Addressing the Missing Heritability Problem With the Help of Regulatory Features

Evolutionary Bioinformatics
Volume 15: 1–2
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1176934319860861

Shan-Shan Dong, Yan Guo and Tie-Lin Yang

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

ABSTRACT: Genome-wide association studies (GWASs) have successfully identified thousands of susceptibility loci for human complex diseases. However, missing heritability is still a challenging problem. Considering most GWAS loci are located in regulatory elements, we recently developed a pipeline named functional disease-associated single-nucleotide polymorphisms (SNPs) prediction (FDSP), to predict novel susceptibility loci for complex diseases based on the interpretation of regulatory features and published GWAS results with machine learning. When applied to type 2 diabetes and hypertension, the predicted susceptibility loci by FDSP were proved to be capable of explaining additional heritability. In addition, potential target genes of the predicted positive SNPs were significantly enriched in disease-related pathways. Our results suggested that taking regulatory features into consideration might be a useful way to address the missing heritability problem. We hope FDSP could offer help for the identification of novel susceptibility loci for complex diseases.

KEYWORDS: Missing heritability, regulatory features, complex diseases, FDSP

RECEIVED: June 3, 2019. **ACCEPTED:** June 10, 2019.

TYPE: Commentary

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study is supported by the National Natural Science Foundation of China (31871264, 31771399, and 31701095), Natural Science Basic Research Program Shaanxi Province (2018JQ3058), Innovative Talent Promotion Plan of Shaanxi Province for Young Sci-Tech New Star (2018KJXX-010), and the Fundamental Research Funds for the Central Universities.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Tie-Lin Yang, Key Laboratory of Biomedical Information Engineering of Ministry of Education and Institute of Molecular Genetics, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an 710049, P. R. China. Email: yangtielin@mail.xjtu.edu.cn

COMMENT ON: Dong SS, Guo Y, Yao S, et al. Integrating regulatory features data for prediction of functional disease-associated SNPs. *Brief Bioinform*. 2019;20:26-32. doi:10.1093/bib/bbx094. PMID: 28968709. https://www.ncbi.nlm.nih.gov/pubmed/28968709.

With the help of genome-wide association studies (GWASs), thousands of susceptibility loci for human complex diseases have been uncovered. However, missing heritability, which refers to the fact that published susceptibility loci could only account for limited proportion of the total heritability of complex diseases, is still a challenging problem. With the stringent genome-wide significance threshold, GWASs might miss the true association signals with modest genetic effect size[1,2]; therefore, new methods for susceptibility loci identification are needed.

Recently, with the regulatory data from Encyclopedia of DNA Elements (ENCODE)[3] and Roadmap Epigenomics Project,[4] it is recognized that susceptibility loci from GWASs usually lie within regulatory elements.[3,5] In addition, we have previously found that promoters of susceptibility genes for complex diseases[6,7] shared similar regulatory features. This prior knowledge reminds us that integrating complex regulatory features data and the whole-genome single-nucleotide polymorphisms (SNPs) may offer a new way to solve the missing heritability problem. However, multiple regulatory data for millions of SNPs result in a large amount of data, which is hard to handle with commonly used statistical methods (eg, regression model in GWASs). Machine learning is widely used to assist humans in analyzing large complex data sets. Specifically, based on the analysis of regulatory data, machine learning has been used to predict enhancer-promoter interactions[8] and chromatin organization.[9] Combining the regulatory features data and genetic variants, machine learning has been used to estimate the effects of human genetic variants.[10,11] In other words, machine learning is applicable of handling such multi-dimensional regulatory data for millions of SNPs. Therefore, we developed functional disease-associated SNPs prediction (FDSP),[12] which uses machine learning to build new susceptibility loci perdition model for complex diseases with known GWASs loci and disease-specific regulatory data as input. For the predicted positive (risk) SNPs, we also assigned a rank score to facilitate future validation experiments.

We successfully applied our model in type 2 diabetes (T2D) and hypertension. In addition to index SNPs ($P < 5 \times 10^{-8}$) obtained from GWAS catalog database, SNPs in strong linkage disequilibrium (LD, $r^2 \geqslant .8$) with each index SNP were also referred as the labeled positive SNPs. Four sets of SNPs with different distances to the labeled positive SNPs were generated to form the labeled negative SNP set. Labeled positive SNPs were generally enriched in regulatory features about transcriptional activation (eg, H3K27ac, H3K36me3, H3K79me2, and H4K20me1) and depleted in regulatory features about transcriptional repression (eg, H3K9me3, H3K27me3, and H2AZ). To assess whether the predicted results could offer help for addressing the missing heritability problem, we confirmed that the predicted positive SNPs could explain additional heritability. First, the predicted positive SNPs cannot be represented by the index SNPs as over 90% predicted positive SNPs were in extremely weak LD ($r^2 < .1$) with the index SNPs. Second, using individual data from the Database of Genotypes and Phenotypes (dbGaP, phs000867.v1.p1 for T2D and phs000297.v1.p1 for hypertension), we confirmed that the explained heritability by index SNPs and predicted positive SNPs was significant higher than the index SNPs. Third, to control the effect of SNP counts on the heritability estimation, we further confirmed that heritability explained by predicted positive tag SNPs and index SNPs was significantly higher than the null expected. Last, the increase in heritability was specific to the predicted positive SNPs as tag SNPs of the predicted positive SNPs explain significantly more heritability than the random negative SNPs. Further annotation results suggested that the predicted positive SNPs might be functional, as all of them were located in at least one regulatory region in at least one disease-specific cell/tissue. Pathway analyses showed that genes that might be affected by the predicted positive SNPs were enriched in T2D/hypertension-related pathways, implicating the effectiveness of our method.

FDSP has several advantages compared with conventional GWAS strategies. First, rather than enlarging sample size to improve statistical power and detect susceptibility loci with modest effects, FDSP aims to predict novel susceptibility SNPs based on integration of the published known index SNPs and regulatory feature data through machine learning. This time- and cost-effective method was proved to be useful for addressing the missing heritability problem. Second, with regulatory data implemented in the prediction model, the predicted positive SNPs by FDSP are all with potential regulatory functions, which may provide more insights into disease biology and facilitate further function validation experiments. Third, with cell/tissue-specific regulatory feature included in the model, genes that might be affected by the predicted novel positive SNPs are enriched in disease-related pathways, providing potential targets for therapeutic studies. Fourth, one of the current struggles researchers have with GWAS results is that it is hard to decide the priority of the association signals. The SNPs prioritization according to their annotation of the regulatory features was supported by FDSP, which might facilitate the selection of SNPs for subsequent cellular/animal studies. Last, except for T2D and hypertension, FDSP could easily be used for other complex diseases with user-defined labeled index SNPs and regulatory features. Users do not need to try all machine learning models one by one; FDSP would automatically test 4 commonly used algorithms (single decision tree, soft independent modeling by class analogy, random forest, and support vector machines with class weights) and select the best one. Users could also try other preferred models or algorithms (detail in the website https://github.com/xjtugenetics/FDSP).

Limitations of our study should be acknowledged. First, as we mentioned in the article, when applied to T2D and hypertension, we did not take trans expression quantitative trait loci into consideration during the model training. Second, we only focused on SNPs without considering other types of variants, such as small insertions and deletions and structure variations. However, FDSP is reasonably user-friendly and researchers could easily add features or variants in practice.

In summary, we developed FDSP to predict new susceptibility loci for complex diseases based on the integration of regulatory feature data and published GWAS results. Application of FDSP to T2D and hypertension proved the effectiveness of our method. We hope FDSP could help to solve the missing heritability problem.

Author Contributions

T-LY designed the study. S-SD wrote the manuscript. YG revised the manuscript.

REFERENCES

1. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461:747–753.
2. Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet*. 2011;88:294–305.
3. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489:57–74.
4. Kundaje A, Meuleman W, Ernst J, et al. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015;518:317–330.
5. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. *Genome Res*. 2012;22:1748–1759.
6. 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 (Lond)*. 2016;40:1170–1176.
7. 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.
8. Whalen S, Truty RM, Pollard KS. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat Genet*. 2016;48:488–496.
9. Huang J, Marco E, Pinello L, Yuan GC. Predicting chromatin organization using histone marks. *Genome Biol*. 2015;16:162.
10. 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. *Nat Genet*. 2014;46:310–315.
11. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods*. 2015;12:931–934.
12. Dong SS, Guo Y, Yao S, et al. Integrating regulatory features data for prediction of functional disease-associated SNPs. *Brief Bioinform*. 2019;20:26–32.