A spectral approach integrating functional genomic annotations for coding and noncoding variants

Iuliana Ionita-Laza[1,8], Kenneth McCallum[1,8], Bin Xu[2] & Joseph D Buxbaum[3–7]

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

The tremendous progress achieved recently in massively parallel sequencing technologies enables investigators to efficiently obtain genetic information down to single-base resolution on a genome-wide scale[1–3]. 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 Genome Evolutionary Rate Profiling (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 Genotype-Tissue Expression (GTEx) project is building a massive biospecimen repository to identify tissue-specific expression quantitative trait loci (eQTLs) and splicing QTLs using more than 40 tissues and over 1,000 samples[8]. Hence, there is now 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 because they can help predict the functional effect of a variant and can be further combined with population-level genetic data to identify the variants at a locus of interest that are more likely to have a causal role in disease[9–12]. 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, the effect of an amino acid change on the protein function or structure in the case of coding variants, or the potential effect on regulatory elements in the case of noncoding variants. 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 (for example, 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[13–15].

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 for loci influencing risk to particular complex diseases[16,17]. Other studies have focused on the integration of many different functional annotations into one score of functional importance. For example, Kircher *et al.*[18] proposed a supervised approach (support vector machine, or SVM) to train a discriminative model. Specifically, their approach begins with two sets of variants, one labeled as deleterious and a second labeled as benign, and a model is fit 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 that are fixed in the human genome are assumed to be mostly benign. These are compared to *de novo* variants generated randomly on the basis of models of mutation rates across the genome. Although the proposed aggregate score, combined annotation-dependent depletion (CADD), has notable advantages as described in ref. 18,

[1]Department of Biostatistics, Columbia University, New York, New York, USA. [2]Department of Psychiatry, Columbia University, New York, New York, USA. [3]Seaver Autism Center for Research and Treatment, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [4]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [5]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [6]Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [7]Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [8]These authors contributed equally to this work. Correspondence should be addressed to I.I.-L. (ii2135@columbia.edu).

Received 17 July 2015; accepted 30 November 2015; published online 4 January 2016; doi:10.1038/ng.3477

it has several potential limitations. In particular, the quality of the resulting model depends on the quality of the labeled data used in the training stage. First of all, the two sets used in the training data are unlikely to be sharply divided into benign and deleterious variants; specifically, the set of simulated *de novo* variants (labeled as deleterious) likely contains a substantial proportion of benign variants. Second, the SVM is trained to distinguish between variants that may be under evolutionary constraint and those that are likely under neutral selection; hence, for disease-associated 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 Genome-Wide Annotation of Variants (GWAVA) for noncoding variants[19], which uses as a training data set the "regulatory mutations" from the public release of the Human Gene Mutation Database as deleterious variants and common (minor allele frequency ≥1%) single-nucleotide variants (SNVs) from the 1000 Genomes Project as benign variants.

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

Here we introduce an unsupervised spectral approach (Eigen) for scoring variants that does not make use of labeled training data. As such, its performance is not sensitive to a particular labeling of the training data set. Instead, the approach we introduce in this paper is based on training using a large set of variants, with a diverse set of annotations for each of these variants but no label as to their functional status (**Supplementary Table 1**). We assume that the variants can be partitioned into two distinct groups, functional and non-functional (although the partition is unknown to us), and that for each annotation the distribution is a two-component mixture, corresponding to the two groups. The key assumption in the Eigen approach is that of blockwise conditional independence between annotations given the true state of a variant (either functional or non-functional). This assumption implies that any correlation between annotations in different blocks is due to differences in the annotation means between functional and non-functional variants (Online Methods). Because of this, the correlation structure among the different functional annotations (**Fig. 1** and **Supplementary Fig. 1**) can be used to determine how well each annotation separates functional and non-functional variants (that is, the predictive accuracy of each annotation). Subsequently, we construct a weighted linear combination of annotations, based on these estimated accuracies. We illustrate the discriminatory ability of the proposed meta-score using numerous examples of disease-associated variants and putatively benign variants from the literature. In addition, we consider a related but conceptually simpler meta-score, Eigen-PC, which is based on eigendecomposition of the annotation covariance matrix and uses the lead eigenvector to weight the individual annotations.

RESULTS

Nonsynonymous variants

For the coding set, all variants with a match in the dbNSFP database[21], a database of all potentially nonsynonymous SNVs in the human genome, were included. Note that this excludes synonymous variants that fall in coding regions but do not alter protein sequences. Annotations for nonsynonymous variants are derived from several sources. In particular, protein function scores (SIFT, PolyPhen (PolyPhenDiv and PolyPhenVar, each trained on different data sets; PolyPhenDiv was trained on the HumDiv data set, whereas PolyPhenVar was trained using the HumVar data set[4]), and Mutation Assessor (MA)) were all taken from dbNSFP v2.7. Evolutionary conservation scores (GERP_NR and GERP_RS[5]); PhyloP primate (PhyloPri), placental mammal (PhyloPla) and vertebrate (PhyloVer); and PhastCons primate (PhastPri), placental mammal (PhastPla) and vertebrate (PhastVer)) were obtained from the UCSC Genome Browser (November 2014). Allele frequencies in four populations (African (1-AF_AFR), European (1-AF_EUR), East Asian (1-AF_ASN) and admixed American (1-AF_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 owing to high missing rates. Using the training data on ~76.7 million coding nonsynonymous variants, we calculated the weights for the different annotations (**Supplementary Table 2**). For Eigen, several protein function scores (PolyPhenDiv, PolyPhenVar and MA) had the highest weights, consistent with the expectation for coding nonsynonymous variants, followed by evolutionary conservation scores and alternate allele frequencies. For Eigen-PC, evolutionary conservation scores had higher weights than the protein function scores. Because the evolutionary conservation block is large in comparison with the other blocks (**Supplementary Fig. 1**), the evolutionary conservation block dominates the first principal component of the covariance matrix, increasing the weights in this block.

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

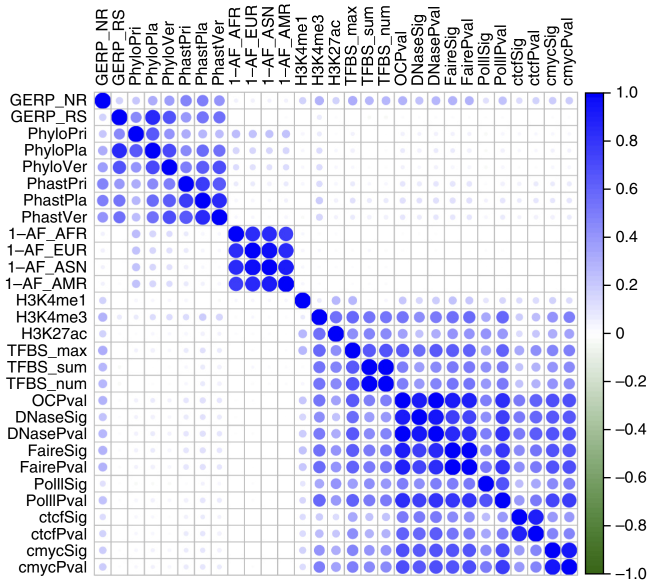

Figure 1 Correlation among different functional annotations for the noncoding variants on chromosome 1 in the training data set. The correlation plot for nonsynonymous coding variants appears in **Supplementary Figure 1**. Detailed information on the annotations is given in **Supplementary Table 1**.

Table 1 Mutations in genes for Mendelian diseases

Gene	*n*	Variant type	Score	*P* value
MLL2	108	Missense and nonsense	Eigen	1.1×10^{-56}
			Eigen-PC	1.6×10^{-50}
			CADD score v1.0	1.2×10^{-42}
			CADD score v1.1	1.3×10^{-49}
	31	Missense	Eigen	3.1×10^{-13}
			Eigen-PC	5.1×10^{-13}
			CADD score v1.0	2.8×10^{-2}
			CADD score v1.1	2.8×10^{-6}
			SIFT	6.8×10^{-15}
CFTR	160	Missense and nonsense	Eigen	1.3×10^{-69}
			Eigen-PC	8.2×10^{-65}
			CADD score v1.0	1.1×10^{-65}
			CADD score v1.1	3.1×10^{-39}
	92	Missense	Eigen	2.8×10^{-37}
			Eigen-PC	9.6×10^{-37}
			CADD score v1.0	7.9×10^{-35}
			CADD score v1.1	1.7×10^{-21}
			PolyPhenVar	4.8×10^{-36}
BRCA1	125	Missense and nonsense	Eigen	2.5×10^{-38}
			Eigen-PC	6.0×10^{-25}
			CADD score v1.0	2.2×10^{-28}
			CADD score v1.1	1.3×10^{-22}
	28	Missense	Eigen	4.0×10^{-3}
			Eigen-PC	1.6×10^{-2}
			CADD score v1.0	5.0×10^{-3}
			CADD score v1.1	1.4×10^{-3}
			SIFT	1.0×10^{-5}
BRCA2	110	Missense and nonsense	Eigen	9.8×10^{-28}
			Eigen-PC	3.3×10^{-14}
			CADD score v1.0	1.5×10^{-46}
			CADD score v1.1	7.7×10^{-40}
	13	Missense	Eigen	2.3×10^{-1}
			Eigen-PC	3.5×10^{-1}
			CADD score v1.0	3.6×10^{-1}
			CADD score v1.1	1.8×10^{-2}
			MA	9.5×10^{-3}

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

ClinVar pathogenic and benign variants

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 the following—'benign', 'likely benign', 'pathogenic' or 'likely pathogenic'—were selected for the validation set. These variants were subdivided into a nonsynonymous coding set and a synonymous coding and noncoding set. The nonsynonymous coding set consisted of all variants that matched an entry in dbNSFP, including missense, nonsense and splice-site variants. This set is intended to capture all variants that alter protein structure. The coding synonymous and noncoding set (discussed in the next section) consists of variants that do not have a match in dbNSFP.

The area under the curve (AUC) values for discriminating between nonsynonymous 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 the **Supplementary Note** for a discussion of the differences between the two versions), and the individual functional scores) are reported in **Supplementary Table 3**. For missense variants, PolyPhenDiv had the highest discrimination power (AUC = 0.903), whereas the proposed Eigen score had an AUC of 0.864 and CADD score v1.0 had an AUC of 0.837.

Mutations in genes for Mendelian diseases

MLL2 (also known as *KMT2D*), *CFTR*, and *BRCA1* and *BRCA2* are four well-known genes carrying pathogenic mutations for Kabuki syndrome, cystic fibrosis and breast cancer, respectively. We selected reported disease-associated mutations (namely pathogenic or likely pathogenic SNVs reported in the ClinVar database) in *MLL2* (n = 108 with 31 missense), *CFTR* (n = 160 with 92 missense), *BRCA1* (n = 125 with 28 missense) and *BRCA2* (n = 100 with 13 missense). *P* values obtained from the Wilcoxon rank-sum test when comparing these with benign variants in the ClinVar database are shown in **Table 1**. The overall results were highly significant for all the different methods, with the Eigen score performing better than the Eigen-PC and CADD scores in most of the cases. In particular, for missense variants in *MLL2*, the *P* value was 3.1×10^{-13} for Eigen and 5.1×10^{-13} for Eigen-PC, whereas for the two versions of CADD score the *P* values were 2.8×10^{-2} (v1.0) and 2.8×10^{-6} (v1.1). Note that, because 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, because they cause a recessive disease (cystic fibrosis), a larger proportion of them are missense as compared to mutations in the other three genes (*MLL2*, *BRCA1* and *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 performed best, although the best performing annotation in each case was a protein function score (SIFT, MA or PolyPhenVar). Results for each individual functional score are reported in **Supplementary Table 4**.

De novo mutations reported in studies of autism, schizophrenia, epileptic encephalopathies and intellectual disability

We identified all *de novo* mutations associated with autism (ASD), schizophrenia, epileptic encephalopathies and intellectual disability

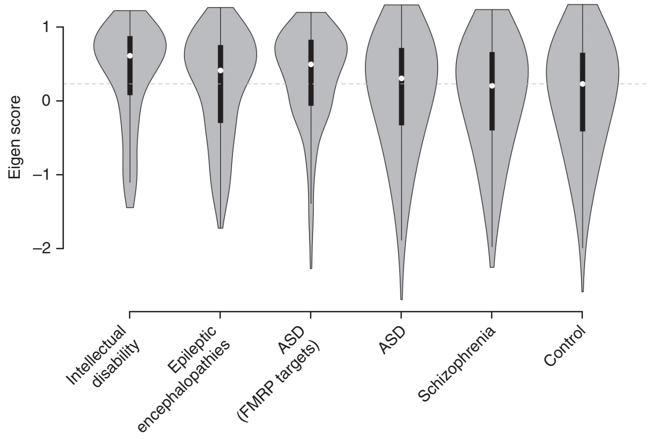

Figure 2 Violin plots showing the distribution of Eigen scores for *de novo* mutations in intellectual disability, epileptic encephalopathies, ASD (FMRP targets), ASD, schizophrenia and controls. The horizontal line corresponds to the median Eigen score for *de novo* mutations in controls (the lowest scoring set).

Table 2 ***De novo* mutations in neuropsychiatric diseases**

Disease	*n*	Variant type	Score	*P* value
ASD	2,027	Missense and nonsense	Eigen	6.0×10^{-3}
			Eigen-PC	1.6×10^{-2}
			CADD score v1.0	8.4×10^{-2}
			CADD score v1.1	3.2×10^{-1}
	1,753	Missense only	Eigen	9.0×10^{-2}
			Eigen-PC	1.5×10^{-1}
			CADD score v1.0	7.4×10^{-1}
			CADD score v1.1	5.8×10^{-1}
			PolyPhenDiv	5.4×10^{-2}
ASD (FMRP targets)	132	Missense and nonsense	Eigen	4.2×10^{-5}
			Eigen-PC	9.4×10^{-6}
			CADD score v1.0	5.5×10^{-3}
			CADD score v1.1	4.7×10^{-3}
	113	Missense only	Eigen	3.2×10^{-4}
			Eigen-PC	9.4×10^{-5}
			CADD score v1.0	4.2×10^{-2}
			CADD score v1.1	1.7×10^{-2}
			MA	1.0×10^{-4}
EPI	210	Missense and nonsense	Eigen	3.1×10^{-3}
			Eigen-PC	5.0×10^{-3}
			CADD score v1.0	4.0×10^{-2}
			CADD score v1.1	2.0×10^{-1}
	184	Missense only	Eigen	6.0×10^{-3}
			Eigen-PC	1.3×10^{-2}
			CADD score v1.0	8.1×10^{-2}
			CADD score v1.1	1.7×10^{-1}
			PolyPhenVar	3.0×10^{-3}
ID	114	Missense and nonsense	Eigen	1.7×10^{-6}
			Eigen-PC	1.1×10^{-6}
			CADD score v1.0	3.7×10^{-6}
			CADD score v1.1	9.5×10^{-3}
	99	Missense only	Eigen	6.7×10^{-5}
			Eigen-PC	6.0×10^{-5}
			CADD score v1.0	3.5×10^{-5}
			CADD score v1.1	3.3×10^{-2}
			MA	1.0×10^{-4}
SCZ	636	Missense and nonsense	Eigen	9.9×10^{-1}
			Eigen-PC	9.8×10^{-1}
			CADD score v1.0	1.5×10^{-1}
			CADD score v1.1	1.8×10^{-1}
	573	Missense only	Eigen	6.3×10^{-1}
			Eigen-PC	5.8×10^{-1}
			CADD score v1.0	9.8×10^{-1}
			CADD score v1.1	2.8×10^{-2}
			PhastPri	9.5×10^{-2}

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

from published studies, along with *de novo* mutations identified in controls in those studies. We selected only mutations with entries in the dbNSFP database. In total, for ASD, we had $n = 2{,}027$ such mutations, of which 1,753 were missense[22–26]. For schizophrenia, we had $n = 636$ mutations, of which 571 were missense[27–31]. For epileptic encephalopathies, we identified $n = 210$ mutations, with 184 missense[32], and for intellectual disability we had $n = 114$ mutations, with 99 missense[33,34]. For controls, we had $n = 1{,}310$ mutations, of which 1,157 were missense[23,25,26,28,31,34]. For ASD, we also performed an analysis based only on the *de novo* mutations that fell into genes encoding FMRP targets, as it has been shown that *de novo* ASD mutations are enriched among genes encoding fragile-X mental retardation protein (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 intellectual disability and ASD (FMRP targets) had the highest Eigen scores, followed by epileptic encephalopathy, ASD, schizophrenia and control mutations. *P* values from the Wilcoxon rank-sum tests comparing scores for *de novo* mutations in cases and controls are reported in **Table 2**. The Eigen-PC score performed similarly to the proposed Eigen score and much better than the CADD score, especially for epileptic encephalopathies and ASD, with the *P* values being orders of magnitude smaller for the Eigen and Eigen-PC scores. Notably, when we considered the small subset of *de novo* variants in ASD that fell into genes encoding FMRP targets, the results became much more significant (even though the number of variants was reduced by 15-fold): in particular, for missense variants, the *P* value was 3.2×10^{-4} for Eigen and 9.4×10^{-5} for Eigen-PC as compared to 4.2×10^{-2} for CADD score v1.0 and 1.7×10^{-2} for CADD score v1.1. We also report the best performing individual annotation for each data set; as before, no single annotation was best in all cases, although the best ones were again protein function scores. Results for each individual functional score are reported in **Supplementary Table 5**.

Noncoding and synonymous coding variants

For noncoding and synonymous coding variants, we used a suite of evolutionary conservation annotations and many regulatory annotations from the ENCODE Project[6]. ENCODE histone modification, transcription factor binding and open chromatin data were downloaded from the UCSC Genome Browser (January 2015). A full list of functional genomic scores is given in **Supplementary Table 1**. For the training data set, all variants in the 1000 Genomes Project data set without a match in dbNSFP and within 500 bp upstream of the gene start site were included, for a total of 418,997 variants. We report the estimated weights for individual annotations in **Supplementary Table 6**. Evolutionary conservation scores tended to have the highest weights for the Eigen score, whereas regulatory annotations had the highest weights for the Eigen-PC score. Note that the regulatory block is large (**Fig. 1**), containing over half the annotations used for calculating weights. Therefore, the regulatory block dominated the first principal component of the covariance matrix, increasing the weights in this block.

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

ClinVar noncoding and synonymous coding variants

We selected noncoding and synonymous coding variants from the ClinVar database. The selected variants included 3′ UTR, upstream, downstream, intergenic, noncoding change, intronic and synonymous coding variants. We identified 111 such pathogenic mutations. For controls, we selected a set of 111 benign variants from ClinVar matched to the pathogenic variants on the basis of functional class (see the **Supplementary Note** for more details). The AUCs for several aggregate scores and individual functional scores are given

Table 3 GWAS and eQTL SNPs

Data set	*n*	Comparison	Score	*P* value
GWAS	2,115	Regulatory GWAS vs. tag SNPs	Eigen	1.2×10^{-5}
			Eigen-PC	4.0×10^{-6}
			CADD score v1.0	5.9×10^{-4}
			CADD score v1.1	2.0×10^{-4}
			GWAVA (TSS)	4.1×10^{-6}
			TFBS num	4.9×10^{-5}
GWAS	2,115	Regulatory GWAS vs. other SNPs	Eigen	1.6×10^{-9}
			Eigen-PC	2.0×10^{-13}
			CADD score v1.0	2.0×10^{-6}
			CADD score v1.1	8.6×10^{-7}
			GWAVA (TSS)	7.4×10^{-13}
			TFBS sum	5.6×10^{-9}
GWAS	10,718	GWAS vs. matched controls	Eigen	6.9×10^{-8}
			Eigen-PC	3.5×10^{-13}
			CADD score v1.0	1.0×10^{-4}
			CADD score v1.1	5.2×10^{-7}
			GWAVA (TSS)	2.5×10^{-9}
			H3K4me1	4.0×10^{-11}
eQTLs	676	Regulatory eQTLs vs. tag SNPs	Eigen	1.8×10^{-10}
			Eigen-PC	7.0×10^{-23}
			CADD score v1.0	3.1×10^{-4}
			CADD score v1.1	4.3×10^{-5}
			GWAVA (TSS)	1.3×10^{-3}
			H3K4me3	2.2×10^{-24}
eQTLs	676	Regulatory eQTLs vs. other SNPs	Eigen	5.9×10^{-13}
			Eigen-PC	2.6×10^{-27}
			CADD score v1.0	2.8×10^{-4}
			CADD score v1.1	2.1×10^{-5}
			GWAVA (TSS)	7.3×10^{-8}
			H3K4me3	3.8×10^{-25}

P values (Wilcoxon rank-sum test) for GWAS SNPs and eQTLs. Comparisons are shown between GWAS index SNPs and tag SNPs mapping to 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 mapping to regulatory elements are shown. The best performing individual annotation is reported.

in **Supplementary Table 3**. Several conservation scores (GERP_RS, PhyloPla and PhyloVer) performed best, followed closely by the Eigen score. Eigen-PC and GWAVA performed rather poorly for this data set, similarly to the regulatory annotations.

Genome-wide significant SNPs

We computed scores for 14,915 genome-wide association study (GWAS) index SNPs that have been found to be genome-wide significant and are reported in the National Human Genome Research Institute (NHGRI) GWAS catalog. We note here that only a small proportion of GWAS index SNPs are expected to be causal (estimated at 5% in ref. 37), with most being in linkage disequilibrium (LD) with the true causal SNPs.

The Eigen score distributions for variants in different functional classes (for example, regulatory, upstream, downstream, intergenic and intronic) are shown in **Supplementary Figure 2a**. GWAS variants mapping to a known regulatory element (2,115 variants) had the highest Eigen scores, as expected. We used the Genome Variation Server (GVS) to extract tag SNPs that had an r^2 (LD) value of at least 0.8 with each GWAS index SNP. GVS divides the SNPs in an LD bin into 'tag SNPs' and 'other SNPs'. The latter group consists of all SNPs for which the r^2 value with any other SNP in the bin is below the threshold of 0.8. We constructed two types of control sets, one consisting of tag SNPs and another consisting of other SNPs, all mapping to a known regulatory element. We compared the various scores (Eigen, Eigen-PC, the two versions of CADD score and GWAVA) for GWAS index SNPs and these control variants. We generated 20 such matched control sets and report the median *P* values from the Wilcoxon rank-sum tests across these 20 comparisons in **Table 3**. Both the Eigen and Eigen-PC scores performed substantially better than the CADD score. Furthermore, Eigen-PC tended to perform best, outperforming all the other meta-scores and the best performing individual functional annotations.

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

Expression quantitative trait loci

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 mapping to a known regulatory element (676 eQTLs) had the highest Eigen scores (**Supplementary Fig. 2b**). We constructed similar control sets to those for GWAS, based on tag SNPs and other SNPs. The *P* values from the Wilcoxon rank-sum tests are reported in **Table 3**. The Eigen and Eigen-PC scores led to more significant results than both the CADD score and the GWAVA score. We report the results for all the individual functional scores in **Supplementary Table 7**.

Noncoding cancer mutations from the COSMIC database

We also compared scores for recurrent and non-recurrent somatic noncoding mutations in the Catalogue of Somatic Mutations in Cancer (COSMIC) database[39] (GWAVA scores were only available for a small number of the COSMIC variants, namely those that have been reported in dbSNP; therefore, we omitted the comparison with GWAVA for this data set). The *P* values from the Wilcoxon rank-sum tests for variants in different functional classes are reported in **Table 4** (**Supplementary Table 8** contains results for individual annotations). The *P* values for the Eigen and Eigen-PC scores were orders of magnitude smaller than those for the CADD score, across different groups of variants. We show the Eigen score distribution for variants in different functional classes in **Figure 3**. Regulatory, 5′ UTR and 3′ UTR variants had the highest scores, whereas intergenic variants had the lowest scores, as expected.

DISCUSSION

The Eigen score proposed here represents both a quantitative improvement in predictive power as compared to existing methods and a qualitative difference in the predictive model. The shift from supervised (CADD and GWAVA) to unsupervised algorithms as discussed

Table 4 Noncoding cancer mutations from the COSMIC database

Variant class	*n* rec	*n* non-rec	Eigen	Eigen-PC	CADD score v1.0	CADD score v1.1	Best individual annotation
Regulatory	21,279	428,398	2.02×10^{-165}	5.13×10^{-264}	1.05×10^{-71}	2.70×10^{-50}	$\leq 2.22 \times 10^{-308}$ (PolIIpval)
Intronic	85,502	2,093,158	2.40×10^{-155}	2.13×10^{-112}	2.89×10^{-61}	1.09×10^{-10}	$\leq 2.22 \times 10^{-308}$ (GERP_NR)
Downstream	15,956	318,967	2.73×10^{-92}	3.04×10^{-128}	4.31×10^{-36}	1.83×10^{-28}	1.01×10^{-155} (GERP_NR)
Upstream	14,636	309,615	1.28×10^{-52}	2.01×10^{-84}	7.90×10^{-24}	3.21×10^{-17}	9.68×10^{-86} (PolIIPval)
Noncoding change	4,903	66,717	2.51×10^{-7}	2.49×10^{-21}	1.51×10^{-1}	4.84×10^{-5}	8.13×10^{-35} (PolIIPval)
3′ UTR	2,236	28,261	6.94×10^{-3}	4.22×10^{-4}	1.06×10^{-5}	3.37×10^{-1}	5.67×10^{-5} (GERP_NR)
5′ UTR	417	3,908	1.14×10^{-2}	2.32×10^{-1}	6.43×10^{-2}	1.15×10^{-1}	2.79×10^{-7} (GERP_NR)
Intergenic	75,327	2,182,466	1.49×10^{-2}	3.97×10^{-6}	1.08×10^{-6}	6.30×10^{-16}	1.19×10^{-18} (H3K4me1)
Synonymous	434	2,388	1.09×10^{-1}	9.69×10^{-1}	8.25×10^{-1}	2.88×10^{-1}	2.16×10^{-3} (PhyloPri)

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

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

In addition, we have studied the performance of a related score, Eigen-PC, based on the eigendecomposition of the annotation covariance matrix and using the lead eigenvector to weight the individual annotations. In our experiments, Eigen-PC performs well across many scenarios, although, as we discuss in the **Supplementary Note**, it is more sensitive than Eigen to component annotations and possible confounding factors; therefore, Eigen is a more robust approach at this point. We have only experimented with the first principal component, but it is possible that the second principal component is also informative (**Supplementary Tables 11**–**13**). Further work is needed to investigate how to incorporate the information in additional principal components.

The 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 noncoding variants. The regulatory block is more than twice the size of the evolutionary conservation block. This causes the regulatory block to dominate the first principal component of the covariance matrix, increasing the weights in this block. With the current set of annotations, the strong weights placed on the regulatory annotations improve the ability of Eigen-PC to discriminate between the different data sets for noncoding variants used here. Changing the set of annotations could disrupt this behavior.

Eigen can incorporate a large number of the correlated functional annotations that are being generated by high-throughput projects such as ENCODE and Roadmap Epigenomics, if these fit the assumed block-structured correlation. We note that the set of annotations used by Eigen is a proper subset of the set used by CADD. In particular, we have excluded several non-numerical annotations. To verify that Eigen's improvement over CADD is not due to this difference in annotation sets, we have retrained CADD on the same set of annotations used by Eigen and have shown that this new version of CADD performs similarly to the full version of CADD (**Supplementary Tables 14**–**16** and **Supplementary Note**). As an additional experiment, we have also considered including CADD as one of the component annotations. The resulting score tends to perform worse than the original Eigen score, largely because of the fact that including CADD violates our assumption of conditional independence for annotations in different blocks (**Supplementary Tables 17**–**21** and **Supplementary Note**).

Although these aggregate scores can be very sensitive for mutations in Mendelian diseases, for the majority of variants associated with disease risk in complex diseases, these scores are expected to be mostly useful when combined with additional population-level genetic data (**Supplementary Fig. 3**).

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

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

Precomputed Eigen and Eigen-PC scores for every possible variant in the human genome (hg19) are available for download at our website (see URLs).

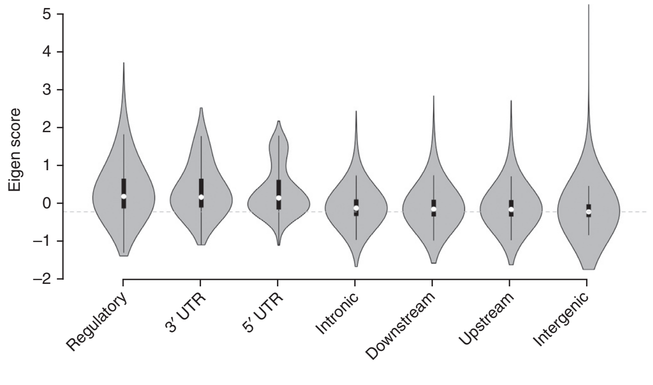

Figure 3 Violin plots showing the distribution of Eigen scores for noncoding variants in the COSMIC database that reside in different functional categories. The horizontal line corresponds to the median Eigen score for intergenic variants (the lowest scoring class).

URLs. Combined Annotation-Dependent Depletion (CADD), http://cadd.gs.washington.edu/; ClinVar, http://www.ncbi.nlm.nih.gov/clinvar/; Catalogue of Somatic Mutations in Cancer (COSMIC) database, http://cancer.sanger.ac.uk/cosmic/; dbNSFP, https://sites.google.com/site/jpopgen/dbNSFP; Eigen score software and score download, http://www.columbia.edu/~ii2135/eigen.html; ENCODE, https://www.encodeproject.org/; Ensembl, http://www.ensembl.org/index.html; GTEx, http://www.gtexportal.org/home/; Genome Variation Server (GVS), http://gvs.gs.washington.edu/GVS141/; GWAS genes, http://www.genome.gov/Pages/About/OD/OPG/GWAS%20Catalog/GWASCatalog112608.xls; National Human Genome Research Institute (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 Project, http://www.1000genomes.org/; UCSC Genome Browser, https://genome.ucsc.edu/; Variant Effect Predictor (VEP), http://www.ensembl.org/info/genome/variation/predicted_data.html#con.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

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

AUTHOR CONTRIBUTIONS

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

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Mardis, E.R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24**, 133–141 (2008).
2. Metzker, M.L. Sequencing technologies—the next generation. *Nat. Rev. Genet.* **11**, 31–46 (2010).
3. Zhang, J., Chiodini, R., Badr, A. & Zhang, G. The impact of next-generation sequencing on genomics. *J. Genet. Genomics* **38**, 95–109 (2011).
4. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
5. Davydov, E.V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6**, e1001025 (2010).
6. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
7. Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
8. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
9. Capanu, M. *et al.* The use of hierarchical models for estimating relative risks of individual genetic variants: an application to a study of melanoma. *Stat. Med.* **27**, 1973–1992 (2008).
10. Capanu, M. & Begg, C.B. Hierarchical modeling for estimating relative risks of rare genetic variants: properties of the pseudo-likelihood method. *Biometrics* **67**, 371–380 (2011).
11. Kichaev, G. *et al.* Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.* **10**, e1004722 (2014).
12. Ionita-Laza, I., Capanu, M., De Rubeis, S., McCallum, K. & Buxbaum, J.D. 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 (2014).
13. Ng, S.B. *et al.* Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome. *Nat. Genet.* **42**, 790–793 (2010).
14. Bamshad, M.J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12**, 745–755 (2011).
15. Meyer, K.B. *et al.* 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 (2013).
16. Pickrell, J.K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).
17. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).
18. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
19. Ritchie, G.R.S., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat. Methods* **11**, 294–296 (2014).
20. Gulko, B., Hubisz, M.J., Gronau, I. & Siepel, A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.* **47**, 276–283 (2015).
21. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum. Mutat.* **34**, E2393–E2402 (2013).
22. Iossifov, I. *et al.* The contribution of *de novo* coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
23. Iossifov, I. *et al.* *De novo* gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285–299 (2012).
24. Neale, B.M. *et al.* Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. *Nature* **485**, 242–245 (2012).
25. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* **485**, 246–250 (2012).
26. Sanders, S.J. *et al.* *De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–241 (2012).
27. Fromer, M. *et al.* *De novo* mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179–184 (2014).
28. Gulsuner, S. *et al.* Spatial and temporal mapping of *de novo* mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* **154**, 518–529 (2013).
29. Girard, S.L. *et al.* Increased exonic *de novo* mutation rate in individuals with schizophrenia. *Nat. Genet.* **43**, 860–863 (2011).
30. McCarthy, S.E. *et al.* *De novo* mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. *Mol. Psychiatry* **19**, 652–658 (2014).
31. Xu, B. *et al.* *De novo* gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat. Genet.* **44**, 1365–1369 (2012).
32. Epi4K Consortium & Epilepsy Phenome/Genome Project. *De novo* mutations in epileptic encephalopathies. *Nature* **501**, 217–221 (2013).
33. de Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* **367**, 1921–1929 (2012).
34. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380**, 1674–1682 (2012).
35. Darnell, J.C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247–261 (2011).
36. Dong, S. *et al.* *De novo* insertions and deletions of predominantly paternal origin are associated with autism spectrum disorder. *Cell Rep.* **9**, 16–23 (2014).
37. Farh, K.K. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
38. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
39. Forbes, S.A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43**, D805–D811 (2015).
40. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45**, 124–130 (2013).
41. Ye, C.J. *et al.* Intersection of population variation and autoimmunity genetics in human T cell activation. *Science* **345**, 1254665 (2014).
42. Ko, A. *et al.* Amerindian-specific regions under positive selection harbour new lipid variants in Latinos. *Nat. Commun.* **5**, 3983 (2014).
43. Gudbjartsson, D.F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015).
44. O'Rawe, J. *et al.* Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med.* **5**, 28 (2013).
45. Lam, H.Y. *et al.* Performance comparison of whole-genome sequencing platforms. *Nat. Biotechnol.* **30**, 78–82 (2012).
46. Fang, H. *et al.* Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome Med.* **6**, 89 (2014).

ONLINE 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.

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 data set consisting of a large number of variants and their functional annotations. For simplicity, we first assume conditional independence among the individual functional scores. A description of the main variables used in this section for ease of reference is provided in **Supplementary Table 22**. Let m be the number of variants and k be the number of functional predictors (for example, PolyPhen, GERP, etc.). Let $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ik})$ be i.i.d. vectors of k functional impact scores for variants $i = 1, \ldots, m$. It is assumed that the scores have been standardized so that for every score j we have $\mu_j = \mathrm{E}[Z_{ij}] = 0$ and $\sigma_j^2 = \mathrm{Var}(Z_{ij}) = 1$. 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} and F_{j0} are the conditional distributions of Z_{ij} given $C_i = 1$ and $C_i = 0$, respectively. Define $\mu_{jl} = \mathrm{E}[Z_{ij} | C_i = l]$ for score j and $l = 0,1$. Note that

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

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

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

This can be expressed in matrix form as

$$\mathbf{Q} = \pi\Sigma_1 + (1-\pi)\Sigma_0 + \mathbf{R} \qquad (3)$$

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

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

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

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

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

Therefore, under the assumption of conditional independence, the off-diagonal entries in the covariance matrix are equal to those of the rank-one matrix **R**. We are interested in μ_0 as the entries in μ_0 can be used to rank the scores because the accuracy of the score depends in part on how far apart the means of the conditional distributions are (that is, $\mu_{j1} - \mu_{j0} = -\frac{1}{\pi}\mu_{j0}$). Normally, we do not know μ_0, but the values of μ_0 can be estimated by first estimating the diagonal entries of **R** (see below) and then computing the leading eigenvector.

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

We note that, if the within-component variances are small as compared to the means, it follows from equation (3) that $\mathbf{Q} \approx \mathbf{R}$. A simple approach then is to take the first principal component of matrix **Q** as an approximation of μ_0, without the need to estimate the rank-one matrix **R**. However, this approach may fail if the within-component variances are not all small. We refer to this approach as Eigen-PC, while the main approach that assumes (blockwise) conditional independence is referred to as Eigen.

Blockwise conditional independence among individual functional scores. The assumption of conditional independence may not be appropriate in the case of functional genomics annotations. For instance, protein functional predictors that use similar information for prediction (for example, multiple-sequence alignments and protein three-dimensional structures) are likely to be correlated even given the true functional status for a variant. In contrast, 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 data set (similarly, **Supplementary Fig. 1** shows the corresponding correlation structure based on the coding variants on chromosome 1 from the training set). A clear block structure can be seen, with different types of annotations forming distinct blocks, with stronger correlations within blocks than between them. The three distinct blocks are an evolutionary conservation block (including several conservation scores such as GERP and PhyloP), a regulatory information block (including open chromatin measures, transcription factor binding and histone modifications) and an allele frequency block.

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

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

Proof. See the **Supplementary Note**.

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 **R**. As discussed previously, the entries in the eigenvector for the rank-one matrix **R** are proportional to the accuracies of the individual predictors and can be used to rank the various predictors. Next, we discuss how we may use these estimates of accuracies to combine the different predictors into one meta-score.

Meta-predictors. Once the blockwise division is chosen, the rank-one matrix **R** can be estimated and the leading eigenvector determined. As discussed above, the entries in the eigenvector can be used to rank and combine annotations. Larger values for the components of the eigenvector indicate greater accuracy for the corresponding annotations, and the component values can be used as weights for combining annotations in a linear combination. This way, we give more weight to the more accurate annotations. If $(e_1, \ldots, e_k)$ is the eigenvector for the matrix **R** and $(Z_{i1}, \ldots, Z_{ik})$ are the functional scores for variant i, then the meta-score for variant i is given by

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

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

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

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

Note that if the Eigen-PC method is used the outline is similar. Steps 3 and 4 will be omitted, as the covariance matrix **Q** is used directly. In step 5, the eigendecomposition is applied to **Q** and in step 6 the lead eigenvector, the one with the greatest eigenvalue, is used (it was not necessary to specify this previously because **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 and for PolyPhen a value of 1 is imputed, whereas 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 pairwise correlations, which allows variants with missing values for some annotations to be used.

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