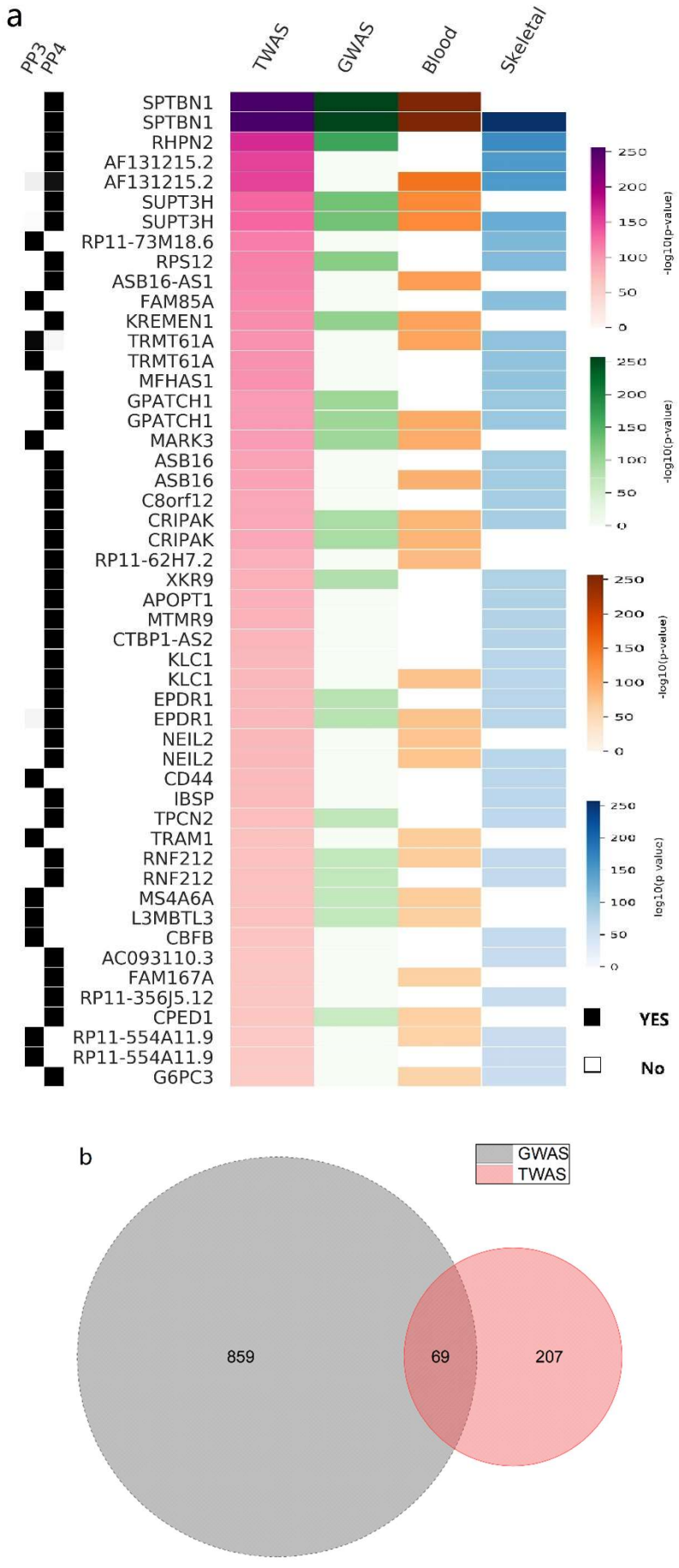
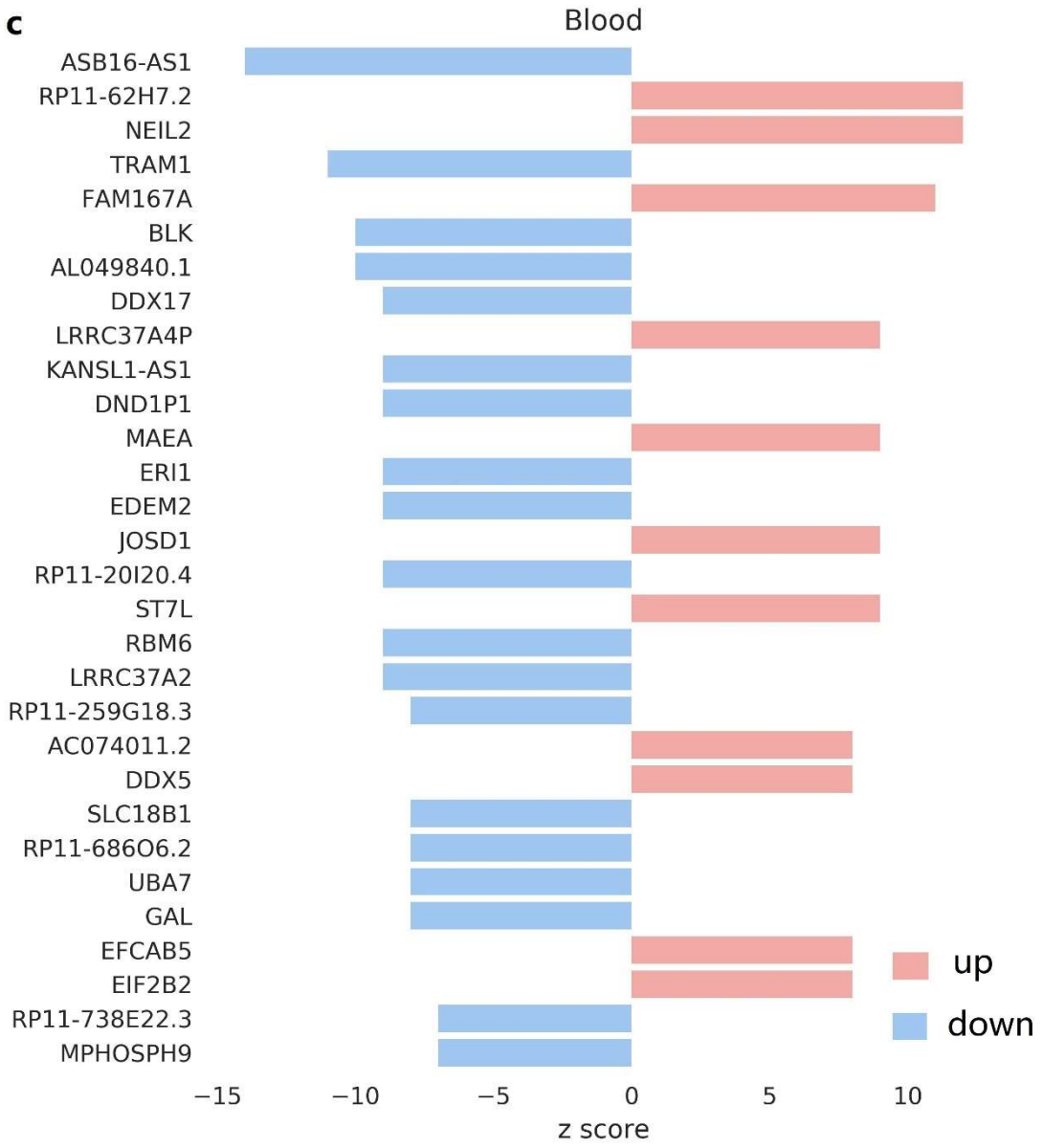


Figure 2. Biological patterns identified by TWAS. a: For significant SNPs in the coding regions, rs10411210 (P_{GWAS} = 1.6E-119) in 19q13.11 is associated with *RHPN2* (P_{TWAS} = 4.4E-73). b: For SNPs in the non-coding regions, rs2785197 (P_{GWAS} =6.5E-44) in 11p13 associated with *PDHX* marked green in GWAS, but The causal gene for rs2785197 is more likely to be *CD44* marked red (P_{TWAS} = 1.1E-32) in our TWAS. c: rs4792909 (P_{GWAS} = 1.5E-74) in 17q21.31 may be associated with *G6PC3* (P_{TWAS} = 4.2E-26). The distance between rs4792909 and *G6PC3* was 387kb, but there is not found gene reported by GWAS near rs4792909. d: rs1003260 (P_{GWAS} = 3.6E-08) in the 6q13 associated with *RIMS1* (P_{TWAS} = 2.1E-8). rs1003260 is not significant in GWAS.

Genes expression difference identified by TWAS may be causally associated with the phenotype of interest, but also can be due to variants LD or co-expressions (Huang et al. 2015; Hormozdiari et al. 2016). To pinpoint causal relationship between the target gene of an eQTL and the complex trait, we performed colocalization analysis by using COLOC method; see Methods section. The results showed that 134 TWAS associations have strong evidence of joint causal variants with PP3 > 0.9 shown in Figure 3a and Supple mentary Figure 1 and Table 1, and 142 have evidence of a single shared causal variant with PP4 > 0.8 shown in Figure 3a and Supplementary Figure 1 and Table 2.

Comparing with previous GWAS studies, we observed that 3 genes located in novel loci and 204 genes have not been reported to be associated with OP risk in previous GWAS loci, 69 genes were previously implicated to be OP risky by literature using either GWAS or functional studies shown in Figure 3a-3b. We also found that 117 (117/276) genes were expressed in skeletal tissue and 71 (71/276) genes were expressed in blood tissue, and 88 (88/276) genes were expressed in both tissues as shown in Figure 3a. For 69 genes found in previous GWAS studies, our results provided additional evidence to support these previous findings. For the rest genes, 79 in blood tissue and 128 in skeletal tissue were considered as novel candidates shown in Figure 3c-3d.





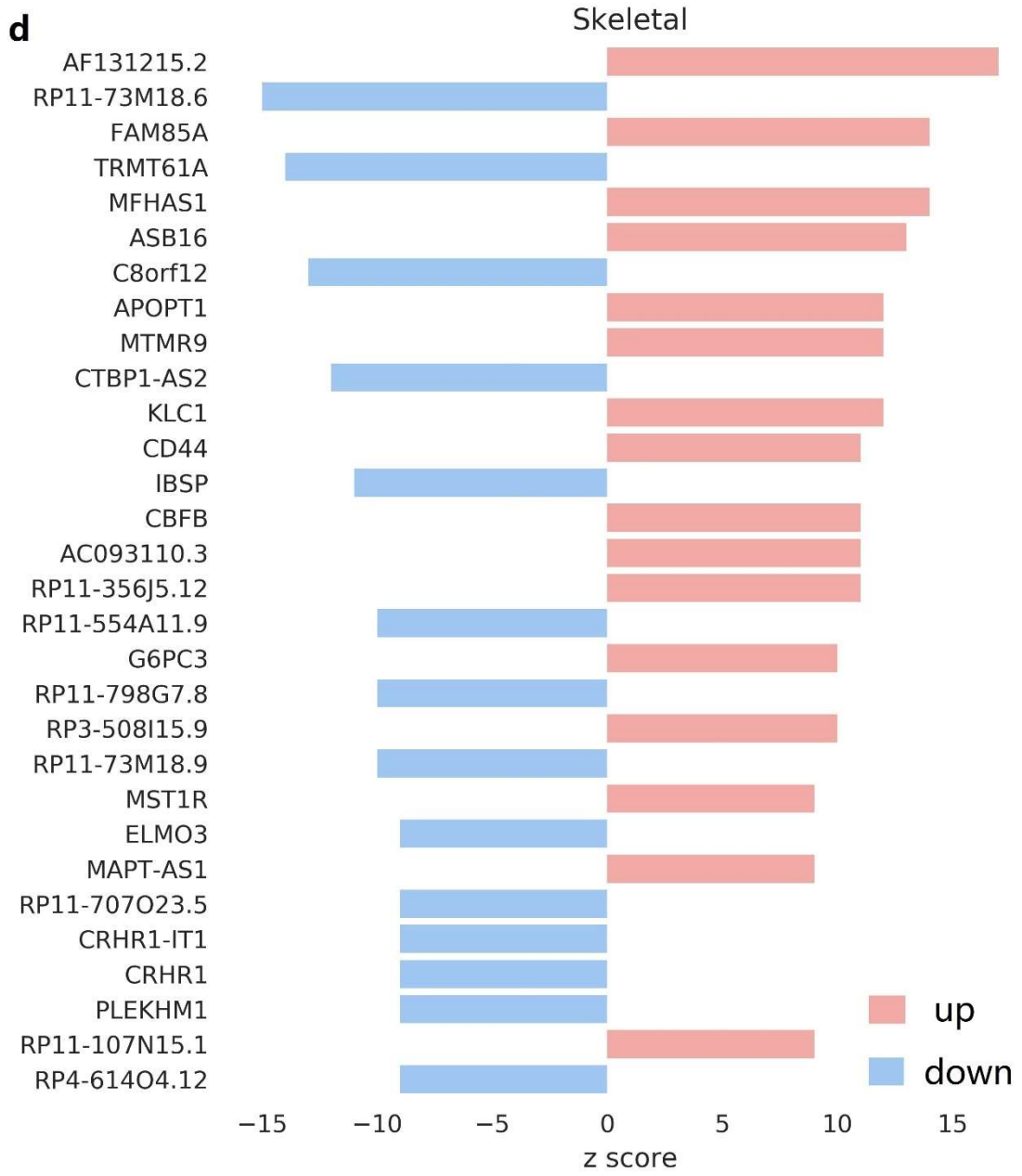


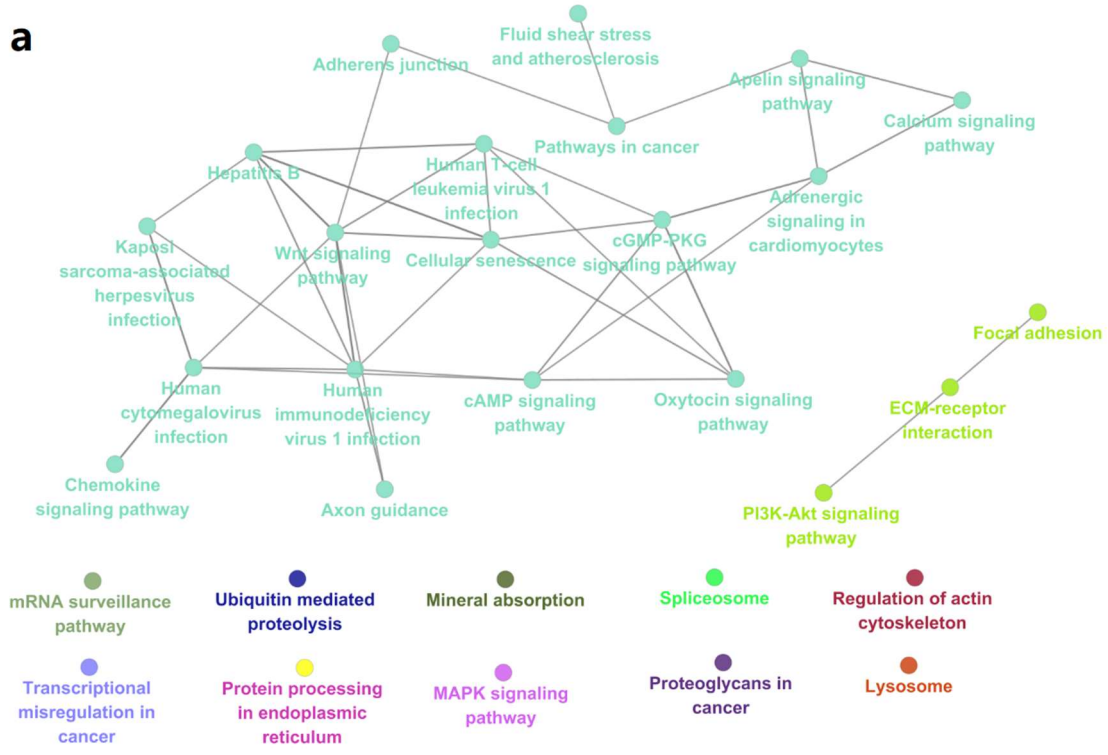
Figure 3. TWAS−significant genes and novel candidate genes in blood and skeletal tissue. PP3: represents trait−gene expression associations are caused by two distinct causal variants. PP4: represents trait−gene expression associations are caused by a signale causal variants. a: Heatmap of top 50 TWAS−significant genes, whether discovered in previous GWAS study, blood tissue or skeletal tissue, whether passed the criteria for colocalization analysis in PP3 > 0.9 or PP4 > 0.8 (full lists can be found in Supplementary Figure 1, Supplementary Table 2 and 3). b: Comparison of associated genes found by TWAS and GWAS methods. c and d: Top 20 novel candidate genes were found in blood and skeletal tissues respectively, the red bars represent gene expression up-regulated, and blue bars indicate down-regulated (full lists can be found in Supplementary Figure 2 and 3, Supplementary Table 2 and 3).

Assessment of the candidate gene−OP association

For 207 novel candidate genes, we evaluated the associations between the candidate genes and OP by implementing VarElect analysis. The analytical results showed that 24 genes (Supplementary Table 3) were 'direct' associations and 129 genes were 'indirect' associations (Supplementary Table 4); the rest were unclassified yet. The direct associations indicated the target genes were supported by rich evidences (the relevant literature, gene function annotation, etc.). The score in Supplementary Table 3 indicated the strength of the association between the gene and OP: the higher score, the stronger evidence. Indirectly associated genes may interact with intermediary to influence the development of OP, though protein interaction netorks and pathways (Supplementary Table 5). The remaining uncharacterized genes are mainly lncRNA, transcripts as the potential disease factors without available evidence requiring further investigations.

Functional pathways of the candidate genes

In order to further verify the associations between the TWAS−significant genes and OP, we explored the biological function pathways of these genes by applying STRING and CluePedia tool. We found the majority of pathways were easy understanding to the occurrence of OP and some of them interact with each other (e.g. focal adhesion and ECM-receptor interaction, PI3K-Akt signaling), as shown in Figure 3a and Supplementary Table 5. We further enriched the functional pathways for the categorized gene lists (direct, indirect, in blood, in skeletal). We found all significant pathways (p−value < 0.5) were enriched in the skeletal tissue while part enriched in blood. 'Direct' genes can be enriched in the critical pathway such as mineral absorption and calcium signaling pathway. These results showed that TWAS−significant genes involved many biological mechanisms in developing OP.



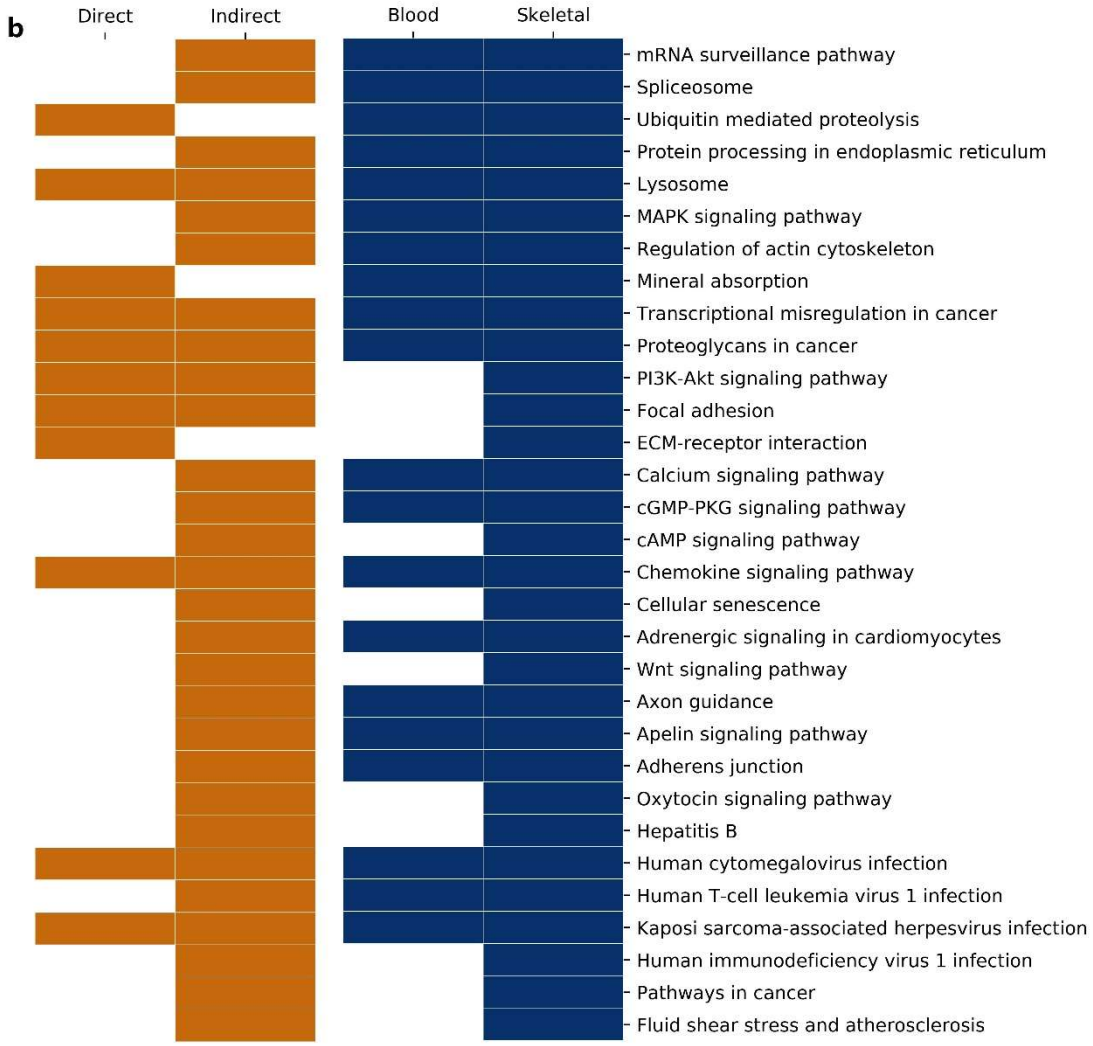
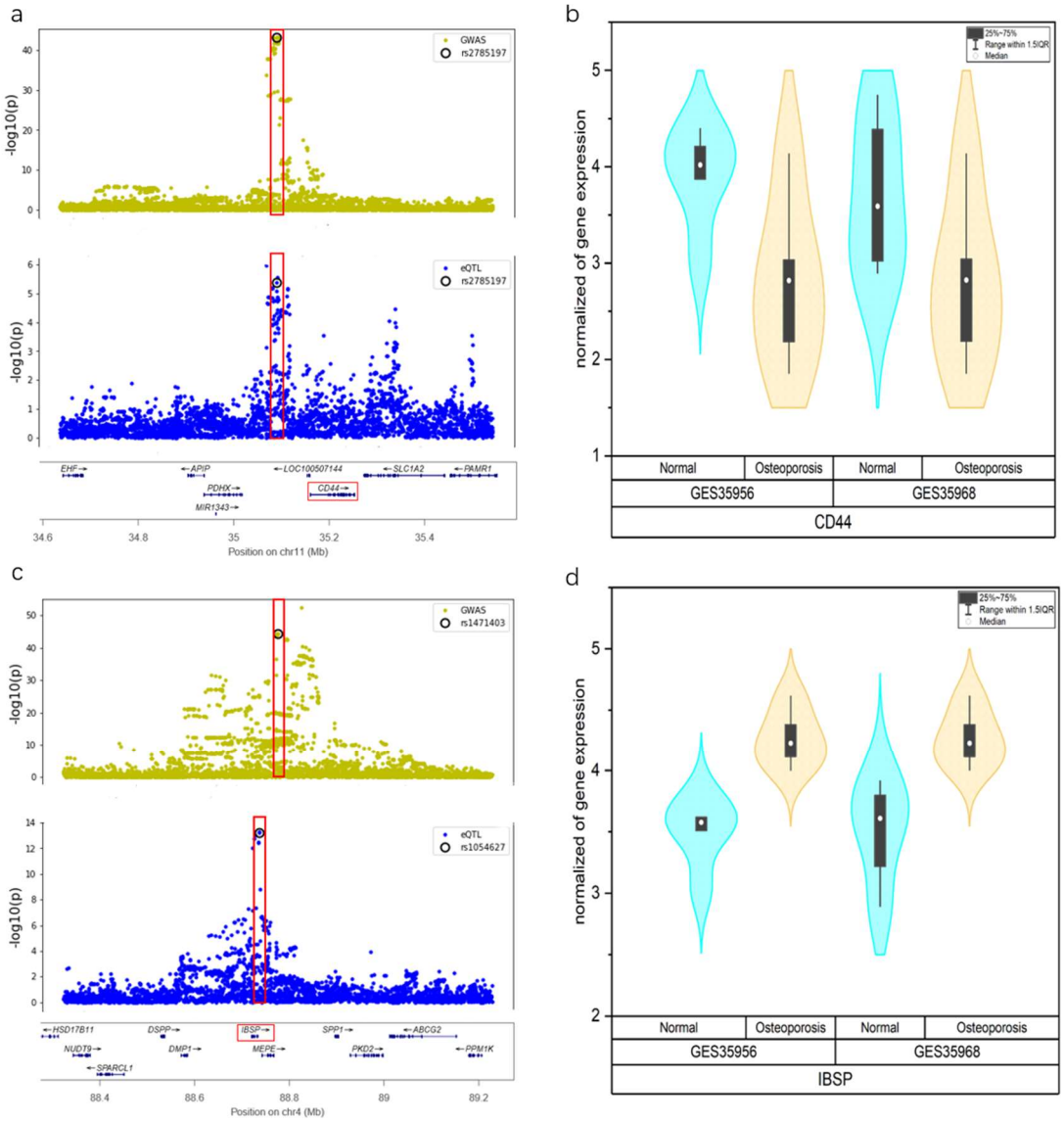


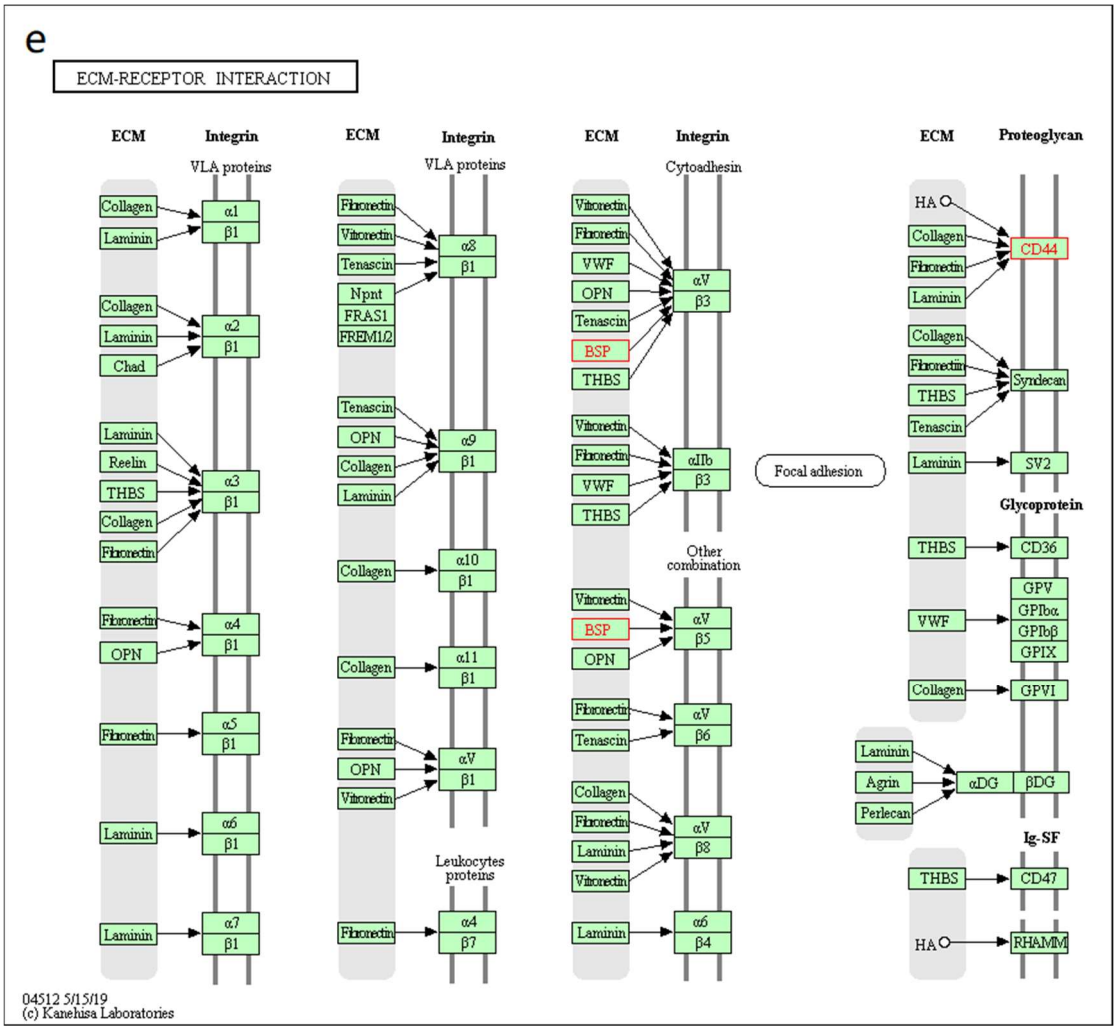
Figure 4. The functional pathways of TWAS−significant genes. They are enriched by applying STRING and CluePedia tool. Significance of pathways is determined by the hypergeometric test (one−sided) followed by Fisher's combined probability test (one−sided) to determine combined pathway significance (p−value < 0.5). a: The functional pathways of TWAS−significant genes, the circles represent functional pathways, and the line represents the interactions between pathways. b: Classification of functional pathways according to the categorized gene lists. Genes in skeletal tissue are enriched in all significant pathways. The genes list for each pathway are found in Supplementary Table 5.

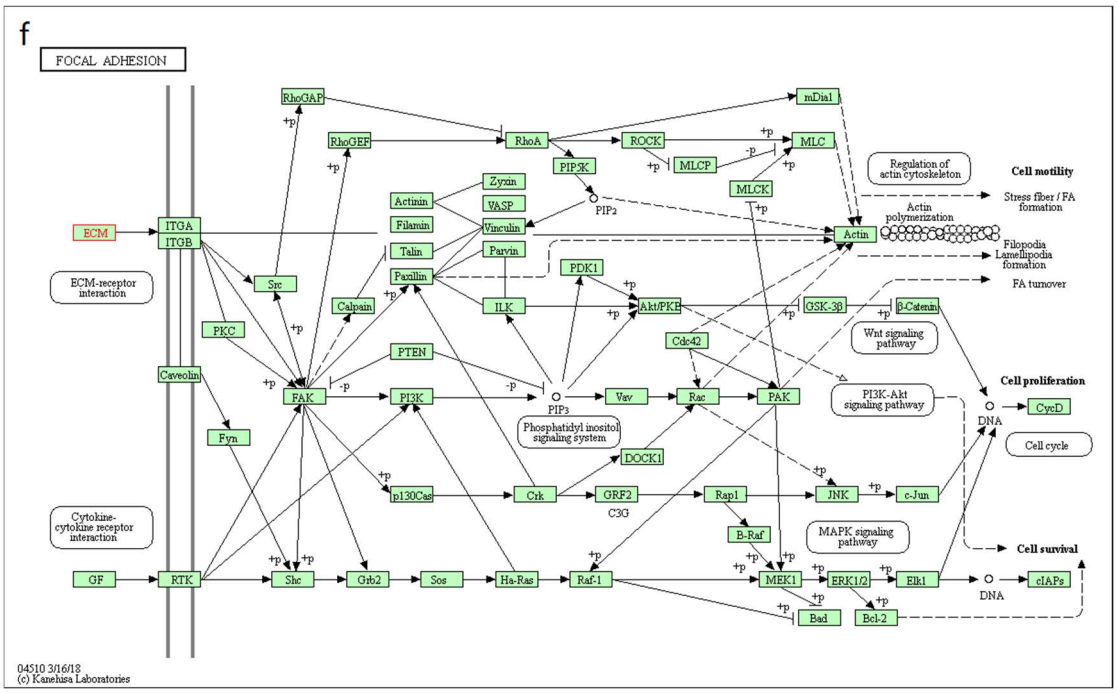
Functional validation for the candidate genes

Previous research, utilizing expression profiling with gene signatures of cellular models to characterize the gene's involvement in bone metabolism and disease processes, found that impaired osteoblastic differentiation reduces bone formation and causes severe OP in animals (Stein et al. 1990; Wu et al. 2003; Misof et al. 2012). We analyzed two gene expression profiles GSE35956 and GSE35959 from GEO, containing two groups people: the primary OP and normal. Based on the cut-off criteria of $p < 0.05$ and logFC > 1 to select DEGs, a total of 156 and 265 DEGs were identified from GSE35956 and GSE35959 datasets. Comparing DEGs with TWAS-significant genes, 5 up-regulated and 2 down-regulated genes overlapped in two type datasets shown in Supplementary

Figure 4. We observed that these genes were quite significant in TWAS, and their expression differences were also consistent with COLOC analysis results, as shown in Figure 4a-4d. Therefore, we inferred that these genes are very likely to be the causal pathogenic gene of OP. The results of functional pathway analysis also supported our findings, as shown in Figure 4e-4g, *IBSP* and *CD44* included in the ECM-receptor interaction pathway, which is a branch of the focal adhesion pathway and acts on PI3K-AKT signaling pathway. Owning to the small samples size of gene expression datasets, more experiments are needed in the future.







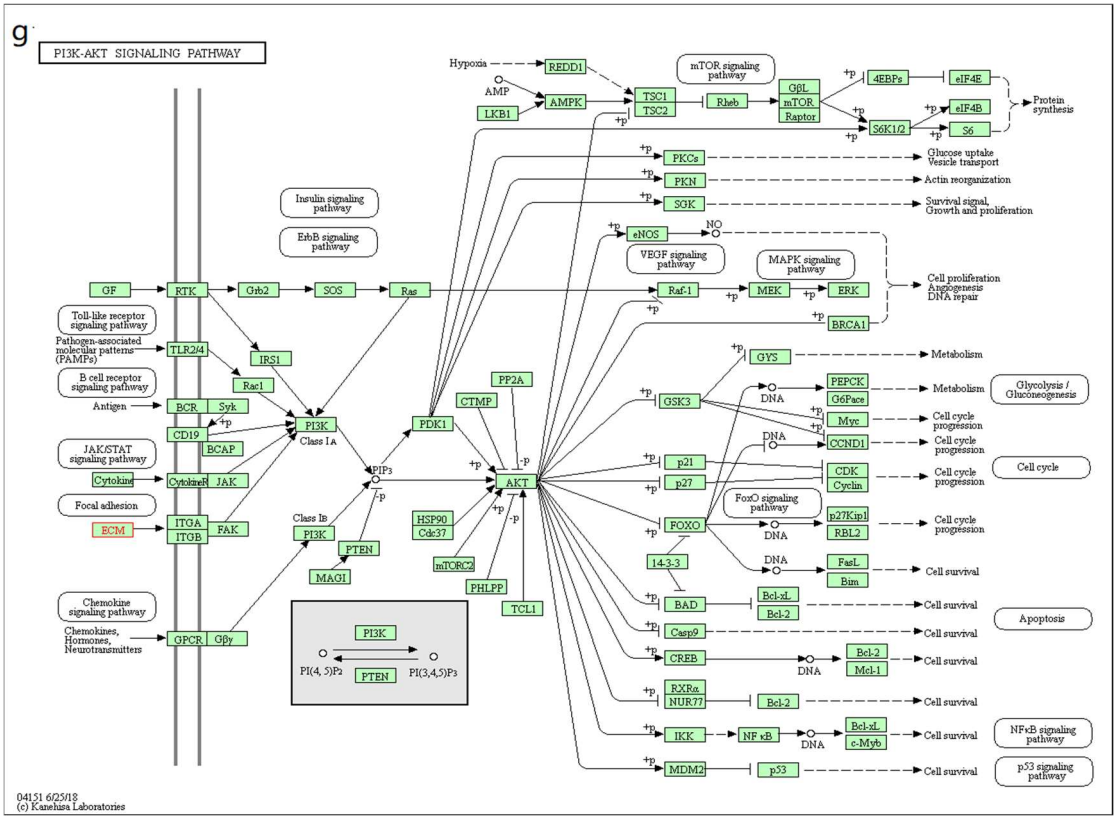


Figure 5. Biological function verification for genes. TWAS−significant gene *CD44* (P_{TWAS} =1.1E-32) and IBSP (P_{TWAS} =1.8E-32) are validated by COLOC method, gene expression profiles, and biological function pathways. a and c: The colocalization analysis results for gene *CD44* and *IBSP*, showing a single shared causal variant rs2785197 and joint causal variants rs1471403 and rs1054627 respecively. b and d: Gene expression for *CD44* and *IBSP* in the GSE35956 and GSE35959 datasets. e-g: *IBSP* and *CD44* are enriched in the ECM-receptor interaction pathway, which is a branch of the focal adhesion pathway and acts on PI3K-AKT signaling pathway, also see Figure 4a.

TWAS for OP identifies new loci

We found 3 genes in novel loci. *RIMS1* (P_{TWAS} = 2.1E-8) associated with rs1003260 (P_{GWAS} = 1.8E-8, MAF = 0.125) and is a *RAS* gene superfamily member in 6q13 that regulates synaptic vesicle exocytosis; *SPESP1* (P_{TWAS} =3.3E-8) in 15q23 associated with rs12917011 (P_{GWAS} = 2.1E-06, MAF = 0.438) code a human alloantigen involved in spermegg binding and fusion; *MAP3K7CL* (P_{TWAS} = 1.1E-9) associated with rs2251381 (P_{GWAS} = 1.4E-06, MAF = 0.367) and is a protein coding gene in 21q21.3. VarElect analysis showed that the biological function of three genes were indirectly associated with OP and provided evidence for causality for OP shown in Supplementary Table 4. However, we did not enrich significant functional pathways for three genes, the causal effect of them on OP needs to be verified by advanced biological experiments. As shown in Figure 2d and Supplementary Figure 6a-6b, we observed that the distance between the three causal SNPs and causal genes is within 500kb, and other significant GWAS SNPs were not found, the results indicated one of the advantage of TWAS method, which can find causal genes in the non-significant GWAS regions.

Discussion

Multiple GWAS studies have been performed with considerable sample sizes to detect OP heredity, yet progress towards understanding disease mechanisms has been limited. Most GWAS hits are in non-coding regions and difficult to understand the downstream biological inference. In most cases, the nearest genes were uaually reported (Smemo et al. 2014; Claussnitzer et al. 2015; Spain and Barrett 2015). In fact, SNPs in the non-coding region did not have to regulate gene based on the distance between SNPs and genes. Intergreting GWAS data and transcriptome data will empower novel discovery and possibly pinpoint the causality. TWAS method calculated local SNPs−gene expression correlations, and further calculated likelihood of genes causality. Therefore, for a significant SNPs in the coding region, the causal genes identified by GWAS and TWAS should be and indeed are consistent, as shown in Figure 2a. For SNPs in non-coding regions, the causal genes may be close to the significant eQTLs but different from the GWAS hits as shown in Figure 2b. TWAS method can even discover causal genes non-significantly associated SNPs with OP shown in Figure 2c, and relatively distant significant SNP shown in Figure 2d. More valuable region plots can be found in Supplementary Figure 5-6.

We totally found 276 candidate genes, of which 69 were replicated in GWAS, and the rest 207 were novel candidates. Among them, 142 target genes are regulated by two distinct causal variants, and 134 target genes share one causal variant. By analyzing the biological functions behind, we found that 24 novel candidate genes directly affect the pathways closely related to the development of osteoporosis in our results: *IBSP*, *EIF2B2, CD44*, *FEN1*, *UBA7*, *MARCO*, *ATF1,CBFB*, *G6PC3*, *SLC11A2*, *GAL*, *CCR3*, *MST1R*, *PLEKHM1*, *ATRIP*, *CCDC36*, *AKAP7*, *EPRS*, *CTSB*, *ASB16-AS1*, *CRHR1*, *FADS1*, *MAP1LC3A*, *MAEA*. For example, *SLC11A2* enriched in mineral absorption pathway regulates the fine-tuned balance between bone resorption and bone formation and thus affects bone density (Xu et al. 2017), shown in Supplementary Figure 7. In the other hand, 129 novel candidate genes seem exerting their biological functions to affect the development of osteoporosis through protein-protein interaction networks. As shown in Supplementary Figure 8, *RAC3* and *NFATC4* were enriched in the MAPK signaling pathway through interacting with genes (*ESR1*, *FOS*, *IGF1*, *TGFB1*, *JUN*, *NFATC1*, *IGF*1, *LRP5*, *TNF*, *PRKACA*) known to be associated with osteoporosis. MAPK signaling pathway is involved in the regulation of many cellular physiological functions such as proliferation, differentiation, inflammation, and apoptosis, and affect bone formation (Peng et al. 2009; Wanachewin et al. 2012). More information on gene interactions can be found in Supplementary Table 5.

We found *RIMS1*, *MAP3K7CL*, *SPESP1* located in new loci and their causal SNPs were non-significantly associated with OP in GWAS. *RIMS1*, regulating synaptic membrane exocytosis 1, is a RAS gene superfamily member and plays a role in the regulation of voltage-gated calcium channels during neurotransmitter and insulin release. *MAP3K7CL*, *MAP3K7* C-terminal like, is a protein coding gene. The GO annotation (GO:0005515) showed *MAP3K7CL* interact selectively and non-covalently with any protein or protein complex. But there is little research on its biological function.

SPESP1 code sperm equatorial segment protein 1 involved in fertilization ability of sperm. The current studies have not supported evidence for the causal association between three genes and OP, so we hope to have follow-up experiments to verify them.

Furthermore, we provided additional evidence by comparing with differential expression genes by analyzing two gene expression profiles in OP and non-OP groups. We found seven significant differential expression genes in our results: *IBSP*, *CD44*, *SPTBN1*, *PAPSS2*, *TRAM1*, *PPP1CB*, *NCKAP1*, shown in Supplementary Figure 4. *IBSP* is remarkably downregulated and associates with OP significantly (TWAS p-value=1.8E-32). SNP rs1471403 and rs1054627 may co-regulate gene expression of *IBSP* (PP3=1, Figure 4c-4d). Previous studies showed that IBSP is expressed in all major bone cells including osteoblasts, osteocytes and osteoclasts (Trošt et al. 2010) and encodes a major non-collagenous bone matrix protein binding to calcium and hydroxyapatite via its acidic amino acid clusters (Mafi Golchin et al. 2016). Another discovery *CD44* is remarkably upregulated. Previous research argued that a linkage synonymous mutation in exon 9 of the *CD44* gene through a cell experiment, may increase the susceptibility of the family to OP by influencing alternative splicing of gene transcription (Vidal et al. 2009). Information about other genes can be found in Supplementary Table 6.

This is as yet largest study integrating GWAS and TWAS to identify susceptibility genes of OP. We used data from the 426,824 individuals GWAS of OP and 860 samples TWAS in our analyses. Many findings were discovered, although there still exist limitations of this research. First, TWAS method cannot explain the variants influencing disease that are independent of cis expression, as it was only trained on cis-eQTL analysis. Second, there may be bias using normal blood and skeletal tissues from GTEx to make predictions. Third, tissue sensitivity and tissue specificity are important issues when running TWAS. Prediction models built on gene expression data from osteoblasts cells in OP patients will help identifying additional candidate genes associated with OP (Orlic et al. 2007).

In summary, we integrated data from GWAS and transcripome expression to identify 276 candidate genes associated with OP; 69 of them were replicated from GWAS, and 204 novel candidate genes in loci reported by GWAS and 3 novel candidate genes in new loci. We analyzed biological patterns of those loci and explained their pathway interactions. We hope that our findings will provide novel insights into the future pathogenetic studies of OP.

Methods

GWAS summary datasets of OP

The GWAS summary statistics for OP was derived from GEFOS Consortium website (URL) in December 2018. The phenotype feature of OP was measured by bone mineral density estimated from quantitative heel ultrasounds. The large scale GWAS analysis for OP were performed in a cohort of 426,824 participants (55% female) from UK Biobank (Morris et al. 2019). Briefly, GWAS analysis was performed based on the HRC imputation panel (hg19) including about 14,000,000 SNPs with MAF $\geq$ 0.05% and acceptable imputation quality (info score > 0.3). A detailed description of sample

characteristics, experimental design, and statistical analysis can be found in the published study (Sudlow et al. 2015).

Integration of GWAS and gene expression

To integrate GWAS results and gene expression, we used TWAS method. We included two relevant reference transcriptome datasets in our analysis: whole blood and muscle-skeletal from GTEx v7. TWAS method integrated information from expression reference panels (SNP−gene expression correlation), GWAS summary statistics (SNP−OP correlation), and linkage disequilibrium (LD) reference panels (SNP−SNP correlation) to assess the association between the cis−genetic component of expression and trait (expression−OP correlation) (Gusev et al. 2016; Gusev et al. 2018), In practice, the effect sizes of cis-SNP−expression in the 500kb loci region were estimated with a sparse mixed linear model (Zhou et al. 2013). TWAS used pre−computed gene expression weights combined with GWAS summary statistics to calculate the association effect for each gene to disease. In this study, the gene expression weights of whole blood and muscle−skeletal were derived from the FUSION website (URL). The genes with significant association signals were identified at p-value < 3.7E-6 after strict Bonferroni correcting.

Evaluation of trait−gene expression associations

To evaluate the reliability of TWAS analysis results and understand the biological mechanisms of trait−gene expression associations, we performed COLOC method (Giambartolomei et al. 2014). COLOC method uses asymptotic Bayes factors with summary statistics and regional LD structure to estimate five posterior probabilities: no association with either GWAS or eQTL (PP0), association with GWAS only (PP1), association with eQTL only (PP2), association with GWAS and eQTL but two independent SNPs (PP3), and association with GWAS and eQTL having one shared SNP (PP4). For each of the GWAS hits, we defined a 500kb region at either side of the index variant and tested for colocalization within the entire cis−region of any overlapping eQTLs (transcription start and end position of an eQTL gene plus and minus 500kb, as defined by GTEx) in two human tissues from GTEx v7. A signal with PP3 > 0.9 was considered the evidence for trait − gene expression associations caused by two distinct causal variants from GWAS and eQTL. A signal with PP4 > 0.8 was considered the evidence for trait−gene expression associations caused by a joint signal from GWAS and eQTL.

Assessment of gene-disease associations

To assess the likelihood of functional genes which are more likely to be causal, VarElect (Stelzer et al. 2016a; Stelzer et al. 2016b), a cutting-edge Variant Election application for disease/phenotype-dependent gene variant prioritization, were used to assess the associations of biological function between the candidate genes and OP. VarElect provides a robust algorithm for ranking genes within a short list, and pointing out their likelihood associated with disease, and produces a list of prioritized, scored, and contextually annotated genes and direct links to supporting evidence and additional information. VarElect utilizes the deep LifeMap Knowledgebase to infer the 'direct'

or 'indirect' association of biological function between genes and phenotypes. 'Direct' association between genes and disease has been supported by many studies that genes can directly affect the development of disease. 'Indirect' association between genes and disease are based on shared pathways, protein-protein interaction networks, paralogy relations, domain-sharing, and mutual publications.

PPI network and pathway enrichment analysis

The functional networks of TWAS−significant genes with OP were further validated by STRING and CluePedia tool. STRING (Search Tool for the Retrieval of Interacting Genes, URL) is an online tool designed to evaluate the protein−protein interaction (PPI) networks (Szklarczyk et al. 2015; Szklarczyk et al. 2017). The CluePedia is a plugin of Cytoscape software and search for potential genes associated with the certain signaling pathway by calculating linear and nonlinear statistical dependencies from experimental data (Shannon et al. 2003; Bindea et al. 2013). The PPI networks of TWAS-significant genes was constructed by STRING. The functioal pathways were detected and visualized by CluePedia. The pathways were identified at p-value < 0.5 (Bindea et al. 2013).

Differential analysis of gene expression

To further validate the functional causslity of candidate genes, we compared the candidate genes with differential expression genes (DEGs) in osteoblasts for osteoporosis sufferer. The original datasets comparing the gene expression profiles between OP and normal controls were downloaded from NCBI GEO databases (URL). Two gene expression profiles GSE35956 and GSE35959 were based on GPL570 (Affymetrix Human Genome U133 Plus2.0 Array, Affymetrix, SantaClara, CA, U.S.A). We performed robust multi−array average approach (Hochreiter et al. 2006) for background correction and normalization. The original GEO data were then converted into expression measures. Limma package (Smyth 2005) was used for determining DEGs between OP samples and non-OP samples ($p < 0.05$ and $log2FC > 1$ as the cut-offcriterion).

Data access

GWAS summary data are available in the Genetic Factors for Osteoporosis (GEFOS) Consortium (http://www.gefos.org/); gene expression weights of whole blood and muscle−skeletal were derived from the FUSION website (https://gusevlab.org/projects/fusion/); Two gene expression profiles are available in NCBI GEO databases under der accession number GSE35956 and GSE35959.

URL

GEFOS: http://www.gefos.org/
Fusion: https://gusevlab.org/projects/fusion/
GEO: https://www.ncbi.nlm.nih.gov/geo/
STRING: https://www.string−db.org/cgi/

Competing interests

The authors declare that they have no competing interests.

Author's contributions

PY and MZ conceived the project and designed the experiments. MZ, HF, JJ, LY, WS and YL analyzed the data. PY and MZ wrote the manuscript. All authors read and approved the final manuscript

Acknowledgements

We are thankful to our institutes who provided their expertise that greatly assisted this research work.

Funding

This research was supported by the National Natural Science Foundation of China (grant numbers 11801542) and the Shenzhen Science and Technology Projects (grant numbers JCYJ20170818164014753 and JCYJ20170818163445670 and JCYJ20180703145002040).

References

Albert FW, Kruglyak L. 2015. The role of regulatory variation in complex traits and disease. *Nat Rev Genet* **16**: 197-212.

Atkins I, Kinnersley B, Ostrom QT, Labreche K, Il'yasova D, Armstrong GN, Eckel-Passow JE, Schoemaker MJ, Nothen MM, Barnholtz-Sloan JS et al. 2019. Transcriptome-wide association study identifies new candidate susceptibility genes for glioma. *Cancer Res* doi:10.1158/0008-5472.CAN-18-2888.

Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, Torstenson ES, Shah KP, Garcia T, Edwards TL et al. 2018. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat Commun* **9**: 1825.

Bindea G, Galon J, Mlecnik B. 2013. CluePedia Cytoscape plugin: pathway insights using integrated experimental and in silico data. *Bioinformatics* **29**: 661-663.

Clark GR, Duncan EL. 2015. The genetics of osteoporosis. *Brit Med Bull* **113**: 73-81.

Claussnitzer M, Dankel SN, Kim K-H, Quon G, Meuleman W, Haugen C, Glunk V, Sousa IS, Beaudry JL, Puviindran V et al. 2015. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *New England Journal of Medicine* **373**: 895-907.

Duncan EL, Brown MA. 2010. Genetic Determinants of Bone Density and Fracture Risk-State of the Art and Future Directions. *J Clin Endocr Metab* **95**: 2576-2587.

Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, Plagnol V. 2014. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *Plos Genetics* **10**.

Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, Jansen R, de Geus EJC, Boomsma DI, Wright FA et al. 2016. Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics* **48**: 245.

Gusev A, Lee SH, Trynka G, Finucane H, Vilhjalmsson BJ, Xu H, Zang C, Ripke S, Bulik-Sullivan B, Stahl E et al. 2014. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am J Hum Genet* **95**: 535-552.

Gusev A, Mancuso N, Won H, Kousi M, Finucane HK, Reshef Y, Song L, Safi A, Schizophrenia Working Group of the Psychiatric Genomics C, McCarroll S et al. 2018. Transcriptome-wide association

study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nat Genet* **50**: 538-548.

Hochreiter S, Clevert D-A, Obermayer K. 2006. A new summarization method for affymetrix probe level data. *Bioinformatics* **22**: 943-949.

Hormozdiari F, van de Bunt M, Segre AV, Li X, Joo JWJ, Bilow M, Sul JH, Sankararaman S, Pasaniuc B, Eskin E. 2016. Colocalization of GWAS and eQTL Signals Detects Target Genes. *American Journal of Human Genetics* **99**: 1245-1260.

Huang YT, Liang LM, Moffatt MF, Cookson WOCM, Lin XH. 2015. iGWAS: Integrative Genome-Wide Association Studies of Genetic and Genomic Data for Disease Susceptibility Using Mediation Analysis. *Genet Epidemiol* **39**: 347-356.

Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira PG et al. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**: 506-511.

Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, Beer MA. 2015. A method to predict the impact of regulatory variants from DNA sequence. *Nat Genet* **47**: 955-961.

Lu Y Beeghly-Fadiel A Wu L Guo X Li B Schildkraut JM Im HK Chen YA Permuth JB Reid BM et al. 2018. A Transcriptome-Wide Association Study Among 97,898 Women to Identify Candidate Susceptibility Genes for Epithelial Ovarian Cancer Risk. *Cancer Res* **78**: 5419-5430.

Mafi Golchin M, Heidari L, Ghaderian SMH, Akhavan-Niaki H. 2016. Osteoporosis: A Silent Disease with Complex Genetic Contribution. *Journal of Genetics and Genomics* **43**: 49-61.

Mancuso N, Shi H, Goddard P, Kichaev G, Gusev A, Pasaniuc B. 2017. Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. *Am J Hum Genet* **100**: 473-487.

Misof BM, Gamsjaeger S, Cohen A, Hofstetter B, Roschger P, Stein E, Nickolas TL, Rogers HF, Dempster D, Zhou H et al. 2012. Bone material properties in premenopausal women with idiopathic osteoporosis. *J Bone Miner Res* **27**: 2551-2561.

Moonesinghe R, Khoury MJ, Liu T, Ioannidis JP. 2008. Required sample size and nonreplicability thresholds for heterogeneous genetic associations. *Proc Natl Acad Sci U S A* **105**: 617-622.

Morris JA, Kemp JP, Youlten SE, Laurent L, Logan JG, Chai RC, Vulpescu NA, Forgetta V, Kleinman A, Mohanty ST et al. 2019. An atlas of genetic influences on osteoporosis in humans and mice. *Nature Genetics* **51**: 258-266.

Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs KV, Li X, Li H, Kuperwasser N, Ruda VM et al. 2010. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**: 714-719.

Nguyen ND, Ahlborg HG, Center JR, Eisman JA, Nguyen TV. 2007. Residual lifetime risk of fractures in women and men. *J Bone Miner Res* **22**: 781-788.

Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C, Barroso I, Dermitzakis ET. 2010. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet* **6**: e1000895.

Nicolae DL, Gamazon E, Zhang W, Duan SW, Dolan ME, Cox NJ. 2010. Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *Plos Genetics* **6**.

Orlic I, Borovecki F, Simic P, Vukicevic S. 2007. Gene expression profiling in bone tissue of osteoporotic mice. *Arh Hig Rada Toksikol* **58**: 3-11.

Peacock M, Econs MJ, Turner CH, Foroud T. 2002. Genetics of Osteoporosis. *Endocrine Reviews* **23**:

303-326.

Peng S, Zhou G, Luk KDK, Cheung KMC, Li Z, Lam WM, Zhou Z, Lu WW. 2009. Strontium promotes osteogenic differentiation of mesenchymal stem cells through the Ras/MAPK signaling pathway. *Cell Physiol Biochem* **23**: 165-174.

Rachner TD, Khosla S, Hofbauer LC. 2011. Osteoporosis: now and the future. *Lancet* **377**: 1276-1287.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**: 2498-2504.

Smemo S, Tena JJ, Kim K-H, Gamazon ER, Sakabe NJ, Gómez-Marín C, Aneas I, Credidio FL, Sobreira DR, Wasserman NF. 2014. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507**: 371.

Smyth GK. 2005. limma: Linear Models for Microarray Data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, doi:10.1007/0-387-29362-0_23 (ed. R Gentleman, et al.), pp. 397-420. Springer New York, New York, NY.

Spain SL, Barrett JC. 2015. Strategies for fine-mapping complex traits. *Human molecular genetics* **24**: R111-R119.

Stein GS, Lian JB, Owen TA. 1990. Relationship of cell growth to the regulation of tissue-specific gene expression during osteoblast differentiation. *FASEB J* **4**: 3111-3123.

Stelzer G, Plaschkes I, Oz-Levi D, Alkelai A, Olender T, Zimmerman S, Twik M, Belinky F, Fishilevich S, Nudel R et al. 2016a. VarElect: the phenotype-based variation prioritizer of the GeneCards Suite. *BMC Genomics* **17 Suppl 2**: 444.

Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Stein TI, Nudel R, Lieder I, Mazor Y et al. 2016b. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current Protocols in Bioinformatics* **54**: 1.30.31-31.30.33.

Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D et al. 2007. Population genomics of human gene expression. *Nat Genet* **39**: 1217-1224.

Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M et al. 2015. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**: e1001779.

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP et al. 2015. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* **43**: D447-452.

Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P et al. 2017. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res* **45**: D362-D368.

Trošt Z, Trebše R, Preželj J, Komadina R, Logar DB, Marc J. 2010. A microarray based identification of osteoporosis-related genes in primary culture of human osteoblasts. *Bone* **46**: 72-80.

Vidal C, Cachia A, Xuereb-Anastasi A. 2009. Effects of a synonymous variant in exon 9 of the CD44 gene on pre-mRNA splicing in a family with osteoporosis. *Bone* **45**: 736-742.

Wanachewin O, Boonmaleerat K, Pothacharoen P, Reutrakul V, Kongtawelert P. 2012. Sesamin stimulates osteoblast differentiation through p38 and ERK1/2 MAPK signaling pathways. *BMC Complement Altern Med* **12**: 71-71.

Westra H-J, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, Christiansen MW, Fairfax BP, Schramm K, Powell JEJNg. 2013. Systematic identification of trans eQTLs as putative drivers

of known disease associations. **45**: 1238.

Wu L Shi W Long J Guo X Michailidou K Beesley J Bolla MK Shu XO Lu Y Cai Q et al. 2018. A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nat Genet* **50**: 968-978.

Wu XB, Li Y, Schneider A, Yu W, Rajendren G, Iqbal J, Yamamoto M, Alam M, Brunet LJ, Blair HC et al. 2003. Impaired osteoblastic differentiation, reduced bone formation, and severe osteoporosis in noggin-overexpressing mice. *J Clin Invest* **112**: 924-934.

Xu X, Jia X, Mo L, Liu C, Zheng L, Yuan Q, Zhou X. 2017. Intestinal microbiota: a potential target for the treatment of postmenopausal osteoporosis. *Bone Research* **5**: 17046.

Zhang X, Joehanes R, Chen BH, Huan T, Ying S, Munson PJ, Johnson AD, Levy D, O'Donnell CJ. 2015. Identification of common genetic variants controlling transcript isoform variation in human whole blood. *Nat Genet* **47**: 345-352.

Zhou X, Carbonetto P, Stephens M. 2013. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet* **9**: e1003264.