Published in final edited form as:

Hum Genet. 2017 August ; 136(8): 963–974. doi:10.1007/s00439-017-1825-4.

Regulatory element-based prediction identifies new susceptibility regulatory variants for osteoporosis

Shi Yao¹, Yan Guo¹, Shan-Shan Dong¹, Ruo-Han Hao¹, Xiao-Feng Chen¹, Yi-Xiao Chen¹, Jia-Bin Chen¹, Qing Tian², Hong-Wen Deng², and Tie-Lin Yang^{1,*}

¹Key Laboratory of Biomedical Information Engineering of Ministry of Education, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi Province, 710049, P. R. China

²School of Public Health and Tropical Medicine, Tulane University, New Orleans, LA, 70112, USA

Abstract

Despite genome-wide association studies (GWASs) have identified many susceptibility genes for osteoporosis, it still leaves a large part of missing heritability to be discovered. Integrating regulatory information and GWASs could offer new insights into the biological link between the susceptibility SNPs and osteoporosis. We generated five machine learning classifiers with osteoporosis-associated variants and regulatory features data. We gained the optimal classifier and predicted genome-wide SNPs to discover susceptibility regulatory variants. We further utilized Genetic Factors for Osteoporosis Consortium (GEFOS) and three in-house GWASs samples to validate the associations for predicted positive SNPs. The random forest classifier performed best among all machine learning methods with the F1 score of 0.8871. Using the optimized model, we predicted 37,584 candidate SNPs for osteoporosis. According to the meta-analysis results, a list of regulatory variants was significantly associated with osteoporosis after multiple testing corrections and contributed to the expression of known osteoporosis-associated protein-coding genes. In summary, combining GWASs and regulatory elements through machine learning could provide additional information for understanding the mechanism of osteoporosis. The regulatory variants we predicted will provide novel targets for etiology research and treatment of osteoporosis.

Keywords

Osteoporosis; Machine learning; GWASs; Regulatory features data; Random forest

*Corresponding author: Tie-Lin Yang, Ph.D., PHONE: 86-29-82668463 yangtielin@mail.xjtu.edu.cn.

Compliance with Ethical Standards

Conflict of Interest

The authors declare that they have no conflict of interest.

Research involving Human Participants and/or Animals

For this type of study formal consent is not required.

This article does not contain any studies with animals performed by any of the authors.

Informed consent

Informed consent was obtained from all individual participants included in the study.

Introduction

Osteoporosis is a common disease characterized by decreased bone mass and increased tendency of fragility fractures (Guo et al. 2010). Bone mineral density (BMD), which is the most widely used predictor for osteoporosis, is highly heritable with heritability estimates between 0.5–0.85 (Ralston and Uitterlinden 2010). Genome-wide association studies (GWASs) is an effective strategy to identify genetic variants for human complex diseases/traits. Using this strategy, hundreds of genetic variants have been identified to be associated with BMD (Estrada et al. 2012; Yang et al. 2008). However, these susceptibility loci together can only explain a relatively small fraction of the estimated heritability for osteoporosis (Estrada et al. 2012), leading many to question where the ‘missing’ heritability is. True association signals with modest genetic effect size may be missed due to the strict significance thresholds (Yang et al. 2010) of GWASs. In addition, epigenetic inheritance (Slatkin 2009) was not considered in GWASs, which is another possible cause of missing heritability. For better prevention, diagnosis and treatment of osteoporosis, it is necessary to detect the underlying genetic variants with new methods.

The vast majority of reported trait/disease-associated SNPs more than expected are located in non-coding regions (Hindorff et al. 2009) and many lay far away from the nearest protein-coding genes. Due to lack of functional verification, little is known about the mechanisms of non-coding genetic variants, besides simply being potential markers. Actually, most of the GWASs SNPs harbor a rich array of regulatory elements that have diverse gene regulatory and other functions (ENCODE Project Consortium 2012), which implies that these variants associated with disease may be important for gene regulation rather than biochemical function (Musunuru et al. 2010). Using epigenomic analyses for promoters of known osteoporosis susceptibility genes, we have recently predicted candidate genes and identified *BDNF* as a new susceptibility gene for osteoporosis (Guo et al. 2016), which reminds us that regulatory features data have the potential to predict disease-associated variants when integrating with GWASs. However, we only focused on promoter regions, which might neglect potentially meaningful regulatory information of other genomic regions. Integrating all SNPs and regulatory data produces a new set of challenges of excessive computational demands with the large number of SNPs and regulatory elements.

Machine learning is a subfield of computer science that gives computers the ability to process big data without being explicitly programmed (L. 1959). Machine learning is also a powerful tool for interpreting large genomic data sets and it has been applied in numerous areas in genetics and genomics (Libbrecht and Noble 2015). For example, it has been used to prioritize the noncoding variants with various genomic and epigenomic annotations (Ritchie et al. 2014) as well as measure of functional importance with an unsupervised approach (Ionita-Laza et al. 2016). Recent studies used support vector machine (Kircher et al. 2014) and a deep learning-based algorithmic framework (Zhou and Troyanskaya 2015) to interpret genetic variation with genomic annotations, estimating the relative pathogenicity of human genetic variants. However, these studies didn’t specify a particular disease, it reminds us that machine learning has the potential to acquire the characteristics of osteoporosis-associated SNPs with regulatory information.

In this study, we integrated GWASs data and regulatory features with machine learning methods for predicting osteoporosis-associated variants. We hypothesized that an optimal algorithm could predict susceptibility SNPs with disease specific regulatory elements. To achieve this goal, we first annotated positive and negative SNP sets with all available regulatory elements. Next, we trained multiple predictors with annotated information and gained the optimal classifier according to the performance of models. At last, we predicted genome-wide SNPs with the optimal classifier and validated the predicted positive SNPs in available GWASs datasets. Our results demonstrate that regulatory elements are effective to depict the regularities of genetic distribution of disease-related variants and be the predictor of osteoporosis-associated variants, which can improve the power of detecting associations and provide valuable insights into genetic architecture of osteoporosis.

Results

Acquisition of osteoporosis-associated labeled SNPs set

We obtained a total of 172 osteoporosis-associated index SNPs in white populations (Online Resource Table S1) from the GWAS Catalog, PheGenI database and two GWASs reported recently. We used a highly stringent LD threshold of $r^2 = 1$ to supplement the positive SNPs set, and the total number of the positive SNPs set was extended to 1,898. We found only 1.37% labeled positive SNPs mapped to coding exon (Online Resource Fig S1a). Most of SNPs located in non-coding regions, of which 63.91% mapped to intron, 17.91% mapped to intergenic region, 14.91% mapped to non-coding RNA, 1.05% mapped to UTR, 0.42% mapped to upstream and 0.42% mapped to downstream (Online Resource Fig S1a). To build a binary supervised classification, we selected negative SNPs set using 20 negative SNPs per positive SNP as shown in Materials and Methods.

Regulatory elements are highly effective for predicting osteoporosis-associated SNPs

To model the interaction status between labeled SNPs and regulatory elements, we built a complete machine learning analysis procedure (Fig 1). The input data are labeled SNPs and regulatory elements annotation. In this study, we focused on 55 osteoporosis-associated cell lines (Online Resource Table S2). The full list of ULRs for annotation elements we used for training models was provided in Online Resource Table S3. The output is a model for predicting whether an unlabeled SNP is associated with osteoporosis susceptibility. We found that the labeled positive SNPs were potentially functional, since all the labeled positive were covered by at least one histone modification in at least one cell line and 76.66% (1,455) were covered by at least one annotation among DHSs, TFBSs, eQTLs in whole blood and evolutionary conserved annotation (Online Resource Fig S1b).

Before inputting the labeled SNPs and annotation element sets to generate model, we need to remove highly correlated features. Although regulatory elements were annotated from different databases and cell lines, they might be highly correlated with each other thus redundant to the model. For example, trimethylation of histone H3 at lysine 36 (H3K36me3) in 13 cell lines was highly correlated (Fig 2). We calculated the correlation of all elements and finally removed 70 highly correlated regulatory elements to avoid redundant information.

To get the best classifier of osteoporosis-associated SNPs, we modeled five common algorithms and evaluated the performance with multiple metrics after removing highly correlated elements. The algorithms mined numerous of regulatory elements to build models. To test whether such a large elements set is needed, we utilized feature selection. Based on the rankings of elements in each algorithm, we built multiple feature subsets with the number of the features increasing from 1 to the maximum (762). We then used appropriate number of features of each model to generate the best algorithm (Table 1). Five models were all evaluated with sensitivity, specificity, precision, accuracy and F1 score. We found that random forest with 40 features performed best among all algorithms with the F1 score of 0.8871 (Table 1, Fig 3a). In contrast, when we performed random forest model with random assignments of labeled annotation as positive or negative, the accuracy and F1 score of the model dropped to ~0.5. Thus, we chose random forest algorithm with the top 40 important features as our final model (Fig 3b).

No single regulatory element distinguishes true osteoporosis-associated variants

The regulatory elements had a complex combination between positive set and negative set. Any given regulatory element can't distinguish true osteoporosis-associated SNPs individually (Fig 4). For example, all labeled positive SNPs and negative SNPs in selected region were covered by H3k27me3 in Monocd14, Gm12878, HSMM and Ezh2_ (39875) in HSMM. No single feature distinguished true osteoporosis-associated variants. This complex condition proved the necessity of the utilization of machine learning.

Osteoporosis-associated SNPs prediction using regulatory elements is robust

Using the final random forest model, we predicted genome-wide common SNPs in white populations and obtained 37,584 candidate osteoporosis-associated loci. Almost 95% of the predicted positive SNPs were in weak LD with labeled positive SNPs in white populations ($r^2 < 0.1$, Fig 5a), indicating that the predicted positive SNPs were independent from the labeled positive SNPs. Similar to the labeled positive SNPs, majority of these predicted positive SNPs were located in non-coding regions, only 0.95% mapped to coding exon (Fig 5b), indicating that the candidate regulatory variants we predicted may be important for gene regulation. Among all predicted SNPs, 99.93% (37,556) were covered by at least one histone modification in at least one cell line and 72.44% (27,224) were covered by at least one annotation among DHSs, TFBSs, eQTLs in whole blood and evolutionary conserved annotation (Fig 5c), suggesting that the predicted positive SNPs were potentially functional. To investigate whether the eQTL target genes of predicted positive SNPs are related to osteoporosis, we performed pathway enrichment analysis. These genes were significantly enriched in eleven pathways in KEGG after multiple testing corrections, including lysosome, viral myocarditis and metabolic pathways (Fig 5d).

Since our ultimate goal is to predict osteoporosis-associated SNPs with regulatory elements, we further tested whether predicted positive SNPs were associated with osteoporosis and may provide novel targets. We checked the predicted positive SNPs for their association signals with meta-analysis results of four GWASs and found 369 SNPs significantly associated with femoral neck or/and spine BMD after multiple testing corrections (P -value $< 1.33 \times 10^{-6}$, Online Resource Table S4). These SNPs located in 41 genes, of which 38 were

overlapped with the labeled positive genes. As for the intergenic SNPs, we assigned the SNP to the closest gene according to their physical position. We identified three new genes that may be associated with osteoporosis, which are *ESPL1*, *AMT* and *GAL* (Table 2). Most of the 369 SNPs assigned to known osteoporosis-associated genes, indicating that our prediction is trustworthy. On the other hand, we found 93 SNPs of the 369 SNPs were annotated as enhancer in at least one osteoporosis-associated cell line according to the HMM annotation (Online Resource Table S5). It emphasized the hypothesis that disease-associated variants may be important for gene regulation rather than biochemical function. These SNPs contributed to the expression of 28 protein-coding genes, including *AMT*, *GAL* and *SPPI* (Online Resource Table S6).

Discussion

Although GWASs have detected hundreds of variants associated with osteoporosis, they can only explain a small fraction of the estimated heritability. Here, we illustrated the feasibility of machine learning to find missing heritability using regulatory elements. In addition, functional annotation, pathway enrichment analysis and GWASs validation suggested that the positive SNPs we predicted can be candidate variants for osteoporosis. We identified three novel susceptibility genes, including *AMT*, *ESPL1* and *GAL*, and a list of susceptibility regulatory variants for osteoporosis.

The variable importance evaluation function of our final optimized model can assess the biological relevance of regulatory elements. The most useful element in our final model was eQTL annotation, indicating that whether the genomic loci contributing to the gene expression variation is vital for predicting osteoporosis-associated loci. As for histone modification marks, H3k27me3 in 5 cell lines were selected in our final model. H3k27me3 was also reported being a negative signature of Wnt signaling (Wang et al. 2010), which is a crucial pathway for bone biology and development. A previous study has reported that EZH2 contributes to the development of osteoporosis through shifted mesenchymal stem cells (MSC) lineage commitment to adipocyte (Jing et al. 2016). EZH2_ (39875) histone mark in two cell lines (HSM and CD20) was important in our optimized model.

For the enriched pathways through genes identified by predicted positive SNPs, multiple pathways have already been reported to be associated with osteoporosis. For example, lysosome (Hsa04142) pathway has been reported to be related to the degradation of *OPG* (osteoprotegerin), a famous gene in the pathogenesis of osteoporosis (Tat et al. 2006). Other pathways, such as metabolic pathways (Hsa01100) (Byers and Pyott 2012), graft-versus-host disease pathway (Hsa05332) (Mitchell et al. 2010), type I diabetes mellitus (Hsa04940) (Khan and Fraser 2015) as well as cysteine and methionine metabolism (Hsa00270) (Sellmeyer et al. 2001) also have been reported to be associated with osteoporosis. As for viral myocarditis (Hsa05416) and herpes simplex infection (Hsa05168), the relationships between them and osteoporosis are unknown. These findings suggested the fact that osteoporosis could result from multiple diseases, showing that our model for predicting osteoporosis-associated variants is robust.

We overlapped the predicted positive SNPs with SNPs reported in NHGRI GWAS Catalog (Online Resource Table S7). The results showed that rs2273061 of the *JAG1* has already been reported to be associated with BMD in populations including Asian and European individuals (Kung et al. 2010). And rs7797976 within *CPED1* have been reported to be significantly associated with pediatric areal bone mineral density (Chesi et al. 2015) in European ancestry which was not included in our labeled positive set since it was reported as a susceptibility locus specific to pediatric trait. These results demonstrate the effectiveness of the algorithm we modeled for predicting new susceptibility loci for osteoporosis. We also observed plenty of SNPs associated with other diseases when overlapping the results with SNPs that have been reported in NHGRI GWAS Catalog, which including diabetes, blood metabolite levels, Parkinson's disease, height, etc. It consistent with the fact that phenotypic outcome of osteoporosis could be caused by multiple diseases. Diabetes mellitus (Hofbauer et al. 2007) and Parkinson's disease (Fink et al. 2005) have been previously published to contribute to lower BMD, and both body mass index (BMI) and height were associated with fracture at specific site (Compston et al. 2014).

We suggested a list of candidate loci for osteoporosis, among which, three novel susceptibility genes (*AMT*, *ESPL1* and *GAL*) were confirmed for association with spine BMD in meta-analysis of four GWAS datasets. *AMT* encodes one of four critical components of the glycine cleavage system, which provides a bypass reaction in glycine metabolism. Glycine was found to enhance the activity of osteoblast and the formation of bone matrix, and insufficient supply of glycine can cause a number of health problems, such as arthrosis and osteoporosis (Kim et al. 2016). Genetic mutations in *AMT* were associated with glycine encephalopathy (Coughlin et al. 2016), and the patients with glycine encephalopathy usually suffered skeletal system problems, including osteoporosis. Another gene also has potential connection with osteoporosis. *ESPL1* encodes separase in human. Although we haven't found direct biological evidence for *ESPL1* involved in bone, two SNPs (rs1318648 and rs17125266) in *ESPL1* showed correlation with *SP7* (Ham and Roh 2014), a transcription factor responsible for regulating osteoblast differentiation. *SP7* is one of the most famous osteoporosis-associated genes, and is the master gene for skeletal development along with *RUNX2* (Timpson et al. 2009), suggesting that *ESPL1* may be related with osteoporosis through interacting with *SP7*. *GAL* encodes a neuroendocrine peptide, which proteolytically processed to generate galanin and galanin message-associated peptide (GMAP). Galanin has been reported to inhibit formation of IL-1 β and TNF- α , both associated with osteoclast formation and osteoporosis, and has been reported to increase bone thickness and osteoblast number in mouse (McDonald et al. 2007). Together, our findings highlighted the important role of three genes to the pathogenesis of osteoporosis and the credibility of our predicted positive SNPs.

We also found 93 novel osteoporosis-associated SNPs were annotated as enhancer (Online Resource Table S5) and contributed to the expression of 28 protein-coding genes, including *AMT*, *GAL* and *SPPI* (Online Resource Table S6). *AMT* and *GAL* were the two of the three genes that we discovered as novel susceptibility genes in this project. The association between these two genes and osteoporosis have been discussed above. As for *SPPI*, the protein encoded by this gene is a natural ligand for the vitronectin receptor and is involved in

mineralizing bone matrix through the attachment of osteoclasts (Reinholt et al. 1990). These findings emphasized the regulatory function of our prediction.

One of the greatest challenges facing genetics of human complex diseases/traits is the reinterpretation of GWASs data. Despite the huge computational burden, integrating GWASs and regulatory features data is an effective approach to identify regulatory variants for complex diseases. Machine learning is a powerful tool for interpreting large GWASs data and regulatory elements, and the random forest classifier performed best with the F1 score of 0.8871 in our project, indicating the reliability of our hypothesis.

Limitations of our study also need to be addressed. For example, we only focused on SNPs that associated with osteoporosis in white populations, but the regulatory mechanisms may be different among races. This limitation is due to the fact that most reported GWASs in osteoporosis were performed in the whites. Secondly, the features we used for our classifier included many regulatory elements in monocytes and lymphocytes (Online Resource Table S2). Although both monocytes and lymphocytes are involved in the osteogenesis, they are not the direct target cell lines for osteoporosis. Additionally, as our current knowledge about the regulatory mechanism is still incomplete, the prediction results would be greatly improved by using additional informative features.

In conclusion, we integrated comprehensive regulatory data and GWASs to build a machine learning algorithm and we identified 37,584 susceptibility regulatory variants for osteoporosis. We also discovered *AMT*, *ESPL1* and *GAL* as susceptibility genes for osteoporosis through meta-analysis results of four GWASs. Our findings demonstrated that combining GWASs and regulatory elements through machine learning could provide additional information for understanding the mechanism of osteoporosis. The regulatory variants we predicted may provide novel targets for etiology research and treatment of osteoporosis.

Materials and Methods

Acquisition of osteoporosis-associated labeled SNPs set

In this study, we first selected a total of 8,550,209 common SNPs with minor allele frequency (MAF) > 0.01 as the whole SNP set from European samples in 1000 Genome Project (<http://www.1000genomes.org/>). The SNP set was further divided into labeled SNP set and unlabeled SNP set (prediction set), of which labeled SNP set was constituted by positive set and negative set.

For the positive set, osteoporosis-associated index SNPs were downloaded from NHGRI GWAS Catalog (Welter et al. 2014), PheGenI database (Ramos et al. 2014) and two GWASs (Styrkarsdottir et al. 2016; Zheng et al. 2015) reported recently with P -value less than 5×10^{-8} in white populations in the discovery stage, using osteoporosis related phenotypes including bone mineral density (BMD), bone density, bone properties, osteoporosis and femoral neck bone geometry. Moreover, susceptibility loci that specific to pediatric trait were not included in the positive SNP set. However, GWASs was confined by low genomic coverage provided by genotyping microarrays (Grant and Hakonarson 2008) which may

overlook the genuine risk-associated SNPs that in LD (linkage disequilibrium) with causal variants (McClellan and King 2010). Therefore, we generated all SNPs that were in strong LD ($r^2 = 1$) with each index SNP by PLINK (Purcell et al. 2007) using data from 1000 Genomes as the complement of the positive SNPs set.

We created several sets of negative SNPs with different distances to positive SNPs. For each positive SNP, the maximum distances allowed in each group were 40kb, 200kb, 1,000kb and 5,000kb respectively, and the MAF difference between the positive SNP and negative SNP was less than 0.05. Moreover, to get osteoporosis-independent genetic loci, we filtered out positive SNPs and SNPs in weak LD ($r^2 > 0.1$) with positive SNPs. Next, we selected the full set of the closest SNPs set, 60%, 40% and 20% random subsets of other negative sets with different distances to positive SNPs as the whole negative set. At last, we used 20 negative SNPs per positive SNP to build a binary supervised classification.

Regulatory elements annotation

The elements annotation were computed through BEDTools (Quinlan and Hall 2010) from Encyclopedia of DNA Elements (ENCODE) project (ENCODE Project Consortium 2012), Roadmap Epigenomics Mapping Consortium (REMC) (Bernstein et al. 2010), expression quantitative trait loci (eQTLs) data and evolutionary information-derived annotation. Cell lines related with osteoporosis were included in the annotation (Online Resource Table S2). We also included blood cell lines from monocytes and lymphocytes, considering monocytes act as precursors of osteoclasts (Udagawa et al. 1990) and lymphocytes have been reported to be involved in the osteogenesis (Kotake et al. 2001; Manilay and Zouali 2014). For each element, a SNP was annotated as 1 if the SNP overlaps with the element and 0 otherwise. The GRCh37/hg19 genome assembly was used for all analyses in our study. Detailed information of the regulatory data is described as follows:

ENCODE—The annotation features from ENCODE can be classified into three groups of regulatory elements, including transcription factor binding sites (TFBSs), histone modifications and DNase I hypersensitive sites (DHSs). A total of 225 regulatory elements were used in the analysis.

REMC—The annotation features from REMC can be classified into three groups of regulatory elements, including TFBSs, histone modifications and chromatin state segmentation by hidden Markov model (HMM). A total of 605 regulatory elements were used in the analysis.

Other elements—We also added cis-eQTLs data of whole blood from GTEx Portal (GTEx Consortium 2015) with a *P*-value cutoff of 0.01 on expression level and evolutionary conservation feature (Davydov et al. 2010).

Model generation and evaluation

High correlation inherent elements were randomly removed with the absolute value of threshold of correlation > 0.7 before generating models. The correlation matrix was implemented by using the corrpilot package (from Taiyun Wei, Viliam Simko) in R. The

labeled SNPs with regulatory features annotation were further separated into training and testing sets randomly. The training SNPs contained 80% of the total labeled SNPs. In this study, five familiar algorithms were used to train the model, including random forest (rf), single decision tree (C5.0), k-nearest neighbors (knn), soft independent modeling of class analogy (CSimca) and support vector machines with radial basis function kernel (svmRadial).

Models were evaluated using five-fold cross-validation. In this study, osteoporosis-associated SNPs were treated as the positive class. True positive (TP) and true negative (TN) were the number of SNPs that have been correctly classified. Whereas, false positive (FP) represented the number of negative SNPs that were predicted as positive, and false negative (FN), on the other hand, was the number of positive SNPs that were predicted as negative. Performance was measured with sensitivity, specificity, precision, accuracy and F1 score. All simulations and analyses were executed by using the caret package (from Max Kuhn) in R.

Model optimization and prediction

After removing highly correlated elements, feature selection was also used to optimize models to improve the performance of prediction and provide more cost-effective elements. Firstly, using the importance evaluation function in caret package, we obtained model-based importance rankings for all elements. Then, we built multiple feature subsets with the number of the features increasing from 1 to the maximum. At last, we determined the appropriate number of features and used the model corresponding to the optimal subsets with the highest prediction performance to predict new susceptibility SNPs using the unlabeled SNPs.

Characteristic of regulatory elements in final model

We showed the distributions of the top ten important elements in a random region of chromosome 2 to test whether a smaller subset was still effective for separating SNPs according to their labels. This procedure was executed using the Gviz package (from Florian Hahne) in R.

Functional validation of predicted positive SNPs

Independence of predicted positive SNPs—To check whether the predicted positive SNPs were independent from the labeled positive SNPs, we first calculated the LD of each predicted positive SNP with labeled positive SNPs.

Gene annotation—We assigned the predicted positive SNPs to genes according to the physical position via ANNOVAR (Yang and Wang 2015). We used RefSeq genes in this study.

Functional annotation—We then checked the annotation of predicted positive SNPs using all regulatory annotation results, including histone modification, TFBSs, DHSs, eQTLs and evolutionary annotation.

Gene set enrichment analysis—We conducted gene set enrichment analysis to evaluate whether the predicted positive SNPs associated with osteoporosis at the pathway level. The gene list, which predicted positive SNPs contribute to its expression level in whole blood, was supplied to pathway enrichment analysis. KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway enrichment analysis was performed by using DAVID (<https://david.ncifcrf.gov/>).

Validation in GWAS datasets

To further validate the predicted positive SNPs at the population level, we took advantage of NHGRI GWAS Catalog and other four GWAS datasets, including GEFOS (Genetic Factors for Osteoporosis Consortium) dataset, and three GWAS samples from in-house studies. We meta-analyzed the four GWAS datasets by using METAL (Willer et al. 2010) and checked the predicted positive SNPs for their association signals with the results. We gained the gene of SNPs that reached the significant threshold after multiple testing corrections by using ANNOVAR (Yang and Wang 2015). The related information of GWAS datasets is described in detail as follows.

NHGRI GWAS Catalog—The predicted positive SNPs were overlapped with SNPs reported in NHGRI GWAS Catalog, a curated collection of all published GWASs, to check whether predicted positive SNPs associated with other diseases/traits.

GEFOS dataset—GEFOS is a large international collaboration comprising various prominent research groups in the bone field, including 17 GWASs and 32,965 samples of European and East Asian ancestry (Estrada et al. 2012).

Three in-house GWAS samples—Our GWAS datasets include three BMD samples, which are Kansas City Osteoporosis Study (KCOS) with 2,286 unrelated European individuals, Omaha Osteoporosis Study (OOS) with 1,000 unrelated European individuals and Chinese Osteoporosis Study (COS) with 1,627 unrelated Han Chinese individuals. The detailed description of the three samples have been reported in our previous studies (Guo et al. 2010; Xiong et al. 2009; Yang et al. 2012). BMD (g/cm^2) at spine and femoral neck (FN) was measured in all samples with dual energy x-ray absorptiometry (DXA). Raw BMD values were adjusted by sex, age, weight and height. The KCOS and COS samples were genotyped with Genome-Wide Human SNP Array 6.0, while the OOS samples were genotyped with Affymetrix Human Mapping 500K array set, according to the manufacturer's protocols. Genotype data were cleaned by applying minimum call rates (95%) and maximum individual missingness (5%). SNPs deviating from Hardy-Weinberg equilibrium ($P < 0.0001$) were also excluded. These data were imputed using IMPUTE2 (Howie et al. 2011) with 1000 Genome Project phase 3 as the reference panel.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (31471188, 81573241, and 31511140285); China Postdoctoral Science Foundation (2016M602797, 2016T90902); Natural Science Basic Research Program Shaanxi Province (2016JQ3026); and the Fundamental Research Funds for the Central Universities. The study was also funded by the grants from National Institutes of Health (P50AR055081, R01AG026564, R01AR050496, and R01AR057049).

References

- Bernstein BE, et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology*. 2010; 28:1045–1048. DOI: 10.1038/nbt1010-1045
- Byers PH, Pyott SM. Recessively inherited forms of osteogenesis imperfecta. *Annual review of genetics*. 2012; 46:475–497. DOI: 10.1146/annurev-genet-110711-155608
- Chesi A, et al. A trans-ethnic genome-wide association study identifies gender-specific loci influencing pediatric aBMD and BMC at the distal radius. *Human molecular genetics*. 2015; 24:5053–5059. DOI: 10.1093/hmg/ddv210 [PubMed: 26041818]
- Compston JE, et al. Relationship of weight, height, and body mass index with fracture risk at different sites in postmenopausal women: the Global Longitudinal study of Osteoporosis in Women (GLOW). *Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research*. 2014; 29:487–493. DOI: 10.1002/jbmr.2051
- Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012; 489:57–74. DOI: 10.1038/nature11247 [PubMed: 22955616]
- Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*. 2015; 348:648–660. DOI: 10.1126/science.1262110 [PubMed: 25954001]
- Coughlin C, et al. The Genotypic Spectrum of Classic Nonketotic. Hyperglycinemia Due to Mutations in Glc6 and Amt *Molecular genetics and metabolism*. 2016; 117:236–236.
- Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++ *PLoS computational biology*. 2010; 6:e1001025.doi: 10.1371/journal.pcbi.1001025 [PubMed: 21152010]
- Estrada K, et al. Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nature genetics*. 2012; 44:491–501. DOI: 10.1038/ng.2249 [PubMed: 22504420]
- Fink HA, Kuskowski MA, Orwoll ES, Cauley JA, Ensrud KE, Osteoporotic Fractures in Men Study G. Association between Parkinson's disease and low bone density and falls in older men: the osteoporotic fractures in men study. *Journal of the American Geriatrics Society*. 2005; 53:1559–1564. DOI: 10.1111/j.1532-5415.2005.53464.x [PubMed: 16137287]
- Grant SF, Hakonarson H. Microarray technology and applications in the arena of genome-wide association. *Clinical chemistry*. 2008; 54:1116–1124. DOI: 10.1373/clinchem.2008.105395 [PubMed: 18499899]
- Guo Y, et al. Integrating Epigenomic Elements and GWASs Identifies BDNF Gene Affecting Bone. Mineral Density and Osteoporotic Fracture Risk *Scientific reports*. 2016; 6:30558.doi: 10.1038/srep30558 [PubMed: 27465306]
- Guo Y, et al. Genome-wide association study identifies ALDH7A1 as a novel susceptibility gene for osteoporosis. *PLoS genetics*. 2010; 6:e1000806.doi: 10.1371/journal.pgen.1000806 [PubMed: 20072603]
- Ham S, Roh TY. A Follow-up Association Study of Genetic Variants for Bone. Mineral Density in a Korean Population *Genomics & informatics*. 2014; 12:114–120. DOI: 10.5808/GI.2014.12.3.114
- Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*. 2009; 106:9362–9367. DOI: 10.1073/pnas.0903103106 [PubMed: 19474294]

- Hofbauer LC, Brueck CC, Singh SK, Dobnig H. Osteoporosis in patients with diabetes mellitus. *Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research*. 2007; 22:1317–1328. DOI: 10.1359/jbmr.070510
- Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes G3. 2011; 1:457–470. DOI: 10.1534/g3.111.001198 [PubMed: 22384356]
- Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature genetics*. 2016; 48:214–220. DOI: 10.1038/ng.3477 [PubMed: 26727659]
- Jing H, et al. Suppression of EZH2 Prevents the Shift of Osteoporotic MSC Fate to Adipocyte and Enhances Bone. *Formation During Osteoporosis Molecular therapy : the journal of the American Society of Gene Therapy*. 2016; 24:217–229. DOI: 10.1038/mt.2015.152 [PubMed: 26307668]
- Khan TS, Fraser LA. Type 1 diabetes and osteoporosis: from molecular pathways to bone phenotype. *Journal of osteoporosis*. 2015; 2015:174186.doi: 10.1155/2015/174186 [PubMed: 25874154]
- Kim MH, Kim HM, Jeong HJ. Estrogen-like osteoprotective effects of glycine in in vitro and in vivo models of menopause. *Amino acids*. 2016; 48:791–800. DOI: 10.1007/s00726-015-2127-6 [PubMed: 26563333]
- Kircher M, Witten DM, Jain P, O’Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*. 2014; 46:310–315. DOI: 10.1038/ng.2892 [PubMed: 24487276]
- Kotake S, et al. Activated human T cells directly induce osteoclastogenesis from human monocytes - Possible role of T cells in bone destruction in rheumatoid arthritis patients. *Arthritis Rheum*. 2001; 44:1003–1012. DOI: 10.1002/1529-0131(200105)44:5<1003::Aid-Anr179>3.0.Co;2-# [PubMed: 11352231]
- Kung AW, et al. Association of JAG1 with bone mineral density and osteoporotic fractures: a genome-wide association study and follow-up replication studies. *American journal of human genetics*. 2010; 86:229–239. DOI: 10.1016/j.ajhg.2009.12.014 [PubMed: 20096396]
- L SA. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*. 1959; 3:210–229.
- Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nature reviews Genetics*. 2015; 16:321–332. DOI: 10.1038/nrg3920
- Manilay JO, Zouali M. Tight relationships between B lymphocytes and the skeletal system. *Trends Mol Med*. 2014; 20:405–412. DOI: 10.1016/j.molmed.2014.03.003 [PubMed: 24726716]
- McClellan J, King MC. Genetic heterogeneity in human disease. *Cell*. 2010; 141:210–217. DOI: 10.1016/j.cell.2010.03.032 [PubMed: 20403315]
- McDonald AC, Schuijers JA, Gundlach AL, Grills BL. Galanin treatment offsets the inhibition of bone formation and downregulates the increase in mouse calvarial expression of TNFalpha and GalR2 mRNA induced by chronic daily injections of an injurious vehicle. *Bone*. 2007; 40:895–903. DOI: 10.1016/j.bone.2006.10.018 [PubMed: 17157570]
- Mitchell SA, et al. Determinants of functional performance in long-term survivors of allogeneic hematopoietic stem cell transplantation with chronic graft-versus-host disease (cGVHD). *Bone marrow transplantation*. 2010; 45:762–769. DOI: 10.1038/bmt.2009.238 [PubMed: 19784078]
- Musunuru K, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*. 2010; 466:714–719. DOI: 10.1038/nature09266 [PubMed: 20686566]
- Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*. 2007; 81:559–575. DOI: 10.1086/519795 [PubMed: 17701901]
- Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010; 26:841–842. DOI: 10.1093/bioinformatics/btq033 [PubMed: 20110278]
- Ralston SH, Uitterlinden AG. Genetics of osteoporosis. *Endocrine reviews*. 2010; 31:629–662. DOI: 10.1210/er.2009-0044 [PubMed: 20431112]
- Ramos EM, et al. Phenotype-Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. *European journal of human genetics : EJHG*. 2014; 22:144–147. DOI: 10.1038/ejhg.2013.96 [PubMed: 23695286]

- Reinholt FP, Hultenby K, Oldberg A, Heinegard D. Osteopontin - a Possible Anchor of Osteoclasts to Bone. *Proceedings of the National Academy of Sciences of the United States of America*. 1990; 87:4473–4475. DOI: 10.1073/pnas.87.12.4473 [PubMed: 1693772]
- Ritchie GRS, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. *Nature methods*. 2014; 11:294–U351. DOI: 10.1038/nmeth.2832 [PubMed: 24487584]
- Sellmeyer DE, Stone KL, Sebastian A, Cummings SR. A high ratio of dietary animal to vegetable protein increases the rate of bone loss and the risk of fracture in postmenopausal women Study of Osteoporotic Fractures Research Group. *The American journal of clinical nutrition*. 2001; 73:118–122. [PubMed: 11124760]
- Slatkin M. Epigenetic inheritance and the missing heritability problem. *Genetics*. 2009; 182:845–850. DOI: 10.1534/genetics.109.102798 [PubMed: 19416939]
- Styrkarsdottir U, et al. Sequence variants in the *PTCH1* gene associate with spine bone mineral density and osteoporotic fractures. *Nature communications*. 2016; 7:10129.doi: 10.1038/ncomms10129
- Tat SK, Padrines M, Theoleyre S, Couillaud-Battaglia S, Heymann D, Redini F, Fortun Y. OPG/membranous-RANKL complex is internalized via the clathrin pathway before a lysosomal and a proteasomal degradation. *Bone*. 2006; 39:706–715. DOI: 10.1016/j.bone.2006.03.016 [PubMed: 16750945]
- Timpson NJ, et al. Common variants in the region around *Osterix* are associated with bone mineral density and growth in childhood. *Human molecular genetics*. 2009; 18:1510–1517. DOI: 10.1093/hmg/ddp052 [PubMed: 19181680]
- Udagawa N, et al. Origin of Osteoclasts - Mature Monocytes and Macrophages Are Capable of Differentiating into Osteoclasts under a Suitable Microenvironment Prepared by Bone. Marrow-Derived Stromal Cells *P Natl Acad Sci USA*. 1990; 87:7260–7264. DOI: 10.1073/pnas.87.18.7260
- Wang L, Jin Q, Lee JE, Su IH, Ge K. Histone H3K27 methyltransferase *Ezh2* represses *Wnt* genes to facilitate adipogenesis. *Proceedings of the National Academy of Sciences of the United States of America*. 2010; 107:7317–7322. DOI: 10.1073/pnas.1000031107 [PubMed: 20368440]
- Welter D, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research*. 2014; 42:D1001–1006. DOI: 10.1093/nar/gkt1229 [PubMed: 24316577]
- Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*. 2010; 26:2190–2191. DOI: 10.1093/bioinformatics/btq340 [PubMed: 20616382]
- Xiong DH, et al. Genome-wide association and follow-up replication studies identified *ADAMTS18* and *TGFBR3* as bone mass candidate genes in different ethnic groups. *American journal of human genetics*. 2009; 84:388–398. DOI: 10.1016/j.ajhg.2009.01.025 [PubMed: 19249006]
- Yang H, Wang K. Genomic variant annotation and prioritization with ANNOVAR and WANNOVAR. *Nature protocols*. 2015; 10:1556–1566. DOI: 10.1038/nprot.2015.105 [PubMed: 26379229]
- Yang J, et al. Common SNPs explain a large proportion of the heritability for human height. *Nature genetics*. 2010; 42:565–569. DOI: 10.1038/ng.608 [PubMed: 20562875]
- Yang TL, et al. Genome-wide copy-number-variation study identified a susceptibility gene, *UGT2B17*, for osteoporosis. *American journal of human genetics*. 2008; 83:663–674. DOI: 10.1016/j.ajhg.2008.10.006 [PubMed: 18992858]
- Yang TL, et al. Genetic variants in the *SOX6* gene are associated with bone mineral density in both Caucasian and Chinese populations. *Osteoporosis international : a journal established as result of cooperation between the European Foundation for Osteoporosis and the National Osteoporosis Foundation of the USA*. 2012; 23:781–787. DOI: 10.1007/s00198-011-1626-x
- Zheng HF, et al. Whole-genome sequencing identifies *EN1* as a determinant of bone density and fracture. *Nature*. 2015; 526:112–117. DOI: 10.1038/nature14878 [PubMed: 26367794]
- Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature methods*. 2015; 12:931–934. DOI: 10.1038/nmeth.3547 [PubMed: 26301843]

Fig 1. Schematic diagram of the model generation

The input data of the model generation were labeled SNPs and regulatory elements annotation. The output was a model for predicting whether a SNP (unlabeled) with the same regulatory elements was associated with osteoporosis susceptibility. Positive SNPs were obtained from NHGRI GWAS Catalog, PheGenI database and two GWASs reported lately. Labeled SNPs which consisted of positive SNPs and negative SNPs were divided into training sets and testing sets randomly with the ratio of 4:1. Regulatory elements included five parts, including histone modification, DHSs, TFBSs, eQTLs and evolutionary conserved annotation. The 'streamline features' part focused on removing the high correlated features. The 'optimize model' part mainly focused on feature selection.

Fig 2. Correlation of all selected regulatory elements

The correlation matrix included all 832 regulatory elements, of which 70 high correlated elements with the absolute value of threshold of 0.7 were removed before training models to avoid redundant information. The amplification region in the correlation was H3K36me3 in 13 cell lines (E026, E029, E033, E034, E041, E042, E043, E044, E045, E049, E062, E089, and E090).

Fig 3. Model evaluation

(a) Multiple feature subsets were used to evaluate appropriate predictor subsets of all five models according to the feature importance, of which, 'rf' represented random forest, 'C5.0' represented single decision tree, 'knn' represented k-nearest neighbors, 'CSimca' represented soft independent modeling of class analogy and 'svmRadial' represented support vector machines with radial basis function kernel, respectively. The sizes of features ranged from 1 up to 762. The performance of models was evaluated by F1 score, the harmonic mean of precision and recall. (b) Feature importance of the 40 regulatory elements in final model. Importance ranking for the regulatory elements in the optimal model was evaluated by using importance evaluation function in caret package in R.

Fig 4. Predicting osteoporosis-associated SNPs in a random region of chromosome 2

The peaks of top ten predictive regulatory elements (Fig 3b) of our final model in labeled positive SNPs are shown, separated by labeled negative SNPs. The most important elements in our final model was eQTL annotation in whole blood which existed partly in both labeled positive SNPs and negative SNPs. The peaks of H3k27me3 in 3 cell lines (Monocd14, Gm12878 and HSMM) and Ezh2_ (39875) in HSMM enriched in selected region which can't distinguish positive SNPs from negative ones. Other elements either covered both labeled SNPs, such as P300kat3b in Osteoblasts and 'Weak Repressed PolyComb' in E049, or didn't covered all specific labeled SNPs, including 'Repressed PolyComb' in E049, 'Quiescent/Low' in E090 and Ezh2_ (39875) in CD20 which can't predicted for positive SNPs alone.

Fig 5. Validation of prediction

(a) The proportion of predicted positive SNPs that in LD with labeled positive SNPs in different degrees. (b) Genomic region annotation distribution of the predicted positive SNPs. (c) Regulatory annotation of predicted positive SNPs, including TFBSs, DHSs, eQTL and evolutionary conservation. About three-quarter predicted positive SNPs covered by at least one annotation among these four kinds of annotations. Histone modifications were also used to annotate predicted positive SNPs, and the majority of SNPs (99.93%) were covered by at least one histone modification in at least one cell line (not shown). (d) Pathway enrichment analysis of genes whose expression level was affected by predicted positive SNPs. Eleven pathways in KEGG were significantly enriched after multiple testing corrections, including lysosome, viral myocarditis, metabolic pathways, herpes simplex infection pathway, etc.

Table 1

Comparison of machine learning

Measure	CSimca	knn	svmRadial	C5.0	rf
Number of Features	80	763	763	80	40
Sensitivity	0.5949	0.7667	0.6297	0.7180	0.8256
Specificity	0.9957	0.9869	0.9988	0.9992	0.9982
Precision	0.8755	0.7513	0.9628	0.9790	0.9583
Accuracy	0.9760	0.9762	0.9817	0.9854	0.9897
F1 Score	0.7084	0.7589	0.7614	0.8284	0.8871

Footnotes: CSimca, soft independent modeling of class analogy; knn, k-nearest neighbors; svmRadial, support vector machines with radial basis function kernel; C5.0, decision tree; rf, random forest.

Significant association results between the new genes of predicted positive SNPs and femoral neck/spine BMD in the GWASs BMD datasets

Table 2

SNP	Chr	Position ^a	A1/A2	Genic position	Gene ^b	Combined meta-analysis ^c	
						P_FN	P_SPINE
rs1464566	3	49459376	T/C	intronic	AMT	0.00137	8.98×10^{-7}
rs2272313	12	53671549	A/G	intronic	ESPL1	3.38×10^{-6}	1.95×10^{-11}
rs2510397	11	68419559	A/G	intergenic	GAL	0.04502	7.89×10^{-8}

Footnotes: Chr: chromosome; FN: femoral neck.

^aPosition was relative to the GRCh37/hg19 version of the human genome.

^bThe predicted positive SNP was assigned to the closest gene.

^cCombined meta-analysis indicates the *P* value was combined by the four GWASs datasets, including GEFOS and three in-house GWASs datasets.