Human Mutation
OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

Regulatory Single-Nucleotide Variant Predictor Increases Predictive Performance of Functional Regulatory Variants

Thomas A. Peterson,[1] Matthew Mort,[2] David N. Cooper,[2] Predrag Radivojac,[3] Maricel G. Kann,[1] and Sean D. Mooney[4]*

[1]Department of Biological Sciences, University of Maryland, Baltimore County, Baltimore, Maryland; [2]Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, CF14 4XN, United Kingdom; [3]Department of Computer Science and Informatics, Indiana University, Bloomington, Indiana; [4]Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, Washington

Communicated by Paolo M. Fortina

Received 4 March 2016; accepted revised manuscript 28 June 2016.
Published online 13 July 2016 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.23049

ABSTRACT: In silico methods for detecting functionally relevant genetic variants are important for identifying genetic markers of human inherited disease. Much research has focused on protein-coding variants since coding regions have well-defined physicochemical and functional properties. However, many bioinformatics tools are not applicable to variants outside coding regions. Here, we increase the classification performance of our regulatory single-nucleotide variant predictor (RSVP) for variants that cause regulatory abnormalities from an AUC of 0.90–0.97 by incorporating genomic regions identified by the ENCODE project into RSVP. RSVP is comparable to a recently published tool, Genome-Wide Annotation of Variants (GWAVA); both RSVP and GWAVA perform better on regulatory variants than a traditional variant predictor, combined annotation-dependent depletion (CADD). However, our method outperforms GWAVA on variants located at similar distances to the transcription start site as the positive set (AUC: 0.96) as compared with GWAVA (AUC: 0.71). Much of this disparity is due to RSVP's incorporation of features pertaining to the nearest gene (expression, GO terms, etc.), which are not included in GWAVA. Our findings hold out the promise of a framework for the assessment of all functional regulatory variants, providing a means to predict which rare or de novo variants are of pathogenic significance.

Hum Mutat 37:1137–1143, 2016. © 2016 Wiley Periodicals, Inc.

KEY WORDS: RSVP; ENCODE; regulatory variant; intergenic; variant prediction; variant predictor

Introduction

Although it is likely that most single-nucleotide variants (SNVs) lack functional significance [Sachidanandam et al., 2001], SNVs are the most common form of human genetic variation [Gibbs et al., 2003] and many have been shown to be associated with, or even causative of, human disease [Buckland et al., 2004; Pastinen and Hudson, 2004; Prokunina and Alarcon-Riquelme, 2004; Campino et al., 2008; Savinkova et al., 2009; Ward and Kellis, 2012]. A substantial body of research has been devoted to characterizing functional SNVs in human genes [Mottagui-Tabar et al., 2005; Buckland, 2006; Pampin and Rodriguez-Rey, 2007; Chorley et al., 2008]. However, it is currently impractical to investigate each variant in vitro given the very large number that have been identified in the human genome. Thus, computational (in silico) approaches for the prediction of functional SNVs represent a promising alternative to laborious large-scale in vitro analyses [Mooney, 2005; Peterson et al., 2013].

Several studies have been conducted to predict functional SNVs within the human genome using information that could be biologically relevant to gene transcriptional regulation. A number of such studies, focusing exclusively on promoter regions, have considered only the disruption of transcription factor-binding sites (TFBS) for prediction [Ponomarenko et al., 2002; Andersen et al., 2008; Lapidot et al., 2008]. However, these methods rely on the completeness of genome-wide TFBS annotation and blithely ignore coding variants that may also influence gene transcriptional regulation. Moreover, these methods lack information pertaining to the distance to the nearest transcription start site (TSS), the presence of a CpG island, and local sequence repetitivity, which were found by Montgomery et al. (2007) to be important features for the prediction of functional regulatory variants. In addition, the integration of information from the Encyclopedia of DNA Elements (ENCODE) has previously been reported to improve the prediction of functional regulatory elements [Torkamani and Schork, 2008], a finding that was recently confirmed by the Genome-Wide Annotation of Variants (GWAVA) tool [Ritchie et al., 2014]. However, all the aforementioned tools utilize only SNV-based information about the variant itself and ignore important information about the gene being regulated (gene-based features) such as gene expression, gene function, interaction complexity of gene products, frequency of optimal codons (FOP), and effective number of codons (ENC). In our previous study, where we developed a classifier to distinguish functional regulatory variants from putatively neutral variants [Zhao et al., 2011], we demonstrated that these gene-based features are very important. The GWAVA tool did not incorporate replication timing information from the ENCODE project, which we found to be a very informative feature in this analysis.

In our regulatory single-nucleotide variant predictor (RSVP), we have employed a supervised machine learning method using a set of 1,999 known functional regulatory SNVs from the Human Gene Mutation Database (HGMD) [Stenson et al., 2014] together with a dataset of putatively neutral SNVs that commonly occur in human genomes. The integration of both SNV-based and

Additional Supporting Information may be found in the online version of this article.

Contract grant sponsors: National Institutes of Health (NIH) (R01LM009722, R01MH105524, and 1K22CA143148); Qiagen Inc.

*Correspondence to: Dr. Sean D. Mooney, University of Washington, Medical School Seattle, Washington. E-mail: sdmooney@uw.edu

gene-based features enabled comparable performance to GWAVA when predicting all types of variant, with both methods achieving an area under the receiver operating characteristic (ROC) curve (AUC) value of 0.97. However, RSVP exhibits improved performance when classifying using only negative variants in close proximity to the true regulatory variants (0.95 AUC as compared with GWAVA's 0.71 AUC). Additionally, we demonstrate that RSVP's performance improves when analyzing only disease-causing (DM) variants from HGMD that are supported by functional evidence from in vitro expression assays (achieving an AUC of 0.98).

In this analysis, we determined that genomic features discovered in the ENCODE project are powerful predictors of regulatory variants while emphasizing the importance of gene-based features and the variant's distance to the TSS. Our findings hold out the promise of an in silico framework for the assessment of all functional regulatory variants and, more importantly, provide a means to predict which rare or de novo variants may be involved in genetic disease. To annotate new variants with the RSVP predictor, a script may be obtained from our FTP site (http://bioinf.umbc.edu/RSVP/ftp/) or precomputed scores for all variants in dbNSFP [Liu et al., 2011] can be obtained upon request.

Materials and Methods

Data Preparation

In order to evaluate features (attributes) with the potential to be useful in identifying sites that play a functional role in gene regulation, two datasets were collected. First, a positive set was obtained from the HGMD database [Stenson et al., 2012] on January 30, 2014, and this was filtered so as to include only "disease-causing mutations" (DM), "functional polymorphisms" (FP), and "disease-associated polymorphisms" with additional supporting functional evidence (DFP), resulting in a total of 1,999 regulatory variants. Second, variants were obtained from the dbSNP database and were filtered so as to include only validated variants with an allele frequency greater than or equal to 5%, yielding a set of 4,804,913 negative variants. To make the analysis computationally viable, downsampling was employed on the dbSNP database with a random selection of 1% of these variants being used, yielding a set of 51,147 variants, which we shall henceforth refer to as the "full" negative set. To control for the different distribution of distances relative to the TSSs between the positive and negative datasets, we created two additional negative datasets to assess the possible impact of this feature. The "±1,000" and "±500-bp" region sets comprise all variants from the dbSNP database set within a window around any positive variant of length 1,000 and 500 bp, respectively. This resulted in a total of 90,628 variants for the ±1,000-bp region set and 18,872 for the ±500 bp region set. To analyze the predictor's performance on independent datasets, we obtained the full genome sequences of 454 individuals from the Wellderly Project (downloaded December 26, 2014), 2,513 individuals from the 1000 Genomes Project [Abecasis et al., 2012) (v5a.20130502), somatic mutations from the tumors of 1,043 patients with breast invasive carcinoma (BRCA) from the TCGA project [Collins and Barker, 2007] (downloaded July 2, 2014), and a list of genes with a known relation to cancer from the Cancer Gene Census (CGC) [Futreal et al., 2004]. In addition, an update to the HGMD regulatory variant dataset was obtained on May 20, 2015, which was used as an unseen dataset for further validation of the method.

Features

Features used in this study were divided into two distinct sets: those directly annotated to the SNV under consideration (SNV based) and those annotated to the gene within whose transcription regulatory region the SNV lies (gene based). A summary of all features used in this analysis can be found in Supp. Table S1. SNV-based features include SNV distance to the nearest TSS, flanking nucleotide GC content, flanking nucleotide conservation, and SNV occurrence within known functional elements. The SNV-based known functional elements used were collected from several independent studies as well as the ENCODE project [Kellis et al., 2014]. These binary SNV-based features describe the variant as being located within functional elements collected from independent studies for enhancers [Pennacchio et al., 2007], insulators [Bell et al., 2001], RNA polymerase II-enriched regions [Barski et al., 2007], conserved noncoding sequences [Wang et al., 2006], and nuclease hypersensitive sites [Crawford et al., 2006]. In addition to the nuclease hypersensitive sites from the independent study, "DNase peaks" were obtained from the ENCODE project and were treated separately from the other nuclease hypersensitive sites. Also obtained from the ENCODE project were continuous-value features for histone peaks, FAIRE peaks, DNA methylation sites, and replication timing. Where a genomic region had not been annotated for replication timing, the average between the two nearest annotated regions was used. Two binary features from ENCODE were used to specify whether the variant was located within a TFBS or a CpG island. Furthermore, two other binary features, one for exonic and one for intronic, were used to describe the location of the variant in the context of the gene. Finally, a database of sequences known to be exonic splicing enhancers (ESE) and exonic splicing silencers (ESS) were obtained from Sterne-Weiler et al. (2011) and one feature for each was used, indicating a gain or loss of an ESE or ESS site.

Gene-based features for a given variant refer to features that are associated with the closest gene on the chromosome, and all variants found near that gene would share this annotation. Gene-based features were further divided into two sets: those pertaining to the function of the associated gene (function based) and those relating to the mRNA expression of the associated gene (expression based). For function-based features, a set of prediction scores for GO biological process (1,788 features) and molecular function terms (344 features) was generated using the FANN-GO (functional annotator that uses multioutput artificial neural networks) predictor of protein GO term annotations [Clark and Radivojac, 2011]. The use of predicted GO terms instead of experimentally determined annotations allowed us to obtain values for all data points and a set of features that was less likely to be biased toward genes frequently studied by biomedical researchers (which could have resulted in an overestimation of performance accuracy). We also included interaction complexity (node degree in a protein–protein interaction network), which is derived from high-throughput experiments in this subset of functional features. Expression-based features were generated using microarray platforms GPL1074 and GPL96 [Su et al., 2004]. A set of 158 features was generated that represent the normalized expression levels of each gene across 79 tissues. Features pertaining to the mean, standard deviation, coefficient of variation, and maximum and minimum expression levels of each gene across tissues were also generated. Finally, two codon-usage features that were not classified as being either expression based or function based were generated using CodonW's [Sharp and Li, 1986] FOP and ENC.

Machine Learning and Cross-Validation

An ensemble of decision trees was employed using the "treefit" function in MATLAB with default parameters; 10-fold cross-validation was used to assess our prediction performance with respect to unseen variants not found in our training set. An ensemble of 1,000 trees was used, which was found to have better performance than ensembles of 100 or 500. A random selection of features was used for each tree with a minimum of five and a maximum of 1,000 features per tree, ensuring that the ensemble is not dominated by a few powerful features. For cross-validation, our dataset was divided into 10 partitions each containing a random selection of 10% of the positive variants and 10% of the negative variants. For each of the 10-fold, one partition was left out for testing and nine partitions were used for training. During this process, each partition was used only once for testing. At each fold, a predictor was trained using 1,000 decision trees. For each tree, we balanced our training data by downsampling our negative variants using random selection until our positive and negative variants were of equal size. The final prediction score was an average of all 1,000 scores output by the ensemble of decision trees; we calculated the AUC of all predicted scores using 1,000 evenly spaced thresholds between 0 and 1. Separately, to examine the effect of training on some of the same genes that are in the test set, we performed a cross-validation using "gene colocation," meaning that each gene could be found in either only the test set or only the training set. In this analysis, the predictor will never train on any gene that is in the test set or vice versa. Classification performance was calculated using the AUC. Spanning the interval between 0 and 1 with 1 being the best possible predictor, the AUC measures the true positive rate (sensitivity) as a function of the false-positive rate (1, specificity).

Results

Cross-Validation Performance

When evaluating the ability of RSVP to discriminate between putatively neutral variants and variants that cause regulatory abnormalities, we achieved an AUC of 0.97. This indicates a marked improvement over our previous model [Zhao et al., 2011], which achieved an AUC of 0.90. To demonstrate that RSVP outperforms a traditional approach for regulatory variants, we compared it with combined annotation-dependent depletion (CADD) [Kircher et al., 2014], a metapredictor that also uses ENCODE features in addition to scores from PolyPhen, SIFT, and several other predictors as features for a SVM. We scored the entire positive and negative dataset created in this study with the CADD Webserver and normalized the PHRED scores between 0 and 1. This resulted in an AUC of 0.84 for the CADD predictor, indicating that RSVP will perform better with respect to identifying variants that play a role in gene regulation. Additionally, we compared RSVP's results with DeepSEA [Zhou and Troyanskaya, 2015], a deep learning-based tool for predicting chromatin effects of disease-associated HGMD variants. We found that DeepSEA achieves an AUC of 0.81 on our training data, suggesting a need for tools that are trained specifically on variants that cause regulatory abnormalities since such variants are expected to have different properties than disease-associated variants with no impact on gene regulation. Finally, we compared RSVP's results with the GWAVA tool, which achieved a similar AUC (0.97) using the same HGMD regulatory variant dataset and a similar negative dataset. However, RSVP performs significantly better when compared with GWAVA's AUC of 0.71 for their "region" negative dataset, which comprises all negative variants in the region of ±500 bases around the positive variants. When training and testing our model on our ±1,000 and ±500 bp region sets, our model achieved AUCs of 0.96 and 0.95, respectively (Fig. 1). In addition to analyzing our performance on all variants from the HGMD regulatory variant dataset, we also performed our analysis on the DM, DFP, and FP subsets of the positive variants (Fig. 2). Of these subsets, training and testing using only the DM subset showed the best performance (AUC of 0.98), followed by DFP (AUC of 0.96), and finally FP (AUC of 0.96). This suggests that RSVP performs better on variants that are causatively involved with disease but the tool still performs well on known functional variants with no known disease association. Furthermore, since the performance of our tool will inevitably rely on the chosen threshold, we provide users with the F-measure and accuracy performance at different thresholds corresponding to estimated 1%, 5%, and 10% false-positive rates (Table 1).

In practice, our predictor will see many of the same genes during training and testing. However, since our predictor uses several thousand gene-based features that can be identical for many variants, it is important that our model performs well on genes that were not found in the training set. Additionally, since our model uses many gene-based features, it is also important that we do not overfit for genes that are already in the HGMD database. Thus, we performed a separate cross-validation using "gene colocation" where no genes appeared in both the training and testing sets in any fold. This analysis resulted in an AUC of 0.94 for all regulatory variants, 0.96 for the DM subset, 0.93 for the DFP subset, and 0.92 for the FP subset. This suggests that our model will still perform well on genes for which there is no known disease association or functional evidence, and does not overfit in relation to genes already logged in the HGMD database.

Individual Feature Performance

For each feature (or group of features, as is the case with the FANN-GO and expression datasets), the classification performance was tested individually to assess predictive power. Here, we obtained AUCs for each feature or group of features by building the ensemble of decision trees using only that feature subset. The results for individual features are reported in Table 2 with similar results found when using both the disease-causing mutation positive subset (Supp. Table S2) and when using the ±500 bp region negative set (Supp. Table S3). These results indicate that the distance to the TSS is important for identifying variants that play a functional role in gene regulation, a conclusion that supports the findings of the previous research. Also highlighted is the importance of incorporating gene-based features (FANN-GO, expression data, FOP, ENC, etc.) in the analysis, a key finding of our previous study. Furthermore, incorporating newly discovered functional genomic sites from the ENCODE project improved the ability of our method to identify variants that play a functional role in gene regulation. Thus, replication timing, histone peaks, TFBS, DNase sensitivity peaks, and FAIRE peaks were found to be "power features" for identifying functional regulatory variants.

Independent Dataset Validation

To assess the model's performance on data not used in training, the genomes of healthy individuals from the 1000 Genomes Project and the Wellderly Project were scored with the RSVP predictor in order to provide a comparison with an independent set of variants that cause regulatory abnormalities (Fig. 3). Labeled "HGMD

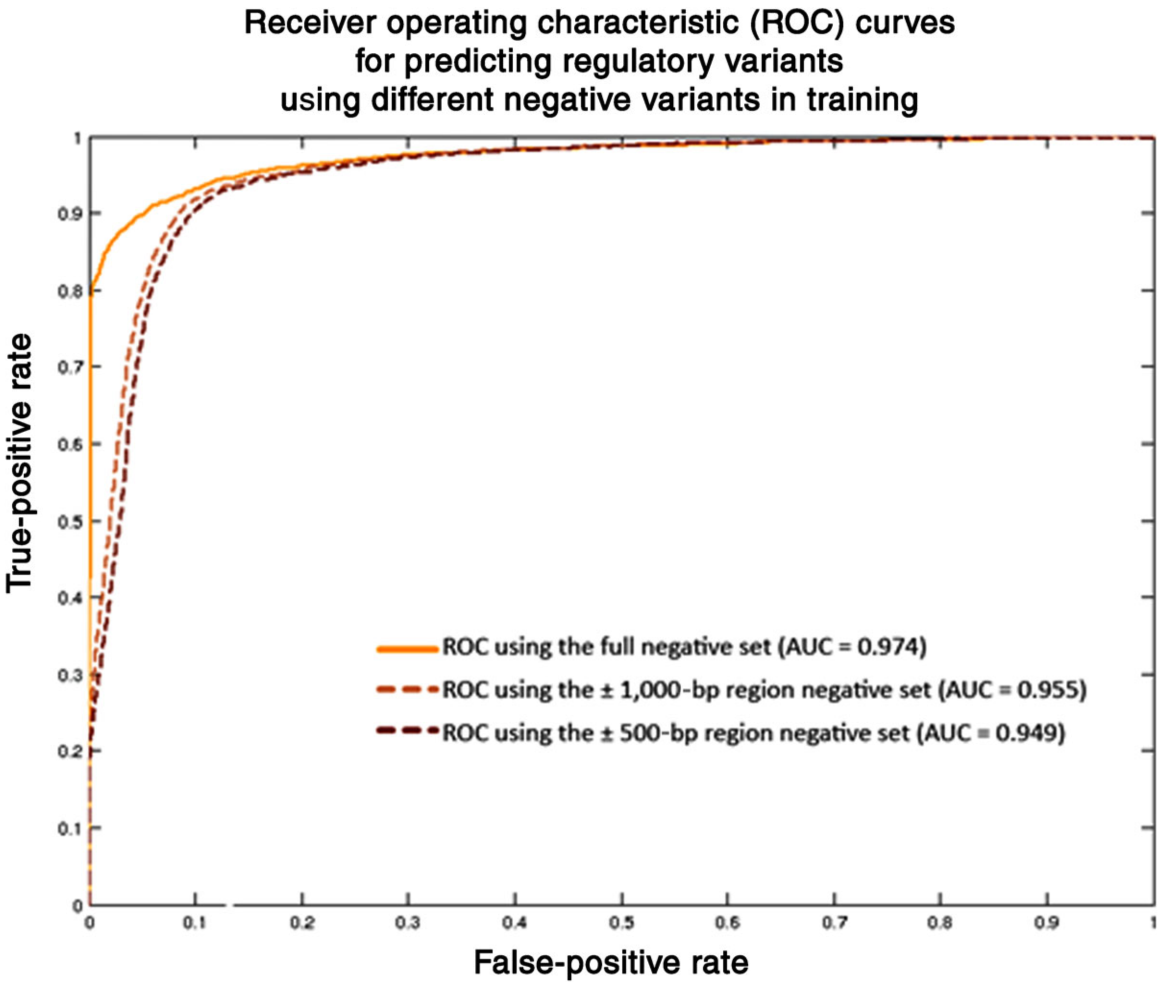

Figure 1. Comparison of cross-validation classification performance for three different models, each trained with a different negative set. The model trained using randomly sampled variants from dbSNP with a minor allele frequency of ≥5% (full negative set), had an area under the receiver operator characteristic curve (AUC) of 0.974. Restricting the negative variants to those within ±1,000 and ±500 bp regions flanking any positive variant resulted in models that achieved AUCs of 0.955 and 0.949, respectively, in cross-validation.

Update" in Figure 3A are scores for 199 new variants in 169 genes that were recently added to the HGMD regulatory variant subset of the database, and hence were not originally used to train the classifier. The second dataset, termed "Wellderly Project Germline Variants" (Fig. 3B), comprised 18,157 variants from the same 169 genes as the HGMD update set found in the genomes of individuals from the Wellderly Project. These variants are assumed to be depleted in deleterious variants as compared with other populations. Finally, the third dataset, labeled "1000 Genomes Germline Variants" (Fig. 3C), comprised 21,203 variants that are present in the 1000 Genomes Project for these 169 genes. We found that variants from healthy individuals in the Wellderly Project and the 1000 Genomes Project exhibit a significant difference in RSVP scores as compared with variants in the independent HGMD dataset (*t*-test *P* value: $<1\times10^{-15}$ for both comparisons).

Somatic Variants Associated with Regulatory Abnormalities in Breast Invasive Carcinoma Patients

In Figure 4, RSVP scores are compared between somatic variants found in tumor samples from TCGA breast invasive carcinoma patients, somatic variants in regulatory/intergenic regions near genes with known cancer relevance from the Cancer Gene Census (CGC) database (Fig. 4A), somatic variants in regulatory/intergenic regions near genes that were not in the CGC (Fig. 4B), and germline variants in regulatory/intergenic regions from the 1000 Genomes Project (Fig. 4C). Overall, in comparison to germline variants located in regulatory/intergenic regions from the 1000 Genomes Project, somatic variants in regulatory/intergenic regions from TCGA breast invasive carcinoma patients tend to have higher RSVP scores (*t*-test *P* value: 1.9×10^{-12}) and similar results were found using variants from the Wellderly Project (*t*-test *P* value: 1.3×10^{-11}). Moreover, we find that somatic variants near genes with known cancer relevance from the CGC tend to have higher RSVP scores (*t*-test *P* value: 2.3×10^{-2}) than genes with no known cancer relevance.

Discussion

In this paper, we have evaluated the performance achieved by our RSVP by incorporating the wealth of information generated in recent years by the ENCODE project. In comparison to our previous study that did not include features from the ENCODE project [Zhao

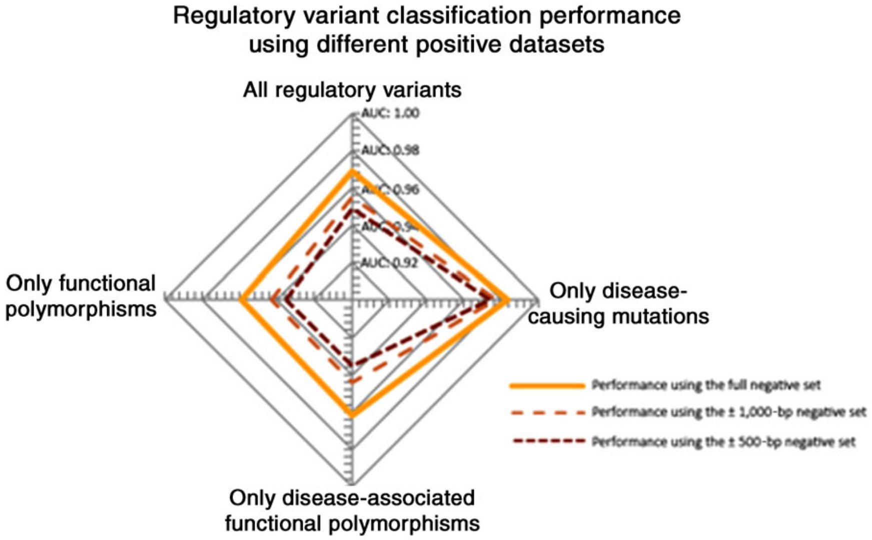

Figure 2. Regulatory variant classification performance using different mutation classes in the positive datasets during cross-validation. The area under the receiver operator characteristic curve (AUC) was calculated during cross-validation for all regulatory variants (AUCs of 0.969, 0.955, and 0.949 for the full, ±1,000, and ±50 0bp negative sets) as well as subsets containing only disease-causing mutations (AUCs of 0.983, 0.977, and 0.974 for the full, ±1,000, and ±500 bp negative sets), only disease-associated functional polymorphisms (AUCs of 0.962, 0.944, and 0.935 for the full, ±1,000, and ±500 bp negative sets), and only functional polymorphisms with no disease association (AUCs of 0.959, 0.943, and 0.935 for the full, ±1,000, and ±500 bp negative sets).

Table 1. RSVP Performance at Different False-Positive Rates as Measured by F-Measure and Accuracy

False-positive rate	Threshold	F-measure	Accuracy
1%	0.90	0.69	0.98
5%	0.57	0.46	0.95
10%	0.40	0.32	0.90

et al., 2011], RSVP's performance (measured in AUC) improved from 0.90 to 0.97 during cross-validation when trained using the full HGMD regulatory variant dataset. Additionally, we demonstrate that our tool also performs well on all types of functional regulatory variant recorded in the HGMD database, with "disease-causing mutations" (DMs) achieving an AUC of 0.98, "disease-associated polymorphisms" with additional supporting functional evidence (DFPs) achieving an AUC of 0.96, and "functional polymorphisms" (FPs) achieving an AUC of 0.96.

To benchmark our tool, we compared RSVP's results with a recently published tool, Genome Wide Annotation of Variants (GWAVA), which also employs a machine learning approach that incorporates data from the ENCODE project as features to predict regulatory variants, and to a traditional variant predictor used to classify coding variants, the CADD tool. We found that our overall performance is comparable to GWAVA, which also achieved an AUC of 0.97, and that both tools perform better on regulatory variants than a traditional tool, CADD, which only achieved an AUC of 0.84. RSVP and GWAVA also outperform a predictor trained with similar ENCODE-based features but which utilized the entire HGMD database, DeepSEA (AUC of 0.81), thereby underscoring the utility of a predictor trained specifically for variants that cause regulatory abnormalities. However, as highlighted in the publication describing the GWAVA tool, it is important to evaluate the predictor's

Table 2. Performance of Individual Features Ranked by the Area Under the Receiver Operator Characteristic Curve (AUC) Achieved by Using Each Feature or Set of Features (in the Case of FANN-GO and Expression Level Features) Individually

Feature name	Type	AUC	Feature name (cont.)	Type	AUC
Distance to the nearest transcription start site	SNV based	0.945	PPINT complexity	Gene based	0.715
FANN-GO category	Gene based	0.881	Flanking nucleotide GC content	SNV based	0.710
Gene expression across 79 tissues	Gene based	0.856	FAIRE peak	SNV based	0.703
Replication timing	SNV based	0.820	Exonic splicing enhancers	SNV based	0.681
Frequency of optimal codons (FOP)	Gene based	0.805	Exonic splicing silencers	SNV based	0.681
Mean of gene expression sets	Gene based	0.800	CpG island	SNV based	0.623
Maximum of expression sets	Gene based	0.800	Conserved noncoding region	SNV based	0.580
Covariance of expression sets	Gene based	0.798	Insulator	SNV based	0.563
Standard deviation of expression sets	Gene based	0.794	DNAse/nuclease region	SNV based	0.533
Histone peak	SNV based	0.786	Methylation peak	SNV based	0.518
Minimum of expression sets	Gene based	0.771	Enhancer	SNV based	0.501
Effective number of codons (ENC)	Gene based	0.757	Phastcons conserved region	SNV based	0.500
TFBS	SNV based	0.752	RNA pol II enriched site	SNV based	0.500
DNase peak	SNV based	0.746			

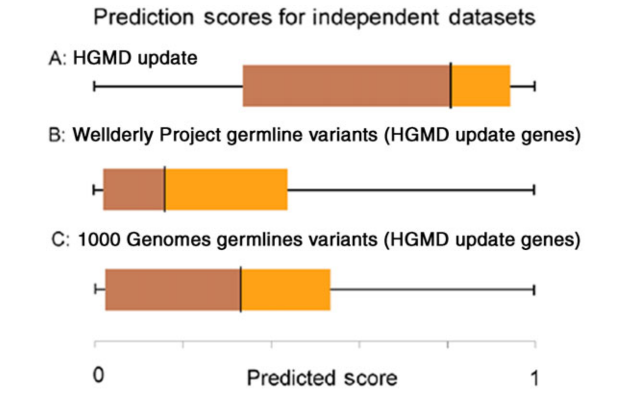

Figure 3. Distribution of RSVP prediction scores for regulatory variants from three unseen and independent variant datasets. **A**: One-hundred ninety-nine variants in 169 genes from the HGMD, which were added to HGMD after January 30, 2014 (HGMD update). **B**: 18,157 variants from patients sequenced in the Wellderly Project from the same 169 genes as the HGMD update. **C**: 21,203 variants from patients sequenced in the 1,000 Genomes Project, also from the same 169 genes as the HGMD Update.

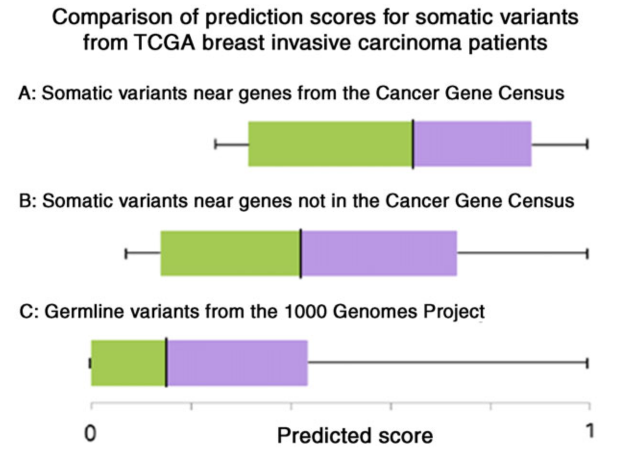

Figure 4. Comparison of RSVP prediction score distributions for somatic variants from TCGA breast invasive carcinoma patients. **A**: RSVP scores for regulatory/intergenic somatic variants near genes with known cancer relevance from the Cancer Gene Census. **B**: RSVP scores for regulatory/intergenic somatic variants near genes that were not in the Cancer Gene Census. **C**: RSVP scores for regulatory/intergenic germline variants from the 1,000 genomes project.

performance on variants displaying similar distances to the TSS as those in the positive set, since this could be considered to be a "power feature" as a single-feature classifier, achieving an AUC of 0.95. Thus, to determine how the GWAVA tool performs on these variants, we performed a separate analysis using a subset of the negative variants during cross-validation that comprised all negative variants within a 1,000-bp window around any HGMD variant. Here, the GWAVA tool achieved an AUC of only 0.71, whereas RSVP achieved an AUC of 0.96. In addition, our method performs well when only considering those negative variants within a 500-bp window around any HGMD variant, achieving an AUC of 0.95. This disparity likely stems from the choice of features used in RSVP since the GWAVA tool does not include any gene-based features, which were determined to be highly predictive features in our previous study [Zhao et al., 2011], as well as in our current study. Intuitively, gene-based features such as predicted GO terms and expression level in various tissues should be key to predicting the effect of variants on the regulation of genes in relation to disease. Indeed, these were found to be highly predictive features with individually assessed AUCs of 0.88 and 0.86, respectively. Not surprisingly, as seen in Supp. Table S4, many of the most predictive FANN-GO features are involved with regulatory activity (e.g., response to stimulus, phosphorylation, transcription factor activity, etc.). Likewise, all other gene-based features also displayed good discriminatory power with AUCs of 0.81 for FOC, 0.76 for EOC, and 0.72 for PPINT complexity.

Since our method employed multiple gene-based features, one possible source of bias could be introduced if our predictor overfits for genes that are already in the HGMD database. To assess this potential bias, we performed a separate cross-validation using "gene colocation" where no genes appeared in both the training and testing sets in any fold. Resulting in an AUC of 0.94 for all regulatory variants, 0.96 for the DM subset, 0.93 for the DFP subset, and 0.92 for the FP subset (Fig. 2), we determined that our model is unaffected by this bias and that it will perform well for new genes for which there are no known disease associations or functional evidence. Overfitting can also become a concern when using large feature vectors such as the FANN-GO feature set. Thus, in addition to using 10-fold cross-validation, we also address this problem in our random forest algorithm, where we randomly select between five and 1,000 features for each tree. This ensures that each tree will be trained using different features and that no tree will have more features than variants in the positive dataset. This also ensures that power features such as distance to the TSS do not dominate the entire ensemble of trees.

The selection of a negative training dataset is crucial for assessing performance and comparing models. Thus, in this study, we employed several negative datasets that were similar to the ones used in our previous study [Zhao et. al., 2011] and to the negative datasets used in the GWAVA tool. Like GWAVA, our "unmatched" dataset was randomly sampled from variants that are not known to cause regulatory abnormalities. Although our performance was unchanged from an AUC of 0.97 for three random samples of variants from dbSNP (data available upon request), we did notice a marked change when analyzing only negative variants within a 1,000- or 500-bp window around any positive variant. Indeed, we find that the most predictive feature, distance to the nearest TSS, achieves an AUC of 0.95 in cross-validation alone for the unmatched dataset, although the AUC decreases to 0.91 and 0.86 for the 1,000- and 500-bp negative datasets, respectively. Thus, it is vital for any comparison between methods to consider the unequal distribution of distances to the TSS in the negative datasets chosen.

To assess RSVP's performance on an independent dataset, we scored and compared variants from a recent update to the HGMD regulatory variant database that were not used to train the model to and germline variants from the Wellderly and 1000 Genomes Projects. As expected, variants from the independent HGMD update dataset that were not used to train the model were scored significantly higher than variants from the same genes found in the Wellderly and 1000 Genomes Projects, indicating that RSVP has the power to detect novel variants that will cause regulatory abnormalities. Additionally, to assess the potential application of the RSVP tool to identify somatic variants with the potential to drive tumor progression, we performed an analysis of RSVP scores for a dataset of somatic variants from the tumors of patients with breast invasive carcinoma from The Cancer Genome Atlas project. We found that, overall, somatic variants in regulatory/intergenic regions are predicted by RSVP to be more likely to be associated with regulatory abnormalities than similar variants from the Wellderly and 1000

Genomes Projects. Moreover, we find that somatic variants located near genes known to play a role in cancer from the Cancer Gene Census tend to have higher RSVP scores than somatic variants near genes with no known cancer relevance. This indicates that the RSVP tool could also be used to identify variants that likely play a role in the regulation of pathways involved in tumor progression.

References

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56–65.

Andersen MC, Engstrom PG, Lithwick S, Arenillas D, Eriksson P, Lenhard B, Wasserman WW, Odeberg J. 2008. In silico detection of sequence variations modifying transcriptional regulation. PLoS Comput Biol 4(1):e5.

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. Cell 129(4):823–837.

Bell AC, West AG, Felsenfeld G. 2001. Insulators and boundaries: versatile regulatory elements in the eukaryotic genome. Science 291(5503):447–450.

Buckland PR. 2006. The importance and identification of regulatory polymorphisms and their mechanisms of action. Biochim Biophys Acta 1762(1):17–28.

Buckland PR, Hoogendoorn B, Guy CA, Coleman SL, Smith SK, Buxbaum JD, Haroutunian V, O'Donovan MC. 2004. A high proportion of polymorphisms in the promoters of brain expressed genes influences transcriptional activity. Biochim Biophys Acta 1690(3):238–249.

Campino S, Forton J, Raj S, Mohr B, Auburn S, Fry A, Mangano VD, Vandiedonck C, Richardson A, Rockett K, Clark TG, Kwiatkowski DP. 2008. Validating discovered Cis-acting regulatory genetic variants: application of an allele specific expression approach to HapMap populations. PLoS One 3(12):e4105.

Chorley BN, Wang X, Campbell MR, Pittman GS, Noureddine MA, Bell DA. 2008. Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: current and developing technologies. Mutat Res 659(1-2):147–157.

Clark WT, Radivojac P. 2011. Analysis of protein function and its prediction from amino acid sequence. Proteins 79(7):2086–2096.

Collins FS, Barker AD. 2007. Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. Sci Am 296(3):50–57.

Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, Margulies EH, Chen Y, Bernat JA, Ginsburg D, Zhou D, Luo S, et al. 2006. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing MPSS). Genome Res 16(1):123–131.

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. 2004. A census of human cancer genes. Nat Rev Cancer 4(3):177–183.

Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, Ch'ang L-Y, Huang W, Liu B, Shen Y. 2003. The international HapMap project. Nature 426(6968):789–796.

Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, Dunham I, Elnitski LL, et al. 2014. Defining functional DNA elements in the human genome. Proc Natl Acad Sci U S A 111(17):6131–6138.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46(3):310–315.

Lapidot M, Mizrahi-Man O, Pilpel Y. 2008. Functional characterization of variations on regulatory motifs. PLoS Genet 4(3):e1000018.

Liu X, Jian X, Boerwinkle E. 2011. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat 32(8):894–899.

Montgomery SB, Griffith OL, Schuetz JM, Brooks-Wilson A, Jones SJ. 2007. A survey of genomic properties for the detection of regulatory polymorphisms. PLoS Comput Biol 3(6):e106.

Mooney S. 2005. Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. Brief Bioinform 6(1):44–56.

Mottagui-Tabar S, Faghihi MA, Mizuno Y, Engstrom PG, Lenhard B, Wasserman WW, Wahlestedt C. 2005. Identification of functional SNPs in the 5-prime flanking sequences of human genes. BMC Genomics 6:18.

Pampin S, Rodriguez-Rey JC. 2007. Functional analysis of regulatory single-nucleotide polymorphisms. Curr Opin Lipidol 18(2):194–198.

Pastinen T, Hudson TJ. 2004. Cis-acting regulatory variation in the human genome. Science 306(5696):647–650.

Pennacchio LA, Loots GG, Nobrega MA, Ovcharenko I. 2007. Predicting tissue-specific enhancers in the human genome. Genome Res 17(2):201–211.

Peterson TA, Doughty E, Kann MG. 2013. Towards precision medicine: advances in computational approaches for the analysis of human variants. J Mol Biol 425(21):4047–4063.

Ponomarenko JV, Orlova GV, Merkulova TI, Gorshkova EV, Fokin ON, Vasiliev GV, Frolov AS, Ponomarenko MP. 2002. rSNP_Guide: an integrated database-tools system for studying SNPs and site-directed mutations in transcription factor binding sites. Hum Mutat 20(4):239–248.

Prokunina L, Alarcon-Riquelme ME. 2004. Regulatory SNPs in complex diseases: their identification and functional validation. Expert Rev Mol Med 6(10):1–15.

Ritchie GR, Dunham I, Zeggini E, Flicek P. 2014. Functional annotation of noncoding sequence variants. Nat Methods 11(3):294–296.

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409(6822):928–933.

Savinkova LK, Ponomarenko MP, Ponomarenko PM, Drachkova IA, Lysova MV, Arshinova TV, Kolcha NA. 2009. TATA box polymorphisms in human gene promoters and associated hereditary pathologies. Biochemistry Mosc 74(2):117–129.

Sharp PM, Li WH. 1986. Codon usage in regulatory genes in Escherichia coli does not reflect selection for 'rare' codons. Nucleic Acids Res 14(19):7737–7749.

Stenson PD, Ball EV, Mort M, Phillips AD, Shaw K, Cooper DN. 2012. The Human Gene Mutation Database HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. Curr Protoc Bioinformatics **Chapter 1**: Unit1 13.

Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. 2014. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet 133(1):1–9.

Sterne-Weiler T, Howard J, Mort M, Cooper DN, Sanford JR. 2011. Loss of exon identity is a common mechanism of human inherited disease. Genome Res 21(10):1563–1571.

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A 101(16):6062–6067.

Torkamani A, Schork NJ. 2008. Predicting functional regulatory polymorphisms. Bioinformatics 24(16):1787–1792.

Wang H, Zhang Y, Cheng Y, Zhou Y, King DC, Taylor J, Chiaromonte F, Kasturi J, Petrykowska H, Gibb B, Dorman C, Miller W, et al. 2006. Experimental validation of predicted mammalian erythroid cis-regulatory modules. Genome Res 16(12):1480–1492.

Ward LD, Kellis M. 2012. Interpreting noncoding genetic variation in complex traits and human disease. Nat Biotech 30(11):1095–1106.

Zhao Y, Clark WT, Mort M, Cooper DN, Radivojac P, Mooney SD. 2011. Prediction of functional regulatory SNPs in monogenic and complex disease. Hum Mutat 32(10):1183–1190.

Zhou J, Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods 12(10):931–934.