Published in final edited form as:

Nat Genet. 2016 February ; 48(2): 214–220. doi:10.1038/ng.3477.

A SPECTRAL APPROACH INTEGRATING FUNCTIONAL GENOMIC ANNOTATIONS FOR CODING AND NONCODING VARIANTS

IULIANA IONITA-LAZA[1,8], **KENNETH MCCALLUM**[1,8], **BIN XU**[2], and **JOSEPH BUXBAUM**[3,4,5,6,7]

[1]Department of Biostatistics, Columbia University, New York, NY 10032

[2]Department of Psychiatry, Columbia University, New York, NY 10032

[3]Seaver Autism Center for Research and Treatment, Icahn School of Medicine at Mount Sinai, New York, NY 10029

[4]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY 10029

[5]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029

[6]Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY 10029

[7]The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029

Abstract

Correspondence should be addressed to: I.I.-L. (; Email: ii2135@columbia.edu)

[8]Equal contribution

URLs
CADD: http://cadd.gs.washington.edu/
ClinVar: http://www.ncbi.nlm.nih.gov/clinvar/
COSMIC database: http://cancer.sanger.ac.uk/cosmic/
dbNSFP: https://sites.google.com/site/jpopgen/dbNSFP
ENCODE: https://www.encodeproject.org/
Ensembl: http://www.ensembl.org/index.html
GTEx: http://www.gtexportal.org/home/
GVS: http://gvs.gs.washington.edu/GVS141/
GWAS genes: http://www.genome.gov/Pages/About/OD/OPG/GWAS%20Catalog/GWASCatalog112608.xls
NHGRI GWAS Catalog: http://www.genome.gov/page.cfm?pageid=26525384&clearquery=1#download
Olfactory genes: http://senselab.med.yale.edu/ordb/info/humanorseqanal.htm
Roadmap Epigenomics: http://www.roadmapepigenomics.org/
1000 Genomes: http://www.1000genomes.org/
UCSC genome browser: https://genome.ucsc.edu/
VEP: http://www.ensembl.org/info/genome/variation/predicted_data.html#con

Competing Financial Interests
The authors declare no competing financial interests.

Author Contributions
I.I.-L. designed the study and wrote the manuscript. I.I-L. and K.M. developed the statistical methods and the software. I.I.-L. and K.M. analyzed the data. B.X. and J.D.B. provided bioinformatics support and contributed to the interpretation of the results. All authors have read and contributed to the manuscript.

Over the past few years, substantial effort has been put into the functional annotation of variation in human genome sequence. Such annotations can play a critical role in identifying putatively causal variants among the abundant natural variation that occurs at a locus of interest. The main challenges in using these various annotations include their large numbers, and their diversity. Here we develop an *unsupervised* approach to integrate these different annotations into one measure of functional importance (**Eigen**), that, unlike most existing methods, is not based on any labeled training data. We show that the resulting meta-score has better discriminatory ability using disease associated and putatively benign variants from published studies (in both coding and noncoding regions) compared with the recently proposed CADD score. Across varied scenarios, the **Eigen** score performs generally better than any single individual annotation, representing a powerful single functional score that can be incorporated in fine-mapping studies.

1. Introduction

The tremendous progress in massively parallel sequencing technologies enables investigators to efficiently obtain genetic information down to single base resolution on a genome-wide scale [1, 2, 3]. This progress has been complemented by numerous efforts to functionally annotate both coding and noncoding genomic elements and genetic variants in the human genome. Examples include computational tools such as PolyPhen [4] and GERP [5] for genetic variant annotation, and large-scale genomic projects such as the Encyclopedia of DNA Elements (ENCODE) [6], Ensembl and Roadmap Epigenomics [7] for genomic element annotation. Furthermore, the GTEx project is building a massive biospecimen repository to identify tissue-specific eQTLs and splicing QTLs using more than 40 tissues and over 1000 samples [8]. Hence, we now have available a rich set of functional annotations for both coding and noncoding variants, and this set will continue to increase in size. These annotations are important since they can help predict the functional effect of a variant, and can be further combined with population level genetic data to identify those variants at a locus of interest that are more likely to play a causal role in disease [9, 10, 11, 12]. As is well-known, although there are now many known genome-wide significant loci for many complex disorders, for the most part the underlying causal variants are unknown.

There are several difficulties in taking full advantage of these diverse functional annotations. One important challenge is that different annotations can measure different properties of a variant, such as the degree of evolutionary conservation, or the effect of an amino acid change on the protein function or structure in the case of coding variants, or, in the case of noncoding variants, the potential effect on regulatory elements. It is not known a priori which of the different annotations is more predictive of the most relevant functional effect of a particular variant. Another problem is that there is a high degree of correlation among annotations of the same type (e.g. evolutionary conservation scores, or regulatory-type annotations). Therefore, despite their potential to be useful for identifying functional variants, most of these annotations tend to be used in a subjective manner [13, 14, 15].

Recent efforts have been made to employ these diverse annotations in a more principled way. In particular, several studies have focused on identifying functional genomic elements enriched with or depleted of loci influencing risk to particular complex diseases [16, 17].

Other studies have focused on the integration of many different functional annotations into one score of functional importance. For example, Kircher et al. [18] proposed a supervised approach (support vector machine or SVM) to train a discriminative model. That is, they begin with two sets of variants, one labeled as deleterious and a second one as benign, and they fit a model that best separates the two sets. Benign variants are selected by comparing the human genome to the inferred genome of the most recent shared human-chimpanzee ancestor. Alleles that are not found in the common ancestor and which are fixed in the human genome are assumed to be mostly benign. These are compared to *de novo* variants generated randomly based on models of mutation rates across the genome. Although the proposed aggregate score, CADD, has notable advantages as described in [18], it has several potential limitations. In particular, the quality of the resulting model depends on the quality of the labeled data used in the training stage. First of all, the two sets used in the training dataset are unlikely to be sharply divided into benign and deleterious variants; specifically, the set of simulated *de novo* variants (labeled as deleterious) likely contains a substantial proportion of benign variants. Second, the SVM is trained to distinguish between variants that may be under evolutionary constraint and those likely neutral, and hence for disease mutations that are under weak evolutionary constraint (such as those influencing risk to complex traits), the trained model may not perform that well. Other supervised methods include GWAVA for noncoding variants [19], that uses as training dataset the 'regulatory mutations' from the public release of the Human Gene Mutation Database as deleterious variants, and common (minor allele frequency ≥1%) single-nucleotide variants from the 1000 Genomes Project as benign.

To the best of our knowledge almost all of the existing methods for integrating diverse functional annotations are supervised, i.e. they are based on a labeled training set as described above. Ideally, the training data would be obtained by sampling variants at random and then applying a gold-standard method to determine deleteriousness (or functionality). Unfortunately, such a gold-standard approach is currently not practical for a large number of variants, and so supervised methods must resort to training data that may be inaccurate or biased. Other approaches such as fitCons [20] are based on assessing evolutionary conservation, and may be suboptimal for weakly selected (or possibly not selected) disease mutations for complex traits.

Here we introduce an unsupervised spectral approach (**Eigen**) for scoring variants which does not make use of labeled training data. As such, its performance is not sensitive to a particular labeling of the training dataset. Instead, the approach we introduce in this paper is based on training using a large set of variants with a diverse set of annotations for each of these variants, but no label as to their functional status (Supplementary Table 1). We assume that the variants can be partitioned into two distinct groups, functional and non-functional (although the partition is unknown to us), and that for each annotation the distribution is a two-component mixture, corresponding to the two groups. The key assumption in the **Eigen** approach is that of blockwise conditional independence between annotations given the true state of a variant (either functional or non-functional). This last assumption implies that any correlation between annotations in different blocks is due to differences in the annotation means between functional and non-functional variants (Methods section). Because of this,

the correlation structure among the different functional annotations (Figure 1 and Supplementary Figure 1) can be used to determine how well each annotation separates functional and non-functional variants (i.e. the predictive accuracy of each annotation). Subsequently we construct a weighted linear combination of annotations, based on these estimated accuracies. We illustrate the discriminatory ability of the proposed meta-score using numerous examples of disease associated variants and putatively benign variants from the literature. In addition we consider a related, but conceptually simpler meta-score, **Eigen-PC**, which is based on the eigendecomposition of the annotation covariance matrix, and using the lead eigenvector to weight the individual annotations.

2. Results

Non-synonymous Variants

Training Data—For the coding set all variants with a match in the dbNSFP database [21], a database of all potentially non-synonymous SNVs in the human genome, were included. Note that this excludes synonymous variants which fall in coding regions but do not alter protein sequences. Annotations for non-synonymous variants are derived from several sources. In particular, the protein function scores (SIFT, PolyPhen - Div and Var scores, Mutation Assessor or MA) are all taken from dbNSFP v2.7. Evolutionary conservation scores (GERP_NR, GERP_RS, PhyloP - primate, placental mammal and vertebrate scores, PhastCons - primate, placental mammal and vertebrate scores) were obtained from the UCSC genome browser (November 2014). Allele frequencies in four populations (African or AFR, European or EUR, East Asian or ASN, Ad Mixed American or AMR) were obtained from the 1000 Genomes project (November 2014). Note that allele frequencies are only used in the training stage, and are not used in calculating the meta-score for specific variants due to high missing rates. Using the training data on $\approx$ 81.4 million coding non-synonymous variants, we calculate the weights for the different annotations (Supplementary Table 2). As shown, for **Eigen** several protein function scores (PolyPhenDiv, PolyPhenVar, and MA) have the highest weights, consistent with the expectation for coding non-synonymous variants, followed by evolutionary conservation scores and alternate allele frequencies. For **Eigen-PC**, evolutionary conservation scores get higher weights than the protein function scores. Since the evolutionary conservation block is large compared with the other blocks (Supplementary Figure 1), the evolutionary conservation block dominates the first principal component of the covariance matrix, increasing the weights in this block.

Once we derive the weights for individual functional scores, we can compute the meta-scores for variants of interest. We show below applications to possible pathogenic and benign variants from disease studies in the literature.

ClinVar Pathogenic vs. ClinVar Benign—The pathogenic and benign variant sets used for validation were obtained from the ClinVar database. Variants on chromosomes 1–22 that were categorized as one of “benign”, “likely benign”, “pathogenic”, or “likely pathogenic” were selected for the validation set. These were subdivided into a non-synonymous coding set, and a synonymous coding and noncoding set. The non-synonymous coding set consisted of all variants which matched an entry in dbNSFP, which included missense, nonsense, and

splice-site variants. This set is intended to capture all variants that alter protein structure. The coding synonymous and noncoding set (discussed in the next section) consists of variants that do not have a match in dbNSFP.

The AUC values for discriminating between non-synonymous pathogenic (n = 16,545) and benign (n = 3,482) variants using different functional scores (including the **Eigen** and **Eigen-PC** scores, v1.0 and v1.1 of the CADD-score (see Supplementary Material for a discussion of the differences between the two versions), and the individual functional scores) are reported in Supplementary Table 3. As shown, for missense variants PolyPhenDiv has the highest discrimination power (AUC=0.903), while the proposed **Eigen** score has an AUC of 0.864, and CADD-score v1.0 has an AUC of 0.837.

Mutations in genes for Mendelian diseases—*MLL2, CFTR, BRCA1* and *BRCA2* are four well-known genes carrying pathogenic mutations for Kabuki Syndrome, Cystic Fibrosis, and breast cancer, respectively. We selected reported disease mutations (namely "pathogenic" or "likely pathogenic" single nucleotide variants reported in the ClinVar database) in the *MLL2* (n = 108 with 31 missense), *CFTR* (n = 160 with 92 missense), *BRCA1* (n =125 with 28 missense) and *BRCA2* (n =110 with 13 missense) genes. P values from the Wilcoxon rank-sum test when comparing with benign variants in the ClinVar database are shown in Table 1. Overall results are highly significant for all the different methods, with the **Eigen** score performing better than the **Eigen-PC** and the CADD-score in most of the cases. In particular for missense variants in *MLL2*, the p value for **Eigen** is 3.1E-13, 5.1E-13 for **Eigen-PC**, whereas for the two versions of CADD-score the p values are 2.8E-02 (v1.0) and 2.8E-06 (v1.1). Note that since only a small proportion of the pathogenic SNVs in *MLL2, BRCA1*, and *BRCA2* are missense (most of them are nonsense), when we restrict consideration to missense variants, the differences between scores for pathogenic and benign variants become far less significant. For *CFTR* mutations, since they cause a recessive disease (cystic fibrosis), a larger proportion of them are missense compared to the other three genes (*MLL2, BRCA1, BRCA2*) which lead to diseases inherited in an autosomal dominant pattern. We also report the best performing individual annotation for each gene in Table 1. Overall, no single annotation performs best, although the best performing annotation in each case is a protein function score (SIFT, MA or PolyPhenVar). Results for each individual functional score are reported in Supplementary Table 4.

De novo mutations reported in ASD, SCZ, EPI and ID studies—We identified all autism (ASD), schizophrenia (SCZ), epileptic encephalopathies (EPI) and intellectual disability (ID) *de novo* mutations from published studies, along with *de novo* mutations identified in controls (CTRL) in those studies. We selected only those mutations with entries in the dbNSFP database. In total for ASD, we have n = 2,027 such mutations among which 1, 753 are missense [22, 23, 24, 25, 26]. For SCZ, we have n =636 mutations of which 571 are missense [27, 28, 29, 30, 31]. For EPI we identify n =210 mutations with 184 missense [32], and for ID we have n =114 mutations with 99 missense [33, 34]. For CTRL, we have n =1,310 mutations, of which 1,157 are missense [23, 25, 26, 28, 31, 34]. For ASD we also performed an analysis based only on those *de novo* mutations that fall into genes encoding

FMRP targets, as it has been shown that *de novo* ASD mutations are enriched among genes encoding FMRP targets [35, 36]. Results for the comparison of **Eigen** scores for mutations in different diseases and controls are shown in Figure 2. *De novo* mutations in ID and ASD-FMRP have the highest **Eigen** scores, followed by EPI, ASD, SCZ and CTRL mutations. P values from the Wilcoxon rank-sum test comparing scores for *de novo* mutations in cases vs. controls are reported in Table 2. The **Eigen-PC** score performs similar to the proposed **Eigen** score, and much better than the CADD-score, especially for EPI and ASD, with the p values being orders of magnitude smaller for the **Eigen** and **Eigen-PC** scores. Notably, when we consider the small subset of *de novo* variants in ASD that fall into genes encoding FMRP targets, the results become much more significant (even though the number of variants is reduced 15-fold), and in particular, for missense variants, the p value for the **Eigen** score is 3.2E-04, 9.4E-05 for **Eigen-PC** vs. 4.2E-02 for CADD-score v1.0 and 1.7E-02 for CADD-score v1.1. We also report the best performing individual annotation for each dataset, and as before no single annotation is best in all cases, although the best ones are again protein function scores. Results for each individual functional score are reported in Supplementary Table 5.

Noncoding and Synonymous Coding Variants

Training Data—For noncoding and synonymous coding variants, we use a suite of evolutionary conservation annotations and many regulatory annotations from the ENCODE project [6]. ENCODE histone modification, transcription factor binding and open chromatin data were downloaded from the UCSC genome browser (January 2015). A full list of functional genomic scores is given in the Supplementary Material (Supplementary Table 1). For the training dataset all variants in the 1000 Genomes Project dataset without a match in dbNSFP and within 500bp 5′ of the gene start site were included, for a total of 1,604,525 variants. In Supplementary Table 6 we report the estimated weights for individual annotations; as reported, evolutionary conservation scores tend to have the highest weights for the **Eigen** score, whereas regulatory annotations get the highest weights for **Eigen-PC**. Note that the regulatory block is large (Figure 1), containing over half the annotations used for calculating the weights. Therefore the regulatory block dominates the first principal component of the covariance matrix, increasing the weights in this block.

Below we show results of applications to possible pathogenic and benign noncoding and synonymous coding variants from disease studies in the literature. In addition to the two versions of CADD-score we also compare with another supervised method, GWAVA [19], specifically designed for noncoding variants.

ClinVar Noncoding and Synonymous Coding Variants—We have selected noncoding and synonymous coding variants from the ClinVar database. The selected variants include 3′UTR, upstream, downstream, intergenic, noncoding change, intronic, and synonymous coding variants. We have identified 111 such pathogenic mutations. For controls we selected a set of 111 benign variants from ClinVar matched for functional class (i.e. see Supplementary Material for more details) to the pathogenic variants. The AUC for several aggregate scores, and individual functional scores are given in Supplementary Table 3. As shown several conservation scores (GERP_RS, PhyloPla and PhyloVer) perform best,

followed closely by the **Eigen** score. **Eigen-PC** and GWAVA perform rather poorly for this dataset, similar to the regulatory annotations.

Genome-wide significant Single Nucleotide Polymorphisms (SNPs)—We computed scores for 14, 915 GWAS index SNPs that have been found genome-wide significant and reported in the NHGRI GWAS catalog. We note here that only a small proportion of the GWAS index SNPs are expected to be causal (estimated at 5% in [37]), with most of them being in linkage disequilibrium with the true causal SNPs.

Eigen score distribution for variants in different functional classes (e.g. regulatory, upstream, downstream, intergenic, intronic) are shown in Supplementary Figure 2A. GWAS variants hitting a known regulatory element (2, 115 variants) have the highest **Eigen** scores, as expected. We used the Genome Variation Server (GVS) to extract tag SNPs that have an r^2 of at least 0.8 with each GWAS index SNP. GVS divides the SNPs in an LD bin into "tag SNPs" and "other SNPs". This latter group consists of all the SNPs for which the r^2 value with any other SNP in the bin is below the 0.8 threshold. We construct two types of control sets, one consisting of "tag SNPs", and another one consisting of "other SNPs", all hitting a known regulatory element. We compare the various scores (**Eigen**, **Eigen-PC**, the two versions of CADD-score, GWAVA) for GWAS index SNPs and these control variants. We generate 20 such matched control sets, and in Table 3 we report the median p values from the Wilcoxon rank-sum test across these 20 comparisons. As shown both the **Eigen** and the **Eigen-PC** scores perform substantially better than the CADD-score. Furthermore the **Eigen-PC** tends to perform best, outperforming all the other meta-scores and the best performing individual functional annotations.

In addition, we have generated control sets matched for frequency, functional class (i.e. regulatory, 3′UTR, upstream, downstream, intergenic, noncoding change, intronic, and synonymous coding; see Supplementary Material for more details), and GWAS chip presence. We matched on SNP presence on four of the most commonly used GWAS platforms (Affymetrix Genome-Wide Human SNP Array 6.0, Illumina Human610-Quad BeadChip, Illumina OmniExpress, Illumina Human1M BeadChip). The matched control SNPs are chosen to be within ±100 kb of each index SNP. We generate 20 such matched control sets (due to the various constraints on the control sets, the number of SNPs in these matched sets, for both GWAS SNPs and control SNPs, is 10,718), and in Table 3 we report the median p values from the Wilcoxon rank-sum test across these 20 comparisons. As before, **Eigen-PC** outperforms all the other scores. In Supplementary Table 7 we report results for all the individual functional scores. As shown, the best performing individual annotations all belong to the regulatory block.

eQTLs—We selected a list of 3, 259 gene eQTLs identified using 373 European samples in Lappalainen et al. [38]. As with GWAS SNPs, eQTL variants hitting a known regulatory element (676 eQTLs) have the highest **Eigen** scores (Supplementary Figure 2B). We have constructed similar control sets to GWAS, based on "tag SNPs" and "other SNPs". The p values from the Wilcoxon rank-sum test are reported in Table 3. As shown, the **Eigen** and **Eigen-PC** scores lead to more significant results compared with both the CADD-score and

the GWAVA score. In Supplementary Table 7 we report results for all the individual functional scores.

Noncoding Cancer mutations from the COSMIC Database—We also compared scores for recurrent vs. non-recurrent somatic noncoding mutations in the COSMIC database [39] (the GWAVA scores are only available for a small number of the COSMIC variants, namely those that have been reported in dbSNP; therefore we omit the comparison with GWAVA for this dataset). The p values from the Wilcoxon rank-sum test for variants in different functional classes are reported in Table 4 (Supplementary Table 8 contains results for individual annotations). The p values for the **Eigen** and **Eigen-PC** scores are orders of magnitude smaller than those for the CADD-score, across different groups of variants. In Figure 3 we show the **Eigen** score distribution for variants in different functional classes. As shown, regulatory, 5′ UTR and 3′ UTR variants have the highest scores, while intergenic variants have the lowest scores, as expected.

3. Discussion

The **Eigen** score proposed here represents both a quantitative improvement in predictive power compared to existing methods, and a qualitative difference in the predictive model. The shift from supervised (CADD, GWAVA) to unsupervised algorithms as discussed here reduces the dependence on existing databases of observed variants, previously characterized elements and existing models of mutation, and allows extensions to cell type/tissue specific scores [16, 17, 37, 40, 41, 42]. Although supervised learning is preferable to unsupervised learning if a large, representative, and correctly labeled training set is available, unsupervised methods may have an advantage when labeled data is limited. We have shown that **Eigen** performs well compared to existing methods in both coding and noncoding regions. We have also shown that compared to individual annotations, the proposed meta-score performs favorably, with **Eigen** being close to optimal across a wide range of scenarios (Supplementary Tables 9 and 10). However **Eigen** should be viewed as complementary to the individual annotations; when possible, modeling each annotation's importance to a particular disease can be very informative [16].

In addition we have studied the performance of a related score **Eigen-PC**, based on the eigendecomposition of the annotation covariance matrix, and using the lead eigenvector to weight the individual annotations. In our experiments, **Eigen-PC** performs well across many scenarios, although, as we discuss in the Supplementary Material, it is more sensitive than **Eigen** to the component annotations and possible confounding factors; therefore **Eigen** is a more robust approach at this point. Although we have only experimented with the first principal component, it is possible that the second principal component is also informative (Supplementary Tables 19, 20, 21). Further work is needed to investigate how to incorporate the information in this second principal component.

Results for **Eigen** and **Eigen-PC** are similar for coding variants, with **Eigen** performing slightly better. In contrast, **Eigen-PC** has a considerable advantage over **Eigen** for the noncoding variants. The regulatory block is more than twice the size of the evolutionary conservation block. This causes the regulatory block to dominate the first principal

component of the covariance matrix, increasing the weights in this block. With the current set of annotations, the strong weights placed on the regulatory annotations improve **Eigen-PC**'s ability to discriminate between the different datasets for noncoding variants used here. Changing the set of annotations could disrupt this behavior.

Eigen can incorporate a large number of correlated functional annotations that are being generated by high-throughput projects such as ENCODE and Roadmap Epigenomics, if they fit the assumed block-structured correlation. We note that the set of annotations used by **Eigen** is a proper subset of the set used by the CADD-score. In particular, we have excluded several non-numerical annotations. To verify that **Eigen**'s improvement over CADD is not due to this difference in annotation sets, we have re-trained CADD on the same set of annotations used by **Eigen** and have shown that this new version of CADD performs similarly to the full CADD scores (Supplementary Material). As an additional experiment, we have also considered including CADD as one of the component annotations. The resulting score tends to perform worse than the original **Eigen** score, largely due to the fact that including CADD violates our assumption of conditional independence for annotations in different blocks (Supplementary Material).

Although for mutations in Mendelian diseases these aggregate scores can be very sensitive, for the majority of disease risk variant in complex diseases, these scores are expected to be mostly useful when combined with additional population level genetic data (Supplementary Figure 6).

Currently the **Eigen** score is defined separately on coding and noncoding variants, because different types of annotations are relevant to the two types of variants. In principle, these could be integrated into a score that encompasses both types. Given that **Eigen** is based on a two-component mixture model, this could be accomplished by converting the scores to posterior component probabilities, which would have the additional advantage of improving the interpretability of the scores.

Although indels represent only a small proportion of sequence variants [43], they represent a class of mutations that are likely to be functionally important, particularly when they cause frameshifts. However it is currently difficult to detect indels with high accuracy from short read sequence data [44, 45]. As methods to improve indel detection become more mature [46], we will take advantage of these new developments in future extensions of the **Eigen** and **Eigen-PC** scores.

Precomputed **Eigen** and **Eigen-PC** scores for every possible variant in the human genome are available for download at our website.

4. Methods

We assume that we have a set of randomly selected variants from the human genome, together with a diverse set of annotations, but no label as to their functional status. We assume that the variants can be partitioned into two distinct groups, functional and nonfunctional (although the partition is unknown to us), and that for each annotation the distribution is a two-component mixture, corresponding to the two groups.

Estimating the accuracy of individual functional annotation scores

Our approach is inspired by a recent paper by Parisi et al. [47] which considered the problem of combining multiple binary classifiers of unknown reliability, and which are conditionally independent (given the true status). The resulting meta-classifier is shown to be more accurate than most classifiers considered. Here we propose generalizations to cover prediction scores with arbitrary continuous distributions, as appropriate for many functional genomics scores. Generalizations to the case of blockwise conditional independence for functional scores are also considered.

Conditional independence among individual functional scores

We start with a dataset consisting of a large number of variants and their functional annotations. For simplicity, we first assume conditional independence among the individual functional scores. Supplementary Table 22 contains a description of the main variables used in this section for ease of reference. Let m be the number of variants, and k be the number of functional predictors (e.g. PolyPhen, GERP, etc). Let $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ik})$ be i.i.d. vectors of k functional impact scores for variants $i=1,\ldots, m$. It is assumed that the scores have been standardized so that for every score j we have $\mu_j = \mathrm{E}[Z_{ij}] = 0$ and $\sigma_j^2 = Var(Z_{ij}) = 1$. Let $\mathbf{C} = (C_1, \ldots, C_m)$ be indicator variables for the true status of the variants, with $C_i = 1$ if variant i is functional and $C_i = 0$ otherwise. Let F_j be the distribution of scores Z_{ij} for functional score j. The general idea is to treat the scores as belonging to a two-component mixture distribution, where the components correspond to a variant either being functional or not. In Parisi et al. the restriction of the predictors to binary outcomes yields a parametric family for the mixture component distributions. For continuous scores we make use of non-parametric mixture models. We have:

$$F_j(Z_{ij}) = \pi F_{j1}(Z_{ij}) + (1-\pi) F_{j0}(Z_{ij}),$$

where $\pi : P[C_i = 1]$, and F_{j1}, F_{j0} are the conditional distributions of Z_{ij} given $C_i = 1$ and $C_i = 0$ respectively. Define $\mu_{jl} = \mathrm{E}[Z_{ij} | C_i = l]$ for score j and $l = 0,1$. Note that

$$\mu_j = \pi \mu_{j1} + (1-\pi)\mu_{j0} = 0 \Rightarrow \mu_{j1} = -\frac{1-\pi}{\pi}\mu_{j0}. \quad (1)$$

It is easy to show that the covariance of any two scores j_1 and j_2 can be expressed as

$$Cov(Z_{ij_1}, Z_{ij_2}) = \pi\, Cov(Z_{ij_1}, Z_{ij_2} | C_i = 1) + (1-\pi)\, Cov(Z_{ij_1}, Z_{ij_2} | C_i = 0) + \frac{1-\pi}{\pi}\mu_{j_1 0}\mu_{j_2 0}. \quad (2)$$

This can be expressed in matrix form as

$$\mathbf{Q}=\pi\sum\nolimits_{1}+(1-\pi)\sum\nolimits_{0}+\mathbf{R}. \quad (3)$$

where $\mathbf{Q} = [q_{ij}]$ is the covariance matrix for $\mathbf{Z}$, Σ_1, Σ_0, are the component specific covariance matrices, and

$$\mathbf{R}=\frac{1-\pi}{\pi}\boldsymbol{\mu}_0^T\boldsymbol{\mu}_0. \quad (4)$$

where $\boldsymbol{\mu}_0 = (\mu_{10}, \ldots, \mu_{k0})$.

Therefore if the scores are conditionally independent given the true functional status for a variant (C_i) then we get that the covariance of any two scores j_1 and j_2 can be written as:

$$Cov(Z_{ij_1}, Z_{ij_2})=\frac{1-\pi}{\pi}\mu_{j_1 0}\mu_{j_2 0}. \quad (5)$$

Therefore under the assumption of conditional independence, the off diagonal entries in the covariance matrix are equal to those of the rank one matrix **R**. We are interested in $\boldsymbol{\mu}_0$ as the entries in $\boldsymbol{\mu}_0$ can be used to rank the scores since the accuracy of the score depends in part on how far apart the means of the conditional distributions are (i.e. $\mu_{j_1}-\mu_{j_0}=-\frac{1}{\pi}\mu_{j_0}$). Normally we do not know $\boldsymbol{\mu}_0$, but the values of $\boldsymbol{\mu}_0$ can be estimated by first estimating the diagonal entries of **R** (see below) and then computing the leading eigenvector.

The assumption of conditional independence is important since it implies that the off diagonal elements of the covariance matrix $\mathbf{Q} = [q_{ij}]$ equal the off diagonal elements of **R**, thereby allowing for the estimation of the rank one matrix **R**. Using the change of variable $|r_{ij}| = |q_{ij}| = e^{t_i}e^{t_j}$, the elements of **R** can be estimated by first solving the system of equations given by $\log|q_{ij}| = t_i + t_j$ for $i \neq j$. This gives a system of $k(k-1)/2$ equations with k unknowns. Since in practice the population covariance matrix **Q** of the functional scores is not known, the sample covariance matrix is used to estimate the population covariance matrix, and so least squares is used to estimate the solution. Then the diagonal elements can be estimated by $\widehat{r_{ii}}=e^{2\widehat{t_i}}$. In the next section we handle the case of blockwise conditional independence.

Note—We note that if the within component variances are small compared to the means, it follows from eq. (3) that $\mathbf{Q} \approx \mathbf{R}$. A simple approach then is to take the first principal component of matrix **Q** as an approximation of $\boldsymbol{\mu}_0$, without the need to estimate the rank one matrix **R**. However, this approach may fail if the within component variances are not all small. We refer to this approach as **Eigen-PC**, while the main approach that assumes (blockwise) conditional independence is referred to as **Eigen**.

Blockwise conditional independence among individual functional scores

The assumption of conditional independence may not be appropriate in the case of functional genomics annotations. For instance, protein functional predictors that use similar information for prediction (e.g. multiple sequence alignments and protein 3D-structures) are likely to be correlated even given the true functional status for a variant. On the other hand it is more plausible that predictors of different types, such as protein function scores and regulatory effect scores, would be independent given the true functional status of a variant. This motivates using the less strict assumption of blockwise conditional independence. Under this assumption the scores can be divided into disjoint, exhaustive blocks, such that predictors from different blocks are conditionally independent, while predictors within a block are still allowed to be conditionally dependent. In Figure 1, we show the correlation structure for 29 different functional annotations using the set of noncoding variants on chromosome 1 from the training dataset (see also the Results section; similarly, Supplementary Figure 1 shows the corresponding correlation structure based on the coding variants on chromosome 1 from the training set). A clear block structure can be seen, with different types of annotations forming distinct blocks, with stronger correlations within blocks than between them. The three distinct blocks are: an evolutionary conservation block (including several conservation scores such as GERP and PhyloP), a regulatory information block (including open chromatin measures, transcription factor binding, histone modifications), and an allele frequency block.

Under the assumption of blockwise conditional independence, we show that as long as there are at least three conditionally independent blocks we can still solve uniquely the system of equations above, and are able to estimate the rank one matrix **R**, and its leading eigenvector. More precisely, we prove the following lemma:

Lemma 1—*Let q_{ij} be the ijth entry of the covariance matrix* **Q***. Suppose that* **Q** *has a block structure, with three or more disjoint, exhaustive blocks, denoted by B_1, B_2, B_3 etc., that are conditionally independent. Then there is a unique solution for the variables $t_1, \ldots t_k$ in the system of equations given by* $\log|q_{ij}| = t_i + t_j$*, for i,j corresponding to different blocks.*

Proof: See Supplementary Material.

We estimate r_{ij} with i and j in the same block by $\widehat{r_{ij}}=e^{\widehat{t_i}}e^{\widehat{t_j}}$. We calculate the leading eigenvector of $\widehat{\mathbf{R}}$. As discussed previously, the entries in the eigenvector for the rank one matrix **R** are proportional to the accuracies of the individual predictors, and can be used to rank the various predictors. Next, we discuss how we may use these estimates of accuracies to combine the different predictors into one meta-score.

4.2. Meta-predictors

Once the blockwise division is chosen, the rank one matrix **R** can be estimated and the leading eigenvector determined. As discussed above, the entries in the eigenvector can be used to rank and combine annotations. Larger values for the components of the eigenvector indicate greater accuracy for the corresponding annotations, and the component values can be used as weights for combining annotations in a linear combination. This way we give

more weight to the more accurate annotations. If (e_1, …, e_k) is the eigenvector for the matrix **R**, and (Z_{i1}, …, Z_{ik}) are the functional scores for variant i, then the meta-score for variant i is given by

$$Eigen(i) = \mathbf{Z_i e}^T = \sum_{j=1}^{k} e_j Z_{ij}.$$

We refer to this method as **Eigen**. For **Eigen-PC** we use as weights the lead eigenvector of the covariance matrix **Q**.

Algorithm Outline—For ease of reference, we summarize here the complete approaches **Eigen** and **Eigen-PC** described above. For **Eigen**:

1. Rescale the functional scores to have mean zero, and variance one.
2. Calculate the covariance matrix, **Q**.
3. Designate the block structure for the set of annotations. In our setting, for non-synonymous coding variants we have three different blocks: one block with protein function scores, a second block with evolutionary conservation annotations, and a third block with allele frequencies. For noncoding and synonymous coding variants, we have one block with evolutionary conservation annotations, a second block with regulatory annotations, and a third block with allele frequencies.
4. Using the entries q_{ij} of **Q** corresponding to between block correlations, solve the system of equations given by $\log|q_{ij}| = t_i + t_j$ and use the variables t_1, … t_k to construct a rank one matrix **R**.
5. Take the eigendecomposition of **R**.
6. Calculate the scores as the weighted sum of the annotations, with the vector of weights equal to the eigenvector from the previous step.

Note that if the **Eigen-PC** method is used, the outline is similar. Steps 3. and 4. will be omitted, since the covariance matrix **Q** is used directly. In step 5. the eigendecomposition is applied to **Q** and in step 6. the lead eigenvector, the one with the greatest eigenvalue, is used (it was not necessary to specify this previously since **R** by construction has only one eigenvector).

Missing Annotations—Not all annotations are available at every variant. In particular, some annotations are only defined for specific classes of variants. For example, protein function scores are only defined in coding regions (for missense variants). This raises the question of how to calculate the meta-score for a variant when one or more annotations for this variant are missing or undefined. We calculate the meta-scores of coding missense, nonsense, and splice site variants, and of the remaining variants (including noncoding, and synonymous coding) separately. When an annotation is not defined for a type of variant, then we do not use it. When a variant is missing a value for an annotation (that is normally defined for that type of variant), we use mean imputation. The exception to this is where

protein function scores, such as SIFT, PolyPhen and MA scores, are missing at nonsense and splice site variants. In these cases, imputing the mean value will tend to underestimate the severity of these mutations. For SIFT a value of 0 is imputed, for PolyPhen a value of 1 is imputed, while for MA a value of 5.37 is imputed (the maximum values for those annotations). Note that we do not perform any imputation in the training stage when we learn the weights for the different annotations; the covariance matrix used to calculate the weights is based on pair-wise correlations, which allows variants with missing values for some annotations to be used.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by US National Institutes of Health grants R01 MH095797, R01 MH100233 and the Seaver Foundation. All analyses were conducted on the Minerva HPC complex at the Icahn School of Medicine at Mount Sinai.

References

1. Mardis ER. The impact of next-generation sequencing technology on genetics. Trends Genet. 2008; 24:133–141. [PubMed: 18262675]
2. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010; 11:31–46. [PubMed: 19997069]
3. Zhang J, et al. The impact of next-generation sequencing on genomics. J Genet Genomics. 2011; 38:95–109. [PubMed: 21477781]
4. Adzhubei IA, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7:248–249. [PubMed: 20354512]
5. Davydov EV, et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++ PLoS Comput Biol. 2010; 6:e1001025. [PubMed: 21152010]
6. Consortium E. P. et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]
7. Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. Nature. 2015; 518:317–330. [PubMed: 25693563]
8. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015; 348:648–660. [PubMed: 25954001]
9. Capanu M, et al. The use of hierarchical models for estimating relative risks of individual genetic variants: an application to a study of melanoma. Stat Med. 2008; 27:1973–1992. [PubMed: 18335566]
10. Capanu M, et al. Hierarchical modeling for estimating relative risks of rare genetic variants: properties of the pseudo-likelihood method. Biometrics. 2011; 67:371–380. [PubMed: 20707869]
11. Kichaev M, et al. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet. 2014; 10:e1004722. [PubMed: 25357204]
12. Ionita-Laza I, et al. Identification of rare causal variants in sequence-based studies: methods and applications to VPS13B, a gene involved in Cohen syndrome and autism. PLoS Genet. 2014; 10:e1004729. [PubMed: 25502226]
13. Ng SB, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet. 2010; 42:790–793. [PubMed: 20711175]
14. Bamshad MJ, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. 2011; 12:745–755. [PubMed: 21946919]

15. Meyer KB, et al. Fine-Scale Mapping of the FGFR2 Breast Cancer Risk Locus: Putative Functional Variants Differentially Bind FOXA1 and E2F1. Am J Hum Genet. 2013; 93:1046–1060. [PubMed: 24290378]
16. Pickrell JK. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am J Hum Genet. 2014; 94:559–573. [PubMed: 24702953]
17. Gusev A, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am J Hum Genet. 2014; 95:535–552. [PubMed: 25439723]
18. Kircher M, et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014; 46:310–315. [PubMed: 24487276]
19. Ritchie GRS, et al. Functional annotation of noncoding sequence variants. Nat Methods. 2014; 11:294–296. [PubMed: 24487584]
20. Gulko B, Hubisz MJ, Gronau I, Siepel A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat Genet. 2015; 47:276–283. [PubMed: 25599402]
21. Liu X, Jian X, Boerwinkle E. dbNSFP v2.0: A Database of Human Non-synonymous SNVs and Their Functional Predictions and Annotations. Human Mutation. 2013; 34:E2393–E2402. [PubMed: 23843252]
22. Iossifov I, et al. The contribution of de novo coding mutations to autism spectrum disorder. Nature. 2014; 515:216–221. [PubMed: 25363768]
23. Iossifov I, et al. De novo gene disruptions in children on the autistic spectrum. Neuron. 2012; 74:285–299. [PubMed: 22542183]
24. Neale BM, et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature. 2012; 485:242–245. [PubMed: 22495311]
25. O'Roak BJ, et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature. 2012; 485:246–250. [PubMed: 22495309]
26. Sanders SJ, et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature. 2012; 485:237–241. [PubMed: 22495306]
27. Fromer M, et al. De novo mutations in schizophrenia implicate synaptic networks. Nature. 2014; 506:179–184. [PubMed: 24463507]
28. Gulsuner S, et al. Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. Cell. 2013; 154:518–529. [PubMed: 23911319]
29. Girard SL, et al. Increased exonic de novo mutation rate in individuals with schizophrenia. Nature Genetics. 2011; 43:860–863. [PubMed: 21743468]
30. McCarthy SE, et al. De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. Mol Psychiatry. 2014; 19:652–658. [PubMed: 24776741]
31. Xu B, et al. *de novo* gene mutations highlight patterns of genetic and neural complexity in schizophrenia. Nat Genet. 2012; 44:1365–1369. [PubMed: 23042115]
32. Epi4K Consortium. De novo mutations in epileptic encephalopathies. Nature. 2013; 501:217–221. [PubMed: 23934111]
33. de Ligt J, et al. Diagnostic exome sequencing in persons with severe intellectual disability. N Engl J Med. 2012; 367:1921–1929. [PubMed: 23033978]
34. Rauch A, et al. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet. 2012; 380:1674–1782. [PubMed: 23020937]
35. Darnell JC, et al. FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. Cell. 2011; 146(2):247–261. [PubMed: 21784246]
36. Dong S, et al. De novo insertions and deletions of predominantly paternal origin are associated with autism spectrum disorder. Cell Rep. 2014; 9:16–23. [PubMed: 25284784]
37. Farh KK, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. 2015; 518:337–343. [PubMed: 25363779]
38. Lappalainen T, et al. Transcriptome and genome sequencing uncovers human functional variation. Nature. 2013; 501:506–511. [PubMed: 24037378]

39. Forbes SA, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucl Acids Res. 2015; 43:D805–D811. [PubMed: 25355519]
40. Trynka G, et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat Genet. 2013; 45:124–130. [PubMed: 23263488]
41. Ye CJ, et al. Intersection of population variation and autoimmunity genetics in human T cell activation. Science. 2014; 345:1254665. [PubMed: 25214635]
42. Ko A, et al. Amerindian-specific regions under positive selection harbour new lipid variants in Latinos. Nat Commun. 2014; 5:3983. [PubMed: 24886709]
43. Gudbjartsson DF, et al. Large-scale whole-genome sequencing of the Icelandic population. Nat Genet. 2015; 47:435–44. [PubMed: 25807286]
44. O'Rawe J, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med. 2013; 5:28. [PubMed: 23537139]
45. Lam HY, et al. Performance comparison of whole-genome sequencing platforms. Nat Biotechnol. 2011; 30:78–82. [PubMed: 22178993]
46. Fang H, et al. Reducing INDEL calling errors in whole genome and exome sequencing data. Genome Med. 2014; 6:89. [PubMed: 25426171]
47. Parisi F, Strino F, Nadler B, Kluger Y. Ranking and combining multiple predictors without labeled data. Proc Natl Acad Sci. 2014; 111:1253–1258. [PubMed: 24474744]
48. Liao BY, Zhang J. Null mutations in human and mouse orthologs frequently result in different phenotypes. Proc Natl Acad Sci USA. 2008; 105:6987–6992. [PubMed: 18458337]
49. MacArthur D, et al. A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. Science. 2012; 335:823–828. [PubMed: 22344438]
50. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet. 2013; 9:e1003709. [PubMed: 23990802]
51. Cohen, J. Statistical Power Analysis for the Behavioral Sciences. 2. Lawrence Erlbaum Associates; 1988.

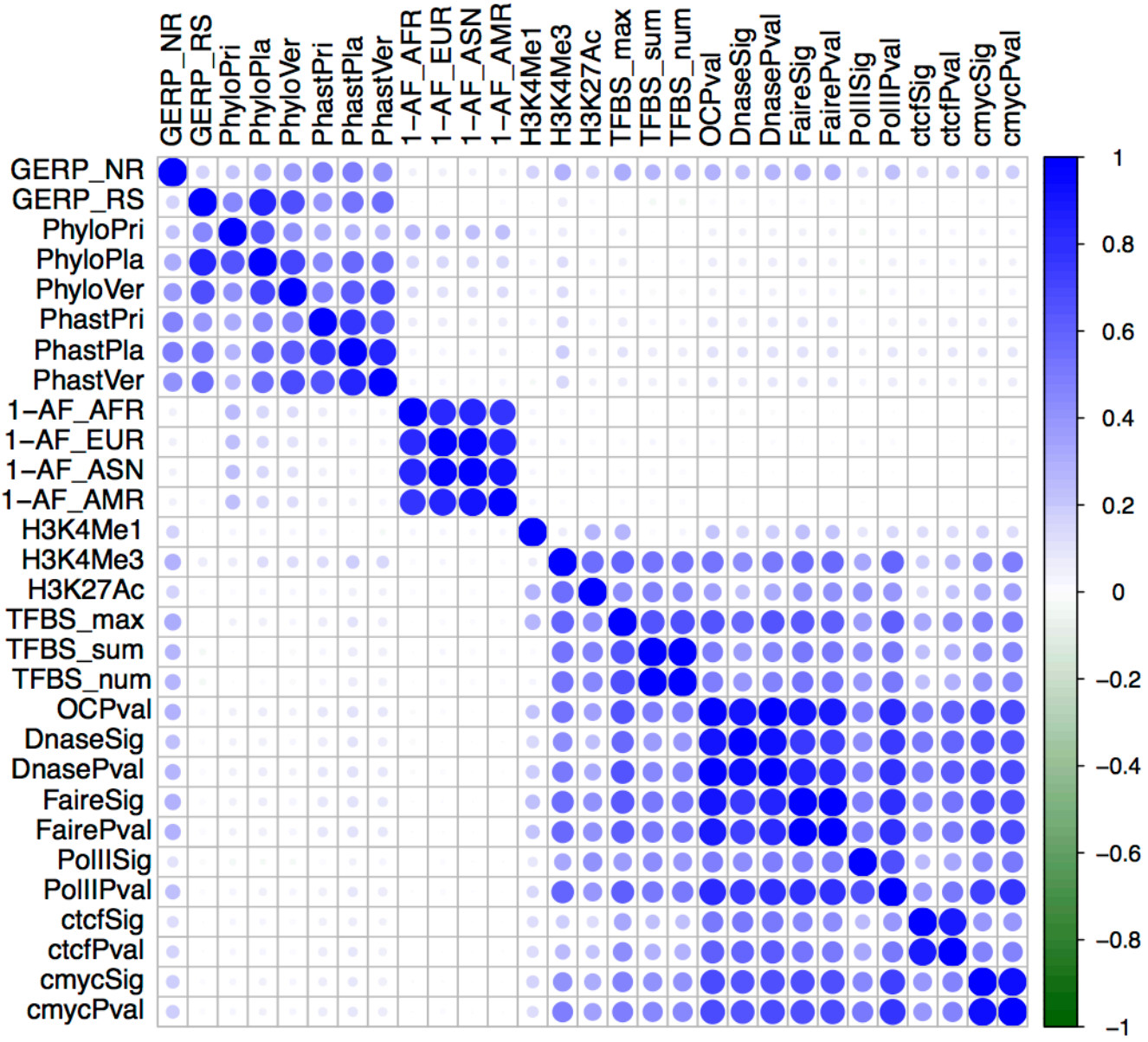

Figure 1.

Correlation among different functional annotations for the noncoding variants on chromosome 1 in the training dataset. Supplementary Figure 1 contains the correlation plot for non-synonymous coding variants.

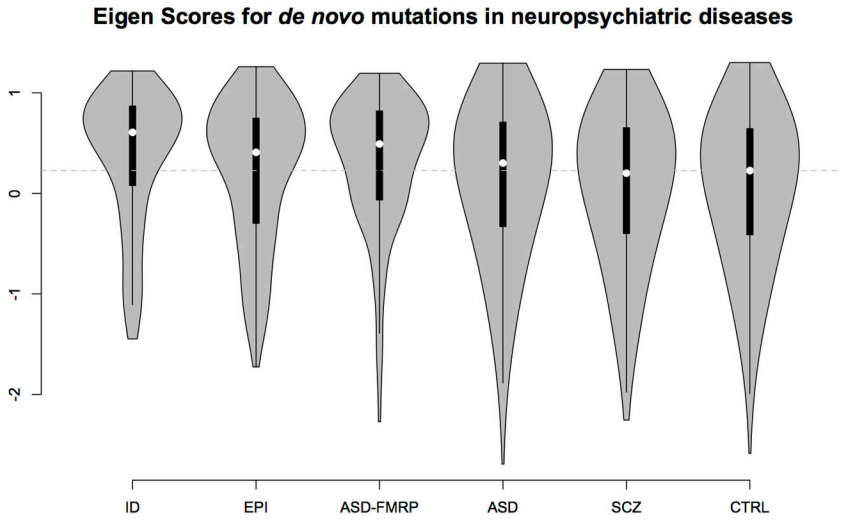

Figure 2.

Violin plots for **Eigen** scores for *de novo* mutations in ID, EPI, ASD-FMRP, ASD, SCZ and CTRL. The horizontal line corresponds to the median **Eigen** score for *de novo* CTRL mutations (the lowest scoring set).

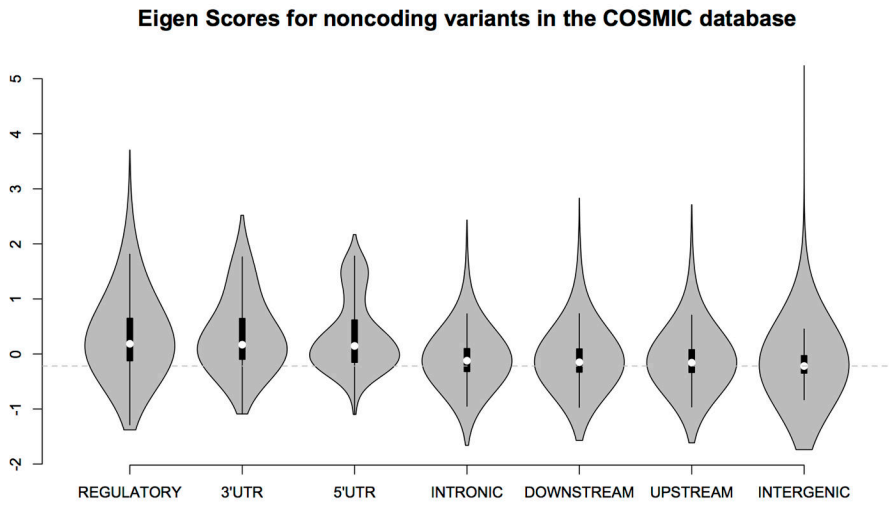

Figure 3.

Violin plots for **Eigen** scores for noncoding variants in the COSMIC database that reside in different functional categories. The horizontal line corresponds to the median **Eigen** score for intergenic variants (the lowest scoring class).

Table 1

P values (Wilcoxon rank-sum test) for *MLL2, CFTR, BRCA1, BRCA2*, contrasting pathogenic variants with benign variants in the ClinVar database. The best performing individual annotation is also reported (for missense variants only).

Gene	n	Variant type	Score	P value
			Eigen-PC	1.6E-50
			CADD-score v1.0	1.2E-42
			CADD-score v1.1	1.3E-49
	31	Missense	**Eigen**	3.1E-13
			Eigen-PC	5.1E-13
			CADD-score v1.0	2.8E-02
			CADD-score v1.1	2.8E-06
			SIFT	6.8E-15
CFTR	160	Missense and Nonsense	**Eigen**	1.3E-69
			Eigen-PC	8.2E-65
			CADD-score v1.0	1.1E-65
			CADD-score v1.1	3.1E-39
	92	Missense	**Eigen**	2.8E-37
			Eigen-PC	9.6E-37
			CADD-score v1.0	7.9E-35
			CADD-score v1.1	1.7E-21
			PolyPhenVar	4.8E-36
BRCA1	125	Missense and Nonsense	**Eigen**	2.5E-38
			Eigen-PC	6.0E-25
			CADD-score v1.0	2.2E-28
			CADD-score v1.1	1.3E-22
	28	Missense	**Eigen**	4.0E-03
			Eigen-PC	1.6E-02
			CADD-score v1.0	5.0E-03
			CADD-score v1.1	1.4E-03
			SIFT	1.0E-05
BRCA2	110	Missense and Nonsense	**Eigen**	9.8E-28
			Eigen-PC	3.3E-14
			CADD-score v1.0	1.5E-46
			CADD-score v1.1	7.7E-40
	13	Missense	**Eigen**	2.3E-01
			Eigen-PC	3.5E-01
			CADD-score v1.0	3.6E-01
			CADD-score v1.1	1.8E-02
			MA	9.5E-03

Table 2

P values (Wilcoxon rank-sum test) for *de novo* mutations in ASD, EPI, ID, and SCZ studies. ASD-FMRP analyses are based on *de novo* mutations in ASD cases that hit FMRP targets. The best performing individual annotation is also reported (for missense variants only).

Disease	n	Variant type	Score	P value
ASD	2,027	Missense and Nonsense	**Eigen**	6.0E-03
			Eigen-PC	1.6E-02
			CADD-score v1.0	8.4E-02
			CADD-score v1.1	3.2E-01
	1,753	Missense only	**Eigen**	9.0E-02
			Eigen-PC	1.5E-01
			CADD-score v1.0	7.4E-01
			CADD-score v1.1	5.8E-01
			PolyPhenDiv	5.4E-02
ASD-FMRP	132	Missense and Nonsense	**Eigen**	4.2E-05
			Eigen-PC	9.4E-06
			CADD-score v1.0	5.5E-03
			CADD-score v1.1	4.7E-03
	113	Missense only	**Eigen**	3.2E-04
			Eigen-PC	9.4E-05
			CADD-score v1.0	4.2E-02
			CADD-score v1.1	1.7E-02
			MA	1.0E-04
EPI	210	Missense and Nonsense	**Eigen**	3.1E-03
			Eigen-PC	5.0E-03
			CADD-score v1.0	4.0E-02
			CADD-score v1.1	2.0E-01
	184	Missense only	**Eigen**	6.0E-03
			Eigen-PC	1.3E-02
			CADD-score v1.0	8.1E-02
			CADD-score v1.1	1.7E-01
			PolyPhenVar	3.0E-03
ID	114	Missense and Nonsense	**Eigen**	1.7E-06
			Eigen-PC	1.1E-06
			CADD-score v1.0	3.7E-06
			CADD-score v1.1	9.5E-03
	99	Missense only	**Eigen**	6.7E-05
			Eigen-PC	6.0E-05
			CADD-score v1.0	3.5E-05
			CADD-score v1.1	3.3E-02
			MA	1.0E-04

Disease	n	Variant type	Score	P value
SCZ	636	Missense and Nonsense	**Eigen**	9.9E-01
			Eigen-PC	9.8E-01
			CADD-score v1.0	1.5E-01
			CADD-score v1.1	1.8E-01
	573	Missense only	**Eigen**	6.3E-01
			Eigen-PC	5.8E-01
			CADD-score v1.0	9.8E-01
			CADD-score v1.1	2.8E-02
			PhastPri	9.5E-02

Table 3

P values (Wilcoxon rank-sum test) for GWAS SNPs and eQTLs. Comparisons are shown between GWAS index SNPs and tag SNPs hitting regulatory elements. Also shown are comparisons between GWAS index SNPs and control SNPs matched for frequency, functional consequence, and GWAS array availability. Additionally, comparisons between eQTLs and tag SNPs hitting regulatory elements are shown. The best performing individual annotation is also reported.

Dataset	n	Comparison	Score	P value
GWAS	2,115	Regulatory GWAS vs. Tag SNPs	**Eigen**	1.2E-05
			Eigen-PC	4.0E-06
			CADD-score v1.0	5.9E-04
			CADD-score v1.1	2.0E-04
			GWAVA (TSS)	4.1E-06
			TFBS num	4.9E-05
GWAS	2,115	Regulatory GWAS vs. Other SNPs	**Eigen**	1.6E-09
			Eigen-PC	2.0E-13
			CADD-score v1.0	2.0E-06
			CADD-score v1.1	8.6E-07
			GWAVA (TSS)	7.4E-13
			TFBS sum	5.6E-09
GWAS	10,718	GWAS vs. Matched Controls	**Eigen**	6.9E-08
			Eigen-PC	3.5E-13
			CADD-score v1.0	1.0E-04
			CADD-score v1.1	5.2E-07
			GWAVA (TSS)	2.5E-09
			H3K4Me1	4.0E-11
eQTLs	676	Regulatory eQTLs vs. Tag SNPs	**Eigen**	1.8E-10
			Eigen-PC	7.0E-23
			CADD-score v1.0	3.1E-04
			CADD-score v1.1	4.3E-05
			GWAVA (TSS)	1.3E-03
			H3K4Me3	2.2E-24
eQTLs	676	Regulatory eQTLs vs. Other SNPs	**Eigen**	5.9E-13
			Eigen-PC	2.6E-27
			CADD-score v1.0	2.8E-04
			CADD-score v1.1	2.1E-05
			GWAVA (TSS)	7.3E-08
			H3K4Me3	3.8E-25

Table 4

P values (Wilcoxon rank-sum test) for somatic mutations (recurrent vs. non-recurrent) in the COSMIC database. Comparisons are done for variants in different functional categories. n-rec is the number of recurrent somatic mutations, and n-nonrec is the number of nonrecurrent somatic mutations. The best performing individual functional annotation is also reported.

Variant Class	n-rec	n-nonrec	Eigen	Eigen-PC	CADD-score v1.0	CADD-score v1.1	Best Individual Annotation
Regulatory	21,279	428,398	2.02E-165	5.13E-264	1.05E-71	2.70E-50	≤2.22E-308 (PolIIpval)
Intronic	85,502	2,093,158	2.40E-155	2.13E-112	2.89E-61	1.09E-10	≤2.22E-308 (GERP NR)
Downstream	15,956	318,967	2.73E-92	3.04E-128	4.31E-36	1.83E-28	1.01E-155 (GERP NR)
Upstream	14,636	309,615	1.28E-52	2.01E-84	7.90E-24	3.21E-17	9.68E-86 (PolIIPval)
Noncoding Change	4,903	66,717	2.51E-07	2.49E-21	1.51E-01	4.84E-05	8.13E-35 (PolIIPval)
3′UTR	2,236	28,261	6.94E-03	4.22E-04	1.06E-05	3.37E-01	5.67E-05 (GERP NR)
5′UTR	417	3,908	1.14E-02	2.32E-01	6.43E-02	1.15E-01	2.79E-07 (GERP NR)
Intergenic	75,327	2,182,466	1.49E-02	3.97E-06	1.08E-06	6.30E-16	1.19E-18 (H3K4Me1)
Synonymous	434	2,388	1.09E-01	9.69E-01	8.25E-01	2.88E-01	2.16E-03 (PhyloPri)

Nat Genet. Author manuscript; available in PMC 2016 July 04.