# Functional annotation of noncoding sequence variants

Graham R S Ritchie[1,2], Ian Dunham[1], Eleftheria Zeggini[2] & Paul Flicek[1,2]

**Identifying functionally relevant variants against the background of ubiquitous genetic variation is a major challenge in human genetics. For variants in protein-coding regions, our understanding of the genetic code and splicing allows us to identify likely candidates, but interpreting variants outside genic regions is more difficult. Here we present genome-wide annotation of variants (GWAVA), a tool that supports prioritization of noncoding variants by integrating various genomic and epigenomic annotations.**



The majority of genetic variants associated with complex traits lie in noncoding regions of the genome, and many of these lie some distance away from the nearest protein-coding locus[1]. This observation implies that many variants that affect the risk of common, complex diseases are likely to exert their effect by altering the regulation of genes rather than by directly affecting gene and protein function. However, the majority of efforts to annotate functional variants to date have focused on variants that directly affect coding sequence, such as missense and nonsense mutations, or those that affect transcript splicing signals[2]. Recently, large-scale efforts such as the Encyclopedia of DNA elements (ENCODE) consortium[3] and the US National Institutes of Health Roadmap Epigenomics project[4] have made available data from a wide range of assays across the genome aimed at identifying functional noncoding elements. These sources of data offer an opportunity to interpret noncoding variants, but it is not yet clear which annotations, or combinations of annotations, will help us discriminate variants likely to be functionally involved in medically relevant phenotypes from the large number of apparently benign variants that occur across the genome.

Existing computational approaches to predict the effect of a coding variant on protein function such as the 'sorting tolerant from intolerant' (SIFT) algorithm[5] and 'polymorphism phenotyping' (PolyPhen) tool[6] are largely based on quantifying constraint on the affected residue from a multiple-sequence alignment. This approach is possible because protein sequences have been highly conserved throughout evolution. Regulatory elements are known to have much higher evolutionary turnover[7], which implies that conservation is a less important signal when interpreting variants in regulatory regions. Effects of regulatory variants are also harder to interpret because they are likely to have quantitative rather than qualitative effects on gene expression, and the same variant may have a larger or smaller effect in different tissues, at different developmental stages and even in different individuals.

In this work, we used a wide range of variant-specific annotations of different classes and at a range of genomic scales to investigate whether a combination of regulatory features, genic context and genome-wide properties can be used to identify variants likely to be functional. We found marked differences in the distribution of several of these annotations for functional regulatory variants compared to controls (**Supplementary Figs. 1** and **2**, and **Supplementary Results**), but on their own these differences were insufficient to allow us to discriminate functional variants from controls with reasonable precision. We built a classifier that integrates the range of annotations and could discriminate functional variants from background. Using several case studies, we demonstrate how this classifier can be used in future association studies.

To identify annotations that can be used to discriminate noncoding variants likely to be involved in disease from benign variants (i.e., those without known pathogenic effect), we compared variants implicated in disease with common control variants. For the disease-implicated set, we used all variations annotated as 'regulatory mutations' from the public release of the Human Gene Mutation database (HGMD)[8]. For all control sets, we used common (minor allele frequency ≥1%) single-nucleotide variants (SNVs) from the 1000 Genomes Project (1KG)[9]. The first control set we constructed was a random selection of SNVs from across the genome in order to sample overall background. The HGMD variants are not distributed randomly across the genome; 75% lie within a 2 kilobase (kb) window around an annotated transcription start site (TSS). To control for this distribution, a second control set was matched for distance to the nearest TSS genome-wide. The third and most stringent control set accounted for the fact that genes near the HGMD variants are unlikely to have been selected in an unbiased way. This final control set was composed of all 1KG variants in the 1 kb surrounding each of the HGMD variants.

We used a modified version of the random forest algorithm[10] to build three classifiers using all available annotations to discriminate between the disease variants and variants from each of the three control sets (Online Methods). We show the average receiver operating characteristic (ROC) curves for the classifiers trained on each of the three training sets, computed using ten-fold cross-validation (**Fig. 1**). The areas under the ROC curves (AUC) demonstrate that the classifier for each training set can usefully discriminate between disease and control variants.

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, UK. [2]Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK. Correspondence should be addressed to E.Z. (eleftheria@sanger.ac.uk) or P.F. (flicek@ebi.ac.uk).

RECEIVED 16 JULY 2013; ACCEPTED 2 JANUARY 2014; PUBLISHED ONLINE 2 FEBRUARY 2014; DOI:10.1038/NMETH.2832

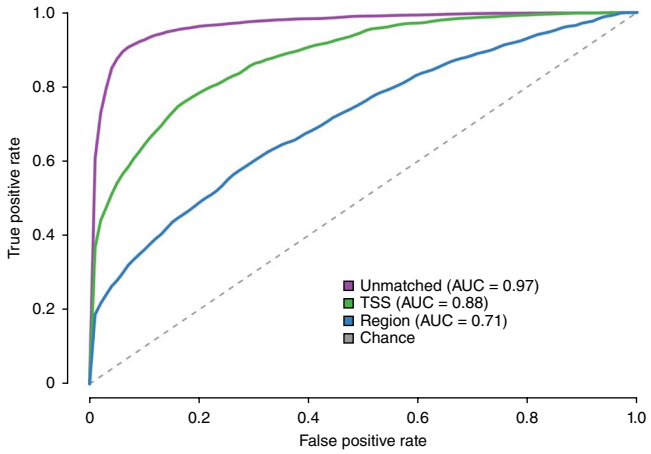


**Figure 1** | Mean ROC curves for tenfold cross-validation experiments on each of the three training sets. Dashed line indicates random chance.

As we expected, relative performance improved as variant sets became less stringently matched.

We also analyzed which of the annotations contribute the most to the discriminative power of each classifier (Online Methods) and found considerable differences according to the training set used (**Supplementary Results** and **Supplementary Fig. 3**). Generally, we observed that annotations available across more of the genome, such as G+C content, evolutionary conservation, DNase I hypersensitivity, some histone modifications and distance to the nearest TSS were among the most informative for all three classifiers. More specific annotations, such as DNase I footprints, were informative for the classifier trained on variants matched by genomic region.

As an independent validation, we annotated a set of 194 noncoding variants classified as pathogenic in the US National Center for Biotechnology Information ClinVar database and not found in HGMD. We compared classifier scores for these variants against the 150 noncoding variants classified in ClinVar as nonpathogenic and also a set of 19,400 1KG variants matched for distance to the nearest TSS. The AUC values for discriminating pathogenic variants in these two sets were 0.75 and 0.84, respectively (**Supplementary Fig. 4**).

To establish whether prediction scores are likely to be generalizable to other data sets, we conducted experiments that demonstrate how classifier scores could be applied to prioritize candidate functional variants.

For the first experiment, we annotated noncoding variants associated with complex disease from genome-wide association studies (GWAS)[1]. Many of the top-ranking variants from these studies are unlikely to be causal and instead are likely to be in linkage disequilibrium with the functional variant(s). Nonetheless, we found that noncoding GWAS SNVs had a slightly but significantly higher average GWAVA score than control variants selected from the same genotyping chips used in GWAS and matched for distance to the nearest TSS (mean GWAVA score 0.268 versus 0.248, $P = 3.6 \times 10^{-29}$, two-sided Mann-Whitney *U* test; **Supplementary Fig. 5**). When we used the strategy from ref. 11 to stratify GWAS signals into those that are unreplicated, replicated in the same study and replicated in an independent study, we found that variants that replicate more robustly had higher average GWAVA scores (not replicated versus independently replicated $P = 3.65 \times 10^{-7}$; **Supplementary Fig. 6**). We also applied GWAVA predictions to three fine-mapping studies that follow up on GWAS signals (**Supplementary Tables 1–3** and **Supplementary Results**) and found that GWAVA consistently ranked the candidate functional variant highly.

To establish whether GWAVA scores might be useful in a personal genomics context, we identified all SNVs called in a single (arbitrarily chosen) individual from the 1000 Genomes Project (NA06984) and limited our analysis to variants on chromosome 22. To simulate putatively functional variants, we then 'spiked in' the 33 HGMD regulatory variants from chromosome 22 to this set and built a version of the classifier trained on variants matched by distance to the nearest TSS, excluding all chromosome 22 data from the training set. We could discriminate the spike-in variants from the background variants with good accuracy (AUC = 0.85, **Supplementary Fig. 7**), though at reasonable score thresholds we would still expect a substantial number of false positives in a whole genome. In the context of personal genomes, we would therefore recommend combining GWAVA scores with other sources of evidence of variant candidacy, such as segregation with disease in a family study, or in combination with prior biological or clinical evidence for specific genes or regions.

Next we established whether the scores might help to identify the functional variant when the relevant gene is suspected (for example, from other evidence such as known disease-implicated coding variants from the same locus). For each of the 24 unique genes annotated as being affected by the HGMD variants, we identified all noncoding variants from NA06984 in the region around each gene (5 kb upstream and downstream of it) and observed where the spike-in variant was ranked according to the GWAVA score (**Supplementary Table 4**). GWAVA ranked the spike-in variant first for five genes and in the top three for ten genes, significantly more often than expected by chance ($P = 0.003$

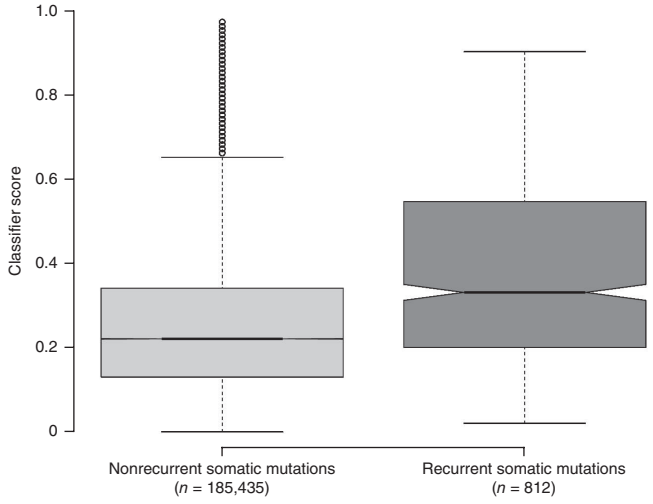


**Figure 2** | Classifier scores for recurrent versus nonrecurrent noncoding somatic mutations from the COSMIC database. The AUC for discriminating between these two classes of mutation is 0.67. $P = 3.1 \times 10^{-61}$, two-sided Mann-Whitney *U* test. Boxplots show the median (center lines), the upper and lower quartiles, also known as the first and third quartiles (boxes), scores within 1.5× the interquartile range of the upper and lower quartiles (whiskers), and the rest of the data (circles).

and $P = 0.0002$, respectively, by simulation, details of which are described in Online Methods).

As an application to cancer studies, we annotated noncoding somatic mutations discovered in whole-genome sequencing studies from the 'catalog of somatic mutations in cancer' (COSMIC) database[12]. We identified all mutations that had been identified in multiple studies ($n = 812$ mutations) and found that these recurrent mutations were assigned a significantly higher average GWAVA score than nonrecurrent mutations ($P = 3.1 \times 10^{-61}$, two-sided Mann-Whitney $U$ test AUC = 0.67, **Fig. 2**). Recurrence of somatic mutations is a widely used proxy of likely function, so this result represents a validation of the classifier from an entirely different domain and suggests that this approach might be useful in prioritizing mutations in the search for cancer driver mutations.

We sought to compare GWAVA scores with those of other tools that can classify noncoding variants. The only such tool we were aware of is MutationTaster[13], which can provide predictions for noncoding variants that can be mapped to a transcript model, such as those in untranslated regions and introns. For comparison, we used noncoding somatic mutations from the COSMIC database in order to avoid the issue that HGMD variants are used to train both tools and that known disease-implicated variants (such as those in ClinVar) are automatically classified as disease-causing by the MutationTaster webserver. We obtained predictions from the MutationTaster webserver for 92,352 noncoding mutations that could be mapped to a transcript model. MutationTaster does not supply prediction scores but rather a qualitative prediction of "disease_causing" or "polymorphic." To compare results, we therefore used a GWAVA score threshold of 0.5, above which mutations were classified as "functional" and below or equal to which mutations were considered "nonfunctional." We then computed contingency tables to test whether mutations identified as functional were also recurrent. Although we found a significant enrichment for recurrent mutations among those called as functional by either tool, the odds ratio for GWAVA results was 5.4 (Fisher's exact $P = 1.3 \times 10^{-56}$) compared to only 2.0 for MutationTaster results (Fisher's exact $P = 6.5 \times 10^{-8}$).

The computational approach we presented can combine information from a wide range of available annotations to predict the functional impact of noncoding variants, and address issues of context dependency and inconsistent signal from evolutionary conservation in regulatory elements. The classifier software and annotation data are freely available for download (**Supplementary Software** and ftp.sanger.ac.uk/pub4/resources/software/gwava/). We also precomputed classifier scores for all known variants from the Ensembl variation database[14] (release 70) and these scores, along with the underlying annotations are available from a web server (http://www.sanger.ac.uk/resources/software/gwava/).

We hope that by incorporating GWAVA predictions for noncoding variants into disease-association studies we will substantially improve the chances of finding variants relevant to disease and other phenotypes.

## 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

G.R.S.R. is supported by European Molecular Biology Laboratory and the Sanger Institute via an EBI-Sanger Postdoctoral Fellowship. This work was funded by the Wellcome Trust (098051 and 095908) and by the European Molecular Biology Laboratory. The research leading to these results has received funding from the EU Seventh Framework Programme (FP7/2007-2013) under grant agreement (282510–BLUEPRINT).

### AUTHOR CONTRIBUTIONS

G.R.S.R. implemented the method, performed all analyses and drafted the manuscript. I.D. assisted with access to ENCODE data and suggested how to construct the control sets. E.Z. and P.F. contributed to the development of the method, manuscript writing and jointly directed the work.

### COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.



1. Hindorff, L.A. *et al. Proc. Natl. Acad. Sci. USA* **106**, 9362–9367 (2009).
2. Cooper, G.M. & Shendure, J. *Nat. Rev. Genet.* **12**, 628–640 (2011).
3. The ENCODE Project Consortium. *Nature* **489**, 57–74 (2012).
4. Bernstein, B.E. *et al. Nat. Biotechnol.* **28**, 1045–1048 (2010).
5. Kumar, P., Henikoff, S. & Ng, P.C. *Nat. Protoc.* **4**, 1073–1081 (2009).
6. Adzhubei, I.A. *et al. Nat. Methods* **7**, 248–249 (2010).
7. Schmidt, D. *et al. Science* **328**, 1036–1040 (2010).
8. Stenson, P.D. *et al. Genome Med.* **1**, 13 (2009).
9. The 1000 Genomes Project Consortium. *Nature* **491**, 56–65 (2012).
10. Breiman, L. *Mach. Learn.* **45**, 5–32 (2001).
11. Maurano, M.T. *et al. Science* **337**, 1190–1195 (2012).
12. Forbes, S.A. *et al. Nucleic Acids Res.* **39**, D945–D950 (2011).
13. Schwarz, J.M., Rödelsperger, C., Schuelke, M. & Seelow, D. *Nat. Methods* **7**, 575–576 (2010).
14. Flicek, P. *et al. Nucleic Acids Res.* **41**, D48–D55 (2013).

## ONLINE METHODS

**Annotation sources.** We acquired a range of annotations at different scales and in a variety of data formats. Here we group data types by class and name their sources.

*Open chromatin*. We used DNase I hypersensitivity assay followed by sequencing (DNase-seq) and formaldehyde-assisted isolation of regulatory elements followed by sequencing (FAIRE-seq) peak calls and DNase footprints from ENCODE.

*Transcription factor binding*. We used ChIP-seq peak calls for 124 transcription factors from ENCODE, JASPAR[15] motifs aligned under corresponding factor chromatin immunoprecipitation–sequencing (ChIP-seq) peaks from Ensembl and bound transcription factor binding motif data from ENCODE.

*Histone modifications*. We used ChIP-seq peak calls for 12 different modifications from ENCODE.

*RNA polymerase binding*. We used ChIP-seq peak calls from ENCODE.

*CpG islands*. We used predictions from Ensembl[14].

*Genome segmentation*. We used Ensembl integration[16] of the ENCODE SegWay[17] and ChromHMM[18] segmentation calls, which identified seven discrete states.

*Conservation*. We used genomic evolutionary rate profiling (GERP) scores from mammalian alignments from the Sidow laboratory at Stanford University, both at the specific variant nucleotide and averaging over 100 base pairs surrounding each variant[19].

*Human variation*. We took variants, allele frequencies and ancestral allele calls from 1000 Genomes Project phase 1 data. Mean heterozygosity and mean derived allele frequency of variants were both calculated in 1 kb windows from global population frequencies.

*Genic context*. We used distance to the nearest TSS from Gencode annotation provided by ENCODE[20], and distance to the nearest splice site and summary gene region annotations (any base annotated as exonic, intronic, coding sequence, 5′ or 3′ untranslated region, splice site, or start or stop codon in any transcript) from Gencode annotation provided by Ensembl.

*Sequence context*. We used sequence context information from the GRCh37 assembly of the human genome produced by the Genome Reference Consortium[21] including G+C content calculated over the 100 bp surrounding each variant, a Boolean variable indicating whether the variant is in a CpG context in the reference assembly, reference nucleotide at the variant position. We also used a Boolean variable indicating if the variant falls in repeat sequence from the University of California at Santa Cruz genome browser.

We developed a pipeline that can apply these annotations to a given set of variant loci. The result was a large matrix with a row for each variant locus and a column for each possible annotation. The column type, which depends on the annotation class, can be: (i) the number of cell lines in which the variant locus overlaps some annotation, such as DNase I hypersensitive sites and ChIP-seq peaks, (ii) a present-absent binary flag for the annotation at the variant locus (for example, whether this region is ever in an annotated intron) or (iii) a continuous value for genome-wide annotations, such as conservation and distance to the nearest TSS.

**Construction of disease and control variant sets.** The disease-implicated set of variants was composed of all variants annotated as 'regulatory mutations' from the April 2012 release of HGMD and downloaded from Ensembl release 70. After we removed variants that has the same position, a set of 1,614 disease-implicated SNVs remained. For all three control sets, we used variants identified in the low-depth whole-genome study in the 1000 Genomes Project (1KG) phase 1 release, downloaded from the project website in December 2012. We limited our analysis to variants with minor allele frequency ≥1% to reduce the chance of including rare functional variants in our control set. (We performed sensitivity analyses by either focusing exclusively on variants from European populations or on rare, singleton variants, and we found qualitatively similar results for cross-validation with the common variant controls; data not shown.)

As we only had SNVs in our 'disease' set, we also limited our analysis to SNVs in our control sets, for a total of 15,730,276 potential control SNVs. The first control set was a random selection of SNVs from across the genome, 100 times the size of the disease set, to get a reasonable sample of the background while making analyses computationally tractable. The second control set included 1KG SNVs matched for distance to the nearest TSS genome-wide, but not necessarily near the same genes as the HGMD variants. This set was 10 times the size of the disease set; for larger sets we could not ensure that the distributions of distances matched those of the HGMD variants. The final control set was composed of all 1KG variants in the 1 kb surrounding each of the HGMD variants ($n$ = 5,027 variants).

**Individual feature analysis.** We investigated whether any of the annotations showed a different distribution in the disease and control sets. Annotations can be grouped into two classes: a large class of regional data indicating whether or not each variant lies in an annotated element, possibly across multiple cell lines, and a smaller class consisting of several continuous variables.

For the regional data, we ignored the number of cell lines in which a variant was found (as these were not independent across cell lines) and just used a single binary variable per feature indicating whether each variant was found in this element in any cell line. Annotations not specific to a cell line are already binary. For each feature we then computed a contingency table identifying how these counts differed in our disease and control sets. We used Fisher's exact test to compute the significance of enrichment or depletion.

For continuous features, we used a two-sided Mann-Whitney $U$ test to establish whether there was a significant difference in the distribution of each feature between the two classes. We used this test (here and throughout this study), as it does not make any assumptions about the underlying distributions of our data. For the measures of the distance to the nearest TSS or splice site, we used absolute values, though we supplied the original signed value to the classifier as it may be informative to take into account whether the variant is upstream or downstream from the nearest TSS. All $P$ values were adjusted using the Bonferroni correction for multiple testing. Unadjusted $P$ values are also reported (**Supplementary Table 5**).

**Classifier algorithm.** Our classifier needs to simultaneously handle a large number of continuous and categorical features. Two of the control sets are also very unbalanced with respect to variant class, in that there are considerably fewer disease-implicated

variants than controls. To address these issues, we used a slightly modified version of the random forest algorithm[10]. Random forests are a robust and widely used approach to classification that can deal with the different feature types that we use. They are also robust to the presence of features that are not predictive (so we did not perform any feature selection). We modified the standard algorithm to address class imbalance by sampling equally from both classes when generating the training set for each component decision tree in the forest. The even class distribution means that each tree is trained on a smaller subset of control variants, but we used enough trees that most of the controls should be used at least once in the full model (subject to the normal random subsampling that is part of the algorithm). The random forest approach also has the advantage that it allowed us to compute the relative importance of each feature from the trained model.

We trained three forests, one for each of the three control variant sets and using the same disease variants as the positive set in each forest. We experimented with different numbers of decision trees in the forest and found that performance seemed to saturate around 100 trees, and this number should also ensure that we sample a good proportion of variants in each of the training sets. We used the mean AUC value across each of the test sets in each fold as our main measure of classifier performance.

One potentially confounding characteristic of the HGMD data is that some genes have multiple associated variants (mean = 2.03, median = 1), some of which are located physically close and may have annotations in common. When performing cross-validation, variants from the same gene that appear in both the training and test sets may inflate performance statistics. To control for this, we created a stringent set of disease variants in which a single variant is randomly selected for each gene, and we observed a similar performance pattern, with slightly reduced AUC values (0.95, 0.82 and 0.64, respectively).

All software was written in the Python language, using a random forest implementation from the Scikit-learn library[22]. The modified source code is available at http://www.sanger.ac.uk/resources/software/gwava/ and as **Supplementary Software**.

**Feature importance.** To identify which features contribute to the discriminative ability of each classifier, we computed the relative Gini importance of each feature across each component tree of the three forests (**Supplementary Fig. 3**). Gini importance measures the mean decrease in impurity at each node in the tree owing to the feature of interest, weighted by the proportion of samples reaching that node.

**Classifier score distribution.** We computed the distribution of scores across all variants from the 1000 Genomes Project on chromosome 16 (with variants included in any training set removed; **Supplementary Fig. 8**). Although the distributions were somewhat different for each classifier, as expected, few variants were assigned high scores by any version. These distributions allowed us to compare scores from any candidate variant with the background distribution to estimate how 'unexpected' any given score is.

**Validation experiments.** *Annotating pathogenic variants from ClinVar*. We downloaded the full ClinVar database in VCF format in early 2013 (file name: clinvar_20130118.vcf), identified all variants annotated as 'pathogenic' (those with US National Center for Biotechnology Information (NCBI) clinical significance code = 5 in the "INFO" field) and extracted them. We first removed all variants that overlapped any coding sequence or essential splice sites (as annotated in Ensembl release 70) and then any variants overlapping with an HGMD variant. The resulting set of 194 variants constitutes the set of pathogenic noncoding variants we used in this analysis. We performed a similar filtering to identify all likely nonpathogenic variants annotated (those with NCBI clinical significance code 2 or code 3) and derived a set of 150 nonpathogenic noncoding variants. We also constructed a control set matched for distance to the nearest TSS from the 1000 Genomes Project data as described above for the HGMD variants, and again we only included 1000 Genomes variants with mean allele frequency ≥1%, and we included 100 control variants for each ClinVar variant, which resulted in a set of 19,400 control variants. We annotated these three sets of variants with the classifier trained on variants matched by distance to the nearest TSS and compared the classification results with ROC curves (**Supplementary Fig. 4**).

*Annotating GWAS SNPs*. We downloaded the GWAS catalog from the US National Human Genome Research Institute website in December 2012 and identified all variants with a "Context" field implying the variant did not fall in coding sequence. For the matched control set, we used a list of SNVs from common GWAS genotyping arrays constructed using information from Ensembl release 70, and overlapping with variants from the 1000 genomes project. We selected ten matching SNVs for each GWAS signal. The genotyping platforms used were Affymetrix GeneChip 100K, GeneChip 500K and SNP6, and Illumina HumanCNV370 Quadv3, HumanHap300v2, HumanHap550v3.0, Cardio Metabo, Human1M-duoV3 and Human660W-quad.

We compared the score distributions of these two sets of variants with a two-sided Mann-Whitney *U* test (**Supplementary Fig. 5**).

We downloaded the replication status annotations available in the supplementary material of ref. 11. We used these annotations to stratify the classifier scores according to whether the annotated SNPs were not validated, were internally validated or were validated in an independent study (**Supplementary Fig. 6**). Comparison of score distributions was performed with a two-sided Mann-Whitney *U* test. The *P* values comparing all pairwise combinations of these three sets of variants are: not replicated versus internally replicated, $P = 2.56 \times 10^{-9}$; not replicated versus independently replicated: $P = 3.65 \times 10^{-7}$ and internally replicated versus independently replicated, $P = 0.024$.

*Application to personal genomics*. We downloaded variant calls for the individual NA06984 from the 1000 Genomes Project website, and identified all variants found on chromosome 22 in this individual. We created a training set for the classifier based on the control set matched for distance to the nearest TSS but with all variants on chromosome 22 removed. We then built a classifier using the same approach described earlier on this reduced training set. We used this classifier to annotate all variants from the NA06984 chromosome 22 and the 33 HGMD variants from the same chromosome, and used a ROC curve to demonstrate how well we can discriminate the HGMD variants from background (**Supplementary Fig. 7**).

For individual gene analysis, we used the 24 unique genes annotated in HGMD as being affected by this set of 33 variants. For genes associated with more than one variant, we randomly




doi:10.1038/nmeth.2832

selected a single variant and disregarded the rest. We downloaded the coordinates from each of these genes from Ensembl and identified all variants from NA06984 that overlapped the gene region ± 5 kb (the distance used by Ensembl to associate a variant with a gene). We removed any variant overlapping coding sequence or an essential splice site. For each gene, we then computed the GWAVA score using the classifier trained on the control set matched for distance to the nearest TSS and identified the rank of the HGMD variant at each locus (**Supplementary Table 4**). To test the significance of this result, we developed some simulation software (available at the FTP site above, along with all other software) to establish how often we would expect to find a result as extreme as or more extreme than that observed if we were ranking the variants around each gene at random. We used this software to derive empirical *P* values for our results based on 1,000,000 random samples.

*Application to somatic mutations.* We downloaded all annotated noncoding somatic mutations from the COSMIC database, release 64, in March 2013 and limited our analysis to those annotated as being discovered in a whole-genome study. We identified all mutation loci that are found in more than one study (according to the COSMIC study identifier) and annotated these as recurrent. Comparison of score distributions was performed with a two-sided Mann-Whitney *U* test (**Fig. 2**).

**Comparison with MutationTaster.** We uploaded all noncoding somatic mutations from whole-genome studies in COSMIC release 64 that did not overlap either coding sequence or essential splice sites to the MutationTaster webserver in October 2013, and we obtained predictions for 93,692 unique mutations that could be mapped to a transcript model. MutationTaster reports multiple predictions for mutations that overlap multiple transcripts, and we computed a unique prediction for each mutation by assigning the prediction "disease_causing" to any mutation with this prediction in any transcript and "polymorphism" otherwise. We discarded variants with a prediction of "polymorphism_automatic" as these are made by database lookup ($n$ = 1,340 variants). We used contingency tables to compare the number of variants predicted as "disease_causing" with whether or not the mutation was recurrent in different studies, and used Fisher's exact test to compute the significance of the enrichment. To compare this result with that of GWAVA, we assigned GWAVA scores to the same 92,352 mutations and threshold the GWAVA score with mutations scoring >0.5 identified as "functional" and all others "nonfunctional," and again used a contingency table to compute the enrichment of recurrent mutations among those called as functional.

**Classifier availability.** The GWAVA web server allows users to retrieve precomputed scores from each of the three classifiers for all known germ-line and somatic SNVs found in Ensembl release 70. All the underlying annotations used by the classifier are also available at http://www.sanger.ac.uk/resources/software/gwava/.

The source code, documentation, set of annotations used, all variant data sets described here and a plugin for the Ensembl Variant Effect Predictor[23] are available from the FTP server linked from the GWAVA webpage.


15. Mathelier, A. *et al. Nucleic Acids Res.* **42**, D142–D147 (2014).
16. Hoffman, M.M. *et al. Nucleic Acids Res.* **41**, 827–841 (2013).
17. Hoffman, M.M. *et al. Nat. Methods* **9**, 473–476 (2013).
18. Ernst, J. & Kellis, M. *Nat. Methods* **9**, 215–216 (2012).
19. Davydov, E.V. *et al. PLoS Comput. Biol.* **6**, e1001025 (2010).
20. Harrow, J. *et al. Genome Res.* **22**, 1760–1774 (2012).
21. Church, D. *et al. PLoS Biol.* **9**, e1001091 (2011).
22. Pedregosa, F. *et al. J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
23. McLaren, W.M. *et al. Bioinformatics* **26**, 2069–2070 (2010).