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

Corneliu A. Bodea[1,2,4], Adele A. Mitchell[1], Aaron G. Day-Williams[1], Heiko Runz[1‡*], Shamil R. Sunyaev[2,3,4‡*]

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

‡These authors jointly supervised this work.

* ssunyaev@rics.bwh.harvard.edu or heiko.runz@merck.com

Keywords: noncoding variant, computational functionality prediction

Abstract

Functional characterization of the noncoding genome is essential for the biological understanding of gene regulation and disease. Here, we introduce the computational framework PINES (Phenotype-Informed Noncoding Element Scoring) which evaluates the functional impact of noncoding variants in an unsupervised setting by integrating epigenetic annotations from diverse sources in a phenotype-dependent manner. A unique feature of PINES is that analyses may be customized towards genomic annotations of the highest relevance to phenotypes of interest. We show that PINES identifies functional noncoding variation more accurately than methods that do not use phenotype-weighted knowledge. To demonstrate its versatility, we show how PINES can be applied to systematically distinguish variants in enhancer regions from background variation, to delineate high-penetrance noncoding variants leading to Mendelian phenotypes, or to pinpoint epigenetic signals underlying noncoding alleles at GWAS loci. We further show how PINES can help interpret the results of fine-mapping approaches and predict novel causal alleles from GWAS on Parkinson's and inflammatory bowel disease. Its flexibility and ease of use through a dedicated web portal establish PINES as a powerful new *in silico* method to prioritize and assign functions to noncoding genetic variants.

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) reside in the noncoding genome and are hypothesized to have regulatory functions. 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, causal SNPs [3]. Functional prioritization of noncoding variants thus holds significant promise to help interpret and prioritize variants identified in GWAS 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 the Encyclopedia of DNA Elements [4] and the NIH Roadmap Epigenomics Mapping Consortium [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-PC [9], GWAVA [10], CADD [11], DANN [12], FATHMM-MKL [13], LINSIGHT [14], and deltaSVM [15]. One drawback of models such as GWAVA and deltaSVM 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 many existing scoring methods globally merge annotations across many different cell types, which ignores the observation that regulatory elements can operate in a very cell type-specific manner [16]. In addition, most methods that allow for a cell type-specific analysis, such as deltaSVM, do so by restricting their annotation sources or training set to solely a cell type of interest. This approach has multiple downsides. First, potentially informative annotations from similar cell types are ignored, which can decrease statistical power. Secondly, if the correct cell type is unknown, then focusing the analysis on a pre-defined cell type may lead to results that are not informative of the underlying biology. Finally, a large computational burden is placed on users by requiring them to provide annotations and training data from cell types of interest in addition to running the scoring method.

We hypothesized that in order to identify noncoding variants of relevance to a specific phenotype, a principled integration of prior biological knowledge around that phenotype should increase prediction reliability. Such prior knowledge can consist of information from phenotype-relevant cell types, tissues, genes or pathways. We further 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 [17, 18].

Here we introduce the Phenotype-Informed Noncoding Element Scoring (PINES) framework. PINES enables a phenotype-dependent scoring of noncoding variants that can be customized towards annotations considered as of highest relevance to a phenotype of interest. During the PINES scoring process, such annotations are assigned the highest weights, while annotations of lower relevance to a phenotype are down-weighted. Moreover, different than most previous scoring tools, PINES uses an unsupervised approach to systematically assess the functional significance of noncoding SNVs and indels.

PINES is a purely epigenetic-based noncoding variant prioritization approach and does not incorporate GWAS summary statistics into its scoring procedure, as opposed to methods such as PAINTOR [19] and fGWAS [20] which are meant as GWAS fine-mapping tools. PINES can thus help interpret the consequences of noncoding variation for complex phenotypes that do not yet benefit from extensive GWAS information, or for Mendelian phenotypes, for instance to shortlist rare noncoding variants of putative large effect that could help explain etiologies of rare disease. PINES can also signal the presence of regulatory elements that are not yet tied to particular phenotypes. To the best of our knowledge, no alternative variant interpretation tool can integrate known functional information into the scoring procedure, and also perform well when no prior phenotypic information is available.

We demonstrate the accuracy and versatility of PINES by addressing frequent challenges across several fields, for instance, predicting functional variants at assumed promoter and enhancer regions, prioritizing noncoding variants at GWAS loci, or assessing the relevance of noncoding variation on Mendelian phenotypes. With its ease of use via a customizable web server (http://genetics.bwh.harvard.edu/pines/) and open source code, we expect PINES to become a valuable resource for the interpretation of the noncoding human genome.

Results

PINES integrates epigenetic information to score non-coding variants in phenotype-centered queries

Under a default setting, PINES integrates data from each of the 127 cell types analyzed by the NIH Roadmap Epigenomics Mapping Consortium and the Encyclopedia of DNA Elements (ENCODE) project, 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 [21]. 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 Fig. 1 and Methods for details).

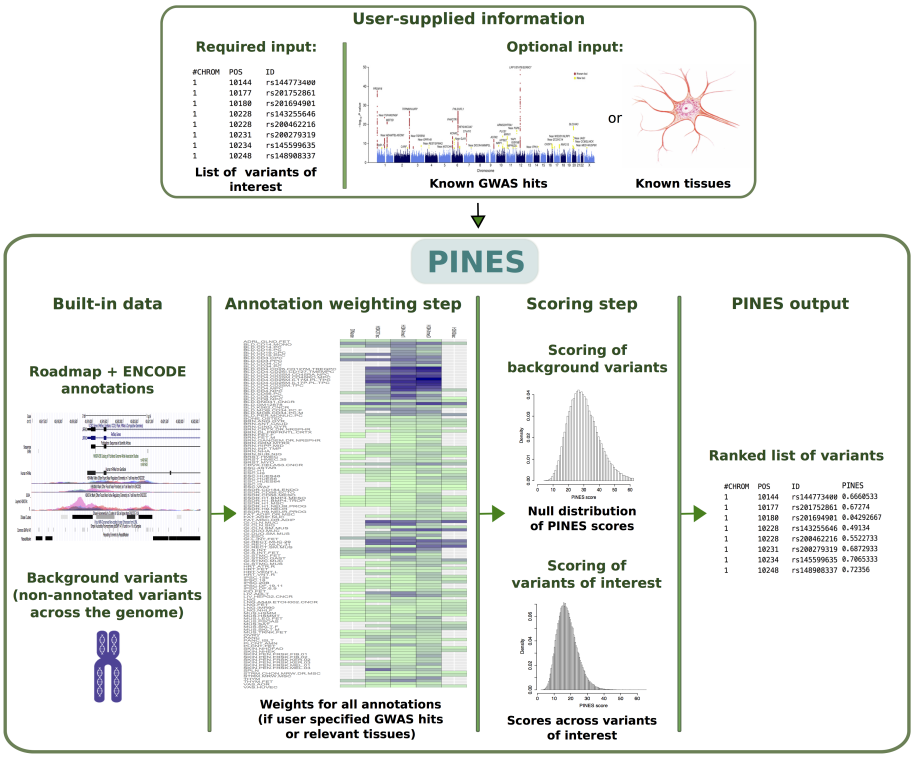

Figure 1. Overview of the PINES framework. PINES aims to systematically predict and rank the functional relevance of noncoding genomic variants. It can either work in a default ("unweighted") mode and compare user-defined variants against the genomic background. Alternatively, users can customize searches 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.

The key methodological advance 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). Prior biological knowledge in the form of lead SNPs or pre-specified tissues and cell types is encoded as a weight matrix in PINES, where each annotation from each cell type receives an individual weight. When GWAS lead SNPs are provided, enrichment within GWAS peaks for each annotation is used to set weights. As an example, the annotation weight matrix that PINES computes based on enrichment of epigenetic marks across inflammatory bowel disease GWAS lead SNPs ([22]) is shown in Supplemental Fig. S9 (see also Methods).

PINES prioritizes causal noncoding variants with high confidence

We first illustrate the potential of PINES to prioritize variants across noncoding genomic regions implicated in Mendelian traits. In order to test whether PINES correctly identifies and prioritizes known functional noncoding SNPs, we applied the algorithm to *in silico* score variants across 20kb regions around six noncoding alleles whose regulatory impact on a nearby gene had been confirmed experimentally: rs12821256 [23], rs6801957 [24], rs356168 [25], rs2473307 [26], rs12350739 [27], and rs227727 [28] (Fig. 2 and Supplemental Table S1). 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-PC in correctly identifying causal noncoding variants (see Supplemental Figs. S1-S6). "Weighted PINES", i.e., when PINES was run in a phenotype-weighted mode, differentiated causal variants from background variants even more clearly than when the software was run in the default ("unweighted") mode, showcasing the additional benefits that come from including phenotype-relevant information into the scoring procedure.

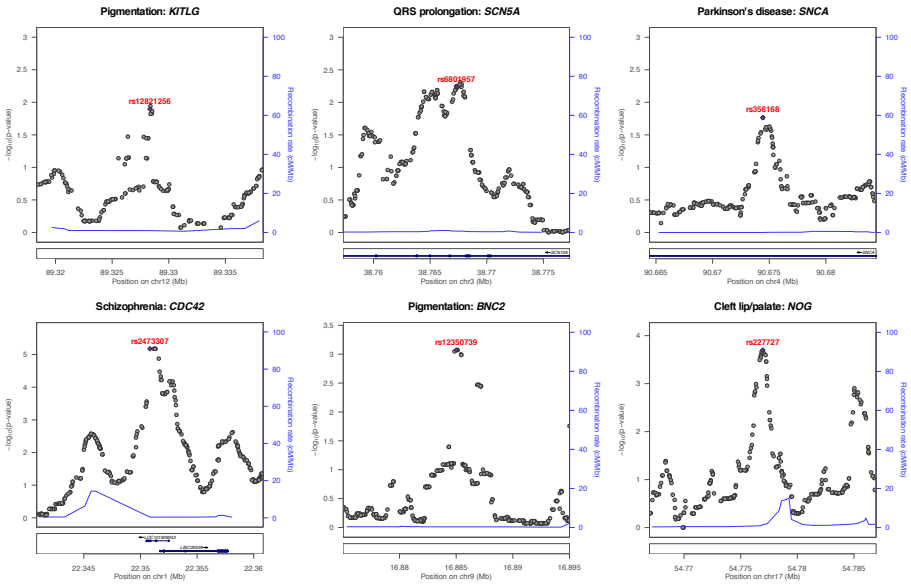

Figure 2. PINES prioritizes experimentally validated functional noncoding variants. We score all variants across 20kb regions surrounding functional noncoding variants (purple dots) and show that all of the variants validated experimentally as regulating expression of a nearby trait-associated gene are also assigned the highest PINES scores. Supplemental Figs. S1-S6 show that PINES outperforms existing methods on all loci.

PINES detects signal at fine mapped GWAS risk loci

We next aimed to evaluate how information on the phenotype can improve 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 [29]. 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 (Supplemental Fig. S7). Comparison of the weighted PINES results with outcomes from GWAVA, Eigen-PC, CADD, LINSIGHT, DANN, and FATHMM-MLK shows that PINES delivers the best predictive performance (Fig. 3 first panel).

Next, we assessed the performance of PINES to nominate causal alleles from loci associated with individual complex traits and diseases, including non-immune phenotypes. We extracted high-confidence (posterior probability ≥ 0.25) fine mapped candidate causal noncoding variants

from [29] related to multiple sclerosis (12 variants), celiac disease (7 variants), inflammatory bowel disease (31 variants), and blood lipid levels (22 variants). We determined weighted and unweighted PINES scores for each of these variants, and compared the outcomes of PINES to those of GWAVA, Eigen-PC, CADD, LINSIGHT, DANN, and FATHMM-MLK. Phenotype-based annotation weights were automatically assigned based on GWAS data ([22, 30–34]). PINES consistently delivered the highest AUROC values, with up to 12% improvement over the other methods when running PINES in the phenotype-weighted mode (Fig. 3). Upweighting of phenotype-relevant annotations enabled weighted PINES to again outcompete unweighted PINES results on all individual phenotypes.

Of note is that we notice a drastic improvement in classifier accuracy when PINES computes annotation weights based on GWAS data in the MS, Celiac disease, IBD, and blood lipids analyses. The panel highlighting all immune-related candidate variants groups together variants across many immune disorders, and thus relies on a uniform manual up-weighting of all cell types involved in immunity. Classification accuracy is increased in this setting as well compared to an unweighted analysis, but an enrichment-based weighting of features is preferred.

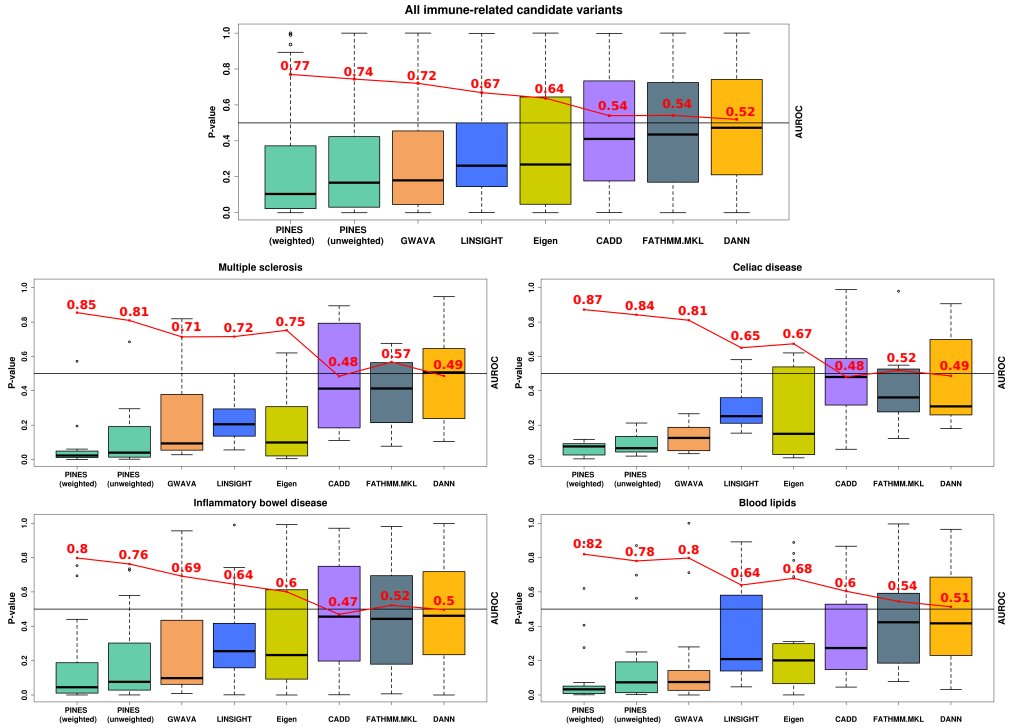

Figure 3. PINES improves statistical power to detect fine mapped variants across common neurologic, immune, and metabolic traits and diseases. AUROC values (red) were computed by selecting 20,000 background variants as negative examples, and the fine mapped variants relevant to each disease as positive examples. Weighted PINES consistently achieves AUROC values above 0.8, up to 78% higher than the other methods, based on its inclusion of weights encoding prior disease knowledge (relevant cell types or GWAS lead SNPs).

Unweighted PINES scores detect signal at expression-modulating variants identified in a multiplexed reporter assay, and for variants residing in FANTOM5 CAGE-defined enhancers

To compare the performance of PINES, Eigen-PC, CADD, and GWAVA to detect functional noncoding variants en masse when no phenotype information is available to guide the scoring, we used 230 variants that were found to directly affect gene expression in a massively parallel reporter assay (MPRA) [35]. 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 54% improvement over other

methods (Fig. 4A).

We next compared the power of PINES, Eigen-PC, CADD, GWAVA, DANN, LINSIGHT, and FATHMM-MKL to correctly prioritize 9,000 variants residing within enhancers that have been identified by the FANTOM5 project through cap analysis of gene expression (CAGE) [6]. In this analysis no phenotype-relevant information is available, so PINES was run in the unweighted mode. PINES correctly assigned noncoding variants residing in these regions the highest relevance scores relative to genomic background variation. Importantly, in doing so unweighted PINES outscored all other methods except for GWAVA, which delivered similar results (Fig. 4B).

These analyses highlight that, even in settings where no phenotype-relevant information is available, none of the methods that we tested were able to perform better than PINES.

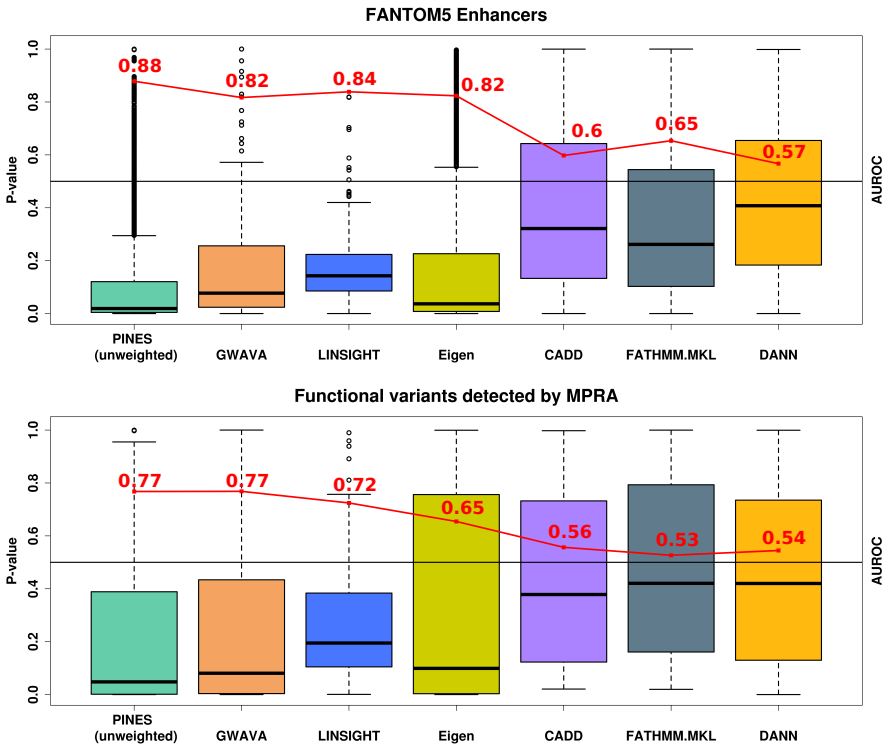

Figure 4. Unweighted PINES scores improve the prioritization of noncoding variants. (A) PINES improves the prioritization of variants residing in experimentally validated enhancer regions. The AUROC values (red) were computed by selecting 20,000 background variants as negative examples, and the variants residing in enhancer loci as positive examples. Based on AUROC values, the unweighted PINES approach performs at least as well as GWAVA, Eigen-PC, CADD, DANN, LINSIGHT, and FATHMM-MKL in its ability to pinpoint enhancer variants. (B) 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 54% higher than the other methods, PINES outperforms GWAVA, Eigen-PC, CADD, DANN, LINSIGHT, and FATHMM-MKL in its ability to detect the functional variants.

PINES identifies variants with highly cell type-specific annotation profiles

A comparison of weighted and unweighted PINES scores on the same set of variants can pinpoint noncoding variants that exhibit epigenetic annotations in a cell type-specific manner. We simulated 5000 background variants and 100 cell type-specific variants across 100 annotations, and randomly selected 10 annotations for up-weighting. For the cell type-specific variants, the up-weighted annotations are drawn from a Bernoulli distribution with p=0.9 and the remaining annotations are drawn from a Bernoulli distribution with $p = 0.1$. For the background variants, all annotations are drawn from a Bernoulli distribution with $p = 0.3$. The resulting cell type-specific variants have low unweighted scores and high weighted scores (Fig. 5A). As a quantitative measure of location in this score space, we compute the angle formed by each variant to the main diagonal. Cell type-specific variants lead to significantly larger angles than background variants (Fig. 5B), and can thus be identified. Applying this approach to lead SNPs from the inflammatory bowel disease GWAS [22] we can detect variants with, for example, GI-specific annotation profiles (Fig. 5C).

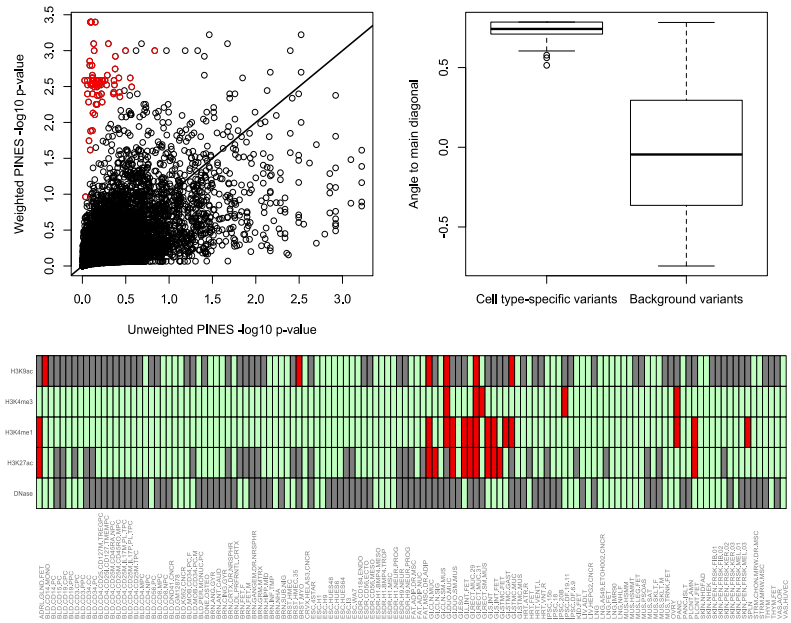

Figure 5. A comparison of weighted and unweighted PINES scores reveals cell type-specific variants. We simulate 5000 background variants (black circles) and 100 cell type-specific variants (red circles) and compute both weighted and unweighted PINES scores. Weighting is based on the annotations that are representative for the cell type-specific variants. The red dots are easily distinguishable from the background variants based on their location in the PINES score space (panel A) as well as the angles they form to the main diagonal (panel B). Applying this approach to variants from the inflammatory bowel disease GWAS in [22] we can detect noncoding variants with putative GI-specific activity. As an example, panel C depicts the annotations present at rs6017342 (red: annotation present, green: annotation absent, gray: missing data), a variant rich in GI-specific annotations that has been implicated by fine mapping of IBD GWAS loci [36].

PINES points to novel causal variants for Parkinson's disease and IBD through epigenetic prioritization at GWAS loci

We next tested whether phenotype-weighted 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 [37] and IBD [22]. 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 prioritize the most likely functional SNP across the 12 selected Parkinson's disease loci and 19 selected IBD loci. With this approach, PINES significantly distinguished 16 novel noncoding alleles from the background, implicating them based on epigenetic grounds as likely 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)(Fig. 6 and Supplemental Tables S2-S3).

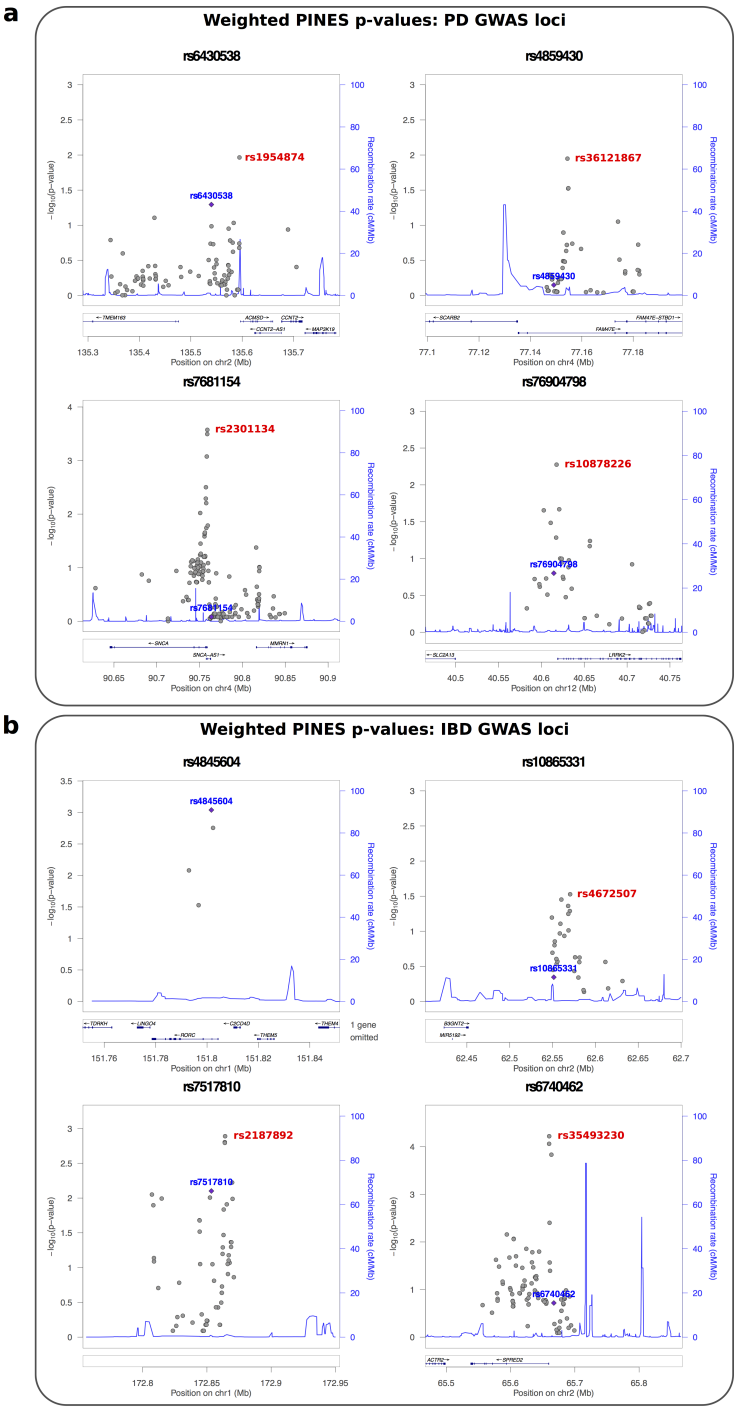

Figure 6. PINES predicts novel noncoding pathogenic variants through epigenetic prioritization of variants in Parkinson's disease and IBD GWAS loci. Loci were extracted from [37] and [22]. For each lead SNP, all variants with $LD \geq 0.4$ were selected, and loci were discarded if this list overlapped 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 prioritization is marked red. For the rs4845604 locus the GWAS and PINES lead SNP overlaps.

Discussion

Human genetics is revealing an ever-increasing number of noncoding variants playing a role in complex and Mendelian disease. However, the biological and pathogenic consequences of the vast majority of these alleles and signals remain elusive. Hence, there is a strong need to shift the focus of the field from identifying novel signals towards better understanding what they actually mean. Yet, the notion of biological function of a variant 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 [29]. 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 relevance in the most appropriate 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 a 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. By performing noncoding variant prioritization based solely on epigenetic annotations, PINES can aid in prioritizing noncoding variants underlying complex phenotypes that do not yet benefit from extensive and well-powered GWAS information, as well as prioritize lists of rare noncoding variants stemming from family studies of Mendelian phenotypes.

For users less experienced in working with the open-access PINES source code, PINES can be easily queried through a web server at http://genetics.bwh.harvard.edu/pines/ in a similar manner to our previous variant interpretation tool PolyPhen [38], which had been tailored to assess the consequences of coding variation. The web 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 systematic identification of regulatory activity specific to a subgroup of cell types or tissues can greatly increase our understanding of the still under-explored noncoding part of the human genome, as well as disease mechanisms linked to it.

Methods

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 [39] and SiPhy [40].

Noncoding background variants were selected randomly across the genome, and we used the ClinVar database [41] and GWAS Catalog [42] to ensure that no overlap exists with known functional loci. For the analysis in Fig. 3 we used GWAS studies of IBD [22, 32], celiac disease [31], blood lipid levels [33, 34], and multiple sclerosis [30] to determine enrichment-based weights for weighted PINES. We then used fine mapped variants on the corresponding phenotypes from [29] as our test set. All immune-related fine mapped variants in [29] with posterior probability ≥ 0.2 were used to generate the first panel in Fig. 3. A list of CAGE-defined FANTOM5 enhancers [6] was used to create Fig. 5. The analysis of all noncoding Parkinson's disease and IBD loci (Fig. 6) was based on regions identified in [37] and [22].

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) [43–45],

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 $\mathbf{\Sigma}$ 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 $\mathbf{\Sigma}$ is a noisy estimate of the true correlation structure of $\mathbf{X}$, we perform the spectral decomposition $\mathbf{\Sigma} = \sum_{i=1}^{639} \lambda_i \mathbf{u_i} \mathbf{u_i}^T$, 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): $\hat{\mathbf{\Sigma}} = \sum_{i=1}^{30} \lambda_i \mathbf{u_i} \mathbf{u_i}^T$. The matrix $\hat{\mathbf{\Sigma}}^{-1}$ 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: $\hat{\mathbf{\Sigma}}^{-1} = \sum_{i=1}^{30} \frac{1}{\lambda_i} \mathbf{u_i} \mathbf{u_i}^T$. If $\mathbf{x}$ is a vector of annotations, then the length of the vector projected into the decorrelated space is $\sqrt{(\mathbf{x}-\mu)^T \mathbf{W} \hat{\mathbf{\Sigma}}^{-1} \mathbf{W} (\mathbf{x}-\mu))}$, which corresponds to the reweighed Mahalanobis distance to the mean vector μ [46]. The cosine of the angle between the projections of two vectors $\mathbf{x}$ and $\mathbf{y}$ into the decorrelated annotation space is given by

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

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 [47], we additionally scale this angle by the length of the projected $\mathbf{x}$ vector, resulting in the following PINES score:

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

Smaller scores indicate greater evidence for functionality. Supplemental Fig. S8 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 **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.

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-PC 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-PC 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-PC directly.

Choice of feature weights

Weighting of features can be performed manually, as was the case for the pigmentation variants presented in Fig. 2, the first panel in Fig. 3, and the simulation in Supplemental Fig. S8, 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 Fig. 2. Such enrichment-based approaches are frequently used in predicting cell types contributing to specific phenotypes when GWAS or fine mapping data is available [29], 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. As an example, the annotation weight matrix that PINES computes based on enrichment of epigenetic marks across inflammatory bowel disease GWAS lead SNPs ([22]) is shown in Supplemental Fig. S9.

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.

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-PC [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. In our comparisons of different methods we extracted pre-computed published scores for GWAVA, CADD, and Eigen-PC, DANN, LINSIGHT, and FATHMM-MKL.

Data access

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.

Acknowledgments

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.

References

1. Visser M, Kayser M, Palstra RJ. HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. Genome research. 2012;22(3):446–455.

2. Eiberg H, Troelsen J, Nielsen M, Mikkelsen A, Mengel-From J, Kjaer KW, et al. 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. 2008;123(2):177–187.

3. Stavropoulos DJ, Merico D, Jobling R, Bowdin S, Monfared N, Thiruvahindrapuram B, et al. Whole-genome sequencing expands diagnostic utility and improves clinical management in paediatric medicine. npj Genomic Medicine. 2016;1:15012.

4. Consortium EP, et al. The ENCODE (ENCyclopedia of DNA elements) project. Science. 2004;306(5696):636–640.

5. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317–330.

6. Arner E, Daub CO, Vitting-Seerup K, Andersson R, Lilje B, Drabløs F, et al. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. Science. 2015;347(6225):1010–1014.

7. Whalen S, Truty RM, Pollard KS. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. Nature genetics. 2016;48(5):488–496.

8. Teng L, He B, Wang J, Tan K. 4DGenome: a comprehensive database of chromatin interactions. Bioinformatics. 2015;31(15):2560–2564.

9. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nature genetics. 2016;.

10. Ritchie GR, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. Nature methods. 2014;11(3):294–296.

11. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nature genetics. 2014;46(3):310.

12. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2014; p. btu703.

13. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day IN, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics. 2015;31(10):1536–1543.

14. Huang YF, Gulko B, Siepel A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nature genetics. 2017;49(4):618–624.

15. Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, et al. A method to predict the impact of regulatory variants from DNA sequence. Nature genetics. 2015;47(8):955–961.

16. Trynka G, Sandor C, Han B, Xu H, Stranger BE, Liu XS, et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. Nature genetics. 2013;45(2):124–130.

17. Liu B, Dai Y, Li X, Lee WS, Yu PS. Building text classifiers using positive and unlabeled examples. In: Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. IEEE; 2003. p. 179–186.

18. Yang P, Li XL, Mei JP, Kwoh CK, Ng SK. Positive-unlabeled learning for disease gene identification. Bioinformatics. 2012;28(20):2640–2647.

19. Kichaev G, Yang WY, Lindstrom S, Hormozdiari F, Eskin E, Price AL, et al. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet. 2014;10(10):e1004722.

20. Pickrell JK. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. The American Journal of Human Genetics. 2014;94(4):559–573.

21. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nature methods. 2012;9(3):215–216.

22. Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nature genetics. 2015;47(9):979–986.

23. Guenther CA, Tasic B, Luo L, Bedell MA, Kingsley DM. A molecular basis for classic blond hair color in Europeans. Nature genetics. 2014;46(7):748–752.

24. van den Boogaard M, Smemo S, Burnicka-Turek O, Arnolds DE, van de Werken HJ, Klous P, et al. A common genetic variant within SCN10A modulates cardiac SCN5A expression. The Journal of clinical investigation. 2014;124(4):1844–1852.

25. Soldner F, Stelzer Y, Shivalila CS, Abraham BJ, Latourelle JC, Barrasa MI, et al. Parkinson-associated risk variant in distal enhancer of α-synuclein modulates target gene expression. Nature. 2016;533(7601):95–99.

26. Gilks WP, Hill M, Gill M, Donohoe G, Corvin AP, Morris DW. Functional investigation of a schizophrenia GWAS signal at the CDC42 gene. The World Journal of Biological Psychiatry. 2012;.

27. Visser M, Palstra RJ, Kayser M. Human skin color is influenced by an intergenic DNA polymorphism regulating transcription of the nearby BNC2 pigmentation gene. Human molecular genetics. 2014; p. ddu289.

28. Leslie EJ, Taub MA, Liu H, Steinberg KM, Koboldt DC, Zhang Q, 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. 2015;96(3):397–411.

29. Farh KKH, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. 2015;518(7539):337–343.

30. Consortium IMSG, 2 WTCCC, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature. 2011;476(7359):214–219.

31. Dubois PC, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A, et al. Multiple common variants for celiac disease influencing immune gene expression. Nature genetics. 2010;42(4):295–302.

32. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012;491(7422):119–124.

33. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010;466(7307):707–713.

34. Consortium GLG, et al. Discovery and refinement of loci associated with lipid levels. Nature genetics. 2013;45(11):1274–1283.

35. Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, et al. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. Cell. 2016;165(6):1519–1529.

36. Huang H, Fang M, Jostins L, Mirkov MU, Boucher G, Anderson CA, et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. Nature. 2017;.

37. Nalls MA, Pankratz N, Lill CM, Do CB, Hernandez DG, Saad M, et al. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. Nature genetics. 2014;46(9):989–993.

38. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic acids research. 2002;30(17):3894–3900.

39. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010;6(12):e1001025.

40. Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics. 2009;25(12):i54–i62.

41. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic acids research. 2016;44(D1):D862–D868.

42. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic acids research. 2014;42(D1):D1001–D1006.

43. Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC. Estimating the support of a high-dimensional distribution. Neural computation. 2001;13(7):1443–1471.

44. Manevitz LM, Yousef M. One-class SVMs for document classification. the Journal of machine Learning research. 2002;2:139–154.

45. Tax DM, Duin RP. Support vector data description. Machine learning. 2004;54(1):45–66.

46. Greevy RA, Grijalva CG, Roumie CL, Beck C, Hung AM, Murff HJ, et al. Reweighted Mahalanobis distance matching for cluster-randomized trials with missing data. Pharmacoepidemiology and drug safety. 2012;21(S2):148–154.

47. Kriegel HP, Zimek A, 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. ACM; 2008. p. 444–452.