Published in final edited form as:
Nat Methods. 2014 March ; 11(3): 294–296. doi:10.1038/nmeth.2832.

Functional annotation of non-coding sequence variants

Graham R. S. Ritchie[1,2], **Ian Dunham**[1], **Eleftheria Zeggini**[2,*], and **Paul Flicek**[1,2,*]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, UK

[2]Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

Abstract

Identifying functionally relevant variants against the background of ubiquitous genetic variation is a major challenge in human genetics. For variants that fall in protein-coding regions our understanding of the genetic code and splicing allow us to identify likely candidates, but interpreting variants that fall outside of genic regions is more difficult. Here we present a new tool, GWAVA, which supports prioritisation of non-coding variants by integrating a range of annotations.

The majority of genetic variants associated with complex traits lie in non-coding regions of the genome, with many lying some distance away from the nearest protein-coding locus1. This observation implies that many variants affecting 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 on functional annotation of variants to date have focused on variants that directly affect coding sequence, such as missense and nonsense mutations, or those that affect transcript splicing signals2. Recently large-scale efforts such as the ENCODE consortium3 and the NIH’s Roadmap Epigenomics project4 have made available data from a wide range of assays across the genome aimed at identifying functional non-coding elements. These sources of data offer a new opportunity to interpret non-coding variants, but it is not yet clear which of these annotations, or combinations of annotations, will help us discriminate variants likely to be functionally involved in medically relevant phenotypes from the significant number of apparently benign variants that occur across the genome.

Existing computational approaches to predicting the effect of a coding variant on protein function, such as SIFT5 and PolyPhen6, are largely based on quantifying constraint on the affected residue from a multiple sequence alignment. This approach is possible because

[*]joint corresponding authors: eleftheria@sanger.ac.uk, flicek@ebi.ac.uk.

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. All authors read, commented on and approved the final manuscript.

Competing financial interests statement

None

protein sequences have been highly conserved throughout evolution. Regulatory elements are known to have much higher evolutionary turnover7 implying that conservation is a less important signal when interpreting variants in regulatory regions. The 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 more or less of an effect in different tissues, developmental stages and even individuals.

In this work we use a wide range of variant-specific annotations of different classes and at a range of genomic scales to investigate if a combination of regulatory annotations, genic context and genome-wide properties can be used to identify variants likely to be functional. We find that annotated functional regulatory variants show marked differences in their distribution with respect to several of these annotations when compared to controls (Online Methods), but that on their own these differences are insufficient to allow us to discriminate functional variants from controls with reasonable precision. We build a classifier that integrates the range of annotations and demonstrate that we can usefully discriminate functional variants from background. We present several case studies that demonstrate how this classifier can be used in next generation association studies.

In order to identify annotations that are helpful in discriminating non-coding variants likely to be functionally involved in disease from benign variants, we compare a set of variants implicated in disease with common control variants. For the disease-implicated set we use all variations annotated as ‘regulatory mutations’ from the public release of the Human Gene Mutation database (HGMD)8. For all 3 control sets we use common (minor allele frequency > 1%) single nucleotide variants (SNVs) from the 1000 Genomes Project (1KG)9. The first control set we construct is a random selection of SNVs from across the genome in order to get a sample of the overall background. The HGMD variants are not distributed randomly across the genome, 75% lie within 1 kilobase (kb) either side of an annotated transcription start site (TSS). To control for this we construct a second control set matched for distance to the nearest TSS genome-wide. The third and most stringent control set is constructed to account for the fact that the genes near the HGMD variants are unlikely to be have been selected in an unbiased way. This final control set is composed of all 1KG variants in the 1kb surrounding each of the HGMD variants.

We used a modified version of the Random Forest algorithm10 to build 3 classifiers using all available annotations to discriminate between the disease variants and variants from each of the 3 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 10-fold cross-validation (Fig. 1). The area under the ROC curves (AUC) demonstrates that for each of our training sets we build a classifier that can usefully discriminate between the disease and control variants. As expected, performance depends on how stringently matched the variant sets are. We also analysed which of the various annotations contribute most to the discriminative power of each classifier (Online Methods), and we found considerable differences according to the training set used (Supplementary Fig. 1).

As an independent validation, we annotated a set of 194 non-coding variants classified as pathogenic in the NCBI’s ClinVar database and not found in HGMD. We compared classifier scores for these variants against the 150 non-coding variants classified in ClinVar as non-pathogenic, and also a set of 19400 1KG variants matched for distance to the nearest TSS. We find that the AUC for discriminating pathogenic variants in these two sets are 0.75 and 0.84 respectively (Supplementary Fig. 2).

To further establish if the prediction scores from the classifier are likely to be generalizable to other data sets, we conducted validation experiments that demonstrate how scores from the classifier could be applied to prioritise candidate functional variants.

The first experiment we perform is to annotate non-coding variants associated with complex disease from genome-wide association studies1. These associations have typically been discovered using genotyping technologies and many of the lead variants are unlikely to be causal, but rather in linkage disequilibrium with the functional variant(s). Nonetheless, we find that non-coding GWAS SNPs are assigned a slightly but significantly higher GWAVA score than control variants selected from the same genotyping chips used in GWAS and matched for distance to the nearest TSS (mean score 0.268 vs. 0.248, $P = 3.6\times10^{-29}$) (Supplementary Fig. 3). If we stratify the GWAS signals using the strategy from Maurano et al.11 into those that have not been replicated, those that replicated in the same study and those that replicated in an independent study, we find that variants that replicate more robustly are assigned higher average GWAVA scores (not replicated vs. independently replicated $P = 3.65\times10^{-07}$) (Supplementary Fig. 4). We have also applied GWAVA predictions to 3 example fine-mapping studies following up on GWAS signals (Online Methods, Supplementary Tables 1,2,3) and we find that GWAVA consistently ranks the candidate functional variant highly.

To establish if 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 some small number of 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 data from chromosome 22 from the training set. We find that we can discriminate the spike-in variants from the background variants with good accuracy (AUC = 0.85, Supplementary Fig. 5), though at reasonable score thresholds we would still expect a substantial number of false positives in a whole genome. In this context 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. To establish if the scores might help identify the functional variant assuming that we know the relevant gene (from other evidence such as known disease-implicated coding variants from the same locus) we carried out a further experiment with this data set. For each of the 24 unique genes annotated as being affected by the HGMD variants, we identify all non-coding variants from NA06984 in the region around each gene (5kb up and downstream) and observe where the spike-in variant is ranked according to the GWAVA score (Supplementary Table 4). We find that we rank the spike-in variant top for 5 genes and in the top 3 for 10 genes, significantly more often than expected by chance ($P = 0.003$ and $P = 0.0002$ respectively, by simulation).

Finally, as an application to cancer studies, we annotate non-coding somatic mutations discovered in whole-genome sequencing studies from the COSMIC database12. We identified all mutations that are found to recur in different studies (n = 812) found that these recurrent mutations are assigned a significantly higher average GWAVA score than the non-recurrent mutations ($P = 3.1\times10^{-61}$, AUC = 0.67, Fig. 2). Recurrence of somatic mutations is a widely used proxy of likely function, and so this result represents a validation from an entirely different domain that the classifier is able to identify likely functional sequence variants, and suggests that this approach might also be useful in prioritising mutations for follow-up in the search for cancer driver mutations.

We sought to compare GWAVA scores with other tools that can classify non-coding variants. The only such tool we are aware of is MutationTaster13 which can provide predictions for non-coding variants that can be mapped to a transcript model, such as those in untranslated regions and introns. In order to address the issue that MutationTaster is trained many of the same HGMD variants that are used to train GWAVA, and that known disease implicated variants (such as those in ClinVar) are automatically classified as disease causing by the MutationTaster webserver, we used the set of non-coding somatic mutations from COSMIC to compare the approaches on those mutations where both tools can make a prediction. We obtained predictions from the MutationTaster webserver for 92,352 non-coding 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”. In order to compare results we therefore threshold the GWAVA score at 0.5, with mutations scored > 0.5 identified as “functional” and those ≤ 0.5 as “non-functional”. We computed contingency tables comparing mutations identified as functional by either tool with whether the mutations are recurrent. We find that, while there is a significant enrichment for recurrent mutations among those called as functional for both tools, the odds ratio for GWAVA was 5.4 (Fisher’s exact $P = 1.3\times10^{-56}$), higher than the result for MutationTaster (odds ratio = 2.03, Fisher’s exact $P = 6.5\times10^{-08}$).

We have presented a computational approach to predicting the functional impact of non-coding variants and have demonstrated that the technique can combine information from a wide range of available annotations, addressing issues of context dependency and inconsistent signal from evolutionary conservation in regulatory elements. The classifier software and annotation data are freely available for download. We have precomputed classifier scores for all known variants from the Ensembl variation database14 (release 70) and these scores, along with the underlying annotations, are available from a webserver.

GWAVA represents the most widely applicable technique currently available for annotation of non-coding variants, and we hope that by incorporating the predictions into disease association studies we will substantially improve chances of finding variants relevant to disease and other phenotypes.

Online methods

Annotation sources

We acquired a wide range of annotations at a range of different scales and in a variety of data formats. Here we identify all the annotations we used in this study grouped by the class of data and with the data sources identified.

- Open chromatin:
  - DNase-seq & FAIRE-seq peak calls (ENCODE)
  - DNase footprints (ENCODE)
- TF binding:
  - ChIP-seq peak calls for 124 transcription factors (ENCODE)
  - JASPAR motifs aligned under corresponding factor ChIP-seq peaks (Ensembl)
  - Bound TF binding motifs (ENCODE)
- Histone modifications:
  - ChIP-seq peak calls for 12 different modifications (ENCODE)
- RNA Polymerase binding:
  - ChIP-seq peak calls (ENCODE)
- CpG islands:
  - Predictions from Ensembl14
- Genome segmentation:
  - Ensembl integration15 of the ENCODE SegWay16 and ChromHMM17 segmentation calls, 7 discrete states identified
- Conservation:
  - GERP scores from mammalian alignments (Stanford), both at the specific variant nucleotide and an average over the 100bp surrounding each variant18
- Human variation:
  - Variants, allele frequencies and ancestral allele calls from the 1000 Genomes Project phase 1 data
    - Mean heterozygosity of variants in 1kb window, calculated from global population frequencies
    - Mean derived allele frequency of variants in 1kb window, again calculated from global population frequencies

- Genic context:
    - Distance to the nearest transcription start site (TSS) (GENCODE/ENCODE)19
    - Distance to the nearest splice site (GENCODE/Ensembl)
    - Summary gene region annotations; any base annotated as exonic, intronic, CDS, 5’ or 3’ UTR, splice site, start or stop codon in any transcript (GENCODE/Ensembl)
- Sequence context:
    - GC content calculated over the 100bp surrounding each variant (GRC)
    - Boolean variable indicating if the variant is in a CpG context in the reference assembly (GRC)
    - The reference nucleotide at the variant position (GRC)
    - Boolean variable indicating if the variant falls in repeat sequence (UCSC)

We developed a pipeline that can annotate a given set of variant loci with all these annotations. The result of this pipeline is essentially a large matrix with a row for each variant locus and a column for each possible annotation. The type of each column depends on the annotation class, but can be one of three classes:

1. a count of the number of cell lines in which the variant locus overlaps some annotation, such as DNase1 hypersensitive sites and ChIP-seq peaks
2. a binary flag where the annotation is simply presence or absence of the annotation at the variant locus (e.g. is this region ever in an annotated intron)
3. 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 is composed of all variants annotated as ‘regulatory mutations’ from the April 2012 release of HGMD, and downloaded from Ensembl release 70. After removing variants at the same positions this left a set of 1614 disease-implicated SNVs. For all 3 control sets we use 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 those variants with a minor allele frequency above or equal to 1% in the global population to reduce the chance of including rare functional variants in our control set (we have performed sensitivity analyses by only focusing on variants from European populations and also to rare, singleton variants and we find qualitatively similar cross-validation results to those with the set of common variant controls, data not shown). As we only had SNVs in our disease set we also limited our analysis to SNVs in each of our control sets. This left us with a total of 15,730,276 potential control SNVs. The first control set was simply a random selection of SNVs from across the genome, 100 times as many as the number of disease implicated variants in order to get a reasonable sample of the background, while making the 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. In this set we include 10 times as many control variants as disease implicated variants as we found that this was as large a set as we could construct while ensuring the distributions of distances matched the HGMD variants. The final control set is composed of all 1KG variants in the 1kb surrounding each of the HGMD variants (n = 5027).

Individual feature analysis

We investigated if any of the annotations show a different distribution in the disease and control sets. These annotations can be grouped into two classes of features: a large class of regional data where the annotation for each variant is a count of whether the variant is an annotated element, possibly across multiple cell lines, and several continuous variables. We used different analysis approaches for each of these classes. For the regional data we ignore the number of cell lines in which a variant is found (as these are not independent across cell lines) and just use a single binary variable per feature indicating if each variant is found in this element in any cell line. Annotations not specific to a cell line are already binary. For each feature we then compute a contingency table identifying how these counts differ in our disease and control sets. We used Fisher's exact test to compute the significance of the enrichment or depletion.

For continuous features we used a two-sided Mann-Whitney U test to establish if there is 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 use the absolute value of the measure in these analyses (though we supply the original signed value to the classifier as taking into account whether the variant is up or downstream from the nearest TSS may be informative). All *P*-values are adjusted using the Bonferroni correction to account for multiple testing. Unadjusted *P*-values are also reported (Supplementary Table 5).

We show the relative proportion of disease variants overlapping all annotated functional elements compared to each of the three control sets (in this discussion we refer to statistics comparing the variants matched by distance to the nearest TSS) (Supplementary Fig. 6). As expected, we find that the disease variants are enriched in number of functional elements, in particular in open chromatin (DNase1 peaks $P = 2.9 \times 10^{-55}$, DNase1 footprints $P = 4.5 \times 10^{-53}$), transcribed DNA (non-coding exonic DNA $P = 2 \times 10^{-51}$, RNA Polymerase II $P = 2.4 \times 10^{-82}$), protein binding sites (JUND $P = 7.2 \times 10^{-51}$, SP1 $P = 3.6 \times 10^{-47}$), several histone modifications which indicate both active gene expression and regulatory activity (H3K4me3 $P = 8.1 \times 10^{-112}$, H3K4me2 $P = 1.2 \times 10^{-101}$) and, perhaps unexpectedly, repressive marks such as H3K27me3 ($P = 1.3 \times 10^{-55}$). We observe that the enrichments generally decrease as the control variants are more tightly matched to the disease variants, and that more specific annotations, such as the DNase1 footprints and TF binding sites are found to be more significant for the control set matched by region than in the other two sets.

We also compared several quantitative genome-wide variables including evolutionary conservation and GC content (Supplementary Fig. 7). These results demonstrate that the disease-associated variants are generally found nearer TSSs than controls (except, as expected, in the control set where we explicitly match the variants by this feature), in GC-rich sequence, in regions with less variation in human populations and in more conserved regions. The differences we find in conservation at the variant positions, both at the specific nucleotide ($P = 1.8\times10^{-21}$) and in the flanking 100 bases ($P = 4.7\times10^{-12}$), are statistically significant but small. We also find significant differences in the two measures of diversity in human populations; the average heterozygosity ($P = 5.7\times10^{-08}$) and derived allele frequency ($P = 2.5\times10^{-11}$) in the kilobase surrounding each variant. Both these classes of annotation are intended to help identify genomic regions under constraint at different timescales, and this result implies that for regulatory elements evidence for constraint at shorter timescales than traditional conservation metrics is also informative.

These analyses demonstrate that a large number of features show significant differences in distribution between the disease variants and our control sets. However, while the differences are statistically significant, these features are not individually predictive of disease status and this motivates the use of an integrated approach that combines these features to provide an overall predictor of likely functionality.

Classifier algorithm

The form of available annotation data drives the need for a technique that can simultaneously handle a large number of continuous and categorical features. In two of the control sets we also have a very unbalanced distribution of classes in that there are considerably fewer disease-implicated variants than controls. To address both of these issues we use a slightly modified version of the Random Forest algorithm10. Random Forests are a robust and widely used approach to classification that can deal with the different feature types we use and are robust to the presence of features that are not predictive (so we do not perform any feature selection). The modification we make to the standard algorithm is to address the class imbalance issue, when sampling the training set for each component decision tree in the forest we sample from the two classes such that there is an even class distribution in the training set. This means that each tree is trained on a relatively smaller subset of the control variants, but we use enough trees in the forest 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 allows us to compute the relative importance of each feature from the trained model.

We train 3 forests, one for each of the different sets of controls, with the same set of disease associated variants as the positive class in each training set. We experimented with a range of different numbers of decision trees in the forest and found that performance seems to saturate when using around 100 trees, and this number should also ensure we should sample a good proportion of variants in each of our 3 training sets. We use the mean area under the

ROC curve (AUC) 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 so may share annotations. When performing cross-validation, if variants from the same gene appear in both the training and test sets this may inflate the performance statistics. To control for this we created a stringent set of disease variants where a single variant is randomly selected for each gene and we observe 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 library20. The modified source code is available at the URL below.

Feature importance

In order to identify which features are contributing to the discriminative ability of each classifier we computed the relative Gini importance of each feature across each component tree of the 3 forests (Supplementary Fig. 1). Gini importance measures the mean decrease in impurity at each node in the tree due to the feature of interest, weighted by the proportion of samples reaching that node. In the first training set distances to the nearest TSS or splice site are clearly the most important features, and this reflects the fact that the HGMD variants are on average much nearer genes than randomly selected controls. We see that these are relatively less important signals for the other two sets, though distance to the nearest TSS is still the third most important feature in the training set matched on this feature. This implies that, conditional on some other annotations, this is still an informative feature. Generally we observe that annotations that are available across more of the genome, and those which are not specific to any particular cell line, are ranked as more important by the classifier and this is expected as the decision tree building algorithm will be able to use them more often. However we also observe that the DNase1 footprints, one of the more specific classes of annotations, only appears in the top 20 for our training set matched by region, suggesting that as the two classes of variants are more closely matched, more specific annotations become more important. It is interesting to observe that conservation scores are consistently identified as an important annotation, despite the fact that there is only a small difference in average scores between any of our datasets. Again, this result implies that conditional on other annotations around the variant, evolutionary conservation is still an important signal.

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). While the distributions are somewhat different for each classifier, as expected, few variants are assigned high scores by any version. These distributions allow 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 (filename: clinvar_20130118.vcf). We identified all variants annotated as "pathogenic" (those with CLNSIG=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 non-coding variants we used in this analysis. We performed a similar filtering to identify all likely non-pathogenic variants annotated (those with CLNSIG = 2 or CLNSIG = 3) and derived a set of 150 non-pathogenic non-coding 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 include 1000 Genomes variants with MAF > 1%, and we included 100 control variants for each ClinVar variant resulting in a set of 19400 control variants. We annotated these 3 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. 2).

Annotating GWAS SNPs

We downloaded the GWAS catalogue from the NHGRI 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 10 matching SNVs for each GWAS signal. The genotyping platforms used were:

Affymetrix GeneChip 100K

Affymetrix GeneChip 500K

Affymetrix SNP6

Illumina HumanCNV370 Quadv3

Illumina HumanHap300v2

Illumina HumanHap550v3.0

Illumina Cardio Metabo

Illumina Human1M-duoV3

Illumina Human660W-quad

We compared the score distributions of these two sets of variants with a two-sided Mann-Whitney U test (Supplementary Fig. 3).

We downloaded the replication status annotations available in the supplementary materials from Maurano et al.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. 4). Comparison of score distributions was performed with a two-sided Mann-Whitney U test. The *P*-values comparing all pairwise combinations of these 3 sets of variants are:

Not replicated vs. internally replicated: $P = 2.56\times10^{-09}$

Not replicated vs. independently replicated: $P = 3.65\times10^{-07}$

Internally replicated vs. independently replicated: $P = 0.024$

Following up disease association signals

The classifier scores can help prioritise a list of candidate variants identified as potentially interesting by other means, for example variants in high linkage disequilibrium with some significantly associated variant from a GWAS or sequence-based association study. There are three case studies we are aware of where a non-coding complex disease-associated variant is deemed to be causal.

In the first study, Musunuru et al.21 investigated a locus on chromosome 1p13 that has been strongly associated in GWAS with plasma low-density lipoprotein cholesterol (LDL-C). The authors performed a fine-mapping study to identify the minimal genomic region responsible for the association and identify 20 single SNVs in the region, of which 6 SNVs with the strongest association cluster in a 6.1kb non-coding region including the 3’ UTRs of *CELSR2* and *PSRC1.* They identify a single variant, rs12740374, that had evidence of association in a distinct population from the original discovery population, and which alters a binding site for the TF CEBPA. The authors also demonstrate that this variant alters hepatic expression of the *SORT1* gene. We annotated the 20 SNVs from the fine-mapping study and find that rs12740374 is ranked fifth highest by classifier score (with a score of 0.58) among these 20. The annotations found for this variant include DNase1 hypersensitive sites and footprints in multiple cell lines, along with binding sites for CEPBA and several other transcription factors (Supplementary Table 1).

Adrianto et al.22 followed up on an association of the *TNFAIP3* locus with systemic lupus erythematosus (SLE), and using a fine-mapping approach identified a ‘risk haplotype’ comprising 28 variants with strong evidence of association with the disease. The authors find that a novel polymorphism among these shows evidence of affecting NFKB binding which appears to have a subsequent effect on *TNFAIP3* expression. We find that this polymorphism is ranked highest among these 29 variants (with a classifier score of 0.77), and the variant is annotated with many of the same annotations used by the authors in selecting this variant for follow-up, including a ChIP-seq peak for NFKB, along with several other TFs and overlapping DNase1 footprints in 2 cell lines, and that this variant is in a highly conserved region (Supplementary Table 2).

In a study focused on type 2 diabetes, Gaulton et al.23 used FAIRE-seq data to identify variants lying in regions of open chromatin in human pancreatic islet cells. The authors identified 5 SNPs in linkage disequilibrium with the rs7903146 GWAS signal at the *TCF7L2* locus. They found that, of these, only rs7903146 maps to a region of islet-selective open chromatin and the risk allele of this variant shows significantly increased enhancer activity in MIN6 cells. We annotated these 6 variants and computed classifier scores as previously (Supplementary Table 3). In this case, the putatively functional variant is ranked second in the list of 6. However, all variants have low scores and it appears that the regulatory element tagged by this variant is acting in a very tissue specific manner.

Application to personal genomics

We downloaded the 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. 5).

For the individual gene analysis, there are 24 unique genes annotated in HGMD as being affected by this set of 33 variants, and for genes where there was more than one variant we randomly selected a single variant and disregard the rest. We downloaded the coordinates from each of these genes from Ensembl and identified all variants from NA06984 that overlapped the gene region ±5kb (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 below, along with all other software) to establish how often we would expect to find a result as extreme or more extreme as 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 non-coding 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 ID) 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 non-coding 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 = 1340). 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 (as above), and used Fisher's exact test to compute the significance of the enrichment. To compare this result with 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 "non-functional" and again used a contingency table to compute the enrichment of recurrent mutations among those called as functional.

Classifier availability

We have termed the method "GWAVA" (for Genome Wide Annotation of VAriants), and we have developed a web server that allows users to retrieve precomputed scores from each of the three classifiers for all known germline and somatic SNVs found in Ensembl release 70. All the underlying annotations used by the classifier are also available. The URL of this resource is:

http://www.sanger.ac.uk/resources/software/gwava

The source code, documentation, set of annotations used, and all variant data sets described here are available for download from the FTP server linked from the GWAVA webpage. A plugin for the Ensembl Variant Effect Predictor24 is also available from this location.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

GRSR is supported by EMBL and the Sanger Institute via an EBI-Sanger Postdoctoral (ESPOD) Fellowship. This work was funded by the Wellcome Trust (098051) and by the European Molecular Biology Laboratory. The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement [nº 282510 - BLUEPRINT].

References

1. Hindorff LA, et al. PNAS. 2009; 106:9362–9367. [PubMed: 19474294]
2. Cooper GM, Shendure J. Nat Rev Genet. 2011; 12:628–640. [PubMed: 21850043]
3. The ENCODE Project Consortium. Nature. 2012; 489:57–74. [PubMed: 22955616]
4. Bernstein BE, et al. Nat Biotechnol. 2010; 28:1045–1048. [PubMed: 20944595]
5. Kumar P, Henikoff S, Ng PC. Nat Protoc. 2009; 4:1073–1081. [PubMed: 19561590]
6. Adzhubei IA, et al. Nat Methods. 2010; 7:248–249. [PubMed: 20354512]
7. Schmidt D, et al. Science. 2010; 328:1036–1040. [PubMed: 20378774]
8. Stenson PD, et al. Genome Med. 2009; 1:13–13. [PubMed: 19348700]
9. The 1000 Genomes Project Consortium. Nature. 2012; 491:56–65. [PubMed: 23128226]
10. Breiman L. Machine Learning. 2001; 45:5–32.
11. Maurano MT, et al. Science. 2012; 337:1190–1195. [PubMed: 22955828]
12. Forbes SA, et al. Nucleic Acids Res. 2011; 39:D945–D950. [PubMed: 20952405]

13. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. Nat Methods. 2010; 7:575–576. [PubMed: 20676075]

14. Flicek P, et al. Nucleic Acids Res. 2012; 41:D48–D55. [PubMed: 23203987]

15. Hoffman MM, et al. Nucleic Acids Res. 2012; 41:827–841. [PubMed: 23221638]

16. Hoffman MM, et al. Nat Methods. 2012; 9:473–476. [PubMed: 22426492]

17. Ernst J, Kellis M. Nat Methods. 2012; 9:215–216. [PubMed: 22373907]

18. Davydov EV, et al. Plos Comput Biol. 2010; 6:e1001025. [PubMed: 21152010]

19. Harrow J, et al. Genome Research. 2012; 22:1760–1774. [PubMed: 22955987]

20. Pedregosa F, et al. J Mach Learn Res. 2011; 12:2825–2830.

21. Musunuru K, et al. Nature. 2010; 466:714–719. [PubMed: 20686566]

22. Adrianto I, et al. Nat Genet. 2011; 43:253–258. [PubMed: 21336280]

23. Gaulton KJ, et al. Nat Genet. 2010; 42:255–259. [PubMed: 20118932]

24. McLaren WM, et al. Bioinformatics. 2010; 26:2069–2070. [PubMed: 20562413]

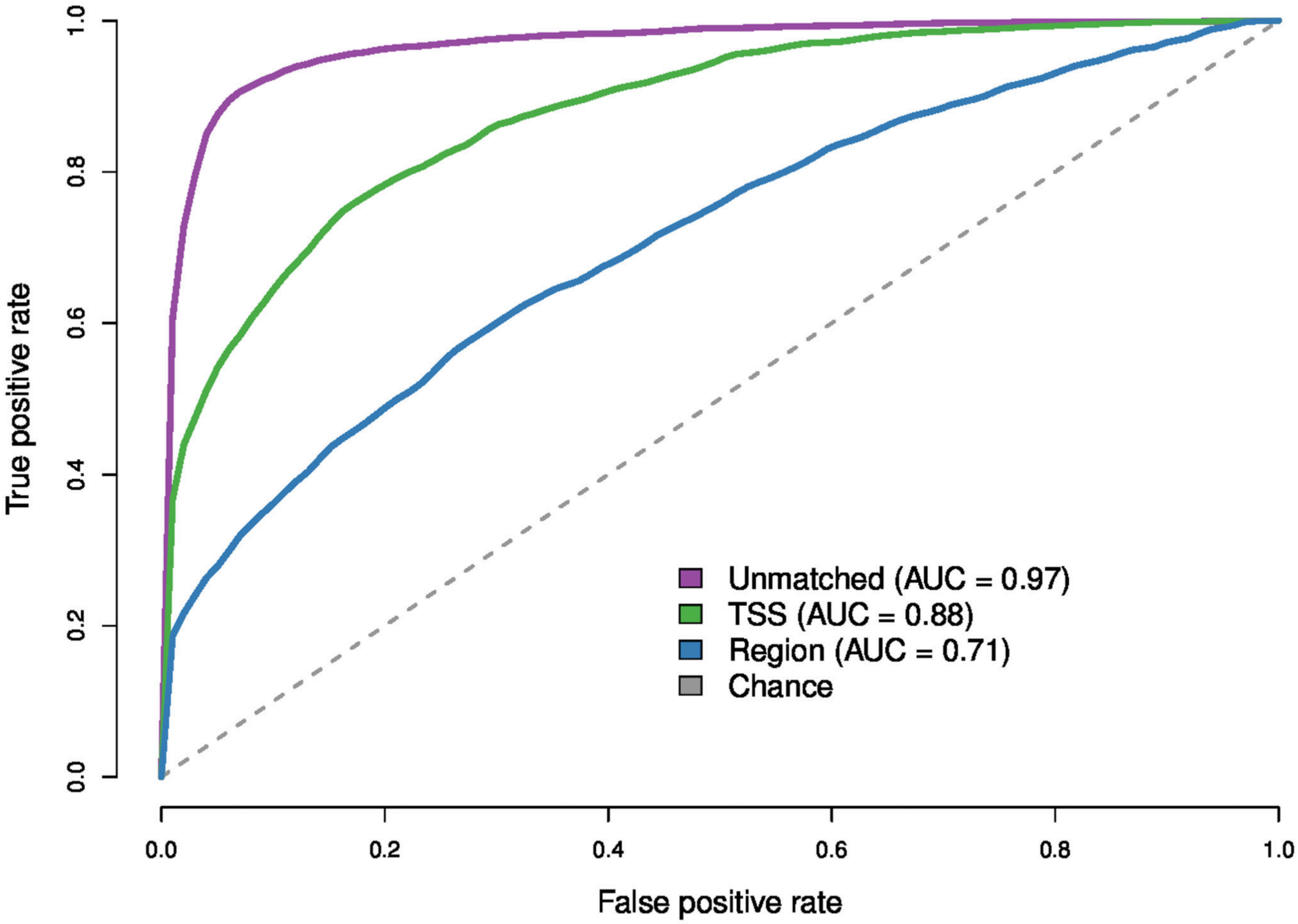

Figure 1.

Mean receiver operating characteristic (ROC) curves for 10 fold cross-validation experiments on each of the three training sets. The area under the curve (AUC) statistics illustrate that all classifiers are able to discriminate between the disease variants and controls, though performance depends on how well the variants in the training sets are matched.

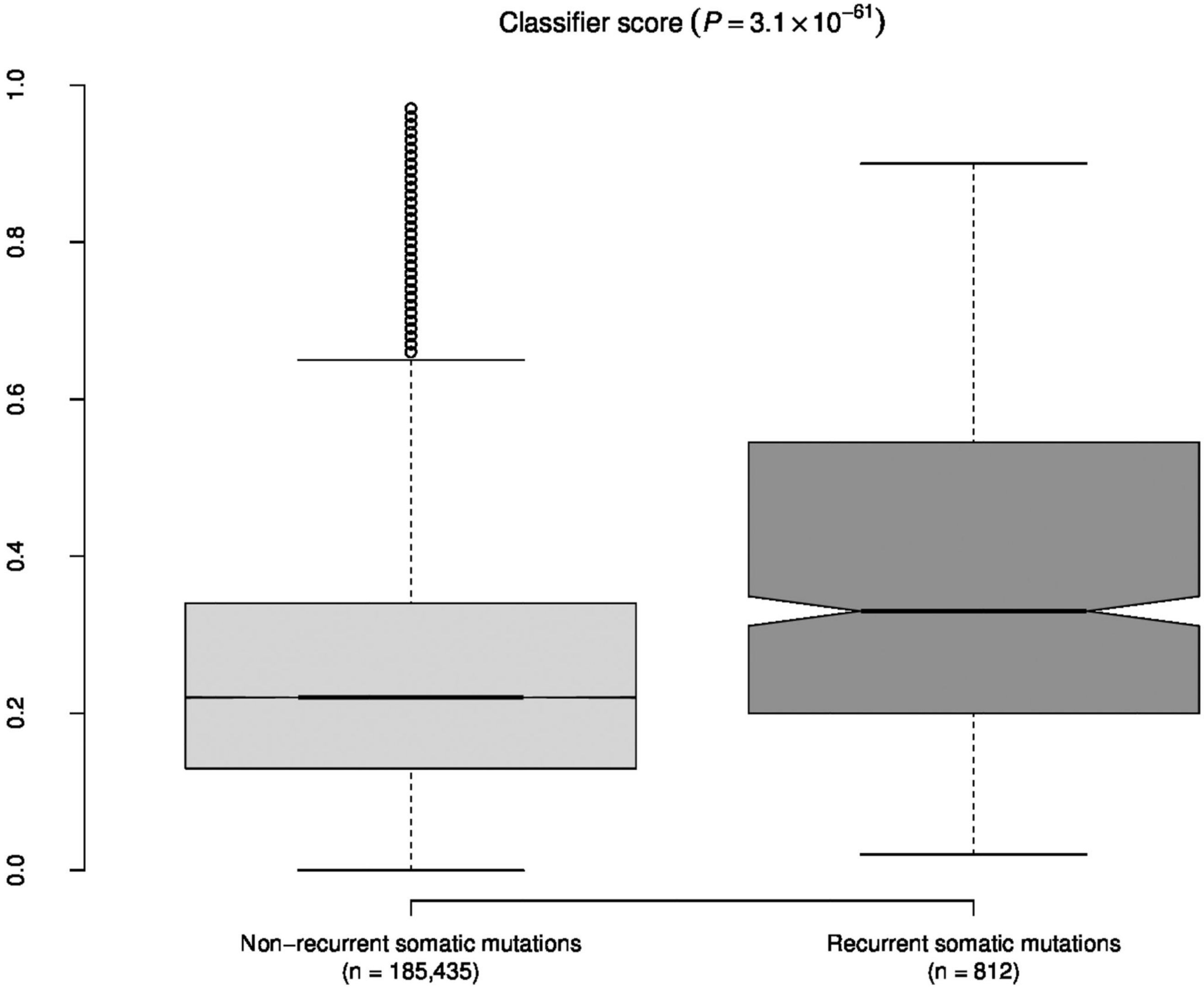

Figure 2.

Classifier scores for recurrent (n = 812) vs. non-recurrent (n = 185,435) non-coding somatic mutations from COSMIC. The AUC for discriminating between these two classes of mutation is 0.67. The *P*-value is calculated using a two-sided Mann-Whitney U test. The whiskers include scores within 1.5 × IQR of the upper and lower quartiles (the default in the R package).