Phenotype-specific information improves prediction of functional impact for noncoding variants

Corneliu A. Bodea[1,2,4], Adele A. Mitchell[1], Heiko Runz[1,5], and Shamil R. Sunyaev[2,3,4,5]

[1]Department of Genetics and Pharmacogenomics, MRL, Boston, Massachusetts, USA.

[2]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA.

[3]Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA.

[4]The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.

[5]These authors jointly supervised this work. Correspondence should be addressed to S.R.S. (ssunyaev@rics.bwh.harvard.edu).

Abstract

A myriad of noncoding genetic association signals are now awaiting the identification of causal alleles and their functional interpretation. We introduce the novel computational framework PINES (Phenotype-Informed Noncoding Element Scoring), which evaluates the functional impact of noncoding variants by integrating diverse epigenetic annotations. A unique feature of PINES is that it directs the analysis towards genomic annotations most relevant to phenotypes of interest. We show that PINES identifies functional noncoding variation more accurately than methods that do not use phenotype-specific knowledge. We apply PINES to fine map noncoding alleles at GWAS loci across a range of diseases, and predict new causal risk alleles for Parkinson's disease and inflammatory bowel disease. We also use PINES to confirm several high-penetrance variants implicated in Mendelian traits, as well as variants residing within known enhancer regions. PINES consistently identifies functional variants in fine mapping analyses, dissecting pathogenic loci while avoiding the resource-intensive traditional fine mapping studies. Due to its flexibility and ease of use through a dedicated web portal, PINES provides a powerful *in silico* method to prioritize and fine map functional noncoding variants.

1 Introduction

A growing body of evidence suggests that DNA variants outside of protein-coding regions of the genome (here termed noncoding variants) impact human phenotypes, including the risk for common diseases. Many signals identified by genome-wide association studies (GWAS) point to regulatory regions as key determinants of complex traits. Likewise, some human Mendelian phenotypes such as hair, skin, and eye pigmentation are under the tight control of individual highly penetrant noncoding alleles [1, 2]. Thanks to whole genome sequencing (WGS), our ability to uncover novel noncoding alleles has increased substantially. However, studies of both common and rare phenotypes almost never resolve findings in noncoding regions to individual functional causal SNPs [3]. Functional prioritization of noncoding variants thus holds significant promise to assist in fine-mapping efforts and identify genetic lesions underlying Mendelian diseases.

To better understand the architecture of the noncoding genome, several large-scale efforts, such as ENCODE [4] and the Roadmap Epigenomics Project [5], have aimed to characterize the diverse landscape of histone modifications and DNA accessibility based on a wide range of assays across 127 cell types. Other efforts such as the FANTOM5 project [6] have identified enhancer elements across the genome, and computational tools such as TargetFinder [7] have been developed to link such enhancers to the relevant gene promoters. Databases of chromatin interactions such as 4DGenome [8] are also helpful in identifying potential regulatory elements. Yet, while most genomic regions are by now annotated with a plethora of epigenetic features, the challenge remains to draw meaningful conclusions from these annotations, especially since the data are highly dimensional and many epigenetic features are correlated.

Several approaches have recently been introduced to prioritize potentially functional noncoding variants and address the complexity of the annotation data in a principled manner. These include Eigen [9], GWAVA [10], and CADD [11], as well as population genetics and conservation-inspired models such as PhastCons [12] and INSIGHT [13]. One drawback of models such as GWAVA is that they require a training dataset of both functional and non-functional examples. However, high-quality and particularly experimentally validated training data sets for noncoding variants are still very scarce and incomplete, thus limiting the prediction accuracy. A second drawback is that current scoring methods globally merge annotations across many different cell types, which ignores the observation that regulatory elements often operate in a very cell type-specific manner [14].

We hypothesized that predicting the functional relevance of noncoding variants could benefit from taking into account the vast number of variants that have not yet been annotated as having any function (background variants), and use these as a baseline to search for variants that deviate significantly from this background. Such an approach falls under the category of PU (Positive and Unlabeled example) learning, or one-class classification, and is a well-studied machine learning method in the field of outlier detection [15, 16]. We further hypothesized that for identifying noncoding variants of potential relevance to a specific phenotype, integration of prior biological knowledge around the phenotype should increase prediction reliability. This prior knowledge can consist of relevant cell types, but also of relevant genes or pathways. To the best of our knowledge, no current method has the ability to integrate general phenotype-relevant knowledge into the scoring procedure in such a principled and flexible manner.

Here we introduce the Phenotype-Informed Noncoding Element Scoring (PINES) framework. PINES uses an unsupervised approach to systematically assess the functional significance of noncoding SNVs and indels, and allows to customize the search towards annotations considered as of highest relevance to a phenotype of interest. We apply PINES to *in silico* fine map noncoding variants at GWAS loci and predict novel causative SNPs at assumed promoter and enhancer regions for several common traits and diseases, including Parkinson's disease, inflammatory bowel disease (IBD), multiple sclerosis, and blood lipid levels among others. We further show that PINES can be applied to assess a potential functional relevance of noncoding variants on Mendelian phenotypes. With its ease of use via a customizable web server (http://genetics.bwh.harvard.edu/pines/) we expect PINES to become a valuable resource for the interpretation of the noncoding human genome.

2 Results

2.1 PINES integrates epigenetic information to score non-coding variants in customizable queries

Under a default setting, PINES integrates data from each of the 127 cell types analyzed in the Roadmap and ENCODE projects, specifically information on histone modifications indicative of promoters or enhancers (H3K4me1, H3K4me3, H3K27ac, H3K9ac), DNase I hypersensitive sites, sequence constraint scores (GERP, SiPhy), and ChromHMM chromatin state segmentations [17].

These data are applied to compare individual or a set of user-defined SNPs relative to the genomic background. Users may either apply PINES in a default mode, or pre-define sets of reference variants that are known to be linked to phenotypes of interest, or tissues of specific relevance to their respective scientific question. This allows that, if available, prior knowledge can be taken into account during the subsequent scoring process (see Figure 1 and Methods for details).

An important feature of PINES is its ability to incorporate prior biological information into the scoring procedure. When no prior knowledge is provided or available, equal weights are assigned to all features to perform an undirected scoring of variants. When prior information is provided (for instance a list of lead SNPs identified as significantly associated with a trait of interest through GWAS), PINES searches for annotations that are enriched in the provided dataset in order to learn which annotations are most relevant for the phenotype under consideration. Alternatively, users can manually specify the most phenotypically relevant tissue, and all annotations relevant to that tissue will be up-weighted by the scoring procedure. Prior information can additionally be specified in the form of separate annotations. By relying on an angle-based distance from the vector of the maximal possible annotation load in a de-correlated annotation space, PINES addresses both the correlation structure as well as the high dimensionality of epigenetic annotation data (see also Methods).

2.2 PINES predicts causal noncoding variants with high confidence

We illustrate the potential of PINES to fine map noncoding genomic regions implicated in Mendelian traits, as well as likely causal noncoding variants from GWAS loci. In order to test whether PINES correctly identifies and prioritizes functional noncoding SNPs, we applied the algorithm to *in silico* fine-map 20kb regions around seven noncoding alleles whose regulatory impact on a nearby gene had been confirmed experimentally: rs12350739 [18], rs356168 [19], rs6801957 [20], rs12821256 [21], rs138557689 [22], rs2473307 [23], and rs227727 [24] (Figure 2 and Supplementary Table 1). Such well-studied noncoding variants, although still rare in the current literature, provide an optimal testbed for scoring methods. Weighted PINES was used when genome-wide significant SNPs were identified by GWAS for the trait under consideration (see Methods). In all cases, PINES scores peak around the experimentally confirmed functional variants, and PINES assigns these causal variants the lowest p-values. Notably, comparative analyses using these regions demonstrate that PINES outcompetes GWAVA, CADD or Eigen in correctly identifying causal noncoding variants (see Supplementary Figures 1-7).

2.3 PINES detects signal at fine mapped GWAS risk loci

We next aimed to evaluate the ability of PINES to predict causal noncoding variants at GWAS loci across a range of conditions. As a first example for this, we extracted 3,625 candidate causal noncoding variants reported by a large statistical fine mapping study spanning 40 autoimmune diseases [25]. This study used densely-mapped genotyping data and the observed pattern of association at the LD locus to estimate each SNP's probability of being a causal variant. Since this collection aggregates multiple immune-related phenotypes we applied weighted PINES scores, where all annotations corresponding to immune cells were equally up-weighted (see Methods). We compared the collective PINES signal on these 3,625 variants with the results obtained on an additional set of 30,000 background variants that we randomly selected across the genome. Q-Q plots detailing the results highlight a well-calibrated PINES null distribution and clear signal on the fine mapped variant set (Supplementary Figure 8). A comparison of the weighted PINES results with GWAVA, Eigen, and CADD show that PINES delivers the best predictive performance (Figure 3 first panel).

Next, we assessed the performance of PINES to nominate causal alleles from loci associated with individual common traits and diseases, including non-immune phenotypes. We extracted high-confidence (posterior probability ≥ 0.5) fine mapped candidate causal noncoding variants from [25] related to multiple sclerosis (15 variants), celiac disease (26 variants), inflammatory bowel disease (29 variants), and blood lipid levels (19 variants). We determined weighted and unweighted PINES scores for each of these variants, and compared the outcomes of PINES to those of GWAVA, Eigen and CADD. Phenotype-based annotation weights were automatically assigned based on GWAS data ([26, 27, 28, 29, 30, 31]). PINES consistently delivered the highest AUROC values, with up to 12% improvement over GWAVA, Eigen, or CADD when running PINES in the phenotype-weighted mode (Figure 3).

2.4 PINES detects signal at expression-modulating variants identified in a multiplexed reporter assay

To test the performance of PINES to detect functional noncoding variants en masse, we used 230 variants that were found to directly affect gene expression in a massively parallel reporter assay (MPRA) [32]. For the selection of variants we used an FDR cutoff (Benjamini-Hochberg) of 1%. MPRA is an extension of the traditional reporter gene setup, whereby the use of unique barcodes in the 3' UTR of the reporter differentiate expression of individual oligos and thus allow for testing of many different sequences simultaneously. Since most variants identified through this approach have not yet been linked to individual phenotypes, we performed an unweighted PINES analysis. PINES delivered the highest AUROC values, with up to 43% improvement over GWAVA, Eigen, or CADD (Figure 4).

2.5 PINES detects functional evidence for variants residing in FANTOM5 enhancers

We next tested the power of PINES to correctly prioritize 9,000 variants residing within enhancers that have been identified by the FANTOM5 project through cap analysis of gene expression [6]. As expected, PINES correctly assigned noncoding variants residing in these regions the highest relevance scores relative to genomic background variation. Importantly, in doing so PINES outscored GWAVA, Eigen, and CADD considerably, as demonstrated by a Wilcoxon signed rank test between the weighted PINES results and those of all other methods, which delivered p-values strictly below 10^{-60} (Figure 5).

2.6 PINES predicts novel causal variants for Parkinson's disease and IBD through fine mapping of GWAS loci

We next tested whether PINES can be applied to predict novel functional noncoding SNPs from GWAS loci. For this, we applied PINES to all GWAS loci associated in recent meta-analyses at genome-wide significance with Parkinson's disease [33] and IBD [29]. We concentrated on those loci where all SNPs in LD of $R^2 \geq 0.4$ to the GWAS lead SNP were either intronic or intergenic. For both Parkinson's disease and IBD we used PINES to determine enrichment-based phenotype-specific weights from the complete set of significantly disease-associated GWAS SNPs. We then ran weighted PINES to fine map the most likely causal SNP across the 12 selected Parkinson's disease loci and 19 selected IBD loci. With this approach, PINES distinguished 16 novel noncoding alleles from the background that can be assumed with a high likelihood of being causal for conferring risk for Parkinson's disease (rs10878226, rs3756063, rs2301134, rs36121867, rs1954874, rs9275373, rs117896735) and IBD (rs35493230, rs2187892, rs4672507, rs4845604, rs2019262, rs10489630, rs12622128, rs55776317, rs7685642)(Figure 6 and Supplementary Tables 2 and 3).

3 Discussion

The field of human genetics has accumulated thousands of linkage and association signals. The focus is now rapidly shifting towards the identification of functional DNA variants underlying these signals, biological interpretation of their roles, and generation of mechanistic hypothesis of disease etiology. However, the notion of biological function of an allele is diffuse. Genetically mediated phenotypic presentations are usually limited to a specific organ system, tissue or even cell type. Some of them are pleiotropic and affect several systems, but very few represent truly systemic disorders or traits that impact every cell in the body. This suggests that, from the genetic perspective, the notion of function only makes sense in the specific phenotypic context defined by cell type, developmental stage, and stimulus response. This is especially true for regulatory variants involved in transcriptional control. A number of recent studies showed that genetic association signals are enriched in putative regulatory elements, and that this enrichment is highly cell type-specific [25]. However, many experimental and almost all computational approaches to probe the functional effects of allelic variants are agnostic about the context of the phenotypic presentation.

Functional genomics is now actively embracing the multitude of contexts, starting from cell type variability in epigenetic annotations [5]. PINES leverages this annotation richness and attempts to predict the actual functional effect in the most relevant context rather than in the abstract framework of ubiquitous functional relevance. We note that simply restricting the analysis to the most relevant cell type is not the optimal approach. From purely statistical perspective, noisy correlated data provide information and should not be completely neglected. More importantly, from the biological perspective, many alleles are pleiotropic and many phenotypes are influenced by different biological processes in different organs, tissues and cells. For example, risk of myocardial infarction is partly influenced by blood lipids, but many genetic contributions are unrelated to blood lipid levels and are likely mediated by the vascular effects. All autoimmune diseases are influenced by the adaptive immune system, but individual conditions are limited to specific organs. PINES addresses this complexity to some degree through its customizable weighting of annotations. Additionally, many cell types that are relevant to a phenotype are currently not represented in the ENCODE and Roadmap datasets. The ability of PINES to leverage information from related cell types and tissues enables the analysis of noncoding variants even for such phenotypes. Finally, the noncoding genome has been consistently linked to human phenotypes through our knowledge of conservation, GWAS peak localization, and eQTLs, yet so far only few noncoding loci have been experimentally validated. This lack of unambiguously-defined functional noncoding loci makes the unsupervised approach used by PINES very versatile.

PINES can be easily queried through a web server at http://genetics.bwh.harvard.edu/pines/ in a similar manner to PolyPhen. This portal allows for scoring of noncoding SNVs based on user-defined weighting schemes, making PINES immediately applicable across a wide range of phenotypes. In addition, since the web server performs all data processing, users can query PINES with minimal computational overhead. PINES allows for the addition of epigenetic annotations as they become available without requiring significant changes to the underlying statistical model or software implementation. Due to this ease of upgrading the underlying annotation database, we aim PINES to become an always-up-to-date resource for the scientific community.

In conclusion, PINES' ability to take advantage of a wide range of prior biological information allows it to improve on the predictive power of other methods, and to provide an enhanced prioritization of phenotype-relevant variants. PINES avoids biases stemming from inaccurate labeling of training datasets, and benefits from increased power when prior information is available to direct analyses towards relevant annotations. There is a great need for such methods since identification of regulatory activity specific to a subgroup of cell types or tissues can greatly increase our understanding of disease mechanisms. We have shown that PINES can assist in identifying functional noncoding variants in fine mapping analyses, both for complex disease and Mendelian traits, without requiring the significant resource expenditure involved in a traditional fine mapping study.

4 Acknowledgements

We would like to thank Dr. Ivan Adzhubey for help with setting up the PINES web server. We also acknowledge Dr. Alex Bloemendal for the suggestion to reduce the noise in the annotation covariance matrix by relying on a low-rank approximation of this matrix.

5 Methods

5.1 Annotation sources

PINES uses a wide range of annotations as part of the scoring algorithm. Open chromatin and histone modifications for 127 cell types and tissues were obtained from ENCODE and Roadmap Epigenomics ChIP-seq and DNase-seq peak sets. In order to capture combinatorial interactions between different chromatin marks in their spatial context, we used ChromHMM chromatin state segmentations from Roadmap Epigenomics computed via the standard 15-state HMM model. Chromatin interaction data from a variety of assays (3C, 4C-Seq, 5C, Hi-C, ChIA-PET, Capture-C) were obtained from the 4DGenome database [8]. Additional DNaseI regions inferred via HMM from ENCODE and Roadmap Epigenomics data were obtained from the Reg2Map database. Conservation was evaluated via GERP [34] and SiPhy [35].

Noncoding background variants were selected randomly across the genome, and we used the ClinVar database [36] and GWAS Catalog [37] to ensure that no overlap exists with known functional loci. For the analysis in Figure 3 we used GWAS studies of IBD [28, 29], celiac disease [27], blood lipid levels [30, 31], and multiple sclerosis [26] to determine enrichment-based weights for weighted PINES. We then used fine mapped variants on the corresponding phenotypes from [25] as our test set. All immune-related fine mapped variants in [25] with posterior probability ≥ 0.2 were used to generate the first panel in Figure 3. A list of FANTOM5 enhancers [6] was used to create Figure 5. The analysis of all noncoding Parkinson's disease and IBD loci (Figure 6) was based on regions identified in [33] and [29].

5.2 Working with a high-dimensional correlated annotation space

Individual variants are assigned a score of 0 or 1 for each of the annotations referenced above. In particular, each variant is characterized by a vector of length 639 composed as follows:

- Presence or absence of H3K4me1, H3K4me3, H3K27ac, H3K9ac, and DNase annotations for each of the 127 epigenomes (635 values).
- Presence or absence of a conserved region as predicted by GERP and SiPhy (2 values).
- Presence or absence of a DHS region as predicted by the ChromHMM 15 state model trained on all epigenomes (1 value).
- Presence or absence of a region involved in chromatin interactions with other regions as reported in the 4DGenome database (1 value).

In our annotation dataset, each variant is thus characterized by a vector of 635 cell type-specific scores and 4 cell type-independent scores. The joint distribution of this vector is difficult to ascertain explicitly due to its complex correlation structure; there are few outlier detection techniques that are robust to correlated data. One such approach is the one-class support vector machine (SVM) [38, 39, 40], which fits a hyperplane or hypersphere to the data in an attempt to isolate outlying points. One-class SVMs however suffer from a few disadvantages, such as difficulty in choosing tuning parameters, and the inability to add user-specified feature weights.

The alternative approach used in PINES is based on angular distances in a de-correlated annotation space. Let $\mathbf{X}$ be the 639-dimensional matrix of annotations with covariance matrix $\boldsymbol{\Sigma}$ and mean vector μ, and let $\mathbf{W}$ be a diagonal matrix of annotation weights (which in an unweighted analysis is the identity matrix). Since $\boldsymbol{\Sigma}$ is a noisy estimate of the true correlation structure of $\mathbf{X}$, we perform the spectral decomposition $\boldsymbol{\Sigma} = \sum_{i=1}^{639} \lambda_i \mathbf{u_i} \mathbf{u_i}^T$, and compute a low-rank approximation of the estimated covariance matrix based on the first 30 eigenvectors (chosen by visual inspection of the scree plot): $\hat{\boldsymbol{\Sigma}} = \sum_{i=1}^{30} \lambda_i \mathbf{u_i} \mathbf{u_i}^T$. The matrix $\hat{\boldsymbol{\Sigma}}^{-1}$ is obtained via the Moore-Penrose pseudo-inverse and is used to project annotation vectors corresponding to individual variants into a decorrelated annotation space via a Cholesky transformation: $\hat{\boldsymbol{\Sigma}}^{-1} = \sum_{i=1}^{30} \frac{1}{\lambda_i} \mathbf{u_i} \mathbf{u_i}^T$. If $\mathbf{x}$ is a vector of annotations, then the length of the vector projected into the decorrelated space is $\sqrt{(\mathbf{x}-\mu)^T \mathbf{W} \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{W} (\mathbf{x}-\mu))}$, which corresponds to the reweighed Mahalanobis distance to the mean vector μ [41]. The cosine of the angle between the projections of two vectors $\mathbf{x}$ and $\mathbf{y}$ into the decorrelated annotation space is given by

$$\frac{(\mathbf{x}-\mu)^T \mathbf{W} \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{W} (\mathbf{y}-\mu))}{\sqrt{(\mathbf{x}-\mu)^T \mathbf{W} \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{W} (\mathbf{x}-\mu))} \cdot \sqrt{(\mathbf{y}-\mu)^T \mathbf{W} \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{W} (\mathbf{y}-\mu))}}$$

which corresponds to the correlation between the projections of $\mathbf{x}$ and $\mathbf{y}$. In PINES we are specifically interested in the angle between the projection of a variant of interest and the projection of the all-1 annotation vector $\mathbb{1}$. This vector is significant since it provides the direction of a point with maximal annotation load, and thus the greatest evidence for functionality. Following the approach presented in [42], we additionally scale this angle by the length of the projected $\mathbf{x}$ vector, resulting in the following PINES score:

$$\text{PINES}(\mathbf{x}) = \frac{\text{acos}\left[\frac{(\mathbf{x}-\mu)^T \mathbf{W} \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{W} (\mathbb{1}-\mu))}{\sqrt{(\mathbf{x}-\mu)^T \mathbf{W} \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{W} (\mathbf{x}-\mu))} \cdot \sqrt{(\mathbb{1}-\mu)^T \mathbf{W} \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{W} (\mathbb{1}-\mu))}}\right]}{\sqrt{(\mathbf{x}-\mu)^T \mathbf{W} \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{W} (\mathbf{x}-\mu))} \cdot \sqrt{(\mathbb{1}-\mu)^T \mathbf{W} \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{W} (\mathbb{1}-\mu))}}$$

Smaller scores indicate greater evidence for functionality. Supplementary Figure 9 presents the relationship between weighted PINES scores and annotation load in the original highly correlated space on a simulated example, as well as the effect of introducing weights in the model. For this simulation we generated 20 correlated Bernoulli random variables (representing the epigenetic annotations). The diagonal weight matrix $\mathbf{W}$ was constructed to assign weight 4 to one of the annotations and weight 1 to the remaining features. The plot shows the relationship between annotation count and PINES score, as well as shift in the score profile of observations that have the up-weighted annotation (red points). Finally, to determine the significance level of a given score we compute an empirical p-value based on a large set of background variants. These background variants represent 150,000 common variant sites (since rare variant sites can potentially harbor more penetrant mutations) selected randomly across the genome and that are not represented in ClinVar or the GWAS Catalogue.

5.3 Background variants and the null distribution

PINES makes use of a set of background SNPs against which to compare the score for new variants. We randomly selected 150,000 common variants across the genome to serve as background, none of which have been previously tied to phenotypes. Based on this background, PINES reports a one-sided p-value for each input variant, but the scores reported by CADD, GWAVA, and Eigen do no have an absolute unit of meaning and are thus not directly comparable. To enable this comparison we use the collection of background variants to determine a null distribution for each scoring method, and transform the raw CADD, GWAVA, and Eigen scores into empirical one-sided p-values based on their respective null distribution. This approach is similar to the one used to compute scaled CADD scores by transforming raw into rank-based scores [11]. Having performed this normalization, we can compare the results of PINES, CADD, GWAVA, and Eigen directly.

5.4 Choice of feature weights

Weighting of features can be performed manually, as was the case for the pigmentation variants presented in Figure 2, the first panel in Figure 3, and the simulation in Supplementary Figure 9, by setting the weight for biologically relevant annotations to a user-specified constant. Alternatively, when GWAS peaks are available, the weights used by PINES to differentiate between the different cell type-specific annotations are automatically computed based on the enrichment of each annotation across the GWAS loci. Enrichment within GWAS peaks for each annotation is used to set weights for the Parkinson's disease, QRS prolongation, schizophrenia, and cleft lip and cleft palate variants presented in Figure 2. Such enrichment-based approaches are frequently used in predicting cell types contributing to specific phenotypes when GWAS or fine mapping data is available [25], with highly enriched annotations indicating potentially relevant cell types and disease mechanisms. In particular, we used the corresponding -log10(enrichment p-value) as weight for every annotation, although different functional forms are possible. Regardless of whether a manual or enrichment-based weighting is employed to construct the matrix $\mathbf{W}$, no annotation will be completely excluded from the model. For example in a study of pigmentation, the objective is for variants that have melanocyte-related annotations as well as exhibit evidence of functionality in other cell types to receive more significant scores than variants that only have melanocyte-related annotations. Another reason to rely on data from multiple cell types, even when the phenotypic effect of variants is limited to a single, well characterized cell type, is to gain statistical power from accumulating noisy correlated datasets.

5.5 Other methods

A few approaches have been recently proposed to score noncoding regions and address the complexity of the annotation data in a principled manner. The Genome-Wide Annotation of Variants method (GWAVA) [10] aims to predict the impact of noncoding genetic variants based on a random forest classifier, using variants reported in the Human GeneMutation Database (HGMD) as deleterious training data, and common SNPs from the 1000 Genomes Project as benign examples. The CADD approach [11] is based on the premise that harmful mutations are edged out of the gene pool over time via natural selection and that variation that has not been selected against is thus less likely to be deleterious. Notable for CADD is that it uses a dataset of simulated mutations for training, which is then compared to observed variants. A score of deleteriousness is assigned to every possible SNP in the human genome. One of the most recent methods, Eigen [9], is an unsupervised scoring framework that uses the eigen-decomposition of the covariance matrix associated with a collection of functional annotations to compute variant scores representing weighted sums of individual annotations.

5.6 Code availability

PINES can be queried through a web interface at http://genetics.bwh.harvard.edu/pines/. The source code and corresponding annotation data will also be available for download on this website, allowing users to customize and run PINES on their own system.

References

[1] Mijke Visser, Manfred Kayser, and Robert-Jan Palstra. Herc2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the oca2 promoter. *Genome research*, 22(3):446–455, 2012.

[2] Hans Eiberg, Jesper Troelsen, Mette Nielsen, Annemette Mikkelsen, Jonas Mengel-From, Klaus W Kjaer, and Lars Hansen. Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the herc2 gene inhibiting oca2 expression. *Human genetics*, 123(2):177–187, 2008.

[3] Dimitri J Stavropoulos, Daniele Merico, Rebekah Jobling, Sarah Bowdin, Nasim Monfared, Bhooma Thiruvahindrapuram, Thomas Nalpathamkalam, Giovanna Pellecchia, Ryan KC Yuen, Michael J Szego, et al. Whole-genome sequencing expands diagnostic utility and improves clinical management in paediatric medicine. *npj Genomic Medicine*, 1:15012, 2016.

[4] ENCODE Project Consortium et al. The encode (encyclopedia of dna elements) project. *Science*, 306(5696):636–640, 2004.

[5] Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.

[6] Erik Arner, Carsten O Daub, Kristoffer Vitting-Seerup, Robin Andersson, Berit Lilje, Finn Drabløs, Andreas Lennartsson, Michelle Rönnerblad, Olga Hrydziuszko, Morana Vitezic, et al. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science*, 347(6225):1010–1014, 2015.

[7] Sean Whalen, Rebecca M Truty, and Katherine S Pollard. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature genetics*, 48(5):488–496, 2016.

[8] Li Teng, Bing He, Jiahui Wang, and Kai Tan. 4dgenome: a comprehensive database of chromatin interactions. *Bioinformatics*, 31(15):2560–2564, 2015.

[9] Iuliana Ionita-Laza, Kenneth McCallum, Bin Xu, and Joseph D Buxbaum. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature genetics*, 2016.

[10] Graham RS Ritchie, Ian Dunham, Eleftheria Zeggini, and Paul Flicek. Functional annotation of noncoding sequence variants. *Nature methods*, 11(3):294–296, 2014.

[11] Martin Kircher, Daniela M Witten, Preti Jain, Brian J O'Roak, Gregory M Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, 46(3):310, 2014.

[12] Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W Hillier, Stephen Richards, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8):1034–1050, 2005.

[13] Ilan Gronau, Leonardo Arbiza, Jaaved Mohammed, and Adam Siepel. Inference of natural selection from interspersed genomic elements based on polymorphism and divergence. *Molecular biology and evolution*, page mst019, 2013.

[14] Gosia Trynka, Cynthia Sandor, Buhm Han, Han Xu, Barbara E Stranger, X Shirley Liu, and Soumya Raychaudhuri. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature genetics*, 45(2):124–130, 2013.

[15] Bing Liu, Yang Dai, Xiaoli Li, Wee Sun Lee, and Philip S Yu. Building text classifiers using positive and unlabeled examples. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 179–186. IEEE, 2003.

[16] Peng Yang, Xiao-Li Li, Jian-Ping Mei, Chee-Keong Kwoh, and See-Kiong Ng. Positive-unlabeled learning for disease gene identification. *Bioinformatics*, 28(20):2640–2647, 2012.

[17] Jason Ernst and Manolis Kellis. Chromhmm: automating chromatin-state discovery and characterization. *Nature methods*, 9(3):215–216, 2012.

[18] Mijke Visser, Robert-Jan Palstra, and Manfred Kayser. Human skin color is influenced by an intergenic dna polymorphism regulating transcription of the nearby bnc2 pigmentation gene. *Human molecular genetics*, page ddu289, 2014.

[19] Frank Soldner, Yonatan Stelzer, Chikdu S Shivalila, Brian J Abraham, Jeanne C Latourelle, M Inmaculada Barrasa, Johanna Goldmann, Richard H Myers, Richard A Young, and Rudolf Jaenisch. Parkinson-associated risk variant in distal enhancer of α-synuclein modulates target gene expression. *Nature*, 533(7601):95–99, 2016.

[20] Malou van den Boogaard, Scott Smemo, Ozanna Burnicka-Turek, David E Arnolds, Harmen JG van de Werken, Petra Klous, David McKean, Jochen D Muehlschlegel, Julia Moosmann, Okan Toka, et al. A common genetic variant within scn10a modulates cardiac scn5a expression. *The Journal of clinical investigation*, 124(4):1844–1852, 2014.

[21] Catherine A Guenther, Bosiljka Tasic, Liqun Luo, Mary A Bedell, and David M Kingsley. A molecular basis for classic blond hair color in europeans. *Nature genetics*, 46(7):748–752, 2014.

[22] Nevena Cvjetkovic, Lorena Maili, Katelyn S Weymouth, S Shahrukh Hashmi, John B Mulliken, Jacek Topczewski, Ariadne Letra, Qiuping Yuan, Susan H Blanton, Eric C Swindell, et al. Regulatory variant in fzd6 gene contributes to nonsyndromic cleft lip and palate in an african-american family. *Molecular genetics &amp; genomic medicine*, 3(5):440–451, 2015.

[23] William P Gilks, Matthew Hill, Michael Gill, Gary Donohoe, Aiden P Corvin, and Derek W Morris. Functional investigation of a schizophrenia gwas signal at the cdc42 gene. *The World Journal of Biological Psychiatry*, 2012.

[24] Elizabeth J Leslie, Margaret A Taub, Huan Liu, Karyn Meltz Steinberg, Daniel C Koboldt, Qunyuan Zhang, Jenna C Carlson, Jacqueline B Hetmanski, Hang Wang, David E Larson, et al. Identification of functional variants for cleft lip with or without cleft palate in or near pax7, fgfr2, and nog by targeted sequencing of gwas loci. *The American Journal of Human Genetics*, 96(3):397–411, 2015.

[25] Kyle Kai-How Farh, Alexander Marson, Jiang Zhu, Markus Kleinewietfeld, William J Housley, Samantha Beik, Noam Shoresh, Holly Whitton, Russell JH Ryan, Alexander A Shishkin, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539):337–343, 2015.

[26] International Multiple Sclerosis Genetics Consortium, Wellcome Trust Case Control Consortium 2, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, 476(7359):214–219, 2011.

[27] Patrick CA Dubois, Gosia Trynka, Lude Franke, Karen A Hunt, Jihane Romanos, Alessandra Curtotti, Alexandra Zhernakova, Graham AR Heap, Róza Ádány, Arpo Aromaa, et al. Multiple common variants for celiac disease influencing immune gene expression. *Nature genetics*, 42(4):295–302, 2010.

[28] Luke Jostins, Stephan Ripke, Rinse K Weersma, Richard H Duerr, Dermot P McGovern, Ken Y Hui, James C Lee, L Philip Schumm, Yashoda Sharma, Carl A Anderson, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422):119–124, 2012.

[29] Jimmy Z Liu, Suzanne van Sommeren, Hailiang Huang, Siew C Ng, Rudi Alberts, Atsushi Takahashi, Stephan Ripke, James C Lee, Luke Jostins, Tejas Shah, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature genetics*, 47(9):979–986, 2015.

[30] Tanya M Teslovich, Kiran Musunuru, Albert V Smith, Andrew C Edmondson, Ioannis M Stylianou, Masahiro Koseki, James P Pirruccello, Samuli Ripatti, Daniel I Chasman, Cristen J Willer, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707–713, 2010.

[31] Global Lipids Genetics Consortium et al. Discovery and refinement of loci associated with lipid levels. *Nature genetics*, 45(11):1274–1283, 2013.

[32] Ryan Tewhey, Dylan Kotliar, Daniel S Park, Brandon Liu, Sarah Winnicki, Steven K Reilly, Kristian G Andersen, Tarjei S Mikkelsen, Eric S Lander, Stephen F Schaffner, et al. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell*, 165(6):1519–1529, 2016.

[33] Mike A Nalls, Nathan Pankratz, Christina M Lill, Chuong B Do, Dena G Hernandez, Mohamad Saad, Anita L DeStefano, Eleanna Kara, Jose Bras, Manu Sharma, et al. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for parkinson's disease. *Nature genetics*, 46(9):989–993, 2014.

[34] Eugene V Davydov, David L Goode, Marina Sirota, Gregory M Cooper, Arend Sidow, and Serafim Batzoglou. Identifying a high fraction of the human genome to be under selective constraint using gerp++. *PLoS Comput Biol*, 6(12):e1001025, 2010.

[35] Manuel Garber, Mitchell Guttman, Michele Clamp, Michael C Zody, Nir Friedman, and Xiaohui Xie. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*, 25(12):i54–i62, 2009.

[36] Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Jeffrey Hoover, et al. Clinvar: public

archive of interpretations of clinically relevant variants. *Nucleic acids research*, 44(D1):D862–D868, 2016.

[37] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, et al. The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic acids research*, 42(D1):D1001–D1006, 2014.

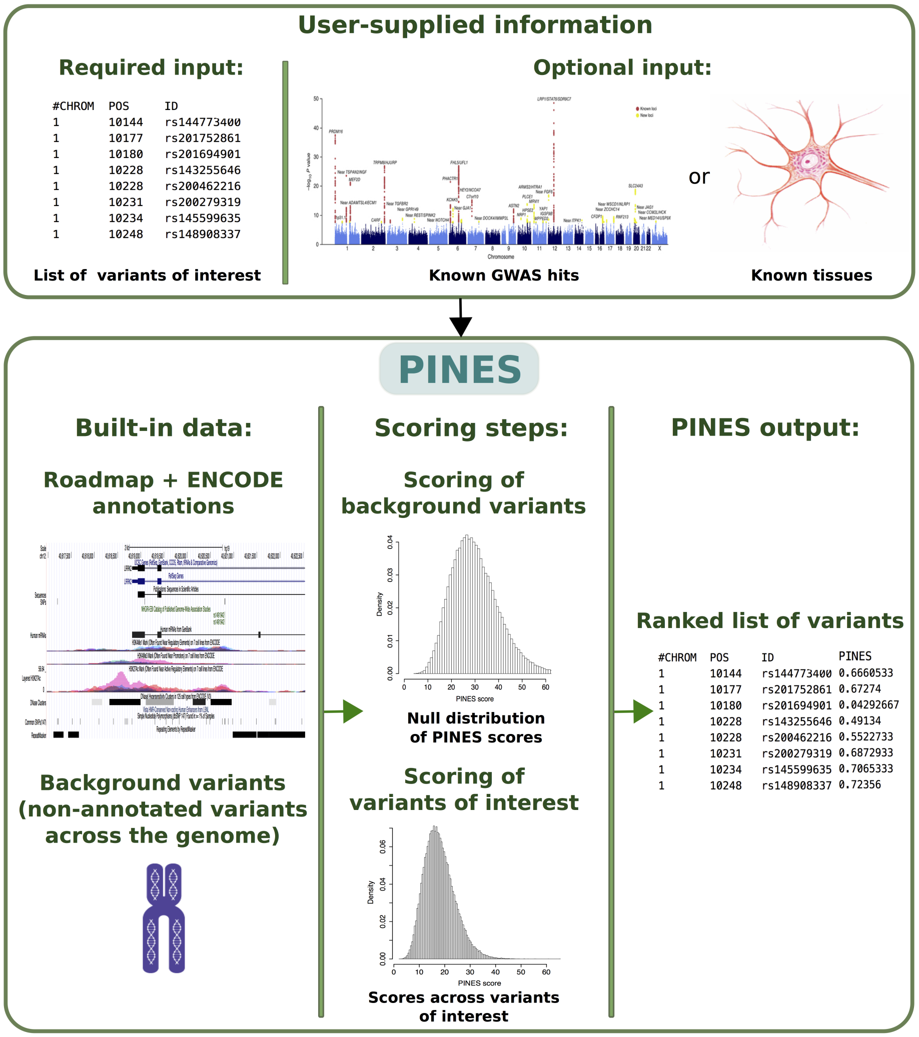
[38] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

[39] Larry M Manevitz and Malik Yousef. One-class svms for document classification. *the Journal of machine Learning research*, 2:139–154, 2002.

[40] David MJ Tax and Robert PW Duin. Support vector data description. *Machine learning*, 54(1):45–66, 2004.

[41] Robert A Greevy, Carlos G Grijalva, Christianne L Roumie, Cole Beck, Adriana M Hung, Harvey J Murff, Xulei Liu, and Marie R Griffin. Reweighted mahalanobis distance matching for cluster-randomized trials with missing data. *Pharmacoepidemiology and drug safety*, 21(S2):148–154, 2012.

[42] Hans-Peter Kriegel, Arthur Zimek, et al. Angle-based outlier detection in high-dimensional data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 444–452. ACM, 2008.


User-supplied information
Required input:
List of variants of interest
Optional input:
or
Known GWAS hits
Known tissues
PINES
Built-in data:
Roadmap + ENCODE annotations
Background variants (non-annotated variants across the genome)
Scoring steps:
Scoring of background variants
Density
PINES score
Null distribution of PINES scores
Scoring of variants of interest
Scores across variants of interest
PINES output:
Ranked list of variants

```
#CHROM  POS      ID
1       10144    rs144773400
1       10177    rs201752861
1       10180    rs201694901
1       10228    rs143255646
1       10228    rs200462216
1       10231    rs200279319
1       10234    rs145599635
1       10248    rs148908337
```

```
#CHROM  POS    ID          PINES
1       10144  rs144773400 0.6660533
1       10177  rs201752861 0.67274
1       10180  rs201694901 0.04292667
1       10228  rs143255646 0.49134
1       10228  rs200462216 0.5522733
1       10231  rs200279319 0.6872933
1       10234  rs145599635 0.7065333
1       10248  rs148908337 0.72356
```

Figure 1: Overview of the PINES framework. PINES aims to systematically predict and rank the functional relevance of noncoding genomic variants. It can either work in a default (”unweighted”) mode and compare user-defined variants against the genomic background. Alternatively, users can customize searches towards annotations considered as of highest relevance to a phenotype of interest, for instance by providing a list of SNPs associated with a disease of interest through GWAS, or by highlighting disease-relevant tissues (”weighted” PINES mode). Scores of genomic background variants serve as an empirical null distribution against which significance levels for each variant of interest are computed and scored in an output file.

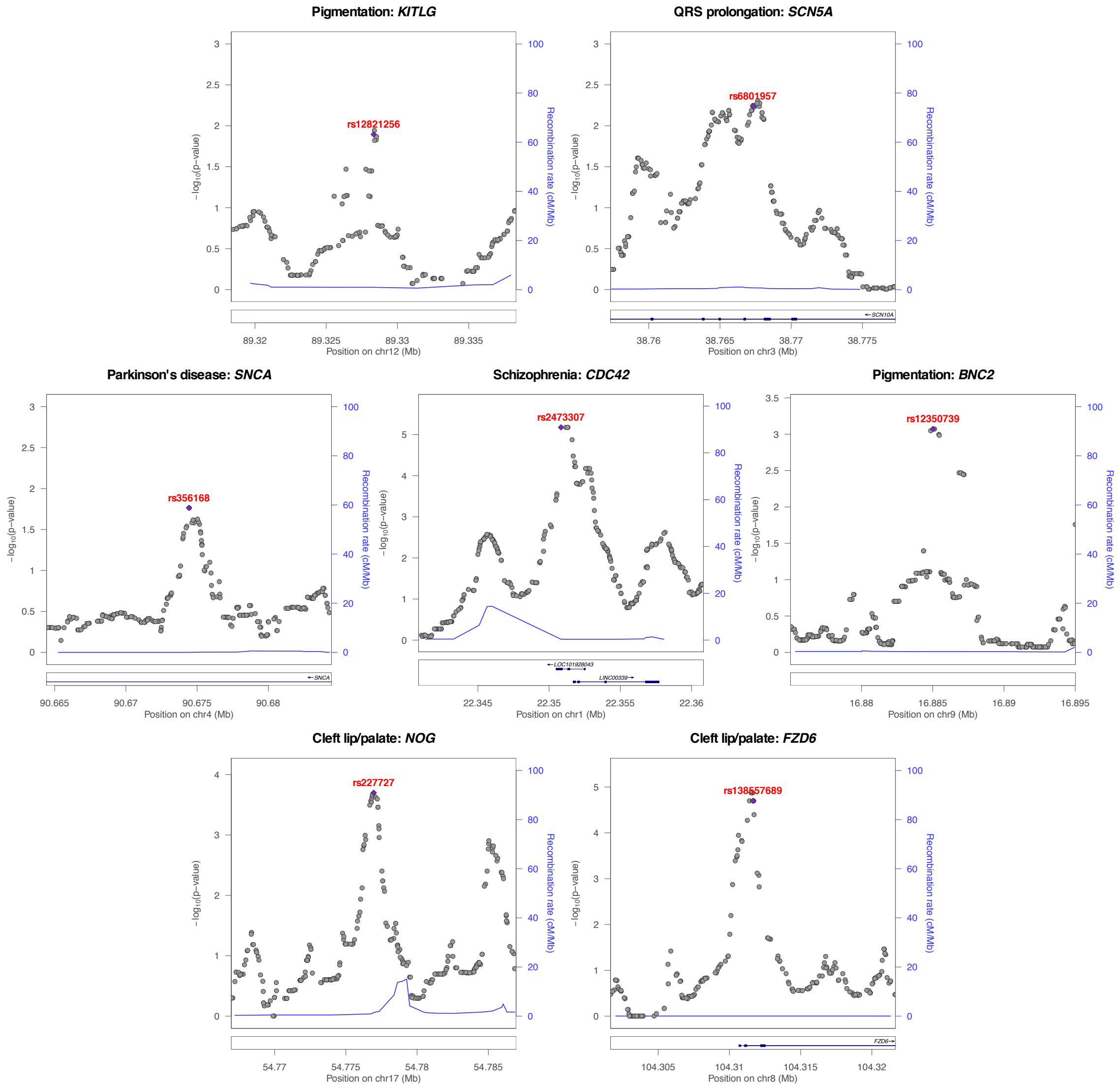


Figure 2: PINES prioritizes experimentally validated functional noncoding variants. We fine map 20kb regions surrounding functional noncoding variants (purple dots) and show that all of the variants validated experimentally as regulating expression of a nearby trait-associated gene are also assigned the highest PINES scores. Supplementary Figures 1-7 show that PINES outperforms existing methods on all loci.

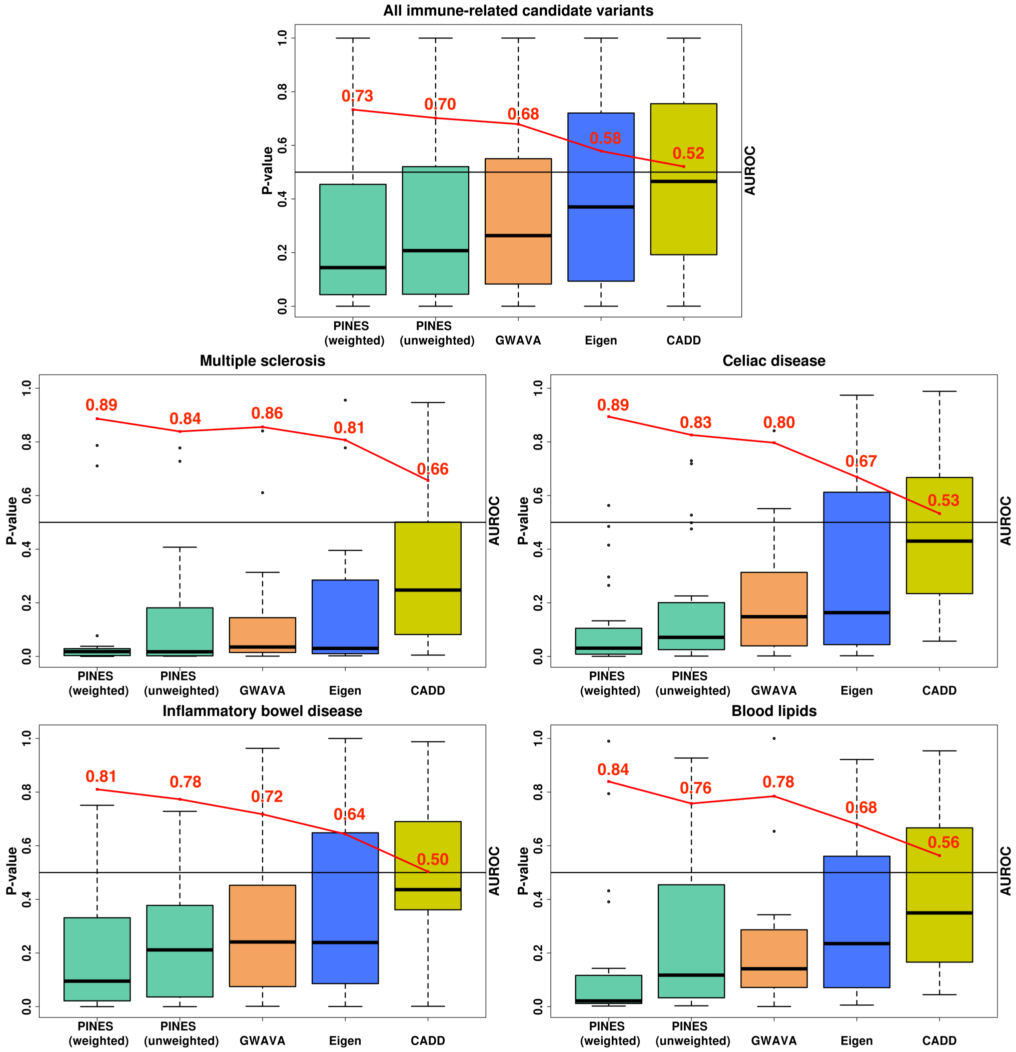


Figure 3: PINES improves statistical power to detect fine mapped variants across common neurologic, immune, and metabolic traits and diseases. AUROC values (red) were computed by selecting 20,000 background variants as negative examples, and the fine mapped variants relevant to each disease as positive examples. PINES achieves the best AUROC values, of to 12% higher than the other methods, based on its inclusion of weights encoding prior disease knowledge (in this case relevant cell types).

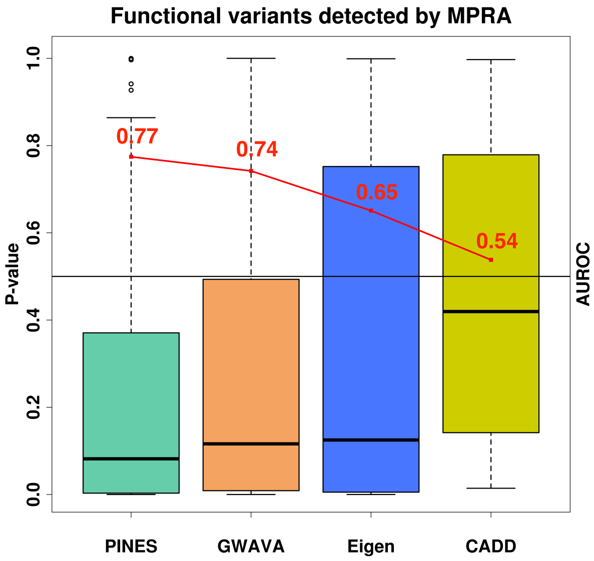


Figure 4: PINES delivers improved statistical power to identify functional noncoding variants detected by a massively parallel reporter assay. The AUROC values (red) were computed by selecting 20,000 background variants as negative examples, and the reported functional variants as positive examples. With AUROC values up to 43% higher than the other methods, PINES outperforms GWAVA, Eigen, and CADD in its ability to detect the functional variants.

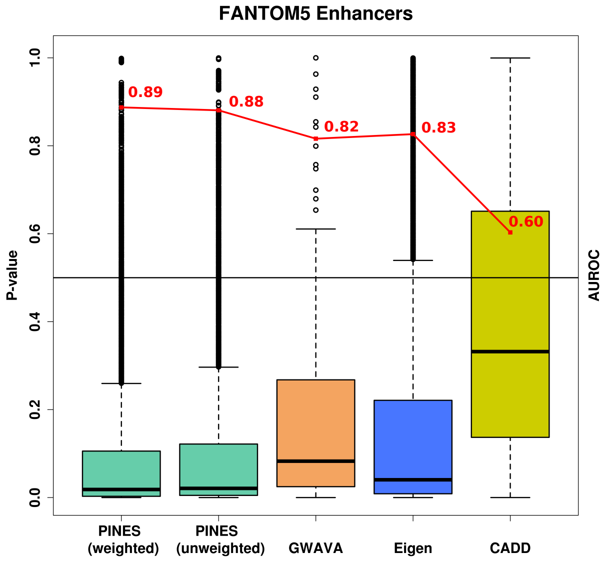


Figure 5: PINES improves the prioritization of variants residing in experimentally validated enhancer regions. The AUROC values (red) were computed by selecting 20,000 background variants as negative examples, and the variants residing in enhancer loci as positive examples. Based on AUROC values, the weighted PINES approach outperforms GWAVA, Eigen, and CADD in its ability to pinpoint enhancer variants. Additionally, testing whether the weighted PINES significance levels are smaller that those of other methods via a Wilcoxon signed rank test delivers p-values that are all below 10^{-60}.

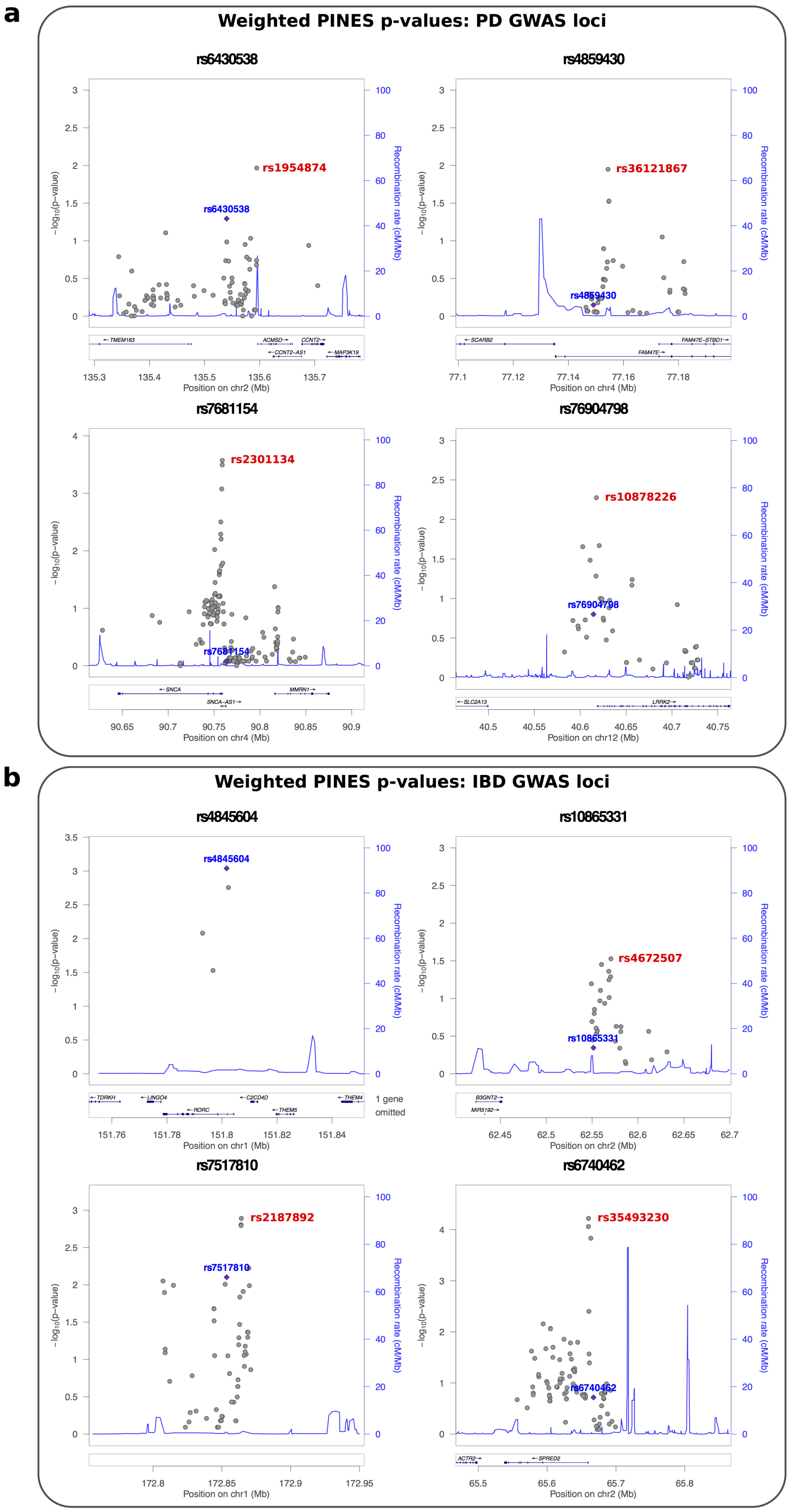


Figure 6: PINES predicts novel noncoding pathogenic variants through fine mapping of Parkinson's disease and IBD GWAS loci. Loci were extracted from [33] and [29]. For each lead SNP, all variants with $LD \geq 0.4$ were selected, and loci were discarded if this list overlapped any coding regions or 3' or 5' UTRs of coding genes. All variants in LD to the lead SNP were scored via weighted PINES. The GWAS lead SNP is marked blue, and the variant predicted as likely causal through PINES fine mapping is marked red. For the rs4845604 locus the GWAS and PINES lead SNP overlaps.