IPEV: a web server for inferring pathogenic enhancers with variants

Guanxiong Zhang

Harbin Medical University

Aimin Xie

Harbin Medical University

Jing Bai

Harbin Medical University

Tao Luo

Harbin Medical University

Huating Yuan

Harbin Medical University

Gaoming Liao

Harbin Medical University

Shangyi Luo

Harbin Medical University

Jianlong Liao

Harbin Medical University

Jinyuan Xu

Harbin Medical University

Jian Shi

Harbin Medical University

Yong Zhang

Harbin Medical University

Xinxin Zhang

Harbin Medical University

Tingting Zhao

First Affiliated Hospital of Harbin Medical University

Yun Xiao

Harbin Medical University

Xia Li (✉ lixia@ems.hrbmu.edu.cn)

Harbin Medical University

Keywords: variants, enhancers, disease, web server

Posted Date: September 8th, 2019

DOI: <https://doi.org/10.21203/rs.2.14112/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background Enhancer has been recognized as an important driver whose genetic alterations contribute to disease progression. However, there is still no easy-to-use tools to identify pathogenic enhancers, allowing for deciphering functional influence of genetic variants on enhancer. Results We developed a user-friendly one-stop shop platform, named inferring pathogenic enhancer with variant (IPEV), only requiring variants as input, to quickly infer the pathogenic enhancers that harbor variants affecting their activities. Results of IPEV are explored in an interactive, user-friendly web environment, which is designed to highlight the most probable pathogenic enhancers and their target genes. Furthermore, IPEV provides intuitive visualizations of how a variant affects the corresponding enhancer activity by mediating TF binding changes. Conclusions IPEV is specially designed to prioritize the potentially pathogenic enhancers with genetic variants, and provides intuitive visualizations how a variant affects the corresponding enhancer activity by mediating which transcription factor binding changes. The use of IPEV does not require any specialized computer skills. We believe that IPEV will be useful in interpreting non-coding variants by the inferring pathogenic enhancers. It is freely available at <http://biocc.hrbmu.edu.cn/IPEV/> or <http://210.46.80.168/IPEV> and supports recent versions of all major browsers.

Background

Large-scale sequencing studies of tumor patients have generated millions of non-coding variants [1, 2]. Importantly, many of these variants occur in enhancers, in turn affecting transcription factor binding [3]. Furthermore, Genome-wide association study (GWAS) discovered many risk loci located in non-coding regions, with a significant enrichment in enhancers [4]. Gradually, genetic alterations of enhancers were recognized as one of the main drivers contributing to tumor progression [5, 6], providing a new understanding of the pathogenesis of cancer. Thus, inferring disease enhancers from substantial non-coding mutations is urgently needed for uncovering novel drivers underlying cancer and deciphering the mechanisms of tumorigenesis.

However, the identification of disease enhancers using non-coding mutations still remain a challenge due to a large number of genomic and epigenetic features required, as well as the binding of many TFs. Furthermore, it is a daunting task to process the large volume and high dimensional data, and build a machine learning model to infer the pathogenic enhancers, especially for biologists and clinicians without a programming background. To date, as there is still no tools for this purpose, we believe it is desirable to develop an easily used online tool to infer pathogenic enhancers from non-coding variants and interpret the potential effect of these variants in the enhancers.

The Inferring Pathogenic Enhancers with Variants (IPEV) is a one-stop shop to quickly infer and return the potential pathogenic enhancers that harbor genetic variants affecting their activities in an interactive and user-friendly web environment. IPEV infers the pathogenic enhancers based on more than 540 genomic and epigenetic features and a random forest classification model, and only requires variant data as input.

Further visualization provided by IPEV allows users to investigate which TF bindings in the enhancer were affected by the non-coding variant. With a rapidly increasing interest in enhancers, we believe that IPEV is a timely and valuable tool for the understanding of enhancer deregulation in tumor pathogenesis.

Methods

Experimental datasets

In this study, we built a random forest classifier based on a range of genetic and epigenetic features to infer the pathogenic enhancers that harbor noncoding genetic variants. A total of 3865 enhancer-variant pairs were used in this study for training. The training set was composed of 438 disease-associated positive pairs and 3,427 negative pairs.

The positive set was compiled from the DiseaseEnhancer database (Version1.0.1) [7], which manually collected 847 disease-associated enhancers and their associated variants. Enhancers with multiple variants or indels were eliminated to restrict our set to single nucleotide substitutions. To acquire the nonpathogenic enhancer-variant pairs, an initial catalog of enhancer came from a combination of the Roadmap Epigenomics Mapping Consortium (REMC) [8], the Encyclopedia of DNA Elements (ENCODE) [9] Project (identified using ChromHMM [10]), and FANTOM5 [11] database as described in ELMER [12]. Target genes of these enhancers were identified based on the chromatin interactions from 4DGenome database [13]. Then enhancers were eliminated if they were included in the positive set, or their target genes were included in OMIM, DisGeNET, MalaCards or DISEASE database [14-17]. For each nonpathogenic enhancer, the associated variant was randomly selected from the 1000 Genomes Project (minor allele frequency $\geq 5\%$) [18].

Feature annotation integration

For each enhancer-variant pair, a 548-dimensional annotations were generated (Supplementary Table S1).

(1) Variant-based features

Mutation pattern. We took five-nucleotide window sequence centered on each variant site from the R package *BSgenome.Hsapiens.UCSC.hg19* which contained full genome sequences for homo sapiens as provided by UCSC;

Conservation. To determine the conservation of the variant base in the case of substitutions, PhastCons, phyloP and GERP++ were used [19-21];

Changes of TF bindings. Based on DeepBind [22], we calculated potential change scores of 515 transcription factor bindings mediated by the variant. First the binding score for the reference and variant

sequence were calculated using the sequence around the variant ($\pm 25\text{bp}$). Then the potential change scores were defined as .

(2) Region-based features

Conservation. Region-based conservation scores (i.e. PhastCons, phyloP and GERP) were evaluated by averaging over all base pairs for each candidate enhancer [19-21].

Negative selection. SNP density was expressed as the average number of SNPs in each candidate enhancer region based on 1000 Genomes Project phase 3 data [18]. Meanwhile, we took 338,198 regions belonging to "sensitive" and "ultrasensitive" categories found by Fu *et al* [23]. Frequency of overlap between the candidate enhancer with sensitive or ultrasensitive regions were then be calculated.

Potential TF binding. We took 161 transcription factor binding sites (TFBS) across 91 cell lines from UCSC [24], and motif instances of more than 1000 known and de novo TF motif in human genome (hg19) derived from encode-motif [25]. Then we calculated frequency of overlap between the candidate enhancer with them, respectively.

Epigenetic Activity. We used Coefficient of variation (CV, defined as the ratio of the standard deviation to the mean) and reads per kilobase million (RPKM) of 5 histone modifications (H3K4me1, H3K4me3, H3K27ac, H3K27me3, H3K36me3) across 98 tissue-/cell-types from NIH Roadmap [8].

Disease features. We calculated frequency of overlap between the candidate enhancer GWAS disease SNP [26];

Recurrent features. We calculated frequency of overlap between the candidate enhancer with 6 classes of recurrent regulatory regions (Transcription factor binding peak; DNase I hypersensitive sites; Segway/ChromHMM predicted enhancers; Enhancer distal regulatory modules; site; COSMIC recurrent regulatory variants) which obtained from FunSeq2 [23];

Feature selection and outlier removal

Feature selection and outlier removal were employed to achieve the best performance. The optimal feature set was selected depend on the largest area under the receiver operating characteristic curve (ROC-AUC) value as described in previous study [27]. Briefly, the confidence of each feature was measured by p values based on Wilcoxon rank sum test. Then we reduced one feature with the largest p-value at a time iteratively and evaluated the performance of classifier based on the mean ROC-AUC of 5-fold cross-validation (5FCV). Further, we remove outlier based on proximities to all other cases within each class as described by Yang *et al.* [28]. For each class, 5% samples with the smallest proximities to all other cases were removed.

IPEV classifier

At last, we built a random forest classifier with 1000 trees, each of which was constructed by randomly selecting the same number of negative pairs as in the positive set, using R package *randomForest*.

Robustness

To be evaluate whether and how much the disproportionate samples affected the model, we firstly divided the variants into two set based on whether the variant were derived from cancers. The corresponding benign variants were selected randomly as the same size of positive set. Using the new sets of variants, we trained cancer-associated (CM) and noncancer-associated model (NCM) with the same features and parameters, and obtain the average ROC-AUC for 5FCV. Furthermore, we also compared the predicted ROC-AUC values and other measures, such as Matthews Correlation Coefficient (MCC), for each model with each test set as input (Table S2). These procedures were repeated 100 times.

Visualization

To help users intuitively understand how variants affect the activity of enhancers, we provided visualizations for the TFs which were both identified as changes of binding by DeepBind [22] and motifbreakR [29]. For results from motifbreakR [29], we only considered the changes of TF bindings with "strong" effect, and defined "gain of motif" if the alleleRef score is greater than alleleAlt score, otherwise, it is "loss of motif".

Results

Overview of the IPEV approach

In this study, we built a random forest classifier (IPEV) based on genetic and epigenetic features to infer the pathogenic enhancers that harbor non-coding genetic variants. A total of 3,865 enhancer-variant pairs was used for training the classification model, including 438 disease-associated positive pairs from the DiseaseEnhancer database (Version 1.0.1)[7] and 3,427 negative pairs between the enhancers without interacting with any disease genes and common SNPs. For each enhancer-variant pair, a 548-dimensional feature vector, encompassing 5 sequence context features, 515 variant-induced transcription factor (TF) binding change scores, 6 conservation scores, 3 features for negative selection, 2 potential TF binding features, 10 features for five histone modifications, 6 recurrent mutation features and 1 feature about enrichment of GWAS disease SNP [26], was generated (Supplementary Table S1). After feature selection and outlier removal, we built a random forest classifier with 1000 trees, each of which was constructed through randomly selecting the same number of negative pairs as in the positive set, using R package *randomForest* (Figure 1A). The detailed procedures are described in Materials and Methods.

Performance and robustness evaluation

The performance of IPEV was evaluated using 5-fold cross validation (5FCV) and leave-one-out cross validation (LOOCV). The average area under the receiver operating characteristic curve (AUC-ROC) was 0.912 for 5-fold cross validation and 0.916 for LOOCV (Figure 2A). Considering the imbalance of positive and negative set, we further evaluated the method using balanced dataset by down-sampling the negative set. In this situation, the average AUC-ROC and MCC was 0.907 and 0.70 for 5FCV (Supplementary Table S2).

Currently, the training set of IPEV consisted of all the known disease associated non-coding variants from DiseaseEnhancer (version 1.0.1). Since the most (61.71%) of the variants were cancer associated, to evaluate whether and how much this affected the model, we conduct two more experiments using the new rule and compare the results with that obtained before (AM, trained using all variants from DiseaseEnhancer version 1.0.1). To be specific, we firstly divided the variants into two set based on whether the variant were derived from cancers. The corresponding benign variants are selected were selected randomly as the same size of positive set. Using the new sets of variants, we trained cancer-associated (CM) and noncancer-associated model (NCM) with the same features and parameters, and obtain the average AUC for 5FCV. The average AUC-ROC values for CM, NCM and AM were 0.9565, 0.9445, 0.9504, respectively. Next, we compared the predicted AUC-ROC values and MCC for each model with each test set as input. We found that despite slightly dampened performance, IPEV AM model performed well with all the three test set, and significantly outperformed all the other models (Figure2B).

IPEV input and interface

As show in Figure 1B, IPEV prompts the user to submit single nucleotide variants in tab-separated values (TSV) format or variant call format (VCF). Optionally, users can consider providing an email address for communicating job status. Once variant data submitted, IPEV performs filtering and maps variants to known enhancers. Next, a large number of genomic and epigenetic features for each enhancer-variant pair will be calculated and the potential impacts of TF bindings mediated by variants will be computed using motifbreakR [29]. Finally, according to these annotation features, IPEV calculates risk scores for all enhancer-variant pairs through the random forest classification model and then prioritizes pathogenic enhancers affected by variants.

Figure 1C shows the inferred result table of the example provided by the website. A summary (top panel) provides basic statistic information after the mapping of variants to known enhancers. The result table displays an interactive spreadsheet, recording the inferred pathogenic enhancers and their associated variants. What's more, the gain or loss of TF bindings mediated by variants and the corresponding visualizations are provided in the column "Changes of TF bindings", which can help users intuitively

understand how variants affect the activity of these enhancers. Any visualization in this table can be enlarged by clicking the images. In Figure 1C, a C>A substitution at chr6:28949469 was inferred to affect a pathogenic enhancer (chr6:28941202-28949600), in which the variant was predicted to create the consensus binding motif for the interferon-regulatory factor (IRF) proteins family and disrupt the binding motif for the transcription factor TFCP2 and Gamma-butyrobetaine dioxygenase (BBOX). The details for the variant affecting TF consensus binding sites are shown in motif sequence logo plots.

In addition, the result table also provides some extra annotation information for each pathogenic enhancer, including the target genes identified by enhancer-promoter looping across tissue-/cell-types, recurrent mutation in cancers, the presence of sensitive or ultra-sensitive regions, the SNP density calculated using data from 1000 Genome project, the coefficient of variation of epigenetic modifications (including H3K4me1 and H3K27ac) across 98 tissue-/cell-types, and the average conservation score (phastCons) [19]. All of the result table and the visualizations could be downloaded as a compressed file.

IPEV could predict deleterious regulatory variants in enhancers

To highlight the usefulness and performance of IPEV, we reanalyzed the experimentally validated pathogenic enhancer-variant pairs in the updated DiseaseEnhancer (version 1.0.2) [7], which were not presented in the training set. Since the majority of enhancer-variant pairs in the updated DiseaseEnhancer were derived from breast cancer, we thus used these variants as independent test set.

We applied IPEV to the 71 single nucleotide variants collected from breast cancer. Among these variants, 3 were filtered out due to the inconsistency between the reference nucleotide and reference genome, and 18 not located within any candidate enhancer regions were also removed. Thus, the remaining 50 variants were further processed by IPEV. As a result, 90% of variants (45/50) were predicted to be located in 26 pathogenic enhancers. Furthermore, IPEV also provides a visual indication of how these variants affect these enhancers by changing TF binding. For example, the down-regulation of ATF7IP mediated by C-T transition at chr12:14410634 (rs11055880) was experimentally validated, but how this variant involved in the regulation was still unknown [30]. IPEV indicated that this variant may lead to gain of FOXK1 motif, recruiting epigenetic modifying complex [31], and thus repressing gene expression. Similarly, applying IPEV to experimentally validated pathogenic enhancer-variant pairs in chronic kidney disease, 71.43% of the variants (10/14) were predicted to be linked with pathogenic enhancers. The above cases indicated that IPEV was highly sensitive and unbiased for disease types.

Conclusions

IPEV is designed to quickly infer the potential pathogenic enhancers by only requiring variants as input. IPEV further provides the target genes of the pathogenic enhancers and intuitive visualizations of which TFs may be affected by the variants. The easy usage of IPEV allows any investigators to easily infer the pathogenic enhancers and interpret the associated variants without additional expertise in computer

programming. It is particular important for biologists and clinicians, as they may reach different insights through hands-on analysis using IPEV rather than working through an intermediary quantitative scientist, due to the different mindset between them.

The next update of IPEV will include more non-coding annotations. A genome browser will be also integrated to support the visualization of multi-omics data, including the chromatin interactions identified by Hi-C, for intuitively deciphering the potential pathogenesis. We believe that IPEV will serve as a powerful platform for inferring the pathogenic enhancers and interpreting the associated non-coding variants.

Declarations

Ethics approval and consent to participate

Not applicable

Consent to publish

Not applicable

Availability of data and materials

Not applicable

Availability and requirements

Project name: IPEV 1.0

Project home page: <http://biocc.hrbmu.edu.cn/IPEV/>

Programming language: JAVA, R and javascripts

License: GNU General Public License

Abbreviations

IPEV: inferring pathogenic enhancer with variant

TFs: transcription factors

GWAS: Genome-wide association study

REMC: Roadmap Epigenomics Mapping Consortium

ENCODE: Encyclopedia of DNA Elements

TFBS: transcription factor binding sites

CV: coefficient of variation

RPKM: reads per kilobase million

ROC-AUC: area under the receiver operating characteristic curve

5FCV: 5-fold cross-validation

LOOCV: leave-one-out cross validation

MCC: Matthews Correlation Coefficient

CM: cancer-associated model

NCM: noncancer-associated model

TSV tab-separated values

VCF variant call format

References

1. Melton, C., et al., *Recurrent somatic mutations in regulatory regions of human cancer genomes*. Nat Genet, 2015. **47**(7): p. 710-6.
2. Forbes, S.A., et al., *COSMIC: somatic cancer genetics at high-resolution*. Nucleic Acids Res, 2017. **45**(D1): p. D777-D783.
3. Fuxman Bass, J.I., et al., *Human gene-centered transcription factor networks for enhancers and disease variants*. Cell, 2015. **161**(3): p. 661-673.
4. Farh, K.K., et al., *Genetic and epigenetic fine mapping of causal autoimmune disease variants*. Nature, 2015. **518**(7539): p. 337-43.
5. Sur, I. and J. Taipale, *The role of enhancers in cancer*. Nat Rev Cancer, 2016. **16**(8): p. 483-93.
6. Herz, H.M., *Enhancer deregulation in cancer and other diseases*. Bioessays, 2016. **38**(10): p. 1003-15.
7. Zhang, G., et al., *DiseaseEnhancer: a resource of human disease-associated enhancer catalog*. Nucleic Acids Res, 2018. **46**(D1): p. D78-D84.
8. Roadmap Epigenomics, C., et al., *Integrative analysis of 111 reference human epigenomes*. Nature, 2015. **518**(7539): p. 317-30.
9. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome*. Nature, 2012. **489**(7414): p. 57-74.
10. Ernst, J. and M. Kellis, *ChromHMM: automating chromatin-state discovery and characterization*. Nat Methods, 2012. **9**(3): p. 215-6.

11. Andersson, R., et al., *An atlas of active enhancers across human cell types and tissues*. Nature, 2014. **507**(7493): p. 455-461.
12. Yao, L., et al., *Inferring regulatory element landscapes and transcription factor networks from cancer methylomes*. Genome Biol, 2015. **16**: p. 105.
13. Teng, L., et al., *4DGenome: a comprehensive database of chromatin interactions*. Bioinformatics, 2015. **31**(15): p. 2560-4.
14. Amberger, J.S., et al., *OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders*. Nucleic Acids Res, 2015. **43**(Database issue): p. D789-98.
15. Pinero, J., et al., *DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes*. Database (Oxford), 2015. **2015**: p. bav028.
16. Rappaport, N., et al., *MalaCards: an integrated compendium for diseases and their annotation*. Database (Oxford), 2013. **2013**: p. bat018.
17. Pletscher-Frankild, S., et al., *DISEASES: text mining and data integration of disease-gene associations*. Methods, 2015. **74**: p. 83-9.
18. Genomes Project, C., et al., *A global reference for human genetic variation*. Nature, 2015. **526**(7571): p. 68-74.
19. Siepel, A., et al., *Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes*. Genome Res, 2005. **15**(8): p. 1034-50.
20. Pollard, K.S., et al., *Detection of nonneutral substitution rates on mammalian phylogenies*. Genome Res, 2010. **20**(1): p. 110-21.
21. Davydov, E.V., et al., *Identifying a high fraction of the human genome to be under selective constraint using GERP++*. PLoS Comput Biol, 2010. **6**(12): p. e1001025.
22. Alipanahi, B., et al., *Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning*. Nat Biotechnol, 2015. **33**(8): p. 831-8.
23. Fu, Y., et al., *FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer*. Genome Biol, 2014. **15**(10): p. 480.
24. Rosenbloom, K.R., et al., *ENCODE data in the UCSC Genome Browser: year 5 update*. Nucleic Acids Res, 2013. **41**(Database issue): p. D56-63.
25. Kheradpour, P. and M. Kellis, *Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments*. Nucleic Acids Res, 2014. **42**(5): p. 2976-87.
26. MacArthur, J., et al., *The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)*. Nucleic Acids Res, 2017. **45**(D1): p. D896-D901.
27. Chen, L., P. Jin, and Z.S. Qin, *DIVAN: accurate identification of non-coding disease-specific risk variants using multi-omics profiles*. Genome Biol, 2016. **17**(1): p. 252.
28. Yang, F., et al., *Using random forest for reliable classification and cost-sensitive learning for medical diagnosis*. BMC Bioinformatics, 2009. **10 Suppl 1**: p. S22.

29. Coetzee, S.G., G.A. Coetzee, and D.J. Hazelett, *motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites*. *Bioinformatics*, 2015. **31**(23): p. 3847-9.

30. Liu, S., et al., *Systematic identification of regulatory variants associated with cancer risk*. *Genome Biol*, 2017. **18**(1): p. 194.

31. Shi, X., D.C. Seldin, and D.J. Garry, *Foxk1 recruits the Sds3 complex and represses gene expression in myogenic progenitors*. *Biochem J*, 2012. **446**(3): p. 349-57.

Figures

Figure 1

Screenshots from IPEV. (A) Schematic presentation of the workflow of IPEV approach. (B) The input fields in the “HOME” page of IPEV. (C) The inferred result table summarized the genomic and epigenetic annotation features, and the visualizations for changes of TF bindings mediated by variants.

Figure 2

Performance and robustness of IPEV. (A) ROC-curves of IPEV for 5-fold and leave-one-out cross validation. (B) Model trained by all variants in training set outperformed all the other models.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryMaterials2.3.docx](#)