OXFORD

Briefings in Bioinformatics, 20(1), 2019, 26–32

doi: 10.1093/bib/bbx094
Advance Access Publication Date: 16 August 2017
Paper

Integrating regulatory features data for prediction of functional disease-associated SNPs

Shan-Shan Dong,* Yan Guo,* Shi Yao, Yi-Xiao Chen, Mo-Nan He, Yu-Jie Zhang, Xiao-Feng Chen, Jia-Bin Chen and Tie-Lin Yang

Corresponding author: Tie-Lin Yang, Key Laboratory of Biomedical Information Engineering of Ministry of Education, School of Life Science and Technology, Xi'an Jiaotong University, No.28 Xianning West Road, 710049, Xi'an, Shaanxi, P. R. China. Tel.: 86-29-82668463; E-mail: yangtielin@mail.xjtu.edu.cn
*These authors contribute equally to this work.

Abstract

Genome-wide association studies (GWASs) are an effective strategy to identify susceptibility loci for human complex diseases. However, missing heritability is still a big problem. Most GWASs single-nucleotide polymorphisms (SNPs) are located in noncoding regions, which has been considered to be the unexplored territory of the genome. Recently, data from the Encyclopedia of DNA Elements (ENCODE) and Roadmap Epigenomics projects have shown that many GWASs SNPs in the noncoding regions fall within regulatory elements. In this study, we developed a pipeline named functional disease-associated SNPs prediction (FDSP), to identify novel susceptibility loci for complex diseases based on the interpretation of the functional features for known disease-associated variants with machine learning. We applied our pipeline to predict novel susceptibility SNPs for type 2 diabetes (T2D) and hypertension. The predicted SNPs could explain heritability beyond that explained by GWAS-associated SNPs. Functional annotation by expression quantitative trait loci analyses showed that the target genes of the predicted SNPs were significantly enriched in T2D or hypertension-related pathways in multiple tissues. Our results suggest that combining GWASs and regulatory features data could identify additional functional susceptibility SNPs for complex diseases. We hope FDSP could help to identify novel susceptibility loci for complex diseases and solve the missing heritability problem.

Key words: complex diseases; machine learning; SNPs; regulatory feature data; missing heritability; FDSP

Shan-Shan Dong is currently working as a lecturer at Key Laboratory of Biomedical Information Engineering of Ministry of Education, School of Life Science and Technology, Xi'an Jiaotong University.
Yan Guo is currently working as an associate professor at Key Laboratory of Biomedical Information Engineering of Ministry of Education, School of Life Science and Technology, Xi'an Jiaotong University.
Shi Yao is a PhD student of the Key Laboratory of Biomedical Information Engineering of Ministry of Education, School of Life Science and Technology, Xi'an Jiaotong University.
Yi-Xiao Chen is a PhD student of the Key Laboratory of Biomedical Information Engineering of Ministry of Education, School of Life Science and Technology, Xi'an Jiaotong University.
Mo-Nan He is currently working as an undergraduate student at Key Laboratory of Biomedical Information Engineering of Ministry of Education, School of Life Science and Technology, Xi'an Jiaotong University.
Yu-Jie Zhang is a postgraduate student of the Key Laboratory of Biomedical Information Engineering of Ministry of Education, School of Life Science and Technology, Xi'an Jiaotong University.
Xiao-Feng Chen is a PhD student of the Key Laboratory of Biomedical Information Engineering of Ministry of Education, School of Life Science and Technology, Xi'an Jiaotong University.
Jia-Bin Chen is a PhD student of the Key Laboratory of Biomedical Information Engineering of Ministry of Education, School of Life Science and Technology, Xi'an Jiaotong University.
Tie-Lin Yang is currently working as a professor at Key Laboratory of Biomedical Information Engineering of Ministry of Education, School of Life Science and Technology, Xi'an Jiaotong University.
Submitted: 4 May 2017; **Received (in revised form):** 26 June 2017

Introduction

Using genomic information to provide new insights into the disease pathophysiology is the goal of current genetic studies. Genome-wide association studies (GWASs) are an effective strategy to achieve this goal, and many susceptibility loci for human complex diseases have been identified by GWASs [1]. However, missing heritability, which refers to the fact that known susceptibility loci identified by GWASs could only account for limited proportion of the observed heritability of diseases, is still a big challenge for GWASs. True association signals might be missed with the stringent genome-wide significance threshold because of the modest genetic effect size and inadequate statistical power [2, 3]. Therefore, new methods are needed to identify such associations.

As most single-nucleotide polymorphisms (SNPs) reported by GWASs are located in intronic or intergenic regions [4], it was challenging to understand their functional significance. Strikingly, with the regulatory data from Encyclopedia of DNA Elements (ENCODE) [5] and Roadmap Epigenomics Project [6], researchers have found that GWASs SNPs usually lie within regulatory elements, suggesting that they might be involved in regulating gene expression [5, 7]. Finding the common regulatory features of susceptibility SNPs has provided new insights into the biological link between SNPs and phenotypes in several diseases, such as breast cancer [8] and prostate cancer [9]. In addition, we have previously found that promoters of known susceptibility genes for complex diseases (such as obesity [10] and osteoporosis [11]) shared similar regulatory features, and prioritizing genes according to the features could identify novel candidate susceptibility genes. However, we only focused on promoters, and information of the other regions was missed, as most susceptibility loci are located in intergenic or intronic regions. This limitation prompted us to find more powerful methods to predict novel risk SNPs from the large amount of SNP data and complex regulatory features data.

Machine learning is concerned with developing computer algorithms to assist humans in the analysis of large complex data sets, and it has been widely used in the area of genetics and genomics [12]. For example, it can be used to predict transcription start sites [13], identify splice sites [14], promoters [15] and enhancers [16]. Of note, using regulatory features data along the genome, machine learning has been used to predict enhancer–promoter interactions [17] and chromatin organization [18]. In addition, recent studies have used machine learning to estimate the effects of human genetic variants based on regulatory data [19, 20], confirming that machine learning is applicable to interpret regulatory features data for large amount of SNPs. Currently, there is no attempt to predict novel susceptibility variants for a specific complex disease using regulatory features data.

In this study, we hypothesized that interpretation of the functional features for known disease-associated variants with machine learning may identify new susceptibility loci. Based on this hypothesis, we developed a package named functional disease-associated SNPs prediction (FDSP). FDSP is able to predict new susceptibility loci for complex diseases based on known GWAS results and public regulatory data. To illustrate the performance of FDSP, the real GWAS results and regulatory data for type 2 diabetes (T2D) and hypertension were analyzed.

Methods

Pipeline of FDSP

Acquisition of labeled SNPs

The outline of FDSP is shown in Figure 1. SNPs with minor allele frequency (MAF) ≥ 0.01 in the European population were obtained from the 1000 Genome project (Phase III, http://www.1000genomes.org/) and 8 550 206 autosomal SNPs were obtained. We only focused on the autosomal SNPs, as loci on the X chromosome may have different regulatory mechanism because of X-chromosome inactivation. This total SNP set was consisted of labeled and unlabeled SNPs. The labeled SNPs were consisted of positive and negative SNPs. For the labeled positive SNPs, first, we obtained index SNPs from the public SNP–trait association databases: GWAS catalog (https://www.ebi.ac.uk/gwas/) [21] with the threshold of P-value $< 5 \times 10^{-8}$. Owing to the low genomic coverage of GWAS genotyping microarrays, risk-associated SNPs are statistically more likely to be in linkage disequilibrium (LD) with causal variants than to be causal themselves [22]. Therefore, second, we obtained SNPs that were in strong LD ($r^2 \geq 0.8$) with each index SNP using the 1000 Genomes Phase III data. The maximum distance for r^2 calculation was set as 1000 kb. The comprehensive collection of all such SNPs was referred as the positive SNPs. Referring to the methods reported by Zhou *et al.* [20], we created four sets of SNPs with different distances to positive SNPs to form the negative SNP set. The maximum distance in each group was 40, 200, 1000 and 5000 kb, respectively. SNPs with similar MAF to the positive SNP were remained (MAF difference < 0.01). The full set, 80, 60 and 25% random subset of the four groups of SNPs, was merged as negative SNPs. All negative SNPs were filtered to remove overlap with positive SNPs. In the final labeled SNP set, there were 20 negatives per positive.

Feature annotation

Functional annotation of all SNPs was carried out based on the epigenomic data of T2D relevant cell lines (Supplementary

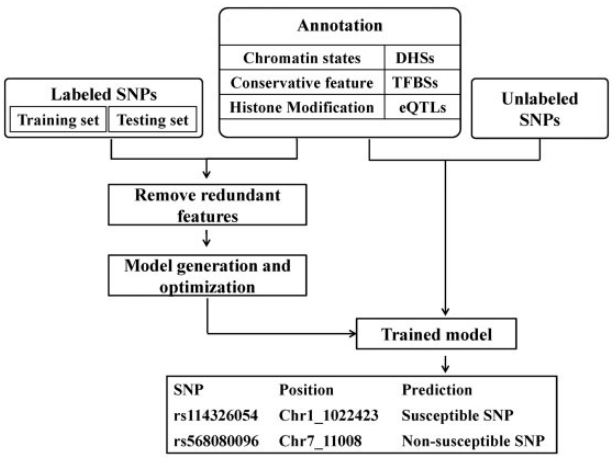

Figure 1. Schematic diagram of FDSP. Labeled positive SNPs were obtained from the GWAS catalog database, and labeled negative SNPs were selected from the whole SNP set according to the procedure described in the 'Methods' section. The whole SNP set was constituted of SNPs with MAF over 0.01 from the 1000 Genome project. All SNPs were annotated with epigenomic data of disease-related cell lines from the ENCODE and Roadmap project, eQTL data in disease-related tissues from the GTEx database (http://www.gtexportal.org/) and conservative genomic regions identified by GERP++. The labeled SNPs were used to train the machine learning model that predicted novel risk SNPs.

Table S1) from the ENCODE [5] and Roadmap Epigenomics Project [6], expression quantitative trait loci (eQTL) data in T2D-relevant tissues (Supplementary Table S2) from the GTEx database (http://www.gtexportal.org/) and conservative genomic regions identified by GERP ++ [23]. All T2D-relevant tissues/cells can be classified into six types according to their source tissues, including adipose, brain, immune cells, liver, muscle and pancreas. The epigenomic data included transcription factor binding sites (TFBSs), chromatin segmentation states, histone modification marks and DNase I hypersensitive sites (DHSs). For each SNP, an epigenomic or conservative feature was labeled 1 if the SNP overlaps with the feature and 0 otherwise. The eQTL feature was labeled 1 if the SNP affects the expression of at least one gene ($P < 0.01$) and 0 otherwise.

Model generation, evaluation and optimization

FDSP automatically tested four widely used machine learning algorithms, including single decision tree (C5.0), soft independent modeling by class analogy (CSimca), random forest (RF) and support vector machines with class weights (svmRadialWeights) to build prediction models. All these algorithms were implemented in the 'caret' package in R (version 3.3.2). For all labeled SNPs, different features may be highly correlated with each other, resulting in redundant information. To address this problem, a correlation matrix was built to remove redundant features. Before training models, highly correlated features ($|r| > 0.7$) were removed. The remained features were used to estimate feature importance values. The labeled SNPs were then randomly divided into training and testing sets according to the proportion of 8:2. Cross-validation was carried out when dividing the labeled SNP set into training and testing set. Next, all models were built by using 5-fold cross-validation with the training set. To correct the imbalance problem, we used the upSample function in the caret package to randomly sample (with replacement) the positive class to be the same size as the negative class. The testing set was further used to evaluate the performance with F1 score [24]. To obtain the best subset of features, we used the recursive feature elimination method as follows:

1. Train the model on the training set using all features
2. Calculate model performance
3. Calculate feature importance and ranking
4. For each subset size S*i*, do
5. Keep the S*i* most important variables
6. Train the model on the training set using S*i* features
7. Calculate mode performance
8. Recalculate the ranking for each feature
9. End
10. Calculate the F1 score for the S*i* features
11. Determine the appropriate number of features
12. Use the model corresponding to the optimal S*i*

Finally, the best performing subset was used to generate the prediction model, and the model was used to predict new susceptibility SNPs. For the predicted positive SNPs, to facilitate future validation experiments, we further ranked them according to the score S, which was defined as follows:

$$S = \sum_{i=1}^{n} Ci * \log_2 FCi.$$

Where n is the total number of features used in the final machine learning model, Ci is the annotation status (0 or 1) of the feature i and FCi is the fold change of the proportion of feature i when compared the labeled positive SNPs and the labeled negative SNPs.

Real data analysis

Acquisition of labeled SNPs

We applied FDSP to analyze the GWAS results and regulatory data of T2D. With P-value $< 5 \times 10^{-8}$, 73 autosomal SNPs associated with T2D in the European population were extracted from GWAS catalog (Supplementary Table S3). To avoid overfitting in the validation process, we excluded eight SNPs from the Finland-United States Investigation of NIDDM Genetics (FUSION) study, and 65 index SNPs were used in subsequent analyses. Of these 65 SNPs, only 4 SNPs mapped to coding exons, with 58 mapping to introns or intergenic regions. The rest three SNPs mapped to 3′ untranslated region (UTR), 5′ UTR and downstream region, respectively. Using the LD cutoff of $r^2 \geq 0.8$, we identified 1769 SNPs in LD with the 65 SNPs. By considering SNPs in LD with the risk-associated SNPs, the number of SNPs mapping to coding exons increased from 4 to 18, and the number of SNPs mapping to intron and intergenic region increased from 37 to 1122 and 21 to 450, respectively. These 1769 SNPs were used as positive risk SNP set, and the negative SNPs were selected using the 1000 genome data accordingly.

Feature annotation

After removing redundant features, 1207 features were remained, including 202 transcription factor (TF) binding profiles, 33 DHSs profiles, 315 histone mark profiles, 639 chromatin states, 17 eQTLs and conservative feature information. The URLs of all features are listed in Supplementary Table S4.

Model generation, evaluation and optimization

Four algorithms were used to get the appropriate prediction model. The performance of the RF algorithm was generally better than others, and the best performance of RF was obtained with the top 60 informative features (Supplementary Figure S1). As shown in Table 1, the best feature numbers for C5.0, CSimca, RF and svmRadialWeights were 60, 300, 60 and 1207, with the best F1 score of 0.8768, 0.6812, 0.9213 and 0.8449, respectively. Therefore, we chose the RF algorithm to predict novel risk SNPs subsequently.

For each feature used in the final model, we performed Fisher's exact test to check whether it is significantly overrepresented or underrepresented in labeled positive SNPs when compared with the labeled negative SNPs. The Benjamini–Hochberg method was used to correct the multiple testing problems. Rank order of the features in the final prediction model is shown in Supplementary Figure S2A. The feature of TFBSs, DHSs and evolutionary conserved region was not included in the model. As

Table 1. Performance of different machine learning algorithms for T2D

Measure	C5.0	CSimca	RF	svmRadialWeights
Number of features	60	300	60	1207
Sensitivity	0.9591	0.5793	0.9736	0.9368
Specificity	0.9681	0.994	0.9852	0.967
Accuracy	0.9677	0.9742	0.9847	0.9655
F1 score	0.8768	0.6812	0.9213	0.8449

Note: C5.0, decision tree; CSimca, soft independent modeling of class analogy; RF, RF; svmRadialWeights, support vector machines with class weights.

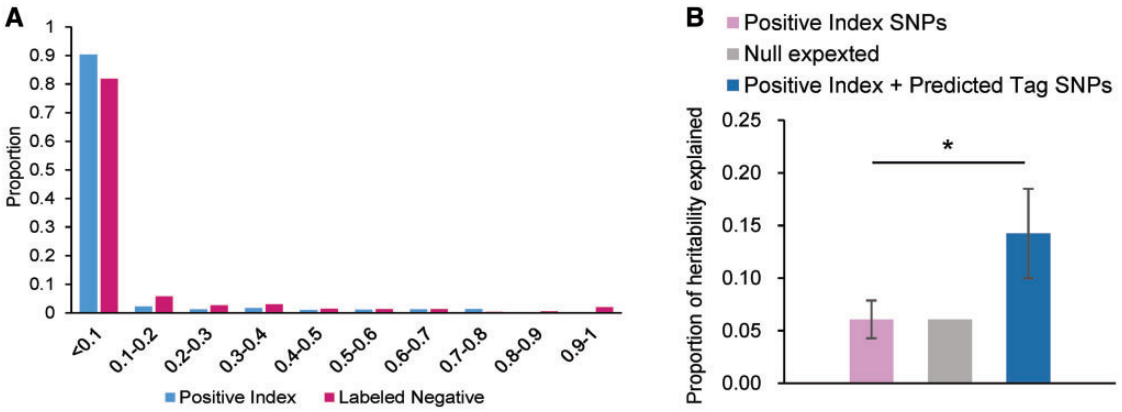

Figure 2. (**A**) Distribution of the LD measurement r^2 between the predicted T2D-positive SNPs and the published GWASs SNPs or labeled negative SNPs. For each predicted positive SNPs, r^2 values were calculated between it and any other index SNPs or labeled negative SNPs. The biggest r^2 for each predicted positive SNP was kept for the summary plot. (**B**) Narrow-sense heritability ($hg^2 \pm$ standard error) explained by index SNPs, null expected and index SNPs + predicted tag SNPs for T2D (*$P < 0.05$).

shown in Supplementary Figure S2B, the eQTL effects from nine tissues (including subcutaneous adipose, five brain tissues, whole blood, muscle and pancreas) were all significantly enriched in labeled positive SNPs. As shown in Supplementary Figure S2C, among the chromatin states, labeled positive SNPs are significantly enriched with the state of 'weak transcription', and this signature is shared in 12 cell lines. Consistently, depletion of the 'quiescent/low' or 'weak repressed polycomb' chromatin state was found in seven cell lines. As shown in Supplementary Figure S2D, among all of the histone marks, enrichment of monomethylation of lysine 20 on histone H4 (H4k20me1) and dimethylation of lysine 79 on histone H3 (H3k79me2) was both found in four cell lines. Enrichment of trimethylation of lysine 36 on histone H3 (H3k36me3) was detected in three cell lines, while enrichment of acetylation of histone H3 at lysine 27 (H3k27ac) was only detected in CD20+. Depletion of the trimethylation of lysine 9 on histone H3 (H3K9me3) repressive mark was found in three cell lines. Depletion of the trimethylation of lysine 27 on histone H3 (H3k27me3) was found in eight cell lines. Significant depletion of H2AZ was detected in two cell lines. As shown in Supplementary Table S1, all these cells are derived from five tissues closely related to T2D, including 3 adipose cells, 2 brain cells, 11 immune cells, 2 liver cells and 3 muscle cells.

New susceptibility SNP prediction

A total of 8 513 057 unlabeled SNPs were subjected to prediction using the RF model, and 15 204 novel potential T2D-associated SNPs were obtained (Supplementary Table S5) with the detection rate of about 0.18%. The ranking scores for each predicted SNPs were also listed in the last column of Supplementary Table S5. The number of SNPs mapping to coding exons was 238, and the number of SNPs mapping to intron and intergenic region was 10 326 and 2 583, respectively. The distribution of the predicted SNPs along the genome is similar to that of the labeled positive SNPs (Supplementary Figure S3). These SNPs could be captured by 5496 tag SNPs.

Predicted SNPs may explain additional heritability

To check whether the predicted SNPs are independent from the published 65 index SNPs or the labeled negative SNPs, we calculated r^2 between each predicted SNP and all index SNPs or labeled negative SNPs on the same chromosome based on the 1000 genome data from the European individuals. For each predicted positive SNP, the biggest r^2 was kept. As shown in Figure 2A, over 90% predicted positive SNPs were in extremely weak LD ($r^2 < 0.1$) with the published GWASs SNPs. Similarly, >80% of the predicted positive SNPs were in weak LD with the labeled negative SNPs. Therefore, the predicted positive SNPs cannot be represented by the published GWASs SNPs or the labeled negative SNPs.

To investigate whether the predicted SNPs could explain additional heritability, SNP-based heritability of T2D was estimated using GCTA-GREML [25]. Data from the Database of Genotypes and Phenotypes (dbGaP) with the accession number of phs000867.v1.p1 were used, including 919 T2D cases and 787 controls. All subjects were independent individuals recruited from Finland. To facilitate the investigation of all predicted SNPs, we used the IMPUTE2 program [26] to impute genotypes of SNPs that based on the 1000 genome data (version 3). The prevalence of diabetes was estimated as 8.5% in Europe, with over 90% of T2D [27]. Therefore, the European T2D prevalence was set as 7.65%. We performed z tests to compare h_g^2 estimates from index SNPs with h_g^2 estimates from index SNPs and the tag SNPs of predicted positive SNPs. As shown in Figure 2B, predicted SNPs significantly increased the proportion of explained heritability ($P < 0.05$).

Considering the results of the heritability calculation may be affected by the number of the predicted positive SNPs, we first calculated the null expected h_g^2 on the basis of the fraction of the genome represented by the tag SNPs of the predicted positive SNPs in the heritability estimates. As previously described [28], $h^2_{null} = h^2_{index\ SNPs} + x \times (\text{total } h_g^2 - h^2_{index\ SNPs})$, where x is the proportion of the genome covered by the tag SNPs of the predicted positive SNPs. z tests were also performed to compare h_g^2 estimates from index SNPs and predicted variants with the null expected heritability estimates. As shown in Figure 2B, the heritability explained by predicted positive tag SNPs and index SNPs was significantly higher than the null expected.

Second, to confirm whether the increase in heritability was specific to the predicted positive SNPs, we also compared predicted positive tag SNPs with random predicted negative SNPs. We selected 1000 random subsets with the same number of predicted positive tag SNPs. As shown in Supplementary Figure S4, tag SNPs of the predicted SNPs explain significantly more heritability than the random negative SNPs.

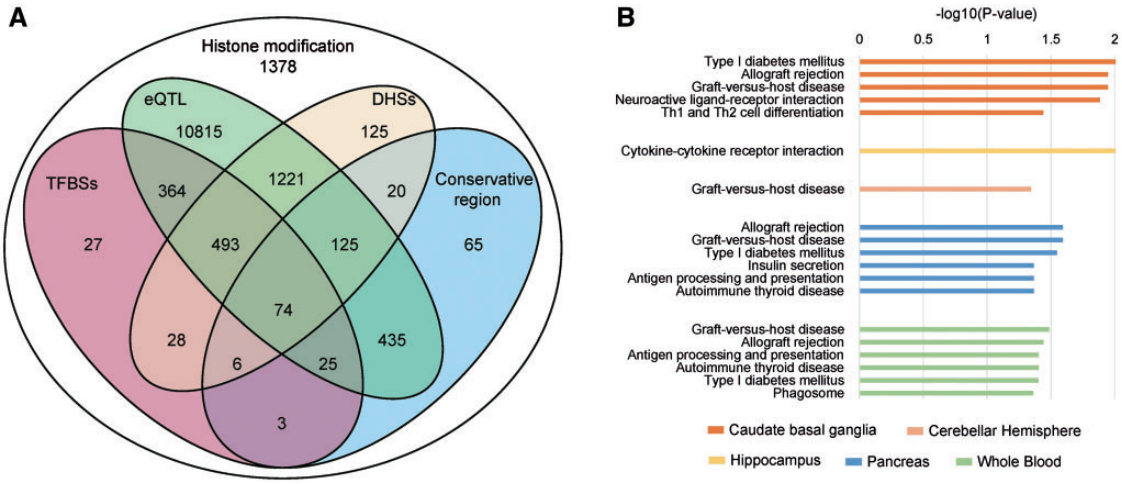

Figure 3. (**A**) Diagram of the annotations results for the predicted T2D susceptibility SNPs. All SNPs were annotated with at least one histone modification in at least one cell line. (**B**) Pathway enrichment analysis results of the genes affected by different genotypes of the predicted positive SNPs for T2D.

Functional annotation of the predicted positive SNPs

The annotation results showed that all predicted positive SNPs were located in at least one histone modification region in at least one cell line (Figure 3A). In total, 1020 SNPs were located in TFBSs, 13 552 SNPs may affect gene expression with $P < 0.01$ in the GTEx data in at least one T2D-related tissue, 2092 SNPs were located in DHSs and 753 SNPs were located in conservative regions.

Characterization of the genes that might be affected by the predicted positive SNPs

Using the eQTL data for nine eQTL tissues among the top 60 features from the GTEx database, we obtained the genes that might be affected by the predicted positive SNPs with the cutoff of $P < 0.01$. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis for the eQTL target genes of the predicted positive SNPs was carried out by using hypergeometric distribution test. The Benjamini–Hochberg method was used to correct the multiple testing problems. In each tissue, background genes were set as all protein-coding genes that might be affected by any SNP with the *P*-value <0.01. As shown in Figure 3B, after multiple testing corrections, pathways enriched with eQTL target genes of the predicted positive SNPs were detected in five tissues. Of note, the type I diabetes mellitus pathway showed significant enrichment in three tissues, and the insulin secretion pathway was also detected in pancreas (adjusted $P = 0.043$). Other pathways were mostly associated with the immune system, such as antigen processing and presentation, graft-versus-host disease, allograft rejection, cytokine–cytokine receptor interaction, etc.

Application of FDSP to hypertension

To further confirm the reliability of FDSP, we applied our pipeline to hypertension. In total, 57 index SNPs (Supplementary Table S6) were obtained from the GWAS catalog database. In total, 2086 labeled positive and 41 720 labeled negative SNPs were then generated. Hypertension-relevant cell lines are listed in Supplementary Table S7, and eQTL tissues are listed in Supplementary Table S8. All features can be obtained from the URLs listed in Supplementary Table S9. After removing redundant features, 1046 features were remained, including 145 TF binding profiles, 46 DHSs profiles, 292 histone mark profiles, 544 chromatin states, 18 eQTLs and conservative feature information. According to the model generation, evaluation and optimization results, the RF model performed best with the F1 score of 0.8881. A total of 84 809 positive SNPs were predicted. To check whether the predicted SNPs could explain additional heritability of hypertension, data from dbGaP with the accession number of phs000297.v1.p1 were used, including 1487 cases and 1366 controls. Individuals without evidences of hypertension (systolic blood pressure, SBP < 140 and diastolic blood pressure, DBP < 90, and no mention of using any antihypertensive drugs) were defined as controls. Individuals with evidences of hypertension (SBP > 140 or DBP > 90 and reported using any antihypertensive drugs) were defined as cases. The results showed that the predicted positive SNPs could also explain additional heritability of hypertension (Figure 4A and Supplementary Figure S5). Pathways enriched with eQTL target genes of the predicted positive SNPs were detected in three tissues, including subcutaneous adipose, hippocampus and putamen basal ganglia. These pathways were associated with the immune system, including cytokine–cytokine receptor interaction, allograft rejection, graft-versus-host disease and viral myocarditis.

Discussion

GWASs have identified many susceptibility loci for complex diseases. However, it is still challenging to find the missing hereditary. The ENCODE [5] and Roadmap Epigenomics Project [6] provide rich sources of regulatory features data, reminding us that integration of the regulatory features data and GWASs results may lead to the identification of new susceptibility SNPs. In this study, we proposed a pipeline, named FDSP, to predict new susceptibility SNPs for complex diseases. We applied this pipeline to T2D and hypertension, and the predicted SNPs could explain additional heritability. Moreover, functional analyses of the novel SNPs suggested that they are potentially associated with T2D or hypertension, implicating the efficiency of finding missing heritability of complex diseases by machine learning with regulatory features data.

In addition to population-level association or linkage studies, researchers have also spent effects to unravel the comprehensive genetic basis of diseases using genome sequence data or gene network information. For example, Guan *et al.* [29] proposed an approach by interrogating high-throughput genomic data in model organisms to functionally associate genes with

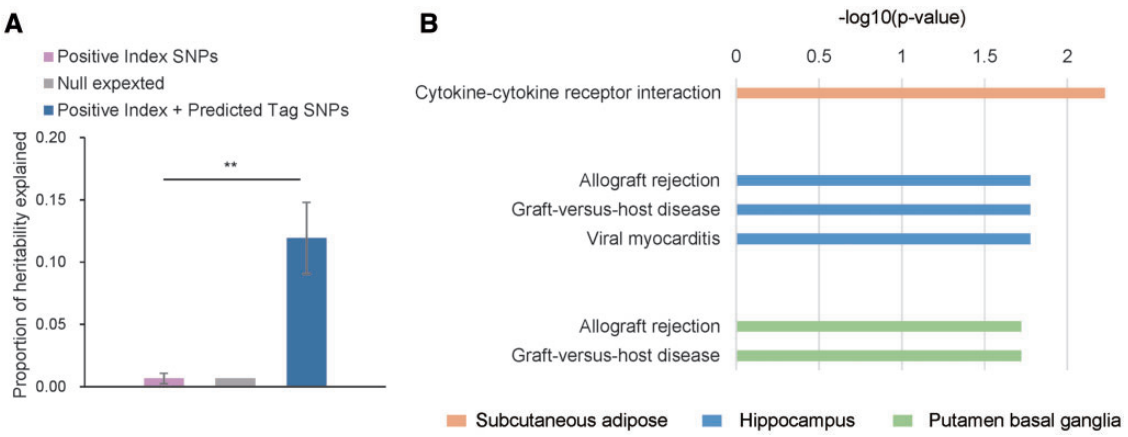

Figure 4. (**A**) Narrow-sense heritability ($hg^2 \pm$ standard error) explained by index SNPs, null-expected and index SNPs + predicted tag SNPs for hypertension (**$P < 0.01$). (**B**) Pathway enrichment analysis results of the genes affected by different genotypes of the predicted positive SNPs for hypertension.

diseases. Vanunu *et al.* [30] used gene network propagation to associate genes with diseases. Further, Krishnan *et al.* [31] developed a complementary approach based on a human brain-specific gene network to predict autism risk genes. However, these methods were all gene-level candidate prediction approaches, and SNP-level prediction methods are still lacking. Here, we proposed a SNP-level prediction pipeline, which might identify new susceptibility loci for complex diseases through interpreting the regulatory features for known disease-associated variants.

Among the features used in the T2D trained model, we observed that labeled T2D SNPs are significantly enriched with chromatin states of transcription and histone modification of H3k27ac, H3k36me3, H3k79me2 and H4k20me1. H3k27ac is the hallmark of active enhancers [32]. Both H3k36me3 and H3k79me2 are elongation-associated histone marks. H4k20me1 is associated with transcriptional activation [33]. In addition, depletion of the chromatin states of quiescent/low state and histone modification of H3K9me3, H3K27me3 and H2AZ were also observed in labeled positive SNPs. H3K9me3 and H3K27me3 are all marks of transcriptional repression [34, 35], and H2AZ is also related to polycomb silencing [36]. Therefore, labeled positive SNPs were generally enriched with regulatory features of transcriptional activation and depleted with features of transcriptional repression. Consistently, labeled positive SNPs were also enriched with eQTL effects, suggesting that although most known positive SNPs were not located in the exonic region, they may be involved in gene regulation through affect the regulatory features. The enrichment or depletion was observed in multiple cell lines, which might be as expected, as T2D is a complex disease, and many tissues are involved in the progression of T2D.

We predicted potential novel susceptibility SNPs for T2D and hypertension. Heritability calculation results confirmed that the predicted SNPs may explain additional heritability. Pathway analyses found that T1D and immune system-related pathways were enriched with eQTL target genes of T2D-predicted SNPs in multiple tissues. The immune system is a key mediator in the development of T2D [37, 38]. For hypertension, immune system-related pathways were also found to be enriched with eQTL target genes of the predicted SNPs in multiple tissues. The immune system plays important roles in the initiation and maintenance of hypertension [39]. Therefore, the results confirmed that predicting novel susceptibility SNPs for complex diseases with our pipeline is effective.

Limitations of our study should be addressed. When we test our pipeline in T2D and hypertension, we only used cis-eQTL results in the annotation of the eQTL feature without considering the trans-eQTL effects. Trans-eQTL is also an important feature for SNPs. Owing to the high density of our test SNPs, it is time and storage costing to get the whole-genome trans-eQTL results. However, if researchers who want to use FDSP have the trans-eQTL results, they can easily add this feature in the annotation process.

In summary, through integrating regulatory features and GWASs data, we developed FDSP to predict new susceptibility loci for human complex diseases. Application of FDSP to T2D and hypertension data demonstrated the effectiveness of our pipeline. We hope that FDSP could provide new insights into the identification of additional susceptibility SNPs for complex diseases.

Key Points

- GWASs are an effective strategy to identify susceptibility loci for human complex diseases. However, missing heritability is still a big problem.
- We developed a pipeline to predict novel disease-associated variants through integrating regulatory features data and GWASs results.
- We applied our pipeline to predict novel susceptibility loci for T2D and hypertension, and the results confirmed the reliability of our pipeline.

Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Acknowledgements

The authors thank the Finland-United States Investigation of NIDDM Genetics (FUSION) study, which aims to identify genetic variants associated with T2D and T2D-related traits. The authors also thank the electronic Medical Records and Genomics (eMERGE) Network consortium, which spent much effect on investigating the genetic variants associated with hypertension. During the preparation

of this manuscript, we did not collaborate with the investigators of the FUSION study or the eMERGE Network consortium. Therefore, our study does not necessarily reflect the opinions of them. The data sets we used were obtained through dbGaP authorized access at http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap with the accession number of phs000867.v1.p1 and phs000297.v1.p1.

Funding

The National Natural Science Foundation of China (grant numbers 31371278, 31471188, 81573241 and 31511140285); China Postdoctoral Science Foundation (grant numbers 2016T90902 and 2016M602797); Natural Science Basic Research Program Shaanxi Province (grant number 2016JQ3026); and the Fundamental Research Funds for the Central Universities.

References

1. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 2005;**6**:95–108.
2. Manolio TA, Collins FS, Cox NJ, *et al*. Finding the missing heritability of complex diseases. *Nature* 2009;**461**:747–53.
3. Lee SH, Wray NR, Goddard ME, *et al*. Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet* 2011;**88**:294–305.
4. Hindorff LA, Sethupathy P, Junkins HA, *et al*. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 2009;**106**:9362–7.
5. Consortium TEP. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74.
6. Roadmap Epigenomics C, Kundaje A, Meuleman W, *et al*. Integrative analysis of 111 reference human epigenomes. *Nature* 2015;**518**:317–30.
7. Schaub MA, Boyle AP, Kundaje A, *et al*. Linking disease associations with regulatory information in the human genome. *Genome Res* 2012;**22**:1748–59.
8. Cowper-Sal lari R, Zhang X, Wright JB, *et al*. Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat Genet* 2012;**44**:1191–8.
9. Guo H, Ahmed M, Zhang F, *et al*. Modulation of long noncoding RNAs by risk SNPs underlying genetic predispositions to prostate cancer. *Nat Genet* 2016;**48**:1142–50.
10. Dong SS, Guo Y, Zhu DL, *et al*. Epigenomic elements analyses for promoters identify ESRRG as a new susceptibility gene for obesity-related traits. *Int J Obes* 2016;**40**:1170–6.
11. Guo Y, Dong SS, Chen XF, *et al*. Integrating epigenomic elements and GWASs identifies BDNF gene affecting bone mineral density and osteoporotic fracture risk. *Sci Rep* 2016;**6**:30558.
12. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet* 2015;**16**:321–32.
13. Ohler U, Liao GC, Niemann H, *et al*. Computational analysis of core promoters in the Drosophila genome. *Genome Biol* 2002;**3**:RESEARCH0087.
14. Degroeve S, De Baets B, Van de Peer Y, *et al*. Feature subset selection for splice site prediction. *Bioinformatics* 2002;**18** (Suppl 2):S75–83.
15. Bucher P. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J Mol Biol* 1990;**212**:563–78.
16. Heintzman ND, Stuart RK, Hon G, *et al*. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* 2007;**39**:311–8.
17. Whalen S, Truty RM, Pollard KS. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat Genet* 2016;**48**:488–96.
18. Huang J, Marco E, Pinello L, *et al*. Predicting chromatin organization using histone marks. *Genome Biol* 2015;**16**:162.
19. Kircher M, Witten DM, Jain P, *et al*. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;**46**:310–5.
20. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* 2015;**12**:931–4.
21. Welter D, MacArthur J, Morales J, *et al*. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 2014;**42**:D1001–6.
22. McClellan J, King MC. Genetic heterogeneity in human disease. *Cell* 2010;**141**:210–7.
23. Davydov EV, Goode DL, Sirota M, *et al*. Identifying a high fraction of the human genome to be under selective constraint using GERP ++. *PLoS Comput Biol* 2010;**6**:e1001025.
24. Rijsbergen V. *Information Retrieval*. London: Butterworths, 1979.
25. Yang J, Lee SH, Goddard ME, *et al*. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 2011;**88**:76–82.
26. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 2009;**5**:e1000529.
27. Tamayo T, Rosenbauer J, Wild SH, *et al*. Diabetes in Europe: an update. *Diabetes Res Clin Pract* 2014;**103**:206–17.
28. Gusev A, Bhatia G, Zaitlen N, *et al*. Quantifying missing heritability at known GWAS loci. *PLoS Genet* 2013;**9**:e1003993.
29. Guan Y, Ackert-Bicknell CL, Kell B, *et al*. Functional genomics complements quantitative genetics in identifying disease-gene associations. *PLoS Comput Biol* 2010;**6**:e1000991.
30. Vanunu O, Magger O, Ruppin E, *et al*. Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol* 2010;**6**:e1000641.
31. Krishnan A, Zhang R, Yao V, *et al*. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat Neurosci* 2016;**19**:1454–62.
32. Creyghton MP, Cheng AW, Welstead GG, *et al*. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci USA* 2010;**107**:21931–6.
33. Wang Z, Zang C, Rosenfeld JA, *et al*. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* 2008;**40**:897–903.
34. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* 2010;**28**:817–25.
35. Arthur RK, Ma L, Slattery M, *et al*. Evolution of H3K27me3-marked chromatin is linked to gene expression evolution and to patterns of gene duplication and diversification. *Genome Res* 2014;**24**:1115–24.
36. Wang H, Wang L, Erdjument-Bromage H, *et al*. Role of histone H2A ubiquitination in Polycomb silencing. *Nature* 2004;**431**:873–8.
37. Shu CJ, Benoist C, Mathis D. The immune system's involvement in obesity-driven type 2 diabetes. *Semin Immunol* 2012;**24**:436–42.
38. Hameed I, Masoodi SR, Mir SA, *et al*. Type 2 diabetes mellitus: from a metabolic disorder to an inflammatory condition. *World J Diabetes* 2015;**6**:598–612.
39. Singh MV, Chapleau MW, Harwani SC, *et al*. The immune system and hypertension. *Immunol Res* 2014;**59**:243–53.