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

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

[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, Mount Sinai School of Medicine, New York, NY 10029

[5] Departments of Genetics and Genomic Sciences, and Neuroscience, and Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029

[6] Mindich Child Health and Development Institute, Mount Sinai School of Medicine, New York, NY 10029

[7] Equal contribution

[*] Corresponding author: ii2135@columbia.edu

Abstract. Over the past few years, substantial effort has been put into the functional annotation of variation in human genome sequence. Indeed, for any genetic variant, whether protein coding or noncoding, a diverse set of functional annotations is available from projects such as Ensembl, ENCODE and Roadmap Epigenomics. 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. In particular, it is not clear a priori which annotation is better at predicting functionally relevant variants. It is therefore desirable to integrate these different annotations into a single measure of functional importance for a variant. Here we develop an *unsupervised* approach to derive such a meta-score (**Eigen**), that, unlike most existing methods, is not based on any labelled training data. Furthermore, the proposed method produces estimates of predictive accuracy for each functional annotation score, and subsequently uses these estimates of accuracy to derive the aggregate functional score for variants of interest as a weighted linear combination of individual annotations. We show that the resulting meta-score has better discriminatory ability using disease associated and putatively benign variants from published studies (for both Mendelian and complex diseases) compared with the recently proposed CADD score. In particular, we show that the proposed meta-score outperforms the CADD score on noncoding variants from GWAS and eQTL studies, noncoding somatic mutations in the COSMIC database, and on *de novo* coding mutations in epilepsy and autism studies. 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 (e.g. case-control frequencies from GWAS or sequencing studies) to identify those variants at a locus of interest that are more likely to play a causal role in a given 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 single 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 labelled 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 labelled 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 (labelled 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 (HGMD) as deleterious variants, and common (minor allele frequency $\geq 1\%$) 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 labelled 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 labelled 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 (Supplemental Table S1). 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 block-wise 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, as we show in the Methods section. Because of this, the correlation structure among the different functional annotations (Figure 1 and Supplemental Figure S1) 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, both coding and noncoding, from the literature. In addition we consider a related, but conceptually simpler meta-score, **Eigen-PC**, which is based on the direct eigendecomposition of the annotation covariance matrix, and using the lead eigenvector to weight the individual annotations. Note that due to difficulties in accurate identification of insertion-deletions (indels), we focus our analyses

below on single nucleotide variants (SNVs), although one can calculate the meta-scores for indels in a similar fashion.

2. Results

2.1. **Non-synonymous Variants.**

Training Data. For the coding set all variants with a match in the dbNSFP database [21], a database of 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, which covers all potentially non-synonymous SNVs in the human genome. 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 ≈ 76.7 million coding non-synonymous variants, we calculate the weights for the different annotations (Supplemental Table S2). 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 (Supplemental Figure S1), 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.

i. 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. This includes 3'UTR, 5'UTR, upstream, downstream, intergenic, noncoding change, intronic, and synonymous coding mutations.

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 Supplemental Material for a discussion of the differences between the two versions), and the individual functional scores) are reported in Supplemental Table S3. 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.

ii. 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 Supplemental Table S4.

iii. 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 epilepsy and autism, 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 Supplemental Table S5.

2.2. **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 obtained is given in the Supplemental Material (Supplemental Table S1). 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 $418,997$ variants. In Supplemental Table S6 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.

i. 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. 3'UTR, upstream, downstream, intergenic, noncoding change, intronic, and synonymous coding; see Supplemental Material for more details) to the pathogenic variants. The AUC for several aggregate scores, and individual functional scores are given in Supplemental Table S3. 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.

ii. 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 (see Web-based Resources). We note here that only a small proportion of the GWAS index SNPs are expected to be causal (estimated at 5% in [37]), and most of them are just 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 Supplemental Figure S2A. 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** score 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 annotation.

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 Supplemental 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 $\pm$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 Supplemental Table S7 we report results for all the individual functional scores. As shown, the best performing individual annotations all belong to the regulatory block.

iii. 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 (Supplemental Figure S2B). 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 Supplemental Table S7 we report results for all the individual functional scores.

iv. Noncoding Cancer mutations from the COSMIC Database. We compared the **Eigen**, **Eigen-PC** and two versions of the CADD-score for recurrent vs. non-recurrent somatic noncoding mutations in the COSMIC database [39] (note that 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 (Supplemental Table S8 contains results for all individual scores). 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 (regulatory, 5' UTR, 3' UTR, upstream, downstream, intronic, intergenic). 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 (e.g. CADD-score, GWAVA) to unsupervised algorithms as discussed here reduces the dependence on existing databases of observed variants, previously characterized elements and existing models of mutation. Furthermore, many existing methods are conservation-based [18, 20], and for complex diseases this may be less than optimal due to the weak (or non-existent) selection against complex disease mutations. Unlike these existing approaches, the proposed method learns from the data the individual functional annotations that are relevant, separately for coding and noncoding variants. We have shown that the proposed score performs well in a wide-range of

scenarios and leads to stronger association signals compared to existing methods for putatively disease and benign variants from published studies, in both coding and noncoding regions. We have also shown that compared to individual annotations, the proposed meta-score performs favorably; while in each specific situation a particular functional annotation may perform best, the **Eigen** score performs close to optimal across a wide range of scenarios, and represents a principled way to combine a large number of annotations (see also Supplemental Tables S9 and S10). We note however that **Eigen** should be viewed as complementary to the many individual annotations; individual annotations provide important information by themselves, in addition to easier interpretability; when possible, modeling each annotation's importance to a particular disease [16] can be very informative.

In addition we have studied the performance of a related score **Eigen-PC**. **Eigen-PC** is based on the direct eigendecomposition of the annotation covariance matrix, and then using the lead eigenvector to weight the individual annotations. In our experiments, **Eigen-PC** performs well across many scenarios, although, as we discuss below, it is more sensitive than **Eigen** to the component annotations and possible confounding factors. Although we have only experimented with the first principal component, it is possible that other principal components are also informative. Further work is needed to investigate potential improvements to the **Eigen-PC** score.

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. A notable difference in the two methods is that **Eigen-PC** uses the entire annotation covariance matrix while **Eigen** uses only the between block entries. The regulatory block is more than twice the size of the next largest one, 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 paired datasets for

noncoding variants used here. Changing the set of annotations could disrupt this behavior. For example, adding new conservation scores could cause weight to be shifted away from the regulatory elements in **Eigen-PC**, which may substantially impact the model's performance. In comparison, adding new annotations to a block in **Eigen** will not cause a shift in the weights for all annotations in that block since it excludes the within block correlations.

The aggregate score proposed here can incorporate a large number of correlated functional annotations, with the condition that they fit the block-structured correlation assumed by **Eigen**. We note that the set of annotations used by **Eigen** is a proper subset of the full set used by the CADD-score. In particular, in the construction of **Eigen** we have excluded several non-numerical annotations, including reference and alternate alleles, and functional consequence. 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 (v1.0 and v1.1), and generally worse than the proposed **Eigen** score (see Supplemental Material for more details).

Although the list of annotations we have currently included in our meta-score calculation is by no means exhaustive, it will be straightforward to include other possible annotations that are being generated by high-throughput projects such as ENCODE and Roadmap Epigenomics to improve the prediction. As an additional experiment, we have also considered including the CADD-score as one of the component annotations. As we show in the Supplemental Material, the resulting score tends to perform worse than the original **Eigen** score, largely due to the fact that including CADD violates our main assumption of conditional independence for annotations in different blocks. Furthermore, when studying particular diseases, it will likely prove essential to incorporate tissue and cell type specific annotations [37, 16, 40, 17, 41, 42]. For example, when studying neuropsychiatric diseases, one might want to incorporate features that are relevant to a neurodevelopmental context. Therefore the general framework we have introduced here can be adapted to construct

disease-specific functional scores. Similarly, it is possible to produce custom scores based on subsets of the annotations used here, if one is interested in investigating specific biological mechanisms. For example, a set of scores could be calculated using only open chromatin measures if chromatin accessibility is of particular interest in a study.

Data artifacts can impact the accuracy of meta-scores as discussed here, especially for **Eigen-PC**. In particular, correlations between annotations that are due to something other than the functional/non-functional mixture can skew the results. However, the block structure we have in the **Eigen** score may help to minimize this problem. The blocks are chosen in such a way that functional annotations derived using the same or similar (experimental) data are grouped together. Since the weights used by the **Eigen** score depend on the **R** matrix, which is derived using between block correlations, the presence of artifacts such as batch effects should be to decrease the weight of the affected annotations. For example, if batch effects distorted several measures of open chromatin, this would likely decrease the discriminative power of those measures when it came to discerning functional and non-functional variants. This in turn would decrease the correlation between those measures and annotations from different blocks, causing them to be down weighted.

Although for mutations involved in Mendelian diseases these aggregate scores can be very sensitive, for the majority of disease risk variants involved in complex diseases, these scores are expected to be mostly useful when combined with additional population level genetic data. We show in the Supplemental Material (Supplemental Figure S6) how the **Eigen** score can be formally combined with population level genetic data in the framework of hierarchical models to help prioritize causal variants for experimental functional studies.

As already mentioned, most of the existing methods are supervised approaches. The accuracy of supervised methods is primarily limited by the quality of the labelled training dataset. If a large, representative, and correctly labelled training set is available, then supervised learning is preferable to unsupervised learning, which usually requires stronger model assumptions. However,

unsupervised methods may have an advantage when labelled data is unavailable, limited, or low quality. The currently available labelled datasets have limitations, as discussed before, which may limit the accuracy of supervised methods for combining functional annotations, and so unsupervised methods such as described here may be preferable at this time.

Currently the **Eigen** score is defined separately on coding and noncoding variants, partly because different types of annotations contribute and are relevant to the two different types of scores. In principle, these could be integrated into a single 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 the posterior component probabilities, which would have the additional advantage of improving the interpretability of the scores. This would require fitting a non-parametric mixture distribution to the set of annotations, which presents non-trivial difficulties. As such, it is left for future work.

Although indels represent only a small proportion of sequence variants (7% in the whole-genome sequencing study in Iceland [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 due to errors in library preparation, biases in sequencing methods, and artifacts in detection algorithms [44, 45]. As methods to improve indel detection become more mature [46], we will take advantage of these new developments in indel identification 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.

Web-based resources

Eigen: http://www.columbia.edu/∼ ii2135/eigen.html

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

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 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.

4.1. **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. Table 5 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 $\mu_j = E[Z_{ij}] = 0$ 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 $\mu_{jl} := E[Z_{ij}|C_i = l]$ 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 the covariance of any two scores j_1 and j_2 can be expressed as

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

This can be expressed in matrix form as

$$(3) \qquad \mathbf{Q} = \pi\mathbf{\Sigma}_1 + (1-\pi)\mathbf{\Sigma}_0 + \mathbf{R}$$

where $\mathbf{Q} = [q_{ij}]$ is the covariance matrix for $\mathbf{Z}$, $\mathbf{\Sigma}_1, \mathbf{\Sigma}_0$ are the component specific covariance matrices, and

$$(4) \qquad \mathbf{R} = \frac{1-\pi}{\pi}\boldsymbol{\mu_0}^T\boldsymbol{\mu_0},$$

where $\boldsymbol{\mu_0} = (\mu_{10}, ..., \mu_{k0})$.

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:

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

Therefore under the assumption of conditional independence, the off diagonal entries in the covariance matrix are equal to those of the rank one matrix $\mathbf{R}$. 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. $\mu_{j1} - \mu_{j0} = -\frac{1}{\pi}\mu_{j0}$). 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 $\mathbf{R}$ (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 $\mathbf{R}$, thereby allowing for the estimation of the rank one matrix $\mathbf{R}$. Using the change of variable $|r_{ij}| = |q_{ij}| = e^{t_i}e^{t_j}$, the

elements of $\mathbf{R}$ 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 $\mathbf{Q}$ 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 $\mathbf{Q}$ as an approximation of $\boldsymbol{\mu}_0$, without the need to estimate the rank one matrix $\mathbf{R}$. 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, Supplemental Figure S1 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 $\mathbf{R}$, and its leading eigenvector. More precisely, we prove the following lemma:

Lemma 1. *Let q_{ij} be the ijth entry of the covariance matrix $\mathbf{Q}$. Suppose that $\mathbf{Q}$ 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 Supplemental 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 $\mathbf{R}$ 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 $\mathbf{R}$ 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 $(e_1, \ldots, e_k)$ is the eigenvector for the matrix $\mathbf{R}$, and $(Z_{i1}, \ldots, Z_{ik})$ are the functional scores for variant i, then the meta-score for variant i is given by

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

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

4.3. **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, $\mathbf{Q}$.
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 $\mathbf{Q}$ 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 $\mathbf{R}$.
5. Take the eigen decomposition of $\mathbf{R}$.
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 $\mathbf{Q}$ 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.

References

[1] Mardis ER (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet* 24: 133–141.

[2] Metzker ML (2010) Sequencing technologies - the next generation. *Nat Rev Genet* 11: 31–46.

[3] Zhang J et al. (2011) The impact of next-generation sequencing on genomics. *J Genet Genomics* 38: 95–109.

[4] Adzhubei IA et al. (2010) A method and server for predicting damaging missense mutations. *Nat Methods* 7: 248–249.

[5] Davydov EV et al. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* 6: e1001025.

[6] Consortium EP et al (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57–74.

[7] Roadmap Epigenomics Consortium (2015) Integrative analysis of 111 reference human epigenomes. *Nature* 518: 317–330.

[8] GTEx Consortium (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348: 648–660.

[9] Capanu M et al. (2008) The use of hierarchical models for estimating relative risks of individual genetic variants: an application to a study of melanoma. *Stat Med* 27: 1973–1992.

[10] Capanu M et al. (2011) Hierarchical modeling for estimating relative risks of rare genetic variants: properties of the pseudo-likelihood method. *Biometrics* 67: 371–380.

[11] Kichaev et al. (2014) Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet* 10: e1004722.

[12] Ionita-Laza I et al. (2014) Identification of rare causal variants in sequence-based studies: methods and applications to VPS13B, a gene involved in Cohen syndrome and autism. *PLoS Genet* 10: e1004729.

[13] Ng SB et al. (2010) Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* 42: 790–793.

[14] Bamshad MJ et al. (2011) Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 12: 745–755.

[15] Meyer et al. (2013) Fine-Scale Mapping of the FGFR2 Breast Cancer Risk Locus: Putative Functional Variants Differentially Bind FOXA1 and E2F1 *Am J Hum Genet* 93: 1046–1060.

[16] Pickrell JK (2014) Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am J Hum Genet* 94: 559–573.

[17] Gusev A et al. (2014) Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am J Hum Genet* 95: 535–552.

[18] Kircher M et al. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* doi: 10.1038/ng.2892.

[19] Ritchie GRS et al. (2014) Functional annotation of noncoding sequence variants. *Nat Methods* 11: 294–296.

[20] Gulko B, Hubisz MJ, Gronau I, Siepel A (2015) A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat Genet* 47: 276–283.

[21] Liu X, Jian X, Boerwinkle E (2013) dbNSFP v2.0: A Database of Human Non-synonymous SNVs and Their Functional Predictions and Annotations. *Human Mutation* 34: E2393–E2402.

[22] Iossifov I et al. (2014) The contribution of de novo coding mutations to autism spectrum disorder. *Nature* 515: 216–221.

[23] Iossifov I et al. (2012) De novo gene disruptions in children on the autistic spectrum. *Neuron* 74: 285–299.

[24] Neale BM et al. (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* 485: 242–245.

[25] O'Roak BJ et al. (2012) Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* 485: 246–250.

[26] Sanders SJ et al. (2012) De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 485: 237–241.

[27] Fromer M et al. (2014) De novo mutations in schizophrenia implicate synaptic networks. *Nature* 506: 179–184.

[28] Gulsuner S et al. (2013) Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* 154: 518–529.

[29] Girard SL et al. (2011) Increased exonic de novo mutation rate in individuals with schizophrenia. *Nature Genetics* 43: 860–863.

[30] McCarthy SE et al. (2014) De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. *Mol Psychiatry* 19: 652–658.

[31] Xu B et al (2012) *de novo* gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat Genet* 44: 1365–1369.

[32] Epi4K Consortium (2013) De novo mutations in epileptic encephalopathies. *Nature* 501: 217–221.

[33] de Ligt J et al. (2012) Diagnostic exome sequencing in persons with severe intellectual disability. *N Engl J Med* 367: 1921–1929.

[34] Rauch A et al. (2012) Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* 380: 1674–1782.

[35] Darnell JC et al. (2011) FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* 146(2):247–261.

[36] Dong S et al. (2014) De novo insertions and deletions of predominantly paternal origin are associated with autism spectrum disorder. *Cell Rep* 9: 16–23.

[37] Farh KK et al. (2015) Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 518: 337–343.

[38] Lappalainen T et al. (2013) Transcriptome and genome sequencing uncovers human functional variation. *Nature* 501: 506–511.

[39] Forbes SA et al. (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucl. Acids Res.* 43: D805–D811.

[40] Trynka G et al. (2013) Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet* 45: 124–130.

[41] Ye CJ et al. (2014) Intersection of population variation and autoimmunity genetics in human T cell activation. *Science* 345: 1254665.

[42] Ko A et al. (2014) Amerindian-specific regions under positive selection harbour new lipid variants in Latinos. *Nat Commun* 5: 3983.

[43] Gudbjartsson DF et al. (2015) Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* in press

[44] O'Rawe J et al. (2013) Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med* 5: 28.

[45] Lam HY et al. (2011) Performance comparison of whole-genome sequencing platforms. *Nat Biotechnol* 30: 78–82.

[46] Fang H et al. (2014) Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome Med* 6: 89.

[47] Parisi F, Strino F, Nadler B, Kluger Y (2014) Ranking and combining multiple predictors without labeled data. *Proc Natl Acad Sci* 111: 1253–1258.

[48] Liao BY, Zhang J (2008) Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc Natl Acad Sci USA* 105: 6987–6992.

[49] MacArthur et al. (2012) A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science*, 335: 823–828.

[50] Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB (2013) Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* 9: e1003709.

[51] Cohen J (1988) Statistical Power Analysis for the Behavioral Sciences (second ed.). Lawrence Erlbaum Associates.

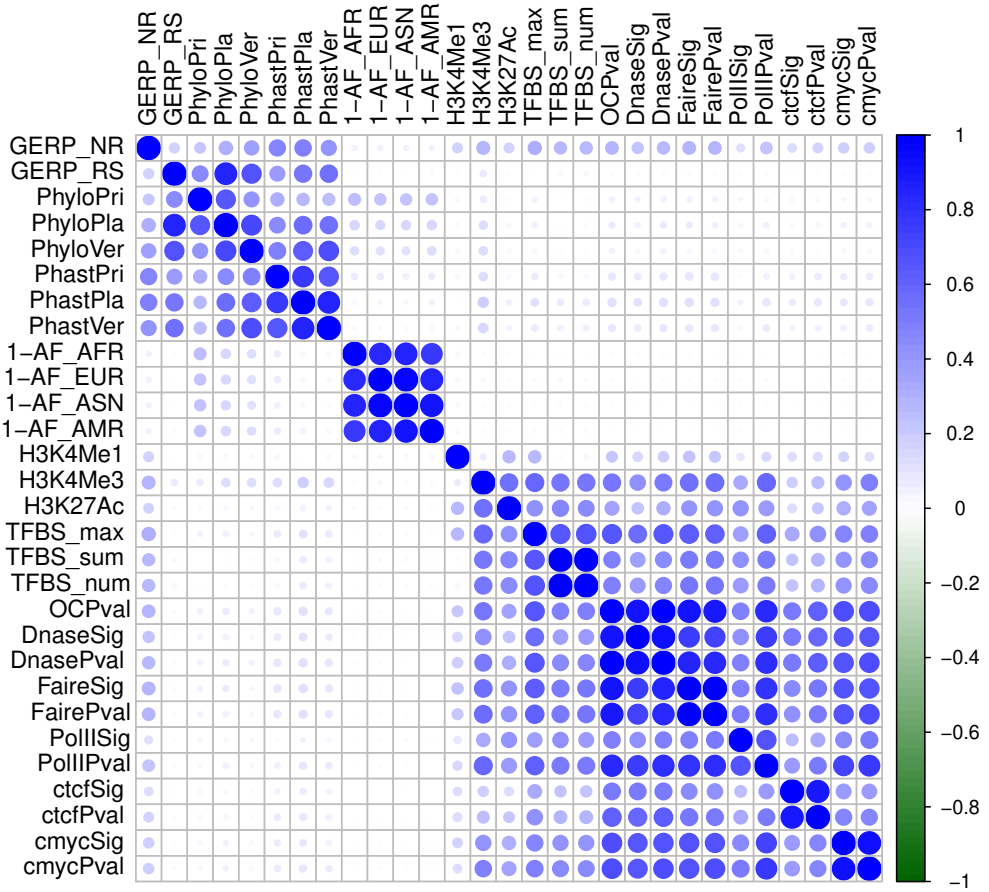


FIGURE 1. Correlation among different functional annotations for the noncoding variants on chromosome 1 in the training dataset. Supplemental Figure S1 contains the correlation plot for non-synonymous coding variants.

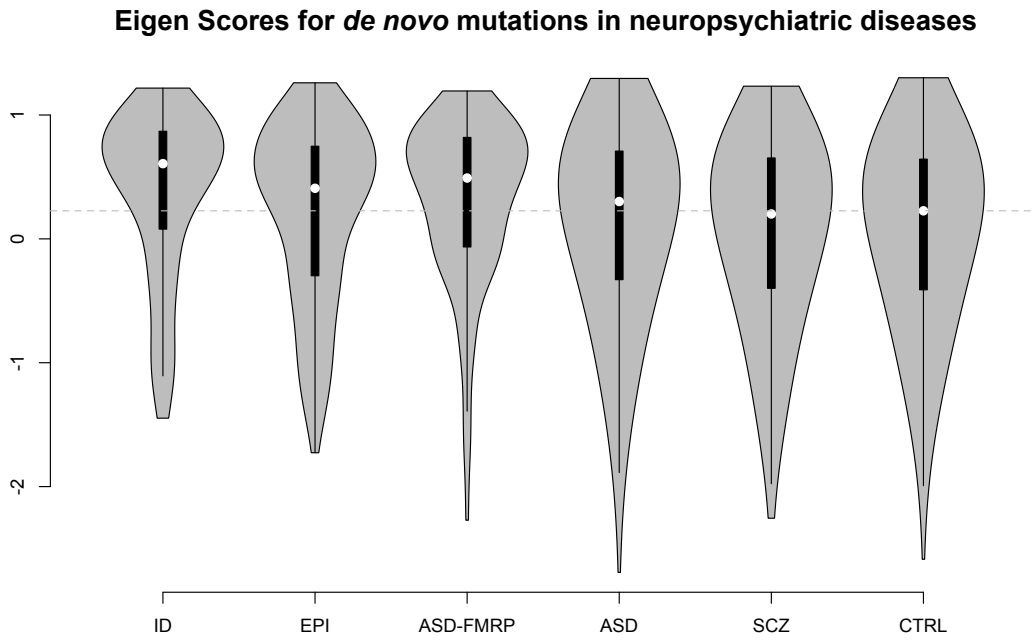


FIGURE 2. Violin plots for **Eigen** scores for *de novo* mutations in ID, EPI, ASD-FMRP, ASD, SCZ and CTRL. The horizontal line corresponds to the median **Eigen** score for *de novo* CTRL mutations (the lowest scoring set).

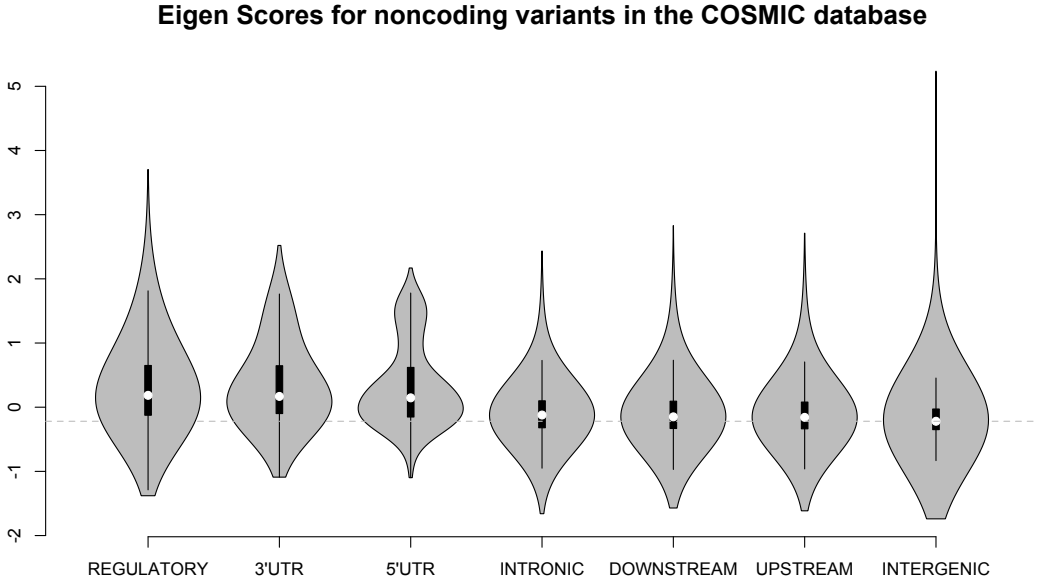


FIGURE 3. Violin plots for **Eigen** scores for noncoding variants in the COSMIC database that reside in different functional categories. The horizontal line corresponds to the median **Eigen** score for intergenic variants (the lowest scoring class).

Gene	n	Variant type	Score	P value
MLL2	108	Missense and Nonsense	**Eigen**	1.1E-56
			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 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).

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
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 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).

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 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.

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	$\leq$ 2.22E-308 (PolIIpval)
Intronic	85,502	2,093,158	2.40E-155	2.13E-112	2.89E-61	1.09E-10	$\leq$ 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 (PolIIIPval)
Noncoding Change	4,903	66,717	2.51E-07	2.49E-21	1.51E-01	4.84E-05	8.13E-35 (PolIIIPval)
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)

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.

Variable	Description
m:	number of variants in the training data
k:	number of functional scores to be combined
C_i:	Indicator variable for the state of variant i (1=functional vs. 0=nonfunctional)
Z_{ij}:	Value of functional score j at variant i
F_j:	Distribution of functional score Z_{ij}
F_{jk}:	For $k = 0, 1$, conditional distribution of Z_{ij} given mixture component $C_i = k$
μ_{jk}:	For $k = 0, 1$, conditional mean of Z_{ij} given mixture component $C_i = k$
π:	Proportion of functional variants in training data
$\mathbf{Q}$:	Matrix of pairwise correlations of standardized annotations
$\mathbf{\Sigma}_1, \mathbf{\Sigma}_0$:	Component specific covariance matrices
$\mathbf{R}$:	Rank one matrix derived from between block values of Q
t_i:	Variables defined such that $\lvert r_{ij} \rvert = e^{t_i + t_j}$

TABLE 5. Definitions of variables used in Methods section.

Supplemental Material

S1. Proof Lemma

Lemma 1. *Let q_{ij} be the ijth entry of the covariance matrix $\mathbf{Q}$. Suppose that $\mathbf{Q}$ 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. We show the proof for the case of three blocks. Let k_1 be the number of functional annotations in the first block, k_2 for the second block, and k_3 for the third block with $k_1 + k_2 + k_3 = M$. For clarity, we rename the variables for the second block as $s_1, \ldots s_{k_2}$, and those for the third block as $u_1, \ldots, u_{k_3}$. Then we have the following systems of equations:

$$\begin{pmatrix} t_1 + s_1 & t_1 + s_2 \ldots t_1 + s_{k_2} \\ t_2 + s_1 & t_2 + s_2 \ldots t_2 + s_{k_2} \\ \cdots & \\ t_{k_1} + s_1 & t_{k_1} + s_2 \ldots t_{k_1} + s_{k_2} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \ldots a_{1k_2} \\ a_{21} & a_{22} \ldots a_{2k_2} \\ \cdots & \\ a_{k_1 1} & a_{k_1 2} \ldots a_{k_1 k_2} \end{pmatrix}. \tag{6}$$

Adding up the elements of this matrix we get:

$$k_2(t_1 + \cdots + t_{k_1}) + k_1(s_1 + \cdots + s_{k_2}) = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} a_{ij}.$$

Similarly, we can get

$$k_3(t_1 + \cdots + t_{k_1}) + k_1(u_1 + \cdots + u_{k_3}) = \sum_{i=1}^{k_1} \sum_{j=1}^{k_3} b_{ij},$$

and

$$k_3(s_1 + \cdots + s_{k_2}) + k_2(u_1 + \cdots + u_{k_3}) = \sum_{i=1}^{k_2}\sum_{j=1}^{k_3} c_{ij}.$$

The matrices **a**, **b**, **c** represent the corresponding sub-matrices of the **Q** matrix. From this system of equations we can solve for $t_1 + \cdots + t_{k_1}$, $s_1 + \cdots + s_{k_2}$ and $u_1 + \cdots + u_{k_3}$. We also have by summing the elements from the first column that

$$t_1 + \cdots + t_{k_1} + k_1 \cdot s_j = \sum_{i=1}^{k_1} a_{ij}$$

for $j = 1 \ldots k_2$ and we can then get $s_1, \ldots s_{k_2}$. Furthermore, from the equations

$$t_i + s_1 = a_{i1}$$

we can get the solution for t_i with $i = 1 \ldots k_1$. We can then easily get the solution for $u_1 \ldots u_{k_3}$.

Since there are more equations than unknowns the only issue remaining is to require that the systems of equations are compatible. Since the exponential of the matrix on the left in eq. (6) is of rank 1, a necessary (and sufficient) condition is that the exponential of the matrix on the right in eqn. (6) is of rank 1. The exponentials of the entries on the right hand side are covariances for pairs of conditionally independent random variables, so by (5) we can write

$$\begin{pmatrix} a_{11} & a_{12} \ldots a_{1k_2} \\ a_{21} & a_{22} \ldots a_{2k_2} \\ \ldots & \\ a_{k_1 1} & a_{k_1 2} \ldots a_{k_1 k_2} \end{pmatrix} = \frac{1-\pi}{\pi} \begin{pmatrix} \mu_{1,0}\lambda_{1,0} & \mu_{1,0}\lambda_{2,0} \ldots \mu_{1,0}\lambda_{k_2,0} \\ \mu_{2,0}\lambda_{1,0} & \mu_{2,0}\lambda_{2,0} \ldots \mu_{2,0}\lambda_{k2,0} \\ \ldots & \\ \mu_{k_1,0}\lambda_{1,0} & \mu_{k_1,0}\lambda_{2,0} \ldots \mu_{k_1,0}\lambda_{k_2,0} \end{pmatrix} \tag{7}$$

where $\mu_{i,0}$ is the conditional mean of functional annotation i in the first block given component 0, and $\lambda_{j,0}$ is the conditional mean of functional score j in the second block. Therefore the matrix can

be written as $\frac{1-\pi}{\pi}\vec{\mu}_0(\vec{\lambda}_0)^T$ where $\vec{\mu}_0$ and $\vec{\lambda}_0$ have dimension $k_1 \times 1$ and $k_2 \times 1$ respectively. Therefore the exponential of the matrix is of rank 1 and the proof is complete. □

S2. Determining the Functional Class of a Variant

The functional class for a variant is retrieved from the CADD database (see Web-based Resources). These were originally produced using the Ensemble Variant Effect Predictor (VEP) with the per_gene option. When a variant matches multiple functional categories - for example, a variant that is synonymous in one splice variant and non-synonymous in another - this option causes VEP to only return the most severe effect for each gene. In most cases this results in a single annotation per variant. The exception is when more than one gene overlaps the variant. If this occurs, CADD will return multiple lines for the annotation, one per gene. In this case the first annotation listed in the CADD output is used here. The severity ranking used by VEP is given in the documentation (see Web-based Resources).

Variants that are classified in the CADD annotations as "Non Synonymous", resulting in an amino acid substitution, "Stop Lost", removing the stop codon, "Stop Gained", producing a premature stop codon, or "Splice Site", or "Canonical Splice", altering the splice junction between exons, are considered to be non-synonymous coding changes. The noncoding annotations are "Regulatory", referring to variants in a sequence with a known regulatory function, "Intronic", referring to variants occurring in introns but not part of a splice site, "Downstream" and "Upstream", referring to variants in genic region either after the last exon or before the first exon, "Noncoding change", referring to variants in noncoding RNAs, "3prime UTR" and "5prime UTR", referring to variants in the untranslated portions of a spliced RNA, "Intergenic", referring to variants outside a known gene region. "Synonymous" refers to variants in a coding region that do not result in an amino acid substitution. This last category is not included in the non-synonymous group since it has no potential to alter the amino acid sequence, and so is not covered by any of the protein function scores.

S3. Gene sets: Olfactory, Non-immune essential, Loss-of-function, GWAS, Tolerant, Intolerant and Random Genes

We selected non-synonymous coding variants from dbNSFP that fall into various gene sets: olfactory genes (see Web-based Resources; $n = 560,522$), non-immune essential genes ([48]; $n = 658,159$), genes with at least one loss-of-function (LoF) variant ([49]; $n = 10,982,549$), GWAS genes (see Web-based Resources; $n = 1,723,191$), tolerant ($n = 6,195,826$) and intolerant ($n = 10,499,312$) genes, defined as being in the upper and lower 5th percentile genes with respect to RVIS [50], respectively, as well as a set of random genes ($n = 1,034,481$). The results are shown in Supplemental Figure S3. For the proposed **Eigen** score, variants in intolerant, essential and GWAS genes have the highest scores, followed by random, LoF, tolerant and olfactory genes. Variants in these gene sets had lower scores than pathogenic variants reported in the ClinVar database, but higher scores than benign variants in the ClinVar database. These results are similar to the ones obtained using the CADD-score (Supplemental Figure S4), with one difference. Namely, variants in tolerant and olfactory genes tend to have lower CADD-scores than benign variants in ClinVar. As shown in Supplemental Figure S5, variants in tolerant and olfactory genes tend to have higher protein function scores (e.g. PolyPhenDiv) compared to benign variants, but lower conservation scores (e.g. PhyloVer). Since the CADD-score focuses on evolutionary selection (it quantifies negative selection at a position), these lower conservation scores for variants in tolerant and olfactory genes are reflected in the lower CADD-scores for these gene sets compared to benign variants in ClinVar.

S4. Hierarchical model to combine sequencing data with functional annotations

The **Eigen**, **Eigen-PC** and other aggregate scores are expected to be most useful when combined with population level genetic data for fine-mapping purposes at loci of interest. Therefore, we have performed simulation studies to investigate the improvement in discriminatory ability by combining

sequencing data from a case-control dataset with the **Eigen** score compared to using the **Eigen** score alone.

We based our simulations on data for one gene, the Vacuolar Protein Sorting 13 homolog B (*VPS13B*, also known as *COH1*, MIM #607817), from a whole-exome sequencing autism spectrum disorders (ASD) case/control dataset with 860 individuals. *VPS13B* is a gene associated with Cohen syndrome (CS, OMIM #216550), a rare autosomal recessive neurodevelopmental disorder, and mutations in this gene have also been reported in individuals with autism and non-syndromic intellectual disability. We simulated the truly causal variants in this gene based on a logistic regression model with the **Eigen** score as the sole predictor, assuming an association between the causal status of a variant and the **Eigen** score of magnitude (relative risk or RR) 1.1, 2 or 4 and assuming the proportion of truly causal variants to be 10%. Only non-synonymous variants were used in the simulations. The case-control status was generated as follows. For carriers of causal variants we generated a continuous phenotype from a normal distribution with a mean of 0.5 and a standard deviation of 0.2, while for non-carriers we used a standard normal distribution (corresponding to a Cohen's d effect size [51] of 0.53 - moderate effect). Cases were defined as the individuals with phenotype values above the median while the remaining individuals were classified as controls. We compare the discriminative performance of the hierarchical model [9, 10, 12] including the **Eigen** score as the functional predictor, with that of using the **Eigen** score alone. ROC curves are shown in Supplemental Figure S6. The average AUC values for the hierarchical model/**Eigen** score are: 0.744/0.508 (RR=1.1), 0.791/0.663 (RR=2), and 0.857/0.766 (RR=4). As shown, combining the **Eigen** score with the case-control frequencies improves the power to identify true causal variants, especially when the RR for the association between the **Eigen** score and the causal status is low (1.1-2), as seems to be the case in many of the examples we looked at.

S5. CADD v1.0 vs. v1.1

In v1.1 Kircher et al. [18] make two changes to the original (v1.0) score. First, the authors add several functional scores not included in v1.0. Second, they use a logistic regression model rather than a SVM as the learner. In the release notes for v1.1 the authors compare the two versions on results from tests used in the original paper. For example, they compare the correlation between the CADD-scores and the change in expression level associated with variants in regulatory regions for three genes. They find that v1.1 has a higher correlation in one of the genes, and in the data pooled from all three; however, v1.0 has higher correlation for the other two genes. They also look at AUC in four comparisons between ClinVar pathogenic and ESP likely benign variants. In three of the four comparisons, v1.1 has a small advantage over v1.0, in the fourth they are essentially equal. These results suggest that v1.1 may be an improvement on balance, but that it does not always dominate v1.0. This is in line with our findings, in which v1.0 sometimes has worse performance than v1.1 and sometimes has better.

S6. Including CADD into the construction of the **Eigen** score

We have performed several experiments with the CADD-score included as one of the component annotations, despite the fact that including the CADD-score violates our main assumption of conditional independence for annotations in different blocks. For both coding and noncoding setting, we included the CADD-score v1.0 in the evolutionary conservation block. As can be seen from the correlation plots (Supplemental Figures S7 and S8), the CADD-score correlates strongly with the conservation scores (as expected), but also with annotations in the other blocks (by way the CADD-score is constructed). Because of these correlations with annotations in other blocks, the CADD-score gets assigned fairly high weight, especially in the noncoding case (Supplemental Tables S11 and S12). However this high weight is not necessarily reflective of the predictive accuracy of CADD, but reflects the natural correlation CADD has with annotations in the other

blocks. Nonetheless we have performed several experiments to investigate how such a combined score (**Eigen+C**) performs. In Supplemental Tables S13 and S14 we show comparisons of **Eigen** and **Eigen+C** when applied to *de novo* mutations in neuropsychiatric diseases, and mutations in genes for Mendelian disorders, respectively. As shown, **Eigen+C** tends to perform worse compared with the **Eigen** score alone in most of the cases. In Supplemental Table S15 we show comparisons between **Eigen**, **Eigen-PC** and **Eigen+C** on the COSMIC dataset. For this dataset, **Eigen-PC** outperforms **Eigen+C** in all cases, whereas **Eigen** sometimes performs better, sometimes worse than **Eigen+C**.

S7. Comparisons with CADD-score with reduced set of annotations

We have performed comparisons with the CADD-score trained on the same set of annotations we have considered in the construction of the **Eigen** score. Specifically, we first re-trained CADD using the exact same set of annotations we used for our own score **Eigen**. Based on the new CADD model, we have calculated new scores for variants in our example datasets (see Results section for more details on these datasets) and the results are summarized in Supplemental Tables S17, S16, and S18. Overall the performance of the new CADD-score is consistent with that of the full CADD-scores (v1.0 and v1.1), and these results show that the **Eigen** score outperforms the CADD-score not simply because of the set of annotations used by **Eigen** (which is a proper subset of the set used by CADD), but rather because of more fundamental differences in the methodologies used to construct the two types of scores, as we explain in the main text. In Table S16 we show results on non-synonymous *de novo* variants in various neuropsychiatric diseases. In Table S17 we show results for the comparisons of the different aggregate scores on missense variants in four Mendelian genes. Note that we have excluded the nonsense mutations in these four genes because the new version of CADD did not work properly for nonsense mutations due to the exclusion of functional consequence from the annotation set. In Table S18 we report results on noncoding variants identified in GWAS and eQTL studies.

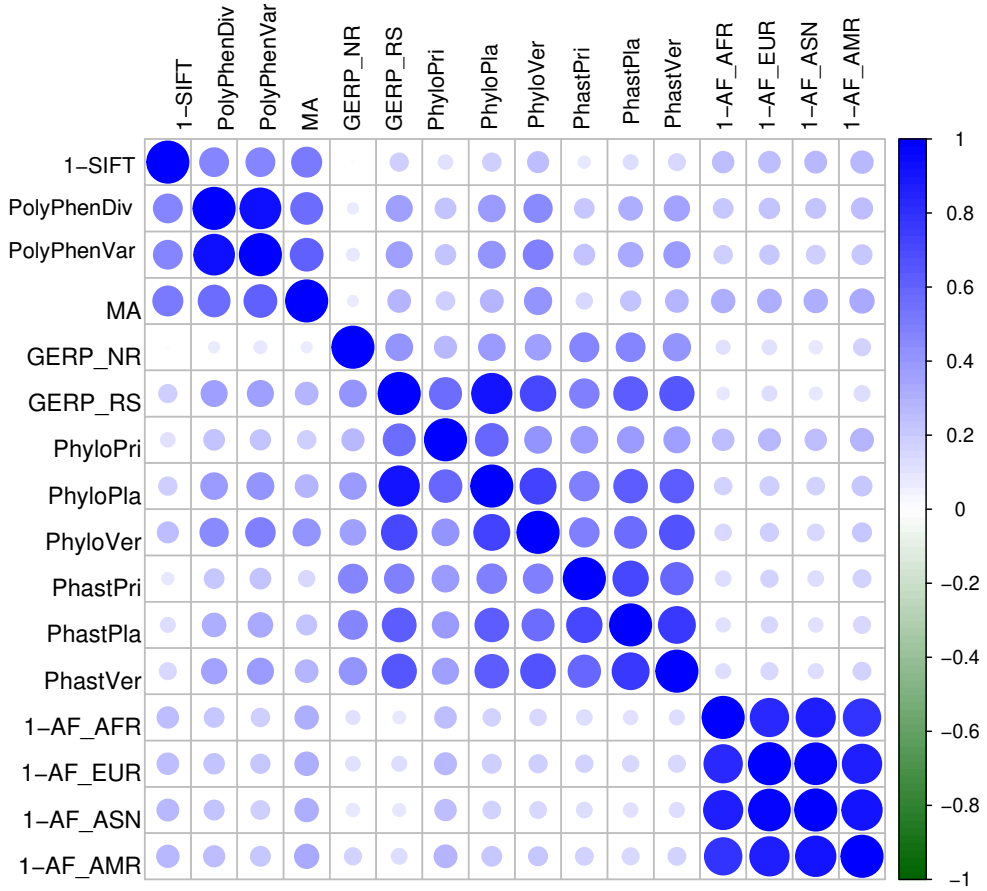
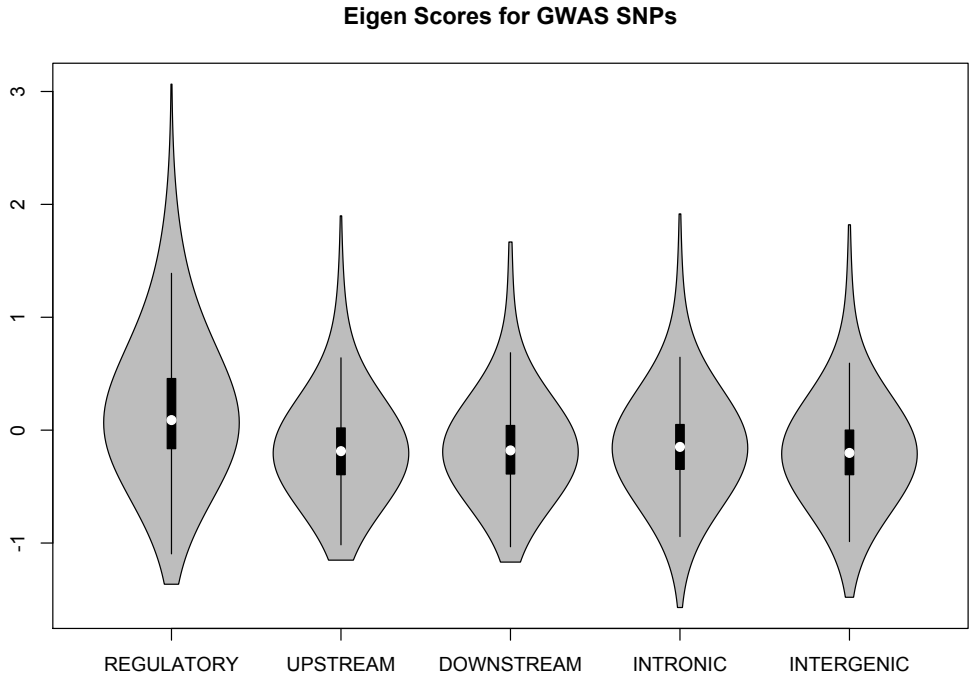
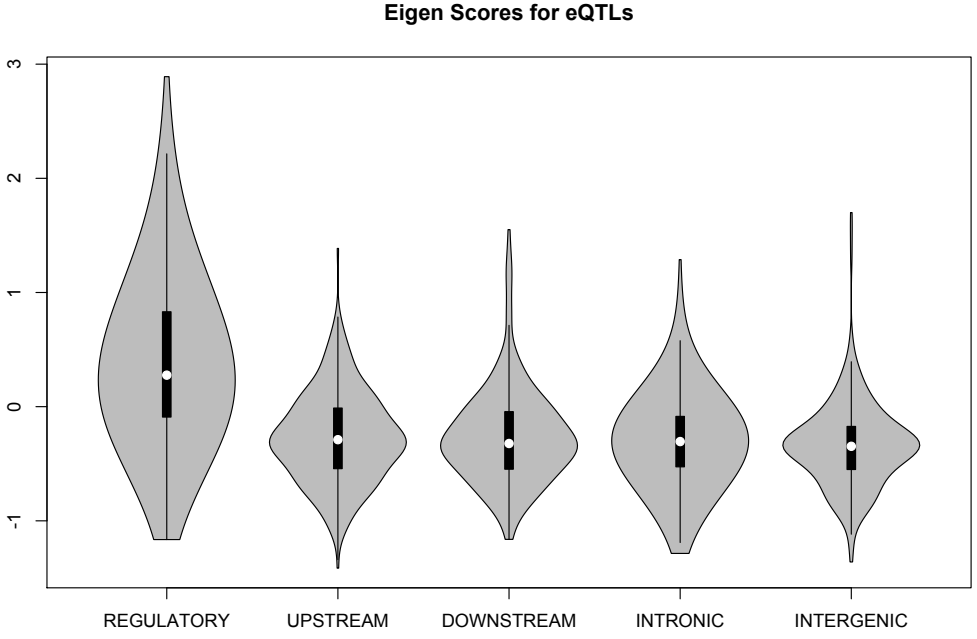


FIGURE S1. Correlation among different functional annotations for the coding variants on chromosome 1 from the training dataset.



(A) GWAS SNPs.



(B) eQTLs.

FIGURE S2. Violin plots for **Eigen** scores for GWAS SNPs and eQTLs for variants in different functional classes.

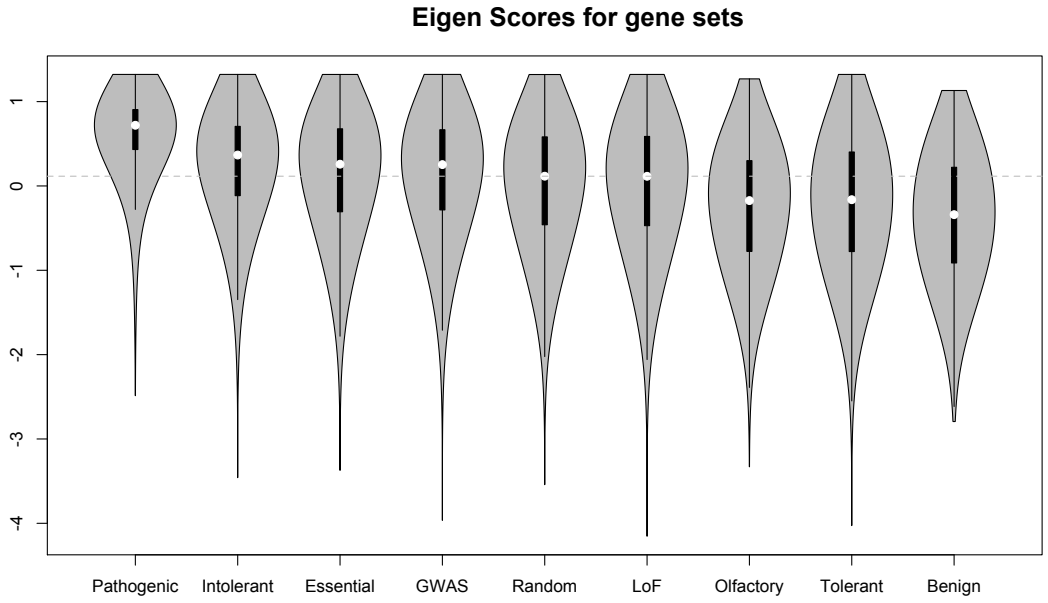


FIGURE S3. Violin plots for **Eigen** scores for non-synonymous variants in several gene sets. The horizontal line corresponds to the median **Eigen** score for variants in random genes.

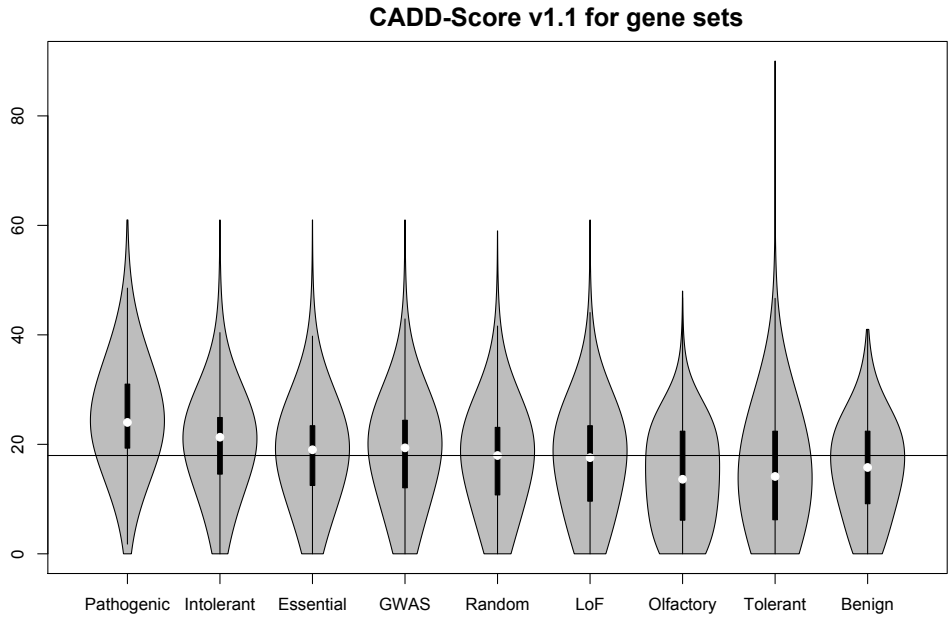
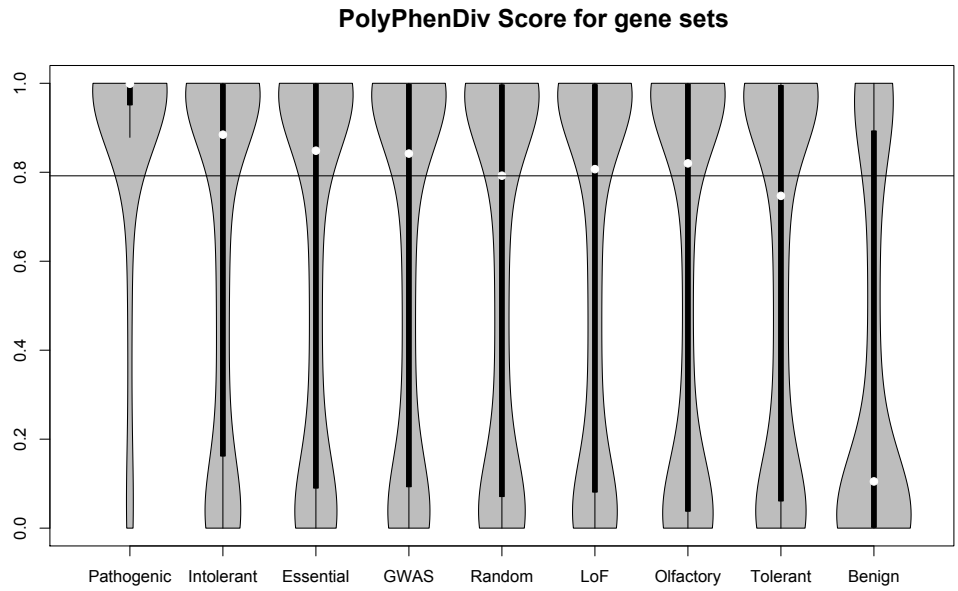
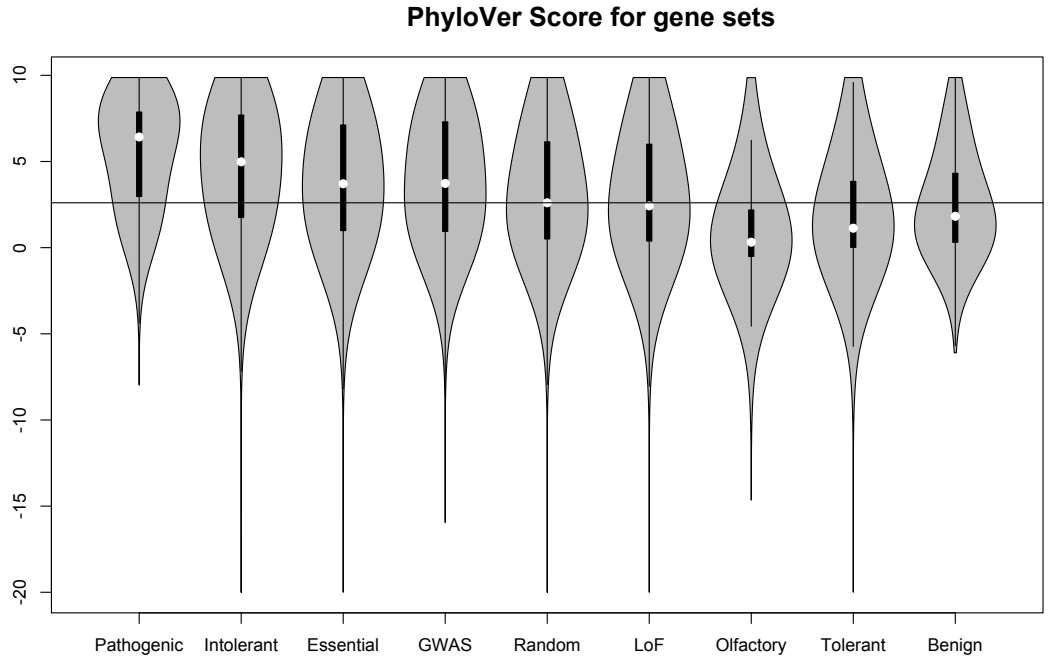


FIGURE S4. Violin plots for CADD-scores (v1.1) for non-synonymous variants in several gene sets.



(A) PolyPhenDiv.



(B) PhyloVer.

FIGURE S5. Violin plots for PolyPhenDiv and PhyloVer scores for variants in several gene sets.

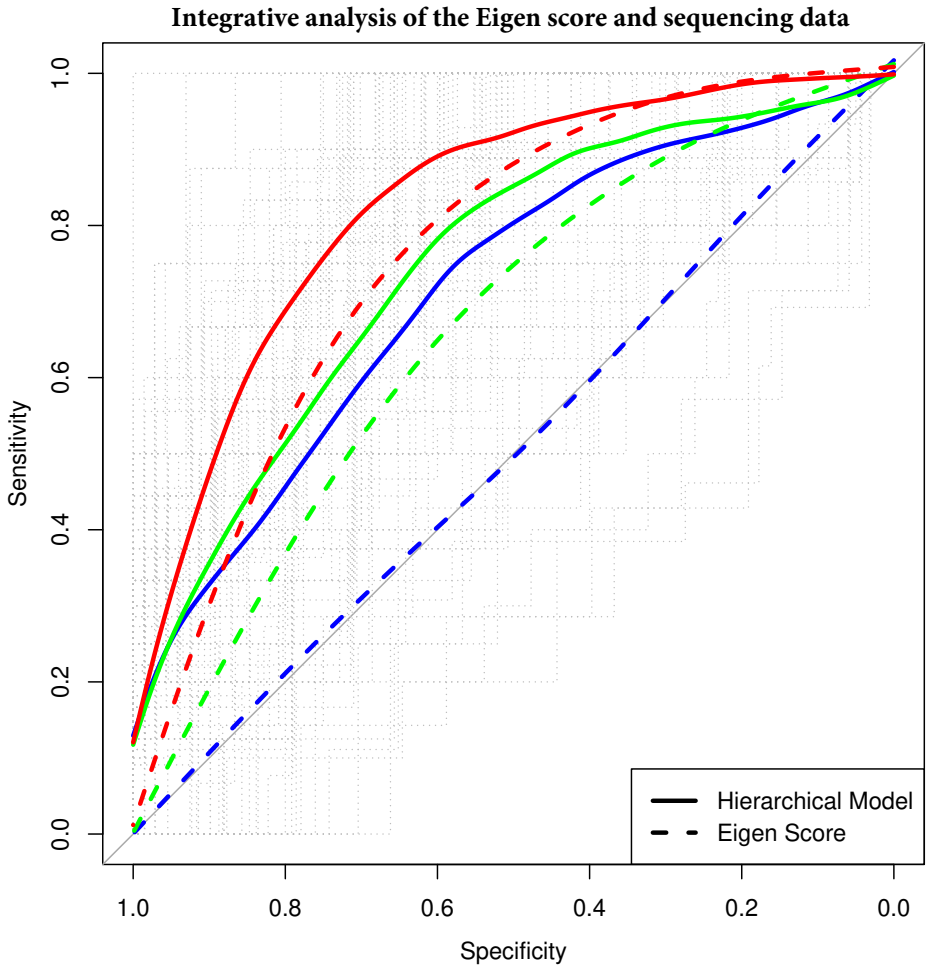


FIGURE S6. ROC curves of the z-values estimated from a hierarchical model with the **Eigen** scores included as a functional predictor (solid curves) and ROC curves based on the ranking of the **Eigen** scores (dashed curves); associations between the **Eigen** score and the causal status of a variant vary, with relative risks of 1.1 (blue), 2 (green), and 4 (red); estimates are averaged across 100 simulations.

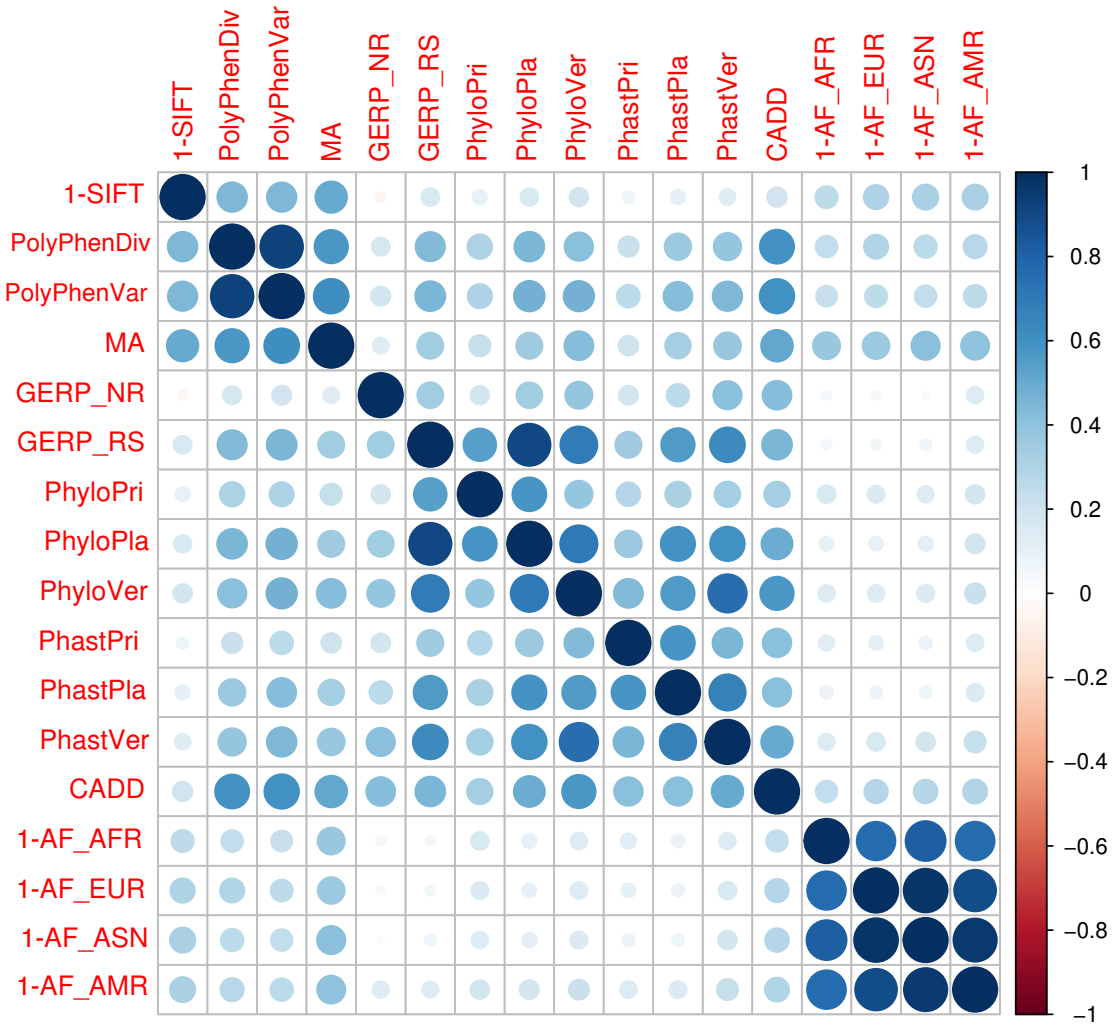


FIGURE S7. Correlation among different functional annotations with CADD score (v1.0) included (non-synonymous coding variants).

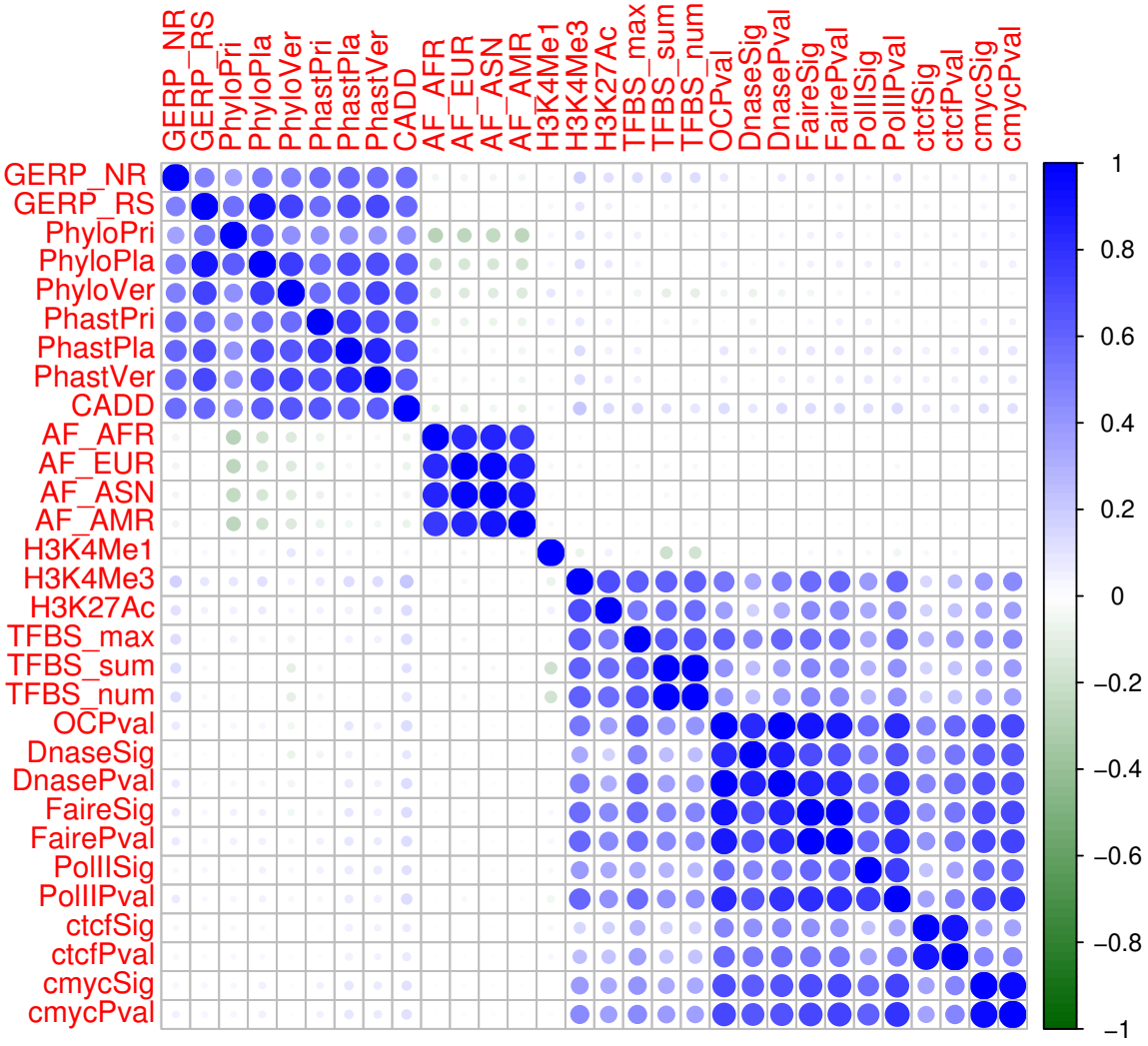


FIGURE S8. Correlation among different functional annotations with CADD score (v1.0) included (noncoding and synonymous coding variants).

Annotation type	Annotations	Source	Variant type
Protein function scores	SIFT, PolyPhenDiv, PolyPhenVar, MA	dbNSFP v2.7	Coding Non-synonymous
Conservation scores	GERP_NR, GERP_RS, PhyloPri, PhyloPla, PhyloVer, PhastPri, PhastPla, PhastVer	UCSC Genome Browser	Coding and Non-coding
AF by population	AF_AFR, AF_EUR, AF_ASN, AF_AMR	1000 Genomes Project	Coding and Non-coding
Histone marks	H3K4Me1, H3K4Me3, H3K27Ac	UCSC Genome Browser	Noncoding
Transcription factor binding sites (TFBS)*	TFBS_max, TFBS_sum, TFBS_num	UCSC Genome Browser	Noncoding
Open chromatin	OCPval, DNaseSig, DNasePval, FaireSig, FairePval, PolIISig, PolIIPval, ctcfSig, ctcfPval, cmycSig, cmycPval	UCSC Genome Browser	Noncoding

TABLE S1. Annotations used in the calculation of the **Eigen** and **Eigen-PC** scores. * The 'Txn Factor ChIP' track in UCSC genome browser gives the ChIP-seq scores for 161 different transcription factors. TFBS_sum is the sum of the scores across all 161 factors at each genomic position, TFBS_max is the maximum score across the factors at each position, and TFBS_num is the number of factors with non-zero scores at each position.

	Weights	
Annotation	**Eigen**	**Eigen-PC**
SIFT	0.049	0.042
PolyPhenDiv	0.092	0.065
PolyPhenVar	**0.093**	0.066
MA	0.092	0.060
GERP_NR	0.025	0.043
GERP_RS	0.053	0.076
PhyloPri	0.064	0.061
PhyloPla	0.070	**0.079**
PhyloVer	0.072	0.076
PhastPri	0.044	0.060
PhastPla	0.052	0.070
PhastVer	0.057	0.073
AF_AFR	0.056	0.054
AF_EUR	0.064	0.059
AF_ASN	0.056	0.057
AF_AMR	0.067	0.059

TABLE S2. Rescaled weights for the individual functional annotations (non-synonymous coding variants) for the **Eigen** and **Eigen-PC** scores. In the case of **Eigen**, PolyPhenVar has the highest weight, while for **Eigen-PC**, PhyloPla has the highest weight.

Variant type	n	Score	AUC
Coding - missense and nonsense	16,545	**Eigen**	0.868
		Eigen-PC	0.839
		CADD-score v1.0	0.861
		CADD-score v1.1	0.776
Coding - missense	12,749	**Eigen**	0.864
		Eigen-PC	0.847
		CADD-score v1.0	0.837
		CADD-score v1.1	0.763
		SIFT	0.770
		PolyPhenDiv	**0.903**
		PolyPhenVar	0.901
		MA	0.789
		GERP_NR	0.552
		GERP_RS	0.694
		PhyloPri	0.598
		PhyloPla	0.719
		PhyloVer	0.803
		PhastPri	0.639
		PhastPla	0.697
		PhastVer	0.722
Noncoding and Synonymous Coding	111	**Eigen**	0.785
		Eigen-PC	0.614
		CADD-score v1.0	0.777
		CADD-score v1.1	0.750
		GWAVA (TSS)	0.690
		GERP_NR	0.610
		GERP_RS	**0.815**
		PhyloPri	0.765
		PhyloPla	0.803
		PhyloVer	0.810
		PhastPri	0.706
		PhastPla	0.751
		PhastVer	0.799
		H3K4Me1	0.650
		H3K4Me3	0.610
		H3K27Ac	0.620
		TFBS_max	0.648
		TFBS_sum	0.653
		TFBS_num	0.655
		OCPval	0.518
		DNaseSig	0.515
		DNasePval	0.516
		FaireSig	0.520
		FairePval	0.550
		PolIISig	0.549
		PolIIPval	0.586
		ctcfSig	0.535
		ctcfPval	0.547
		cmycSig	0.517
		cmycPval	0.573

TABLE S3. AUC values for discriminating between ClinVar pathogenic and benign variants, using the proposed **Eigen** and **Eigen-PC** scores, two versions of the CADD-score, the GWAVA score (for noncoding variants) and individual functional scores. The best single annotation is highlighted.

Score	*MLL2*	*CFTR*	*BRCA1*	*BRCA2*
SIFT	**6.80E-15**	4.28E-17	**1.04E-05**	2.11E-01
PolyPhenDiv	6.95E-11	1.30E-29	1.58E-03	6.20E-02
PolyPhenVar	1.00E-10	**4.77E-36**	2.00E-04	1.14E-01
MA	4.60E-10	4.35E-24	7.78E-05	**9.53E-03**
GERP_NR	3.01E-04	1.09E-01	1.69E-01	8.02E-01
GERP_RS	3.23E-04	3.46E-16	1.07E-01	7.24E-01
PhyloPri	3.44E-05	1.81E-11	1.19E-02	6.36E-01
PhyloPla	4.18E-06	1.48E-16	5.32E-03	9.20E-01
PhyloVer	1.35E-10	7.03E-32	8.26E-01	8.76E-01
PhastPri	2.06E-11	3.30E-16	7.55E-01	7.64E-01
PhastPla	1.61E-07	1.20E-20	8.74E-01	5.96E-01
PhastVer	3.09E-06	1.59E-19	6.55E-01	9.92E-01
Eigen	3.10E-13	2.80E-37	4.00E-03	2.30E-01
Eigen-PC	5.10E-13	9.62E-37	1.64E-02	3.53E-01
CADD-score v1.0	2.80E-02	7.90E-35	5.00E-03	3.60E-01
CADD-score v1.1	2.80E-06	1.70E-21	1.40E-03	1.80E-02

TABLE S4. P values (Wilcoxon rank-sum test) for missense variants in four Mendelian genes. Results are shown for individual functional annotations, as well as for the meta-scores, **Eigen**, **Eigen-PC** and two versions of the CADD score. The best single annotation is highlighted for each gene.

Score	ASD	ASD-FMRP	EPI	ID	SCZ
SIFT	2.94E-01	1.28E-02	2.35E-01	1.28E-02	8.09E-01
PolyPhenDiv	**5.40E-02**	7.20E-03	1.10E-02	7.20E-03	6.97E-01
PolyPhenVar	9.50E-02	9.40E-03	**3.00E-03**	9.40E-03	5.34E-01
MA	7.20E-02	**1.00E-04**	8.00E-03	**1.00E-04**	4.33E-01
GERP_NR	8.60E-02	3.53E-01	5.50E-02	3.53E-01	5.68E-01
GERP_RS	9.08E-01	1.35E-02	3.39E-01	1.35E-02	6.50E-01
PhyloPri	9.91E-01	1.97E-01	8.62E-01	1.97E-01	1.78E-01
PhyloPla	3.62E-01	9.70E-03	1.29E-01	9.70E-03	7.52E-01
PhyloVer	5.15E-01	4.30E-03	9.00E-03	4.30E-03	4.37E-01
PhastPri	9.82E-01	8.00E-04	7.03E-01	8.00E-04	**9.50E-02**
PhastPla	9.27E-01	1.30E-02	4.55E-01	1.30E-02	6.27E-01
PhastVer	1.95E-01	9.20E-03	1.78E-01	9.20E-03	6.12E-01
Eigen	9.00E-02	3.20E-04	6.00E-03	6.70E-05	6.30E-01
Eigen-PC	1.50E-01	9.41E-05	1.32E-02	6.01E-05	5.84E-01
CADD-score v1.0	7.40E-01	4.20E-02	8.10E-02	3.50E-05	9.80E-01
CADD-score v1.1	5.80E-01	1.70E-02	1.70E-01	3.30E-02	2.80E-02

TABLE S5. P values (Wilcoxon rank-sum test) for *de novo* missense mutations in ASD, EPI, ID, and SCZ studies. ASD-FMRP analyses are based on *de novo* mutations in ASD cases that hit FMRP targets. Results are shown for individual functional annotations, as well as for the meta-scores, **Eigen**, **Eigen-PC** and two versions of the CADD score. The best single annotation is highlighted for each dataset.

	Weights	
Annotation	**Eigen**	**Eigen-PC**
GERP_NR	**0.208**	0.028
GERP_RS	0.027	0.008
PhyloPri	0.079	0.010
PhyloPla	0.087	0.012
PhyloVer	0.059	0.011
PhastPri	0.083	0.014
PhastPla	0.105	0.019
PhastVer	0.080	0.015
AF_AFR	0.013	0.001
AF_EUR	0.011	0.002
AF_ASN	0.010	0.001
AF_AMR	0.012	0.002
H3K4Me1	0.009	0.018
H3K4Me3	0.019	0.047
H3K27Ac	0.010	0.038
TFBS_max	0.015	0.054
TFBS_sum	0.008	0.047
TFBS_num	0.008	0.047
OCPval	0.015	**0.066**
DnaseSig	0.014	0.060
DnasePval	0.014	0.065
FaireSig	0.012	0.064
FairePval	0.012	0.064
PolIISig	0.010	0.052
PolIIPval	0.012	0.063
ctcfSig	0.008	0.036
ctcfPval	0.011	0.044
cmycSig	0.011	0.057
cmycPval	0.010	0.058

TABLE S6. Rescaled weights for the individual functional annotations (noncoding and synonymous coding variants). In the case of **Eigen**, GERP_NR has the highest weight, while for **Eigen-PC**, OCPval has the highest weight.

Score	GWAS_tag	GWAS_other	GWAS_matched	eQTL_tag	eQTL_other
GERP_NR	2.46E-03	1.39E-03	1.37E-03	4.60E-03	4.04E-04
GERP_RS	1.12E-01	1.19E-01	1.90E-03	3.39E-01	2.18E-01
PhyloPri	6.62E-01	1.19E-01	2.71E-01	6.07E-01	4.95E-01
PhyloPla	2.00E-01	2.45E-02	4.65E-03	6.87E-01	2.31E-01
PhyloVer	4.73E-01	1.87E-01	1.70E-02	5.84E-01	5.42E-01
PhastPri	1.53E-02	6.14E-03	9.99E-05	1.62E-01	2.17E-01
PhastPla	8.05E-02	9.67E-02	6.90E-02	6.03E-03	5.99E-04
PhastVer	2.47E-01	5.03E-01	1.14E-01	1.10E-01	2.15E-02
H3K4Me1	8.82E-03	4.11E-04	**4.02E-11**	7.07E-07	4.99E-06
H3K4Me3	1.06E-04	4.06E-08	1.97E-05	**2.29E-24**	**3.84E-25**
H3K27Ac	2.56E-03	2.87E-05	3.05E-07	3.86E-06	1.53E-06
TFBS_max	5.40E-04	1.16E-06	1.06E-06	5.61E-18	4.27E-18
TFBS_sum	5.05E-05	**5.66E-09**	7.25E-07	3.44E-21	1.20E-20
TFBS_num	**4.92E-05**	6.74E-09	9.15E-07	7.25E-21	7.76E-21
OCPval	3.31E-03	9.22E-08	2.70E-06	1.42E-18	3.11E-23
DnaseSig	3.42E-03	4.85E-08	2.09E-06	1.06E-18	4.91E-24
DnasePval	3.42E-03	4.30E-08	4.12E-06	9.13E-19	5.70E-24
FaireSig	5.38E-03	7.95E-07	3.30E-06	9.94E-13	1.19E-17
FairePval	1.26E-02	4.35E-06	7.01E-05	6.81E-13	2.99E-17
PolIISig	7.82E-03	6.00E-06	1.77E-05	6.74E-14	1.48E-18
PolIIPval	5.80E-03	6.70E-06	3.13E-04	6.16E-16	3.14E-21
ctcfSig	5.07E-03	3.06E-06	2.16E-05	4.05E-13	1.17E-19
ctcfPval	3.71E-02	1.12E-04	4.87E-04	2.30E-11	8.39E-17
cmycSig	9.32E-02	3.58E-05	2.65E-04	2.10E-11	4.22E-16
cmycPval	5.16E-02	1.31E-04	4.13E-04	3.34E-13	3.97E-18
Eigen	1.28E-05	1.66E-09	6.92E-08	1.89E-10	5.90E-13
Eigen-PC	4.06E-06	2.09E-13	3.51E-13	7.08E-23	2.61E-27
CADD-score v1.0	5.96E-04	2.08E-06	1.08E-04	3.13E-04	2.80E-04
CADD-score v1.1	2.01E-04	8.58E-07	5.25E-07	4.36E-05	2.13E-05
GWAVA (Region)	1.17E-03	2.82E-09	8.30E-10	1.24E-01	2.73E-02
GWAVA (TSS)	4.10E-06	7.45E-13	2.56E-09	1.39E-03	7.34E-08

TABLE S7. P values (Wilcoxon rank-sum test) for GWAS SNPs and eQTLs, for individual annotations and several meta-scores. Comparisons are shown between GWAS index SNPs and tag SNPs hitting regulatory elements (GWAS_tag and GWAS_other). Also shown are comparisons between GWAS index SNPs and control SNPs matched for frequency, functional consequence, and GWAS array availability (GWAS_matched). Additionally, comparisons between eQTLs and tag SNPs hitting regulatory elements are shown (eQTL_tag and eQTL_other). The best single annotation is highlighted for each dataset.

Annotation	Regulatory	Intronic	Downstream	Upstream	Noncoding Change	3'UTR	5'UTR	Intergenic	Synonymous
GERP_NR	1.14E-69	**≤2.22E-308**	**1.01E-155**	3.83E-80	2.69E-04	**5.67E-05**	**2.79E-07**	3.61E-01	7.12E-02
GERP_RS	1.94E-04	1.27E-13	5.44E-01	9.60E-01	5.31E-01	9.64E-01	6.98E-01	5.86E-01	3.48E-01
PhyloPri	3.98E-04	4.98E-02	1.14E-03	3.00E-01	7.76E-18	1.71E-01	7.21E-01	7.93E-04	**2.16E-03**
PhyloPla	2.19E-07	1.17E-29	8.29E-03	1.50E-02	1.74E-04	4.49E-01	5.04E-01	4.00E-04	1.27E-01
PhyloVer	5.85E-05	9.88E-12	6.55E-01	4.11E-01	1.20E-03	5.92E-01	7.42E-01	1.93E-02	3.64E-02
PhastPri	3.52E-02	6.79E-01	2.18E-03	3.48E-03	1.64E-07	1.69E-01	3.14E-01	6.00E-04	1.54E-01
PhastPla	1.68E-03	1.19E-83	7.01E-17	2.30E-12	2.73E-11	2.19E-01	4.76E-01	3.29E-03	1.50E-01
PhastVer	5.69E-08	1.10E-74	3.98E-20	5.22E-13	6.60E-01	4.81E-01	4.76E-01	1.27E-02	1.53E-01
H3K4Me1	3.54E-10	8.49E-01	3.32E-37	5.74E-43	5.04E-10	9.25E-01	1.66E-01	**1.19E-18**	7.80E-03
H3K4Me3	1.52E-201	1.17E-06	3.21E-17	1.19E-47	2.49E-15	3.24E-02	8.07E-01	4.34E-03	2.31E-02
H3K27Ac	8.78E-79	7.38E-13	6.91E-68	3.05E-46	5.41E-09	6.44E-03	5.88E-01	6.58E-10	5.12E-01
TFBS_max	6.84E-112	1.13E-04	4.45E-50	1.59E-31	6.95E-14	1.93E-02	6.52E-01	1.94E-03	1.81E-01
TFBS_sum	5.83E-116	1.16E-04	7.47E-50	1.53E-31	1.07E-13	1.82E-02	6.50E-01	1.89E-03	1.76E-01
TFBS_num	1.05E-106	1.33E-04	2.16E-48	8.85E-31	3.52E-13	2.08E-02	6.54E-01	1.78E-03	2.13E-01
OCPval	9.16E-203	1.20E-68	2.87E-118	1.50E-64	1.45E-20	1.06E-01	6.65E-01	4.78E-15	4.27E-01
DnaseSig	2.97E-216	1.02E-69	1.89E-121	7.22E-67	1.09E-21	9.14E-02	6.52E-01	1.75E-15	5.33E-01
DnasePval	6.12E-214	7.37E-69	5.52E-123	1.53E-67	7.49E-21	1.53E-01	6.11E-01	4.93E-16	4.51E-01
FaireSig	4.37E-162	2.74E-67	1.10E-116	1.05E-64	1.74E-20	8.77E-02	7.45E-01	3.40E-15	6.13E-01
FairePval	2.90E-148	2.39E-31	7.07E-81	3.07E-47	3.82E-19	2.76E-01	5.35E-01	1.01E-06	5.48E-01
PolIISig	2.74E-247	3.26E-61	1.67E-120	1.68E-68	9.54E-31	3.77E-01	7.47E-01	7.20E-09	1.32E-01
PolIIPval	**≤2.22E-308**	1.17E-44	1.00E-136	**9.68E-86**	**8.13E-35**	7.37E-02	7.80E-01	1.31E-02	2.35E-01
ctcfSig	1.36E-195	3.72E-85	4.70E-135	1.91E-75	3.84E-25	2.05E-03	4.41E-01	1.57E-13	3.49E-01
ctcfPval	4.02E-222	1.36E-47	7.91E-92	3.33E-64	7.65E-26	2.19E-02	6.40E-01	1.35E-06	1.17E-01
cmycSig	2.60E-195	7.89E-62	2.16E-112	4.44E-67	7.17E-28	2.44E-01	9.45E-01	7.26E-09	1.96E-01
cmycPval	1.59E-219	1.60E-34	3.66E-78	7.07E-43	4.31E-23	4.55E-01	7.91E-01	3.40E-03	3.14E-01
Eigen	2.02E-165	2.4E-155	2.73E-92	1.28E-52	2.51E-07	6.94E-03	1.14E-02	1.49E-02	1.09E-01
Eigen-PC	5.13E-264	2.13E-112	3.04E-128	2.01E-84	2.49E-21	4.22E-04	2.32E-01	3.97E-06	9.69E-01
CADD-score v1.0	1.05E-71	2.89E-61	4.31E-36	7.90E-24	1.51E-01	1.06E-05	6.43E-02	1.08E-06	8.25E-01
CADD-score v1.1	2.7E-50	1.09E-10	1.83E-28	3.21E-17	4.84E-05	3.37E-01	1.15E-01	6.30E-16	2.88E-01

TABLE S8. P values (Wilcoxon rank-sum test) for somatic mutations (recurrent vs. non-recurrent) in the COSMIC database, for individual functional scores and several meta-scores. 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 single annotation is highlighted for each dataset.

Score	ClinVar	*MLL2*	*CFTR*	*BRCA1*	*BRCA2*	ASD	ASD-FMRP	EPI	ID	SCZ	Median
Eigen	3	2	1	6	6	4	3	2	3	11	3
Eigen-PC	4	3	2	10	7	6	1	6	2	8	5
CADD v1.0	5	16	4	7	8	12	14	8	1	16	8
CADD v1.1	9	10	8	4	2	11	13	10	14	1	9.5
SIFT	8	1	11	1	5	8	10	12	11	15	9
PolyPhenDiv	1	5	6	5	3	1	6	5	7	13	5
PolyPhenVar	2	6	3	3	4	5	8	1	9	6	4.5
MA	7	8	7	2	1	2	2	3	4	4	3.5
GERP_NR	16	14	16	12	13	3	16	7	16	7	13.5
GERP_RS	13	15	14	11	11	13	12	13	13	12	13
PhyloPri	15	13	15	9	10	16	15	16	15	3	15
PhyloPla	11	12	12	8	15	9	9	9	10	14	10.5
PhyloVer	6	7	5	15	14	10	5	4	6	5	6
PhastPri	14	4	13	14	12	15	4	15	5	2	12.5
PhastPla	12	9	9	16	9	14	11	14	12	10	11.5
PhastVer	10	11	10	13	16	7	7	11	8	9	10

TABLE S9. Ranks for meta-scores and individual functional annotations, across all the datasets used in Section 2.1 on non-synonymous coding variants.

Score	ClinVar	GWAS_tag	GWAS_other	GWAS_matched	eQTL_tag	eQTL_other	Regulatory	Intronic	Downstream	Upstream	Noncoding Change	3'UTR	5'UTR	Intergenic	Synonymous	Median
Eigen	5	3	3	4	17	17	12	2	11	12	22	6	2	26	6	6
Eigen-PC	17	1	1	1	2	1	2	3	4	2	8	3	6	14	29	3
CADD v1.0	6	9	13	19	21	22	19	12	20	21	27	1	3	12	28	19
CADD v1.1	9	7	11	6	20	21	21	22	21	22	23	22	4	3	18	20
GWAVA	11	2	2	3	22	18	NA	NA	NA	NA	NA	NA	NA	NA	NA	7[1]
GERP_NR	18.5	10	23	24	23	23	20	1	1	3	25	2	1	28	5	18.5
GERP_RS	1	26	27.5	25	27	27	26	19	28	29	28	29	21	29	20	27
PhyloPri	7	30	27.5	30	29	29	27	27	25	27	13	18	22	17	1	27
PhyloPla	3	27	25	26	30	28	24	18	27	26	24	24	11	15	8	24
PhyloVer	2	29	29	27	28	30	25	21	29	28	26	27	23	27	4	27
PhastPri	10	21	24	18	26	26	29	28	26	25	21	17	7	16	12	21
PhastPla	8	24	26	28	24	24	28	5	24	24	18	19	9.5	21	10	24
PhastVer	4	28	30	29	25	25	23	6	22	23	29	26	9.5	24	11	24
H3K4Me1	14	19	22	2	18	20	22	29	19	16	19	28	5	1	2	19
H3K4Me3	18.5	6	6	15	1	2	9	23	23	13	14	11	28	23	3	13
H3K27Ac	16	11	18	5	19	19	18	20	15	15	20	5	13	8	24	16
TFBS_max	15	8	12	9	8	12	16	24	16	19	15	8	17.5	20	14	15
TFBS_sum	13	5	4	7	3	8	15	25	17	18	16	7	16	19	13	13
TFBS_num	12	4	5	8	4	7	17	26	18	20	17	9	19	18	16	16
OCPval	27	12	9	11	7	5	8	9	8	10	10	15	20	6	22	10
DNaseSig	30	13.5	8	10	6	3	6	7	6	8	7	14	17.5	4	25	8
DNasePval	29	13.5	7	13	5	4	7	8	5	6	9	16	14	2	23	8
FaireSig	26	16	10	12	14	13	13	10	9	9	11	13	24	5	27	13
FairePval	22	20	15	17	13	14	14	17	13	14	12	21	12	11	26	14
PolIISig	23	18	16	14	10	10	3	13	7	5	2	23	25	9	9	10
PolIIPval	20	17	17	21	9	6	1	15	2	1	1	12	26	25	17	15
ctcfSig	25	15	14	16	12	9	10	4	3	4	5	4	8	7	21	9
ctcfPval	24	22	20	23	16	15	4	14	12	11	4	10	15	13	7	14
cmycSig	28	25	19	20	15	16	11	11	10	7	3	20	29	10	15	15
cmycPval	21	23	21	22	11	11	5	16	14	17	6	25	27	22	19	19

TABLE S10. Ranks for meta-scores and individual functional annotations, across all the datasets used in Section 2.2 on noncoding and synonymous coding variants.

[1]based only on non-missing values

Annotation	Weights
SIFT	0.058
PolyPhenDiv	0.117
PolyPhenVar	**0.121**
MA	0.119
GERP_NR	0.018
GERP_RS	0.035
PhyloPri	0.043
PhyloPla	0.046
PhyloVer	0.055
PhastPri	0.032
PhastPla	0.037
PhastVer	0.053
CADD	0.079
AF_AFR	0.042
AF_EUR	0.043
AF_ASN	0.041
AF_AMR	0.062

TABLE S11. Rescaled weights for the individual functional annotations, when CADD v1.0 is included as one of the scores (non-synonymous coding variants). In the case of **Eigen+C**, PolyPhenVar has the highest weight.

Annotation	Weights
GERP_NR	0.161
GERP_RS	0.024
PhyloPri	0.066
PhyloPla	0.070
PhyloVer	0.050
PhastPri	0.067
PhastPla	0.085
PhastVer	0.064
CADD	**0.176**
AF_AFR	0.009
AF_EUR	0.009
AF_ASN	0.008
AF_AMR	0.010
H3K4Me1	0.007
H3K4Me3	0.016
H3K27Ac	0.008
TFBS_max	0.012
TFBS_sum	0.008
TFBS_num	0.008
OCPval	0.013
DnaseSig	0.011
DnasePval	0.012
FaireSig	0.012
FairePval	0.011
PolIISig	0.008
PolIIPval	0.010
ctcfSig	0.008
ctcfPval	0.010
cmycSig	0.010
cmycPval	0.009

TABLE S12. Rescaled weights for the individual functional annotations, when CADD v1.0 is included as one of the component annotations (noncoding and synonymous coding variants). In the case of **Eigen+C**, CADD has the highest weight.

Disease	n	*De novo* Variant type	Score	P value
ASD	2,027	Missense and Nonsense	**Eigen**	6.0E-03
			Eigen+C	5.0E-02
	1,753	Missense only	**Eigen**	9.0E-02
			Eigen+C	5.5E-01
ASD-FMRP	132	Missense and Nonsense	**Eigen**	4.2E-05
			Eigen+C	2.6E-03
	113	Missense only	**Eigen**	3.2E-04
			Eigen+C	1.7E-02
EPI	210	Missense and Nonsense	**Eigen**	3.1E-03
			Eigen+C	1.5E-02
	184	Missense only	**Eigen**	6.0E-03
			Eigen+C	6.3E-02
ID	114	Missense and Nonsense	**Eigen**	1.7E-06
			Eigen+C	6.4E-07
	99	Missense only	**Eigen**	6.7E-05
			Eigen+C	7.8E-06
SCZ	636	Missense and Nonsense	**Eigen**	9.9E-01
			Eigen+C	1.9E-01
	573	Missense only	**Eigen**	6.3E-01
			Eigen+C	9.8E-01

TABLE S13. 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. Results for the **Eigen** score and the **Eigen+C** score are shown.

Gene	n	Variant type	Score	P value
MLL2	108	Missense and Nonsense	**Eigen**	1.1E-56
			Eigen+C	4.8E-48
	31	Missense	**Eigen**	3.1E-13
			Eigen+C	1.4E-04
CFTR	160	Missense and Nonsense	**Eigen**	1.3E-69
			Eigen+C	1.1E-69
	92	Missense	**Eigen**	2.8E-37
			Eigen+C	3.2E-33
BRCA1	125	Missense and Nonsense	**Eigen**	2.5E-38
			Eigen+C	2.0E-36
	28	Missense	**Eigen**	4.0E-03
			Eigen+C	1.2E-02
BRCA2	110	Missense and Nonsense	**Eigen**	9.8E-28
			Eigen+C	5.1E-50
	13	Missense	**Eigen**	2.3E-01
			Eigen+C	3.0E-01

TABLE S14. P-values (Wilcoxon rank-sum test) for *MLL2, CFTR, BRCA1, BRCA2*, contrasting pathogenic variants with benign variants in the ClinVar database. Results for the **Eigen** score and **Eigen+C** score are shown.

Variant Category	n-rec	n-nonrec	**Eigen**	**Eigen-PC**	**Eigen+C**
Regulatory	21,279	428,398	2.02E-165	5.13E-264	7.51E-181
Intronic	85,502	2,093,158	2.40E-155	2.13E-112	6.75E-78
Downstream	15,956	318,967	2.73E-92	3.04E-128	3.16E-82
Upstream	14,636	309,615	1.28E-52	2.01E-84	6.46E-59
Noncoding_Change	4,903	66,717	2.51E-07	2.49E-21	1.16E-11
3Prime_UTR	2,236	28,261	6.94E-03	4.22E-04	1.68E-05
5Prime_UTR	417	3,908	1.14E-02	2.32E-01	1.31E-01
Intergenic	75,327	2,182,466	1.49E-02	3.97E-06	1.34E-05
Synonymous	434	2,388	1.09E-01	9.69E-01	6.98E-01

TABLE S15. 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. Results are shown for **Eigen**, **Eigen-PC** and **Eigen+C** scores.

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
			CADD-reduced	5.7E-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
			CADD-reduced	3.5E-01
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
			CADD-reduced	4.2E-02
	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
			CADD-reduced	2.4E-02
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
			CADD-reduced	1.2E-02
	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
			CADD-reduced	1.9E-02
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
			CADD-reduced	2.3E-04
	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
			CADD-reduced	1.0E-04
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
			CADD-reduced	6.6E-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
			CADD-reduced	3.5E-01

TABLE S16. 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. CADD-reduced corresponds to the CADD-score trained on the same set of annotations as **Eigen**.

Gene	n	Variant type	Score	P value
MLL2	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
			CADD-reduced	3.8E-06
CFTR	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
			CADD-reduced	4.3E-33
BRCA1	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
			CADD-reduced	3.9E-03
BRCA2	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
			CADD-reduced	3.7E-02

TABLE S17. P values (Wilcoxon rank-sum test) for *MLL2, CFTR, BRCA1, BRCA2*, contrasting pathogenic variants with benign variants in the ClinVar database. CADD-reduced corresponds to the CADD-score trained on the same set of annotations as **Eigen**.

Dataset	n	Comparison	Score	P value
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
			CADD-reduced	2.9E-06
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
			CADD-reduced	4.0E-04
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
			CADD-reduced	1.3E-04

TABLE S18. P values (Wilcoxon rank-sum test) for GWAS SNPs and eQTLs. CADD-reduced corresponds to the CADD-score trained on the same set of annotations as **Eigen**. Comparisons are shown 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.

Gene	n	Variant type	Score	P value
MLL2	108	Missense and Nonsense	**Eigen**	1.1E-56
			Eigen-PC1	1.6E-50
			Eigen-PC2	2.1E-26
	31	Missense	**Eigen**	3.1E-13
			Eigen-PC1	5.1E-13
			Eigen-PC2	3.9E-11
CFTR	160	Missense and Nonsense	**Eigen**	1.3E-69
			Eigen-PC1	8.2E-65
			Eigen-PC2	1.4E-45
	92	Missense	**Eigen**	2.8E-37
			Eigen-PC1	9.6E-37
			Eigen-PC2	3.0E-31
BRCA1	125	Missense and Nonsense	**Eigen**	2.5E-38
			Eigen-PC1	6.0E-25
			Eigen-PC2	2.0E-03
	28	Missense	**Eigen**	4.0E-03
			Eigen-PC1	1.6E-02
			Eigen-PC2	1.8E-01
BRCA2	110	Missense and Nonsense	**Eigen**	9.8E-28
			Eigen-PC1	3.3E-14
			Eigen-PC2	8.8E-01
	13	Missense	**Eigen**	2.3E-01
			Eigen-PC1	3.5E-01
			Eigen-PC2	9.1E-01

TABLE S19. P values (Wilcoxon rank-sum test) for *MLL2, CFTR, BRCA1, BRCA2*, contrasting pathogenic variants with benign variants in the ClinVar database. **Eigen-PC1** corresponds to the score derived using the first principal component as a weight, while **Eigen-PC2** is based on the second principal component.

Disease	n	Variant type	Score	P value
ASD	2,027	Missense and Nonsense	**Eigen**	6.0E-03
			Eigen-PC1	1.6E-02
			Eigen-PC2	1.8E-01
	1,753	Missense only	**Eigen**	9.0E-02
			Eigen-PC1	1.5E-01
			Eigen-PC2	3.0E-01
ASD-FMRP	132	Missense and Nonsense	**Eigen**	4.2E-05
			Eigen-PC1	9.4E-06
			Eigen-PC2	4.0E-06
	113	Missense only	**Eigen**	3.2E-04
			Eigen-PC1	9.4E-05
			Eigen-PC2	1.6E-05
EPI	210	Missense and Nonsense	**Eigen**	3.1E-03
			Eigen-PC1	5.0E-03
			Eigen-PC2	2.2E-02
	184	Missense only	**Eigen**	6.0E-03
			Eigen-PC1	1.3E-02
			Eigen-PC2	1.4E-02
ID	114	Missense and Nonsense	**Eigen**	1.7E-06
			Eigen-PC1	1.1E-06
			Eigen-PC2	1.2E-05
	99	Missense only	**Eigen**	6.7E-05
			Eigen-PC1	6.0E-05
			Eigen-PC2	1.6E-04
SCZ	636	Missense and Nonsense	**Eigen**	9.9E-01
			Eigen-PC1	9.8E-01
			Eigen-PC2	9.7E-01
	573	Missense only	**Eigen**	6.3E-01
			Eigen-PC1	5.8E-01
			Eigen-PC2	6.8E-01

TABLE S20. 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. **Eigen-PC1** corresponds to the score derived using the first principal component as a weight, while **Eigen-PC2** is based on the second principal component.

Dataset	n	Comparison	Score	P value
GWAS	2,115	Regulatory GWAS vs. Tag SNPs	**Eigen**	1.2E-05
			Eigen-PC1	4.0E-06
			Eigen-PC2	2.8E-04
GWAS	2,115	Regulatory GWAS vs. Other SNPs	**Eigen**	1.6E-09
			Eigen-PC1	2.0E-13
			Eigen-PC2	9.6E-09
GWAS	10,718	GWAS vs. Matched Controls	**Eigen**	6.9E-08
			Eigen-PC1	3.5E-13
			Eigen-PC2	2.8E-07
eQTLs	676	Regulatory eQTLs vs. Tag SNPs	**Eigen**	1.8E-10
			Eigen-PC1	7.0E-23
			Eigen-PC2	8.8E-13
eQTLs	676	Regulatory eQTLs vs. Other SNPs	**Eigen**	5.9E-13
			Eigen-PC1	2.6E-27
			Eigen-PC2	9.1E-18

TABLE S21. P values (Wilcoxon rank-sum test) for GWAS SNPs and eQTLs. Comparisons are shown 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. **Eigen-PC1** corresponds to the score derived using the first principal component as a weight, while **Eigen-PC2** is based on the second principal component.